An-Najah National University

Faculty of Graduate Studies

The Impact of Social Network Analysis on the Palestinian Telecom Industry, Jawwal - Palestine

By

Wael Khalil Abu Rezeq

Supervisor

Dr. Ramiz Assaf

This thesis is Submitted as a Partial Fulfillment of the Requirements for the Degree of Master of Engineering Management, Faculty of Graduate Studies An-Najah National University, Nablus-Palestine.

2017

The Impact of Social Network Analysis On the Palestinian Telecom Industry, Jawwal - Palestine

By

Wael Khalil Abu Rezeq

This thesis is defended successfully in 05/02/2017 and approved by

Defense committee Members	**Signature**
– **Dr. Ramiz Assaf/Supervisor**	..……
– **Dr. Khaled Rabayah /External Examiner**	……..
– **Dr. Yahya Saleh / Internal Examiner**	……..….

Acknowledgement

First and foremost, I offer my sincere gratitude to my supervisor, Dr. Ramiz Assaf, who has supported me throughout my thesis with his patience and knowledge. I attribute the level of my Master degree to his encouragement and without him this thesis would not have been completed or written.

Also, I would like to thank the faculty at An-Najah National University in general and department of Engineering Management in specific for the full support and facilities I have needed to produce and complete my thesis.

My greatest appreciation goes to Jawwal Company for the outstanding support and amenities to accomplish this research.

The success of this study required the help of various individuals. Without them, the researcher might not be able to meet his objectives in this study. The researcher wants to give gratitude to the following people for their invaluable help and support, my father and mother, my brother and sisters, and my friends.

I am grateful to Ahmed M. Hashim, a consultant in data mining and a PhD holder from the Faculty of Engineering of Mina University in Egypt, for the support he provided and the great collaboration, which I really enjoyed.

Most importantly, thank Almighty God for granting me the wisdom, health and strength to undertake this research task and enabling me to complete it.

Respectfully,

Wael Abu Rezeq

الإقرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان

The Impact of Social Network Analysis on the Palestinian Telecom Industry, Jawwal – Palestine

أقر بأن ما شملت عليه الرسالة هو نتاج جهدي الخاص, باستثناء ما تمت الإشارة إليه حيثما ورد, وأن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل لنيل أي درجة أو لقب علمي أو بحثي لدى أي مؤسسة علمية أو بحثية

Declaration

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degrees or qualifications.

Student's Name: **اسم الطالب:**

Signature **التوقيع:**

Date **التاريخ:**

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

Abbreviation	***Meaning***
ARPU	Average Revenue Per User
CDR	Call Detail Record
CEO	Chief Executive Officer
CIF	Customer Influence Factor
CRISP	Cross Industry Standard Process for Data Mining
DBMS	Database Management System
DDL	Data Definition Language
DM	Data Mining
DML	Data Manipulation Language
DWH	Data Warehouse
GUI	Graphical User Interface
HITS	Hyperlink-Induced Topic Search
ICT	Information and Communication Technology
ISDN	Integrated Services Digital Network
IVR	Interactive Voice Response
MMS	Multimedia Message Service
ONA	Organizational Network Analysis
PALTEL	Palestine Telecommunications Company
ROI	Return On Investment
SMS	Short Message Service
SN	Social Network
SNA	Social Network Analysis
SQL	Structured Query Language
Subs	Subscribers
VSMS	Voice Short Message Service
XML	Extensible Markup Language
URL	Uniform Resource Locator

The Impact of Social Network Analysis on the Palestinian Telecom Industry, Jawwal - Palestine

By
Wael Khalil Abu Rezeq
Supervisor
Dr. Ramiz Assaf

Abstract

Telecommunication industry has proven itself not only as an emerging economic sector but as a rapidly growing sector with a huge chain of economic and social impact. Jawwal, as the market leader in mobile telecom industry in Palestine, should take advantage of datamining science consecutively as the need to sustain customers the longest possible time. This research is based on community detection analysis through identifying group members and influencers. It is motivated by three major research questions: (1) who are the influencers in Jawwal subscribers for churn and new products adoptions? (2) who are subscribers whose loyalty is threatened by churn around them? and (3) What is the ROI percentage for Jawwal's subscribers based on their impact on the community? (their degrees of influence).

The use of datamining algorithms and datamining standard process in today's challenges was set to increase the efficiency of solving not only technical problems but business as to move toward insights and analysis based exercises. This research proposes a model based on the Similarity, Authority and Hub algorithms used in Social Network Analysis (SNA) to group members who share same patterns together, to extrapolate network leaders in terms of authority and dissemination leaders and to extract group

level insights that can be used in understanding subscribers' patterns and ways of communications (in-degrees and out-degrees) between them using IBM SPSS Modeler as the base of datamining process and Oracle as the base of corporate data warehouse. In addition, Monte-Carlo simulation using Matlab has been applied to prove the hypothesis of mean ROI on selected communities has significantly increased (with α=0.05) when incorporating SNA compared to traditional targeting techniques.

This study estimated how much reliable is to use SNA in order to shift from the traditional ways of targeting subscribers, understanding their pattern, enhancing offering efficiency and improving the process of launching new product by selecting the right target. It has contributed to shifting marketing leaders' way of thinking from individual view to network view, that is analysing the 360 degrees of network in a comprehensive manner instead of singular. It has also contributed to providing decision makers with a systematic and scientific approach as to focus on network influencers, to save as much as they can and to build their offers based on well-studied ROIs.

Chapter One

Introduction

In our daily life, we used to communicate with each other starting from our family at the morning, our boss at work or teacher at an academic institution, our colleagues and even friends. During those communication activities, that is – in the simplest way –the bridge between a person and another (or a person to a community) we figure out many communication aspects and skills. Every group starting from two persons should have at least one influencer and the rest are considered to be followers. A manager in our work is most probably an influencer as s/he influences us in her/his decisions, a teacher is also an influencer as s/he influences us in her/his educational experience that s/he is teaching.

Social life is defined as the relations between people that set up networks. Those people have their own behaviors and these behaviors are shaped into a community that includes social structures. More specifically, within every social structure inside community, particular behaviors can be observed at the level of every individual in which only one of them is considered to be the influencer of community and the one who shapes the overall social structure. (Barry, et al., 1988)

1.1 Definitions of Social Network Analysis (SNA)

Social network analysis -related to the network theory- was emerged as a key technique in modern sociology. It has also obtained a significant

impact in economics, information science, biology, anthropology, communication studies, organizational studies, geography, sociolinguistics, and social psychology and it has become a popular topic of speculation and study. (Freeman, 2007)

Social networks were also used to study how companies interact with each other, characterizing the many informal connections that link CEOs together, as well as connections and associations between individual workers at different companies. (Wasserman & Faust, 1994).

In telecommunications, it is quite straightforward to define calls as links and customers as nodes especially in mobile operations, even though there is more than one type of link in mobile operations, such as calls, text messages, multimedia messages, e-mail or even Internet access. (See Figure 1-1).

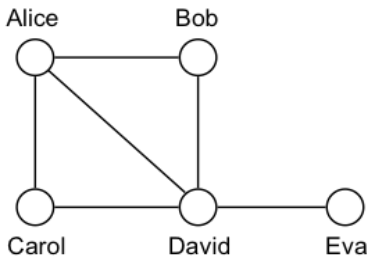


Figure (1.1) Nodes & Relationships

According to Krebs (2013), Social network analysis (SNA) is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. The nodes in the network - the circles- are the people and groups while the links - lines- show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships. Management consultants use this methodology with their business clients and they call it Organizational Network Analysis (ONA). (Krebs, 2013)

According to Evelien Otte and Ronald Rousseau (2002), “Social network analysis (SNA) is a strategy for investigating social structures through the use of network and graph theories.”. Social network analysis uses information about the relationships between people in order to predict their patterns, interests and influencer.

Wasserman and Faust (1994) argued that “social network analysis is based on an assumption of the importance of relationships among interacting units. The social network perspective encompasses theories, models, and applications that are expressed in terms of relational concepts or processes”. (Wasserman & Faust, 1994)

The researcher agrees more with the Kreb’s definition as it’s comprehensive, includes all types of entities that SNA can work with. It also clarifies the nodes and relationships, the presentation of social network analysis as visual or mathematical and it also describes some of the uses by management consultants.

SNA is developed through the calculations of both individual and group characteristics. Individual characteristics are based on self/nodal factors related to attributes and behaviors such as in-degree, out-degree, authority and dissemination scores where group characteristics are more related to group/community level insights such as group size and density. All of them are explained at chapter two, section of ‘How SNA Works?’. (IBM Corporation , 2012)

This research will be applied on Jawwal, the Palestinian Cellular Telecommunication Company and the first and largest operator in

Palestine, as the researcher is part of this respectful family and do has have direct connections with key persons in that firm. Due to Jawwal large base and distribution over Palestinian regions, the researcher believes that the use of Social Network Analysis (SNA) helps Jawwal to define their subscribers' patterns in communicating with each other. Obviously, some have huge communities, i.e a phone shop dealer and others have small ones, i.e: a child aged 10 years.

1.2 Problem Statement

Jawwal is aware of competition and its impact on its revenue stream from day one of Wataniya, the competitor, operational launch. The threat that competitors play is customers' acquisition from other emerging companies through aiming to increase its subscribers' base. The risk is coming through targeting old operator influencers who have high impact of other communities as a tool to win them and their community members.

The nonexistence of SNA model, which is more focused on determining the influencers inside the communities, forced Jawwal to invest in all predicted to churn subscribers regardless if these subscribers Return on Investment (ROI) is high or not and regardless the impact of them on their community.

This has forced Jawwal to think about a solution that helps to identify influencers and to target them aggressively or even engage them into incentive loyalty programs. The same concept is applied on customers' acquisition from other competitors based on their importance and using

marketing techniques such as ‘Member Gets Member’ technique. This technique is being rapidly used by most of telecom operators through targeting specific members – in a community – who has the influence on other members within same community. This technique is also used by most of telecom operators’, i.e: ‘Invite your friends to join Network X and get £50 Amazon voucher’. In addition, this technique is used by Vodafone UK too. (Vodafone, 2013)

1.3 Research Objectives

The aim of this research is to address the importance of SNA and its effect on Jawwal subscribers’ base. In principles, this research aims to answer the following questions:

1. Who are the influencers in Jawwal subscribers’ base? It is going beyond individual value to network value in the areas of:
   A- Who, when they churn, would take few friends with them?
   B- Who, when they adopt, would push a few friends to do the same?
2. Who are subscribers whose loyalty is threatened by churn around them?
3. What is the ROI percentage for Jawwal’s subscribers based on their impact on the community? (their degrees of influence).

1.4 Research Limitations

The main limitations for this research can be summarized by the following points:

- SNA tool: The availability of trial tool, i.e: student version, to implement SNA. The cost of a fully professional tool to implement this model ranges between $61,280 for SAS Enterprise Miner and $11,300 for IBM SPSS Modeler (per user per year). (Executive Information Systems, 2016). SAS and IBM are ranked as top two predictive and social network analysis tools in 2014 (TopPredictiveSoftware, 2014). Of course, there are open sourced tools available in the market. However, most of these tools have no support, have no Graphical User Interface (GUI) and all of them require some programming language experiences to accomplish the task such as R-Language.
- Big data processing: SNA tool deals with huge amount of data on subscriber level, these processing tasks require desktop workstation with high processing power.
- Access: as proposed research may need confidential data, the access is denied or otherwise limited; it will be complicated to gain data from Jawwal Call Detailed Records (CDRs) database in addition to confidentiality of results.

1.5 Research Impact

Within telecommunications, the phone calls and messages among the subscribers provide insight into the social connections among them. Social Network Analysis (SNA) provides a way to leverage this information to improve marketing effectiveness which will not only enhance revenues but

also help deliver a better customer experience. SNA model will identify and quantify INFLUENCE in a calling circle. Once the influencers were identified, they can be targeted them with disproportionate benefits. After all, it is better to target one and allow their influence to naturally do its job rather targeting everyone. (Sonamine, 2010)

This study will help Jawwal to reduce its marketing campaigns non favorable costs and will let it focus more on high ROI subs than low ROI (in terms of targeting them). Moreover, this is expected to help Jawwal expand its base through acquiring new subscribers from competitor using Member gets Member campaigns and prevent key players churn within its customers' base who will affect other community members.

1.6 Thesis Structure

Chapter 1	Introduction: Definitions of Social Network Analysis (SNA), Problem Statement, Research Objectives, Research Limitations and Research Impact on the business.
Chapter 2	Theoretical Background: Business Intelligence, Data Mining, Conceptual Framework, Knowledge Gap, Theoritical Framework and State of the Art of SNA.
Chapter 3	Methodology: CRISP-DM framework, Algorithms of HITS, Modularity and Similarity, Groups and Individual Analysis, Interpretations of Results and IBM SPSS Modeler.
Chapter 4	Research Results: Applying SNA on Jawwal Sample, and ROI Analysis using Monte-Carlo Simulation.
Chapter 5	Discussion of Results: Answering research questions of defining the influencers, followers and calculating ROI
Chapter 6	Conclusions and Recommendations: Managerial Insights and suggestions for future research.

Chapter Tow

Literature Review

In Chapter One, some definitions of Social Network Analysis, a brief about the uses of this technique and the selected company that SNA will be applied on were introduced. This chapter includes conceptual framework, knowledge gap and theoretical framework for the research. The review presents and discusses issues in Business Intelligence, Data Mining, and Social network analysis that were written in a way that simplifies the concepts of social network analysis for non-technical background readers.

2.1 Business Intelligence

In highly competitive and challenging industries of telecommunications, retails, banking sectors and much more, companies are finding it increasingly difficult to maximize their profit. They either have to decrease their cost or to increase their revenues that will result in an increase of net profit. One of the most promising aspects of investments is technology and data investments. Through applying the concepts of Business intelligence (BI), companies could gain a competitive advantage of collecting the data collected from internal and external sources, analyzing it, creating online and automated results driven dashboards that visualize the data to make the analytical results available to corporate decision makers as well as operational workers for more informed business decisions. That is, it is an umbrella term that refers to a variety of software applications used to analyze an organization’s raw data. BI as a discipline is made up of several

related activities, including data mining, online analytical processing, querying and reporting. (Rouse, 2014)

2.2 Data Mining

Data mining is “a nontrivial process from a large amount of Data to obtain valid, novel, potentially useful and ultimately understandable patterns” (Huang, et al., 2013). From the generalized point of data mining, it is stored the "dig" interesting knowledge process a large amount of data in a database, or data warehouse, or other information.

The journey of data mining models starts from a very important word called ‘Data’. When it is available, understandable, reliable and reasonable it will lead to very rich insights and will guide decision makers to the right path. Figure (2-1) shows the lifecycle of data starting from fetching it and ending at transforming the knowledge to actions. These actions could be new strategies, correcting strategies or even guidelines for the current and existing strategies.

