RELATIONSHIP BETWEEN PAYMENTS MADE TO PHYSICIANS BY HEALTHCARE COMPANIES AND THEIR RETURNS

Master's Thesis

Chitra Balasubramanian

@2020, Chitra Balasubramanian. All rights reserved.

RELATIONSHIP BETWEEN PAYMENTS MADE TO PHYSICIANS BY HEALTHCARE COMPANIES AND THEIR RETURNS

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in COMPUTER SCIENCE
track Data Science and Technology

at the DELFT UNIVERSITY OF TECHNOLOGY

by

Chitra Balasubramanian

4742907

Thesis Committee: Prof. Dr. M.A. Larson, EEMCS, Delft University of Technology
Dr. C. Hauff, EEMCS, Delft University of Technology
Dr. H. Wang, EEMCS, Delft University of Technology (supervisor)

Multi Media Computing Group
Faculty of Electrical Engineering, Mathematics and Computer Science,
Delft University of Technology
Delft, Netherlands
http://mmc.tudelft.nl/

Preface

This thesis has been a wonderful journey of finding my passion and discovering my potential. And for this thesis to reach completion would have been not possible without assistance, guidance and faith of many people. For this first, I would like to thank Royal Philips, Amsterdam, for giving me the opportunity to pursue my thesis. The business problem provided by them was not only interesting but also challenging in various manners. I extend gratitude to my supervisor and committee members for their constructive feedback, without which this thesis would have not reached its zenith. Next, I would like to thank all the members of my research group (Multimedia Computing) for their support and contribution. I would also like to thank DAF, for all the support and confidence they had in me.

To my mom, a constant guide, a relentless mentor, I dedicate my thesis to my mom. She has been whiteboard on which I chalked all my ideas. Thank you is not enough to express the amount of gratitude and love I would like to express for her unfathomable belief in me. Finally, i would like to thank Husku, Acchu and Sid for being there for me, in my thesis journey.

Chitra Balasubramanian

Abstract

Healthcare industry is an ever-emerging field in the 21st century. The statistics from Centers for Medicare services (CMS) [15] website shows that in 2017, in USA, the healthcare industry has invested USD 7.4 billion for research collaborations with physicians. These research collaborations in CMS is spread across multiple facets ranging from contributing towards research, developing new products, running clinical trials, royalty, licences, patents, providing innovative ideas etc. In this thesis, we make an assumption that, a relationship between investment made by the healthcare company and the research profile of a physician exists. We aim to answer, what could possibly be the relation between payments made by the healthcare company; on the physicians and the research profile of the physicians. The research profile of a physician includes factors like h-index, years of research experience, citation count, physician citation network, etc. To validate this relationship we use the data corresponding to returns of the healthcare company. Some of the measures of returns, from a research collaboration between physicians, include, innovation, good will, fame, market share, Return on Investment (ROI). We choose ROI, as a measure of return, due to the availability of data and to determine the relationships mentioned above.

To understand the above mentioned relationship, we explore two types of relationships, i.e., direct and indirect relationship. In the direct relationship, we use multiple regression model to understand the direct relationship between the research payments and the research profile of the physicians, by making an assumption that the research profile of the physician describes the research quality of the physician. In the indirect relationship, we make use of a weighted physician co-author citation network, to investigate the relationship between his/her co-author interactions and the research payments he/she received from the healthcare company. To accomplish this, we developed a spreading process that models influence diffusion in a physician citation network. The diffusion of influence is dependent on the topological property of the node in the network.

Our models are an exemplification of the direct and indirect relationships, which exists in the real world. To evaluate our models, we use metrics such as coefficient of determination, Pearson correlation coefficient and Spearman's rank correlation. Once the models were evaluated, we inferred that the model for indirect relationship, explains the relationship between research profile, investments, and return 96.3% more than the model for direct relationship. We also perform a deep analysis, by investigating the nature of the distributions of the variables and scatter plots to understand the relationship between the variables used in understanding the direct and indirect relationship. Lastly, we propose two different redistribution methods, where the original payments made to physicians are redistributed to a potential group of physicians in the physician citation network. These potential physicians are identified based on their topological property. In consequence, our redistribution methods may inspire the healthcare companies, to design their future investments made to physicians.

Thesis Committee: *Prof. Dr. M.A. Larson, EEMCS, Delft University of Technology*
Dr. C. Hauff, EEMCS, Delft University of Technology
Dr. H. Wang, EEMCS, Delft University of Technology (supervisor)

Contents

LIST OF TABLES

LIST of FIGURES

LIST OF ACRONYMS

ROI	- Return on Investment
CROI	- Coauthoring Return on Investment
PROI	- Publication Return on Investment
EBM	- Evidence Based Medicine
API	- Application Program Interface
RESET	- Regression Specification Error Test
HCCM	- Heteroscedasticity Consistent Covariance Matrix
CMS	- Centers for Medicare and Medicaid services
CRM	- Customer Relationship Management
BLUE	- Best Linear Unbiased Estimate
OLS	- Ordinary Least Squares

LIST OF SYMBOLS

(for chapter 4)

Symbol	Meaning
x	independent/explanatory variable
y	dependent variable
z	regressors
ϵ	random disturbance
α	regression parameter (intercept of the regression line)
β	regression parameters (slope)
γ	regression parameters
i	to represent a single observation
j	to represent a single explanatory variable
g	degrees of freedom
$E()$	mean value/expected value
n	total number of observations
m	total number of explanatory variables
X	Explanatory variable matrix
h_{ii}	diagonal elements of the Hessian matrix ()
e_i	estimated error term of the regression equation
$s_{(i)}$	root mean square error, excluding the x_i^{th} observation
s^2	mean square error of regression
$D_i^2(s)$	Cook's distance
F	F statistics
$\hat{y}$	estimated value of y
H_0	null hypothesis
χ^2	Chi-Square statistics
P	P-value (measure of significance)
R^2	coefficient of determination

LIST OF SYMBOLS

(for Chapter 5)

Symbol	**Meaning**
G	- Physician citation network
N	- set of nodes in the physician citation network
L	- set of links in the physician citation network
W	-set of weights (link strength) in the physician citation network
n_k	-node in a network
P_k	-payment attribute of the node n_k
i	-total number of nodes
j	- total number of links, weights
m, k	- a number between 1 and i, used to represent a node
S_k	- topological property of the node
Source	- nodes that receive a payment from healthcare company ABC
p	- total number of source nodes in the physician citation network
q	- remaining nodes in the physician citation network
x	- node in set O
y	- node in set L
$set\ O$	- set of nodes that contain influence
$set\ V$	- set of nodes that are one hop neighbour distance from an element in set L
θ	- spreading parameter
t	- represents the current stage of the spreading process
α	- scaling parameter
H	-set of hospitals/affiliations of N
a	- total number of hospitals

1. INTRODUCTION

Healthcare industry is an ever-emerging field in the 21st century. Over the last decade, we have seen many innovations and advancements in the technology pertaining to healthcare. This leads to an enormous amount of investments towards research and development of healthcare technology. Statistics obtained from Centers for Medicare services (CMS) [15], a U.S. government website with open data, show that in 2017, in USA, the healthcare industry has invested USD 7.4 billion to collaborate with physicians. This collaboration is spread across multiple facets ranging from contributing towards research, developing new products, running clinical trials etc. The amount invested on research and development contribute to USD 4.6 billion. In the same year, Healthcare company ABC, invested about USD 7.4 million towards research payments across 10,000 physicians [15]. The CMS/Open Payments Data makes the payments data available to the patients, analysts, physicians, citizens of the U.S.A for various purposes like transparency, fairness, equality etc. This website also classifies the research payments based on the nature of research payments. For healthcare company ABC, the nature of research payments is towards funding for a new study, coordination and implementation of existing research, royalty, licences, patents, providing innovative ideas, etc.

1.1 Motivation

There is a lack in understanding, if the enormous investment (60-65% of total investment) made by the healthcare industries, towards physicians for research purposes, is meaningful and worthwhile. This knowledge gap drives the motivation of this thesis.

1.2 Problem Statement

To address the motivation in Section 1.1, we aim to understand, what the relationship between the investment made by the healthcare companies on physicians, research profile[1] of these physicians and the different types of return. In this thesis, we make an assumption that there exists a relationship between payments made to physicians, the research profile of the physician and the returns on investment and this relationship is unknown. The motivation behind using the research profile of the physician is, it describes the research quality of the physician. It includes factors like h-index, years of research experience, citation count, physician citation network, etc., which is elucidated in Chapter 3.

The motivation behind using the return, is to validate the outcomes of the relationship between research payments and the research profile of the physicians. There are different types of return to a healthcare company, some of which are, being innovation leaders, good will, fame, being research leaders, ROI (Return on Investment) etc. We choose ROI because it is measurable, data is available and to illustrate the relationships mentioned above.

To estimate the relationship mentioned above, we propose two types of relationship which is an imitation of the real world scenarios. First, the direct relationship, where physicians interact with healthcare companies for research, and the outcomes of the research generate ROI to the healthcare company.

Second, the indirect relationship, where physicians who interact with the healthcare company for research, also collaborate with other physicians, who do not receive any research investment from the healthcare company. The other physicians have research collaborations with physicians who receive investment from a healthcare company, in a research collaboration network. This could lead to the other physicians getting inspired by research or having a research collaboration. The outcome of this collaboration generates ROI from multiple sources to the healthcare company. These sources are hospitals, clinics or universities. Since, the return does not directly come from the physician on whom the healthcare company has a collaboration, we call this relationship indirect. The indirect relationship is represented in Figure 1.1, where a healthcare company interacts with a physician for a research collaboration, indicated as Phy1, and this physician in turn has collaborations with other physicians in a citation network, represented by Phy2-Phy10. The outcome of this collaboration generates multiple returns from hospitals, universities and clinics.

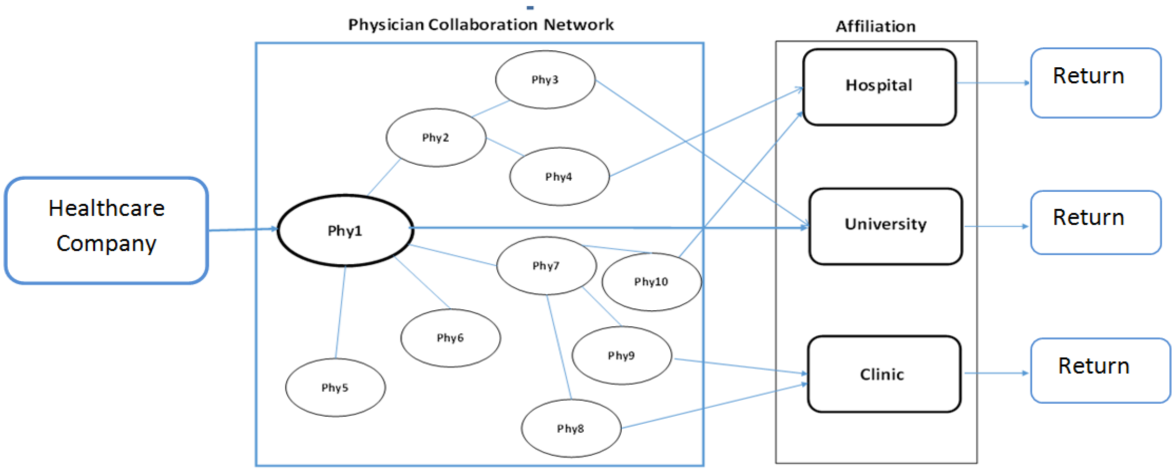


Figure 1.1: A representation of indirect relationship

After understanding the direct and indirect relationship, ***the problem statement is to find a relationship between the payments made by the healthcare company, the return on investment and the research profile of the physicians.***

1.3 Research Questions

From the problem statement mentioned in section 1.2, the following research questions are constructed.

RQ1: What is the relationship between investments on physicians, ROI generated by the healthcare company and the research profile of a physician on whom the healthcare company invests?

To answer this research question, two types of relationships have been proposed in this thesis. First, a direct relationship, which makes use of a machine learning model, known as regression technique. Second, an indirect relationship, where we designed a network spreading process that models how influence diffuses into a physician citation network. This spreading process is evaluated to explain the effectiveness of spread of influence on the ROI generated to the healthcare company.

RQ2: Do the three healthcare companies follow a strategy while making payments?

Two other peer companies were considered to compare the payment strategy of healthcare company ABC with other companies in the healthcare industry. We attempt to answer RQ2, by performing Exploratory Data Analysis to estimate what are the research profile metrics that drive physician payments.

RQ3: What is the effect of redistribution methods on the relationship between investment and its return?

We propose two alternative payment redistribution methods in order to understand how the payment redistribution method affects the properties of the resultant nodal influence and their relationship between investment and nodal influence. We also investigate the relationship between investment and its return after the redistribution of payments to understand the effect of the redistribution methods.

1.4 Contributions

This section highlights the key findings from this thesis.

- Captured the, direct and indirect relationship, between the investment made to physicians, ROI generated by the healthcare company and the research profile of a physician. We estimated the direct relationship by a recursive regression technique and we estimated the indirect relationship by modelling influence diffusion using a network spreading process.
- We concluded that the indirect relationship can better explain the relationship between payments, research profile and ROI.
- We developed network topological metrics based on topological properties that imitate real world influence spreading in a physician citation network.
- Proposed a novel spreading process, which is used to model the diffusion of influence.
- Decoded the payment strategy of the healthcare company ABC and its counter healthcare companies.
- Developed a payment redistribution method which recommends different ways of investments to physicians in the future.

1.5 Thesis outline

This thesis is structured in the following manner, chapter 2 focuses on the background study and terminologies that are required to understand the remaining chapters. Chapter 3 focuses upon data understanding, preliminary data exploration and analysis. Chapter 4 explains the regression analysis technique, outlier detection and elimination, effectiveness of regression model and the relationship between ROI and investment for healthcare company ABC. Chapter 5 explains the design of network metrics based on topological measures and the spreading process which is used to model influence. Chapter 6 discusses the evaluation of the spreading process discussed in chapter 5 by conducting experiments to evaluate the best fit parameters for the spreading process. Chapter 7 focuses if there are other ways in which we can invest on physicians research by developing and analysing two redistribution methods. Chapter 8 provides conclusions, limitations that were overcome in this thesis and future recommendations.

2. BACKGROUND and LITERATURE SURVEY

In this chapter, the required background knowledge to understand the design and implementation of a machine learning model and network spreading process is discussed. Section 2.1 provides a brief of the machine learning approach used in this thesis. This is followed by section 2.2, which introduces the network science and the main terminologies used for the same. Further, topological measures of network are discussed in section 2.3. This is followed by section 2.4 which describes the nature of power law distributions as the degree distribution of the physician citation network follows power law. The last section 2.5 consists of related work that supports the research involved in this thesis.

2.1 Machine Learning Approach

One of the most important and broadly used machine learning and statistical tool is the regression technique. Regression analysis helps in understanding the relationship between an outcome variable with one or more response or predictor variables. It can be represented as a mathematical equation, that defines y (output variable) as a function of x (input variables). If this relationship is linear, it is termed as a linear regression model. In some situations, the relationship between the outcome and the predictor variables may not be linear, in such cases a non linear regression model such as a polynomial regression model is necessary to understand the relationship. A regression coefficient is estimated to understand the relationship between outcome and predictor variables, using Ordinary Least Squares (OLS) technique. In order to obtain estimates which satisfies all the required statistical properties, such as Best Linear Unbiased Estimate (BLUE), diagnostic tests are performed. If the tests detect the presence of heteroscedasticity or specification bias, then accordingly appropriate remedial measures are used such that the estimates do not suffer from the lack of these statistical properties. A detailed explanation of the diagnostic tests, remedial measures etc, are explained in chapter 4.

2.2 Network Science Approach

2.2.1 Terminologies

This section discusses a set of terminologies that are required for further understanding:

2.2.1 Network

A network $G = (N, L)$ is a collection of nodes that are connected together where, nodes are represented as $N = \{n_1, n_2, n_3, \dots., n_i\}$ and i is the total number of nodes in G. These nodes are connected to each other with a set of links, represented as $L = \{l_1, l_2, l_3, \dots., l_j\}$, where j is the total number of links in G. A network can be categorized based on its homogenous and heterogeneous nature of nodes [18] where in a homogenous network all nodes in the network contain the same attribute, for example, in an airport network all nodes are airports. On the other hand a heterogeneous network can carry an assortment of nodes such as hospitals, physicians, healthcare companies etc.

2.2.3 Social Network Analysis

Social Network Analysis is a study that maps the relationship between people, organizations, computers, URLs or any set of entities that are connected to one another. The nodes in the network are representations of people, organizations etc and the links in the network show relationships or flows between these nodes [22]. The physician citation network is a representation of a social network of physicians where the nodes represent physicians and the links represent the research collaboration between them. Hence, the physician citation network can be viewed from a socio-centered perspective [21]. Here we look for ties between nodes that indicate cohesive social groups amongst physicians that reflect social interactions and social stratification amongst these groups.

2.3 Topological Measures of Networks

Topology refers to the manner in which nodes and links are arranged in a network. Topological measures are used to capture various properties mainly the distance, spectra and connection [44]. This section explains three main topological measures which are used to develop network topological metrics in Chapter 5 of this thesis.

2.3.1 Degree of a node

The degree $d_{k,}$ for a node n_k, is the total number of links a node possesses in a network [20, 25]. This is one of the fundamental properties of a network which is used in the computation of other metrics. When the network is a weighted network, then the weighted degree of a node is the sum of weights of all the edges of the node and is represented as wd_k, further in this thesis. The weighted degree of a node is also called the strength of the node in the network.

2.3.2 Hopcount between two nodes

The shortest hopcount $H_{m\to k}$ between any two nodes n_m and n_k is the minimum number of hops required to reach n_k from n_m .

2.3.3 Closeness Centrality of a node

Closeness Centrality of a node is a measure of how close a node is to other nodes in the network. A closeness centrality is often noted as the reciprocal of hopcount of a node [44]. It is represented in Eq. (1)

$$C_k = \frac{1}{\sum_{m\in G} d(m,k)} \tag{1}$$

Where, n_m and n_k are two nodes in the network G and $d(m,k)$ is the distance (number of citations in context to physician citation network) between the two nodes.

2.3.4 Clustering Coefficient of a node

The clustering coefficient, CC_k , of a node, n_k , is a measure of the degree, d_k, to which nodes in a graph cluster together. The coefficient is a real number lying between 0 and 1, where higher the coefficient higher the tendency to cluster together. It can be mathematically represented in Eq. (2):

$$CC_k = \frac{2l_k}{d_k(d_k-1)} \tag{2}$$

Where, d_k is the degree of the node and l_k is the number of links between nodes within the neighbourhood of a node n_k .

2.4 Power law distribution

Degree distribution of a node captures the difference in degree connectivity of a network. It was observed that the physician payment network follows a power law distribution. The power law degree distribution is represented in Eq. (3), where i represents the number of nodes that a given node n_k is connected to other nodes and γ (degree exponent) represents a scale who's value lies between $2 < \gamma < 3$ in most cases and very rarely the values go out of scale [23, 24]. According to Alert-Laszlo Barabasi, in his book Network Science, “a scale-free network is a network whose degree distribution follows a power law". These networks are commonly observed in networks such as the airport traffic network, railway network, social media network etc.

$$P(i) \sim i^{-\gamma} \tag{3}$$

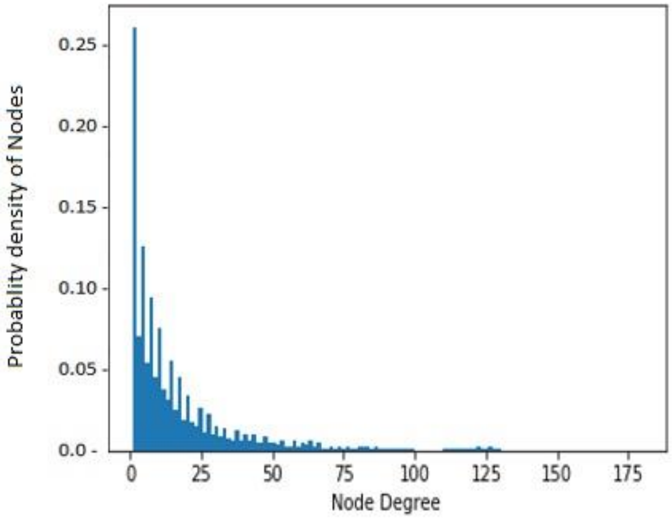


Figure 2.1: Degree distribution of a Scale free physician citation network

Figure 2.1 shows the degree distribution of the physician citation network that follow a power law distribution. In such networks, it can be noted that some nodes have a much higher degree than other nodes in the network. In Figure 2.1, there are some nodes in the network that have a node degree greater than 100, but most nodes have a degree between 0-25. The nodes with a higher degree are mostly the hub nodes which on removal will disconnect the network, hence making them the most crucial nodes in the network. We could imagine this property in a physician citation network as the hub nodes represent the most influential physicians and the periphery nodes represent not so influential physicians.

2.5 Related Work

There are different ways to determine the research profile of a physician based on the healthcare requirement. Current day research [34, 35] associates a physician's profile, to their ability to provide patient centric care. This is indeed one of the primary attributes of a physician but other behaviors of a physicians such as participation in research, is worthy of understanding the physician's importance [36].

