

**NETWORK ANALYSIS OF DRUG SIDE EFFECTS AND
INDICATIONS**

YOUSOFF EFFENDY BIN MOHD ALI

**FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR**

2018

**NETWORK ANALYSIS OF DRUG SIDE EFFECTS AND
INDICATIONS**

YOUSOFF EFFENDY BIN MOHD ALI

**DISSERTATION SUBMITTED IN FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF MASTER OF
SCIENCE**

**INSTITUTE OF MATHEMATICAL SCIENCES
FACULTY OF SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR**

2018

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: (I.C./Passport No.:)

Registration/Matric No.:

Name of Degree:

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Field of Study:

I do solemnly and sincerely declare that:

- (1) I am the sole author/writer of this Work;
- (2) This work is original;
- (3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
- (4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
- (5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
- (6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature

Date:

Subscribed and solemnly declared before,

Witness's Signature

Date:

Name:

Designation:

NETWORK ANALYSIS OF DRUG SIDE EFFECTS AND INDICATIONS

ABSTRACT

This thesis aims to understand drug relationship based on drug side effect and indication through network analysis. Two drug networks are constructed using the SIDER 4.1 dataset which are (a) drug-side effect network and (b) drug-indication network. These networks have been analyzed using network analysis to describe the element-level and network-level properties. Various measurements in network analysis are used to describe those properties including centralities, HITS, PageRank and Burt's constraint. Based on the node measurements in network analysis, all drugs in the networks were ranked and a few prominent drugs have been identified in both networks. The prominent drugs were used to explain the application of network analysis on finding or predicting potentially new uses of drugs called drug repositioning. Interestingly, some of the prominent drugs were appeared in list of successful drug repositioning and some of the predicted new uses were already appeared in current clinical studies.

Keywords: Network Analysis, Drug Network, Drug Development

ANALISIS RANGKAIAN KESAN SAMPINGAN DAN INDIKASI DADAH

ABSTRAK

Tesis ini bertujuan untuk memahami kaitan antara dadah berdasarkan kesan sampingan dan indikasi dadah. Dua rangkaian dadah telah dibina menggunakan dataset daripada SIDER 4.1 iaitu (a) rangkaian dadah-kesan sampingan dan (b) rangkaian dadah-indikasi. Kedua-dua rangkaian ini telah dianalisa menggunakan analisis rangkaian untuk menerangkan dua sifat-sifat rangkaian iaitu di tahap elemen dan rangkaian. Pelbagai ukuran di dalam analisis rangkaian telah digunakan untuk menerangkan sifat-sifat ini termasuklah *centralities*, HITS, PageRank, *constraint score* dan lain-lain. Berdasarkan ukuran nod di dalam analisis rangkaian, kesemua dadah di dalam rangkaian telah disenaraikan dan segelintir dadah yang menonjol telah dikenalpasti di kedua-dua rangkaian. Dadah-dadah yang menonjol itu digunakan untuk menerangkan aplikasi analisis rangkaian dadah dalam mencari atau meramal kegunaan baru yang berpotensi bagi dadah dipanggil reposisi dadah. Yang menariknya, ada diantara dadah yang menonjol tersenarai di dalam senarai dadah yang berjaya dalam reposisi dadah dan ada sebahagian daripada kegunaan baru yang diramal turut muncul dalam kajian klinikal yang terkini.

Kata Kunci: Analisis Rangkaian, Rangkaian Dadah, Pembangunan Dadah

ACKNOWLEDGEMENTS

I am grateful to God for everything that happened in these past few years. I never thought that I could do my study while having multiple commitments.

To my family, wife and friends, thanks for every support that you gave me. Splitting my life into study, work, family, friends and personal interests were never easy during my study. I hope that I can spend enough time with everyone of you.

To my supervisors, Prof. Dr. Kurunathan A/I Ratnavelu and Dr. Kwa Kiam Heong, thanks for the knowledge, time and opportunities given to me during my study. Thank you for the guidance on the presentation, research paper and this thesis. I always thought thesis is like a book.

To my colleague, thanks for the support in the office. Thank you for understanding my life as a student and thank you for coming to my wedding. My very special thanks for the flexi hours, it allows me to split my life in a efficient way.

To my local IT communities, thanks for the knowledge that you share with everyone. Sometimes I wonder how do you guys manage to organize events for free.

To Lim Lian Tze, thank you for the thesis template that you made for a lot of local universities. Her templates is available here <http://liantze.penguinattack.org/>.

Finally, thanks to MyBrain by Ministry of Higher Education Malaysia which helps fund my tuition fees since 2016. I hope this funding programme will continue and make a better Malaysia in the future.

TABLE OF CONTENTS

Abstract	iii
Abstrak	iv
Acknowledgements	v
Table of Contents	vi
List of Figures	viii
List of Tables	xi
List of Symbols and Abbreviations	xiii
List of Appendices	xiv
CHAPTER 1: INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	4
1.3 Objectives	5
1.4 Outline and Structure of Thesis	5
CHAPTER 2: LITERATURE REVIEW	7
2.1 Network Analysis	7
2.2 Drug repositioning using network approach	18
CHAPTER 3: METHODOLOGY	20
3.1 Data	20
3.2 Analyzing drug networks	23
3.2.1 Forming network dataset	24

3.2.2	Constructing drug networks	28
3.3	Drug repositioning using network analysis.....	29
3.3.1	Choosing drugs.....	32
3.3.2	Choosing pairs.....	33
3.3.3	Prediction.....	34
3.3.4	Validation	36

CHAPTER 4: RESULTS AND DISCUSSION 39

4.1	Data analysis on SIDER dataset.....	39
4.2	Network analysis on drug networks	44
4.3	Drug repositioning on top drugs	50

CHAPTER 5: CONCLUSION 56

References	59
List of Publications and Papers Presented	63
Appendices.....	65

LIST OF FIGURES

Figure 1.1: Construction of a simple network. A circle is a convention to represent a node and a line between two nodes represents an edge.	2
Figure 1.2: Example of networks	2
(a) Krackhardt Kite	2
(b) A centralized network.....	2
Figure 1.3: Another network example. Group of nodes within red square are called clusters. Largest group of nodes is called the giant component of the network. A node which does not form any group is called an isolated node.	3
Figure 2.1: Krackhardt kite.	9
Figure 2.2: Krackhardt kite colored by degree score.	10
Figure 2.3: Krackhardt kite colored by closeness score.....	11
Figure 2.4: Krackhardt kite colored by betweenness score.....	12
Figure 2.5: Krackhardt kite colored by HITS scores.	15
Figure 2.6: Krackhardt kite colored by PageRank scores.	16
Figure 2.7: Krackhardt kite colored by Burt's constraint score.	18
Figure 2.8: A <i>de novo</i> drug development may took between 10 to 17 years while drug repositioning might only took between 3 to 12 years.	18
Figure 3.1: Data flow diagram. Green boxes represent data sources, blue boxes represent milestones and line labels represent software packages used. ...	20
Figure 3.2: A typical dataset of meddra_all_se.tsv. The first and the second columns contain the STITCH flat and stereo compound IDs respectively. The third column contains the UMLS concept IDs which represent side effects for this dataset. The fourth column contains MedDRA concept types where LLT and PT are acronyms for lowest level term and preferred term respectively. The fifth column contains UMLS concept IDs for MedDRA term. The last column contains the side effect.	21

Figure 3.3: A typical dataset of meddra_freq.tsv. The first and the second columns contain STITCH flat and stereo compound IDs respectively. The third column contains the UMLS concept IDs which represent side effects for this dataset. The fourth column will show placebo if the side effect information comes from the placebo administration, an empty value otherwise. The fifth column describes the frequencies of the side effects. The sixth and seventh columns describe the lower and upperbound of the frequency. The last three columns describe the MedDRA information similar to the ones in meddra_all_se.tsv.	22
Figure 3.4: A typical dataset of meddra_all_indications.tsv. The first column contains STITCH flat compound IDs. The second column contains UMLS concept IDs which represent indications for this dataset. The third column contains the methods of detection. The fourth column contains the MedDRA concept names. The last three columns describe the MedDRA information similar to those in meddra_all_se.tsv.	22
Figure 3.5: Flowchart to obtain network dataset.	25
Figure 3.6: Example on creating an edge in the drug-side effect network. The first step is to list all side effects for each drugs and set the edge weight equivalent to the number of side effect similarities.	27
Figure 3.7: Transforming adjacency table into adjacency list.	28
Figure 3.8: Flowchart for prediction of a single drug.	30
Figure 3.9: Our extended hypothesis based on the hypothesis by Duran-Frigola and Aloy. (a) Suppose two drugs share similar side effects, then we may infer indications between these two drugs - the one that we already knew and the one that is yet to be discovered. (b) Similarly, we may also infer side effects based on two drugs that share similar indications.	31
(a) First implication.	31
(b) Second implication	31
Figure 3.10: Example on guilt-by-association for potentially new drug indications. ...	35
Figure 4.1: Giant components of both drug networks. Each grey dot represents a drug and each dark grey line represents an edge between two drugs.	45
(a) Drug-side effect network	45
(b) Drug-indication network.	45
Figure 4.2: Visualization for Table 4.15.	46
Figure 4.3: Visualization for Table 4.17.	49

Figure A.1: Example of ClinicalTrials.gov search result with everolimus as its intervention parameter.....	69
--	----

University of Malaya

LIST OF TABLES

Table 1.1: Example of nodes and edges in different networks.....	3
Table 1.2: Example of directed and undirected network.....	4
Table 3.1: Council for International Organizations of Medical Science (CIOMS) frequency convention.	24
Table 3.2: Examples of extracted datasets.	26
(a) Extracted side effect dataset	26
(b) Extracted indication dataset	26
Table 3.3: Node Measurements.....	32
Table 3.4: Sample of clinical studies.....	38
Table 4.1: Information on SIDER 4.1 dataset	39
Table 4.2: Summary on SIDER 4.1 dataset	40
Table 4.3: Summary of VCC side effect dataset.	40
(a) Drug.....	40
(b) Side Effect	40
Table 4.4: Top five drugs by number of side effects	41
Table 4.5: Top five side effects by number of drugs	41
Table 4.6: Summary of CID100005064 (Ribavirin)	41
Table 4.7: Summary of C0027497 (Nausea).....	41
Table 4.8: Summary of indication dataset.....	42
(a) Drug.....	42
(b) Indication.....	42
Table 4.9: Top five drugs by number of indications.....	42
Table 4.10: Top five indications by number of drugs.....	42
Table 4.11: Summary of CID100003003 (Dexamethasone).....	43
Table 4.12: Summary of C0009450 (Communicable Diseases)	43

Table 4.13: Network clusters.....	44
Table 4.14: Summary on network-level properties.	44
Table 4.15: Summary of node measurements for drug side effect network.....	46
Table 4.16: Top ten scorers in drug-side effect network by node measurement. The scores measured are degree centrality C_D , betweenness centrality C_B , closeness centrality C_C , HITS score V_H , inverse of Burt's constraint score V_B ($m = 10^{-3}$) and PageRank score V_P ($m = 10^{-3}$).	48
Table 4.17: Summary of node measurements for drug indication network.	49
Table 4.18: Top ten scorers in drug-indication network by node measurement. The scores measured are degree centrality C_D , betweenness centrality C_B ($m = 10^3$), closeness centrality C_C , HITS score V_H , inverse of Burt's constraint score V_B ($m = 10^{-3}$) and PageRank score V_P ($m = 10^{-3}$).	51
Table 4.19: Drugs in the drug-side effect network that appear at least once in the top ten lists of all six node scores.....	52
Table 4.20: Indications of five prominent drugs predicted based on the drug-side effect network.....	52
Table 4.21: Top five predicted indications based on first neighbours for everolimus. Concept Unique Identifier (CUI) is the UMLS concept ID from SIDER dataset. Frequency is the number of redundancies based on the prediction list from Step 4. Neighbour frequency is the frequency divided by the total number of neighbours for Everolimus in the drug-side effect network. Concept name is the given indication name based on CUI. Number of clinical studies is the number of records found in ClinicalTrials.gov between drug and concept name (ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US), n.d.)	53
Table 4.22: Drugs in the drug-indication network that appear at least once in the top ten lists of all six node scores.....	54
Table 4.23: Side effects of five prominent drugs predicted based on the drug-indication network.....	54
Table A.1: List of importance Python packages for this study	65
Table A.2: Column description for meddra_all_se.tsv	65
Table A.3: Column description for meddra_freq.tsv	66
Table A.4: Column description for meddra_all_indications.tsv	66

LIST OF SYMBOLS AND ABBREVIATIONS

API	: Application program interface.
ATC	: Anatomical Therapeutic Chemical.
CID	: Compound identification number.
CIOMS	: Council for International Organizations of Medical Science.
CLI	: Command-Line Interface.
CMap	: Connectivity Map.
CSV	: Comma-Separated Values.
CUI	: Concept Unique Identifier.
DRoSEf	: Drug Repositioning based on the Side-Effectome.
FDA	: Food and Drug Administration.
GBA	: Guilt by Association.
HITS	: Hyperlink-Induced Topic Search.
HTML	: HyperText Markup Language.
HTS	: High Throughput Screening.
IDE	: Integrated Development Environment.
IUPAC	: International Union of Pure and Applied Chemistry.
MANTRA	: Mode of Action by Network Analysis.
MedDRA	: Medical Dictionary for Regulatory Activities.
MOA	: Mechanism of action.
NCBI	: National Center for Biotechnology Information.
NIH	: National Institutes of Health.
NLM	: National Library of Medicine.
PUG	: Power User Gateway.
REST	: Representational State Transfer.
SIDER	: Side Effect Resource.
SOAP	: Simple Object Access Protocol.
STITCH	: Search Tool for Interacting Chemicals.
TSV	: Tab-Separated Values.
UMLS	: Unified Medical Language System.
UTS	: UMLS Terminology Services.
VCC	: Very common or common.
WHO	: World Health Organization.
XML	: Extensible Markup Language.

LIST OF APPENDICES

Appendix A: Methods.....	65
--------------------------	----

University of Malaya

CHAPTER 1: INTRODUCTION

In this chapter, we will describe the background and the motivation of this study. We will also describe the problem statement and the objectives, followed by the structure of this thesis.

1.1 Background

Network analysis is a branch of graph theory which aims at describing quantitative properties of networks of interconnected entities by means of mathematical tools. The history of graph theory itself can be traced back to 1736 where Leonhard Euler attempted to solve the Königsberg bridge problem (Biggs et al., 1986). The problem concerns on finding ways to not crossing the bridges twice from the seven bridges of Königsberg. The most fundamental concept in graph theory is to define the node and edge for the graph. By definition, a node represents an individual object that exist in the graph. While for edge, it is simply a connection between two nodes. In Königsberg bridge problem, Leonhard Euler defined each land as a node and each bridge connecting the lands as an edge. From there, Euler has created one of the earliest theorem in graph theory in which he also concluded that the problem has no solution (Newman, 2003).

In 1759, Euler had also analyzed a knight's tour puzzle such that a knight on a chessboard only visit each square once. He used graph theory for this puzzle and it has been evolved from a standard 8 x 8 board size to different sizes. Another famous problem that uses graph theory is the travelling salesman problem where a salesman is required to visit n cities and at the end, he needs to return to his original location. Each link between cities have a travel cost and the objective is to find the optimal distance and cost as long as the salesman is able to visit all cities and return to the original location. All of these and many other topological problems that use graph theory have been widely discussed across many

of the graph theory books (Biggs et al., 1986; Bollobás, 2013).

The terms "network" and "graph" can be used interchangeably and in this thesis, we prefer to use the first term rather than the latter. A network can be created in its simplest form when there are two nodes connected by an edge (See Figure 1.1). Examples of network are shown in Figure 1.2.



Figure 1.1: Construction of a simple network. A circle is a convention to represent a node and a line between two nodes represents an edge.