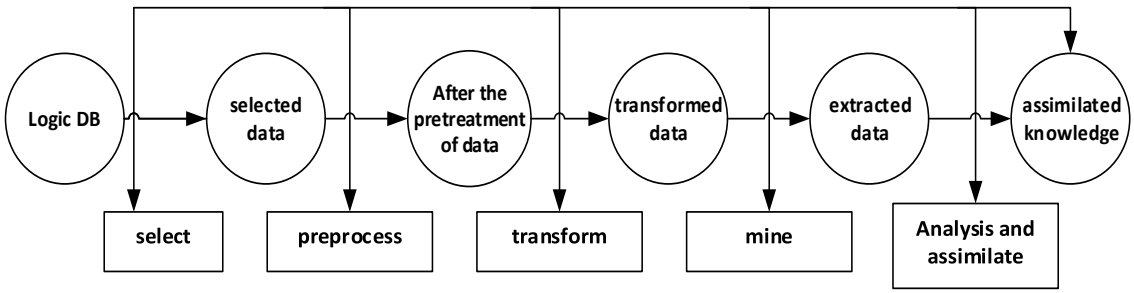


Figure (2.1): The basic processes and the main steps of data mining (Feng, 2007).

2.3 Conceptual Framework

As we are living in a complex social community, we influence our friends and associates and they influence us. Music, politics, opinions and ideas all circulate and evolve in these networks of influence. People are also influenced by each other's commercial decisions, the advent of social media has proven that word-of-mouth is a very powerful thing. (Whitler, 2014)

Each individual belongs to a community that shares with him some common interests or relationships. From business to customer side, companies have recently started to focus on the relationships with their customers in order to strengthen those relations and maintain the customers the longest period they can. This research will offer a similar approach of understanding relationships not only between business to customer, but within customers' communities themselves. In another meaning, it will focus on giving insight into the various roles and groupings in all communities within companies' network that will result in identifying influencers and coming up with strategies to sustain or even retain them. The idea of "social network" was used loosely for over a century to point out complex sets of relationships between members of social systems at all levels, from interpersonal to network.

Social networks were used to study how companies interact with each other, characterizing the many informal connections that link CEOs together, as well as connections and associations between individual workers at different companies. For example, the power within

organizations often comes more from the degree to which an individual within a network is at the middle of many relationships than actual job title. Social networks also play a key role in staffing, in business success, and in job performance. Networks provide ways for organizations to collect information, being able to handle over the competition, and collude in setting policies or even prices. (Wasserman & Faust, 1994)

Social Network Analysis (SNA) uses the interaction patterns of telecom operator, banking sector, government institutions and much more customers' network to identify groups of similar individuals. Characteristics of these groups influence the behavior of the individual group members. For example, small groups having any inter-member relationships and strong leaders have an increased risk of churn – switching from a current network to another, even if no member of the group has actually churned. (Ingen, 2013)

2.4 Knowledge Gap

Social Network Analysis (SNA) brings the power into the business domain, identifies trends and calling networks that companies can use to reduce churn -customers moving from an operator/company to another- or to propagate their own desired behaviors e.g. product adoption. (Passmore, 2011)

How do companies define an individual as a leader of follower, what makes him an influencer? What is the impact when companies lose an influential customer within a social network? Are Influencers able to lead other customers to purchase new products and bundles or to consume new

services? Within the Palestinian context, SNA is considered to be a chance for the researcher to introduce it due to the fact that it hasn't been introduced yet.

2.4.1 SNA in Telecommunication Industry

In telecommunication industry, phone calls and messages among the subscribers provide insight into the social connections between them. Companies need to understand the new perspective of social relations through extracting, understanding and classifying those relations. It will help them to focus the customer relationship management in a more accurate way.

Every generated call in any telecom operator refers to Call Detailed Record (CDR) in which it includes details about the caller (A-Number), the receiver (B-Number), the duration of the call (normally in seconds), the amount of money (or minutes in case of free minutes) deducted in the call and some other relative attributes - is a goldmine of complex social behavior.

The state and shape of a social network help determine a network's usefulness to its individuals. That is, more tightly systems can be less useful to their individuals than networks with heaps of free associations (weak ties) to people outside the networks. More open networks, with many social connections and weak ties, will probably help introduce new ideas and opportunities to their members than tightly (closed) networks with many excessive ties. In other words, a group of friends who only do

things with each other already share the same knowledge and opportunities. A group of individuals with connections to other social worlds is likely to have access to a more extensive scope of information. That is, it is better for individual success to have connections to a variety of networks rather than many connections within a single network. Similarly, individuals can practice influence or act as brokers within their social networks by bridging two networks that are not straightforwardly connected which referred as filling structural holes. (Scott, 1991).

Theoretically, SNA works on the Call Detailed Records (CDRs) data generated from subscribers behaviors (Passmore, 2011). Due to the large CDRs generated every day from a base of more than 2.8 Million, which might lead to more than 60 Million records per day. It's preferable to group the data on subscriber level instead of CDR level (i.e: A-Number might have been calling a B-Number 5 times a day with a total of 10 minutes, instead of having 5 records for this A-Number we will have only 1 aggregated record). Another important variable is the type of B-Number; in another meaning, it's not necessary for the A-Number to call only subscribers from their same network operator, this is called OnNet call type, as A-number might be calling OffNet – subscribers from other operators, like Wataniya, or INTL - International-destinations. Moreover, CDR doesn't only involve calls, it might also be:

- SMS: Short Message Service.
- MMS: Multimedia Message Service.
- VSMS: Voice Short Message Service.

- Data: Mobile data knowing that this is not involving B-number.

After that, the main goal of this research is to calculate the influence of each node by calculating its in-degree, out-degree, degree, authority score, hub (dissemination) score, network density and size. Below are the definitions of each where how will they be calculated and used are discussed in methodology section:

- **In-degree:** a node is considered to be the target involving how many connections were made to it. (IBM Corporation , 2012)
- **Out-degree:** is the number of connections made where the node is the source instead of being the target. (IBM Corporation , 2012)
- **Degree:** How many people can this person reach directly? The count of the number of ties to other actors in the network. (Passmore, 2011)
- **Authority Score:** Authority scores measures the importance of an individual corresponding the number of relationships ending at him or her indicating how much an individual receives from other members of a group. (IBM Corporation , 2012)
- **Dissemination Score:** In contrast to authority, dissemination score measures the importance of an individual with the number of relationships originating from him or her indicating how much an individual connects to other members of a group. (IBM Corporation , 2012)

- **Similarity Measure:** Similarity is measured through examining the neighbors of each member and the strength of relation between members and their neighbors.
- **Network Density:** the proportion of actual connections (relationships) between nodes in a network (group) divided by total possible connections in that group (IBM Corporation , 2012)
- **Network Size:** Total number of nodes in a single network after grouping all similar members and strength ties together.

These measures describe how influential and central the nodes are and how important they are in the social network for problem solving. (Penheiro, 2011).

The relation between each of these measures compared to each subscriber is determined using a set of algorithms shaped into macros. Macro is a programming language statement that when processed generates a sequence of more detailed language statements (Rouse, 2005). Table (2-1) indicates some of these algorithms and the description of each where more details are mentioned in the chapter of Methodology.

Table (2.1): Algorithms in social network analysis

Name	Description
Coverage Threshold	It's responsible on omitting weak relationships between members before splitting them into network. (IBM Corporation, 2012)
Modularity	It's defined as a value between -1 and 1 that measures the density of links inside communities compared to links between communities. (Maryland, 2009)
HITS	The HITS algorithm is a link analysis algorithm used to rate Webpages (nodes). It's an iterative algorithm developed to quantify each page's value as an authority and as a hub. (Devi, et al., 2014)

2.4.2 Social Network Analysis and Data Analysis

Wellman & Marin (2009) argued that SNA is a very powerful tool due to the fact that is different from traditional social scientific studies. This is true because the latter focuses on the traits of every single individual while SNA examines the relationship and ties among the different members within specific social networks.

The study of social networks is formally defined as a set of nodes which consist of network members where those nodes are connected by different types of relations defined as links. Network analysis study takes all those connections as the primary building blocks of the social world. It does not collect just unique types of data, but traditional analysis takes into consideration individual attributes. Social network analysis also considers the attributes of relations from a fundamentally different perspective than that adopted by individualist or attribute - based social science. (Wellman & Marin, Jun 2009).

Traditional methods of data analysis usually consider individual attributes from all observations in order to analyze the available information. What is the average characteristic from a particular population of companies, employees, customers, or markets? Particularly in telecommunications, it is quite common to analyze the individual data to understand customer behavior, such as the average billing - total customers' invoices-, type of payment, frequency and amount of service usage, and so on.

Besides the individual attributes, social network analysis considers all information about the relationships among the network members (nodes).

As a matter of fact, the information about the relations among the individuals within a social network is usually more relevant than the individual attributes of the individuals. The relations among the individuals can tell more about customers than their individual attributes. This is the basic difference between data analysis and social network analysis. What "you are" is not as important as "how you behave and connect with others".

2.4.3 Network Types

Social network analysis was moved from being a suggestive representation to an expository approach and then to a worldview. With its own particular techniques, hypothetical articulations, scientists and invented social network software, analysts reason from aggregate to part, from substance to connection to individual and from behavior to attitude. They typically either ponder personal networks (also known as egocentric systems) or entire systems (also called complete systems) as presented in Figure (2-2). (Barry, et al., 1988)

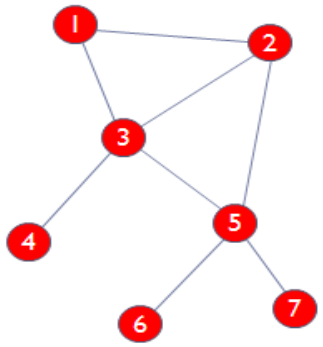


Figure (2.2) Whole Network

The distinction between whole networks and individual/egocentric networks had depended to a great extent on how analysts were able to collect and gather the data. That is, the analyst was expected to have complete information about who was in the network for groups such as companies, schools, or membership societies. All participants being both potential egos and alters as in Figure (2-3).

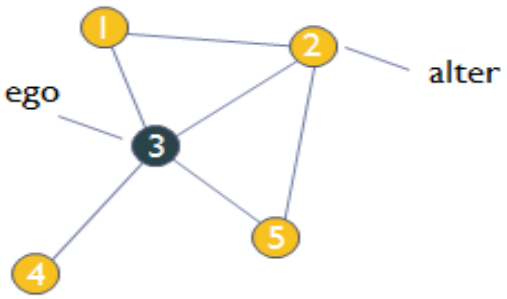


Figure (2.3): '3' is ego, rest are their

Ego is a person who influences others or is considered to be a central point for them and his existence is necessary to connect the whole network together where alters are the rest of people around. Egocentric / individual studies were typically conducted when identities of egos were recognized but not their alters. These studies depend on the egos to supply information about the identities of alters and there is no likelihood that the various egos or groups of alters will be tied to each other as in the next Figure (2-4). (Cheliotis, 2010)

The hybrid network may be useful for studying whole networks that are expected to include important influencers beyond those who are formally identified. For example, employees of a company often work with

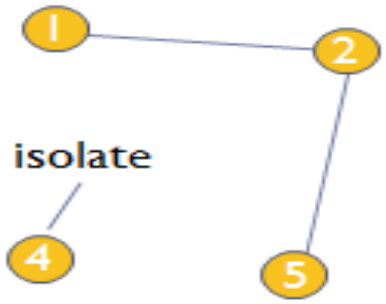


Figure (2.4): A network without ego resulted in connections cut

consultants whom are out of company (non-company consultants) and they might be part of a network that cannot fully be defined prior to data collection.

Freeman argued that “there are several analytic tendencies that distinguish social network analysis: 1) There is no assumption that groups are the building blocks of society: the approach is open to studying less-bounded social systems, from nonlocal communities to links among websites. 2) Rather than treating individuals (persons, organizations, states) as discrete units of analysis, it focuses on how the structure of ties affects individuals and their relationships. 3) In contrast to analyses that assume that socialization into norms determines behavior, network analysis looks to see the extent to which the structure and composition of ties affect norms.” (Freeman, 2007)

2.4.4 Network Analytic Tools

Network analytic tools are used to identify, represent, analyze, visualize, or simulate nodes (e.g. agents, organizations, or knowledge) and edges (relationships or connections) from various types of input data (relational and non-relational) and including mathematical models of social networks. Like other software tools, the data can be saved in external files where there are various data input formats used by network analysis software packages such as Text-Based, XML-Based, Object-Oriented, and some packages support multiple files format and the output data can be saved in external files. (DataFormats, 2006)

Network analytic tools allow researchers to investigate representations of networks of different size - from small (e.g. families, project teams) to very large (e.g. the Internet, disease transmission). Visual representations of social networks are important to understand network data and convey the result of the analysis. Visualization is often used as an additional or standalone data analysis method. With respect to visualization, network analysis tools are used to change the layout, colors, size and other properties of the network representation.

On the other hand, predictive analytic tools include both network analysis and individual level analysis and prediction so that tools provide mathematical and statistical routines which can be applied to the network model. In this research. Predictive network analytic software is used. Some examples of these predictive analytics tools are: SAS Predictive Analytics, IBM Predictive Analytics, Rapid Miner, Angoss Predictive Analytics, Oracle Data Mining (ODM), TIBCO Analytics and much more. (TopPredictiveSoftware, 2014)

In this work, IBM SPSS Modeler will be used to implement SNA on Jawwal. In the next paragraphs, researcher discusses three of predictive analytics software along with some examples of using them to implement social network analysis.

Firstly, SAS (pronounced "sass") stands for "statistical analysis system." It began at North Carolina State University as a project to analyze agricultural research. Demand for such software capabilities began to grow, and SAS was founded in 1976 to help customers in all sorts of industries – from

pharmaceutical companies and banks to academic and governmental entities.

SAS – both the software and the company – thrived throughout the next few decades. Development of the software attained new heights in the industry because it could run across all platforms, using the multivendor architecture for which it is known today. While the scope of the company has spread across the globe, the encouraging and innovative corporate culture has remained the same. (SAS, 2015).

A Product Manager in SAS Institute Inc, Ken King, in his book Social Network Analysis in Telecommunications which he wrote during a postdoctoral at Dublin City University, implemented SNA model using SAS. He provided a step by step details on how to implement this model; his case study focused on a telecommunication company in order to understand the customers' relationship, and hence, to identify influential customers from the perspective of distinct businesses.