The article by Daniele et al. [32] aims to understand the relationship between the professional collaboration of the physician and evidence based medicine (EBM) [32]. EBM assists in the decision making process of medical practices based on evidence and research. They made use of a collaboration network of physicians to obtain a set of core and peripheral nodes in a network by using network centrality metrics. Results obtained indicated that the core nodes in the network are negatively associated with EBM and the peripheral nodes in the network show strong association. The behavior shown by core and professional nodes can be used by policy makers and healthcare organizations to address the right set of physicians [32].

Studies also show that Social Network Analysis was not studied in the field of healthcare until 2012 [31]. The article by Chambers et al. [31] makes use of a minimum spanning tree on a physician-patient network to determine the cost involved in transporting information between patients-patients, patients-physicians and physicians-physicians [31]. An inspiration to construct a physician-physician network was acquired from this paper.

Agneessens et al. [43] introduces a tuning parameter to centrality metrics which regulates the impact of resources that are received by the nearby nodes and the nodes that are distant. This paper makes use of multiple centrality metrics such as closeness, degree, betweenness centrality which act as estimates of a geodesic distance. These centrality metrics are then used to distribute resources to the remaining nodes in the network. It was found that more nodes are influenced at a geodesic distance closer to the seed node than nodes that are at a farther geodesic distance. The rank correlation is computed for different values of δ and plotted for comparison. The metric of evaluation used is R^2, which is the explanatory power of the variable resources. This is computed for different values of δ based on the reciprocal of closeness centrality. This paper also focuses on estimation of output measures such as behaviour of δ in a diverse environment and resource richness which are of primary concern for future improvements.

We take inspiration from the idea of introducing a control or regulatory parameter, δ, and introduce a similar parameter while designing our metric.

This paper [45], proposes models for influence propagation through social networks. They developed a greedy model and compared it with; a threshold model and cascade model to compare the performance. From this paper, we capture the idea of maximizing the influence propagation within nodes in a social network. We also take inspiration from an idea of activating and deactivating nodes dynamically in the spreading process. This paper also introduces an operational model to spread an idea or innovation through the network, by setting controls on nodes moving from inactive to active and not the other way around. The model used is said to be a progressive model, since only a single idea or innovation is spread through the network and once a node in a network is influenced with some information, it cannot be influenced again.

Overall, most of the research performed are either narrowed down on geographical locations [31], field of study, such as radiology, cardiology etc. [26, 32, 33, 34] or based on specific requirements such as understanding the physician and patient interaction, information spread, innovation spread etc. [35]. It can be noted that research is mostly narrowed down either due to lack of data or requirement. However, we derive ideas and inspirations from this literature survey that can be developed with the current availability of data.

2.5.1 PAYMENT as a form of INFLUENCE

In this thesis, we conceptualize the idea of the investment made by a healthcare company as a form of influence. In simple terms it can be stated that, when a healthcare company invests on a physician he/she is influenced by the healthcare company and in return he/she could influence other physicians and also recommend to the hospitals/universities they are connected to. To support this idea, we researched if these ideas were used in the past. A study performed by Faden et al. [39] states that "Monetary payments are often used as inducements; they motivate people to do something that is preferred by the sponsor". To support this study he surveyed 57 pharmacists, with a questionnaire of which 2/3 of the sample responded positive to the idea of monetary payments being used as inducements, and the remaining 1/3 were put into the neutral or negative category.

Another study conducted by Bentley et al. [40] states that monetary payments increase a person's willingness to participate in research that is started by the sponsor. They also mentioned that their survey shows that, people tend to talk about their sponsor in social events like conferences, public events etc [40]. To validate the prior statement, a survey was conducted with 326 participants of which 279 participants responded positively towards expressing the popularity of their sponsor. It was also seen that these participants received monetary payments below the third quartile range. The participants who received greater than the third quartile payments range did not express any significant popularity towards the sponsor [40]. This proves that there is certainly an upper bound on the monetary payments, beyond which the influence remains constant.

A survey conducted by, Cornett et al. [41] shows a relationship between the payment levels by the sponsor and the relative willingness of the receiver to take risk involved in research. Here, risk refers to how much effort a medical practitioner can put in factors of time, research and survey. It was shown that higher levels of payment make the respondents more willing to participate.

From the above three surveys we can conclude that monetary investment from the healthcare company is a keen driver of influence for the physicians.

2.6 Discussion

In this chapter, we first discussed terminologies that are required to understand further chapters in this thesis. It is followed by a discussion on how a physician citation network follows a power law distribution. This chapter concludes with a deep dive on the related work from which we gain inspiration for designing a network spreading process. A study confirming our assumption of how payment is a form of influence is also made in this chapter.

3. DATA COLLECTION and EXPLORATION

A well-prepared dataset is the most important requirement for a data science project and it involves cleaning up unstructured data and combining this data from multiple data sources into one. This chapter discusses the data required for the two types of relations, explained in section 1.2. The data corresponding to direct relation, where regression analysis is used, is explained in section 3.1, 3.2 and 3.3. The data required for indirect relation, which makes use of network analysis and is explained in section 3.1, 3.2, 3.4, 3.5, 3.6 and 3.7. At the end of the chapter, we perform exploratory data analysis on the datasets obtained from the three healthcare companies in section 3.5 and answer the RQ2. This will also be used to explore company-wise payment strategy in chapter 4.

3.1 Data requirement

In this thesis, to answer the two types of relations mentioned in section 1.2, we required data from different sources. First, financial data which contains records of amounts invested, ROI, win rate and the physician's affiliation such as a university or hospital he/she is currently associated to. Second, to explain the academic profile of a physician, research metrics such as citation count, total articles/documents published and some advanced metrics like h-index would be of utmost interest. The next section explains in detail different data sources from which we extract data for this thesis.

3.2 Data Sources

The three data sources used in this thesis are as follows,

- **Research data** – Scopus is one of the largest data source for citation and author data of peer reviewed research articles. It is available on the internet with restricted access, consisting of over 16 million profiles of researchers from all over the world [16]. It consists of data from various fields of technology, medicine, science etc. and provides a broader scope to understand collaboration networks.

  There are various factors that contribute to the research profile of the researcher. Some of these factors are calculated metrics from citation data and the other factors, mentioned in Table 3.1, are derived from the calculated metrics, previously mentioned. Table 3.1, explains all the metrics that are available from Scopus which are used in this thesis.

Metric Name	**Description**
Citation count	Citation count for an author is the measure of number of other authors in the Scopus database that cite the current author's articles.
H-Index	H-Index is a formulated metric, measuring the productivity of an author. It is a function f that corresponds to the maximum of number of citations verses maximum number of papers published. Higher the h-index greater the productivity of an author [27].
Total number of articles published	This metric represents the total number of all the publications the author possess in the Scopus database
Total number of co-authors	This metric counts the total number of co-authors a physician is connected to in the Scopus database
Years of experience	This metric counts the number of years the physician is involved in research publications.

Table 3.1: Metrics that determine the research profile of a physician

Figure 3.1, demonstrates the structure of the Scopus dataset. The primary key is the author and all the other entities are attributes of the author, as explained in Table 3.1.

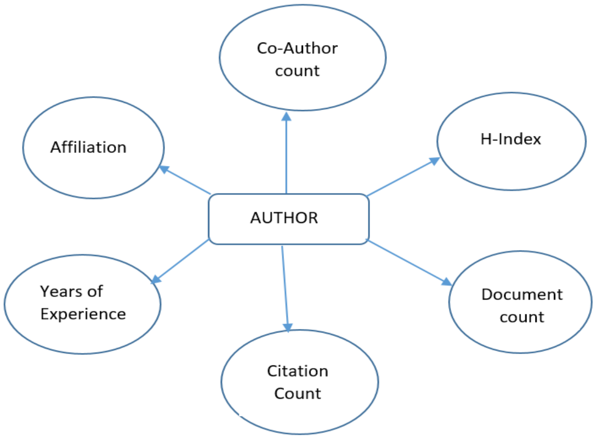


Figure 3.1: Structure of the Scopus dataset

- **Payments data** - OPENPAYMENTS is the outcome of the Physician Payment Sunshine Act passed by the US Congress in 2010, which aims to bring transparency in payments between physicians and healthcare companies [17]. The purpose is to promote affordable care to improve the overall health and well beings of individuals. This database is Open Data for public use and can be freely downloadable without much hassle. The data source is available for over 5 years from 2013 to 2017. As of 2017, the dataset consists of 11.27 million records of physicians and the payments received from 1500 healthcare companies in the USA. Figure 3.2, represents the design of the Author/Physician datasets and its attributes. The physician/author is the primary key and its attributes are payments made to the physician, his/her address recorded when the payment being made and the field in which the physician is working.

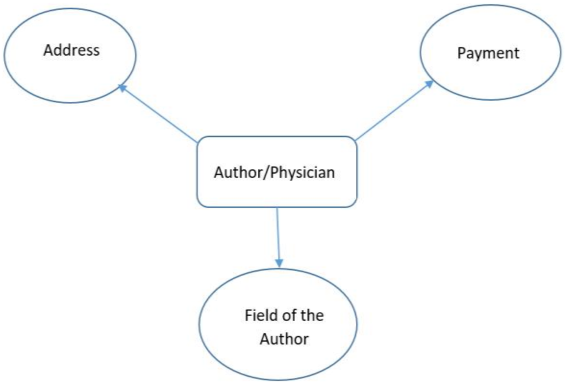


Figure 3.2: Structure of the OpenPayments dataset.

- **Return data** – SALESFORCE is a customer relationship management (CRM) tool that records transactions between companies and their customers [29], which is leveraged by the company to improve returns, analyze win rates etc. This being a licensed software the ROI data is available only for healthcare company ABC. Figure 3.3 represents the structure of the Salesforce dataset. The affiliation (primary key) is either a hospital, clinic or a university making a purchase from a healthcare company. The attributes of the primary key are the affiliation's address and the payment made.

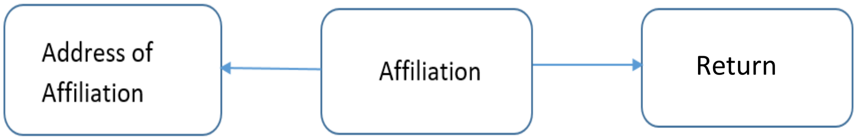


Figure 3.3: Structure of Return Dataset

3.3 Crawling the web

Data scraping also known as web crawling, is a process of importing information from a website to a local storage file on the computer, through a communicative medium of an API (application programming interface) provided by the website. It combines data that is scattered across multiple websites into a single structured form that is stored in one location of the dataset. The three main components required for web scraping are client, server and a communication link to interact between the two. The client is nothing but a script that defines the required information to be extracted. The server is a program that serves files from web pages to users. Communication takes place between the client and server using an API that provides a secure channel for interaction between the client and the server. These three components always need to remain synchronous to each other for data extraction. Once a secure connection is established between the client and the server, the client then requests the server for the required information which is extracted in restricted quantities.

In this thesis, two API keys are used to provide two levels of secure authentication. The first key is the Scopus My API Key who's primary function is to authenticate the IP address. It is also used to limit the number of times the server is hit to prevent service capacity overflow. This API key has a limitation of 5000 requests made to the server in one week, hence around 3 million server requests are made to complete the data extraction process. The second authorization key made compulsory by Scopus is the Institution Token Key, provided by TU Delft and is accessible by connecting to the university's VPN. This ensures that the data extracted is used only for research purposes and not for commercial use. With the help of these two authorization keys a scrape script is written over the Scopus API [30] to extract the required data. The Scopus API provides real time data, built under a RESTful architecture, thereby making the scraping process secure and trustworthy. It also allows for modification of data into the required data structure and shape.

3.4 Combining data sources

The three distinct data sources explained in section 3.2, needs to be combined to a single unique dataset. This is done in two steps, at first we combine the research data and payments data, based on the common primary key, which is the author/physician. In the second step, the affiliation from the ROI data is matched against the affiliation from the intermediate data obtained in the first step.

From the return data it is observed that multiple physicians are associated to one affiliation resulting in overstatement of ROI. To fix this, the ROI of an affiliation is equally distributed over the physicians associated with that affiliation. The combined dataset now consists of author as a primary key and every author has research metrics, payments he/she received, the hospital or university he/she is affiliated to and the distributed ROI. This is represented in Table 3.2.

Author	**Citation Count**	**H-Index**	**# No. of article published**	**# No. of coauthors**	**Years of experience**	**Payments (USD)**	**Affiliation**	**Return (<u>USD</u>)**
A	10	12	30	20	9	2000	KL	100000
B	30	15	35	25	10	3500	MN	150000
C	15	17	70	30	12	8000	KL	100000
D	25	10	22	35	8	15000	KL	100000
E	45	25	30	40	13	2500	ST	50000

Table 3.2: An example of the combined dataset

3.5 Data Insights

In this section, we attempt to answer the RQ2 mentioned below by performing exploratory data analysis on the available data of the three companies.

RQ2: Do the three healthcare companies follow a strategy while making payments?

The data extracted and combined into datasets from three healthcare companies show some variation in their strategy towards making payments explained in Table 3.3. Before we understand the strategy followed by the three companies, it is important to note that we study these three companies as they are peer companies and it is useful for healthcare company ABC to be aware of the strategies followed by the peer companies. From Table 3.3 it can be noted that healthcare company ABC addresses almost twice the number of physicians compared to healthcare company XYZ and healthcare company LMN. It is interesting to note that the total number of publications from these three companies is almost uniform yet the number of coauthors being addressed by healthcare company ABC is significantly higher compared to the other two companies. Hence, the higher CROI (Coauthor Return on Investment) for the same. Another absorbing insight being, the average payment per physician in healthcare company ABC is significantly less. This shows that healthcare company ABC addresses physicians whose nature is to collaborate with a wider audience and have half the number of publications. Whereas, healthcare company XYZ and healthcare company LMN collaborate with fewer researchers but have more number of publications. This variation in behavior from the peer companies provides a good base to understand and analyze the difference in payment strategy amongst these companies.

	Healthcare Companies			**Total**
	ABC	**XYZ**	**LMN**	
Number of physicians paid by company	5096	2605	2657	10,358
Number of coauthors of physicians	851,719	105,730	41,046	998,495
Total Number of Publications of physicians paid by company	197,628	197,348	203,220	598,198
Amount invested by company on these physicians in Millions of USD	12,5	16	14,5	43
Average coauthors per physician	167	41	16	224
Average publications per physician	39	76	77	192
Amount invested per physician in USD	2453	6142	5457	14,052
PROI (publication return on investment) as publications per USD 1000 invested	15,8	15,78	16,26	47,84
CROI (Coauthoring return on investment) i.e. size of network covered per each USD 1000 invested	68,14	8,46	3,28	79,88

Table 3.3: Assessment of strategy for healthcare companies viz., ABC, XYZ and LMN

3.6 Data required for network construction

There are two main elements required to construct a network of physicians, represented as $G = (N, L)$, where N is the set of nodes and L is the set of links, discussed in section 2.2.1. The node can be represented as a physician node or co-author node. Co-author nodes can be nodes that represent a physician in practice or a researcher. In order to maintain simplicity we refer to these nodes as co-author nodes further in this thesis. A link between any two nodes is formed if they have published research articles together. To enrich the network with more information, weights are added to the links which represent the strength of that link. These weights between any two nodes can be defined as the total number of articles the two nodes have collaborated together.

3.7 Network Data Structure

For the network construction, data is restricted to healthcare company ABC. This is because the data pertaining to ROI is only available for this company which is used to evaluate the effectiveness of the network. The network granularity can be broken down into two levels, a physician level and an affiliation level. At the physician level every node in the network is a physician or his/her co-author, and the edges represent research collaboration. From section 2.1.1 a set of nodes is represented as $N = \{n_1, n_2, n_3, \ldots., n_i\}$ where i is the total number of nodes that are present in the network. Each node consists of four attributes, i.e, name of the physician, current affiliation of the physician, the payment received by the physician, represented as P_k for node n_k and the topological metric , S_k generated for a physician, n_k. If a healthcare company invests on a physician the payment attribute is updated with the amount, measured in USD, else the attribute is updated with a zero. These set of nodes are connected to each other by a link, if the two nodes have research collaboration. They are represented as $L = \{l_1, l_2, l_3, \ldots., l_j\}$, where j is the number of links connecting all the nodes in the network. These links also carry information and they are defined in an attribute $W = \{w_1, w_2, w_3, \ldots.., w_j\}$. The weight on a link, equates to the number of articles published together by the two nodes present on either side of the link.

At the affiliation level, nodes represent affiliations and can be defined as $H = \{h_1, h_2, h_3, \ldots., h_a\}$, where a is the total number of affiliations in the dataset. Each node has three attributes, name of the affiliation, a list of physicians who are currently associated with this affiliation, represented as U, and the ROI generated by the affiliation. The affiliation nodes are independent of each other and they are not connected to each other by links, the main reason being they are used only at the evaluation stage which does not involve interaction between these nodes.

Figure 3.4, shows a detailed representation of the physician citation network at a physician level, where the physicians are interconnected to each other, whose names are labelled in black and the weights are, displayed in red and represent the weight of the link. Links with higher weights are displayed with thicker blue line.

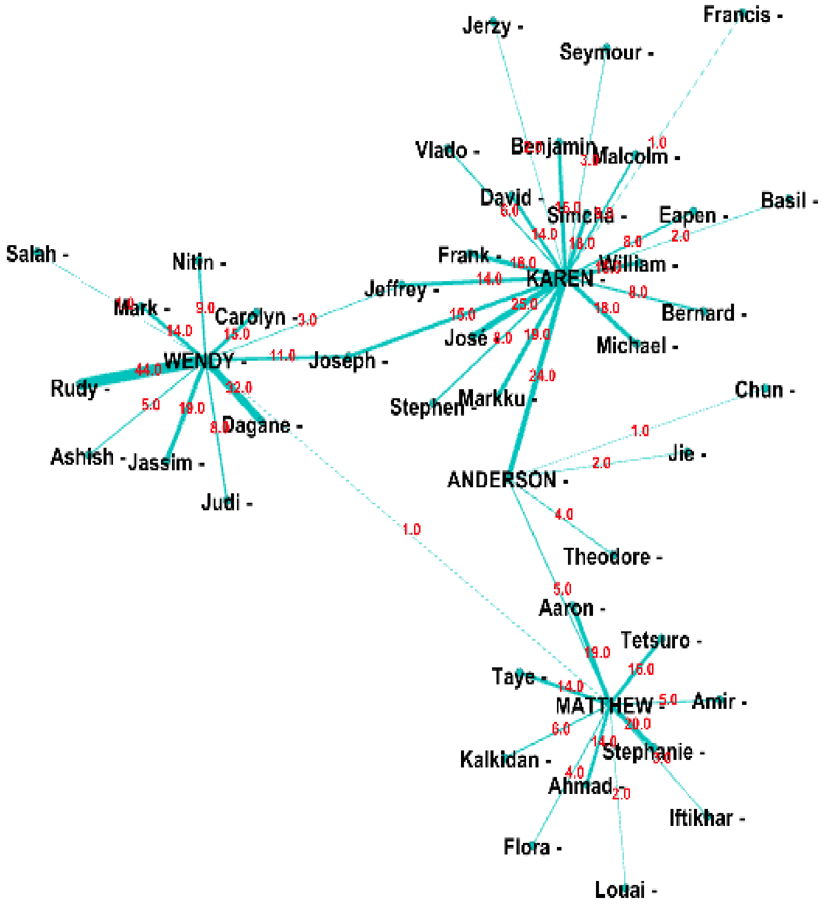


Figure 3.4: Physician citation network with nodes and weighted links

3.8 Discussion

In this chapter we discussed the different data sources and how they were extracted and combined. The data that was analysed and extracted in this chapter is used to answer the two types of relations (direct and indirect relations) proposed in section 1.2. Exploratory data insights were also provided for better understanding of the datasets. A network data structure was discussed with an illustrative example of the physician data that can be used to understand the network thoroughly. The granularity of the network at a physician level and affiliation level is also discussed.

4. STATISTICAL DATA EXPLORATION and MODELING

In this chapter, we use statistical data analysis to analyse the direct relation, explained in detail in section 1.2. By analysing the direct relationship between payments made to the physicians by the healthcare company and the ROI to the healthcare company we will answer the following research questions:

RQ1: What is the relationship between investments on physicians, ROI generated by the healthcare company and the research profile of a physician on whom the healthcare company invests?

RQ2: Do the three healthcare companies follow a strategy while making payments?

The motivation to use regression analysis is to explain the variation in the payments made to the physicians by using their research profile as an explaining factor, for the three healthcare companies. Then we decode the underlying payment strategy used by the three healthcare companies and compare them against each other to answer RQ2.

Further we build a recursive regression model, who's main purpose is to understand the relationship between ROI generated by the healthcare company, the money invested on the corresponding physicians and the research profile of the physicians to answer RQ1 using the direct relationship mentioned in section 1.3.

This chapter begins by describing the distribution of payments made to the physicians followed by determining the relationship between payments made to physicians and their research profile. This is explained in section 4.1. After understanding the univariate and bivariate structure of data, a more systematic approach of explaining the variation in payments made to the physicians is analyzed using the regression technique, which is explained in section 4.2. Section 4.3 attempts to understand if the investment made by the healthcare companies is related to the ROI earned by the company. In this context, the analysis is restricted to only healthcare company ABC as the ROI data is not available for the other healthcare companies. The chapter ends with a discussion and interpretation of results from regression models.