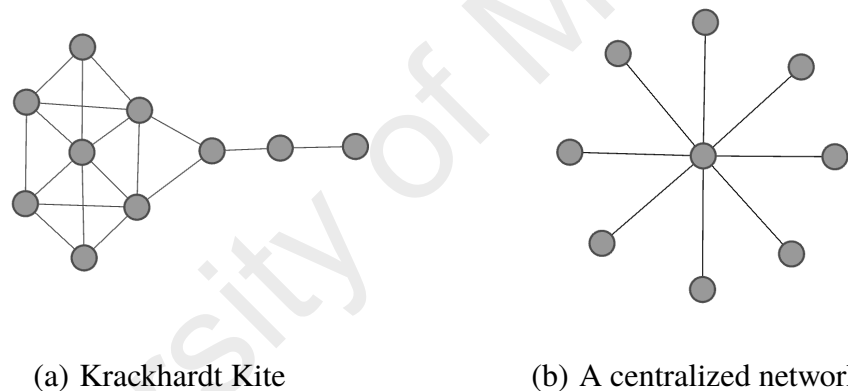


Figure 1.2: Example of networks

Nodes may represent different objects while edges may indicate different relations in different networks. As shown in Table 1.1, each of the combinations has a different application. A network helps us visualize and see things in different perspectives. In a social network, it can be used to detect influential or suspicious individual in the network. In a transportation network, it can be used to detect possible congested roads.

In a larger network, there is a possibility to find nodes that are not connected to all other nodes in the network. We called these nodes as isolated nodes. In a road network where each node represents a place for example, an isolated node means that there are no

Table 1.1: Example of nodes and edges in different networks.

Network	Node (Object)	Edge (Relation)
Social Network	Person	Friend of
Social Network	Organization	Partnership
Transportation Network	Junction	Road
Computer Network	Server	Can connect to
Electrical Network	Electronic Component	Wired to

accessible roads to go to that particular place. It is also possible for nodes in the network to connect few other nodes forming groups of nodes. These groups are called clusters. For any size of cluster, suppose we choose any two nodes A and B in a network, there exists a path between those two nodes in the network. The largest cluster in the network that captures the most nodes is called the giant component of the network (See Figure 1.3).

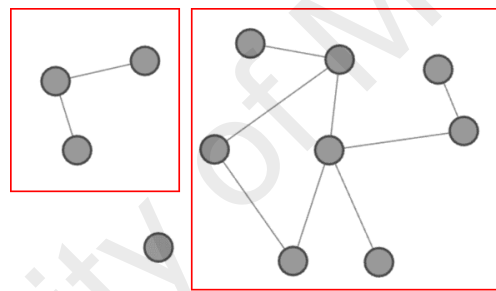
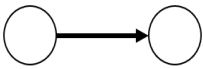
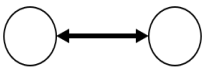
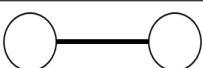


Figure 1.3: Another network example. Group of nodes within red square are called clusters. Largest group of nodes is called the giant component of the network. A node which does not form any group is called an isolated node.

Networks can be further split in two types of networks which are directed and undirected network. Table 1.2 shows example of these networks based on social media concepts. A network in certain situations may also contain self-loops where an edge connects a node to itself. For example, in a citation network where each node represents an author and each edge represents the citation between authors. A node can have a self-loop if an author cites his own publications. A network without self-loops is often called as a simple network.

While graph theory is an abstract field that concerns more on the graph classes, structures and algorithmic solutions, network analysis on the other hand is more focused on problems

Table 1.2: Example of directed and undirected network.

Network	Type	Example
	Directed Network (one-way)	Person A follows Person B on Twitter
	Directed Network (two-way)	Person A and B follow each other on Twitter
	Undirected Network	Person A is a friend of Person B on Facebook

that are modeled by networks and the interactions of nodes (Zweig, 2016). Network analysis methodology has been widely used in social or transportation networks, but only some research has been done on drug networks. By analyzing drug networks, we may visualize how drugs are related to each other and we might discover something new from the relationship as well.

This thesis attempts to create and analyze drug networks. We have chosen drug phenotypic profiles which are drug side effects and indications as the attributes to link drugs in our two drug networks. These profiles may be used to complement existing drug repositioning approaches and might overcome some of the issues in drug development. This will be discussed in the next section.

1.2 Problem Statement

Pharmaceutical industries are facing major productivity issues on drug development. The issues are mainly due to the high cost and time-consuming processes in standard drug development. For a single drug, it may cost between \$500 to more than \$2,000 million and may consume between 10 - 15 years (Adams & Brantner, 2006; DiMasi et al., 2016).

To ease these productivity issues, drug repositioning is considered one of the alternatives for a cost and time-effective drug development. However, incomplete understanding on drugs causes drug repositioning to be highly dependent on luck and does not always work (Wu et al., 2013). Therefore, we propose to analyze drug networks to help understand the

phenotypic profiles of drugs with the hope to make such analysis as a useful part of drug repositioning.

We are interested in these two drug networks i.e., (a) the drug-side effect network and (b) the drug-indication network. The first network is constructed based on side effect similarities where two drugs are connected if they share at least one side effect. The second network is constructed based on indication similarities where two drugs are connected if they share at least one indication. Both networks are undirected and simple networks. Additionally, based on the outcomes of the network analysis, we will predict some potentially new uses of existing drugs.

1.3 Objectives

Based on the problem statement stated in the previous section, we have divided the objectives of this thesis into two :

1. To study drug relationships through network analysis based on drug side effects and indications.
2. To implement a drug repositioning algorithm using drug network properties.

1.4 Outline and Structure of Thesis

This thesis is organized into a few chapters and we will focus more on the methodology and results.

Chapter 2 will describe the literature review for this study. We will also explain some arguments that will be the foundation of our algorithm.

Chapter 3 describes the methodology for this study. We will provide step-by-step explanation for data collections, network analysis, and finally the prediction of potentially new uses of existing drugs.

Chapter 4 describes the findings from the network analysis and a few predictions based on our proposed algorithm.

Chapter 5 concludes the findings of this research and propose what can be done to improve this study.

University of Malaya

CHAPTER 2: LITERATURE REVIEW

In this chapter we provide an overview of network analysis and its application in drug repositioning. We will also explain how we can make use of two drug phenotypic profiles with network analysis for a potential drug repositioning.

2.1 Network Analysis

Network theory has been widely used across many fields of studies with the purpose to understand interconnected objects or parts of a system. As a network increases in size, it will be difficult to visualize and identify prominent nodes in the network. Thus, network analysis is usually used to understand these large networks. The most common use of network analysis is to find nodes that resides in certain positions in the network. There are various quantitative measurements from network analysis that can be used to detect prominent nodes, usually by ranking nodes in each measurements. These measurements usually have different interpretations for describing the positions of the nodes. In general, network analysis can be separated into three levels (Baur et al., 2009; Brandes & Erlebach, 2005):

1. **Element-level** analysis focuses on individual node and edge properties.
2. **Group-level** analysis focuses on specific subsets of nodes
3. **Network-level** analysis focuses on the global properties of the network.

For a more comprehensive description of the application of network analysis, the reader is referred to Wasserman and Faust (1994), Borgatti et al. (2009), Bell and lida (1997) and Ahuja et al. (1993).

In recent years, network analysis has been used to solve even more complex problems in the real world (Kranakis, 2012; Barabási et al., 2010). Zhou et al. have created a

human symptoms–disease network to understand association between diseases in the field of system medicines (Zhou et al., 2014). The analysis of the disease network can be used to identify new disease genes and drug targets. Another interesting application of network analysis is on both network pharmacology and system pharmacology - two areas that analyze drug networks (Hopkins, 2008; Berger & Iyengar, 2009; Danhof, 2016). Nacher and Schwartz used centralities in their drug-therapy network to find drugs with high centralities that might act on multiple targets (Nacher & Schwartz, 2008). DrugComboRanker uses a centrality score by combining betweenness, closeness and also PageRank (Huang et al., 2014). Iorio et al. created Mode of Action by Network Analysis (MANTRA) that analyzes drug network based on transcriptional response similarity (Iorio et al., 2010). All these studies use network analysis for the same purpose i.e., to understand and gain more information from the networks. A better understanding on drugs may help us improve drug repositioning. We will illustrate this further in the next section.

In this thesis, we are interested in 6 node measurements of element-level analysis. These measurements are degree centrality, betweenness centrality, closeness centrality, HITS, PageRank and Burt's Constraint (Freeman, 1978; Burt, 2004; Page et al., 1999). To illustrate these measurements, we will use a network called Krackhardt's Kite graph as an example (See Fig 2.1). A node that achieves high score is colored in red.

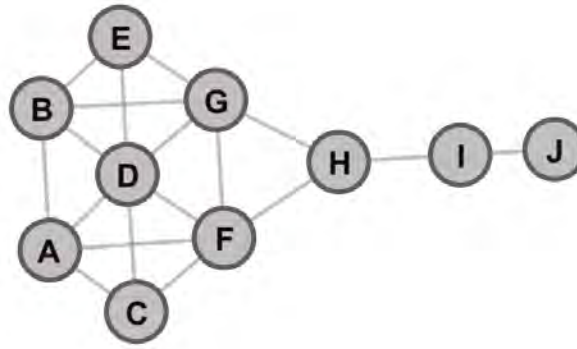


Figure 2.1: Krackhardt kite.

Degree Centrality

One of the simplest node measurements is degree centrality and it is used to identify important nodes in the network. In this measurement, a node is considered important if they are connected to many nodes in the network. For example, a person with a lot of friends can be considered an important person in a local community network. By definition, the degree centrality of a node v is the number of ties that v has with other nodes and is denoted by:

$$C_D(v) = \deg(v) \quad (2.1)$$

where $\deg(v)$ is the degree of vertex v . For a normalized degree centrality score, the following equation will be applied:

$$C_D(v)' = \frac{\deg(v)}{n-1} \quad (2.2)$$

where n is the number of nodes in the network. For example, in Figure 2.1 we can calculate degree score for node D where $C_D(D) = \deg(D) = 6$ and $C_D(D)' = \frac{\deg(D)}{n-1} \approx 6/9$. The

degree distribution is visualised in Figure 2.2.

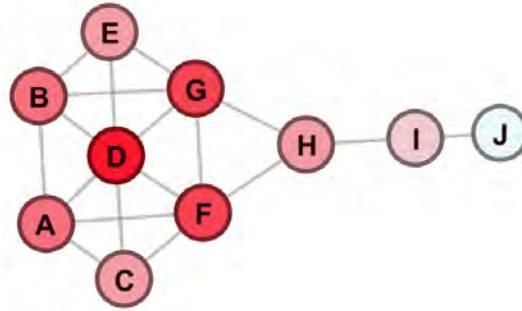


Figure 2.2: Krackhardt kite colored by degree score.

Closeness Centrality

Another node measurement to identify important nodes is closeness centrality. This differs from the previous centrality measure, as it considers a node to be important if the node is closer to all other nodes in the network. For example, in a residential network, a house located at the city centre is considered important since it can visit all other houses at a shorter distance compared to the houses at the suburbs. The closeness centrality of a node v is defined as the reciprocal of the sum of geodesic distances between v to all other nodes in the graph :

$$C_C(v) = 1 / \sum_{i \neq v} d(i, v) \quad (2.3)$$

where each i is a node other than v in the graph and $d(i, v)$ is the distance between i and v .

For a normalized closeness centrality score, the following equation will be applied :

$$C_C(v)' = (n - 1) / \sum_{i \neq v} d(i, v) \quad (2.4)$$

where n is the number of nodes in the network. For example, in Figure 2.1, the closeness score for node D is $C_C(D) = 1/\sum_{i \neq D} d(i, D) = 1/14 \approx 0.071$ and the normalize score is $C_C(D)' = (n - 1)/\sum_{i \neq D} d(i, D) = 9/14 \approx 0.643$. The distribution for this measurement is visualised in Figure 2.3.

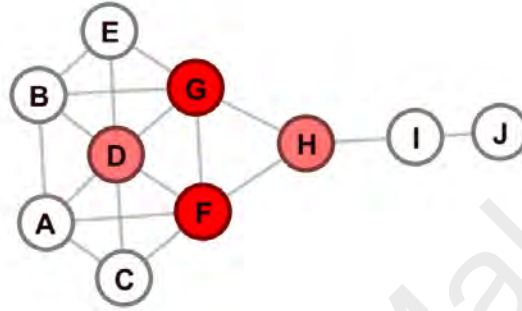


Figure 2.3: Krackhardt kite colored by closeness score.

Betweenness Centrality

Betweenness centrality was introduced by Linton Freeman in 1978 where node is considered important if it serves as a bridge between other nodes in the network (Freeman, 1978). In a road network for example, if there is only one way to go from Point A to Point C which is through Point B, then we consider Point B to be important. Betweenness centrality of a node v is defined by the number of geodesics passing through v :

$$C_B(v) = \sum_{i \neq v \neq j \in V, i \neq j} \frac{\sigma_{ij}(v)}{\sigma_{ij}} \quad (2.5)$$

where $\sigma_{ij}(v)$ is the number of geodesics from node i to j that pass through v and σ_{ij} represents the number of geodesics from i to j . For a normalized betweenness centrality

score of undirected network, the following equation will be applied :

$$C_B(v)' = \frac{2C_B(v)}{n^2 - 3n + 2} \quad (2.6)$$

where n is the number of nodes in the network. For example, the betweenness score for node H is $C_B(H) = \sum_{i \neq H \neq j \in V, i \neq j} \frac{\sigma_{ij}(H)}{\sigma_{ij}} = 14$ since $\sigma_{ij}(H) = \sigma_{ij} = 1$ when $i \in \{A, B, C, D, E, F, G\}$ and $j \in \{I, J\}$. The normalize score is $C_B(H)' = \frac{2C_B(H)}{72} \approx 0.389$. The distribution is visualised in Figure 2.4.

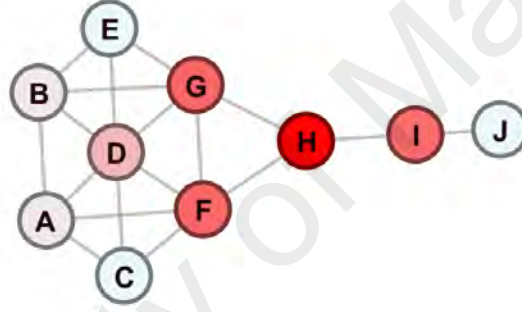


Figure 2.4: Krackhardt kite colored by betweenness score.

HITS

Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm introduced by Jon Kleinberg in 1999 (Kleinberg, 1999). Originally, it was used to rate web pages by calculating two scores iteratively called authority and hub scores. A good hub represents a web page that points to many other web pages, while a good authority represents a page that is linked by many hubs. In general network cases, a node is considered important if it is connected to important nodes. Let h be the n -dimension vector of hub weights and a be

the n -dimensional vector of authority weight, then

$$a_i = \sum_{j:(j,i) \in E} h_j \quad \text{and} \quad h_i = \sum_{j:(i,j) \in E} a_j$$

This equation will be iteratively updated by using the I ("In") operation to update the a -weight (authority score) and O ("Out") operation to update the h -weight (hub score).

The previous equation can be also be represented in matrix form as :

$$I(.) = A^T \quad \text{and} \quad O(.) = A$$

where A is the adjacency matrix. At t th iteration for $t > 0$ we obtain the following expression :

$$a^{(t+1)} = I(O(a^t)) = A^T A a^t \quad \text{and} \quad h^{(t+1)} = I(O(h^t)) = A A^T h^t \quad (2.7)$$

To calculate the score for any node, first we must construct an adjacency matrix for the graph as follows :

$$A = \begin{matrix} & \begin{matrix} D & B & E & G & F & C & A & H & I & J \end{matrix} \\ \begin{matrix} D \\ B \\ E \\ G \\ F \\ C \\ A \\ H \\ I \\ J \end{matrix} & \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{matrix} \quad (2.8)$$

Then, by applying Equation 2.7, we can get these two matrices :

$$a^{(4)} = \begin{bmatrix} 0.4810 \\ 0.3552 \\ 0.2858 \\ 0.3977 \\ 0.3977 \\ 0.2858 \\ 0.3522 \\ 0.1959 \\ 0.0481 \\ 0.0112 \end{bmatrix}, \quad h^{(4)} = \begin{bmatrix} 0.4810 \\ 0.3552 \\ 0.2858 \\ 0.3977 \\ 0.3977 \\ 0.2858 \\ 0.3522 \\ 0.1959 \\ 0.0481 \\ 0.0112 \end{bmatrix}$$

For an undirected graph, both hub and authority scores will converge to the same value.