Secondly, IBM SPSS Modeler is one of IBM Analytics products; these predictive analytics portfolio from IBM includes IBM SPSS Modeler, IBM SPSS Data Collection, IBM SPSS Analytic Server, IBM SPSS Analytic Answers, IBM SPSS Statistics, IBM Social Media Analytics and IBM Analytical Decision Management. IBM SPSS Modeler is an extensive predictive analytics platform that is designed to bring predictive intelligence to decisions made by individuals, groups, systems and the enterprise. By providing a range of advanced algorithms and techniques that include text analytics, entity analytics, decision management and

optimization, SPSS Modeler can help you consistently make the right decisions—from the desktop or within operational systems. (SPSS, 2013)

IBM SPSS Modeler was used by Erik Van Ingen, an assistant professor in Tilburg university department of sociology. Erik implemented SNA to answer one main question about the effect of social network on smoking at a secondary school "Do social network affect the frequency of smoking among pupils in secondary school?" (Ingen, 2013) .

Thirdly, Sonamine is a prediction and analytics company founded in 2009. They focused on helping marketers optimize revenues by predicting their customers' behavior. For example, they assumed that they helped game developers predict which players will churn or buy more virtual goods, they had three telecommunication case studies related to predictions using SNA model starting from the effect of SNA in churn, acquisition using member gets member technique – which researcher previously discussed – and age estimation based on subscriber's community.

Researcher has chosen IBM Modeler as the base environment to work with because:

1. The cost of IBM modeler is relatively low compared to SAS (around $11,300 compared to $67,000).
2. IBM Modeler supports group analysis (in addition to individual analysis) where SAS supports the latest only.
3. The modeler is a more user friendly and requires no hard coding skills as in SAS.

Finally, a free version of IBM modeler can be used to demonstrate the results of analysis done. This version is one month free. Other free software could be used also but they lack GUIs and require a lot of coding.

2.4.5 Previous Studies

Social network analysis -related to the network theory- was emerged as a key technique in modern sociology. It has also obtained a significant impact in economics, information science, biology, anthropology, communication studies, organizational studies, geography, sociolinguistics, and social psychology. SNA has

Social Network Analysis had been also utilized as a part of the study of disease transmission to help see how examples of human contact help or hinder the spread of infections, for example, HIV in a populace. (Passmore, 2011) The development of social networks can sometimes be demonstrated by the utilization of agent based models, giving understanding into the exchange between communication principles, gossips spreading and social structure. SNA was also used in mass surveillance- for example to determine whether or not US citizen were political threats, Total Information Awareness program was doing in-depth research on strategies to analyze social networks.

Robin Dunbar argued that because of the limits in the capacity of the human communication, the maximum number of members an egocentric network can reach is 150 members. In fact, this is proven from multiple sociological and anthropological studies. In theory, this number is

considered as the average ability to track members' different emotions and be able to determine needs of all the members. (Passmore, 2011).

Mark Granovetter pointed out in one study that more numerous weak ties play an important role in seeking information and innovation. In contrast, Clique intended to have more homogenous opinions as well as sharing many common traits. In his opinion, this hemophilic tendency was the reason for the members to be attached together firstly. However, and due to similarity, each member of the Clique group knows more or less what the other members did. As a conclusion, to find new insights or information, members of the Clique do have to look beyond the group to other friends and acquaintances. This is what Granovetter called "the strength of weak ties". (Passmore, 2011)

Guanxi is a central concept in Chinese society (and other East Asian cultures) that can be summarized as the use of personal influence on others. Guanxi can be studied from a social network approach. (Wellman, et al., 2002)

A limited number of related prior studies have proposed approaches to use social network information in order to predict customer churn. "Dasgupta et al., 2008 were the first of its kind in predicting customer churn using social ties between the subscribers of a telecom operator. They analyzed the structure and evolution of a massive telecommunication or call graph for a single mobile operator for four different regions in India with different socio-demographic, urbanization, and cultural characteristics, and with the number of nodes for the regions ranging up to 1.25 million". The

study focused on the prepaid segment of customers, for which CDRs data are the only available source of information. They indicated the possibility to apply relational classifiers to predict customer churn using CDR data, he confirmed the existence and relevance of social network information for customer churn prediction. (Dasgupta, et al., 2008)

Richter et al., 2010 presented the group-first churn prediction approach to predict customer churn based on the analysis of social groups or communities derived from CDR data. The presented approach assigns a churn score to each subscriber based on the churn score of the social group as well as personal characteristics. The results of the study reconfirm the potential of improving the current generation of customer churn prediction models by adding information that captures the social interactions of a subscriber, and indicates that group structure and membership are determinants of churn behavior. This study opens interesting alternative modeling approaches to exploit the information contained within the network structure of the customers of a telco operator. (Richter, et al., 2010)

Sonamine, an analytics and predictions company founded in 2009, was able to predict 25% of real churners when incorporating the SNA with Churn model. The exercise was done in a European mobile operator; their prepaid churn model was not accurate due to lack of customer demographic information. SNA Accurately predicted 25% of churners using only 5% of population lift of more than 5X of accuracy compared to previous results. (Sonamine, 2010)

On the other hand, SNA proved its ability to increase the conversion rate of cross-selling campaigns. According to (Harding 2002), "Cross-selling is the action or practice of selling an additional product or service to an existing customer" Sonamine case study was applied on a US Telecommunication Company to implement Cross selling promotion. The US telecom company aimed to increase a cross sells of a new product, Sonamine indicated that the target list generated by SNA had 340% higher conversion rate rather than other random selection group (called Control Group). Conversion rate is the percentage of the target list that purchased the product. (Sonamine, 2010)

Another case study, also by Sonamine , focused on using subscriber age for product planning and development. The project was performed at an Asian mobile operator called Age estimation by leveraging the concept that similar people form social groups. Sonamine declared that for the test group of 450,000 subscribers, they accurately predicted age for 73% with a (+/-) 1-year error margin. (Sonamine, 2010)

2.5 Theoretical Framework

This section includes the main hypothesis that will be proven in this research starting from elaborating data mining concept to identify influencers and define them, define followers and to prove that incorporating SNA will have a significant effect in Return on Investment (ROI) compared to without. It will also include some previous hypothesis related to SNA.

One of the main outcomes for this research is that researcher, using IBM SPSS Modeler, will cluster Jawwal base of subscribers into communities, extract the influencers among those communities and flag them along with other normal community members. First assumption is that community size ranges from two to forty two persons, and secondly it will have two types of influencers called 'Authority' and 'Disseminator' influencers. Thirdly, the mean of return on investment when incorporating SNA will significantly differ than when SNA is not incorporated. ROI will include the factors of old revenues per network member, new expected revenues, cost of offers offered to those network members. Researcher assumption will take into consideration the fact of opt in (engagement) rate of offers and will be focused on the fact that network/group influencers will have direct influence on other group members that will force them to opt in to the same offer they have been offered, that is instead of creating more than one offer for the whole group (depending on the count of group members) the hypothesis for incorporating the SNA will assume that there will be one offer for the leader that will be shared with other network members.

Erik, using IBM SPSS Modeler, implemented a case of social network effect in smoking among secondary students at a school where the hypothesis he assumed was: The frequency with which a pupil smokes is positively correlated to the frequency of smoking among his/her friends. He started with computing average smoking among friends, he identified the types of persons as two: 1- Ego: A person who is connecting to other and 2. Alter is a person that who is being connected by ego. He collected a data

of 2,914 students from a school in Netherlands where every student had up to 12 friends. These 12 friends are not necessarily to be smokers. Then, Erik computed the correlation between smoking ego and smoking network and found that there is a strong correlation between them.

One study found that happiness tends to be correlated in social networks. When someone is happy, near friends have a 25% higher chance of being also happy themselves. In addition, people at the middle of a social network tend to become happier in the future than those at the borders. Groups of happy and unhappy people were recognized within the studied networks, with a reach of three degrees of separation: a person's happiness was associated with the level of happiness of their friends' friends' friends. (Fowler & Christakis, 2008)

2.6 State-of-the-Art of SNA

Linton Freeman wrote in the development of Social Network Analysis, (Freeman, 2007) about the progress of social networks and social network analysis. The summary of his points are as below:

In the late 1800s, precursors include Emile Durkheim and Ferdinand Tonnies first argued about SNA. According to Tonnies, “he argued that social groups can exist as personal and direct social ties that either link individuals who share values and belief or impersonal, formal and instrumental social links”. A non-individualistic explanation of social facts was given by Durkheim. He distinguished between traditional society, where individuals’ differences are minimized, and modern society, that

develops cooperation between non-similar individuals following independent roles.

In 1908/1971, George Simmel was the first scholar to weigh directly in Social Network terms, his essays focused on the nature of network size on interaction and to the probability of this interaction in individuals rather than groups.

“After a hiatus in the first decade of the twentieth century”, (Freeman, 2007) argued, three main traditions in SN appeared. J.l Moreno, in 1930s, piloted the systematic analysis and recording of social interaction in small groups especially classrooms and sociometry (work groups). Another group, Harvard group, was led by W. Lloyed Warner and Elton Mayo explored interpersonal relations at work.

In 1940, “A.R. Radcliffe-Brown's presidential address to British anthropologists urged the systematic study of networks”

In 1950s- 1960s, urbanization studies in the University of Manchester group of anthropologists investigated community networks in South Africa, UK and India. On parallel, the British anthropologist S.F. Nadel codified a theory of social structure that was influential in later network analysis.

In 1954, the term SNA was used by J. A. Barnes to describe what is commonly known by the public and social scientists as bounded groups. (Freeman, 2007). In the 1960s – 1970s, a growing number of scholars worked to combine different tracks and traditions: Harvard University group represented by Harrison White, Charles Tilly – who focused on networks in political and community sociology and social movements -,

Stanley Milgram – who developed the “six degrees of separation” thesis. (Hogan, 2009) .

2.7 Chapter Summary

Telecommunication industry has proved itself to be one of the top growing and innovated industries starting from cellular calls, data bundles and the internet of things applications like Whats app and Viber. As industry grows, the data also becomes richer and the value behind it switched to be one of the key successes in any commercial and noncommercial organizations. The need of analyzing this data and understanding it inspired data mining skilled employees to start deploying business models to satisfy growing need.

SNA has proven its ability to study the pattern of group members and the effect of its influencers among other members starting from the relations between companies' CEOs together, job performance, disease transmission, the smoking hypothesis and Sonamine business cases. A similar case for this research was Sonamine in telecom sectors, however it didn't specify the network details as much as it focused on only incorporating SNA results with other business data mining models namely churn and up-selling. This research focuses on community characteristics and how to analyze, utilize and conclude them.

Chapter Three

Methodology

3.1 Overview

The first three chapters discussed SNA introduction, theories and literature review. Starting by defining the SNA and its impact on the business in the section of introduction, then by discussing telecom theories, how SNA works and the needed tools to implement in theoretical background section. Literature review was more about focusing on the previous studies related to the topic.

Jawwal has an experienced business intelligence and data mining team. Researcher holds a position of Data Mining and Business Intelligence Section Head with more than five years of experience in this field. The research methodology will be conducted in six steps; the steps of data mining modelling. They are grouped into 'CRISP-DM' (CRoss Industry Standard Process for Data Mining) which means the standard process of any data mining model. CRISP-DM was conceived in late 1996. In 1997 it got underway as a European Union project under the ESPRIT (European Strategic Program on Research in Information Technology) funding initiative. The project was led by five companies: SPSS, Teradata, Daimler AG, NCR Corporation and OHRA, an insurance company. (SPSS Inc., 2000) .

CRISP-DM is another aspect of data mining process mentioned earlier (Figure 2-1) but is more dedicated to business users instead of pure

technical users. CRISP-DM is vendor-independent so it can be used with any data mining tool and it can be applied to solve any DM problem. CRISP-DM defines the phases to be carried out in a DM project. CRISP-DM also defines for each phase the tasks and the deliverables for each task.

In the methodology chapter, the steps of data understanding and extraction, data preparation and data modelling will be explained where the "Evaluation and Deployment" will be separated in other two sections, they are more about results and actions rather than statistics and data mining.

The step of data understanding and extraction comes first; it deals with the data on a call detailed record (CDR) level that will be extracted from the corporate data warehouse. Second step is the data preparation which focuses more on aggregating the data on customer level and validating it to ensure a clean and correct data before transferring it to SNA tool. Third, the modelling phase will focus on applying SNA algorithms to derive decision making variables that will be used by business users.

All data will be extracted from the corporate data warehouse (DW/DWH) and for the most recent three cycles. Customer influence factor (CIF) is calculated on monthly basis, that is all subscribers base will have an influence factor update on each month in order to establish a good sense of customer behavior. All measures related to CIF are established using the mean of the last three months. Mean is used to discard outliers, to reduce the impact of special peaks and to control the average curve of customer behavior.

To successfully define a social network analysis in any community, there will be two main algorithms inside according to IBM Knowledge Center related to the selected tool of IBM SPSS Modeler as mentioned in the section of 'Network Analytics Tools'. The first is the one that will split the groups and assign its members (partitioning into groups) which is a group level exercise that is based on similarity measures (Newman, 2004). Second algorithm is more responsible for studying the pattern inside each divided group of members in terms of members' relations using Hyperlink-Induced Topic Search (HITS) algorithm (Stanford, 2008)

3.2 Data Understanding and Extracting Phase

The first step in social network analysis modelling is to understand and extract the required data. A part of data understanding is to define customers' profiles data and behavioral data. Profile data are customers gender, location, age and profession. Behavioral data stands for a kind of variables that are being updated frequently such as customer's activity on the network, activity duration, time and destination which are called Customer Detail Records (CDRs).

Data will be extracted on CDRs level after applying the filters of tenure which is defined as the period in months a subscriber has been active on the network, CDR type as to exclude machines CDRs which is defined as the calls generated by the subscriber to Interactive Voice Response (IVR) or calls toward call center agents.

Due to the huge amount of data contained in the call-detail records, it's necessary to extract the data either on monthly basis or to calculate the means in the corporate data warehouse which reduces the amount of data. Moreover, and due to the large base that Jawwal has across the region, a specific region will be taken as a prototype for implementing SNA model.

To summarize, data will be extracted for a specific region with subscribers count around 25,000. It's important to notice that region should be selected carefully as it shouldn't include direct connections to its neighbors. For example, middle region normally connects south and west together, the probability of a relationship between a person in south and north is less than a person of south and middle, the same for north. Each region has its own characteristics in terms of community size or usage pattern. Researcher will select one of the small and closed cities in north to implement SNA model on them that includes a set of 25,000 customers. Again, this is a prototype on how to use and implement the model. Researcher can't generalize the results on all other regions unless the filter used to include north region is disabled and the model will be executed again, which will consume a lot of processing time and requires super machines to handle the huge amount of CDRs. The 25,000 representative subs resulted in around 0.353 Million records, this is a real case of how huge are the data in telecoms (although this dataset is an aggregated CDRs). Data types (CDRs and Profile) are listed below:

CDRs Data

1. A number (Originator)

2. B number (Terminator)
3. Call type (Peer to Peer, IVR, Free numbers, emergency numbers)
4. Call duration
5. Count of Calls

Profile Data (Information about customers' attributes)

1. Mobile No.
2. Tenure.
3. Connection type: Prepaid or Postpaid.
4. Region.

Social network analysis is a concept that is closely related to customer influence; thus, the analytic data model was developed according to residential customers only. These customers have the ability to exert influence over others.