4.1 Basic Description of the Data

The average payments made to the physicians are varied across the three healthcare companies. It was observed from the data that the payments made by healthcare company ABC was the lowest i.e., on average paid USD 862.5 to a physician whereas healthcare company XYZ on average paid the highest, USD 1039. The healthcare company LMN has an average payment of USD 953, making the overall average payment of the three healthcare companies combined to USD 950 per physician. We also noticed a wide variation in payments made to physicians across three health care companies reflecting a high standard deviation and a high coefficient of variation[1]. The coefficient of variation measures the inequality amongst the payments made to physicians and the coefficient of variation is greater than 1 for all the three companies which confirms high level of inequality in the investment structure.

It was also noticed that the distribution of payments made to physicians is highly positively skewed and appears as *Skewness* in Table 4.1. It can also be noted that nearly 50% of the physicians are getting less than USD 50. Only around 8% to 17% of the physicians receive payments more than USD 500.

	Percentage of Physicians receiving the payment from healthcare companies		
Payment in USD	**ABC**	**XYZ**	**LMN**
Upto 50	51.3	55.3	50.7
51 to 100	22.3	13.7	13.4
101 to 200	12.9	10	15
201 to 500	5.5	4.4	5.4
501 to 1000	1.9	4.5	4.2
1001 to 5000	3.1	9.2	5.4
Above 5001	3.1	2.9	6
Average payment in USD	862.5	1039	953
Standard Deviation in USD	6218.7	7752.4	3930.5
Coefficient of Variation	7.2	7.5	4.1
Skewness	15.1	17.4	10.6

Table 4.1: Distribution of Payments across Healthcare Companies

Here we attempt to explain the variation in payment through variables such as total number of articles published, number of co-authors, research experience of a physician and h-index [12]. All of these variables are previously explained in detail in section 3.2. To explain the above variation in payment, the associations between payments and the variables that represent research profile is measured using correlation coefficient[2] and the computed results are presented in Table 4.2.

[1] $Coefficient\ of\ variation = \frac{Standard\ deviation}{Average}$

[2] $Correlation\ Coeffient,\ r_{xy} = \frac{cov(x,y)}{s_x s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

Healthcare Company ABC					
	Total number of articles published	Number of co-authors	Research experience	h-index	Payments
Total number of articles published	1	0.620	0.012	0.830	0.012
Number of co-authors		1	-0.039	0.540	0.006
Research experience			1	0.070	0.026
h-index				1	0.027
Payments					1
Healthcare Company XYZ					
Total number of articles published	1	0.370	0.037	0.680	0.034
Number of co-authors		1	0.100	0.590	0.019
Research experience			1	0.120	-0.016
h-index				1	0.100
Payments					1
Healthcare Company LMN					
Total number of articles published	1	0.800	0.049	0.520	0.068
Number of co-authors		1	0.140	0.710	0.080
Research experience			1	0.190	0.046
h-index				1	0.068
Payments					1

Table 4.2: Correlation Matrix across Health Care Companies

The low correlation values show that the payments made to physicians do not have any linear relationship with other variables, as seen in the last column corresponding to **Payments** in the Table 4.2. The possibilities for low association could be either because the relationship is non-linear or the payments may not really depend upon the physician's research profile or experience. These are some of the possible reasons we could conclude, but they are limited, as correlation is not always a case of causation.

In reality, many variables may jointly contribute to explaining the variation in the payments made to the physicians and hence to account for the impact of two or more variables in a more general way, we resort to multiple regression model and is explained in the next section.

4.2 Regression Analysis

The relationship between two variables x and y, defined as $y = f(x)$, is a set of all values of x and y that are characterized by an equation which can be linear or non-linear. Here, x is an independent variable and y is a variable dependent on x. The relationship between these variables can be deterministic or stochastic. It is deterministic, if for every value in x there is only one corresponding value in y. For example, if y is the payment made to a physician and x is the h-index, and assume all the physicians with h-index 10 receive a payment of USD 500 then the relationship is deterministic. However, in reality, this cannot be true due to a number of reasons like unpredictable element of randomness in human response, effect of large number of omitted variables, measurement of errors in variables etc. [47]. Hence, this makes all the relationships, in general, stochastic in nature. From the previous example, if all the physicians have a h-index of 10, then the payments made to physicians is a complete distribution around the mean value $E(y|x = 10)$. Thus the relationship is specified as $y = f(x) + \varepsilon$, where ε is a random disturbance.

In this thesis, we make use of regression analysis, which is defined as a statistical technique to find the relationship between a dependent variable and one or more explanatory (independent) variables by quantifying it in a single equation. In a one-dimensional case, the linear equation can be represented as:

$$y_i = \alpha + \beta\, x_i + \epsilon_i \qquad (4)$$

Where y_i is the dependent variable, x_i is an explanatory/independent variable and ϵ_i is the stochastic or random disturbance, α and β are the regression parameters, which are unknown parameters and are estimated from the model using the values of x_i and y_i. Here the subscript i refers to the i^{th} observation, i.e., i^{th} physician. Hence, the full specification of the regression model includes the regression equation and the probability distribution of the disturbance term [7, 10]. *(Refer Appendix [A.1] for details of the assumptions on the error term and the parameter estimation technique)*

When the relationship between x and y is non-linear in nature, physicians with a higher h-index get higher payments until a certain threshold after which payments decline or remain constant. In such cases, the regression takes a quadratic form as seen in Eq. (5), which is a particular case of a polynomial regression. Another non linear form of expression, is exponential or semi log form, i.e., with a change in h-index, the payment changes exponentially and corresponding regression equation is termed as exponential regression or log linear regression and is approximated in Eq. (6).

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \qquad (5)$$

$$\ln(y_i) = \alpha + \ \beta\ x_i + \epsilon_i \tag{6}$$

In reality, a single indicator may not be able to explain the variation in the dependent variable. Hence, a set of explanatory variables are necessary to explain the variation of the dependent variable for linear or non-linear forms. In such situations, the model can be generally specified as

$$y_i = \ \alpha + \ \sum \beta_j x_{ji} + \varepsilon_i \tag{7}$$

where, i=1,2,...n are observations in the dataset and j=1,2,...m are explanatory variables.

Here y_i is the dependent variable, x_{ji} is a set of explanatory variables, which are either linear or non linear (i.e., $x_j\ or\ x_j^2\ or\ \ e^{x_j}$) and with the assumption that, there is no linear dependences within them. The regression coefficients, $\alpha, \beta_1, \beta_2, \dots, \beta_m$ are estimated by the method of least squares.

The objective of regression is to estimate the regression coefficients, statistical inference on the estimated coefficients and to determine the strength of their relationship. These coefficients are usually estimated by the method of ordinary least squares[3]. Though this method provides optimum coefficients, they can be affected by issues such as outliers, misspecification of the regression equation and inefficient estimates due to heteroscedasticity which are common problems for cross sectional data. These problems are explained as follows:

(a) **Role of outliers:** We observed that estimates of the regression parameters are influenced by extreme observations or outliers, termed as influential observations. In simple regression, i.e., regression with one explanatory variable, detecting outliers are relatively easier through residual plots obtained after regression analysis. However, in the case of multivariate regression, it is impossible to visualize multi dimensional data to analyze the estimated errors, $\widehat{\epsilon_i}$ which are used to identify outliers. Tests such as DFFITS [4] and Cook's distance [5] are used to identify outliers and are defined as follows.

The DFFITS measure can be mathematically represented as,

$$DFITS_i = \left(\frac{h_{ii}}{1 - h_{ii}}\right)^{1/2} \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}} \tag{8}$$

Where h_{ii} is the diagonal elements of the matrix $[X(X'X)^{-1}X']$, where X is the explanatory variable matrix, e_i is the estimated residual and $s_{(i)}$ is the root mean square error from the regression that is computed without considering the corresponding i^{th} observation. The DFITS

[3] *The method of least squares is the automobile of modern statistical analysis; despite its limitations, occasional accidents, and incidental pollution, it and its numerous variations, extensions and related conveyances carry the bulk of statistical analysis, and are known and valued by all [13].*

Stephen M Stigler(1981)

measure is computed for all physicians and if the DIFFITS value is greater than $2\sqrt{\frac{(m+1)}{n}}$, then the payment made to the physician is considered as an outlier.

The Cook's distance can be mathematically represented as,

$$D_i^2 = \frac{e_i^2}{m * s^2}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right) \tag{9}$$

Where, D_i^2 is the Cook's distance measure, s^2 is the mean square error of regression; h_{ii} is the diagonal element of the matrix $[X(X'X)^{-1}X']$ of explanatory variable matrix X, e_i is the estimated residual and $(m+1)$ are the number of regression coefficients. The Cook's distance is computed for all physicians and if the value is greater than $F_{0.5}(m, n-m-1)$ (or for large sample if $D_i^2 > 1$) then the physician is considered as an outlier.

(b) **Misspecification:** A multiple regression model undergoes a functional form misspecification when it does not properly account for the relationship between the dependent and explanatory variables [6, 8, 9]. For example, if the data takes a log-linear form and it is estimated using a linear regression then this model is misspecified, which leads to coefficients being biased [8]. The test used to detect the misspecification of a model was suggested by Ramsey [1] and termed as RESET (Regression Specification Error Test).

This test augments the multiple linear regression, as specified in Eq. (7), with a set of regressors, Z, as $y_i = \alpha + \beta_1 x_i + \beta_2 x_2 + \cdots + \eta_1 z_1 + \eta_2 z_2 + \eta_3 z_3 + \epsilon_i$, and are tested if the hypothesis $H_0: \eta_1 = \eta_2 = \eta_3 = 0$ is true. Here, the regressors take the form of squares, cubes and fourth power of the fitted value i.e., $z_1 = \hat{y}^2$; $z_2 = \hat{y}^3$; $z_3 = \hat{y}^4$.

(c) **Heteroscedasticity:** Heteroscedasticity can be defined when the error terms , $\epsilon_{i\,,}$ do not have a common variance, σ^2. This problem arises when cross sectional data is used. When heteroscedasticity is not accounted, it leads to a biased estimate of standard errors of the regression coefficient thereby; making the t-test invalid. This might lead to major blunders in drawing conclusions of an explanatory variable. Thus, checking for the presence of heteroscedasticity is important and if it is present, appropriate measures should be taken to fix this issue [9, 11].

The generalized test used to check for the presence of heteroscedasticity was introduced by Breusch and Pagan [2] and is termed as Breusch-Pagan test (BP test) explained below.

Breusch-Pagan test (BP test):

The assumption is that, the heteroscedasticity is a function of one or more independent variables, and it is applicable to a linear function assuming all variables in the model are independent. Consider $var(\varepsilon_i) = \sigma_i^2 = f(\alpha_0 + \alpha_1 z_{1i} + \cdots + \alpha_g z_{gi})$, where Z are set of regressors of the form x, x^2, e^x etc. The BP test, tests for the hypothesis $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_g = 0$.

The test statistics, $\frac{S_0}{2\hat{\sigma}^4}$ has a χ^2 distribution with g degrees of freedom.

If the test reveals the presence of heteroscedasticity, then the errors are converted to its homoscedastic form by performing certain transformations, which is only possible only if the functional form and magnitude of these errors are known. In a multiple regression analysis, it is difficult to identify the functional form and magnitude of errors, hence we use Heteroscedasticity Consistent Covariance Matrix (HCCM) to convert the heteroscedastic nature of errors to its homoscedastic form, as suggested by White et al. [3]. This approach removes the bias which arises in the standard error of the coefficient.

4.3 Discussion of Regression Results

Multiple regression, as specified in Eq. (7), was estimated for the three healthcare companies separately. The basic model showed the presence of outliers, misspecifications and heteroscedastic errors.

The outliers identified statistically through DFFIT and Cook's distance are cross verified subjectively to justify why these payments were considered as outliers. The number of outliers from healthcare companies ABC, XYZ and LMN are 18, 39 and 6 respectively. These payments were identified as royalty, license or payments made towards research. The payments corresponding to these outliers were several hundred folds higher than the mean payments and hence identified statistically and subjectively as outliers. The detected outliers were not further considered in the regression model.

Next, we checked for misspecification using the Ramsey RESET, as explained in section 4.2. Results indicate the presence of nonlinearity for healthcare companies ABC and XYZ whereas, healthcare company LMN fail to show the presence of misspecification, which is observed from Table 4.3. Hence we introduced polynomial terms to the regression model, to eliminate misspecification.

	Healthcare Companies		
	ABC	**XYZ**	**LMN**
Ramsey RESET F	3.27	178.40	2.02
Prob>F	0.0212	0.0000	0.1085

Table 4.3: Ramsey RESET F statistics of the three healthcare companies

The final issue to be handled is the presence of heteroscedasticity. The results obtained from BP Test, presented in Table 4.4, show the presence of heteroscedasticity. To overcome this issue we make use of

White’s solution, as explained in section 4.2, to obtain heteroscedastic consistent standard error which validates the t-test.

	Healthcare Companies		
	ABC	**XYZ**	**LMN**
$Breusch - Pagan\ \chi^2(1)$	318.20	179.96	300.58
$prob > \chi^2(1)$	0.0000	0.0000	0.0000

Table 4.4: Heteroscedasticity test statistics of the three healthcare companies

The regression models, represented in Eq. (10), Eq. (11) and Eq. (12), correspond to the healthcare companies ABC, XYZ and LMN respectively.

To come up with these regression models, we tried various forms of regression models for each healthcare company, by including/excluding explanatory variables and various polynomial forms of the explanatory variables in the regression model, Eq. (7). We also experimented with exponential and log forms of the explanatory variables. All these combinations were used in an attempt to improve the explanatory power of the dependent variable. After multiple trials we ended up with different regression models for the three healthcare companies. This result is shows that there are different payment strategies used by the three healthcare companies.

$$Healthcare\ company\ ABC:\ payment = \alpha_0 + \alpha_1\, h - index + \alpha_2\, h - index^2 + \alpha_3\, experience \quad (10)$$

$$Healthcare\ company\ XYZ: payment = \beta_0 + \beta_1\, h - index + \beta_2\, h - index^2 + \beta_3\, experience \quad (11)$$

$$Healthcare\ company\ LMN: payment = \gamma_0 + \gamma_1\, h - index + \gamma_2\, experience \quad (12)$$

In Table 4.5, the R^2 values indicate that only a small variation in payments could explain the research profile and the experience of the physician. However, the F values indicate that the R^2 is significant. Hence, we deep dive into the results to understand the relationship between payments, research profile and experience of physicians for the three healthcare companies.

Healthcare Company ABC:

The results, as seen in Table 4.5, show a positive significant h-index coefficient and negative significant h-index^2 coefficient indicating that with every unit increase in the h-index, the payments made to physicians increases until a point where the h-index is 48 after which the payments starts decreasing while keeping the experience variable constant. The coefficient corresponding to experience is positive and significant implying that with every additional year of experience, the payments go up by about USD

16. Therefore the payment strategy used by healthcare company ABC is stated as, **physicians with higher years of experience and higher h-index (up to the cut-off) are paid higher.**

Healthcare Company XYZ:

From Table 4.5, we can see that the results from healthcare company XYZ also showed a similar pattern as that of healthcare company ABC with respect to h-index and experience variable. The h-index variable peaks at 54 which is roughly the same as that for healthcare company ABC. The coefficient corresponding to experience is positive and significant implying that with every additional year of experience, the payments go up by about USD 15. Therefore the payment strategy used by healthcare company XYZ is stated as, **physicians with higher years of experience and higher h-index generally tend to receive higher payments.**

Healthcare Company LMN:

On the contrary, the h-index was found to be linear for healthcare company LMN and the coefficients corresponding to both h-index and experience were observed to be positive and significant. Further the results indicate that with every unit increase in h-index, the corresponding investment on a physician is increased by USD 2.17 and similarly for every additional year of experience, the physicians were paid around USD 26 more. The payment strategy used by healthcare company LMN is stated as, **physicians with higher years of experience and higher h-index are paid higher.**

Healthcare Company ABC			
	Coefficients	t value	P>t
h-index	13.49	4.63	0.000
h-index^2	-0.14	-4.45	0.000
Experience	15.95	3.57	0.000
Constant	68.46	2.48	0.016
R^2	0.0085	F=9.67	P>F=0.000
Healthcare Company XYZ			
	Coefficients	t value	P>t
h-index	21.62	3.85	0.000
h-index^2	-0.20	-2.41	0.016
Experience	15.35	2.71	0.014
Constant	303.29	3.99	0.000
R^2	0.0121	F=9.40	P>F=0.000
Healthcare Company LMN			
	Coefficients	t value	P>t
h-index	2.17	4.48	0.000
Experience	25.95	4.40	0.000
Constant	20.76	2.34	0.017
R^2	0.0155	F=18.38	P>F=0.000

Table 4.5: Regression Results

Overall, **the three healthcare companies follow their own strategy to invest on their physicians which answers the RQ2**. Among all the variables used to define the research profile of a physician, we conclude that two variables i.e h-index and years of experience are the most crucial variables.

4.4 Return Analysis

In this section, an attempt is made to understand the relationship between ROI generated to the healthcare company by the hospitals and investments made to the physicians affiliated to these hospitals, which answers RQ1. The relationship between ROI generated and investments is represented in Eq. (13).

$$y_1 = \alpha_0 + \alpha_1 y_2 + \epsilon_1 \tag{13}$$

where, y_1 is the ROI earned by the company
y_2 is the payments made to physicians
α's represents the coefficients of explanatory variables in Eq. (13)

In the above equation, investments made to physicians cannot be treated as a pure exogenous variable; unlike in other equations. Since these payments depends on the research profile and years of experience, which is inferred from Table 4.5. Thus we specify another regression models that explains the investment better and is represented in Eq. (14). As we can see Eq. (13) recursively depends on Eq. (14), these two equations together form a recursive regression model and hence should be jointly estimated. It can also be noted that if Eq. (13) is estimated by itself then the coefficients obtained would be biased and inefficient.

$$y_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_2 \tag{14}$$

Where, y_2 is the payments made to the physicians
x_1 is h-index of the physician
x_2 is a polynomial function of h-index (h-index^2)
x_3 is the years of research experience of a physician
β represent the coefficients of explanatory variables in Eq. (14)

Table 4.6 demonstrates the results of the recursive regression model where, Eq. (13) corresponds to the return equation and Eq. (14) the investment equation. The results indicate that the investment made to physicians has no impact on the ROI generated to the company. The payment equation, i.e., Eq. (14) shows that the h-index has a quadratic relationship with payments made to the physicians, i.e., with increase in h-index the payments also increases up to a certain level and beyond which it decreases with the increase in h-index. Experience also has a positive impact on the payments made to the physicians. All the coefficients in Eq. (14) are significant at 1% .

Return Equation			
	co-efficient	Z-value	P>Z
Payment	-2014.61	-1.28	0.199
Constant	3066199.00	5.71	0.000
R^2	0.0004	χ^2=1.65	P>χ^2=0.1993
Payment Equation			
	co-efficient	Z-value	P>Z
h-index	14.06	4.74	0.000
h-index^2	-0.14	-3.82	0.000
Experience	13.79	2.17	0.030
Constant	87.05	2.23	0.021
R^2	0.0084	χ^2=30.04	P> χ^2=0.000

Table 4.6: Recursive regression results

4.5 Discussion

In this chapter there were many key observations and conclusions that are useful for the remaining chapters. First, includes a discovery of a huge variation in investment made to physicians, which have been explained by two main research metrics, i.e., h-index and experience. The second conclusion made from return analysis shows that there is no strong direct relationship between investment made to physicians and ROI to healthcare company ABC. The last observation includes the investment strategy that is established by healthcare company ABC is different from its two peer healthcare companies and answers the RQ2. Overall, we attempted to build a direct relationship between payments and ROI in this chapter and we also answered RQ1 by explaining the relationship between investment, ROI and the research profile of the physician in detail, by first performing an exploratory data analysis which is followed by an advanced regression analysis and finally terminates with results that assists in making the above conclusions.

In the next chapter, we attempt on analysing our second type of relation, i.e. indirect relation by using network science to explain the complex and indirect relationship between payments made to the physicians by the healthcare company and the ROI to the healthcare company. We then compare if direct or indirect relationship can best explain the relationship between payments made by the physicians and the ROI received by the healthcare company, which will be presented in chapter 6.

5. NETWORK SPREADING PROCESS

In this chapter, we attempt to address the indirect relation, explained in section 1.2, to explain the relationship between payments made to physicians by the healthcare company, the ROI to the healthcare company and the research profile in a physician citation network.

We answer RQ1 by developing a network spreading process to model how influence diffuses through the physician citation network. In our spreading process the diffusion of influence is dependent on the topological property of the node.

To measure the effectiveness of the influence diffusion, we use a metric termed *Pearson's correlation coefficient* [48], which measures the relationship between the influence a healthcare company has on a physician and the ROI from the hospital that the physician is employed. We also use visuals to understand the relationship between influence and ROI.

What is the relationship between investments on physicians, ROI generated by the healthcare company and the research profile of a physician on whom the healthcare company invests?