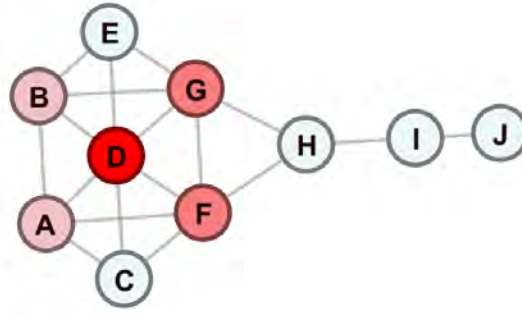


Figure 2.5: Krackhardt kite colored by HITS scores.

PageRank

PageRank is one of the famous link analysis algorithms created by Lawrence Page and Sergey Brin for Google (Page et al., 1999). The PageRank score of node v is a measurement to score v based on random surfer model :

$$PR(v) = (1 - d) + d \sum_i^n \frac{PR(T_i)}{C(T_i)} \quad (2.9)$$

where $PR(v)$ is the PageRank of node v , $PR(T_i)$ is the PageRank of nodes T_i which is linked to node v , $C(T_i)$ is the number of outbound links from node T_i , d is a damping constant between 0 to 1 with a default value of 0.85 and n is the number of nodes in the network. For a normalized PageRank score, the following equation will be applied.

$$PR(v)' = (1 - d)/n + d \sum_i^n \frac{PR(T_i)}{C(T_i)} \quad (2.10)$$

By setting initial $P(v) = 1$ for all $v \in V$ and applying the equation up to t th iteration , we will obtain the following results and visualized in Figure 2.6:

$$PR(D) = 1.471, PR(B) = 1.019, PR(E) = 0.794, PR(G) = 1.289, PR(F) = 1.289,$$

$$PR(C) = 0.794, PR(A) = 1.019, PR(H) = 0.954, PR(I) = 0.857, PR(J) = 0.514$$

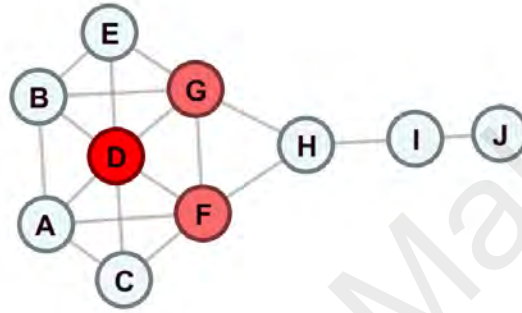


Figure 2.6: Krackhardt kite colored by PageRank scores.

Burt's Constraint

In 2004, Ronald Stuart Burt developed a node measurement to find structural holes, a gap between two nodes which may have complementary sources of information (Burt, 2004). A node with more structural holes may receive non-redundant information, compared to nodes with lesser structural holes. Burt's constraint is defined by

$$C_i = \sum_{j \in V_i \setminus \{i\}} (p_{ij} + \sum_{q \in V_i \setminus \{i,j\}} p_{iq} p_{qj})^2 \quad (2.11)$$

where p_{ij} is the proportional strength between nodes i and j defined by

$$p_{ij} = \frac{a_{ij} + a_{ji}}{\sum_{k \in V_i \setminus \{i\}} (a_{ik} + a_{ki})}$$

and a_{ij} are the elements of adjacency matrix A . Using the adjacency matrix in Equation 2.8, we obtain the matrix for P

$$P = \begin{matrix} & \begin{matrix} D & B & E & G & F & C & A & H & I & J \end{matrix} \\ \begin{matrix} D \\ B \\ E \\ G \\ F \\ C \\ A \\ H \\ I \\ J \end{matrix} & \begin{bmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 1/4 & 0 & 1/4 & 1/4 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & 1/5 & 1/5 & 0 & 0 & 1/5 & 1/5 & 1/5 & 0 & 0 \\ 1/5 & 0 & 0 & 1/5 & 1/5 & 0 & 1/5 & 1/5 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 0 & 1/4 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{matrix}$$

and finally we can compute the constraint score for each node

$$C_i = \begin{matrix} & \begin{matrix} D & B & E & G & F & C & \dots & J \end{matrix} \\ \begin{matrix} D \\ B \\ E \\ G \\ F \\ C \end{matrix} & \begin{bmatrix} 0.4746 & 0.5783 & 0.7059 & 0.4701 & 0.4701 & 0.7059 & \dots & 1.2500 \end{bmatrix} \end{matrix}$$

Nodes that achieve higher scores have more structural holes while lower scores are more constraint in the network. The scores are visualized in Figure 2.7.

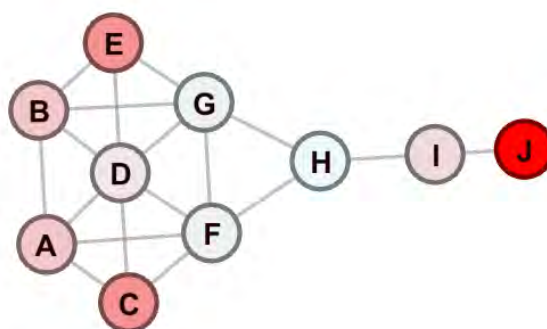


Figure 2.7: Krackhardt kite colored by Burt's constraint score.

2.2 Drug repositioning using network approach

Drug repositioning (also known as drug re-purposing or re-profiling) is the identification and use of existing or failed drugs to treat new indications (Langedijk et al., 2015). Ashburn and Thor have rigorously discussed the advantages of drug repositioning over existing methods for drug development, the main advantage of which is involvement of existing drugs that have passed phases of development for their originally intended indications and provide a faster development process (Ashburn & Thor, 2004) (See Figure 2.8).

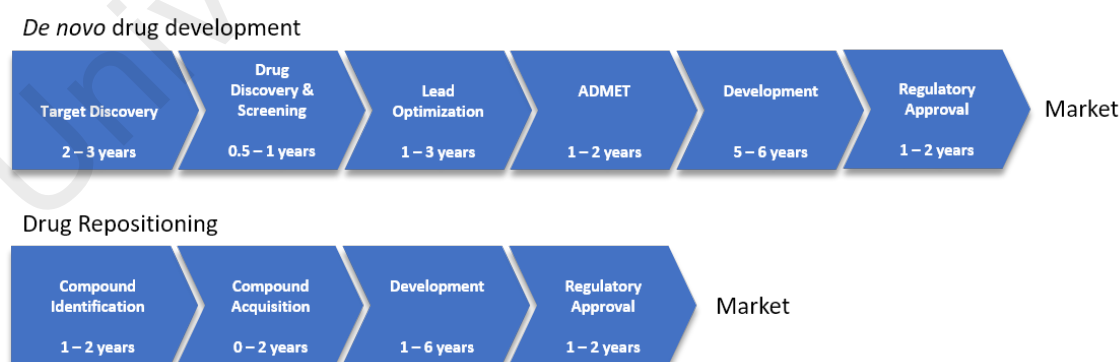


Figure 2.8: A *de novo* drug development may took between 10 to 17 years while drug repositioning might only took between 3 to 12 years.

One of the earliest attempts to enhance drug repositioning using network approach

was proposed by Lamb et al. (2006). They used the Connectivity Map (CMap) approach to provide deeper understanding of the mechanism of action (MOA) of drugs. CMap allows researchers to reveal possible connections among genes, drugs and diseases. Other researchers also suggested several computational ideas including creating a drug-disease relationships or similarities (Dudley et al., 2011), a genome-based method (Lussier & Chen, 2011), a pathway profile-based method (Ye et al., 2012) and an integrative network-based method (Wu et al., 2013). The outcomes were, nevertheless, not always consistent with therapeutic effectiveness in drug development (Ye et al., 2014).

The research mentioned above can be categorized into two computational strategies i.e, the drug-based strategy and disease-based strategy (Dudley et al., 2011). The former strategy looks at the drug perspective and the latter at the disease perspective. These strategies can be further categorized into primary modes of direct inference and indirect inference (Dudley et al., 2011). There are many other recent strategies that can be used to enhance drug repositioning (Kidd et al., 2016; Hodos et al., 2016).

CHAPTER 3: METHODOLOGY

In this chapter, we will describe the research methodology of this thesis and it will be organized into few sections. The first section will describe the data source followed by the network analysis and drug repositioning.

3.1 Data

Our data come from various sources on the Internet. We have a total of 5 data which are Side Effect Resource (SIDER) 4.1, PubChem, ClinicalTrials.gov, Unified Medical Language System (UMLS) and eHealthMe.com. SIDER 4.1 is the main data source used in this study. The data flow is shown in Figure 3.1. More description on the software packages used and data sources are given in Appendices A.1 and A.2.

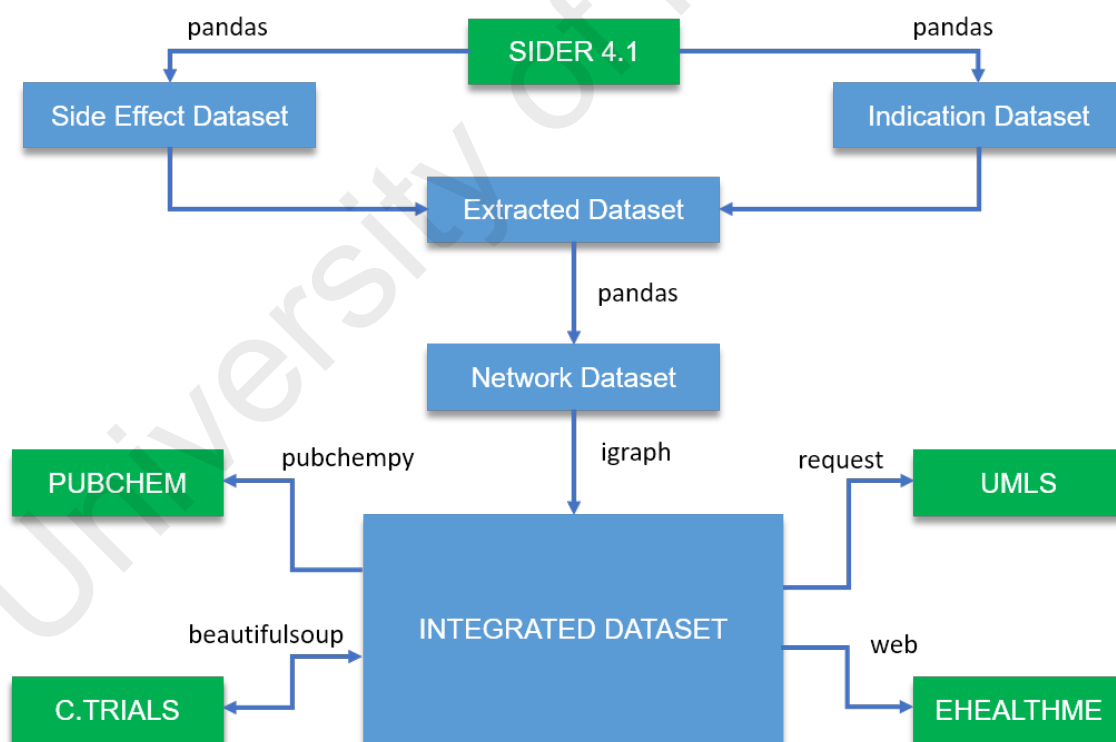


Figure 3.1: Data flow diagram. Green boxes represent data sources, blue boxes represent milestones and line labels represent software packages used.

SIDER 4.1

The main datasets were retrieved from SIDER website¹. The latest SIDER version 4.1 contains 1430 drugs and 5868 side effects with a total of 139756 drug-side effect pairs (Kuhn et al., 2015). Figures 3.2 - 3.4 show the typical datasets of the 3 files used. Detailed information on the files can be found in the download section of SIDER website.

SIDER used Search Tool for Interacting Chemicals (STITCH) compound IDs to uniquely identify drugs in their datasets. STITCH is a database of protein-chemical interactions and the STITCH compound IDs are based on PubChem (Kuhn et al., 2014). There are two types of compounds : stereo-specific compound and flat compound. The flat compounds are stereo-isomers that have been merged into one single compound. Whether a compound is a flat or stereo compound can be identified by looking at the compound ID. Flat compound has a compound ID that starts with CID1 and stereo compound ID starts with CID0.

CID100000085	CID000010917	C0000729	LLT	C0000729	Abdominal cramps
CID100000085	CID000010917	C0000729	PT	C0000737	Abdominal pain
CID100000085	CID000010917	C0000737	LLT	C0000737	Abdominal pain
CID100000085	CID000010917	C0000737	PT	C0687713	Gastrointestinal pain
CID100000085	CID000010917	C0000737	PT	C0000737	Abdominal pain
CID100000085	CID000010917	C0002418	LLT	C0002418	Amblyopia
CID100000085	CID000010917	C0002418	PT	C0002418	Amblyopia
CID100000085	CID000010917	C0002871	LLT	C0002871	Anaemia
CID100000085	CID000010917	C0002871	PT	C0002871	Anaemia
CID100000085	CID000010917	C0003123	LLT	C0003123	Anorexia

Figure 3.2: A typical dataset of meddra_all_se.tsv. The first and the second columns contain the STITCH flat and stereo compound IDs respectively. The third column contains the UMLS concept IDs which represent side effects for this dataset. The fourth column contains MedDRA concept types where LLT and PT are acronyms for lowest level term and preferred term respectively. The fifth column contains UMLS concept IDs for MedDRA term. The last column contains the side effect.

PubChem

PubChem contains databases on chemical molecules and is maintained by National Center for Biotechnology Information (NCBI), a component of National Library of

¹ As of 4 January 2018, the data can be downloaded from <http://sideeffects.embl.de/download/>

CID100000085	CID000010917	C0000737		21%	0.21	0.21	LLT	C0000737	Abdominal pain
CID100000085	CID000010917	C0000737		21%	0.21	0.21	PT	C0000737	Abdominal pain
CID100000085	CID000010917	C0000737		21%	0.21	0.21	PT	C0687713	Gastrointestinal pain
CID100000085	CID000010917	C0000737		5%	0.05	0.05	LLT	C0000737	Abdominal pain
CID100000085	CID000010917	C0000737		5%	0.05	0.05	PT	C0000737	Abdominal pain
CID100000085	CID000010917	C0000737		5%	0.05	0.05	PT	C0687713	Gastrointestinal pain
CID100000085	CID000010917	C0000737		6%	0.06	0.06	LLT	C0000737	Abdominal pain
CID100000085	CID000010917	C0000737		6%	0.06	0.06	PT	C0000737	Abdominal pain
CID100000085	CID000010917	C0000737		6%	0.06	0.06	PT	C0687713	Gastrointestinal pain
CID100000085	CID000010917	C0000737		9%	0.09	0.09	LLT	C0000737	Abdominal pain
CID100000085	CID000010917	C0000737		9%	0.09	0.09	PT	C0000737	Abdominal pain
CID100000085	CID000010917	C0000737		9%	0.09	0.09	PT	C0687713	Gastrointestinal pain
CID100000085	CID000010917	C0000737	placebo	17%	0.17	0.17	LLT	C0000737	Abdominal pain
CID100000085	CID000010917	C0000737	placebo	17%	0.17	0.17	PT	C0000737	Abdominal pain
CID100000085	CID000010917	C0000737	placebo	17%	0.17	0.17	PT	C0687713	Gastrointestinal pain

Figure 3.3: A typical dataset of meddra_freq.tsv. The first and the second columns contain STITCH flat and stereo compound IDs respectively. The third column contains the UMLS concept IDs which represent side effects for this dataset. The fourth column will show placebo if the side effect information comes from the placebo administration, an empty value otherwise. The fifth column describes the frequencies of the side effects. The sixth and seventh columns describe the lower and upperbound of the frequency. The last three columns describe the MedDRA information similar to the ones in meddra_all_se.tsv.