Structured Query Language (SQL) presents the code required to extract information about customers. The structure of telecommunications data is usually established through different layers of information, in which one customer can have more than one line, product, or service. It is necessary to combine different tables in order to collect the right information about customers and their current services. Appendix 1 clarifies the statements executed to extract CDRs and profile for the selected region and listed as 'A' and 'B' respectively.

It is important to notice that the corporate data warehouse in Jawwal is Oracle based. IBM Modeler is well integrated with this technology that allows a direct extraction from the database to IBM environment after

buying a connection license –to allow IBM applications to connect with Oracle databases easily -. Instead, users can export the needed data from oracle database and import it again to IBM environment.

3.3 Data Preparation

A data preparation process is required to prepare all data extracted from the corporate data warehouse (DW/DWH), and to ensure that it is used properly in the social network analysis. The data preparation process consists of gathering all information collected from the data warehouse, observing them from a unique perspective and aggregating them on subscriber level. At this point, information about customers, accounts, fixed lines, CDRs, products, and billing history is collected from the data warehouse. It is necessary to organize the data in a single dataset. Also, data processing takes a place in this phase where data checkups and validation occur. Some examples are replacing the null values in the data set and how to deal with them, eliminating the outliers using, ensuring that data was successfully uploaded – or integrated- from the corporate DWH to the SNA tool. The following three steps are processed during the phase of data preparation:

3.3.1 Removing Non Eligible Subscribers

So far, two datasets were generated in the previous step of data extraction. Profile dataset was labelled as ‘sna_sample_25000_profile’ where behavioral/CDRs dataset was labelled as ‘sna_sample_25000_aggr_cdrs’. Both statements resulted in a set of subscribers with no validation on date

of joining the network, subscriber's status, calls to IVRs and call centers. The first dataset resulted in a group of subscribers from the DWH based on any transactions in the specified period. And as this is about only transactions, this means that there are no details regarding their current status (either left the network in the second or third month – but they had at least one transactions in the first month – or they are new activated – joining- subscribers who joint the network in the second or third month). In both cases, this means that their call detailed records are missing at least one month out of the three months. In addition, the rest of subscribers might have had CDRs to machines (IVR, call center and other free numbers like emergency calls). The below statement filters the data from the first dataset to ensure that the calls are only for peer to peer transactions.

```
DELETE FROM sna_sample_25000_aggr_cdrs a
WHERE  a.call_type IN ( 9, 10, 21 );
COMMIT;
```

The call types of (9, 10 and 21) are classified as IVR, call center and free numbers in the DWH. The commit command committed the transaction of deleting the data from the specified table (Oracle, 2016)

The other statement of removing churners subscribers or new activated subscribers is indicated below, again the reason of deleting these subscribers is that they don't have at least a three months of data. Three months of data is a benchmark used by most of telecom operators and is currently applied in Jawwal to build a pattern of their customers' base and it will be executed on the second dataset.

```
DELETE FROM sna_sample_25000_profile a
WHERE  a.status != 'Active'
      OR a.activation_date >= '01-Jan-2016'
```

3.3.2 Eliminating the Outliers

In order to guarantee having normalized dataset and as a part of data preparation phase to exclude the extremes (outliers) that might affect the model output, the data preparation process includes some steps in respect to the related content. The three attributes (Total Durations, Calls count and average call duration) is analyzed in terms of distribution, and the outliers are removed from the dataset of CDRs using the percentiles, min and max functions.

The process itself is about doing the averages and percentiles related to each attribute. For instance, the average community size in Jawwal's network is 14 persons, this indicates that each subscriber has a unique numbers of connections, to themselves, equals to 14 and in which there are some subscribers below or above this number. Percentiles provide a more distribution of customers' pattern divided into the percentages of 1-99% as shown in Figure (4-2) indicates.

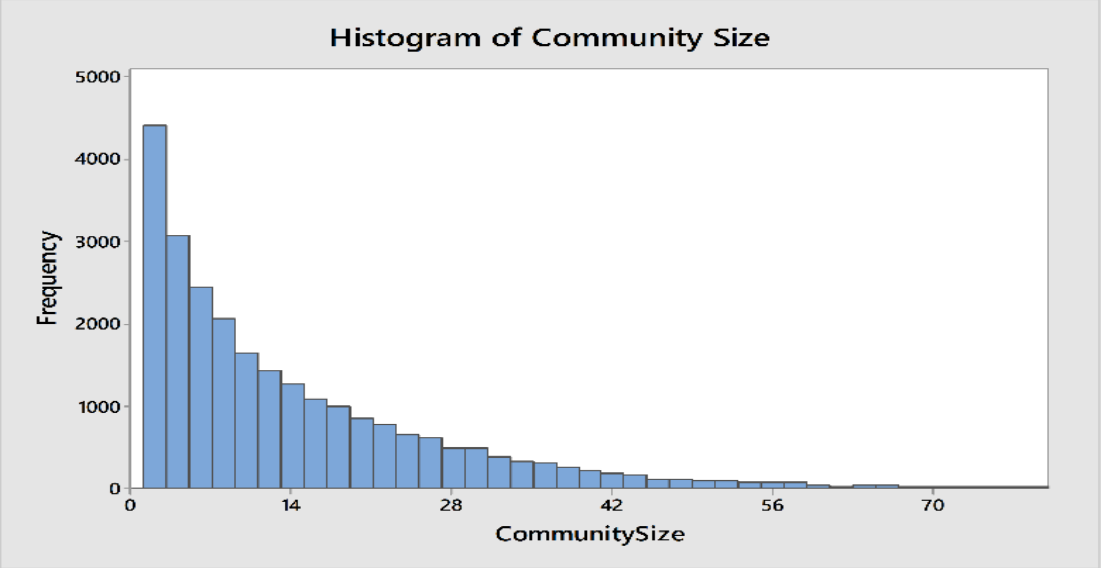


Figure (3.1): Community Size Per Percentile

Table (3-1) below shows that 25% of Jawwal's subscribers has an average unique numbers of connections up to 4 subscribers, this means that the rest of 75% do have unique connections above this number. When doing this exercise, researcher has found that top 1% of percentiles are subs have unique connections of 67 where the rest of 99% do have a number of 42 and less. On the other hand, we can conclude that either the remaining 1% of subscribers are either fraud subscribers since they have a very wide range of connections or they might be traffic dealers who resell minutes to other customers in order to let them make their calls on charged based manner. The same should be implemented on durations. The average duration for each subscriber during a call is 7.13 minutes (which is 427 seconds) for all of his communities, the lowest 1% percentiles call count is 0.66 call during the month. This means that removing the lowest 1% of subscribers will improve the accuracy of the model. Table (3-1) shows the

distribution of average duration for the selected sample of subscribers, the statement of performing this exercise is listed in Appendix (1,C)

Table (3.1): The distribution table for community size, duration per month, count of calls and duration per call attributes.

	1st Pctl	10 Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl	95th Pctl	99th Pctl
Community Size (Number of nodes)	1	2	4	9	20	32	42	67
Duration Per Month (Minutes)	2.33	17.66	42.33	107.00	230.66	383.66	497.00	809.33
Count of Calls (Number of calls)	0.66	2.66	6.66	16.66	36.00	59.33	76.66	125.66
Duration Per Call (Minutes)	1.15	4.22	5.42	6.44	7.64	9.57	12.06	27.50

3.3.3 Merging the Two Datasets

After removing the non-eligible and the outliers' subscribers, the targeted base went down to 24,300 subscribers instead of the original 25,000. A process of merging the two datasets together (Aggregated CDRs and Profile) will take a place. This is needed to extract more profile data rather than only statistical measures. For example, if we found that Jawwal has a number of 20,000 subscribers considered as influencers we might need to get more details about their region place, average revenues (is it high or not?) and are they prepaid or postpaid subscribers. This adds another layer of validating the results with what was expected and comparing them with the current perspective (Region X generates more revenues than Y, does it include more influencers also?).

```
CREATE TABLE data_for_modelling
PARALLEL AS
 SELECT *
 FROM   sna_sample_25000_aggr_cdrs a,
        sna_sample_25000_profile b
 WHERE  a.a_number = b.a_number;
```

The create table command is used to extract the data from other tables in the database into a new table, parallel is a command used for the session to speed up the process where the values in "WHERE" clause is the join key between the two mentioned tables in "FROM" clause.

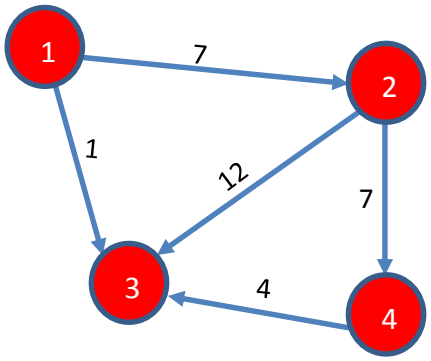


Figure (3.2): Example Social Network

Once the dataset for the model development is created, it's possible to start the modelling activity. The previous procedures related to understanding, cleaning and profiling of data is important to ensure a derived and reliable output.

3.4 Data Modeling

A social network is typically illustrated using a sociogram (Dasgupta, et al., 2008). In this type of visual display, individuals correspond to points, or nodes, in a space. Lines, or edges, connecting the points represent relationships between the individuals. If the relationships are directional, the edges include an arrow to indicate the direction. If the relationships have weights, the labels for the edges denote the values. Figure (3-3)

displays a network of four individuals (1,2,3 and 4) along with their relationship strengths (represented as total duration between them in 7,1,12,7 and 4).

Suppose the network represents the phone calls made by individuals with the relationship weights indicating the length of the calls (duration). In this case, Node #1 called two nodes, spending the majority of time talking to node # 2.

It's all about communities (a group of people who have a shared pattern that distinguishes them from other communities). The sample of 25,000 subscribers is, with no doubt, a combination of different communities indicating influencers in each group. The top five percentiles have a group size of 43 – 67 will be divided into more sub communities. It's clear that there is a clear gap between the latest 90^{th} and 95^{th} percentiles compared to 95^{th} and 99^{th}, a gap from 10 to 24 calculated by subtracting (42-32) and (67-42) respectively! The group will be set to include members from 2 to 42 (minimum and 95^{th} percentile) .

To summarize the process, the next steps are about dividing the sample into groups (Partitioning into groups), describing networks (network properties) and calculating influence factor (describing groups and group members).

3.4.1 Partitioning into Groups

Individuals who have high relationship weights (frequency of calls or total duration) between each other are grouped together, they are similar to each other and this similarity of individuals in a group is measured by weights.

As a result, group identification begins by omitting the weaker relationships in a network. The process of omitting those weak relationships is called the coverage threshold and this threshold is examined as theta (θ) where it is defined as the fraction of the strongest relationships to retain. For example, a coverage threshold value of 0.3 results in the strongest 30% of relationships being used for group identification with omitting the remaining 70% of the relationships.

In addition, group size limits will help in portioning the groups to either split those who have large number of members into separate groups or to completely omit those who have smaller members. Normally, the group limit is a minimum of two members and will be, in this case study, a maximum of 42 members. The groups remaining are called 'core groups' (IBM Corporation, 2012) . Table 3-2 illustrates the relation between coverage threshold and average group size by carrying out experiment with different θ from 10% to 95%.

Table (3.2): Coverage threshold factor distribution

Coverage Threshold (Θ)	**10%**	**15%**	**30%**	**50%**	**60%**	**80%**	**90%**	**95%**
Total Nodes in Network	16,222	17,184	19,085	20,581	20,720	22,509	22,509	22,509
Total Links in Network	34,382	38,416	43,856	44,532	44,978	41,392	41,392	41,392
Total Number of Groups	1,903	1,862	1,633	1,487	1,471	1,197	1,197	1,197
Mean Group Size	8.52	9.23	11.69	13.84	14.09	18.8	18.8	18.8
Mean Group Density	0.3	0.31	0.3	0.29	0.29	0.25	0.25	0.25
Mean In-Degree	2.12	2.24	2.3	2.16	2.17	1.84	1.84	1.84
Mean Out-Degree	2.12	2.24	2.3	2.16	2.17	1.84	1.84	1.84

"Total Nodes in Network" refer to total subscribers resulted after applying each coverage threshold and omitting weaker relationships, "Total Links in Network" is number of connections that a network member originated to

another network member, “Total Number of Groups” stands for count of output communities inside the dataset, “Mean Group Size” is important to judge about the optimal community size in terms of number of groups and nodes count where this will assist in understanding the size of groups, “Mean Group Density” is how the relations within each group exists in terms of actual vs potential (permutations) relations. Finally, “Mean Out-Degree” and “Mean In-Degree” is the average outgoing and incoming connections inside each community. The next sections will present in details on how to calculate these variables.

Table 3-2 shows that the more coverage threshold is, the less tight relationships weight (which might include weak relationships), the more average group members (including non-core members). The last three coverage thresholds (80%, 90% and 95%) resulted in same distribution of group members which concludes, for our sample dataset, that applying any of these thresholds will result in the same. In another meaning, it’s is not significantly different to apply 80%,90% or 95% as coverage threshold.

Splitting members into groups is based on the similarity between those members within the group and the dissimilarity between the groups and where nodes inside the group connected with many more edges than between groups (Newman, 2004). Similarity is measured through examining the neighbors of each member and the strength of relation between members and their neighbors, two nodes are similar if when you start random walk from those two nodes, those random walks will meet soon. Random Walk is “an algorithm provides a good relevance score

between two nodes in a weighted graph, that is how closely two nodes are related and it has been successfully used in numerous settings, like automatic captioning of images, generalizations to the connection subgraphs, personalized PageRank, and many more" (Tong, et al., 2006) .

There are two types of similarity measures considered:

1- Structural Equivalence: Members within the group have a structural similarity compared to each other, they share the same neighbors and approximately the same relation strength. Figure (3-4) shows that nodes #26 and #29 share the same neighbors (shared neighbors are

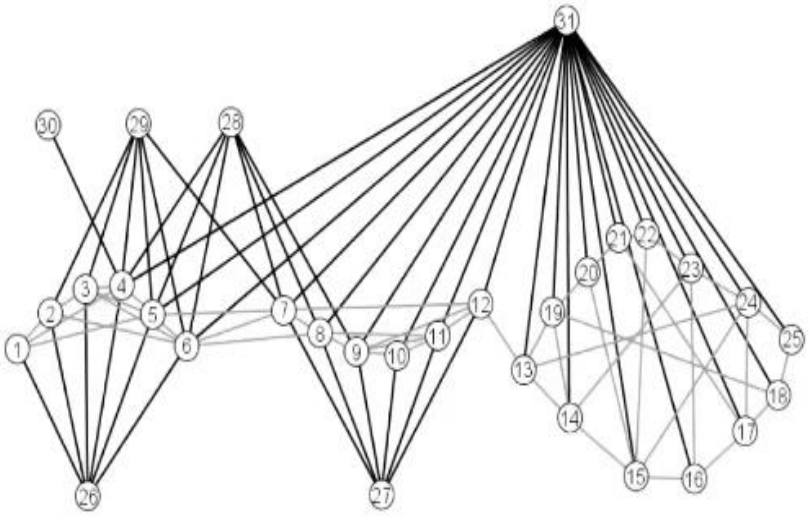


FIGURE (3.3) SPLITTING MEMBERS INTO GROUPS: STRUCTURAL SIMILARITY BETWEEN #26 AND #29

nodes 1,2,3,4,5,6) out of six neighbors for node #26 (1,2,3,4,5,6) and seven for the node #29 (1,2,3,4,5,6 and 7).