The organization of this chapter is as follows, section 5.1 discusses the spreading process to diffuse influence followed by section 5.2 where we developed three metrics based on degree, clustering coefficient and closeness centrality based topological properties. These properties determine the amount of influence that is diffused from one node to another during the spreading process in section 5.1. Finally section 5.3 discusses the reason we choose these three topological properties to develop the network topological metrics.

5.1 Spreading Process

It is inferred from section 2.5.1, that the money invested can be represented as a form of influence. Here the investment made by the healthcare company to a set of physicians is considered as an influence the healthcare company has on these physicians. In this section, we propose a network spreading process that models how influence diffuses through a physician citation network. For the evaluation of the spreading process the association between the influence diffused at every node and the ROI from the hospital associated to that node is computed.

To design the spreading process we use two attributes of a node, payment attribute represented as P_k and the topological property represented as S_k of each node n_k, which is explained in detail in section 5.2.

At t=0, the nodes that receive a payment from the healthcare company are assigned to the source node set, represented as Source = $\{n_{s1}, n_{s2}, \dots, n_{sp}\}$, and are initialised with their respective P_k values. All the remaining nodes in the network are initialised to 0 as they do not receive any payments from the healthcare company. This is represented in Eq. (15). We initiate the spreading process with the source nodes as these nodes have received influence from the healthcare company, inferred from section 2.5.1.

$$P_k(t=0) = \begin{cases} P_k & for\ Source\ nodes \\ 0 & for\ remainaing\ nodes \end{cases} \tag{15}$$

Before initiating the spreading process, we first discuss the mechanism of the spreading process and the motivation behind these mechanisms.

The mechanism of the spreading process determines what kind of nodes can influence another nodes in the network, and is as follows,

1. A node can influence another node in the network **only once,** for all nodes in the network. The reason being the information/influence has already been transferred the first time a node influences another node.

2. Influence cannot flow from any node in the network to the Source nodes. The motivation being the source nodes have already been influenced by the healthcare company and receive payments for the same.

The basic principle behind the spreading process is, every eligible node will keep θ (defined as the ***spreading parameter)*** times the payment attribute and diffuse the remaining to its one-hop neighbours in proportion to the topological property of the node. This value of θ lies between 0 and 1.

Now, we mathematically deduce the spreading process for all t ≥ 1,

At any given time t some number of nodes, in the physician citation network, **participate** in the spreading process. These nodes can either diffuse influence, receive influence or diffuse and receive influence at the same time. The nodes that diffuse influence at time t, are **only** the nodes that receive influence in it's previous time t-1. The nodes that **only** receive influence at time t, are the one-hop neighbours of the eligible nodes which received influence at time t-1 and are subject to the constraints mentioned in the mechanism above. The nodes that diffuse and receive influence, at the same time t, are the eligible nodes, which received influence at time t-1, that diffuse influence to the one-hop neighbouring nodes subject to the constraints from the mechanism, and receive influence from other eligible nodes diffusing at time t. The computation of the payment attributes of these three nodes are represented in Eq. (16)

$$P_k(t) = \begin{cases} \theta * P_k(t-1) + \sum\limits_{x\,\epsilon\,O_{\{k,t\}}} (1-\theta) * P_x(t-1) * \left(\frac{S_k}{\sum_{y\,\epsilon\,V_{\{x,t\}}} S_y}\right) & \forall\ nodes\ which\ diffuse\ and\ receive\ influence\ at\ time\ t \\ P_k(t-1) + \sum\limits_{x\,\epsilon\,O_{\{k,t\}}} (1-\theta) * P_x(t-1) * \left(\frac{S_k}{\sum_{y\,\epsilon\,V_{\{x,t\}}} S_y}\right) & \forall\ nodes\ which\ only\ receives\ influence\ at\ time\ t \\ \theta * P_k(t-1) & \forall\ nodes\ which\ only\ diffuse\ influence\ at\ time\ t \end{cases} \tag{16}$$

Where we denote **set $\boldsymbol{O_{\{k,t\}}}$** as the set of all nodes from which n_k receives influence at time t and for every element in **set $\boldsymbol{O_{\{k,t\}}}$** we denote **set $\boldsymbol{V_{\{x,t\}}}$** as a set of all nodes that influences its one hop neighbour at time t. Here, $\theta * P_k(t-1)$ represents the amount of influence remaining in the node n_k after diffusion at time t, $\sum_{x \in O_{\{k,t\}}}(1-\theta) * P_x(t-1) * \left(\frac{S_k}{\sum_{y \in V_{\{x,t\}}} S_y}\right)$ represents the influence received from its one-hop neighbouring nodes at time t. The first equation of Eq. (16) is used to compute the P_k for all nodes that participate in diffusing influence and also receiving influence at the same time t.

On the other hand, for nodes that only receive influence, $P_k(t-1)$ represents the amount of influence already present in the node and $\sum_{x \in O_{\{k,t\}}}(1-\theta) * P_x(t-1) * \left(\frac{S_k}{\sum_{y \in V_{\{x,t\}}} S_y}\right)$ represents influence received from its one-hop neighbouring nodes at time t. It is to be noted that, the spreading process diffuses influence to its one-hop neighbours in proportion to the topological property of the node, represented as S, which is explained in detail in section 5.2. The reason behind using the topological property of the node is to capture various properties like distance, spectra and connections of the node [44], which unravel information like importance of the nodes, structure of the network etc.

The other nodes that are **not participating** at time t are computed as,

$$P_k(t) = \begin{cases} \theta * P_k(t=0) & for\ Source \\ P_k(t-1) & for\ remaining\ nodes \end{cases} \tag{17}$$

This process is repeated for t ≥ 1 until $P_k(t) = P_k(t-1)$, i.e no nodes in the physician citation network are influencing one another.

Evaluation of the spreading process: At the end of the spreading process, every node in the network has some amount of influence stored in P_k for node n_k. Nodes that belong to the same hospital are grouped together and their corresponding P_k are aggregated, which is a representation of the total influence the healthcare company has on a hospital. The details for which are specified in Eq. (26) of section 6.1.

From a business standpoint we assume that, if the hospital has been influenced by the healthcare company ABC, then the hospital purchases equipments from the healthcare company ABC, which in turn generates a ROI for the healthcare company ABC. For reference, this explanation was pictorially depicted in Figure 1.2. An evaluation metric termed Pearson correlation coefficient [46, 47, 48] is used to compute the relationship between the total influence the healthcare company has on hospitals and the ROI from the hospitals. The spreading process is calculated multiple times by varying the values of α , θ and topological property S, which is explained in detail in section 5.2.

We illustrate the spreading process with an example represented in Figure 5.1. Nodes A and B represent Source nodes and nodes C, D, E and F represent the remaining nodes. The topological measure used in this example is, degree centrality metric, computed from Eq. (19) where the scaling parameter $\alpha = 1$, and the spreading parameter, $\theta = 0.7$. At t=0, the payment attribute of the source nodes, P_A and P_B

are initialized with USD 1000 and USD 2000 respectively, as seen in Figure 5.1 (a). The payment attributes of all the nodes after initialization can be represented as,

$P_A(0) = 1000;\ P_B(0) = 2000;\ P_C(0) = 0;\ P_D(0) = 0;\ P_E(0) = 0;\ P_F(0) = 0$

At t=1, node A diffuses its influence to node D and C, and node B diffuses its influence to nodes E in proportion to their topological values depicted inside the nodes, which is represented in Figure 5.1(b). The amount of influence a node receives is computed using Eq. (16) and Eq. (17). For example, node C receives influence from node A. Hence, set O = {A} and set V = {C, D}. We compute $S_C(d)$ and $S_D(d)$ using Eq. (19). The following computations of Eq. (16) and Eq. (17) are:

$$Node\ A\colon\ P_A(1) = 0.7 * 1000 = 700$$

$$Node\ B\colon\ P_B(1) = 0.7 * 2000 = 1400$$

$$Node\ C\colon\ P_C(1) = 0.7 * P_C(0) +\ 0.3 * P_A(0)\left(\frac{S_C(d)}{(S_C(d) + S_D(d))}\right) = 0.7 * 0 +\ 0.3 * 1000\left(\frac{0.41}{(0.41 + 0.45)}\right) = 0 + 143 = 143$$

$$Node\ D\colon\ P_D(1) = 0.7 * P_D(0) +\ 0.3 * P_A(0)\left(\frac{S_D(d)}{(S_C(d) + S_D(d))}\right) = 0.7 * 0 +\ 0.3 * 1000\left(\frac{0.45}{(0.41 + 0.45)}\right) = 0 + 157 = 157$$

$$Node\ E\colon\ P_E(1) = 0.7 * P_E(0) +\ 0.3 * P_B(0)\left(\frac{S_E(d)}{S_E(d)}\right) = 0.7 * 0 +\ 0.3 * 2000\left(\frac{0.54}{0.54}\right) = 0 + 600 = 600$$

$$Node\ F\colon\ P_F(1) = P_F(t-1) =\ P_F(0) = 0$$

Similarly, the payment attributes of the participating nodes can be calculated for remaining time t =2, 3 and 4 and is terminated at t=4, since $P_k(4) =\ P_k(3)$.

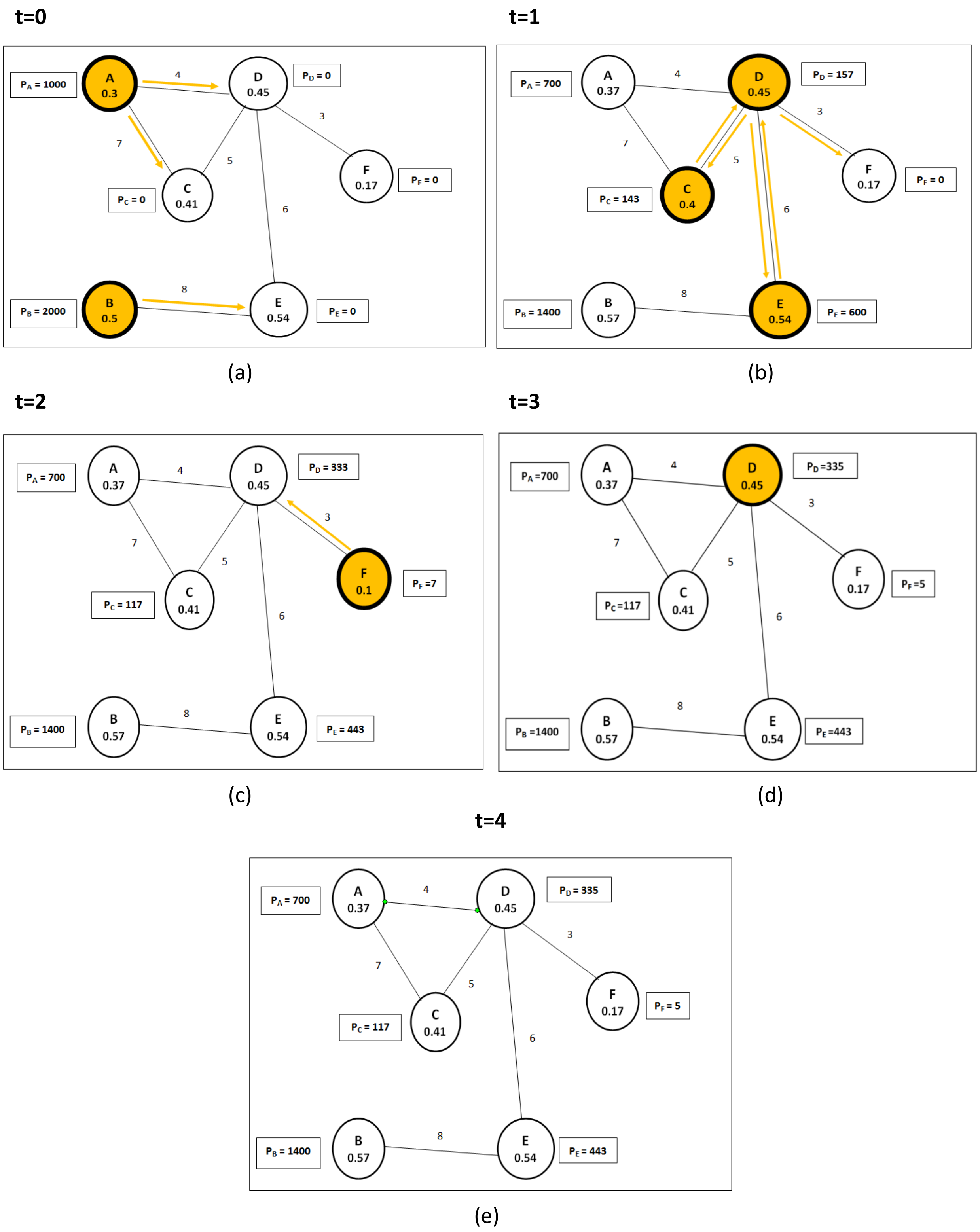


Figure 5.1 Example of influence spreading process (a) influence diffusion from source to remaining nodes at t=0; (b) diffusion of influence at t=1; (c) diffusion of influence at t=2; (d) diffusion of influence at t=3; (e) diffusion of influence at t=4

5.2 Network Topological Metrics

In the previous section we discussed the spreading process in detail, where we understood how influence can be diffused into a physician citation network.

In this section, we design three topological metrics, namely degree centrality metric, clustering coefficient metric and closeness centrality metric. These metrics are based on topological properties of the node, mentioned in section 2.1, and are used in the spreading process, previously mentioned in section 5.1, where the influence diffused to a node is proportional to its topological property. The reason behind using the topological property of the node is to capture various properties like distance, spectra and connections of the node [44], which unravel information like importance of the nodes, structure of the physician citation network etc. In the context of this thesis, it is used to capture the strength of research citations between physicians, number of physicians a physician is connected to, the structure of research citations, etc. Hence we state our objective as:

The objective behind designing these network topological metrics is to diffuse influence to the nodes in proportion to the topological property of the node.

Each of the metrics are built upon an assumption that imitates the behaviour of a physician in his/her physician citation network.

5.2.1 Degree Centrality Metric (S(d))

Degree centrality of a node, d_k , as defined in section 2.1, is the total number of connections a node has but does not indicate the importance or strength of the connection. In order to account for the importance of a node we use a weighted degree centrality metric, wd_k , computed as the sum of all the strengths of the neighbouring nodes in the network. In a physician citation network the strength is the number of articles he/she has co-authored with the neighbouring node. Using this property of a network, we make an assumption as follows,

Assumption1: Nodes with high weighted degree receive higher influence compared to the nodes with lower weighted degree.

The two features of the weighted degree centrality from Assumption1 are, the **strength of the links** and the **number of nodes** that can be varied for a node. This can be captured by computing the weighted degree of the nodes w.r.t. its neighbouring nodes and is represented as:

$$S_k(d) = \frac{wd_k}{\sum_{\mathrm{r} \in \mathrm{R}} wd_r} \ \forall \ k \ \ in \ \ i \tag{18}$$

where,
$S_k(d)$, is the weighted degree centrality metric of a node n_k
wd_k is the weighted degree of the node n_k
wd_r is the weighted degree of the neighbouring nodes of a given node n_k
i is the set of nodes in a given network
R is the set of all one-hop neighbours for every node n_k

To generalize this topological measure we introduce a *scaling parameter* α, which regulates the relative impact of the number of citations between physicians. The choice of α *can be made both theoretically and empirically,* but in this thesis we make use of the empirical approach which is calculated by maximizing the *correlation*, mentioned in Eq. (27), with an outcome variable. By varying the scaling parameter α we identify the optimal value [24, 25] .

The generalized weighted degree centrality metric is defined as

$$\mathrm{S}'_{\mathrm{k}}(d) = \frac{\mathrm{wd}_{\mathrm{k}}^{\alpha}}{\sum_{\{\mathrm{r}\}} \mathrm{wd}_{\mathrm{r}}^{\alpha}} \ \forall \ k \ in \ i \tag{19}$$

It was found that $\mathrm{S}'_{\mathrm{k}}(d)$ is more sensitive for lower values of α , i.e. $\alpha < 1$, compared to higher values of α.

5.2.2 Clustering Coefficient Metric(S(cc))

In section 2.3.4, we discussed that clustering coefficient of the node, measures the ability of a node to form clusters. Nodes with high clustering coefficient diffuse influence much faster amongst nodes than, the nodes with a low clustering coefficient [37]. It is also noted that, Peres et al. mentions diffusion of information within the clusters, is much faster than diffusion of influence in a linear arrangement [38]. Thus we make the assumption

Assumption2 : A node with higher clustering coefficient receives higher influence compared to the nodes that have a lower clustering coefficient.

Limitation of the clustering coefficient metric is that, it cannot distinguish between physicians with a higher number of citations and physicians with a lower number of citations. In order to account for this, several weighted clustering coefficient metrics have been introduced in literature [19, 42] and we adopt this weighted clustering coefficient in our thesis.

The weighted clustering coefficient for n_k is

$$WCC_k = \frac{2 \ \sum(wd_k \ wd_o \ wd_m)^{1/3}}{wd_k(wd_k - 1)} \tag{20}$$

Where, n_k , n_o and n_m are the three inter-connected nodes and $wd_{k,}$ $wd_{o,}$ and wd_m are the respective weighted degrees of the nodes. The normalized weighted clustering coefficient of the neighbouring nodes is defined as

$$S_k(cc) = \frac{WCC_k}{\sum_{\mathrm{r} \in \mathrm{R}} WCC_r} \ \ \forall \ k \ in \ i \tag{21}$$

where,
$S_k(cc)$, is the normalized weighted clustering coefficient metric of n_k
WCC_k is the weighted clustering coefficient of n_k

WCC_r is the weighted clustering coefficient of the neighbouring nodes of a given node n_k
i is the set of nodes in a given network
R is the set of all one-hop neighbours for every node n_k

Similar to the degree centrality metric, we generalize the weighted clustering coefficient by introducing a *scaling parameter α*, which regulates the relative impact of the number of citations between physicians. The generalized weighted clustering coefficient is defined as

$$\mathrm{S}'_{\mathrm{k}}(cc) = \frac{\mathrm{WCC}_{\mathrm{k}}^{\alpha}}{\sum_{\mathrm{r} \in \mathrm{R}} \mathrm{WCC}_{\mathrm{r}}^{\alpha}} \qquad \forall \ \ k \ \ in \ \ i \tag{22}$$

5.2.3 Closeness Centrality Metric(S(c))

Revisiting the definition of closeness centrality of a node from section 2.2.2, which measures how close the source node is with respect to any node in the network. To incorporate information on the number of citations we compute the metric as

$$C_k = \frac{1}{\sum_{m \in G} d(m,k)} \tag{23}$$

Where, n_m and n_k are two nodes in the network G and $d(m,k)$ is the distance (number of citations) between the two nodes. From the closeness property of the node we make an assumption:

Assumption3: Nodes that are closer to the source nodes receive more influence than the nodes that are farther away.

Using the computed closeness metric C_k, we compute the relative closeness of the node w.r.t the neighbouring nodes, since this property was not factored in while computing C_k. Hence the closeness centrality metric is mathematically represented as:

$$S_k(c) = \frac{C_k}{\sum_{\mathrm{r} \in \mathrm{R}} C_r} \ \forall \ k \ in \ i \tag{24}$$

where,
$S_c(k)$, is the weighted closeness centrality metric of a node n_k
C_k is the closeness centrality of the node n_k
C_r is the closeness centrality of the neighbouring nodes of node n_k
i is the set of nodes in a given network
R is the set of all one-hop neighbours for every node n_k

This metric is generalized by introducing a *scaling parameter α*, which regulates the relative impact of the number of collaborations between physicians.

The generalized closeness coefficient metric is defined as

$$\mathrm{S}'_{\mathrm{k}}(c) = \frac{\mathrm{C}_{\mathrm{k}}^{\alpha}}{\sum_{\mathrm{r} \in \mathrm{R}} \mathrm{C}_{\mathrm{r}}^{\alpha}} \; \forall \, k \; in \; i \tag{25}$$

5.3 Reasons for choosing these three topological measures

In section 5.2 we designed three topological metrics based on topological properties, mentioned in section 2.1. As these metrics are built upon different assumptions that, we wanted to understand the empirical relationship between these three topological metrics. Hence, we computed the correlation coefficient between these metrics and it is represented in Table 5.1. It can be noted that, the correlations between the topological metrics strongly depend on the network under study [46]. If the correlation between any two topological measures is low then, they are mutually exclusive to one another. From Table 5.1, it can be seen that the three topological metrics are mutually exclusive to one another.

	Degree	*Clustering*	*Closeness*
Degree	1.0		
Clustering	-0.0130	1.0	
Closeness	0.0029	0.2770	1.0

Table 5.1: Correlation coefficient between metrics

5.4 Discussion

In this chapter, we proposed a network spreading process that is used to diffuse influence. The spreading process is evaluated to answer RQ1. We also introduced the ***scaling parameter* $\boldsymbol{\alpha}$** and the ***spreading parameter* $\boldsymbol{\theta}$** to assist in the estimation of how much influence is diffused to the neighbouring nodes in the spreading process. We also propose that, the nodes diffuse influence in proportion to their topological property, for which, we proposed three generalized network topological metrics. We also proved that the topological measures are mutually exclusive to one another.