CID100000085	C0015544	text_mention	Failure to Thrive	LLT	C0015544	Failure to thrive
CID100000085	C0015544	text_mention	Failure to Thrive	PT	C0015544	Failure to thrive
CID100000085	C0020615	text_mention	Hypoglycemia	LLT	C0020615	Hypoglycaemia
CID100000085	C0020615	text_mention	Hypoglycemia	PT	C0020615	Hypoglycaemia
CID100000085	C0022661	NLP_indication	Kidney Failure, Chronic	LLT	C0022661	Renal failure chronic
CID100000085	C0022661	NLP_indication	Kidney Failure, Chronic	PT	C0022661	Renal failure chronic
CID100000085	C0025521	NLP_indication	Inborn Errors of Metabolism	LLT	C0025521	Inborn error of metabolism
CID100000085	C0025521	NLP_indication	Inborn Errors of Metabolism	PT	C0025521	Inborn error of metabolism
CID100000085	C0026827	text_mention	Muscle hypotonia	LLT	C0026827	Hypotonia
CID100000085	C0026827	text_mention	Muscle hypotonia	PT	C0026827	Hypotonia

Figure 3.4: A typical dataset of meddra_all_indications.tsv. The first column contains STITCH flat compound IDs. The second column contains UMLS concept IDs which represent indications for this dataset. The third column contains the methods of detection. The fourth column contains the MedDRA concept names. The last three columns describe the MedDRA information similar to those in meddra_all_se.tsv.

Medicine (NLM) which is in turn part of United States National Institutes of Health (NIH).

We used one of their databases called PubChem Compounds which contains the information about 92064620 compounds² (Kim et al., 2016). The information of particular drugs can be obtained using their compound identification numbers (CIDs) based on SIDER dataset. As guided by SIDER, to get the PubChem compound ID for a flat compound ID, we need to remove the substring "1000000000". Take aspirin (CID100002244) for example, the PubChem compound ID would be 2244. Therefore, we can use the number 2244 to search the drug information in PubChem. PubChem has provided a number of

² Number of compounds as of October 2017

ways to retrieve drug's information. We can always use their web interface for textual search or use Python interface called PubChemPy, which is based on PubChem Power User Gateway (PUG)-Representational State Transfer (REST), to retrieve information of the drug compounds (Kim et al., 2015).

ClinicalTrials.gov

Similar to PubChem, it is maintained by NLM at the NIH. ClinicalTrials.gov provides information of clinical studies, diseases, conditions and more through their web interface³.

UMLS

UMLS is also maintained by NLM and NIH. From the SIDER dataset, we can use the UMLS concept ID or UMLS Concept Unique Identifier (CUI) in UMLS Terminology Services (UTS) to locate the definitions of the conditions, which is useful for a person with no medical background. To access the dataset, we need to go to UTS website⁴ and from there we are required to create a UTS account and obtain a UMLS Metathesaurus License. UTS provides both web interfaces and web services, i.e, Simple Object Access Protocol (SOAP) and REST application program interface (API).

eHealthMe.com

eHealthMe.com continuously collects and analyzes data from Food and Drug Administration (FDA) and social media. This website is useful for finding reported side effects of particular drugs.

3.2 Analyzing drug networks

This section will describe the analysis of drug networks. We will first describe the extraction of the raw data to form network datasets, followed by the construction and the

³ The web interface for searching information is at <https://clinicaltrials.gov/ct2/home>

⁴ <https://uts.nlm.nih.gov/home.html>

analyses of the networks.

3.2.1 Forming network dataset

The network datasets contain rows of source nodes, target nodes and edge weights. It is a convention to represent the relationships between objects using adjacency tables in network theory. The datasets will be used to construct our networks. The flow to obtain network datasets is simplified and shown in Figure 3.5.

To obtain both network datasets, first we must extract and list down all indications and side effects found in the raw dataset. We will call the results as the extracted datasets. The extracted indication dataset was obtained by simply choosing the flat compound ID column and UMLS concept ID in meddra_all_indications.tsv. The extracted side effect dataset was extracted in a similar manner but with a few additional steps from meddra_freq.tsv.

The first step was to remove placebo administration from the dataset as we only consider side effects from “real” drugs. The second step was to choose side effects that happen within a certain range of occurrences. However, the frequencies in meddra_freq.tsv contain various values of side effect frequencies, from empty, exact frequencies to ranges of frequencies. To categorize the frequency values, we used a convention by CIOMS that describes side effect frequencies as given in Table 3.1.

Table 3.1: CIOMS frequency convention.

Standard	Frequency
Very Common	$\geq 1/10$
Common (Frequent)	$\geq 1/100$ and $< 1/10$
Uncommon (Infrequent)	$\geq 1/1000$ and $< 1/100$
Rare	$\geq 1/10000$ and $< 1/1000$
Very rare	$< 1/10000$

There are few cases on side effect frequency alterations for existing frequency values :

1. When value is "Frequent" or "Infrequent", then we convert it into "Common" or "Uncommon" respectively.

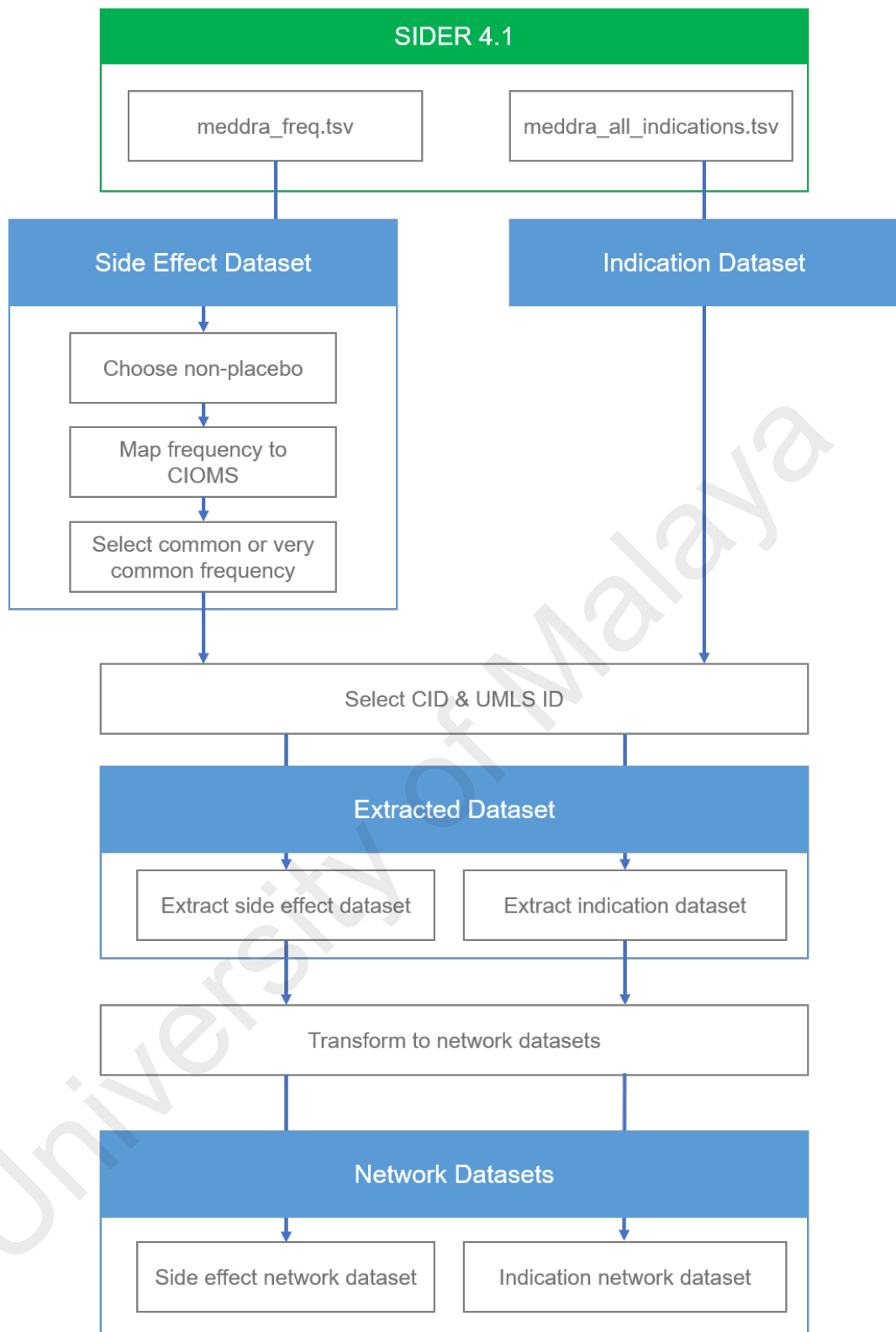


Figure 3.5: Flowchart to obtain network dataset.

2. When the exact frequency is given, then it was mapped to its category.
3. When a range of frequencies is given, then the median frequency was used and mapped to its category.

4. When no frequency is given, then the row was ignored.

After we catered for all these cases, we filtered it again to choose only side effects with either common or very common frequencies. For this thesis, we only chose these frequencies based on our assumption that higher frequencies may lead to better relationships. For example, if both drugs *A* and *B* have the same side effects that happen frequently, then in our assumption these two drugs might have something in common.

We were then able to obtain the extracted dataset by selecting both drug compound ID and the UMLS concept ID columns to be used in getting the network datasets. Examples for both extracted indication and side effect datasets are shown in Table 3.2.

Table 3.2: Examples of extracted datasets.

(a) Extracted side effect dataset

Drug CID	UMLS ID
CID100000085	C0000729
CID100000085	C0000737
CID100000085	C0002418
CID100000085	C0002871
CID100000085	C0003123
CID100000085	C0003467
CID100000085	C0003811
CID100000085	C0004093
CID100000085	C0004238
CID100000085	C0004604

(b) Extracted indication dataset

Drug CID	UMLS ID
CID100000085	C0015544
CID100000085	C0020615
CID100000085	C0022661
CID100000085	C0025521
CID100000085	C0026827
CID100000085	C0085584
CID100000085	C0151786
CID100000085	C0878544
CID100000085	C1142132
CID100000119	C0001768

These extracted datasets were then transformed into network datasets. We connected any two drugs with at least one similarity where the edge weight is equivalent to the number of similarities. An example to create an edge is shown in Figure 3.6.

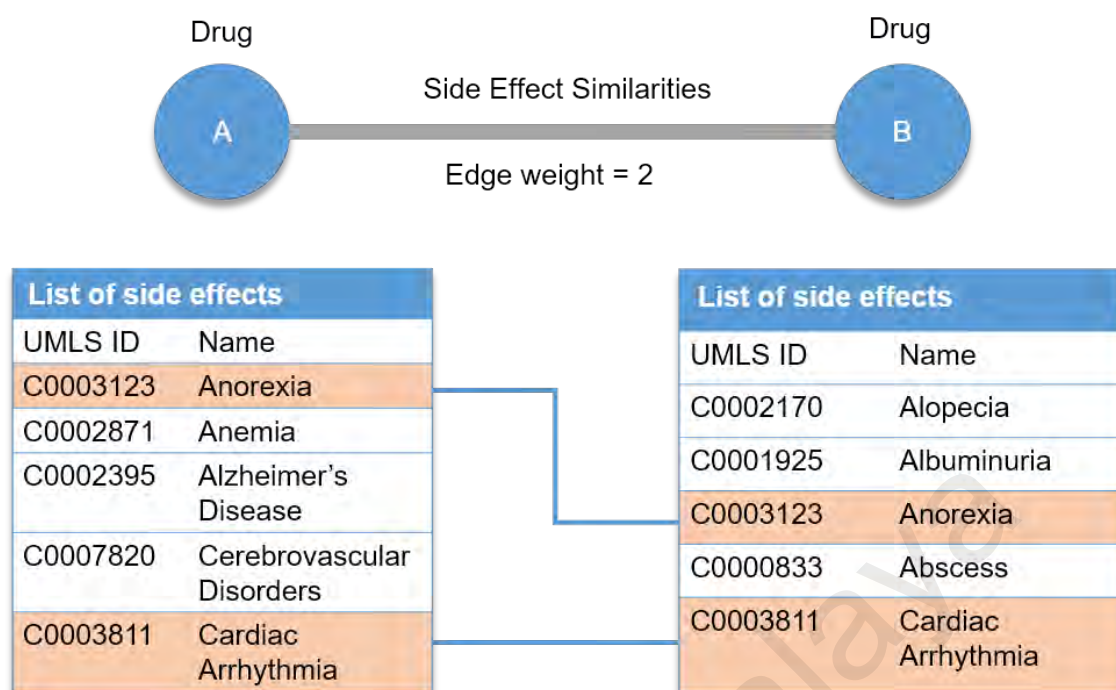


Figure 3.6: Example on creating an edge in the drug-side effect network. The first step is to list all side effects for each drugs and set the edge weight equivalent to the number of side effect similarities.

To create an edge between drugs A and B , the first required step is to list out every possible side effects for both drugs based on extracted side effect dataset. Let $se(A)$ be the set of side effects for drug A and $se(B)$ be the set of side effects for B . Then the set of shared side effects is equivalent to $se(A) \cap se(B)$. The number of shared side effects will be the edge weight between those two drugs. We repetitively applied this method to all other drugs in the extracted side effect dataset using pandas package in Python to obtain an adjacency table. Since our network is a simple undirected network which does not contains any self loops (node that linked to itself), the adjacency table can be further simplified into an adjacency list (See Figure 3.7).

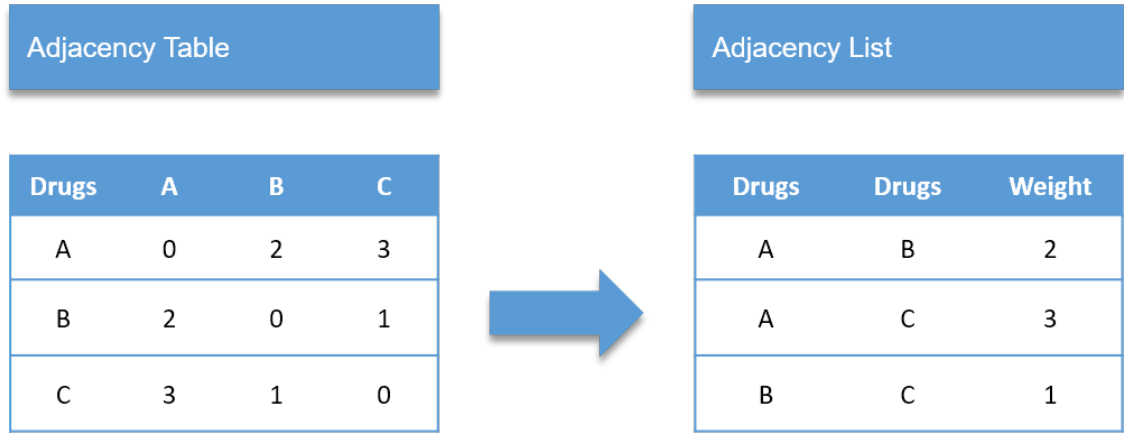


Figure 3.7: Transforming adjacency table into adjacency list.

3.2.2 Constructing drug networks

We constructed two networks :

1. **Drug-side effect network** - constructed using side effect network dataset. This network contains nodes that represent drugs and edge weights that represent the number of side effect similarities with either common or very common frequencies. It would be used to predict potentially novel indications of drugs in the network. We denote this network with g_{se} .
2. **Drug-indication network** - constructed using indication network dataset. This network contains nodes that represent drugs and edge weights that represent the number of indication similarities. It would be used to predict potentially novel side effects of drugs in the network. We denote this network with g_{ind} .