In order to calculate the structural similarity, we have to use one of the below three methods depending on type of network graph (direct or non-directed) and the availability of relationship weights (Weighted or non-Weighted):

A- Jaccard Similarity: Studies the intersection between two individuals' neighbors divided by the union of those individual neighbors sets where the weights are not available:

$$J(v_i, v_j) = \frac{|N(v_i) \cap N(v_j)|}{\left||N(v_i) \cup N(v_j)|\right|} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(3.1)}$$

Where N stands for the neighbors. For example, the Jaccard Similarity between nodes #26 and #29 is

$$J(26,29) = \frac{|[1,2,3,4,5,6] \cap [1,2,3,4,5,6,7]|}{|[1,2,3,4,5,6] \cup [1,2,3,4,5,6,7]|} = 6/7 = 0.85$$

The similarity between nodes 26 and 29 are 0.85 which is strong enough to group them into a community (remember the coverage threshold that above 15%). Another example is J(26,27) which equals Zero.

B- Cosign Similarity (Adjacency Matrix): The graph can be waited in which the relation could have a weight instead of only representing the neighbors count.

$$\sigma(Vi, Vj) = \cos(\theta ij) = \frac{Vi^T\, Vj}{|Vi||Vj|} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum A_{ik}^2}\sqrt{\sum A_{jk}^2}} \quad \ldots\ldots\ldots\ldots \text{(3.2)}$$

Where T is the transpose of the Adjacency matrix, K is the node degree. In a simple form (for unweighted graph as in Jaccard), the Numerator represents number of shared neighbors where the Denominator represents nodes degree for i and j.

C- Pearson Correlation Coefficient:

$$r_{ij} = \frac{\sum_k (A_{ik} - \overline{A_i})\,(A_{jk} - \overline{A_j})}{\sqrt{\sum_k (A_{ik} - \overline{A_i})^2}\,\sqrt{\sum_k (A_{jk} - \overline{A_j})^2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(3.3)}$$

Pearson Correlation will not equal zero if there was no overlap between the nodes (In contrast with the Cosign method). Other than that, we can use either Cosign of Pearson for computing the similarity Pearson Correlation Coefficient.

2- Regular Equivalence: Two nodes are regularly equivalent if they are equally related to equivalent others, in another meaning, they have the same pattern of connecting to others. Example of nodes #26 and #27 in the above figure, they are both connected to seven neighbors in which also those neighbors are shared with two other nodes. Regular equivalence provides nothing for the group identification in terms of community, but it does when it comes to extract the similar pattern within all subscribers' base. Another example is when there are two office managers who have a team of five employees reporting to them, these office managers have the same number of followers but they are at the same time separated into two different communities. (Newman, 2004).

3.4.2 Describing Networks

Information about networks, groups, and individuals needs to be extracted into descriptive characteristics that allow cross-comparisons and inclusion in predictive models based on a set of key performance indicates. One of the most used examples is comparing a group of nodes to another within the network as to describe the characteristics of the former compared to the

latest, other example is compare individuals in the network to others to characterize them or to identify the most important ones.

In order to describe social networks, density and degree are the most common used measures. They both reflect connectivity, density focuses on the entire network or network subs groups where degree describes the individuals in the network and characterizes them. (IBM Corporation , 2012)

3.4.2.1 Network Density

Network density is defined as the proportion of actual connections (relationships) between nodes in a network (group) divided by total possible connections in that group (IBM Corporation , 2012). The more actual connections, the denser is the network and the more cohesive are the nodes in the network. Density ranges from 0 to 1 taking into considerations that connections are either one way or two ways between the nodes (directed or undirected). Some nodes may not have direct connections to other nodes, some nodes may have only outgoing connection to other nodes (the relation is not reciprocated). Figure(3-5) stimulates a group with actual connections (lines in blue) compared to all the possible connections (lines in yellow).

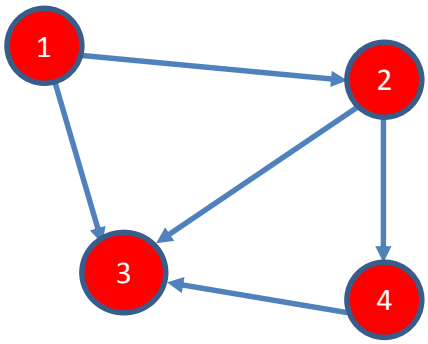


Figure (3.4) A: Actual connections

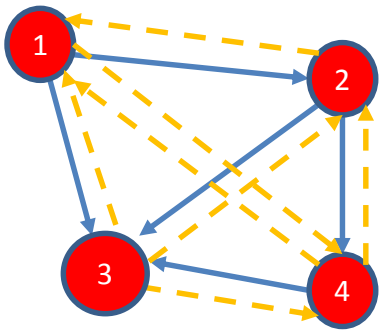


Figure (3.4) B: Possible Connections

In Figure (3-5A), the count of connections between the nodes is 5 (and it's clear that the connections are all 1 way directed). However, Figure (3-5B) shows that a possible of 7 more connections can exist and regardless if they are 1 or 2 ways. For example, the connections between nodes one and four are reciprocated. This means that the total possible connections in this group is 12. Another term of the figure A is 'directed' as it doesn't contain any reciprocal connection, Figure B is undirected. (IBM Corporation, 2012)

In a tabular format, Table(3-3) represents all possible connections for this group of nodes.

Table (3.3): Possible connections (Permutations)

Node A	Node B
1	2
1	3
1	4
2	1
2	3
2	4
3	1
3	2
3	4
4	1
4	2
4	3

So, possible connections calculation is somehow the permutation where the size in the set is always two (a connection is made between two nodes).

The formula for Density:

$$Network\ Density = \frac{Actual\ Connections\ (AC)}{Potential\ Connections\ (PC)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (3.4)$$

The formula for Potential Connections:

$$Potential\ Connections\ (PC) = \frac{n!}{(n-r)!} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (3.5)$$

where n! (n factorial) is count of nodes in the network, r = 2, n>1 and 0! = 1. The network density for the above sample of Figure(3-5) =

$$Network\ Density = \frac{5}{\frac{4!}{(4-2)!}} = \frac{5}{12} = 0.42$$

3.4.2.2 Nodal Degree

Nodal stands for individuals, degree for connections (relationships). The more connections that an individual has (both in and out), the more this individual is important in the network as s/he is involved in the most relationships. This individual acquires information from a variety of sources and spreads it to a large number of other group members. So, degree is defined as total number of connections involving each node in a network and is classified to either 'in-degree' or 'out-degree'. In the first, a node is considered to be the target involving how many connections were made to this node. Conversely, out-degree is the number of connections made where the node is the source instead of being the target. Table (3-4)

summarizes Figure (3-5) A in terms of actual connections set to the node in the cases of target vs. source.

Table (3.4): Representing Figure (3.5 A) connections pattern

	Count of connections (In Vs Out)		
Node #	In-degree	Out-degree	Total Degrees
1	0	2	2
2	1	2	3
3	3	0	3
4	1	1	2

In-degree is often treated as a measure of prestige. Higher in-degree values correspond to more relationships ending at that node. In other words, those individuals are contacted by a high number of other individuals. Many other nodes are initiating relationships with the node. Conversely, out-degree is treated as a measure of centrality. Higher values correspond to more relationships originating from that node. Those individuals contact a high number of other individuals. (IBM Corporation, 2012)

For nodes in the network of Figure (3-5A), the out-degree values reveal that node #3 is the most central. Based on the in-degree values, nodes #2 and #4 have more prestige than #1 and #3. Although, Node #2 is more active compared to node #4 (it has a total degree of three compared to two).

3.4.3 Describing Groups and Groups Members

In addition to the density, in-degree, and out-degree, other statistics describe group dynamics. In particular, authority and dissemination scores offer measures of the social status of the individuals within the groups. The role of each individual in a group is vitally important when trying to predict the behavior of both the group and its members.

Figure (3-6) indicates the weights of relationships between group nodes and in terms of duration. For example, node #2 called node #3 by 12 minutes, and at the same time node #3 received 12 minutes from node #2. Who are authority and dissemination influencers?

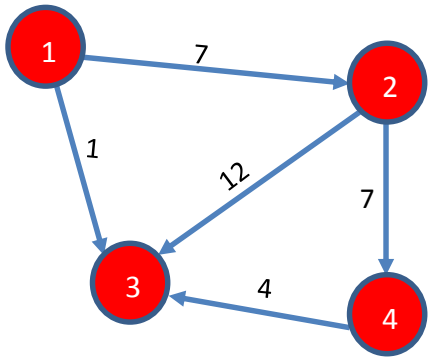


Figure (3.5): Connections Weights

Authority and dissemination algorithms are based on the weights of relationships between group members. The more an individual receives from others, the more they ask him or her for information or opinions. This is called Authority. On contrary, the more an individual connects to other members, the more he has an ability to spread the word. This is called dissemination. Table (3-4) summarizes the previously mentioned group on the level of each node.

Table (3.5): illustrating Figure (3.6) details

Node #	Indegree	Outdegree	Incoming Duration	Outgoing Duration
1	0	2	0	8
2	1	2	7	19
3	3	0	17	0
4	1	1	7	4

There is two widely used algorithms to calculate the authority and dissemination, either to use the PageRank or HITS algorithms and taking

into considerations that both algorithms are based on the right understanding of adjacency matrix (Adj) resulting in the score of importance of each node inside the network. (Devi, et al., 2014) The adjacency matrix (stochastic probability) of Figure (3-6) is as below:

$$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{1} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{1} & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & \frac{7}{8} & \frac{1}{8} & 0 \\ 0 & 0 & \frac{12}{19} & \frac{7}{19} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{1} & 0 \end{bmatrix}$$

Adjacency Matrix (Adj. Binary) — Stochastic Adj. Matrix (Relationships) — Stochastic Adj. Matrix (Weighted)

Each row represents node number, the same for each column. R_{ij} is equal to 1 if i had any outgoing transaction to j (i→j exists). Row#1 & Col #1 represent Node#1, the same for other nodes. To describe the above matrix, Node #1 originated some transactions to nodes #2 & #3 where it didn't originate any outgoing transaction to Node #4. On contrary, Node #3 had no outgoing transaction to any other neighbors which explains the straight zero in the third row.

Probability Adjacency Matrix measures the power of relationship between the nodes, node #2 originated 1 transaction to node #3 out of total two transactions originated by node #2 (to node #3 and node #4). The representation of this relation is ½. We should also note that the sum of each row should be 1.

According to Figure (3-6), it indicates the relationship weight between nodes in this group. For example, Node #1 originated a total of 8 minutes to its neighbors divided into two relations (Node#1→Node#2 and

Node#1→Node#3). The first relation power is 7/8 (7 minutes originated from Node#1 toward node #2 out of total 8 minutes) compared to the latest 1/8.

$$\begin{bmatrix} 0 & 0.875 & 0.125 & 0 \\ 0 & 0 & 0.632 & 0.368 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

3.4.4 PageRank and HITS

Pooja Devi, Ashlesha Gupta and Ashutosh Dixit in their published journal article titled *'Comparative Study of HITS and PageRank Link based Ranking Algorithms'* defined both algorithms, summarized the advantages and disadvantages of each algorithm and exercised them on a network of webpages (represented as nodes A and B).

Link analysis algorithms are used to calculate webpage rank. An example of these algorithms are Pagerank and HITS algorithm, they are different link analysis that employ different models to calculate the rank. Both algorithms give importance to links rather than the content of pages. In the Pagerank, rank score of a page is divided evenly over the pages to which it links. But, HITS algorithm rank pages according to their Authority and Hub of a page. Google search engine uses Pagerank algorithm due to the features of link feasibility, less query time cost, efficiency and less susceptibility to localized links. HITS is used by IBM in all of its applications and search engine. Results demonstrate that HITS calculates authority nodes and Hub correctly. HITS may also be combined with other information retrieval based rankings. (Devi, et al., 2014)

One of the main advantages of HITS over PageRank is that the first uses the input of three important parameters (In-degree, Out-degree and relationship weights) where the latest is based on 'In-degree' only. Regarding the disadvantages, the first requires more processing time (due to the variability of inputs and relations) where the latest is considered to be more efficient in time manner. What to use is based on the context. IBM Modeler uses HITS algorithm as in the same for IBM Search Engine.

A Simplified version of PageRank is defined in the below Equation:

$$PR(A) = (1-d) + d[\frac{PR(T_1)}{C(T_1)} + \ldots\ldots\ldots + \frac{PR(T_n)}{C(T_n)}] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.6)$$

Where PR(A) is the PageRank of page A, $PR(T_i)$ is the PageRank of pages T_i which link to page A, $C(T_i)$ is the number of outbound links on page T_i and d is a damping factor which can be set between 0 and 1. (Sobek, 2003)

As mentioned earlier, HITS algorithm takes into considerations both In-degree and out-degree (which means the Authority and Hub scores). The algorithm is as below:

$$\text{Authority: } a_i^{(t+1)} = \sum_{j:j\to i} h_j^{(t)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.7)$$

$$\text{Hub: } h_i^{(t+1)} = \sum_{j:i\to j} a_j^{(t+1)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.8)$$

Where "i → j" stands for that page i links to page j and a_i is authority of i^{th} page and h_i is the hub representation of i^{th} page. HITS algorithm forms the community adjacency matrix A, whose m (i, j) element is 1 if page i links to page j and 0 otherwise. Number of iterations (t) is used to reach the converge state of the scores. (Manning, et al., 2008)

3.4.4.1 Authority Score

Authority scores measure the importance of an individual corresponding the number of relationships ending at him or her indicating how much an individual receives from other members of a group. The more incoming connections to this node, the more s/he has authority compared to the group members who are asking for opinions or information. Authority scores ranges from 0 to 1, the closer to 1 indicates a closer authority and the higher potential of being the authority leader of that group. (IBM Corporation, 2012). The confidence of authority for the leader (overall strength of the authority leader) is computed through dividing highest authority score in the group by lowest authority score, following the below equation:

$$Authority\ Leader\ Confidence\ = \frac{Max\ (Authority\ Score)}{Min\ (Authority\ Score)} \quad \ldots\ldots\ldots\ldots\ldots (3.9)$$

3.4.4.2 Dissemination Score

In contrast, dissemination score measures the importance of an individual with the number of relationships originating from him or her indicating how much an individual connects to other members of a group. If a particular person contacts many other people in the group, that person can significantly affect the opinions of the entire group. The closer the dissemination score is to 1, the more the node connects to the other group members. The dissemination leader for the group is the one who has the maximum dissemination score among other members (IBM Corporation, 2012). The confidence of dissemination for the leader (overall strength of

the dissemination leader) is computed through dividing highest dissemination score in the group by lowest dissemination score, following the below equation:

$$Dissemination\ Leader\ Confidence\ = \frac{Max\ (\text{dissemination}\ Score)}{Min\ (\text{dissemination}\ Score)} \ldots. (3.10)$$

3.4.5 Calculating HITS for Figure (3-6) Network

To summarize the previous sections, hub and authority measure the importance of the nodes in a network, hub score measures the ability to disseminate (how frequent does a specific node communicate to other nodes in a network, which is the spread of dissemination of info) where authority measures the prestige of nodes (how frequent do other nodes in a network communicate to a specific node). The two scores algorithms were discussed as well as the use of adjacency matrix and the relationship between the nodes. Let's not forget that good hubs are those who always point to good authorities and good authorities are those who were pointed from good hubs (the algorithms indicate that the relationship between the two scores are recursive).