6. RESULTS

The purpose of this chapter is to evaluate the modelling of influence in the spreading process to measure the relationship between nodal influence and the ROI of the healthcare company. Revisiting section 1.2, we proposed two relationships to measure the link between payments made by the healthcare company, the ROI of the healthcare company and the research profile of the physician. Regression technique is used to explain the direct relationship between payments, ROI and research profile of the physicians. Spreading process is proposed to explain the indirect relationship between payments and ROI in a physician citation network.

We evaluate the modelling of influence from the spreading process by estimating the relationship between the nodal influence resulted from the spreading process and ROI of the healthcare company. For which, we use three different ways to evaluate the spreading process.

- We use Pearson's Correlation Coefficient to estimate the linear relationship between nodal influence and ROI [49].
- We also use of Spearman's Rank Correlation to estimate the non linear relationship between ordinal values of nodal influence and ROI.
- Visualize the distributions of payments, ROI and nodal influence, to compare and understand the relation between their distributions.

The organization of this chapter is as follows: section 6.1 discusses the evaluation process where we use three ways to evaluate influence diffusion. This is followed by different experiments to evaluate the network spreading process in section 6.2. Section 6.3 discusses the results obtained from the experiments that were explained in section 6.2. Finally, section 6.4 discusses the time taken to influence physicians in a physician citation network. This chapter terminates, with a comparative study between the two relationships, i.e., direct and indirect relationship, mentioned in section 1.2 of this thesis to conclude best relationship between payments made by the healthcare company and the ROI to the healthcare company.

6.1 Evaluation Process

In this section, we elaborate on the evaluation process and the three different ways in which we evaluate the modelling of influence diffusion by the spreading process. Going back to section 5.1, we see that after the termination of the spreading process, every node in the network has a payment attribute (P_k), which is greater than zero. Since, the ROI for the healthcare company is at the hospital level we group all the physicians belonging to the same hospital and aggregate their corresponding P_k values. The set of physicians that belong to the same hospital is stored as an attribute of the hospital node, it is represented as U_b for every hospital node h_b. For every hospital node h_b, we also have another attribute, called influence at hospital level, represented as I_b , which is the sum of all the payments received by the physicians in the list U_b. These attributes of the network have been explained in detailed in section 3.7 of this thesis. Now, we compute the influence at a hospital level, from Eq. (26).

$$I_b = \sum_{u \epsilon U} P_u \qquad \forall\ h_b\ \in H \tag{26}$$

where, I_b represents the total amount of investment made by the healthcare company on a given hospital h_b .

At this stage, we have influence I_b for every hospital h_b and we have the ROI_b for every hospital h_b. Using this data we visualize the distributions, see the relation using scatter plots, estimate Pearson's correlation coefficient and estimate Spearman's correlation coefficient, which is discussed in detail as follows.

- Estimate the relationship between Influence and ROI at a hospital level, using Pearson correlation coefficient, r(I,ROI). The Pearson correlation co-efficient, r(I,ROI), is a measure of the strength and direction of the linear relationship between influence and ROI [46, 47, 48]. The range of r(I,ROI) if from +1 to -1 and the sign of r(I,ROI) indicates the direction of the correlation between influence and ROI. The closer the absolute value of r(I,ROI) to 1 the strong the correlation between the influence and ROI. We represent r(I,ROI) mathematically as,

$$r(I, ROI) = \frac{\sum_b (I_b - \bar{I})(ROI_b - \overline{ROI})}{\sqrt{\sum (I_b - \bar{I})^2}\sqrt{\sum (ROI_b - \overline{ROI})^2}} \tag{27}$$

  Where $\bar{I}$ and $\overline{ROI}$ are the mean values of the influence and ROI and I is the influence at the hospital level.

- Estimate the relationship between Influence and ROI at a hospital level, using Spearman Rank Correlation $\rho(I, ROI)$) [10][4]. This is a non parametric measure of non-linear dependency of two variables and is more suitable for ordinal or interval data. However we converted the absolute values into ranks and computed the non parametric measure using the following measure. We represent $\rho(I, ROI)$ mathematically as,

$$\rho(I, ROI) = \frac{\sum_b \left(R(I_b) - \overline{R(I)}\ \right)\left(R(ROI_b) - \overline{R(ROI)}\right)}{\sqrt{\sum \left(R(I_b) - \overline{R(I)}\right)^2}\sqrt{\sum \left(R(ROI_b) - \overline{R(ROI)}\right)^2}} \tag{28}$$

  Where $R(I_b)$ and $R(ROI_b)$ are the ranks of Influence and ROI respectively and $\overline{R(I)}$ and $\overline{R(ROI)}$ are the mean ranks of Influence and ROI respectively.

[4] Kendall Concordance measure is another non parametric measure used to understand the association between any two variables, which again depends on the ranks. This measure has more similarity with spearman rank correlation coefficient and hence it was not considered here.

- Visually observe the distribution of Influence for different values of α and θ. Compare it to the distributions of ROI and payments. We also visually understand the relation between influence and ROI by the use of scatter plots. While visualizing the data, we also make sure that the data corresponding to payments, ROI, Influence, etc. are protected. We ensure this by applying one of the most prominent ways of encrypting data in visuals, i.e., by changing the scale of the axis of the graphs to ensure maximum protection of sensitive data. The act of protecting sensitive information comes from Art. 4, 9, 13, 14, 15 and 52 of the GDPR rulebook [50].

6.2 Experiments Conducted

In this section, we demonstrate the different experiments conducted to estimate the Pearson's correlation coefficient and Spearman's rank correlation, defined in section 6.1. The experiments are listed below.

- To compute the **Pearson's correlation coefficient** and **Spearman's rank correlation** we perform experiments by using the below steps.
  1. As explained in section 5.1, we vary the spreading parameter θ between 0 (exclusive) and 1 (exclusive) with a step size of 0.1. It can be inferred that when $\theta = 1$, all the payments remain with the source node and the remaining influence, i.e. (1 - θ) = 0 which indicates no diffusion takes place.
  2. We vary the scaling parameter α, which was introduced in section 5.2, between 0 (exclusive) and 2 in an interval of 0.1. It can be inferred that when $\alpha = 0$, any topological metric will be 1, since $x^0 = 1$.
  3. For the three topological metrics, i.e, degree centrality metric, closeness centrality metric and clustering coefficient metric, we vary the spreading and scaling parameters, i.e, θ and α, which generate multiple experiments/trails. We then compare these experiments with one another to obtain the optimum topological metric, spreading parameter (θ) and scaling parameter (α). The values of θ and α are obtained when the correlation between investment and ROI is maximum, i.e, when the relationship between investment and ROI is the strongest. The maximum correlation indicates the maximum explainablity of payments w.r.t ROI, which is the main aim of this thesis, which can be inferred from section 1.2.

  To illustrate an example of a single experiment let us consider α as 0.5 and θ as 0.7, i.e. 70% of the invested amount is kept within the node and the remaining 30% is diffused to the nodes in the network. Let us consider the degree centrality metric as a topological property in this experiment. These parameters are used into the spreading process and the resulting correlation between I and ROI was calculated to be 0.43, which indicates that the strength between ROI and I is 0.43, measured on a scale of -1 to 1.

- To **visualize the distribution** of Influence for different values of α and θ, we keep α constant and vary θ and vice versa. To visually understand the relationship between Influence and ROI we plot a scatter plot .

6.3 Discussion of results from the experiments conducted

In this section we discuss all the results of the experiments obtained from the experiments performed from section 6.2. We first jointly explain the results obtained for Pearson's correlation coefficient and Spearman's rank correlation. Then we move on to the discussion on the distributions of Influence and ROI and the relation between Influence and ROI using scatter plots.

First, the behaviour of the three topological metrics are captured in Figure 6.1a, 6.1b and 6.1c respectively. The X and Z axis represent varying α and θ values and the Y axis corresponds to the Pearson's correlation coefficient computed for a combination of values on the X and Z axis. Figure 6.1a, demonstrates the spreading process using the degree centrality metric, $S'(d)$ where, it can be observed that when α takes the value 0.5 and θ takes the value 0.7, a maximum correlation was attained at 0.432.

Similarly, Figure 6.1b demonstrates the spreading process using clustering coefficient metric as the topological property, $S'(cc)$, where it is observed that when α takes the value 0.5 and θ takes the value 0.7, a maximum correlation attained was 0.482, which is slightly higher than the correlation from Figure 6.1a.

Figure 6.1c demonstrates the spreading process using closeness centrality metric as the topological property, $S'(c)$ where two peaks can be noted when $\alpha = 0.5\ and\ \alpha = 0.7$ when the parameter θ takes the value 0.7. These two peaks have their correlation at 0.55 and 0.38 respectively. However, the second peak has a lower correlation compared to the other two topological properties (in Figure 6.1a and 6.1b) due to which we do not consider the result of the second peak further.

Overall, from Figures 6.1a, 6.1b and 6.1c we can conclude that all the experiments conducted attain a maximum correlation when the spreading parameter θ is 0.7, i.e., when the nodes in the network keep majority (70%) of the investment with the source nodes and diffuses the remaining 30% of the investment to the remaining nodes in the network to attain a maximum explanation of ROI through investments made.

From figures 6.1a, 6.1b and 6.1c we can see that all the three topological metrics follow a pattern, where they all peak with a maximum correlation when the scaling parameter α is at 0.5. This means that at $\alpha = 0.5$, a maximum explainablity of relationship strength is attained between payments and ROI. We can also interpret the physical meaning of α =0.5 as alpha increases in a concave parabolic function when P_k increases linearly.

From Table 5.1, we see that the three topological metrics are not correlated with each other, i.e. we can say that they are independent of each other, exhibiting different properties of the network. But from Figure 6.1a, 6.1b and 6.1c we can see that the closeness centrality metric attains the highest Pearson's correlation coefficient compared to clustering coefficient metric and degree centrality metric. On the other hand, it can also be inferred that even though degree centrality metric has a relatively lower strength of relationship between payments and ROI, it cannot be ignored, since it represents important properties like strength of links and number of links. Overall, all the three metrics show a similar pattern in correlation when we vary α and θ parameters, which implies that all the three network topological metrics behave the same with a slight variation in correlation.

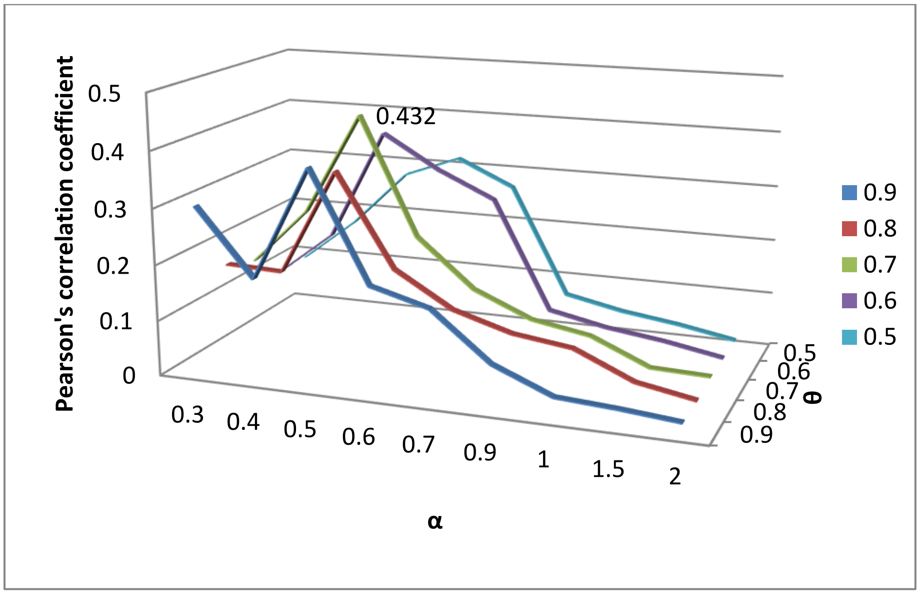


Figure 6.1a: Pearson's correlation coefficient for varying α and θ when the topological metric used is degree centrality metric

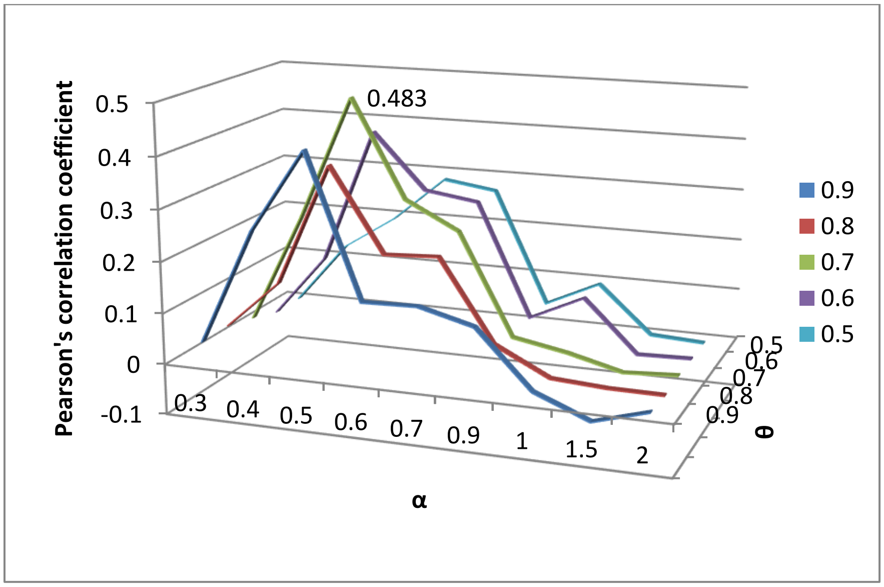


Figure 6.1b: Pearson's correlation coefficient for varying α and θ when clustering coefficient metric is used.

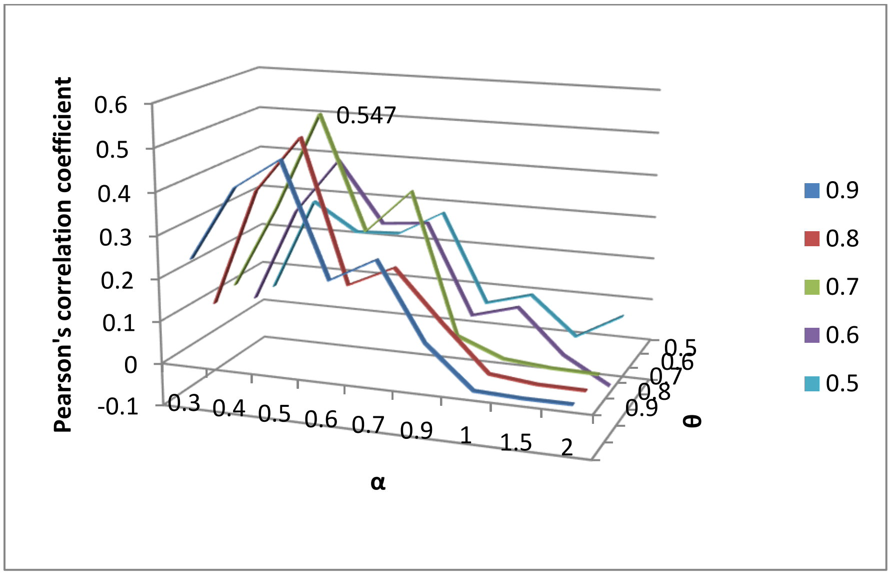


Figure 6.1c: Pearson's correlation coefficient for varying α and θ when closeness centrality metric is used

The second correlation measure used is the Spearman's rank correlation. The behaviour of the three topological metrics are captured in Figure 6.2a, 6.2b and 6.2c respectively. The X and Z axis represent varying α and θ values and the Y axis corresponds to Spearman's rank correlation. It can be seen in Figure 6.2, that there is a similar pattern of Spearman's correlation coefficient, when α and θ were varied, compared to that of Pearson's correlation coefficient, in Figure 6.1. It can be noted that the parameters α and θ were also found to have highest Spearman's rank correlation at α=0.5 and θ=0.7. The highest Spearman's rank correlation was also observed with the closeness centrality metric. We conclude that the overall magnitude of correlation was marginally lower than of Pearson's correlation coefficient. Hence, we will proceed with Pearson's correlation coefficient for further analysis.

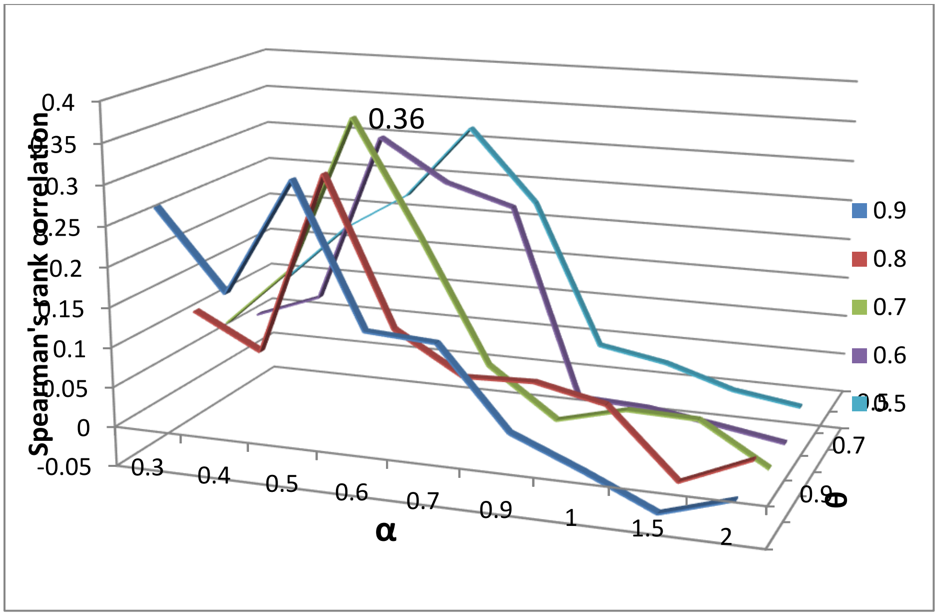


Figure 6.2 (a): Spearman's rank correlation for varying α and θ when the topological metric used is degree centrality metric

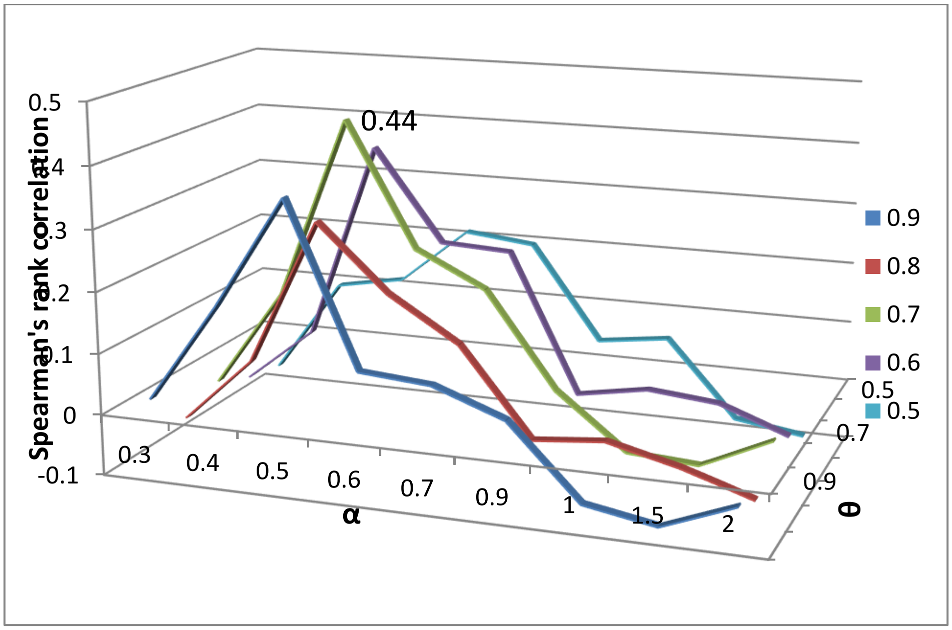


Figure 6.2 (b): Spearman's rank correlation for varying α and θ when the topological metric used is clustering coefficient metic