To construct both networks, we used a Python package called python-igraph (version 0.7.1.post6) by providing the network datasets (Csardi & Nepusz, 2006). The detailed description on network construction and visualization is provided in Appendix A.3.

We analyzed and obtained the network properties of both g_{se} and g_{ind} by using built-in functions in python-igraph (Csardi & Nepusz, 2006). For network-level properties, we

used some of the standard measures such as vertex and edge counts, average transitivity, density, diameter and average path length. For element-level properties, we used various node measurements as mentioned in previous chapter on the giant components of both networks such as degree centrality, betweenness centrality, closeness centrality, Burt's constraint, HITS and PageRank (Freeman, 1978; Burt, 2004; Page et al., 1999). Drugs in both networks were ranked in each of the node measurements. For ease of reanalyzing, we saved all computations using `shelve` package in Python into our integrated dataset. This dataset was used to check information on drugs, side effects or indications from our data sources using software packages from Python.

3.3 Drug repositioning using network analysis

In this section, we will describe the application of network analysis in drug repositioning. We proposed an algorithm for drug repositioning that predicts potentially new uses of existing drugs based on the network and its properties. The overall flow for the prediction is shown in Figure 3.8.

We used g_{se} to predict potentially new indications of drugs. This is based on a hypothesis in (Duran-Frigola & Aloy, 2012) whereby drugs with similar side effects profiles may also share similar therapeutic properties through related MOA. Suppose that we have two drugs that share similar side effects. Then by using this hypothesis we can say that those drugs may share similar therapeutic uses. Supported by the strong association between side effects and indications, we further extend the hypothesis whereby two drugs with similar indications may also share similar side effects (Wang et al., 2014; Duran-Frigola & Aloy, 2012). The extended hypothesis have two implications and they are visualized in Figure 3.9.

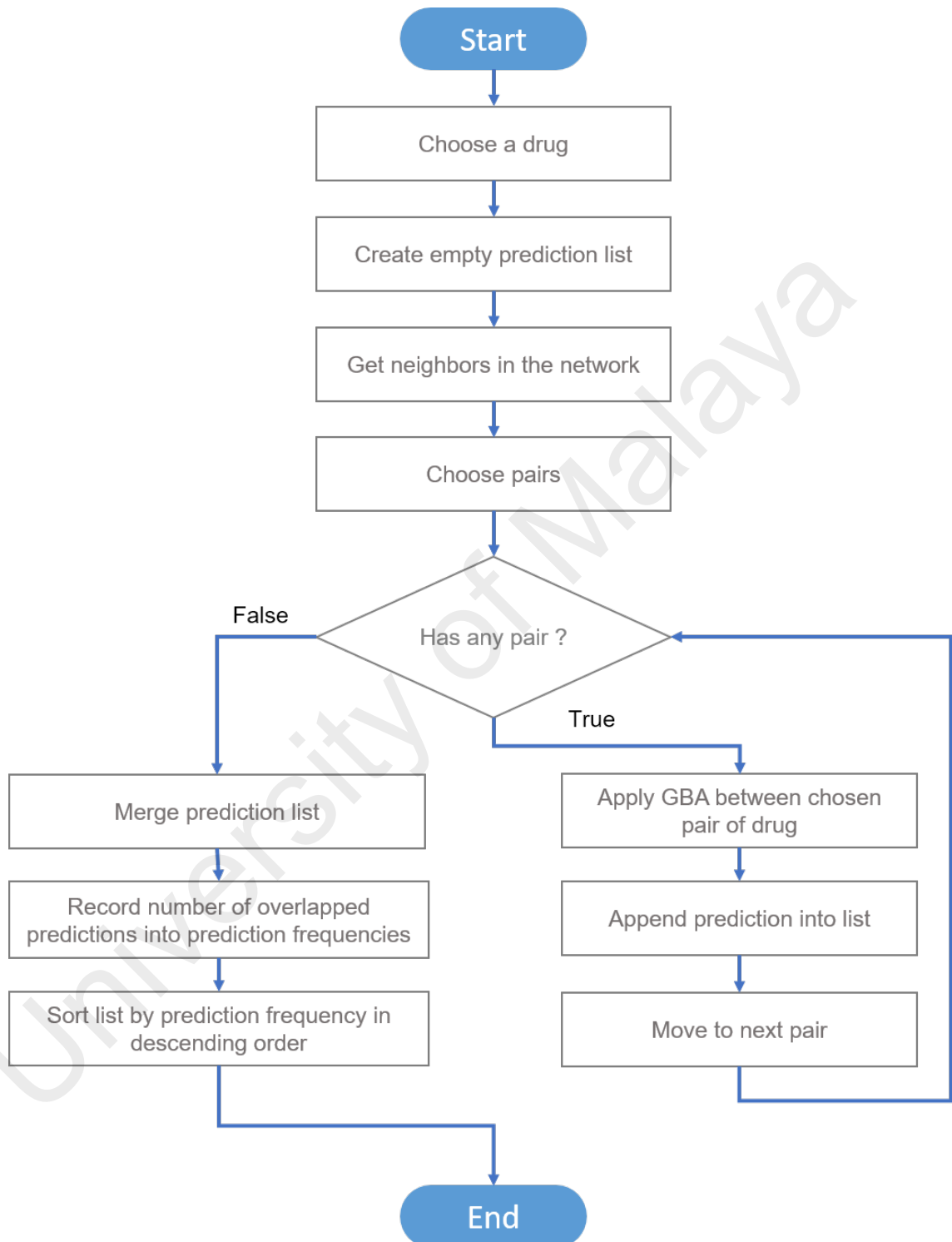
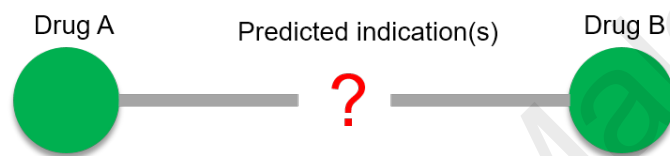
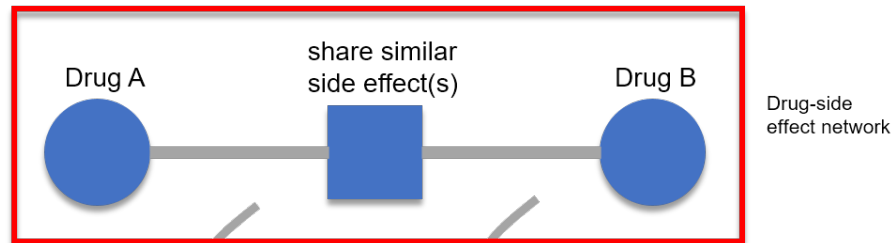
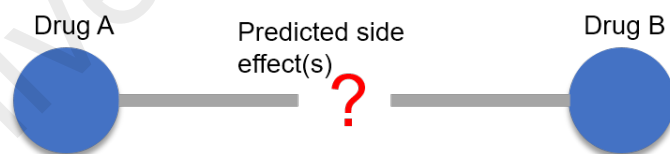
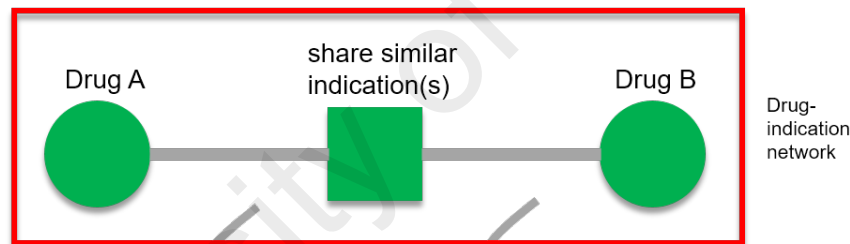


Figure 3.8: Flowchart for prediction of a single drug.



(a) First implication



(b) Second implication

Figure 3.9: Our extended hypothesis based on the hypothesis by Duran-Frigola and Aloy. (a) Suppose two drugs share similar side effects, then we may infer indications between these two drugs - the one that we already knew and the one that is yet to be discovered. (b) Similarly, we may also infer side effects based on two drugs that share similar indications.

3.3.1 Choosing drugs

Based on the previous section, it may seem like we can predict potentially new uses of any drugs in the network provided that those drugs have common side effects or indications. Our first question is, among those drugs in the network, which drugs have better predictability compared to the other drugs. We believe the answer relies on the drug positions in the network. Therefore, we chose drugs based on node measurements mentioned in Section 3.2.2. The interpretations for each of the measurements is shown in Table 3.3.

Table 3.3: Node Measurements.

Measurement	Definition and Interpretation
Degree Centrality	A measurement to calculate the number of neighbouring nodes of a node. A high degree indicates more links that can lead to potentially new discoveries.
Betweenness Centrality	A measurement to calculate the number of shortest paths that pass through a node. Drug with relatively high betweenness may be used to treat more than two categories of diseases (Nacher & Schwartz, 2008; Spiro et al., 2008).
Burt's Constraint	A measurement to find structural holes in a network, each of which indicates a gap between two nodes who have complementary sources to information. A structural hole in the drug-side effect network implies an opportunity to discover unnoticed indications of drugs.
Closeness Centrality	The reciprocal of the total distance of a node to all other nodes; a high closeness implies that the node is close to all other nodes. In the drug-side effect network, a high closeness implies that the node is similar to drugs including those not among its neighbours.
PageRank	A measurement computed recursively to rank a node based on its number of links and the ranks of adjacent nodes. In the drug-side effect network, a highly ranked node represents a drug sharing side effects with many others that are in turn sharing numerous side effects with their neighbours.
HITS	A measurement to rank a node using its importance in providing information on a topic and in giving links to other nodes providing information on the same topic. In the drug-side effect network, a highly ranked node indicates that the node provides useful information and is linked significantly to other drugs also providing useful information on side effects.

3.3.2 Choosing pairs

The prediction was conducted based on chosen drugs in previous section from their neighbouring nodes in the network. However, it is possible for a chosen drug to have a high number of neighbouring nodes. Based on our extended hypothesis, we have another curiosity. Suppose that we have these two situations :

1. $|se(A)| = |se(B)| = 1$ and $|se(A) \cap se(B)| = 1$
2. $|se(A)| = |se(B)| > 10$ and $|se(A) \cap se(B)| = 1$

Drugs in both situations share the same number of side effects. In which situations that both drugs can be considered similar ? We believe that drugs can be considered similar if they share certain percentage of similarity. To quantify this, we will use Jaccard indexes. (See Definitions 1 and 2).

Definition 1 [*Jaccard Index for Shared Side Effects*] Let $se(m)$ and $se(n)$ be the sets of side effects of drugs m and n respectively. Then the Jaccard index for shared side effects of the drugs is given by

$$J_A(m, n)_{se} = \frac{|se(m) \cap se(n)|}{|se(m)| + |se(n)| - |se(m) \cap se(n)|}. \quad (3.1)$$

Definition 2 [*Jaccard Index for Shared Neighbours*] Let $nb(m)$ and $nb(n)$ be the collections of neighbours of two nodes m and n respectively in g_{se} . Then the Jaccard index for shared neighbours of the drugs is given by

$$J_B(m, n)_{se} = \frac{|nb(m) \cap nb(n)|}{|nb(m)| + |nb(n)| - |nb(m) \cap nb(n)|}. \quad (3.2)$$

The above definitions have the same objectives – to measure similarity between drugs. For Definition 1, we consider drugs to be similar if they share a relatively high percentage

of shared side effects. For Definition 2, we consider drugs to be similar if they share certain percentage of neighbours in the network. We further combine these two Jaccard indexes to form a drug similarity score by simply multiplying each other.

Definition 3 [*Drug Similarity Score*] *The similarity score of two nodes m and n in g_{se} is given by*

$$DSC_{se}(m, n) = J_A(m, n)_{se} \times J_B(m, n)_{se}. \quad (3.3)$$

To have a better predictability, we need to choose two drugs that are high in similarity score but not too high as we are likely to lose some potential predictions. However, we also cannot choose drugs that are too low in similarity score as it might lead to a false prediction. To cater this situation, we wanted to set a threshold above which such drugs are considered similar in the network. However, a high threshold may result in fewer to no drugs being similar to the chosen drug. Conversely, a low threshold may lead to the inclusion of most or all neighbours of any chosen drug as similar to the latter. There must be a way to justify the selection of such a threshold of similarity. This we settled in subsequent steps by first predicting new indications of prominent drugs and then comparing the network based predictions to clinical studies using different levels of similarity. Choosing pairs in drug indication network is similar by using sets of drug indications and drug indication network.

3.3.3 Prediction

The prediction of potentially new indications or side effects of drugs was implemented by applying a technique called guilt-by-association (GBA) between chosen drugs and its neighbouring nodes in the network (Chiang & Butte, 2009). Suppose we have two drugs m and n in g_{se} such that are connected in the network. Based on our hypothesis, these two drugs may have indication similarity. Let $ind(m)$ and $ind(n)$ be respectively their sets of

indications. The collection of potentially new indications is formulized in Equation 3.4 and visualized in Figure 3.10.

$$NPI(m, n) = ind(n) - ind(m) \quad (3.4)$$

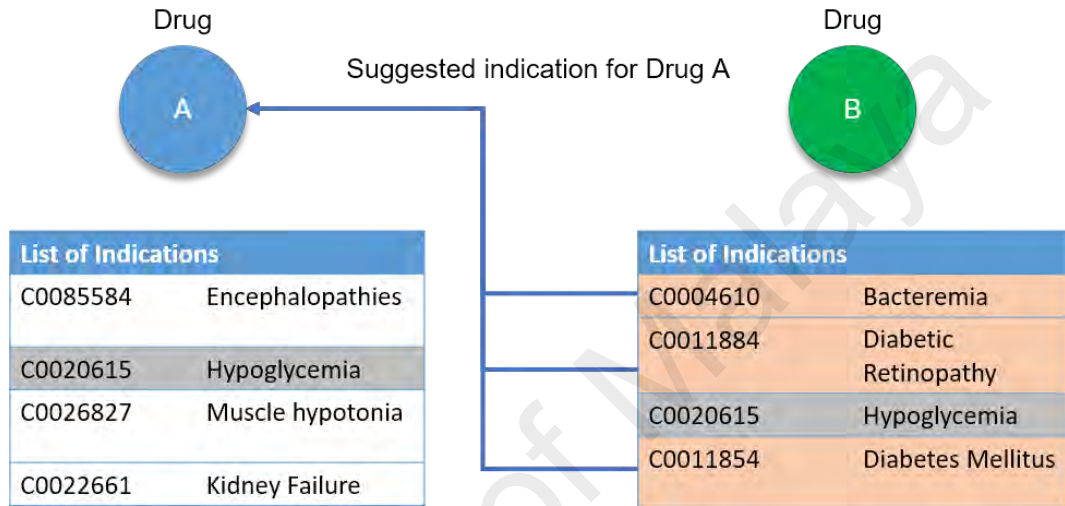


Figure 3.10: Example on guilt-by-association for potentially new drug indications.

In network perspective, instead of applying GBA on one connection, we can apply this method to the neighbours of selected drugs. Suppose we choose drug m to be the selected drug, we can apply Equation 3.4 to each drug n_i where n_i is a neighbour of drug m . This way we get a collection of sets of potentially new indications of drug m , one from each neighbours of m . Taking this into account, we kept track of the frequency of occurrences of each potentially new indication of a prominent drug while looping over the neighbours of the drug which equal or surpass the minimum threshold of similarity scores. More explicitly, we introduced the following set of weighted new indications of drug m , i.e.,

$$NPIF(m) = \{(i, f) \mid f = \text{no. of neighbours } n \text{ of } m \text{ such that } i \in NPI(m, n)\} \quad (3.5)$$

As a natural guiding principle, the higher the frequency f for each $(i, f) \in NPIF(m)$, the more likely the indication i is indeed a novel one of the drug m . We adhered to this principle in the process of predicting new indications of drugs in drug-side effect network. We used the similar technique to predict potentially new side effects of drugs in drug-indication network.