Researcher will start by initializing hub scores, calculate the authority scores and recalculate the hub scores. Let 'X' be the authority score, 'Y' is Hub score, '$\bar{X}$' (X bar) is the normalization of the results of authority scores to the sum of 1 and '$\bar{Y}$' (Y bar) is same as 'X' but for the dissemination. 'A' is the adjacency matrix and 'A^T' is the transpose of this matrix (in order to multiply the matrices).

The adjacency matrix {A} based on Relationships weight:

Node # (i→j)	1	2	3	4
1	0	7	1	0
2	0	0	12	7
3	0	0	0	0
4	0	0	4	0

A^T	1	2	3	4
1	0	0	0	0
2	7	0	0	0
3	1	12	0	4
4	0	7	0	0

Initialization of Y:

$Y_{(0)} = \{1, 1, 1, 1\}$ as there are *Four* nodes in the group.

The Authority Score (X):

$$X_i = \{A^T\} * \{Y_{(i-1)}\} \quad \text{(3.11)}$$

The Hub Score (Y):

$$Y_i = \{A\} * \{X_{(i)}\} \quad \text{(3.12)}$$

The Normalization of (X):

$$\bar{X} = \frac{x}{\sqrt{\Sigma_{x \varepsilon X}(X^2)}} \quad \text{(3.13)}$$

The Normalization of (Y):

$$\bar{Y} = \frac{y}{\sqrt{\Sigma_{y \varepsilon Y}(y^2)}} \quad \text{(3.14)}$$

All iterations till converge are listed in Appendix C as to present the change of values on every iteration along with calculation steps for Authority (x), Hub (h) and the normalized authority ($\bar{X}$) and hub ($\bar{Y}$). Tables (3-6, 3-7) summarize the normalized authority and hub scores per iteration.

The process of calculating Authority and Hub scores is based on Link analysis book section published at Stanford university website. (Manning, et al., 2008) . Number of trials is an indicator of when the results are going

to converged – there are no major changes in the scores after certain number of trials in which difference between two iterations is very small (smaller than an epsilon we define) -. Moreover, the use of Eigenvectors will speed up the process of reaching converge. An example of handmade calculations is below where it took around 10 trials to reach the state of converge.

Table (3.6): Authority Scores Results

Node #	Iteration	1	2	3	4
Authority Scores ($\bar{X}$)	1	-	0.356	0.864	0.356
	2	-	0.120	0.879	0.461
	3	-	0.059	0.881	0.469
	4	-	0.044	0.881	0.471
	5	-	0.044	0.881	0.471
	6	-	0.040	0.881	0.471
	7	-	0.040	0.881	0.471
	8	-	0.039	0.881	0.471
	9	-	0.039	0.881	0.471
	10	-	0.039	0.881	0.471

Table (3.7): Dissemination (Hub) Scores Results

Node #	Iteration	1	2	3	4
Hub Scores ($\bar{Y}$)	1	0.244	0.936	-	0.252
	2	0.120	0.962	-	0.245
	3	0.090	0.965	-	0.245
	4	0.090	0.965	-	0.245
	5	0.081	0.966	-	0.245
	6	0.081	0.966	-	0.245
	7	0.081	0.966	-	0.245
	8	0.081	0.966	-	0.245
	9	0.081	0.966	-	0.245
	10	0.081	0.966	-	0.245

3.4.6 Interpretations of results

Examining the above tables results, the converged state was reached after 8 trials (the result of trials 8 to 10 had no change). On the other hand, converge state in Hub was reached after 5 trials. The reason behind this is that structure of the network. In terms of authority, Nodes #4 and #2 had the same indegree which will require more trials to set the authority scores and based on their relationship strengths (they both had the same incoming relationship strength, which is equal to 7). On the other hand, the hub converge state was reached faster explaining that structure of out-degree is more simple than in-degree, for the above given sample of network.

It's clear that Node #1 had no authority score as it has no in-degree relationships compared to other group members. In addition, Node #3 ranked at the top of authority scores for two reasons, the first is it has the highest in-degree relationships and the strength of these relationships (the sum of incoming traffic weights) is the highest toward all other targeted members compared to others. When it comes to nodes #2 and #4, they both have the same in-degree but they differ in terms of incoming relationship weights. Node #2 had more power to connect to the top authority (node #3) than node #1, this will reflect the strength of relationship between node #2 and node #4 compared to node #1 and node #2. This is why Node #4 ranked higher than node #1 in terms of authority.

The case of hub is not completely different from the authority scores results; they both share the same concept however the latest is about incoming where the earliest is about outgoing. When it comes to hub – the

spreading of the word – we can notice that node #2 ranked at the top, followed by node #4 then node #1. In addition, node #3 had no dissemination score since it had no outgoing degree. As a conclusion, good hubs always point to good authorities. (Stanford, 2008)

3.5 Using SPSS Modeler

A complete guide on how to use SPSS Modeler software is listed in Appendices, Appendix 2. Researcher explained the process of loading the nodes of data and group analysis, importing the datasets into modeler environment, configuring SNA model coverage threshold and defining relationship operators (originator, receivers and weight factors), using the export of results node and finally establishing connections between all of these nodes.

Chapter Four

Research Results

4.1 Applying SNA on Jawwal sample

The concept of SNA in the previous sections was discussed starting from Group analysis measures such as number, density, size & confidence of groups and Individual analysis like nodal degrees, authority and hub scores. In order to apply the SNA on the projected sample, IBM SPSS Modeler is used applying HITS algorithm through IBM SPSS Modeler SNA Group Analysis Node. Below is a snapshot for two groups community (group numbers of #59 & #91) result generated by this Node. The term GAG stands for 'Group Analysis Group', which is for group analysis, and GAI for 'Group Analysis Individual' – for nodal analysis.

4.1.1 Group #59 Results

In this sub section, a visual representation of group #59 relationships weights and directions in addition to the group analysis and Individual analysis tables output will be discussed.

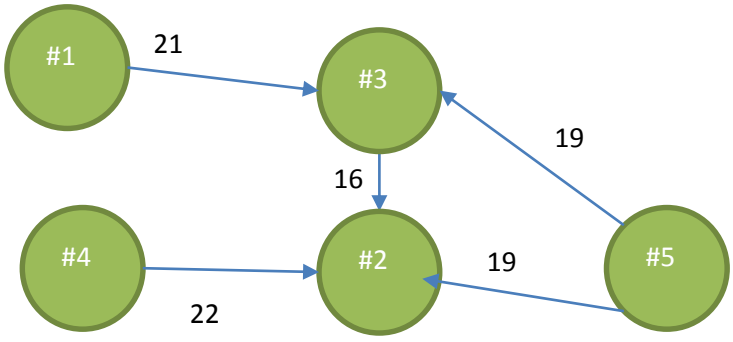


FIGURE (4.1): GROUP #59 AND ITS MEMBERS' RELATIONSHIPS WEIGHTS (TRAFFIC)

Table (4.1): GAG Analysis

GAG_GroupNumber	59
GAG_Size	5
GAG_Density	0.25
GAG_MaxRankType1	0.455877
GAG_MinRankType1	0.101223
GAG_MaxMinRankRatioType1	4.50368
GAG_MaxRankType2	0.26307
GAG_MinRankType2	0.145487
GAG_MaxMinRankRatioType2	1.80821

Rank Type 1: Authority Score; Rank Type 2: Dissemination (Hub) Score; Rank Ratio: Confidence Score.

Table (4.2): GAI Analysis

GAI_NodeNumber	Member #1	Member #2	Member #3	Member #4	Member #5
GAI_RankType1	0.101223	0.455877	0.240453	0.101223	0.101223
GAI_RankOrderType1 * (Authority)	5	1	2	3	4
GAI_RankType2	0.22567	0.145487	0.182887	0.182887	0.26307
GAI_RankOrderType2 * (Dissemination)	2	5	3	4	1
GAI_InDegree	0	3	2	0	0
GAI_OutDegree	1	0	1	1	2
GAI_GroupLeaderType1	0	1	0	0	0
GAI_GroupLeader ConfidenceType1	0	4.50367	0	0	0
GAI_GroupLeaderType2	0	0	0	0	1
GAI_GroupLeader ConfidenceType2	0	0	0	0	1.80821

* Rank Order: Used to rank group members based on their authority or dissemination score;

Examining Table (4-2), Member #2 has the highest rank type 1 (Authority Score) which means that all other network members, as mentioned earlier, refer to him before taking any decision. Having member #2 outside the network means that there will be no reference for other group members

which will lead to a higher probability for others to leave the network as we can notice that member #2 had the highest in-degree as well (although highest in-degree will not guarantee having a member as influencer since the weight is traffic based), this is the churn influence of leader on other members. For those who will adopt, Member #5 has the highest rank type 2 which is more responsible to disseminate (spread the word) to other network members. Having member #5 convinced in a service and buying it will lead to a higher probability for others to do so as we also can notice that member #5 has the highest out-degree, this is effect of dissemination of other network members. Group members 1,3 and 4 will be more probable to have their loyalty threatened as result of having both members #2 and #5 not satisfied.

4.1.2 Group #91 Results

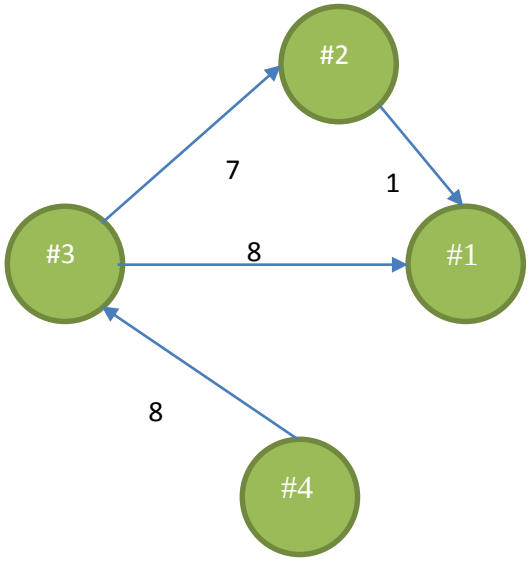


Figure (4.2): Group #91 and its members' relationships weights (traffic)

Table (4.3): GAG Analysis

GAG_GroupNumber	91
GAG_Size	4
GAG_Density	0.333333
GAG_MaxRankType1	0.427861
GAG_MinRankType1	0.126891
GAG_MaxMinRankRatioType1	3.37187
GAG_MaxRankType2	0.372787
GAG_MinRankType2	0.121375
GAG_MaxMinRankRatioType2	3.07136

Rank Type 1: Authority Score; Rank Type 2: Dissemination (Hub) Score; Rank Ratio: Confidence Score.

Table (4.4): GAI Analysis

GAI_NodeNumber	Member #1	Member #2	Member #3	Member #4
GAI_RankType1	0.427861	0.212539	0.232709	0.126891
GAI_RankOrder Type1	1	3	2	4
GAI_RankType2	0.121375	0.177943	0.327894	0.372787
GAI_RankOrder Type2	4	3	2	1
GAI_InDegree	2	1	1	0
GAI_OutDegree	0	1	2	1
GAI_GroupLeader Type1	1	0	0	0
GAI_GroupLeader ConfidenceType1	3.37187	0	0	0
GAI_GroupLeader Type2	0	0	0	1
GAI_GroupLeader ConfidenceType2	0	0	0	3.07136

Rank Order: Used to rank group members based on their authority or dissemination score;

Group (#91) is a structurally different group compared to the previous one (#59) as to prove that not having in-degree might lead to an authority score (depends on hub values pointing to the node since member #4 has authority score although he has no in-degree, it is related to IBM Modeller that subtracts one, for non in-degree nodes, from the rest of authority nodes

who have in-degree), the same for hub values and out-degrees that the solution iterations converge at. Member#1 has the highest RankType1, he is the authority leader and he has the highest in-degree and the highest total of incoming weight (duration). On the other hand, Member #4 was ranked as the hub leader. Although he has one out-degree relationship compared to member #3 who has two out-degrees, member #4 has a relationship weight which equal to 8 and totally focused toward one targeted member. There is no distribution of relationship weight as the case of member #3, when he talks we use all his strength to convince! For the same reason, member #4 has ranked as top compared to member #2 who has strength toward his target equal to one as well. Moreover, authority group leader confidence is higher than disseminator leader confidence which concludes that this network tends to be influenced by its authority leader higher than its disseminator leader (Member #1 has higher in-degrees compared to Member#4 out-degrees)

4.2 ROI Monte-Carlo Simulation

According to the previous analysis, the average group size of selected samples ranges from 2 to 42. Let's assume that these group members daily revenues range from $0.3 to $2 on average. Any commercial company should have a vision to sustain its customers the longest period they can through serving them better and satisfying their needs. This will also guarantee a positive word of mouth that will help company expands its base through the feedback of its current customers. As a result, the more

customer satisfaction a company has, the more customer age is expected and the more revenue this company gains. This section proves that incorporating SNA with targeting process will have a higher and positive impact on ROI compared to the traditional ways of targeting.

According to both Vodafone & Jawwal benchmarks, the trend of revenue uplift in the most of given offer is ranged between 15% and 25% on individual level. But, when it comes to SNA we start to think about group level instead. Those leaders have the power to influence others to buy a company product, to subscribe for an offer or to stay connected to the same network of those leaders. In the normal life, the percent of accepting offers given by telecom operators do range between 7% and 12% - It's also called the opt in ratio. But the case is different when start looking at the whole view through using influencers. It may reach up to 30% specially when utilizing SNA power to target group leaders. In addition to opt in rate and revenue uplift, company should be aware of the cost paid. It's the cost of advertising, systems, employees' payrolls, tools and so on. All of these variables are combined in one important term called Return On Investment (ROI), at the end each activity is an investment.