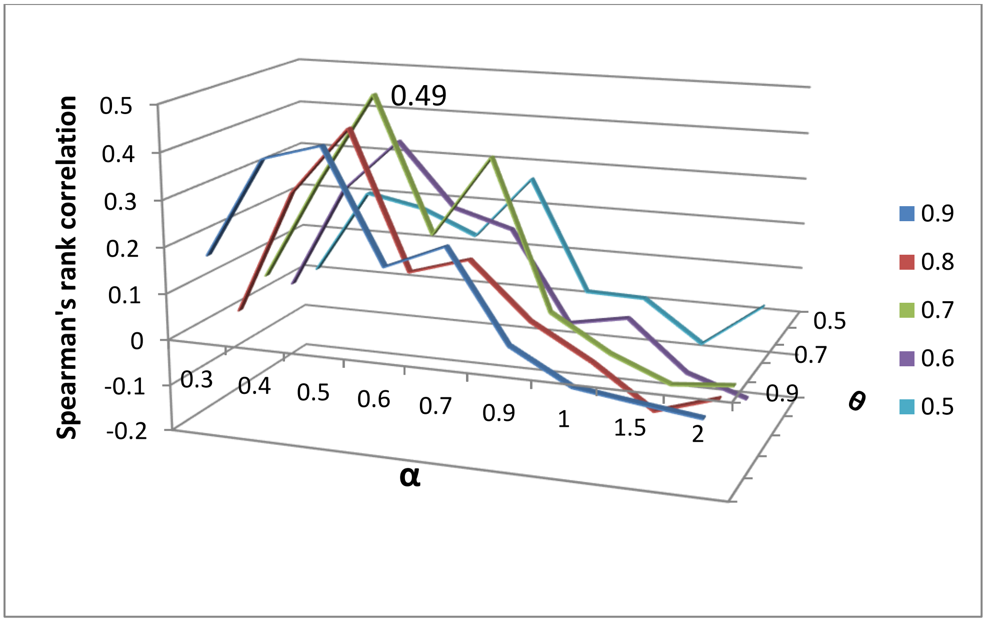
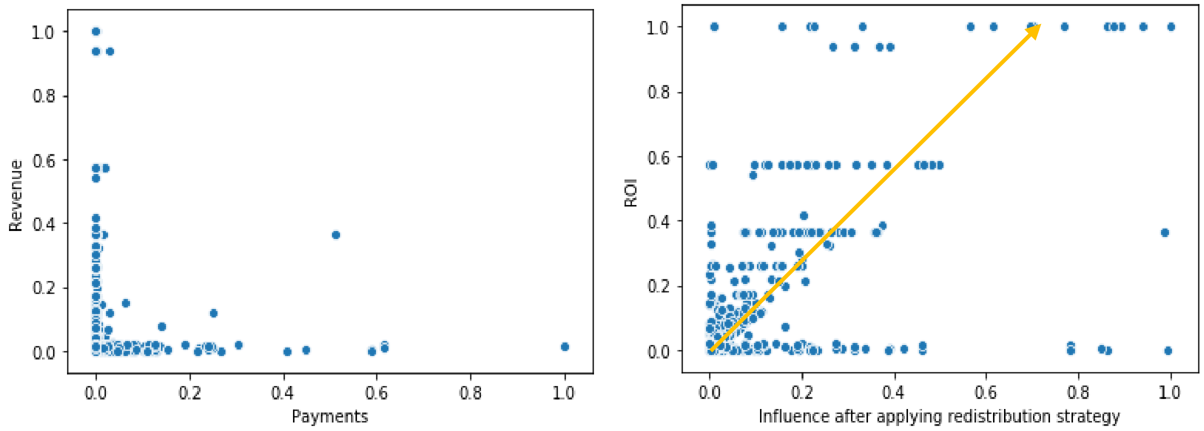


Figure 6.2 (c): Spearman's rank correlation for varying α and θ when the topological metric used is closeness centrality metric

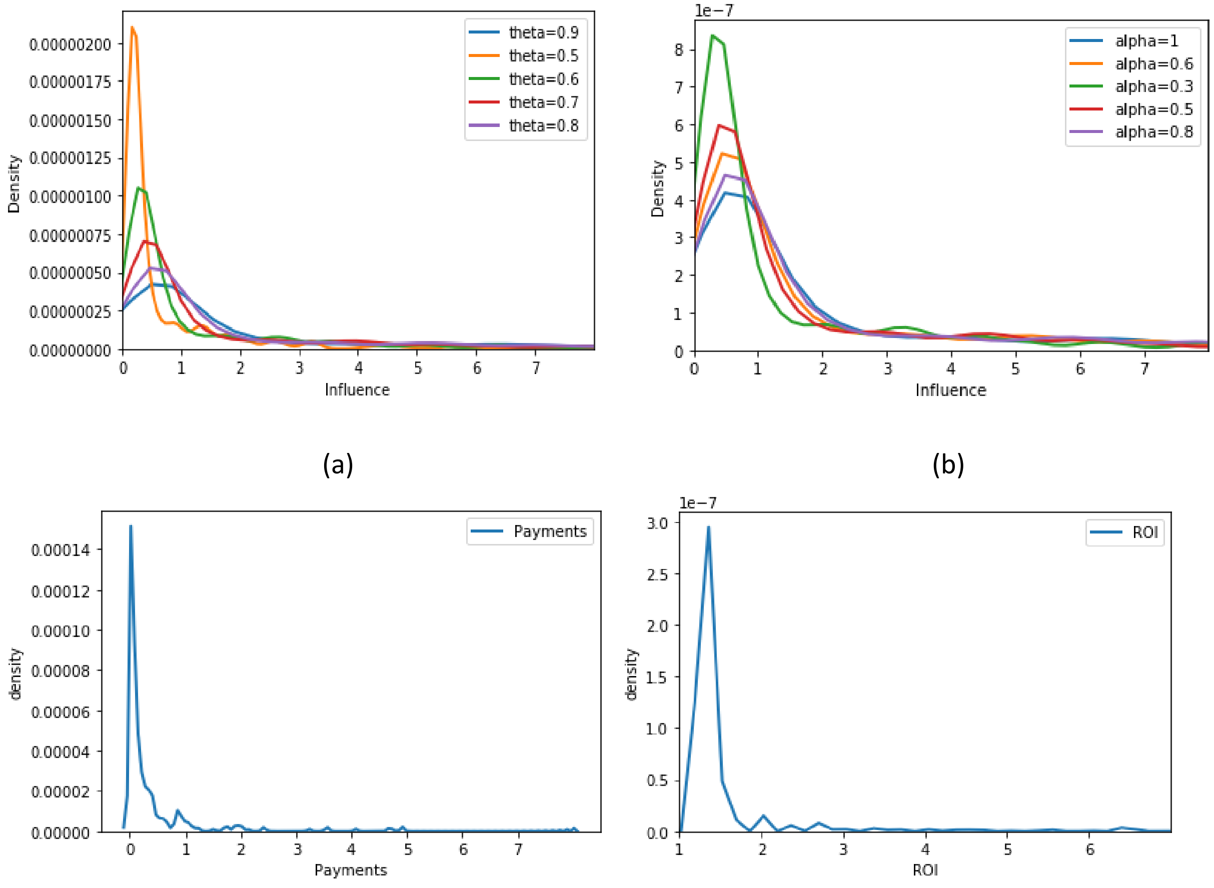
From the Spearman's rank correlation and Pearson's correlation coefficient we concluded that maximum correlation was attained using closeness centrality metric and when α=0.5, θ=0.7. Hence we visualise influence using these parameters. Figure 6.3a and 6.3b shows the scatter plot between payments vs. ROI and Influence (closeness centrality metric, α=0.5, θ=0.7) vs. ROI. In Figure 6.3a we see no linear relationship between payments and ROI whereas in Figure 6.3b we see a comparatively stronger linear relationship between Influence and ROI.

To understand the relation in scatter plot better, we look into the distributions of each of the variables used in Figure 6.3, i.e., influence, ROI and Payments in Figure 6.4. Figure 6.4a shows, the distribution of influence for different values of θ, given the value of α=0.5. It can be seen that there is a lower variation in distribution of θ=0.5 compared to θ=0.7 or θ=0.8. This indicates that when θ=0.5, most nodes have a smaller range of influence. Similarly, Figure 6.4b shows, the distribution of influence for different values of α, given the value of θ=0.7, where the variation in influence is smaller for lower values of α. This indicates that for lower values of α, most nodes have a smaller influence range. Overall, all the distributions look similar, implying that payments, ROI and Influence follow a very similar, positively skewed distribution.



(a) (b)

Figure 6.3 (a) Scatter plot between payments and ROI; (b) Scatter plot between influence after spreading process (closeness centrality metric, α=0.5, ϑ=0.7)



(a) (b)

(c) (d)

Figure 6.4 (a) Distribution of influence for different ϑ values when α=0.5, closeness centrality metric; (b) Distribution of influence for different α values when ϑ=0.7, Closeness centrality metric; (c) Distribution of Payments; (d) Distribution of ROI

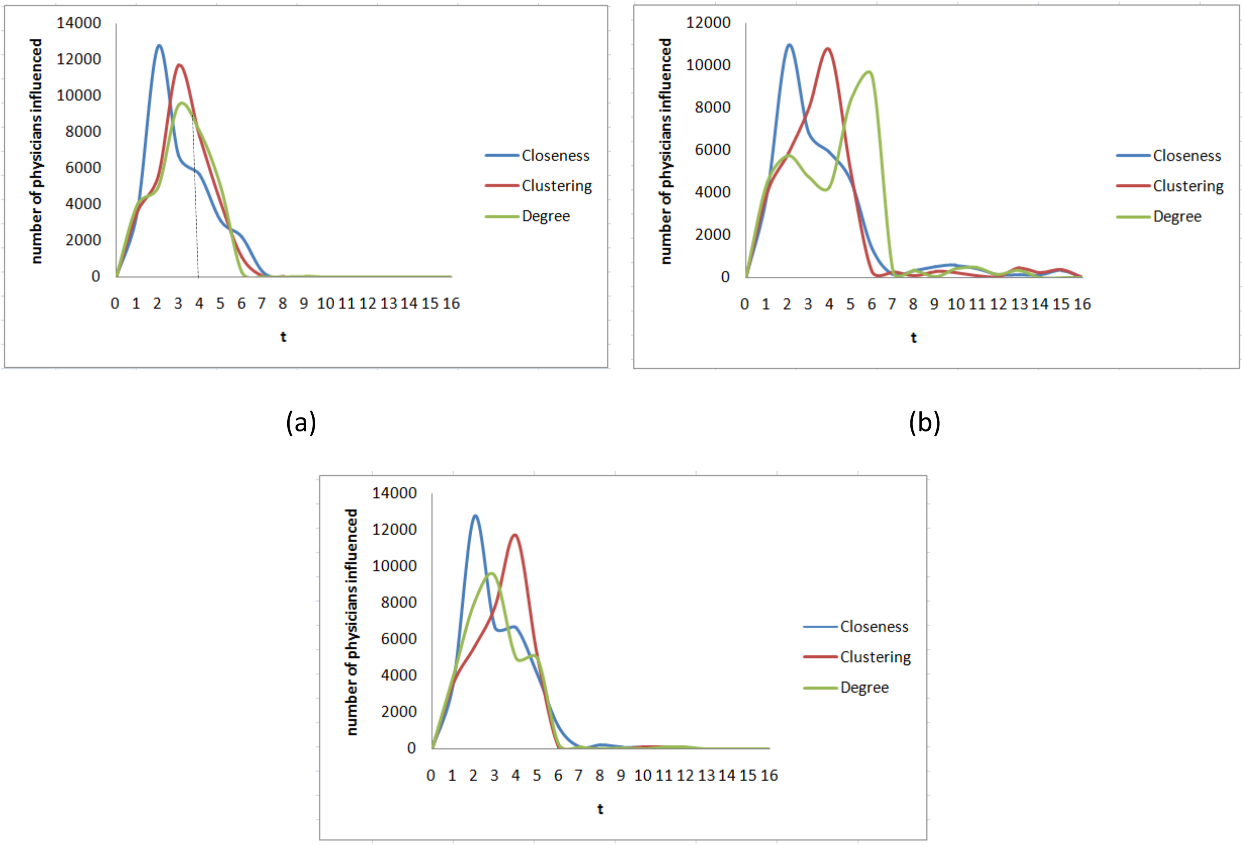
6.4 Time required to influence the physician citation network

The number of nodes influenced at time t is represented in Figure 6.5, where the x-axis represents the time t of the spreading process and y-axis represents the number of nodes influenced during the spreading process. Figure 6.5 (a) represents the time taken by the spreading process for three topological metrics, as explained in section 5.2, when $\alpha = 0.5 \; and \; \theta = 0.7$. On the other hand, figure 6.5 (b) represents the time taken by the spreading process for the same three topological metrics with different parameters, i.e. $\alpha = 2$ and $\theta = 0.5$. The reason behind the choice of parameters comes from the maximum Pearson's correlation coefficient and minimum Pearson's correlation coefficient obtained from the experiments conducted in section 6.3. Finally, Figure 6.5 (c) represents the average behaviour of nodes influenced across a time t, by varying α and θ. The average performance is computed by averaging all the values of α and θ for every timestamp t, across the three topological properties. The reason behind this averaging is to represent the adequate behaviour of the three topological metrics.

It can be seen from Figure 6.5 (a) that more number of nodes are influenced for the closeness centrality metric compared to the other two topological metrics, by $t = 4$. This indicates that with closeness centrality metric, the spreading process diffuses influence to more number of nodes, which can be interpreted as, information spread within the physician citation network is faster with closeness centrality metric than the degree centrality metric and clustering coefficient metric.

From Figure 6.5 (a) and Figure 6.5 (b) we can observe that the spreading process converges at t=8 and t=16 respectively. It can be inferred that the parameter values, i.e. $\alpha \; and \; \theta$, not only contribute to estimating the strength of the relationship between the payments and ROI, but also contribute to the performance of the spreading process.

Overall, from Figure 6.5, we conclude that closeness centrality metric influences more nodes in a short interval of time compared to degree centrality metric which takes more time to influence all the nodes. To understand why the closeness centrality metric performs better than the other two metrics, we look into the distributions of the three topological metrics in Figure 6.5. From section 5.2, it can be recalled that all these metrics are normalized and their values lie between 0 and 1. If they are closer to 1, it means that the topological property has a higher centrality or clustering tendency, i.e., higher potential for information to spread. From Figure 6.6 (c), the distribution of closeness centrality metric shows that there are more number of nodes with a higher closeness centrality value compared to the distribution of the degree centrality metric, from Figure 6.6 (a). The higher values of closeness centrality metric provide an explanation of why the closeness centrality metric diffuses influence to more number of nodes in a shorter time interval.

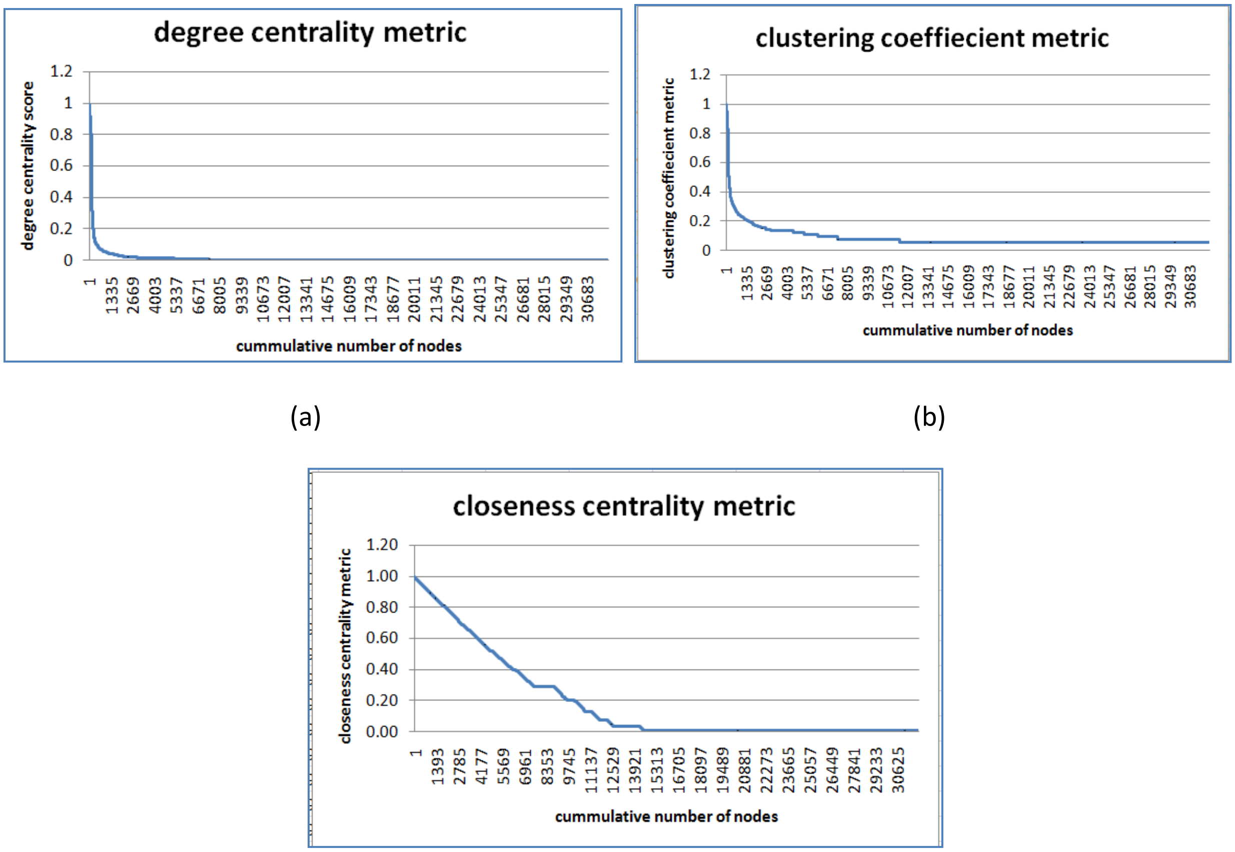


(a)

(b)

(c)

Figure 6.5: Number of nodes influenced over time; (a) number of nodes influenced with the best performing parameters; (b) number of nodes influenced with the worst performing parameters; (c) number of nodes influenced based on the average performance over parameters.



(a)

(b)

(c)

Figure 6.6: Distributions of topological properties before the spreading process; (a) Distribution of degree centrality metric; (b) Distribution of clustering coefficient metric; (c) Distribution of closeness centrality metric.

6.5 Comparative study between Direct and Indirect relationships

In this thesis, two types of relations, direct and indirect, were solved as mentioned in section 1.2. The direct relationship between ROI, payments and the research profile of a physician is addressed through a regression technique. The metric used to measure this relationship is R^2 (coefficient of variation). The R^2 in the regression technique is used to find the relationship between payments and ROI , which is explained in detail in section 4.5, and was found to be 0.0004. To make the direct relation comparable with the indirect relation, we compute r (Pearson's correlation coefficient), which is the square root of R^2. Hence, $r = 0.02$.

To address the indirect relationship, we designed a spreading process to diffuse influence through a physician citation network/physician citation network. The purpose of using the spreading process is to capture the complex indirect relationship which regression technique failed to capture. For which, multiple experiments were performed by varying the topological properties and parameters to estimate the relationship between payments and ROI in a physician citation network. The metric used to capture this relationship is Pearson's correlation coefficient, which is also represented as r was estimated to be 0.55 for the spreading process using the parameters $= 0.5\ and\ \theta = 0.7$, as seen in Figure 6.1.

To compare direct relationship with indirect relationship between payments made to the physicians and ROI earned by the healthcare company, we use the evaluation metric, r, correlation coefficient. The difference in relationship strength can also be seen in Figure 6.3 (a) and Figure 6.3 (b), between payments and ROI for the direct and indirect relationships respectively. From the above two paragraphs, we conclude that the indirect relationship could explain the relationship between payments and ROI 96.36% more than the direct relationship. Hence, it is important for the healthcare company to promote more collaborations amongst physicians in the physician citation network. We also state that the spreading process designed captured the complex indirect relationship between payments and ROI.

6.6 Discussion

In this chapter we discussed about the experiments that we conducted to measure the indirect relationship between investment and the return on investment from the healthcare company, in three different ways, thereby answering the RQ1 using a network spreading process. We also made important conclusions from the experiments conducted in section 6.3, where we found the optimal topological property is closeness centrality metric with optimal parameters $\alpha = 0.5\ and\ \theta = 0.7$ for the physician citation network used. These optimal parameters can be used on any physician citation network in the future, since the network is a scale free network. The parameter $\theta = 0.7$, indicates that when a physician keeps 70% of his/her payment with himself/herself and uses the remaining 30% of the payment to influence his/her neighbours we see maximum influence. The parameter $\alpha = 0.5$, indicates a scaling factor of the topological property. It can be inferred that the influence diffused follows a concave parabola when the payment is linear.

Overall we can state that we saw a significant 96.36% increase in the explainablity of the relationship between payments and ROI, which confirms that our influence diffusion model using a spreading process is a very good model. We concluded the chapter by performing a comparative study between the direct and indirect relationships used to solve the RQ1 of this thesis.

7. PAYMENT REDISTRIBUTION

In this chapter we propose two redistribution methods, to understand if there is a change in the distribution of nodal influence before and after the redistribution of payments amongst the physicians in a physician citation network. The aim of redistribution is to explore if there are other ways of investments made to physicians. We also aim to understand how different are these distributions from the distribution of the original payments made to physicians. By understanding the properties of payments and influence we propose different ways of investments to the healthcare company ABC.

From Chapter 6, we know that maximum correlation between nodal influence and ROI was achieved when α=0.5, θ=0.7 and using closeness centrality metric. In this chapter, we use the same parameters, by making an assumption that the results obtained are true. Using this assumption, we propose two redistribution methods and understand their property of return, to answer RQ3.

RQ3: What is the effect of redistribution methods on the relationship between investment and its return?

Two redistribution methods are proposed, first method, we redistribute payments amongst physicians who have already received an investment from the healthcare company, based on the closeness centrality metric. In the second method, we redistribute payments to nodes who have or have not received payments from the healthcare company in the past, based on their closeness centrality metric.