As an illustration, we predicted new indications of chosen drugs using different levels of similarity. Practically, we used the predictions to obtain a reasonable threshold of the level of similarity for the whole network by comparing predicted indications of chosen drugs to those available in the clinical studies. This was done by searching through the ClinicalTrials.gov dataset and we managed to retrieve the number of clinical studies that have been performed to uncover possibly new indications of chosen drugs. Likewise, we may identify a similarity threshold for the drug-indication network by employing the data of reported side effects based on feedback from end users available at eHealthMe.com. For our work, we used the threshold of 75% for this network as for the drug-side effect network.

It should be noted that once a similarity threshold for the whole network has been selected based on the predictions for prominent drugs, prediction of new indications can then be carried out for any drugs similar to each other even if they are not the prominent ones. In a sense, the chosen similarity threshold acts as a gauge of the desired accuracy of our predictions.

3.3.4 Validation

This section will explain the validation step for both predicted indications and side effects.

Validating predicted indication

To validate our predictions on drug indications, we compared our predictions based on clinical studies currently listed in ClinicalTrials.gov. While ClinicalTrials.gov did provide clinical data for analysis in Extensible Markup Language (XML), the data is a bit too detailed and contain tremendous irrelevant data that we wanted for this thesis. We can also manually search clinical studies drug by drug in the webpages but it maybe time consuming and the information might change frequently. Therefore, we used web scraping to create our own data from the webpages itself. Web scraping, which also known as web harvesting or web data extraction, simply means taking data from the webpages programmatically. This is done by using beautifulsoup4 package from Python to parse the HyperText Markup Language (HTML) from the webpages by using the following URLs :

1. <https://clinicaltrials.gov/search?intr=<intervention>&cond=<condition>>

Example : <https://clinicaltrials.gov/search?intr=everolimus&cond=cancer>

2. <https://clinicaltrials.gov/search?intr=<intervention>>

Example : <https://clinicaltrials.gov/search?intr=everolimus>

where the <condition> is the disease that is being studied and the <intervention> is the field for the given drugs to the clinical trials. The first option is useful for a single check between a single drug and disease, but to parse it everytime we validate may also be time consuming. Therefore, another strategy is to use the second URL, browse by topics and parse all the conditions and studies for all the pages. The results are compiled for all other drugs in the network and used for every indication validations. The second option however may not contain updated list but we proceeded this way to reduce time consumption at the risk of slightly outdated list. Table 3.4 is the example of the extracted data from ClinicalTrials.org. The detail description is provided in Appendix A.4.

Table 3.4: Sample of clinical studies.

Drug CID	Drug Name	Condition	No of clinical studies
CID100000085	Carnitine	Abdominal Obesity Metabolic Syndrome	2
CID100000085	Carnitine	Acid-Base Imbalance	2
CID100000085	Carnitine	Acidosis	2
CID100000085	Carnitine	Acidosis, Lactic	1
CID100000085	Carnitine	Acne Vulgaris	1
CID100000085	Carnitine	Acquired Immunodeficiency Syndrome	16
CID100000085	Carnitine	Acute Kidney Injury	3
CID100000085	Carnitine	Adenocarcinoma	1
CID100000085	Carnitine	Adenocarcinoma of the Ap- pendix	1
CID100000085	Carnitine	Adnexal Diseases	1

Validating predicted side effects

To validate our predictions on drug side effect, we compared our predictions based on reported side effects in eHealthMe.com. We used a programmatical ways of validations by using these URLs :

1. <http://www.ehealthme.com/ds/<drug-name>/<side-effect-name>/>

Example : <http://www.ehealthme.com/ds/dexamethasone/dyspnea/>

2. <http://www.ehealthme.com/drug/<drug-name>/>

Example : <http://www.ehealthme.com/drug/dexamethasone/>

The first URL is used to validate between chosen drugs and predicted side effect. The webpage will display the number of reported side effects along with some statistical analysis. The second URL is used to view all associated side effects of the specified drug.

CHAPTER 4: RESULTS AND DISCUSSION

This chapter will present the results of our study. We will first show the analysis on SIDER dataset, followed by the analysis on drug networks and drug repositioning using our proposed algorithm.

4.1 Data analysis on SIDER dataset

As indicated in Table 4.1, there are 1465 unique drugs based on STITCH compound ID in SIDER datasets and not all drugs were recorded to have indications or side effects. Out of 1465 drugs, 1430 drugs were recorded to have at least 1 side effects and 1437 drugs were recorded to have at least 1 indications. We found that only 887 drugs were having at least 1 side effects with common or very common frequencies. These number of side effects and indications were based on the UMLS concept ID available in the SIDER dataset.

Table 4.1: Information on SIDER 4.1 dataset

No of drugs	1465
No of side effects	5868 for 1430 drugs
No of indications	2714 for 1437 drugs
No of side effects with common/very common frequencies	1957 for 887 drugs

As shown in Table 4.2, on average, each drug was used for 10 different indications and it can have up to 172 distinct indications. For side effects, the mean is around 100 side effects per drug and a drug have at most 769 side effects varying over the range of all frequencies. For side effects with very common or common (VCC) frequencies, the mean is around 28 side effects per drug and a drug have at most 227 unique side effects. Also on average, each drug is belongs to 1 Anatomical Therapeutic Chemical (ATC) or drug's category according to World Health Organization (WHO). Drug such as Dexamethasone (CID100003003) falls in 22 different ATCs and may be used for 22 different areas of treatment.

Table 4.2: Summary on SIDER 4.1 dataset

	Indication	Side Effect	Side Effect (VCC)	ATC
count	1465	1465	1465	1465
mean	9.987030717	94.60273038	17.56518771	1.038225256
std	14.13311891	99.89862618	27.86543122	1.246815448
min	0	0	0	0
25% pctl	2	27	0	0
50% pctl	5	61	6	1
75% pctl	12	127	23	1
max	172	769	227	22

We have analyzed the side effect dataset and as we were only interested in side effects with VCC frequencies, we will only show the analysis of side effects with those frequencies and we named it as VCC side effect dataset. As shown in Table 4.3, VCC side effect dataset consists of 25733 pairs of drugs and side effects. We have listed the top five drugs and side effects in Tables 4.4 and 4.5. Out of 877 drugs, CID100005064 (Ribavirin) has the most VCC side effects and the side effect that appears the most is C0027497 (Nausea). The summary on both CID100005064 (Ribavirin) and C0027497 (Nausea) are shown in Tables 4.6 and 4.7.

Table 4.3: Summary of VCC side effect dataset.

(a) Drug		(b) Side Effect	
Pairs	25733	Pairs	25733
Unique drugs	887	Unique SE	1957
Mean	29.011274	Mean	13.149208
Std. Deviation	30.831734	Std. Deviation	43.997400
Min side effect	1	Min drug	1
At 25% pctl	9	At 25% pctl	1
At 50% pctl	18	At 50% pctl	2
At 75% pctl	38	At 75% pctl	7
Max side effect	227	Max drug	634
Top drug	CID100005064	Top side effect	C0027497

Out of 887 drugs, 75% of the drugs were having less than 38 side effects and on average, each drug has around 29 to 30 side effects with a standard deviation of 30.83. From 1957 number of side effects, 75% of the side effects are belong to less than 7 drugs and on average, each side effect is belongs to 13 or 14 number of drugs with a standard deviation

of 43.99. We found that the top 5 side effects were mostly common and well-known side effects.

Table 4.4: Top five drugs by number of side effects

Drug Compound ID	Drug Name	No of side effects
CID100005064	Ribavirin	227
CID100005514	Topiramate	217
CID100005095	Ropinirole	213
CID100005372	Tacrolimus	196
CID100005073	Risperidone	173

Table 4.5: Top five side effects by number of drugs

UMLS Concept ID	Concept Name	No of drugs having this side effect
C0027497	Nausea	634
C0018681	Headache	624
C0011991	Diarrhea	521
C0012833	Dizziness	502
C0042963	Vomiting	493

Table 4.6: Summary of CID100005064 (Ribavirin)

Drug Compound ID	CID100005064
Drug Name	Ribavirin
Description	Ribavirin is a nucleoside antimetabolite antiviral agent that blocks nucleic acid synthesis and is used against both RNA and DNA viruses.
List of ATCs	['J05AB04']
Total ATCs	1
Total Indications	25
Total Side Effects	404
Total Side Effects (VCC)	227

Table 4.7: Summary of C0027497 (Nausea)

UMLS Concept ID	C0027497
Name	Nausea
Definition	NCI : 'Upper abdominal discomfort associated with an urge to vomit.
No of drugs having this indication	1207
Example of drugs	CID100000085 - Carnitine CID100000137 - 5-aminolevulinic acid CID100000143 - Leucovorin CID100000158 - PGE2 CID100000159 - Prostacyclin

For indication dataset, as shown in Table 4.8, it consists of 14631 pairs of drugs and indications. We have listed the top five drugs and indications in Tables 4.9 and 4.10. In this dataset, CID100003003 (Dexamethasone) has the most indications and C0009450 (Communicable Diseases) is the most common indications among all of the drugs. The summary on both CID100003003 (Dexamethasone) and C0009450 (Communicable Diseases) are shown in Tables 4.11 and 4.12.

Table 4.8: Summary of indication dataset.

(a) Drug		(b) Indication	
Pairs	14631	Pairs	14631
Unique drugs	1437	Unique indication	2714
Mean	10.181628	Mean	5.390936
Std. Deviation	14.200602	Std. Deviation	10.265912
Min indication	1	Min drug	1
At 25% pctl	2	At 25% pctl	1
At 50% pctl	6	At 50% pctl	2
At 75% pctl	13	At 75% pctl	5
Max indication	172	Max drug	210
Top drug	CID100003003	Top indication	C0009450

Table 4.9: Top five drugs by number of indications.

Drug Compound ID	Drug Name	No of indications per drugs
CID100003003	Dexamethasone	172
CID100004900	Prednisone	124
CID100004159	Methylprednisolone	120
CID100003640	Cortisol	119
CID100004894	Prednisolone	111

Table 4.10: Top five indications by number of drugs.

UMLS Concept ID	Concept Name	No of drugs having this indication
C0009450	Communicable Diseases	210
C1565489	Renal Insufficiency	150
C0020538	Hypertensive disease	136
C0006826	Malignant Neoplasms	121
C0030193	Pain	89

Table 4.11: Summary of CID100003003 (Dexamethasone)

Drug Compound ID	CID100003003
Drug Name	Dexamethasone
Description	MeSH : Dexamethasone is an anti-inflammatory 9-fluoro-glucocorticoid.
List of ATCs	['A01AC02', 'A07EA04', 'C05AA05', 'C05AA09', '...
Total ATCs	22
Total Indications	172
Total Side Effects	214
Total Side Effects VCC	25

Table 4.12: Summary of C0009450 (Communicable Diseases)

UMLS Concept ID	C0009450
Name	Communicable Diseases
Definition	NCI :A disorder resulting from the presence and activity of a microbial, viral, or parasitic agent. It can be transmitted by direct or indirect contact.
No of drugs having this indication	210
Example of drugs	CID100000119 - Gamma-aminobutyric acid CID100000175 - Acetate CID100000206 - Glucose CID100000298 - Chloramphenicol CID100000401 - D-cycloserine

Out of 1437 drugs, 75% of the drugs have less than 13 indications and on average, each drug has 10 to 11 indications with a standard deviation of 14.20. It means that on average each drug can be used in multiple disease treatments. CID100003003 (Dexamethasone) itself can be used in 172 different indications which is the highest in the dataset and it is also relatively high compared to CID100004900 (Prednisone). From 2714 indications, 75% of the indications in the dataset are belong to less than 5 drugs. It has an average between 5 to 6 drugs per indication with a standard deviation of 10.25. It means that on average only 5 to 6 drugs have a particular indication that able to treat a certain disease.

4.2 Network analysis on drug networks

This section shows the result from the analysis of the two drug networks.

Network-level properties

When constructing networks, it is common to find groups of connected nodes, commonly called clusters in network analysis. The initial drug-side effect network and drug-indication network contain 6 and 13 clusters respectively (See Table 4.13). Both networks contain a noticeable giant components and small nodes with size of 1 or 2. The giant component is the largest cluster in the network capturing the most interconnected nodes while the smaller clusters contain isolated nodes that did not share any common side effects or indications with other drugs in the network.

Table 4.13: Network clusters

Network	Cluster size
Drug-side effect network	882, 1, 1, 1, 1, 1
Drug-indication network	1424, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1

Network analysis measurements would be meaningful if there exists a path between nodes in the network. Therefore, we were only interested in the giant component and all of the measurements were conducted in the giant components for both networks. Table 4.14 shows the network-level properties of the giant components in both networks and Figure 4.1 shows the visualization of the networks using ForceAtlas2 projections (Jacomy et al., 2014) in Gephi.

Table 4.14: Summary on network-level properties.

Drug-side effect network		Drug-indication network	
Node Count	882	Node Count	1424
Edge Count	318147	Edge Count	106274
Avg. Transitivity	0.937	Avg. Transitivity	0.711
Density	0.819	Density	0.105
Diameter	3	Diameter	5
Avg. path length	1.182	Avg. path length	2.071

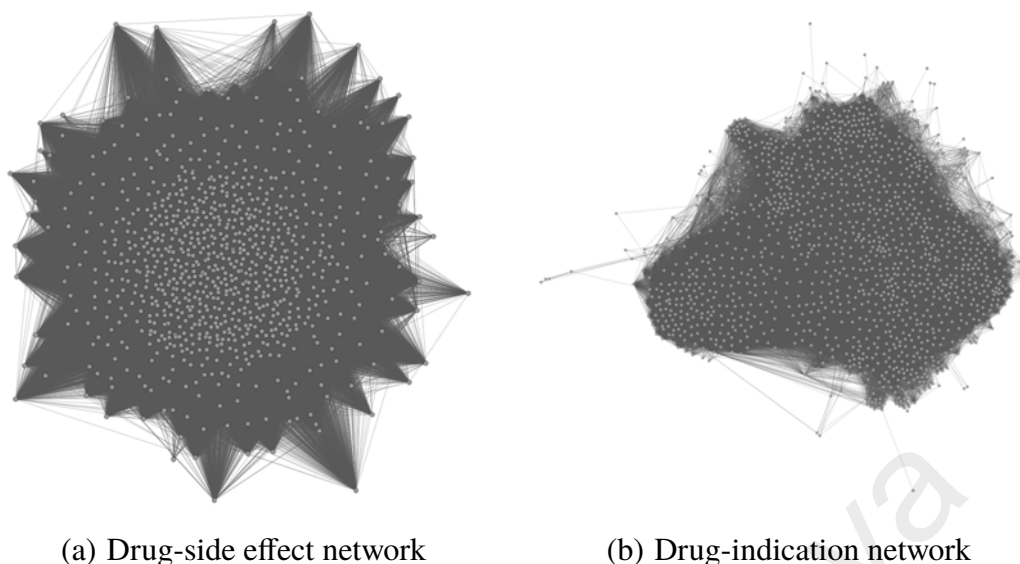


Figure 4.1: Giant components of both drug networks. Each grey dot represents a drug and each dark grey line represents an edge between two drugs.

The giant component of the drug-side effect network in Figure 4.1(a) consists of 882 nodes and 318147 edges. It can be observed that most drugs were concentrated at the center of the network. High network density, low diameter and average path length lead to this which also suggest that most drugs were interconnected. The nodes forming the diameter are CID100000861 (Triiodothyronine) and CID100004436 (Naphazoline) which suggest that these two drugs have the least possible similarities. The mean degree score is relatively high, i.e., each drug has 721 to 722 neighbouring drugs, and this suggests that most drugs share at least 1 side effect with other drugs in the network.

The giant component of drug-indication network in Figure 4.1(b) consists of 1424 nodes and 10624 edges with a low density at 0.105. It can be observed that there are few concentrated areas in the network and each area would most likely contains drugs that share similar indication in particular therapeutic group. The nodes forming the diameter are CID100000896 (Melatonin) and CID100041744 (Valrubicin).