The cost of targeting a group of 5 members individually - which means they might have five different offers according to their daily revenue tiers - is higher than targeting one member and let him influence others. Other important factors are offer utilization and customer satisfaction. Let's imagine a family that consists of five members again, one of them generates a daily revenue equal to $2 and another equal to $0.5. In

traditional ways, two offers will be sent to these members - assuming the concept of buy X minutes for $Y - as the following:

- For the first member, buy 10 minutes for $2.5 (here the goal is to stretch his daily revenue by $0.5)
- For the second member, the goal is to stretch by $0.20 since he got used to spend only $0.5 on a daily basis.

It's more feasible for the first member to buy the offer than second member in the traditional way. But, when it comes to value, offer owners will try to show the second member the value he would take by offering him 10 minutes for the sake of changing his behavior and increasing his daily revenues although they know that he will not abuse the offer to talk all 10 minutes. What will be the feedback of the first member? Well, he will feel the bias and this might force him to either complain or think about switching to another company. The most favorable action from his side is to utilize the free minutes of the second member and save his money! This might cause the company the total loss of $2.5 revenues generated by in addition to the probability of losing him. This section includes a spreadsheet simulation using Monte Carlo to prove how will company save or stretch its current subscribers base of revenues when incorporating SNA following the below hypothesis:

H_0: $\mu 1 = \mu 2$

H_1: $\mu 1 \neq \mu 2$

where

H_0 = the null hypothesis,

H_1 = the alternative hypothesis,

μ_1 = the mean of ROI without incorporating SNA,

μ_2 = the mean of ROI when incorporating SNA.

Null hypothesis assumes that there is no difference in group ROI mean when incorporating SNA vs not. The simulation of the two cases was done using Monte Carlo simulation taking into considerations the following parameters and for 10,000 trials.

Table (4.5): General parameters used in offers targeting

	From	To
Group Size	2	42
Daily Revenues	$0.30	$2.00
Cost per offer: Time, System & Communication	$0.05	$0.10
Active Days	1	12

Table 4-5 above shows the ranges obtained from the main variables of the study. As such, we can obtain from the sample a group with number of members varying from 2 members to 42. Similarly, the possible outcome of the daily revenues varies from 0.3 to 2, and the cost per offer varies from 0.05 to 0.1. Finally, active days – stands for how many days a subscriber will be active on the offer during a month- varies on average from 1 to 12.

Table (4.6): Assumptions used to compare ROI when incorporating SNA vs not

With Vs Without SNA	With SNA	Without SNA
Offer Acceptance Ratio	30%	10%
# of Offers	2	# of members per group
Revenue Uplift	15 to 25%	15 to 25%

Table 4-6 above states the assumptions used in the model. It should be taking into consideration that in the case of SNA only one offer will be sent to the leader and the rest of the group may or may not feel tempted to join the offer, while in the case where SNA is not applied the offer will be sent to all subscribers. Therefore, it was assumed that the leader would have a higher effect on the opt in rate of his followers.

4.2.1 Independent Samples T-test: Compare Two Means

The results of the independent sample T test, for ROI without SNA and ROI with SNA with 10k observation per each is below:

Table (4.7): Group Statistics

Group Statistics

	With SNA	N	Mean	Std. Deviation	Std. Error Mean
ROI	No	10000	94.676	241.644	2.416
	Yes	10000	4749.084	5929.574	59.2950

Table (4.8): Independent Samples Test

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	T	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
ROI	Equal variances assumed	9956.36	.000	-78.4	19998	.000	-4654.408403	59.344	-4770.72943	-4538.0873
	Equal variances not assumed			-78.4	10032.21	.000	-4654.408403	59.344	-4770.73642	-4538.0803

From table above, we conclude that since the p-value is less than 0.05, then the groups' means are significantly different. This means that ROI with or without SNA do significantly differ from each other. As a conclusion, we will reject the null the null hypothesis that they are both equal with a degree of freedom (df) of 19,998. Using SNA may result to increase ROI up to 50 times compared to traditional ways, below are the histograms from both scenarios.

Histograms

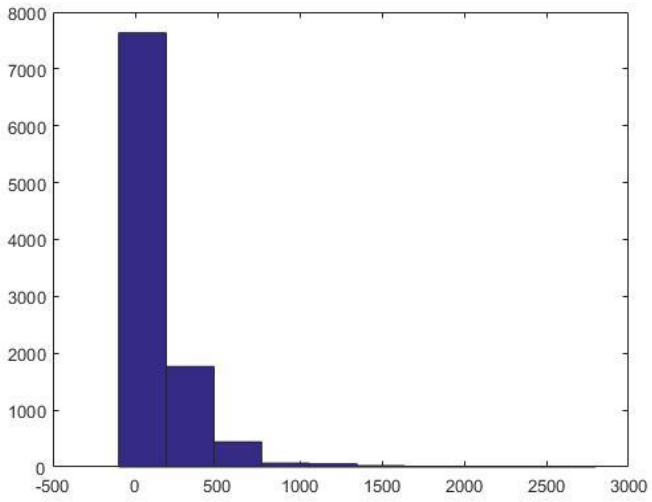


FIGURE (4.3): ROI HISTOGRAM WITHOUT APPLYING SNA

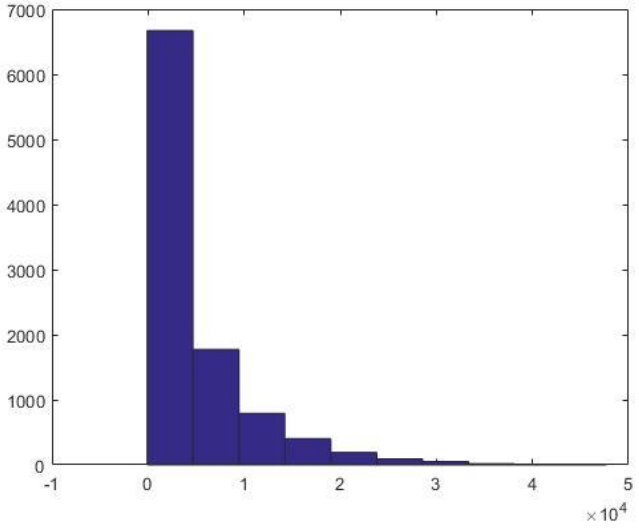


FIGURE (4.4): ROI HISTOGRAM WHEN APPLYING SNA ($\times 10^4$)

4.3.2 Verification of Variables and Validations of Results

All of the numbers that are shared in this exercise are the results of multiple meetings with key employees at Jawwal who confirm the validity of each of the model inputs (daily revenues, revenues uplift, active days, costs and number of members in the different obtained groups). Moreover, the

obtained ROIs are very similar to ROI obtained from similar designed offers. For the results obtained by incorporating SNA, management at Jawwal agrees that results will be very beneficial to improving the performance of offers over time.

For further clarification, the following provides a case extracted from the 10k sample of subs of which SNA was applied and another case of the 10k sample of which SNA was not applied. In both cases, the opt in rates were 100% among the leaders and their members. For the purpose of clarification, two members who actually subscribed were omitted to demonstrate the effect of applying SNA. From Table (4-10) below we can notice that even without these two members, SNA still makes a higher ROI compared to the model without SNA as in Table(4-9).

Table (4.9): Shaping offers for all group members individually (Singular view)

Member #	Tier	Daily Revenues	Cost	Offer Fees	Revenue Uplift	ROI
Member #1	Tier 1	2	0.09	2.5	0.41	456%
Member #2	Tier 2	2.5	0.09	3	0.41	456%
Member #3	Tier 3	3	0.09	3.5	0.41	456%
Member #4	Tier 4	3.5	0.09	4	0.41	456%
Member #5	Tier 5	4	0.09	4.5	0.41	456%
Leader #6	Tier 6	4.5	0.09	5	0.41	456%
Group Summary		**19.5**	**0.54**	**22.5**	**2.46**	**456%**

Tier: 0.5 NIS in difference; Revenue Uplift: (Fees - Cost - Daily ARPU); ROI: Revenue Uplift/ Daily ARPU (Old)

Table (4.10): Shaping only one offer for group leader and let the influence do the job (Network view)

Member #	Tier	Daily Revenues	Cost	Offer Fees	Revenue Uplift	ROI
Member #1	Tier 1	2	0	N/A	N/A	N/a
Member #2	Tier 2	2.5	0	N/A	N/A	N/a
Member #3	Tier 3	3	0	5	2	N/a
Member #4	Tier 4	3.5	0	5	1.5	N/a
Member #5	Tier 5	4	0	5	1	N/a
Leader #6	Tier 6	4.5	0.09	5	0.41	456%
Group Summary		**19.5**	**0.09**	**20**	**4.91**	**5456%**

Table (4-9) shows that we have 5 different offers customized for every individual in the network where table (4-10) includes only 1 offer targeted to the leader in which the latter will influence at least three members to subscribe. SNA saved our time in targeting, increased the ROI through reducing the cost of targeting other subscribers which was estimated to be 0.09 per sub.

Chapter Five

Discussion of Results

The previous four chapters discussed theoretical parts of the thesis along with methodology section which focused more on the drivers related to Social Network Analysis model in terms of their definitions, calculations and algorithms applied. This chapter will answer the research main questions mentioned in section 1.4 under research objectives section. Back to research main questions which were focused on identifying the influencers and influenced in terms of:

1- Identify influencers in Jawwal subscribers' base in terms of:
    A- Who, when they churn, would take few friends with them?
    C- Who, when they adopt, would push a few friends to do the same?
2- Identify subscribers whose loyalty is threatened by churn around them.
3- Calculate the Return on Investment (ROI) from offers given to subscribers based on their impact on the community (their degrees of influence).

A model of subscribers' network was built to determine the connections and the strength of those connections by which one can conclude who is influenced and who is influencing. With this map of the social community within the grasp, SNA can help telecom operators to make better marketing decisions. It can also help predict churn and retain customers by

understanding how influencers will cause a ripple effect through communities, acquire new customers through member get member campaigns, and help cross sell and up-sell with targeted viral marketing. Most importantly, this SNA model will allow telecom operators to change the way they target customers to include those with secured ROI which means less offering cost.

Since the cost of acquiring a new subscriber is high, it makes sense to start by retaining our existing customers. SNA uses community circles to pinpoint these “contagious churners”. These are subscribers who have not churned yet, but are being influenced by their friends and social group to churn. Once we have found the influencers we can target them with disproportionate benefits, after all it is better to target one and allow their influence to naturally do its job rather targeting everyone but the key here is who is that ‘one’.

The final result of SNA came up with very rich insights that need to be studied and analyzed by Jawwal. Those insights will help them understand their subscribers’ patterns in terms of number of communities, average community size, number of connections and how many influencers are there. The use of SNA results will also assist in defining the type of subscriber that Jawwal needs to target based on their KPIs. So if it was meant to introduce a new product in the market, then the strategy here is ‘to spread the word’ by using the dissemination factor. But if it was meant to acquire new members or to increase sales on a new product then the

strategy is ‘to target the influencers’ by using the authority factor. SNA will provide those two key players inside every community!

As an example, in a traditional approach which is used by most of companies through targeting its current customers, they will start looking at how much those customers generate revenues? How much do they utilize the traffic? When do they talk? And start to build an offer expecting uplift equal to 5%. However, on an individual level when incorporating SNA, their target will be more focused on a community view instead of individual which will have an impact of all of community members instead of one individual who might not be the influencer. Regardless if the KPI was set to reduce the churn, stretch the revenue or increase the base, the cost will be also focused on the community key players! Those who will for sure response and let other members do, those who we can expect an ROI of them and their affected followers.

Table (5-1) shows a summary of the previously selected sample in regards of total members (nodes), links (connections), groups (communities) and other group details.

Table (5.1): Summary Statistics

Name	Value
Total Nodes in Network	17,184
Total Links in Network	38,416
Total Number of Groups	1,862
Mean Group Size	9.23
Mean Group Density	0.31
Mean In-Degree	2.24
Mean Out-Degree	2.24

At a coverage threshold equal to 15% -comparing other thresholds mentioned previously in Table (3-2) and out of 24,000 remaining subscribers after the data preparation process that took a place in section (3.3.1)- SNA model indicated that there are 1,862 final groups for the dataset of size equal to 17,184 subscribers (the difference between 24,000 and 17,184 is due to coverage threshold, remember that the ties between group members control which members belong to the same groups and weak ties will be excluded), this is referred as "Total Number of Groups". "Total Links in Networks" (in-degree and out-degree) were 38,416 links with an average of 2.24 of links per group member. Finally, the "Average Group Size" was 9.23 (and again, remember that the minimum group size was set to equal 2 and the max to equal 42) with an "Average Group Density" equal to 0.31.

5.1 Who, when they churn, would take few friends with them?

A group of 9 members will have two influencers, the authority leader and the dissemination leader. If the authority leader has churned the rest of members will have no authority leader to ask for his opinions – due to his highest in-degree level and weight. So, they will simply follow him to a new place. Network operators are keen to sustain their subscribers' base the longest possible period of time. SNA will assist to understand their base structure and to avoid the churn of network communities key players which when they churn will take their friend with them. Jawwal should offer a

very aggressive offers for those influencers who are likely to churn in order to prevent them, and their community members, from doing so.

5.2 Who, when they adopt, would push a few friends to do the same?

On the other hand, if the dissemination leader has adopted to a new product and it's known that 'the word of mouth' he has – due to his highest out-degree- is a powerful tool, then the rest of network members are expected to adopt. Instead of targeting all base and paying a lot of money to 'spread the word' which will affect their campaigns or product release to the market, branding and awareness costs, operators can hire the network community dissemination leaders to do the job!

5.3 Who are Subscribers whose Loyalty is Threatened by Churn around them?

Finally, the rest of 7 members are the group followers that their loyalty will be threatened by the churn of the influencers. In their authority and dissemination scores, they will have lower scores compared to their leaders, they are flagged as '2-5' in their 'Type1' and 'Type2' of a group equal to five members as the below table of group 59 discussed in section 4.4.1

Table (5.2): Type 1 & Type 2 HITS Scores

GAI_NodeNumber	Member #1	Member #2	Member #3	Member #4	Member #5
GAI_RankOrderType1	5	1	2	3	4
GAI_RankOrderType2	2	5	3	4	1

5.4 What is the ROI percentage for Jawwal's subscribers based on their impact on the community?

It is clearly obvious by now that SNA would bring an added value to the company that will adopt it, mainly from the fact that is a new process that highly focuses on the influence that one person may have over his peers. This influencer acts as the alpha male in the pack and pushes his peers to follow his leads. In both the simulation and the provided case, it was demonstrated that SNA would in fact cause the base to increase the ROI through the influencer. As long as the spread between the influencer's and the members' revenues be high, the ROI would be high.