The redistributed payments are then used as Source nodes to diffuse influence using the spreading process, mentioned in section 5.1. At the end of the spreading process, every node in the network contains some amount of nodal influence. The variation in distribution of the nodal influence after redistribution is compared with the variation in the original investment distribution. We also compute the Pearson correlation coefficient between the nodal influence before redistribution and the nodal influence after redistribution of payments, when the influence is generated from the spreading process with parameters α=0.5, θ=0.7 and using closeness centrality metric. We drop out the Spearman's rank correlation, since we concluded that Pearson's correlation coefficient and Spearman's rank correlation follow the exact same pattern, but produces marginally lower scores compared to Pearson, which is explained in Chapter 6.

This chapter is organized as follows, in section 7.1 we discuss the method of redistribution of the payments and how it is useful for the healthcare company. This is followed by section 7.2 which illustrates the design, analysis and outcomes of the first method and section 7.3 demonstrates the design, analysis and outcomes of the second method. In section 7.4, we compare the two methods and provide a recommendation to the healthcare company.

7.1 Methods to redistribute payments

We elucidate the redistribution of investments in a physician citation network as a method of assigning payments to different set of physicians in the network, based on their closeness centrality metric. For which, we propose two different methods to illustrate the possibilities of redistribution of investment.

- **Redistribution of payments amongst the Source nodes in the network.**
  In this redistribution method, we rank all the Source nodes based on their closeness centrality metric. The payment attributes of these Source nodes are updated with the highest payments made by the healthcare company, based on their ranks, i.e., the source node with the highest rank receives the highest payment.

- **Redistribution of payments amongst all the nodes in a physician citation network.**
  In this method, we rank all the nodes in the physician citation network, i.e., Source and the remaining nodes in the network, based on their closeness centrality metric. The payments made by the healthcare company are also ranked. In this redistribution method, the highest payments are assigned to the nodes that have the highest topological property. This implies that only 3620 nodes in the network will receive payments since the healthcare company has only made 3620 payments in the past, and we can only redistribute the already existing payments.

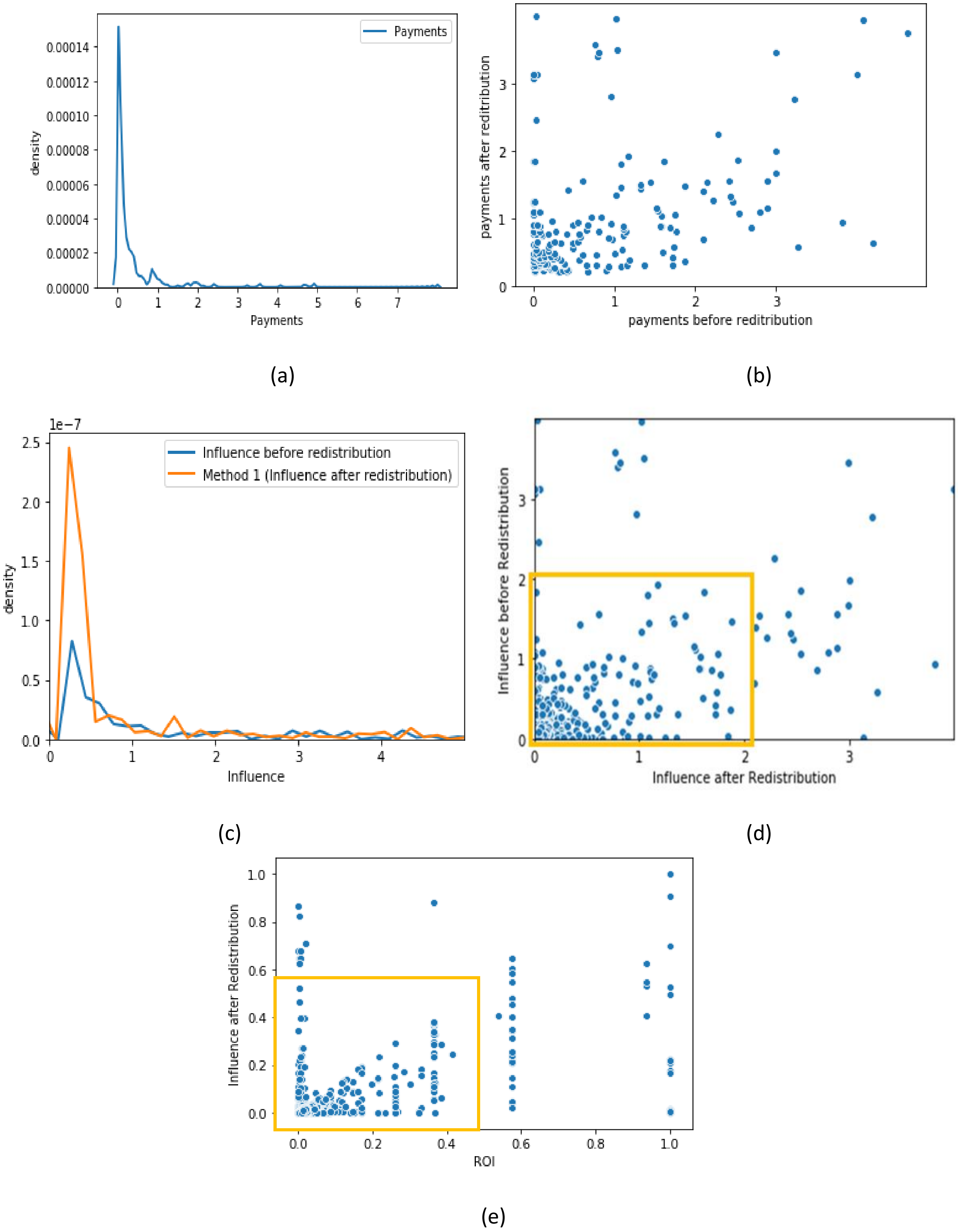
These two methods are discussed and analyzed in the following sections.

7.2 Redistribution of payments amongst the source nodes

As mentioned in section 7.1, the payments are redistributed amongst the source nodes based on their closeness centrality metric, which is mentioned in section 5.2.3. The purpose of this redistribution is to understand the properties of nodal influence before and after the redistribution of investments, by analyzing their respective distributions. To address this purpose, we perform the redistribution of payments amongst source nodes and evaluate the redistribution by understanding the relationship between payments before and after redistribution, influence before and after redistribution, ROI and influence after redistribution. These relationships can be understood with the help of scatter plots, distributions and Pearson correlation coefficient. The redistribution method is performed in the following steps:

- The source nodes are ranked based on their closeness centrality metric, since the healthcare company ABC wants a recommendation on the best performing topological property.
- Applying the redistribution method, the payment attribute of the source nodes with the highest rank are reinitialized with the highest payments made by the healthcare company.
- We initiate the spreading process with reinitialized payment attributes and with the best performing metric, i.e., closeness centrality metric and the best performing parameters $\alpha = 0.5$ and $\theta = 0.7$, as mentioned in section 6.3.

- After the completion of the spreading process, as an evaluation step, we attempt to understand three relationships using scatter plots, distributions and Pearson correlation coefficient as discussed below.
    1. We find the relationship between original payments made to physicians and the payments after redistribution and before the spreading process is completed. Figure 7.1a shows the scatter plot between original payments and payments after redistribution, from which we see a strong relationship between the two variables. Figure 7.1b shows the distribution of payments which is a positively skewed distribution with a long tail and is heterogeneous in nature. To understand this relationship, we cannot use Pearson correlation coefficient because original payments and payments after redistribution are not two independent variables.
    2. After the spreading process is completed, we find the relationship between nodal influence before the redistribution of payments made to physicians and the nodal influence generated after the redistribution, with the optimum parameters, as mentioned in the third bullet point. Figure 7.1c shows the distribution of nodal influence before and after redistribution, and when the spreading process is completed. We infer that both the distributions are heterogeneous in nature, and are positively skewed with a long tail. Figure 7.1d shows a scatter plot where we see a strong relationship between the two variables only in the third quadrant. Similar to the previous evaluation, we do not use Pearson correlation coefficient because influence is a single variable.
    3. We find the relationship between ROI and influence obtained after redistribution, after the spreading process. To evaluate this proposal we would have to put the proposal into action to measure the ROI for the following year, which is not possible. Hence, we make an assumption that, the healthcare company ABC received the same ROI the following year. Given the assumption is true, we compute the aggregated influence at the hospital level from Eq.(26) since ROI is at the hospital level. We then compute the Pearson correlation coefficient to estimate the relationship between ROI and influence after redistribution, which was found to be 0.587. We observe that, this correlation is slightly higher than the correlation obtained before redistribution of influence using the same spreading parameters. Figure 7.1e shows the scatter plot between ROI and aggregated influence after redistribution, from which we see strong correlation only in the third quadrant and higher values of return are independent of influence.



(a)

(b)

(c)

(d)

(e)

Figure 7.1: (a) Distribution of original payments; (b) Scatter plot between original payments and payments after redistribution; (c) Distribution of influence before and after the redistribution of influence; (d) Scatter plot between influence before and after redistribution; (e)Scatter plot between aggregated influence after the redistribution strategy and ROI;

To investigate the impact of the proposed method further, we divided the payments into 3 sections as shown in the below table.

Number of ***physicians receiving higher than their original payments***	1748 (57%)
Number of ***physicians receiving lower than their original payments***	965 (32%)
Number of ***physicians receiving the same payments as before***	347 (11%)

Table 7.1 Distribution of physicians receiving higher or lower payments

As we can see from the above table that more than 57% of physicians receive higher than their original payments and 32% of physicians receive lower than their original payments. It is indeed uncertain, by how much percentage higher or lower the redistributed payments are, compared to the original payments. This deviation in payments is represented in percentages and is demonstrated in Figure 7.2.

In Figure 7.2, the blue line represents payments that are higher than the original payments. It can be noted that more than 1400 physicians of the 1748 physicians receive an increment of up to 10% of the original payments. Similarly, the red line in Figure 7.2 represents payments that are lower than the original payments. It can be noted that over 700 physicians of the 965 physicians receive a decrement of up to 10% of the original payments.

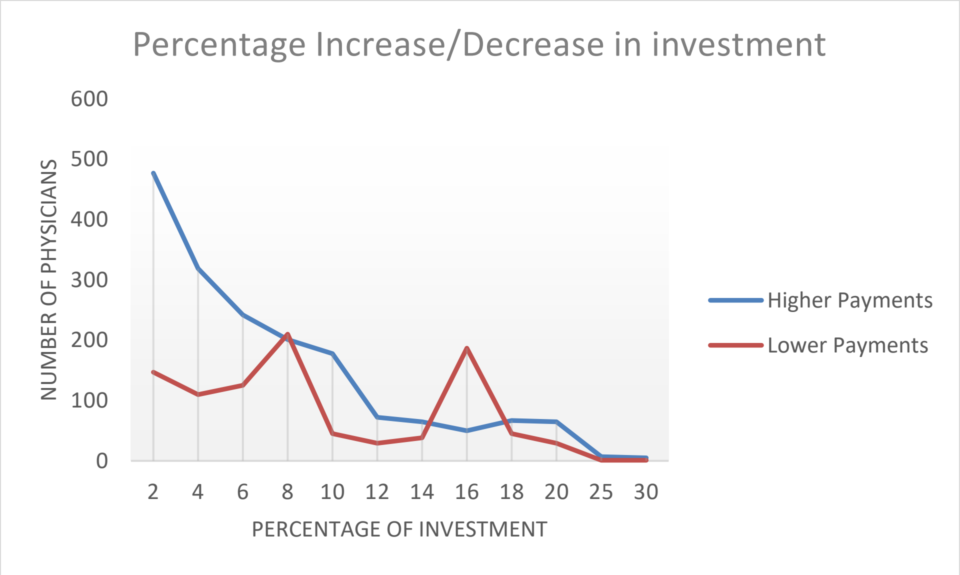


Figure 7.2: Percentage increase or decrease in the investments made to physicians using the redistribution method

It can also be noted that the $x\%$ increase is w.r.t to the payments and from Table 3.1, we can take note that there is a wide variation in payments. Hence it is irrational to recommend the healthcare company to increase payments to all physicians by $x\%$. So we divide the payments along the median set with a percentage threshold of 10%, which is obtained from a business knowledge/input.

Table 7.2 represents two contingency tables, where the table with variables displayed in blue represent payments that correspond to the blue line in Figure 7.2 and the table with variables displayed in red represent payments that correspond to the red line in Figure 7.2.

	Below 10%	Above 10%
Investment below median	1096	267
Investment above median	320	65
	Below 10%	Above 10%
Investment below median	567	73
Investment above median	147	78

Table 7.2 Contingency tables for higher and lower payments

It is observed from Table 7.2 that majority of the increased and decreased payments lie below the median, which indeed confirms the low amount of increments or decrements in the payments required to improve the relationship between ROI and recommended payments made to physicians.

7.3 Redistribution of payments amongst all nodes

As mentioned in section 7.1, the payments are redistributed amongst all the nodes in the physician citation network, based on its closeness centrality metric, which is mentioned in section 5.2.3. The purpose of this redistribution is to understand the properties of nodal influence before and after the redistribution of investments, by analyzing their respective distributions. To address this purpose, we perform the redistribution of payments amongst all nodes and evaluate the redistribution by understanding the relationship between payments before and after redistribution, influence before and after redistribution, ROI and influence after redistribution. These relationships can be understood with the help of scatter plots, distributions and Pearson Correlation Coefficient. The redistribution method is performed in the following steps:

- All the nodes are ranked based on their closeness centrality metric, since the healthcare company ABC wants a recommendation on the best performing topological property.
- Applying the redistribution method, the payment attributes, of only the nodes with the highest ranks are reinitialized with the highest payments made by the healthcare company. We initiate the spreading process with the reinitialized payment attributes and using the best performing metric, i.e. closeness centrality metric and the best performing parameters $\alpha = 0.5$ and $\theta = 0.7$, as mentioned in section 6.3. This implies that only 3620 nodes in the network will receive

payments since the healthcare company only made 3620 payments in the past, and we can only redistribute the already existing payments.

- After the completion of the spreading process, as an evaluation step, we attempt to understand three relationships using scatter plots, distributions and Pearson correlation coefficient as discussed below.

    1. We find the relationship between original payments made to physicians and the payments after redistribution and before the spreading process. Figure 7.3a shows the scatter plot between original payments and payments after redistribution, from which we see a strong relationship between the two variables. Figure 7.3b shows the distribution of payments which is a positively skewed distribution with a long tail. The variation in distribution shows the heterogeneous nature of payments. To understand this relationship, we cannot use Pearson correlation coefficient because the original payments and payments after redistribution are not two independent variables.
    2. After the spreading process is completed, we find the relationship between nodal influence before the redistribution of payments made to physicians and the nodal influence generated after the redistribution, with the optimum parameters, as mentioned in the second bullet point. Figure 7.3c shows the distribution of nodal influence before and after redistribution, when the spreading process is completed. We infer that both the distributions are heterogeneous in nature, and are positively skewed with a long tail. Figure 7.3d shows a scatter plot where we see a strong relationship between the two variables only in the third quadrant. Similar to the previous evaluation, we do not use Pearson correlation coefficient because influence is a single variable.
    3. We find the relationship between ROI and influence obtained after redistribution, when the spreading process is completed. To evaluate this proposal we would have to put the proposal into action to measure the ROI for the following year, which is not possible. Hence, we make an assumption that, the healthcare company ABC receives the same ROI the following year. Given the assumption is true, we compute the aggregated influence at the hospital level from Eq.(26) since ROI is at the hospital level. We then compute the Pearson correlation coefficient to estimate the relationship between ROI and influence after redistribution, which was found to be 0.679. We observe that, this correlation is slightly higher than the correlation obtained before redistribution of influence using the same spreading parameters. Figure 7.3e shows the scatter plot between ROI and aggregated influence after redistribution, from which we see strong correlation only in the third quadrant and higher values of return are independent of influence.

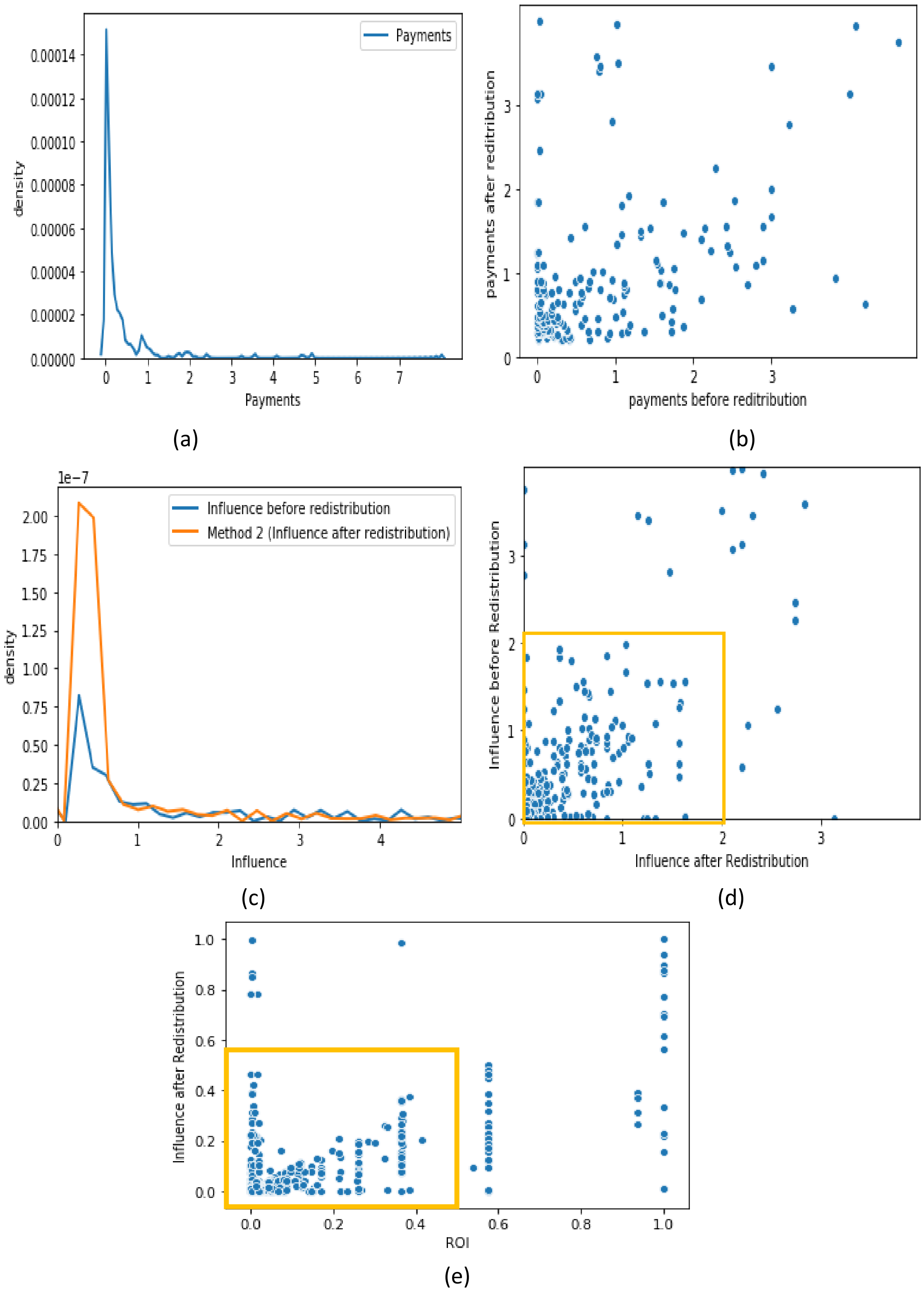


Figure 7.3: (a) Distribution of original payments; (b) Scatter plot between original payments and payments after redistribution; (c) Distribution of influence before and after the redistribution of influence; (d) Scatter plot between influence before and after redistribution; (e) Scatter plot between aggregated influence after the redistribution strategy and ROI;

To investigate the impact of the proposed method, we divided the payments into three sections as shown in Table 7.3.

Number of **physicians receiving higher than their original payments**	1882(61.5%)
Number of **physicians receiving lower than their original payments**	869(28.4%)
Number of **physicians receiving the same payments as before**	309(10%)

Table 7.3 Distribution of physicians receiving higher or lower payments

As we can see from the above table that more than 61% of physicians receive higher than their original payments and 28% of physicians receive lower than their original payments. It is indeed uncertain by how much percentage higher or lower the redistributed investments are compared to the original investments. This deviation in payments is represented in percentages and is demonstrated in Figure 7.3.

In Figure 7.3, the blue line represents payments that are higher than the original payments. It can be noted that more than 1500 physicians of the 1882 physicians receive an increment of up to 10% of the original payments. Similarly, the red line in Figure 7.3, represents payments that are lower than the original payments. It can be noted that over 660 physicians of the 869 physicians receive a decrement of up to 10% of the original payments.

Another interesting insight is that 294 physicians who received investments due to this redistribution method haven't received any payments by the healthcare company ABC before. It can be recommended that these physicians can be approached for future collaborations.