Element-level properties

As shown in Table 4.15 and Figure 4.2, on average, each drug in the drug-side effect network has a high degree score with more than 75% of the drugs were connected to other 723 drugs. Based on the degree distribution in Figure 4.2, there are more drugs with more neighbours. This may contribute to a low path length for drugs to have similarities and thus produce a high value of closeness, i.e., drugs are very close to each other.

Table 4.15: Summary of node measurements for drug side effect network.

	auth	betweenness	burt	closeness	degree	pgrank
Mean	0.8732	79.9683	0.0076	0.8664	721.4218	0.0011
Std. Deviation	0.2208	89.5464	0.0250	0.1128	186.5961	0.0003
Min	0.0024	0.0000	0.0047	0.4894	2	0.0002
At 25% pctl	0.8911	23.3851	0.0048	0.8479	723	0.0011
At 50% pctl	0.9633	55.9972	0.0049	0.9082	792	0.0012
At 75% pctl	0.9851	104.4715	0.0050	0.9372	822	0.0013
Max	1.0000	919.4074	0.5013	0.9789	862	0.0014

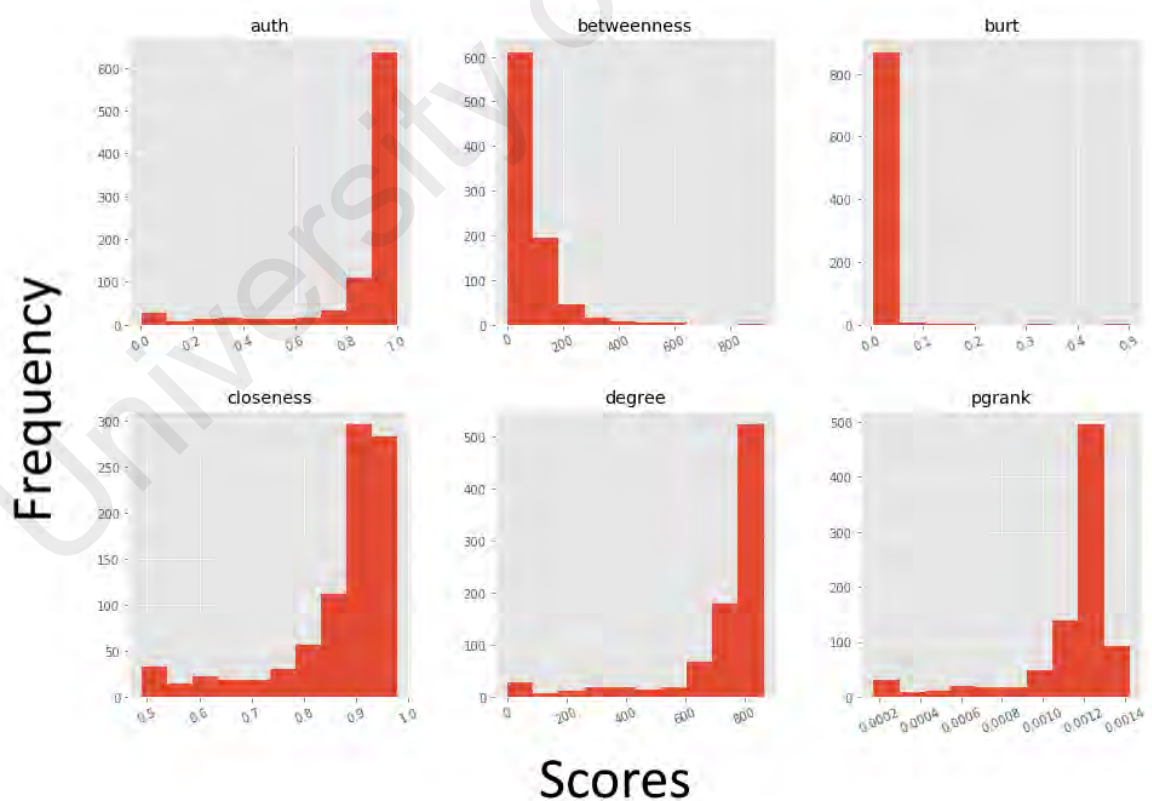


Figure 4.2: Visualization for Table 4.15

The pattern of authority (HITS) and PageRank scores might also indicate that there are more important drugs that are connected among themselves. Most drugs have a very low Burt's constraint score which indicates that most drugs have freedom to share information between drugs. We have listed the top ten scorers for each of the 6 node measurements in Table 4.16. The top scorers for degree centrality obtained score above 850 with CID100004666 (Paclitaxel) and CID100005372 (Tacrolimus) ranked the highest score in this measurement. The Pearson's correlation coefficient between number of VCC side effect and degree centrality score for drugs in this network is around 0.43. This indicates that drugs with more side effect are not necessarily connected to more drugs.

Overall, CID106442177 (Everolimus) scores the highest betweenness centrality, leaving CID100005372 (Tacrolimus) in the second place. These suggest that both CID106442177 and CID100005372 are possibly located between groups of drug in the network. For closeness centrality, both CID100004666 (Paclitaxel) and CID100005372 (Tacrolimus) obtained the highest score which indicates that these drugs located at the center of the network and require the least path to relate with the other drugs. One of the top scorers for HITS or authority score is CID100004666 (Paclitaxel). This drug acts as the hub in the network that may contain valuable source of information. Note that the top scorers obtained authority score near to 1. Based on network visualization, the top scorers were actually connected among each other, therefore produce slightly similar scores.

For PageRank, CID106442177 (Everolimus) again attains the highest score which suggests this drug to be an important drug in the drug-side effect network. Finally, the top scorer for the inverse of Burt's constraint is CID100005372 (Tacrolimus) which indicates that this drug is less constrained and has less structural holes compared to other drugs. We inverted the Burt's constraint so that the top scorer will have a score closer to 0. The top scorer will most likely be associated with more favorable outcomes (Burt, 2004).

Table 4.16: Top ten scorers in drug-side effect network by node measurement. The scores measured are degree centrality C_D , betweenness centrality C_B , closeness centrality C_C , HITS score V_H , inverse of Burt's constraint score V_B ($m = 10^{-3}$) and PageRank score V_P ($m = 10^{-3}$).

No.	Drug	C_D
1	CID100004666	862
2	CID100005372	862
3	CID100644241	858
4	CID100005514	857
5	CID100005064	857
6	CID106442177	857
7	CID100000444	857
8	CID100005073	856
9	CID100005538	856
10	CID100005095	855

No.	Drug	C_B
1	CID106442177	919.407
2	CID100005372	629.140
3	CID100071273	591.049
4	CID100004634	574.990
5	CID100005538	563.390
6	CID100002474	524.419
7	CID100057469	490.400
8	CID100005408	464.443
9	CID100003032	462.043
10	CID100002764	449.465

No.	Drug	C_C
1	CID100004666	0.979
2	CID100005372	0.979
3	CID100644241	0.975
4	CID100005514	0.973
5	CID100005064	0.973
6	CID106442177	0.973
7	CID100000444	0.973
8	CID100005073	0.972
9	CID100005538	0.972
10	CID100005095	0.971

No.	Drug	V_H
1	CID100004666	1.000
2	CID100005514	1.000
3	CID100005064	1.000
4	CID100005073	1.000
5	CID100644241	0.999
6	CID100005372	0.999
7	CID106442177	0.999
8	CID100001690	0.999
9	CID100005095	0.999
10	CID100000444	0.999

No.	Drug	V_B (m)
1	CID100005372	4.686
2	CID100004666	4.691
3	CID100005538	4.698
4	CID106442177	4.705
5	CID100644241	4.705
6	CID100000444	4.708
7	CID100005064	4.710
8	CID100005514	4.710
9	CID100005073	4.714
10	CID100005095	4.716

No.	Drug	V_P (m)
1	CID106442177	1.424
2	CID100005372	1.392
3	CID100005538	1.376
4	CID100004634	1.372
5	CID100071273	1.371
6	CID100000444	1.362
7	CID100003032	1.360
8	CID100004666	1.358
9	CID100002474	1.354
10	CID100001690	1.351

For drug-indication network, as shown in Table 4.17 and Figure 4.3, each drug has an average degree score around 149 to 150 and there are more drugs having less degree score in the network. The closeness distribution for this network follows the characteristic of a scale free network with a bell-shaped slightly skewed to the left. The scale free network also mentioned the existence of few drugs with very high betweenness scores, normally called the hubs of the network.

Table 4.17: Summary of node measurements for drug indication network.

	auth	betweenness	burt	closeness	degree	pgrank
Mean	0.2298	761.6699	0.0573	0.4903	149.2612	0.0007
Std. Deviation	0.2156	1853.4502	0.1298	0.0572	131.7427	0.0005
Min	0.0001	0.0000	0.0048	0.3061	1	0.0001
At 25% pctl	0.0466	4.3307	0.0114	0.4551	38	0.0003
At 50% pctl	0.1607	116.7873	0.0177	0.4975	116	0.0006
At 75% pctl	0.3922	660.1730	0.0419	0.5263	229	0.0010
Max	1.0000	24556.0052	1.0000	0.6894	789	0.0035

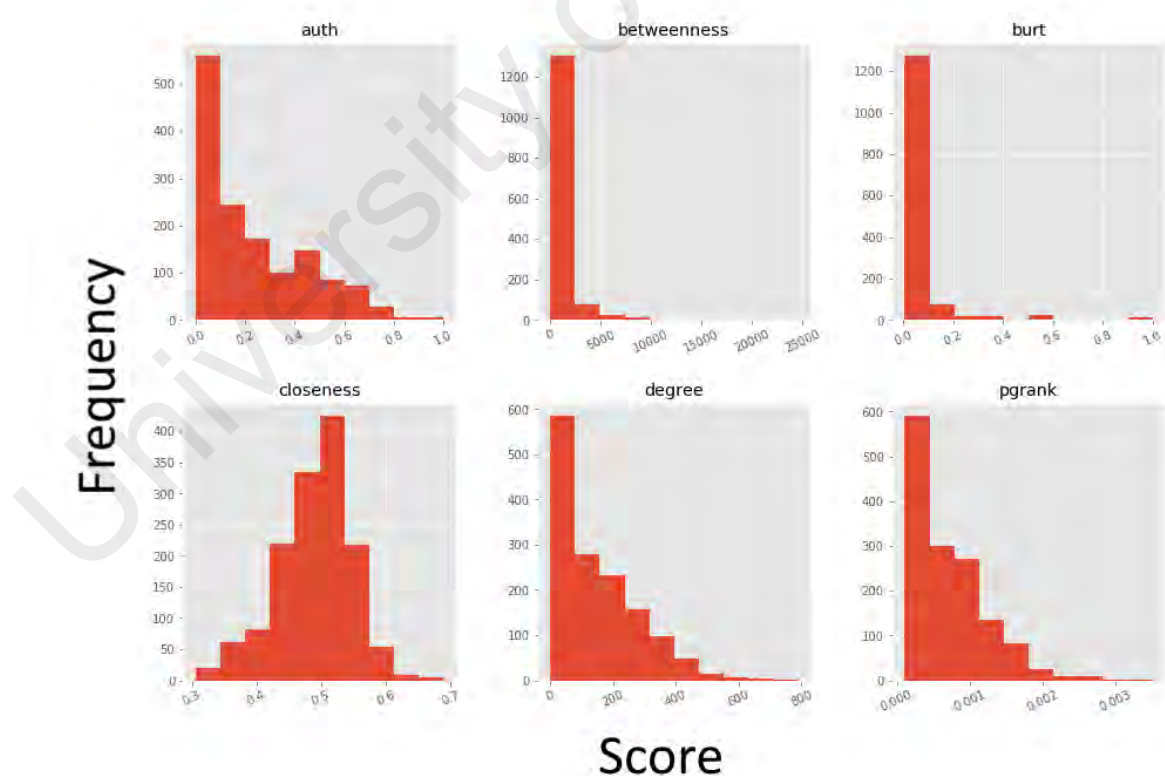


Figure 4.3: Visualization for Table 4.17

The list of top ten scorers for drug-indication network is shown in Table 4.18. Surprisingly, out of 1424 drugs in the drug-indication network, CID100003003 (Dexamethasone) achieved the highest score in all node measurements. For degree centrality, CID100003003 (Dexamethasone) obtained the highest score and connected to 789 drugs in the network leaving CID100004920 (Progesterone) in the second place. These two drugs also obtained the highest betweenness score which suggest that these drugs were located between groups of drug or therapy.

For closeness centrality score, CID100003003 (Dexamethasone) obtained the highest score at 0.689 which indicates that this drug requires the least path to relate with other drugs in drug-indication network. A highest HITS or authority score for CID100003003 (Dexamethasone) suggest that this drug acts as the hub of the network and may contain valuable source of information. This drug is also the least constrained drug in the network and obtained the highest score for PageRank.

4.3 Drug repositioning on top drugs

In this section, we discuss the results of applying network analysis into drug repositioning.

Predicted indication

Based on the network analysis result, we have identified 18 out of 882 drugs in drug-side effect network that appear at least once in the top ten lists of all six node scores. (See Table 4.19). Encouragingly, some of these drugs were successful in previous drug repositioning (Mehndiratta et al., 2016). On average, each of them exhibits 129 known side effects with common or very common frequencies.

Table 4.18: Top ten scorers in drug-indication network by node measurement. The scores measured are degree centrality C_D , betweenness centrality C_B ($m = 10^3$), closeness centrality C_C , HITS score V_H , inverse of Burt's constraint score V_B ($m = 10^{-3}$) and PageRank score V_P ($m = 10^{-3}$).

No.	Drug	C_D
1	CID100003003	789
2	CID100004920	715
3	CID100004900	702
4	CID100005372	687
5	CID100004168	640
6	CID100003032	620
7	CID100003325	606
8	CID100003640	605
9	CID100002367	601
10	CID100004159	598

No.	Drug	C_B (m)
1	CID100003003	24.556
2	CID100004920	21.974
3	CID100004900	17.342
4	CID100004253	15.481
5	CID100000838	14.694
6	CID100005372	14.068
7	CID100003640	13.589
8	CID100003325	12.816
9	CID100000450	11.898
10	CID100003032	11.535

No.	Drug	C_C
1	CID100003003	0.689
2	CID100004920	0.664
3	CID100004900	0.660
4	CID100005372	0.655
5	CID100004168	0.640
6	CID100003032	0.635
7	CID100003325	0.632
8	CID100003640	0.630
9	CID100004159	0.628
10	CID100002367	0.628

No.	Drug	V_H
1	CID100003003	1.000
2	CID100005372	0.999
3	CID100004920	0.963
4	CID100004168	0.950
5	CID100003032	0.916
6	CID100003440	0.888
7	CID100004900	0.851
8	CID100002367	0.842
9	CID100002022	0.839
10	CID100004583	0.836

No.	Drug	V_B (m)
1	CID100003003	4.842
2	CID100004920	4.924
3	CID100004900	5.031
4	CID100005372	5.202
5	CID100003325	5.230
6	CID100004168	5.400
7	CID100004253	5.436
8	CID100003032	5.512
9	CID100003640	5.598
10	CID100004159	5.627

No.	Drug	V_P (m)
1	CID100003003	3.511
2	CID100004900	3.084
3	CID100004920	3.067
4	CID100003640	2.777
5	CID100005372	2.769
6	CID100004253	2.678
7	CID100003325	2.650
8	CID100004159	2.649
9	CID100003032	2.583
10	CID100000838	2.547

Table 4.19: Drugs in the drug-side effect network that appear at least once in the top ten lists of all six node scores.

Drug	Name	Frequency of Appearance
CID106442177	Everolimus	6
CID100005372	Tacrolimus	6
CID100000444	Bupropion	5
CID100004666	Paclitaxel	5
CID100005538	Retinoic acid	5
CID100005514	Topiramate	4
CID100005064	Ribavirin	4
CID100005073	Risperidone	4
CID100644241	Nilotinib	4
CID100005095	Ropinirole	4
CID100004634	Oxybutynin	2
CID100003032	Diclofenac	2
CID100001690	Doxorubicin	2
CID100002474	Bupivacaine	2
CID100071273	Ropivacaine	2
CID100002764	Ciprofloxacin	1
CID100005408	Testosterone	1
CID100057469	Imiquimod	1

Based on our proposed algorithm, we have chosen some of these prominent drugs for drug repositioning. The indications of the five prominent drugs predicted based on their side effects are summarized in Table 4.20. We will briefly describe the top drug that obtained the highest node score in drug-side effect network.