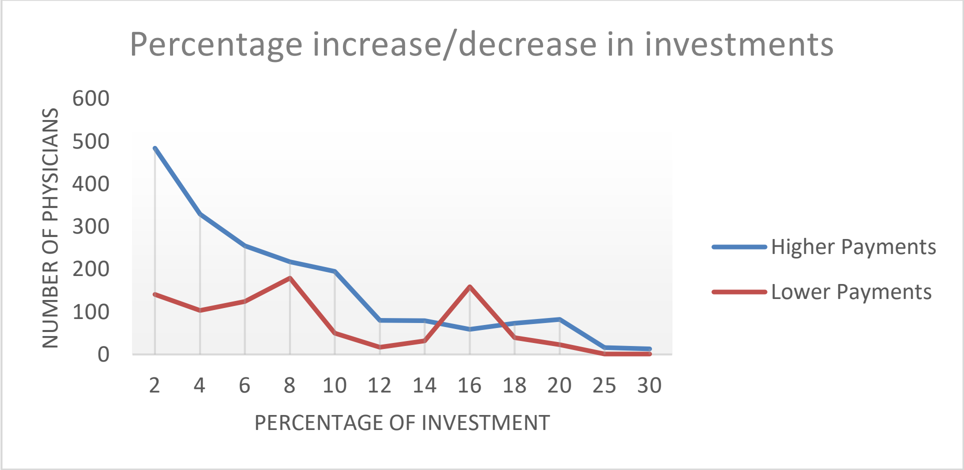


Figure 7.4: Percentage increase or decrease in the investments made to physicians using second method

Similar to the reason mentioned in section 7.2, we divided the investments along its median amount and percentage threshold by 10% from business knowledge.

Table 7.4 represents two contingency tables, where the table with variables displayed in blue represent payments that correspond to the blue line in Figure 7.3 and the table with variables displayed in red represent payments that correspond to the red line in Figure 7.3.

	Below 10%	Above 10%
Investment below median	1168	312
Investment above median	375	27
	Below 10%	Above 10%
Investment below median	428	232
Investment above median	177	32

Table 7.4 Contingency table for higher and lower payments

It is observed from the above table that majority of increased and decreased payments lie below the median, which indeed confirms the low amount of increments or decrements in the payments required to improve the relationship between ROI and recommended payments made to physicians.

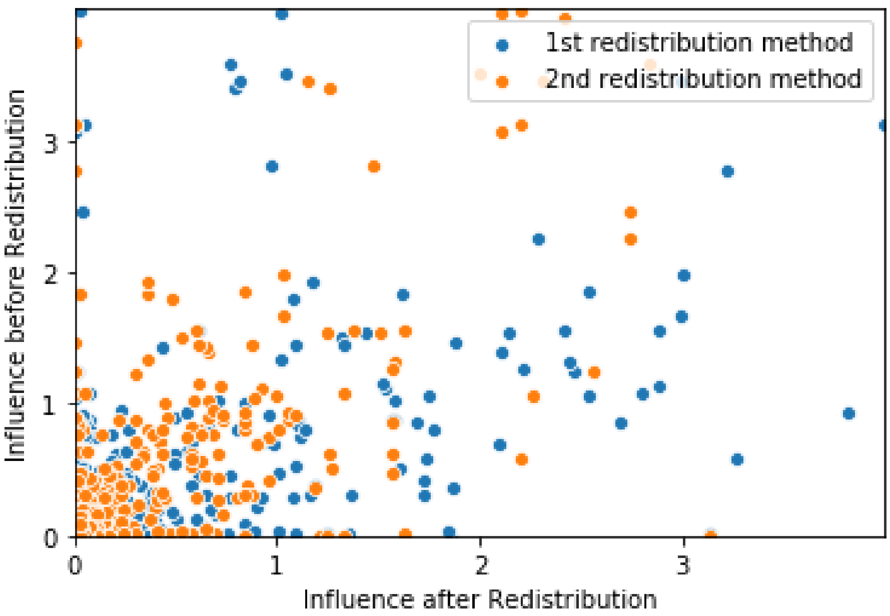
7.4 Comparison of the two methods

From section 7.2 and 7.3 we summarize that, the two proposed methods have a heterogeneous distribution of payments, ROI and influences generated by the two methods before and after the redistributions.

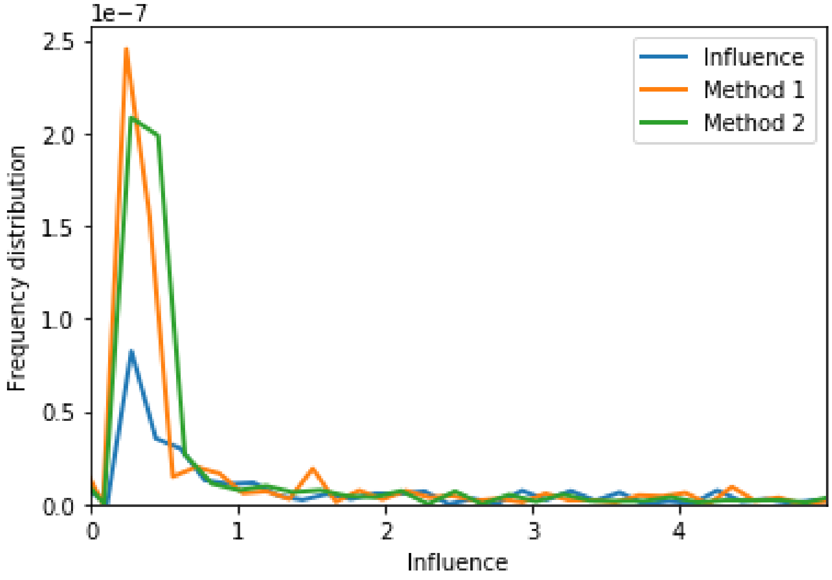
In section 7.2 we suggest the first redistribution method, where we propose investments made to physicians which are different from the original investments done in the past. On the other hand in section 7.3 we suggest another method where we not only propose investments to physicians which are different from the original investments but also recommend physicians, on whom future investments can be made.

Figure 7.5 (a) shows a scatter plot, from which we see a difference in influence scatter in the two different redistribution methods. We infer that the two relationships differ from each other in influence diffusion. Figure 7.5 (b) shows the distribution of nodal influence of the two redistribution methods and the distribution of influence before the redistribution. It can be observed that the distributions corresponding to the two redistribution methods have lower variance compared to the distribution of the distribution of influence before redistribution. This makes the nature of influence after redistribution less heterogeneous compared to the influence before redistribution.

In addition, the correlation computed from second method is 5.4 percentage points more than that of the first method. Hence, we can conclude that the redistribution of the second method is more useful than the redistribution of the first method and we recommend the same to the healthcare company.



(a)



(b)

Figure 7.5:(a) Difference in scatter plot between influence after first and second redistribution with the influence generated before redistribution with parameters, closeness centrality metric, α=0.5, ϑ=0.7; (b) Distribution of influence generated from the spreading process (closeness centrality metric, α=0.5, ϑ=0.7) (blue), distribution of influence after the first redistribution method from section 7.2 (orange) and influence after the second redistribution method from section 7.3 (green)

7.5 Discussion

In this chapter, to answer **RQ3**, we proposed two redistribution methods to understand if the properties of nodal influence before and after the redistribution of investments are the same or different, by analyzing their respective distributions. We also compared the two redistribution methods in section 7.4 and observed the difference between the two methods.

8. CONCLUSIONS, FUTURE WORK and LIMITATIONS

This chapter highlights the key findings from this thesis. It also explains major limitations and challenges that were overcome to complete this thesis.

8.1 Conclusions

We successfully addressed the problem statement, mentioned in section 1.2, by developing a regression model and influence diffusion model for the direct and indirect relationship respectively. For which we came up with three research questions to answer the problem statement.

The first research question (RQ1) is to estimate the relationship between investments made to physicians, return on investment and the research profile of the physician. This has been successfully achieved and concluded in section 4.3 and section 6.3 by using two different models to explain the direct and indirect relationship.

In the direct relationship, we used regression analysis to estimate the relationship between investment, research profile and ROI for the direct influence scenario, as seen in section 1.2. We discovered that there is no direct relationship between payments, ROI and the research profile of the physician.

In the indirect relationship, we proposed a network spreading process to estimate the indirect relationship, as seen in section 1.2, between investment, research collaboration and ROI generated to the healthcare company. A comparative study between the models for direct and indirect relationship was performed and it was found that the model for indirect relationship explains the relationship between payments and ROI 96.3% more than the model for direct relationship.

The second research question (RQ2) is to identify the payment strategies of the three healthcare companies. To answer this question we made use of a regression analysis to decode the underlying payment method used by these companies. We discovered that, healthcare company ABC, XYZ and LMN have their own payment strategies with varying influx points of h-index. We also discovered that the three healthcare companies invest on physicians based on the physician's years of experience.

To answer the third research question (RQ3) we propose two alternative payment redistribution methods in order to understand how the payment redistribution method affects the properties of the resultant nodal influence and their relationship between investment and nodal influence. Our findings may inspire the healthcare companies to design their future investments made to physicians. We also suggest potential physicians, who previously did not receive payments, and are currently contributing to the research as they have strong research collaborations with the physicians who are already receiving payments from the healthcare company.

Last but not the least, our methodologies exemplified in this thesis, such as the regression technique and influence diffusion modelling on a physician citation network to perform deep analysis can be widely applied to other systems to explain the direct and indirect relationship between payments and return at a hospital level and in general, between input and output.

8.2 Future Work

In this thesis, we designed a network spreading process to model influence diffusion through a physician citation network. This serves as a baseline model, to understand the effectiveness of the network. However there are numerous possibilities that are worth studying in the future and here are a few to mention.

- The physician citation network has been restricted to healthcare company ABC in this thesis. It can be extended by adding physicians who are being paid by other peer healthcare companies. This provides room to multiple comparative research and business problems that can be answered by analyzing the network spreading process. One such research problem is, clusters of the dominant healthcare company can be identified, from which the choice of the physicians on whom the company invests or divest in the future could be decided. In short, the network spreading process can be used as a payment validation tool.
- The physician citation network used in this thesis carries minimal and crucial information. Adding more information into nodes in the network, like categorizing the physicians based on their field of work, investment etc or adding more useful and independent research metrics that help explain the relationship between the investment and profile of physicians better is another future study.
- In this thesis, the θ parameter, introduced in chapter 5, is kept constant across the physicians, which is not a true imitation of the real world scenario. Hence, in future we aim to develop dynamic θ values to spread the influence to physicians based on different criteria like performance, strength, etc.

8.3 Limitations

Following are some limitations that were overcome in different phases of this thesis:

- Multiple anomalies exist in missing data. Combining data from three different sources has a lot of missing values, dummy values and duplicates. Hence during the data cleaning process a lot of data points were eliminated.
- Scopus has a restriction limit for data scraping and the process is highly time consuming.
- Lack of return on investment data for all companies, limits the research abilities.
- Lack of return on investment data for hospitals leads to a confusion if they are actual missing values or the healthcare company made an investment and did not procure a return on investment. This is an important drawback since it has a major impact on the company's investment.

APPENDIX

A1. Model Specification including the assumptions made on the error/random variable and Estimation technique of parameters

Basic regression model in its linear form is specified as $y_i = \alpha + \beta\, x_i + \epsilon_i$, where y_i is the dependent variable, variation in which is explained by an explanatory variable x_i and ϵ_i is the disturbance term as the specification is not deterministic in nature.

Since the disturbance term is unobservable, assumptions made about ε are termed as basic assumptions of regression. These assumptions are

(i) $E(\varepsilon_i)=0$ (Zero mean);

(ii) $var(\varepsilon_i)= \sigma^2$(homoscedastic errors) where σ^2 is the variation in the dependent variable for a given value of x;

(iii) $E(\varepsilon_i,\varepsilon_j)=0$for all $i \neq j$ (no autocorrelation);

(iv) $E(x_i,\varepsilon_i)=0$ (non stochastic x);

(v) ε_i are normally distributed for all values of i.

The objective of least squares is to choose $\hat{\alpha}\ and\ \hat{\beta}'s$ as estimates of α and β's so that error sum of squares is minimized. i.e., $Q = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ attains minima.

The estimates $\hat{\beta} = \sum_{i=1}^{n}\frac{x_i y_i}{x_i^2}$ and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ obtained through the method of least squares are best linear unbiased estimates.

In the multiple regression framework the error sums of squares minimized is $= \sum_{i=1}^{n}(y_i - \hat{\alpha} - \widehat{\beta_1}x_1 - \cdots - \widehat{\beta_k}x_k)^2$. In matrix notation, the parameter estimates are $\hat{\beta} = (X'X)^{-1}(X'Y)$, where $\hat{\beta}$ is a vector of parameters. Here as stated above the vector of parameters $\hat{\beta}$ satisfy requiredproperties of being unbiased, consistent and efficient.

Since ε values are independently normally distributed with mean zero and variance σ^2, the y values are also independent normally distributed with mean $\hat{y}$and variance σ^2. Further $\hat{\beta}$, which is a linear function of yfollows the distribution of y. Hence the distribution of $\hat{\beta}$ is normal with $E(\hat{\beta}) = \beta$ and $var(\hat{\beta}) = (X'X)^{-1}\sigma^2$ i.e., $\hat{\beta}$~N(β, $(X'X)^{-1}\sigma^2$). Since the distribution of$\hat{\beta}$ is known, statistical inference on the estimated coefficient is done through t-test, $\frac{\widehat{\beta_i}}{se(\widehat{\beta_i})}$, which helps to check if the variable attached to the coefficient has any influence in explaining the variation in y.

Further,$R^2 = \frac{Explained\ sums\ of\ square}{Total\ sums\ of\ square}$, indicates the amount of variation explained by the set of explanatory variables i.e., it would give a measure of goodness of fit. Value of R^2 lies between zero and one. Higher the value better is the strength of the equation. For example if R^2 is say, 0.96 can be interpreted as that 96% of the variation in the dependent variable is explained by the explanatory variables included in the linear model.

REFERENCES

[1] Ramsey, J.B., (1969), “Tests for Specification errors in Classical Linear Least Squares Regression Analysis, ”Journal of the Royal Statistical Society, Series B, Vol31, pp 350-371.

[2] Breusch, T S and Pagan, A.R (1979), Ä Simple test for Heteroscedasticity and Random coefficient variation”, Econometrica, Vol 47, pp 1287-1294.

[3] White, H., (1980), “A Heteroscedastic Consistent Covariance Matrix Estimator and a Direct test of Heteroscedasticity,” Econometrica, Vol 48, pp817-838.

[4] D. A. Belsley, E. Kuh and R. E. Welsch, “Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,” John Wiley & Sons, Ltd., New York, 1980.

[5] Cook, R.D (1977), “Detection of Influential Observations in Linear Regression”, Technometrics, American Statistical Association, 19(1): 15-18.

[6] Zeller, A., Kmenta, J., and Dreeze, J.,(1966), “Specifications and estimation of Cobb-Douglas Production Function Models”, Econometrica, 34, No.4, pp.784-795.

[7] Maddala, G S., (1977), Econometrics, , New York: McGraw Hill.

[8] Godfrey, L.G and Wickens, M R, (1981) “Testing Linear and Log Linear Regression Functional Form’, Review of Economic Studies, Vol 48, No3, pp 487-496.

[9] Gilstein C Z and Leamer, E E., 1993, “Robust Sets of Regression Estimates”, Econometrica, Vol 51, No.2, p330.

[10] Rao., C R, 1973, Linear Statistical Inference and its Application, 2nd Ed., New York, John Wiley & Sons, Inc.

[11] Rao., C R, 1970, “Estimation of Heteroscedastic Variances in Linear Models,” Journal of American Statistical Association, Vol 65, pp 161-172.

[12] Hirsch, J.E., 2005, An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, *102*(46), pp.16569-16572.

[13] Stephen M Stigler (1981), "Gauss and the Invention of Least Squares", The Annals of Statistics, vol9, No3, pp 465-474.

[14]https://www2.deloitte.com/global/en/pages/life-sciences-and-healthcare/articles/global-health-care-sector-outlook.html

[15] https://openpaymentsdata.cms.gov/summary

[16] https://www.elsevier.com/solutions/scopus

[17] https://www.cms.gov/openpayments/

[18] Shi, C., Li, Y., Zhang, J., Sun, Y. and Philip, S.Y., 2016. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, *29*(1), pp.17-37.

[19]ToreOpsahl and PietroPanzarasa (2009). "Clustering in Weighted Networks". Social Networks. 31 (2): 155–163. doi:10.1016/j.socnet.2009.02.002

[20] Newman, M.E., 2004. Analysis of weighted networks. *Physical review E*, *70*(5), p.056131.

[21] Ganis, Matthew; Kohirkar, Avinash (2015). *Social media Analytics: Techniques and insights for Extracting Business Value Out of Social Media*. New York: IBM Press. pp. 40–137. ISBN 978-0-13-389256-7.

[22] Bonacich, Phillip (1987). "Power and Centrality: A Family of Measures". *American Journal of Sociology*. **92** (5): 1170–1182. doi:10.1086/228631

[23] Onnela, J.P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J. and Barabási, A.L., 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, *104*(18), pp.7332-7336.

[24] Barabási, A.L. and Albert, R., 1999. Emergence of scaling in random networks. *science*, *286*(5439), pp.509-512.

[25] Tang, X., Wang, J., Zhong, J. and Pan, Y., 2013. Predicting essential proteins based on weighted degree centrality. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *11*(2), pp.407-418.

[26] Kline, R.M., Bazell, C., Smith, E., Schumacher, H., Rajkumar, R. and Conway, P.H., 2015. Centers for Medicare and Medicaid Services: using an episode-based payment model to improve oncology care. *Journal of oncology practice*, *11*(2), pp.114-116.

[27] Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, *102*(46), pp.16569-16572.

[28] Davis, K. and Rowland, D., 1986. Medicare policy. *New Directions for Health and Long-Term Care (Baltimore: Johns University Hopkins Press, 1986)*.

[29] https://www.salesforce.com/

[30] https://dev.elsevier.com/tips/AuthorSearchTips.htm

[31] Chambers, D., Wilson, P., Thompson, C. and Harden, M., 2012. Social network analysis in healthcare settings: a systematic scoping review. *PloS one*, *7*(8), p.e41911.

[32] Mascia, D., Cicchetti, A. and Damiani, G., 2013. "Us and Them": a social network analysis of physicians' professional networks and their attitudes towards EBM. *BMC health services research*, *13*(1), p.429.

[33] Bridewell, W. and Das, A.K., 2011. Social network analysis of physician interactions: the effect of institutional boundaries on breast cancer care. In *AMIA Annual Symposium Proceedings*(Vol. 2011, p. 152). American Medical Informatics Association.

[34] Royak-Schaler, R., Passmore, S.R., Gadalla, S., Hoy, M.K., Zhan, M., Tkaczuk, K., Harper, L.M., Nicholson, P.D. and Hutchison, A.P., 2008, September. Exploring patient-physician communication in breast cancer care for African American women following primary treatment.In *Oncology nursing forum* (Vol. 35, No. 5).

[35] Bridewell, W. and Das, A.K., 2011. Social network analysis of physician interactions: the effect of institutional boundaries on breast cancer care. In *AMIA Annual Symposium Proceedings*(Vol. 2011, p. 152). American Medical Informatics Association.

[36] Sabot, K., Wickremasinghe, D., Blanchet, K., Avan, B. and Schellenberg, J., 2017. Use of social network analysis methods to study professional advice and performance among healthcare providers: a systematic review. *Systematic reviews*, *6*(1), p.208.

[37] Soffer, S.N. and Vazquez, A., 2005. Network clustering coefficient without degree-correlation biases. *Physical Review E*, *71*(5), p.057101.

[38] Peres, R., Muller, E. and Mahajan, V., 2010. Innovation diffusion and new product growth models: A critical review and research directions. *International journal of research in marketing*, *27*(2), pp.91-106.

[39] Faden, R.R. and Beauchamp, T.L., 1986. *A history and theory of informed consent*. Oxford University Press.

[40] Bentley, J.P. and Thacker, P.G., 2004. The influence of risk and monetary payment on the research participation decision making process. *Journal of medical ethics*, *30*(3), pp.293-298.

[41] Cornett, M.M. and Saunders, A., 2003. *Financial institutions management: A risk management approach*. McGraw-Hill/Irwin.

[42] Kalna, G. and Higham, D.J., 2007. A clustering coefficient for weighted networks, with application to gene expression data. *Ai Communications*, *20*(4), pp.263-271.

[43] Agneessens, F., Borgatti, S.P. and Everett, M.G., 2017. Geodesic based centrality: unifying the local and the global. *Social Networks*, *49*, pp.12-26.

[44]Hernández, J.M. and Van Mieghem, P., 2011. Classification of graph metrics. *Delft University of Technology: Mekelweg, The Netherlands*, pp.1-20.

[45] Muchnik, L., Pei, S., Parra, L.C., Reis, S.D., Andrade Jr, J.S., Havlin, S. and Makse, H.A., 2013. Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Scientific reports*, *3*, p.1783.

[46] I. Jolliffe, Principal Component Analysis. Wiley Online Library, 2002.

[47] Tryfos, P., 2013. *The measurement of economic relationships* (Vol. 41). Springer Science & Business Media.

[48] Liebetrau AM., 2008: Measures of Association. Beverly Hills, CA, Sage.

[49] Landy, S.D. and Szalay, A.S., 1993. Bias and variance of angular correlation functions. *The Astrophysical Journal*, *412*, pp.64-71.

[50] https://gdpr-info.eu/