Table 4.20: Indications of five prominent drugs predicted based on the drug-side effect network.

Drug	Name	Top Predicted Indications
CID106442177	Everolimus	Acquired immunodeficiency syndrome, major depressive disorder, epilepsy, diabetic, Parkinson's disease
CID100005372	Tacrolimus	Malignant neoplasm of breast, liver diseases, epilepsy, major depressive disorder, decreased interest
CID100000444	Bupropion	Malignant neoplasms, renal insufficiency, liver diseases, diabetes mellitus, epilepsy
CID100004666	Paclitaxel	Liver diseases, diabetes mellitus, epilepsy, major depressive disorder, lymphoma
CID100005538	Retinoic acid	Neoplasms, malignant neoplasm of breast, hepatic impairment, epilepsy, major depressive disorder

Everolimus

CID106442177 or everolimus (Afinitor[®]) is one of the antineoplastic agent with two therapeutic codes i.e., L01XE10 (Protein kinase inhibitors) and L04AA18 (Selective immunosuppressants). It appears in the top ten lists of all measurements. In particular, it tops the betweenness scores whereby a high score of which indicates a potential to treat multiple diseases or to reside in different therapeutic groups. Everolimus is one of the derivatives of sirolimus and currently has 15 known indications and a total of 375 known side effects. One indication of everolimus resulting from our prediction based on the drug-side effect network is C0014544 (Epilepsy). Incidentally, we found that it has been studied as a possible treatment for epilepsy these few years (Krueger et al., 2013; Miller, 2014). Some other indications predicted from the drug-side effect network are C0001175 (Acquired Immunodeficiency Syndrome or AIDS) and C0241863 (Diabetic) (See Table 4.21).

Table 4.21: Top five predicted indications based on first neighbours for everolimus. Concept Unique Identifier (CUI) is the UMLS concept ID from SIDER dataset. Frequency is the number of redundancies based on the prediction list from Step 4. Neighbour frequency is the frequency divided by the total number of neighbours for Everolimus in the drug-side effect network. Concept name is the given indication name based on CUI. Number of clinical studies is the number of records found in ClinicalTrials.gov between drug and concept name (ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US), n.d.)

CUI	Freq.	Neigh. Freq	Concept Name	No. of Clinical Studies
C0001175	17	7.943925	AIDS	5
C0014544	13	6.074766	Epilepsy	5
C0241863	13	6.074766	Diabetic	6
C0030567	13	6.074766	Parkinson Disease	0
C0023418	12	5.607477	Leukemia	859

Predicted side effect

Based on the network analysis result, we have identified 15 prominent drugs that appear at least once in the top ten lists of all six node scores (See Table 4.22). As in the case of

predicting indications, we summarized the predicted side effects of the five prominent drugs in Table 4.23. We will briefly describe the top drug that obtain the highest node score in drug-indication network.

Table 4.22: Drugs in the drug-indication network that appear at least once in the top ten lists of all six node scores.

Drug	Name	Frequency of Appearance
CID100003003	Dexamethasone	6
CID100003032	Diclofenac	6
CID100004900	Prednisone	6
CID100004920	Progesterone	6
CID100005372	Tacrolimus	6
CID100003325	Famotidine	5
CID100003640	Cortisol	5
CID100004159	Methylprednisolone	4
CID100004168	Metoclopramide	4
CID100002367	Dexamethasone sodium phosphate	3
CID100004253	Morphine	3
CID100000838	Epinephrine	2
CID100000450	Estradiol	1
CID100002022	Acyclovir	1
CID100003440	Furosemide	1
CID100004583	Ofloxacin	1

Table 4.23: Side effects of five prominent drugs predicted based on the drug-indication network.

Drug	Name	Top Predicted Side Effects
CID100003003	Dexamethasone	Hypersensitivity, pain, dyspnea, arthralgia, asthenia
CID100003032	Diclofenac	Neutropenia, agitation, connective tissue diseases, muscle weakness, hypokalemia
CID100004900	Prednisone	Thrombocytopenia, dyspnea, pain, hypotension, leukopenia
CID100004920	Progesterone	Thrombocytopenia, leukopenia, Stevens-Johnson syndrome, erythema, agranulocytosis
CID100005372	Tacrolimus	Angioedema, eosinophilia, erythema multiforme, upper respiratory infections, erectile dysfunction

Dexamethasone

CID100003003 or dexamethasone such as Decadron[®] is a corticosteroid useful to treat different categories of inflammatory and autoimmune conditions. According to SIDER 4.1,

it has 172 unique indications and is listed in 22 different ATC groups. It is worth noting that dexamethasone tops all six node scores in the drug-indication network. Our predictions indicate that some unknown side effects of this drug are C0020517 (Hypersensitivity), C0013404 (Dyspnea) and C0003862 (Arthralgia). It has been reported that out of 55,367 cases of side effects reported by patients during their intake of dexamethasone, 3,821 or 5.87% of the cases were associated with dyspnea (EHealthMe.com, n.d.). However, it is worth noted on the contrary that a recent clinical study showed that dexamethasone was significantly associated with the reduction of dyspnea in cancer patients (Hui et al., 2016).

CHAPTER 5: CONCLUSION

Summary of findings

In this thesis, we have used network analysis to understand drugs from the perspectives of drug side effects and indications. Two drug networks were constructed, which are drug-side effect network and drug-indication network, using SIDER datasets. Drugs in the drug-side effect network were linked based on side effect similarity with either common or very common frequencies. Meanwhile for drug-indication network, drugs were linked based on indication similarity. These networks were analyzed using some of the measurements in network analysis to obtain the properties of the network. Based on the network properties, we have shown the possibility to apply network analysis in drug repositioning by proposing an algorithm to predict potentially new uses of drugs.

From side effect similarity perspective, drugs tend to share at least 1 similarity with other drugs due to common side effects such as communicable diseases, pain and nausea. This situation leads to a dense drug-side effect network where most of the drugs manifest high degree centralities. However, from indication similarity perspective, drug-indication network have more nodes with relatively lower degree centralities. This suggests that most indications have only a few drug alternatives.

We have obtained a few sets of prominent drugs which are the top scorers of node measurements in both networks. Interestingly, some of the prominent drugs had been reported to be successful in yielding new uses (Mehndiratta et al., 2016). In principle, we could use any drugs in the network for drug repositioning. However, in the context of our work, for practicality, we use only the selected prominent drugs as it is plausible they could have a better practicality for drug repositioning. We have validated our predictions through the list of clinical studies from ClinicalTrials.org and the list of reported side effects from

eHealthMe.com. Some of our predictions were even already in early clinical trials phases (Krueger et al., 2013; Miller, 2014).

Limitation of study

There are a number of limitations in this study. The first limitation comes from the drug side effects. It is known that each person using a drug may experience different side effects due to factors such as age, gender, lifestyle, dosage or even genetic factors. This might affect the predictability of the proposed algorithm. The second limitation comes from the over-reporting of side effects which may disrupt our side effects validations. The third limitation of this study is due to the old drugs in the dataset. The end users of the drugs tend to use newer drugs which may have better advantages than the old drugs.

Suggestion for future study

There are few areas that can be improved in our studies :

- Create specific measurements for drug network - The measurements that we used are basic network analysis measurements and creating measurements specifically for drug network may improve our understandings on drugs.
- Create a better similarity score - Our similarity score is based on Jaccard indexes for shared phenotypic profiles and shared neighbours. An improved scoring mechanism to quantitatively describe the similarity between drugs will be definitely useful in the future. Adding new edge attributes might also benefit the scoring mechanism.
- Create a better prediction and validation algorithm - The proposed prediction algorithm were mainly based on guilt-by-association method between chosen drugs and its neighbouring nodes. Meanwhile, the validations were only based on existing clinical studies and reported side effects. Prediction based on chosen drugs and

non-neighbouring nodes might be interesting and advanced validation such as k-fold cross validation might be needed for future drug repositioning.

We hope that this work will help researchers especially from chemistry or pharmaceutical background to understand drugs from network perspectives.

University of Malaya

REFERENCES

- Adams, C. P., & Brantner, V. V. (2006). Estimating the cost of new drug development: is it really \$802 million? *Health Affairs*, 25(2), 420-428.
- Ahuja, R. K., Magnanti, T. L., & Orlin, J. B. (1993). *Network flows: Theory, algorithms, and applications*. New Jersey: Pearson.
- Ashburn, T. T., & Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8), 673-683.
- Barabási, A.-L., Gulbahce, N., & Loscalzo, J. (2010). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1), 56–68.
- Baur, M., Brandes, U., Lerner, J., & Wagner, D. (2009). *Group-level analysis and visualization of social networks*. Berlin, Heidelberg: Springer, Berlin, Heidelberg.
- Bell, M. G. H., & Iida, Y. (1997). *Transportation network analysis*. United Kingdom: John Wiley & Sons, Ltd.
- Berger, S. I., & Iyengar, R. (2009). Network analyses in systems pharmacology. *Bioinformatics*, 25(19), 2466–2472.
- Biggs, N., Lloyd, E. K., & Wilson, R. J. (1986). *Graph theory, 1736-1936*. New York: Oxford University Press.
- Bollobás, B. (2013). *Modern graph theory* (Vol. 184). New York: Springer Science & Business Media.
- Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, 323(5916), 892–895.
- Brandes, U., & Erlebach, T. (Eds.). (2005). *Network analysis: Methodological foundations*. Berlin: Springer Science & Business Media.
- Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology*, 110(2), 349-399.
- Chiang, A. P., & Butte, A. J. (2009). Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clinical Pharmacology & Therapeutics*, 86(5), 507.
- ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). (n.d.). *Clinical studies on everolimus*. Retrieved on 31 October 2016 from <https://clinicaltrials.gov/ct2/results/browse?intr=everolimus>.
- Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research [Journal Article]. *InterJournal, Complex Systems*, 1695(5), 1-9.

- Danhof, M. (2016). Systems pharmacology – towards the modeling of network interactions. *European Journal of Pharmaceutical Sciences*, 94, 4–14.
- DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of r&d costs. *Journal of Health Economics*, 47, 20–33.
- Dudley, J. T., Deshpande, T., & Butte, A. J. (2011). Exploiting drug–disease relationships for computational drug repositioning. *Briefings in Bioinformatics*, 12(4), 303311.
- Duran-Frigola, M., & Aloy, P. (2012). Recycling side-effects into clinical markers for drug repositioning. *Genome Medicine*, 4(3), 3.
- EHealthMe.com. (n.d.). *Dexamethasone and shortness of breath - from fda reports*. Retrieved on 31 December 2016 from <https://www.ehealthme.com/ds/dexamethasone/shortness-of-breath/>.
- Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3), 215–239.
- Hodos, R. A., Kidd, B. A., Shameer, K., Readhead, B. P., & Dudley, J. T. (2016). In silico methods for drug repurposing and pharmacology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 8(3), 186–210.
- Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology*, 4(11), 682.
- Huang, L., Li, F., Sheng, J., Xia, X., Ma, J., Zhan, M., & Wong, S. T. C. (2014). Drugcomboranker: drug combination discovery based on target network analysis. *Bioinformatics*, 30(12), i228–i236.
- Hui, D., Kilgore, K., Frisbee-Hume, S., Park, M., Tsao, A., Guay, M. D., . . . Eapen, G. (2016). Dexamethasone for dyspnea in cancer patients: A pilot double-blind, randomized, controlled trial. *Journal of Pain and Symptom Management*, 52(1), 8–16. e1.
- Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaekar, P., Ferriero, R., . . . others (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, 107(33), 14621–14626.
- Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE*, 9(6), e98679.
- Kidd, B. A., Wroblewska, A., Boland, M. R., Agudo, J., Merad, M., Tatonetti, N. P., . . . Dudley, J. T. (2016). Mapping the effects of drugs on the immune system. *Nature Biotechnology*, 34(1), 47–54.
- Kim, S., Thiessen, P. A., Bolton, E. E., & Bryant, S. H. (2015). Pug-soap and pug-rest:

- web services for programmatic access to chemical information in pubchem. *Nucleic Acids Research*, 43(W1), W605-W611.
- Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., . . . Bryant, S. H. (2016). Pubchem substance and compound databases. *Nucleic Acids Research*, 44(D1), D1202-D1213.
- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604-632.
- Kranakis, E. (2012). *Advances in network analysis and its applications*. Berlin: Springer, Berlin, Heidelberg.
- Krueger, D. A., Wilfong, A. A., Holland-Bouley, K., Anderson, A. E., Agricola, K., Tudor, C., . . . Franz, D. N. (2013). Everolimus treatment of refractory epilepsy in tuberous sclerosis complex. *Annals of Neurology*, 74(5), 679-687.
- Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2015). The sider database of drugs and side effects. *Nucleic Acids Research*, 44(D1), D1075-D1079.
- Kuhn, M., Szklarczyk, D., Pletscher-Frankild, S., Blicher, T. H., von Mering, C., Jensen, L. J., & Bork, P. (2014). Stitch 4: integration of protein–chemical interactions with user data. *Nucleic Acids Research*, 42(D1), D401-D407.
- Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., . . . Ross, K. N. (2006). The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795), 1929-1935.
- Langedijk, J., Mantel-Teeuwisse, A. K., Slijkerman, D. S., & Schutjens, M.-H. D. B. (2015). Drug repositioning and repurposing: terminology and definitions in literature. *Drug Discovery Today*, 20(8), 1027-1034.
- Lussier, Y. A., & Chen, J. L. (2011). The emergence of genome-based drug repositioning. *Science Translational Medicine*, 3(96), 96ps35-96ps35.
- Mehndiratta, M. M., Wadhwa, S. A., Tyagi, B. K., Gulati, N. S., & Sinha, M. (2016). Drug repositioning. *International Journal of Epilepsy*, 3(2), 91-94.
- Miller, J. W. (2014). Treating epilepsy in tuberous sclerosis with everolimus: Getting closer. *Epilepsy Currents*, 14(3), 143-144.
- Nacher, J. C., & Schwartz, J.-M. (2008). A global view of drug-therapy interactions. *BMC Pharmacology*, 8(1), 5.
- Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2), 167–256.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web* (Tech. Rep.). Stanford InfoLab.

- Spiro, Z., Kovacs, I. A., & Csermely, P. (2008). Drug-therapy networks and the prediction of novel drug targets. *Journal of Biology*, 7(6), 1.
- Wang, F., Zhang, P., Cao, N., Hu, J., & Sorrentino, R. (2014). Exploring the associations between drug side-effects and therapeutic indications. *Journal of Biomedical Informatics*, 51, 15-23.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). New York: Cambridge university press.
- Wu, Z., Wang, Y., & Chen, L. (2013). Network-based drug repositioning. *Molecular BioSystems*, 9(6), 1268-1281.
- Ye, H., Liu, Q., & Wei, J. (2014). Construction of drug network based on side effects and its application for drug repositioning. *PLoS ONE*, 9(2), e87864.
- Ye, H., Yang, L., Cao, Z., Tang, K., & Li, Y. (2012). A pathway profile-based method for drug repositioning. *Chinese Science Bulletin*, 57(17), 2106-2112.
- Zhou, X., Menche, J., Barabási, A.-L., & Sharma, A. (2014). Human symptoms–disease network. *Nature Communications*, 5, 4212.
- Zweig, K. A. (2016). Graph theory, social network analysis, and network science. In *Lecture notes in social networks* (pp. 23–55). Vienna: Springer Vienna.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Mohd Ali, Y. E., Kwa, K. H., & Ratnavelu, K. (2017). Predicting new drug indications from network analysis. *International Journal of Modern Physics C*, 28(09), 1750118.

University of Malaya