Back to section 3.2 of SNA related papers and to the case of Sonamine(2010), they were able to increase a churn prediction model efficiency in a European mobile operator by 25% through incorporating SNA authority influencers with the model, they were also able to increase the conversion of a product by 340% higher through incorporating SNA dissemination influencers inside a US telecom operator. In addition, Erik in his research hypothesis "The frequency with which a pupil smokes is positively correlated to the frequency of smoking among his/her friends." found that there is a high correlation between the smoking ego and the smoking networks who were affected by them. ROI using SNA can stretch revenues up to 50 times compared to traditional ways, chapter four section #6 discussed this point in deeper details.

As a conclusion gained from Table (5-1), 71% of Jawwal original network ties were ranked as strong ties that connected members together (17,184

members out of 24,000 members) compared to 68% in Vodafone Egypt – Dr. Ahmed said, a Data Mining Consultant in the company-. On the other hand, network density which is defined as actual connections divided by potential connections inside a group of members was 31%, this is an indicator for Jawwal that they need to enhance the density of their subscribers' groups as their current pattern in terms of connectivity was 31% between community members. However, this percentage is a bit higher in Vodafone Egypt to hit 48% (as also indicated by Dr. Ahmed) which means a potential to increase number of transactions between them through shaping more pre-studied and targeted offers, they will be able to predict the abusers of any offers (obviously, subscribers will utilize the offer to their closed circles and the way they connect). All comparison has been made between Jawwal & Vodafone Egypt due to the fact that researcher has access to a key person contact there in addition to the offers between both operators are likely similar.

Chapter Six

Conclusions and Recommendations

6.1 Conclusions

Dynamicity environment of telecommunication industry, high-level of competition and increased customers' expectations have made necessity of getting awareness of increasing offering mechanism, predicting influencers power overall other community members and utilizing subscribers pattern to understand network structures. Jawwal as a market leader in Palestine, should take advantage of methods and patterns consecutively with the aim of consecutive evaluation and improvement of their performance.

This study defines SNA as a method that helps operators to plan, monitor, analyse, and manage customers network more effectively by providing a network view for the decision makers. With a shared purpose, a consistent data model, real-time information, easy-to-use tools, and streamlined processes, it’s much simpler to align data driven insights with strategy that will support faster decisions and boost performance to achieve business goals.

SNA drives organizations to focus on the key drivers of value as they relate to corporate strategy and scientific offering and targeting processes. SNA provides fact-based guidance for value-based decision making.

This research has contributed to providing decision makers with a systematic approach for targeting operator’s customers base with a consideration of the involved relationships among its members. The

similarity algorithm used to analyse subscribers pattern of similarity and relationship weight as to group them together and eliminate weak relationships. Inside every group, there are two types of influencers used for managing other group members, the first is the authority leader who is defined as the reference for every decision other members take and the dissemination (hub) leader who is responsible for 'spread the word' to other group members. The two influencers should be used smartly by marketing owners or decision makers, the analysis of their strength and compare the traditional ways of targeting vs the new projected and scientific based model is also measurable.

A well-organized SNA model is constructed to facilitate the solving process. It is our belief that SNA has reached the compromise and will be useful for many other cases as it has been in the past. In particular, SNA has broken through the management community to be widely used by CEOs. This widespread use is certainly due to its ease of applicability and the structure of SNA that follows the intuitive way in which decision makers solve problems. SNA is based on DWH input of data and the use of SNA algorithms; the more accurate data and available we have is the more SNA model is reliable. It's the data mining models basis!

6.1.1 Managerial Insights

Based on this research, one can see that SNA will solve existing cases in Jawwal that were proposed in this research for: (1) Reducing churn through targeting community leader and focusing the investment of offers on him

instead of all community members. (2) For increasing product launch efficiency when introducing new product and testing it on a specific members- those who have the power to share the details- instead of random test. (3) For focusing on whole network ROI and network leader instead of other normal group members. At the end, it's better to target specific subscribers instead of targeting whole network.

6.2 Recommendation

In this research, it is recommended to run SNA model every three months as a starting point, then the marketing owner will have the decision either to stretch or shorten this period. It's based on how customer base is increasing or decreasing, the acquisition of new subscribers or the churn of the existing. It differs in a company with a stable and mature state of base and with no changes in the market related to network expansion or launching new products than a company with a challenging market. This could be a reason behind the success of SNA model as well.

6.3 Suggestion for Future Research

As stated before, the purpose of this research is to develop a data mining model that will assist telecom network operators to understand their subscribers' base pattern and the way they connect. It's moving from individual view to network view which will enhance the offering methodology, reduce churn and increase new product launch efficiency. The analysis has focused on the process and ways of mapping the result of SNA model with Jawwal's strategy.

Incorporating traffic from other destinations, like competitors, and focusing on how Jawwal subscribers interact with competitors' base will assist in understanding the strategy of the latest and could be a future research objective along with incorporating more demographic variables like gender. Through shaping a competitor SNA model, Jawwal will be able to measure competitors influence on its network along with their regional distribution which might help in predicting if, at someday, the competitors might acquire its base.

References

- Barry, W., Berkowitz, S. & eds., 1988. ***Social Structures: A Network Approach..*** **Cambridge, Cambridge University Press**.
- Cheliotis, G., 2010. ***Social-network-analysis***. [Online]
  Available at: http://www.slideshare.net/gcheliotis/social-network-analysis-3273045
  [Accessed 5 June 2016].
- Dasgupta, K. et al., 2008. ***Social ties and their relevance to churn in mobile telecom networks,*** **India: University of Maryland Baltimore County.**
- DataFormats, 2006. ***"DataFormats".*** [Online]
  Available at:
  http://netwiki.amath.unc.edu/DataFormats/Formats
  [Accessed 2 October 2015].
- De-Shuang Huang, Vitoantonio Bevilacqua, Juan Carlos Figueroa, Prashan Premaratne, 2013. ***Intelligent Computing Theories.*** **Nanning, China,** ICIC.
- Devi, P., Gupta, A. & Dixit, A., 2014. Comparative Study of HITS and PageRank Link. ***International Journal of Advanced Research in Computer and Communication Engineering,*** III(2), p. 6.
- Executive Information Systems, 2016. ***SAS GSA Price List.*** [Online]
  Available at:
  http://www.execinfosys.com/SAS%20GSA%20PriceList.pdf
  [Accessed 28 Februrary 2016].

- Feng, Y., 2007. ***Application of Data Mining in CRM System.*** s.l.:s.n.
- Fowler, H. & Christakis, A., 2008. **Dynamic spread of happiness in a large social network: longitudinal analysis over** 20. 5 December.
- Freeman, L., 2007. ***the Development of Social Network Analysis,*** **Vancouver:** s.n.
- Hansen, W. B. & Reese, E. L., 2009. ***Network Genie User Manual.*** [Online] Available at: https://secure.networkgenie.com/admin/documentation/Network_Genie_Manual.pdf [Accessed 20 September 2015].
- Harding, F., 2002. ***Cross-Selling Success.*** **s.l.:Adams Media.**
- Hogan, B., 2009. ***The Networked Individual: A Profile of Barry Wellman..*** [Online]
  Available at: http:/ /www.semioticon.com/semiotix/semiotix14/sem-14-05.html [Accessed 25 Aug 2015].
- IBM Corporation , 2012. ***Describing networks.*** [Online] Available at: http://www.ibm.com/support/knowledgecenter/api/content/nl/en-us/SS3RA7_15.0.0/com.ibm.spss.sna.doc/sna_overview_statistics.htm [Accessed 05 April 2016].
- IBM Corporation , 2012. ***Network density.*** [Online] Available at: http://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.sna.doc/sna_overview_statistics_density.htm [Accessed 04 April 2016].
- IBM Corporation, 2012. ***About social network analysis.*** [Online] Available at:

http://www.ibm.com/support/knowledgecenter/api/content/nl/en-us/SS3RA7_15.0.0/com.ibm.spss.sna.doc/sna_overview.htm [Accessed 05 April 2016].

- IBM Corporation, 2012. ***Describing groups and group members.*** [Online] Available at: http://www.ibm.com/support/knowledgecenter/api/content/nl/en-us/SS3RA7_15.0.0/com.ibm.spss.sna.doc/group_analysis_describe.htm [Accessed 05 April 2016].
- IBM Corporation, 2012. ***Nodal degree.*** [Online] Available at: http://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.sna.doc/sna_overview_statistics_degree.htm [Accessed 04 April 2016].
- IBM Corporation, 2012. ***Partitioning into groups.*** [Online] Available at: http://www.ibm.com/support/knowledgecenter/api/content/nl/en-us/SS3RA7_15.0.0/com.ibm.spss.sna.doc/group_analysis_partition.htm [Accessed 04 April 2016].
- IBM, 2016. ***IBM SPSS Modeler - Premium.*** [Online] Available at: https://www.ibm.com/marketplace/cloud/XConfigureProductView?catalogId=12301&langId=-1&storeId=18251&partNumber=SW-Modeler+Client+Premium&krypto=EBRbTD2Ug%2BR7enzq2srw6UzwVKsMQWNkpii1RK0LEhsJFyMS%2FGrpZXEt49fl0SjyOxicvqqogKRQu%2Bo9C35jpeR2nYaDxe6J3SoRFZJ3d [Accessed 19 April 2016].
- Ingen, E. V., 2013. ***Social Network Analysis.*** [Online] Available at: http://erikvaningen.nl/?page_id=47 [Accessed 01 March 2016].

- International Telecommunication Union, 2010. ***World Telecommunication/ICT Development Report,*** **Geneva Switzerland: Place des Nations.**
- Jawwal, 2015. ***Jawwal Ltd..*** [Online] Available at: http://www.jawwal.ps/index.php?page=section&pid=1214&section_parent=1208&catid=3&langid=2 [Accessed 01 09 2015].
- Krebs, V., 2013. ***Social Network Analysis: An Introduction.*** [Online] Available at: http://www.orgnet.com/sna.html [Accessed 15 February 2016].
- Manning, C. D., Raghavan, P. & Schütze, H., 2008. ***Hubs and Authorities.*** [Online] Available at: http://nlp.stanford.edu/IR-book/html/htmledition/hubs-and-authorities-1.html [Accessed 17 Jun 2016].
- Manning, C. D., Raghavan, P. & Schütze, H., 2008. Link analysis. In: ***Introduction to Information Retrieval.*** **Cambridge: Cambridge University Press,** pp. 461-481.
- Maryland, U. o., 2009. ***Lec10-modularity.*** [Online] Available at: https://www.cs.umd.edu/class/fall2009/cmsc858l/lecs/Lec10 modularity.pdf[Accessed 12 August 2016].
- Moreno, J. L., 1934. Who Shall Survive?. In: ***Foundations of Sociometry, Group Psychotherapy, and Sociodrama.*** **DC: Nervous and Mental Disease Publishing Co.**
- Newman, M. E. J., 2004. ***Detecting community structure in networks,*** **Michigan: University of Michigan.**

- Oracle, 2016. ***Database SQL Reference.*** **[**Online] Available at: https://docs.oracle.com/cd/B19306_01/server.102/b14200/statements_4010.htm [Accessed 19 March 2016].
- Oracle, 2016. ***ROWNUM Pseudocolumn.*** [Online] Available at: https://docs.oracle.com/cd/B19306_01/server.102/b14200/pseudocolumns009.htm [Accessed 6 March 2016].
- Otte, E. & Rousseau, R., 2002. ***a powerful strategy, also for the information sciences.*** **In: Social network analysis.** s.l.:s.n., pp. 441453.
- Passmore, D. L., 2011. ***Social network analysis: Theory and applications.*** s.l.:s.n.
- Penheiro, C. A. R., 2011. ***Social Network Analysis in Telecommunications.*** **Hoboken,** New Jersey: John Wiley & Sons Inc..
- Richter, Y., Yom-Tov, E. & Slonim, N., 2010. *Predicting customer churn in mobile networks through the analysis of social groups.* s.l., s.n., p. 732–741.
- Rosenblatt, G., 2014. ***NetworkDensity.*** [Online] Available at: http://www.the-vital-edge.com/what-is-network-density/ [Accessed 04 April 2016].
- Rouse, M., 2005. ***macro.*** [Online] Available at: (http://whatis.techtarget.com/definition/macro) [Accessed 10 08 2015].
- Rouse, M., 2014. ***Business Intelligence (BI).*** [Online] Available at: http://searchdatamanagement.techtarget.com/definition/business-intelligence[Accessed 10 February 2017].

- SAS, 2015. ***History of SAS.*** [Online] Available at: https://www.sas.com/en_ae/company-information.html#history [Accessed 2 March 2016].
- Scott, J., 1991. ***Social Network Analysis.*** London: Sage.
- Sobek, M., 2003. ***The PageRank Algorithm.*** [Online] Available at: http://pr.efactory.de/e-pagerank-algorithm.shtml [Accessed 30 July 2016].
- Sonamine, 2010. ***Whitepapers.*** [Online] Available at: http://www.sonamine.com/home/index.php?option=com_phocadownload&view=category&id=2:whitepapers&download=5:social-network-analysis-for-telecommunications-marketing&Itemid=1
- Sonamine, 2010. ***Whitepapers.*** [Online] Available at: http://goo.gl/5nHexg
- SPSS Inc., 2000. ***CRISP-DM 1.0.*** s.l.:SPSS.
- SPSS, I., 2013. ***IBM SPSS Modeller.*** [Online] Available at: http://www-01.ibm.com/software/analytics/spss/products/modeler/ [Accessed 2 October 2015].
- Stanford, 2008. ***Hubs and Authorities.*** [Online] Available at: http://nlp.stanford.edu/IR-book/html/htmledition/hubs-and-authorities-1.html[Accessed 24 Jun 2016].
- Tong, H., Faloutsos, C. & Pan, J.-y., 2006. ***Random Walk with Restart and Its Applications.*** **Pittsburgh, Carnegie Mellon University.**

- TopPredictiveSoftware, 2014. ***top-predictive-analytics-software.*** [Online] Available at: http://www.predictiveanalyticstoday.com/top-predictive-analytics-software/[Accessed 2 October 2015].
- Vodafone, 2013. ***introduce-a-friend.*** [Online] Available at: http://www.vodafone.co.uk/terms-and conditions/consumer/mobile/other/previous-terms-and conditions/vodafone-introduce-a-friend/ [Accessed 27 February 2016].
- Wasserman, S. & Faust, K., 1994. ***Social Network Analysis: Methods and Applications.*** **Cambridge , Cambridge University Press.**
- Wellman, B., Chen, W. & Weizhen, D., 2002. ***Networking Guanxi.*** **Cambridge, Cambridge University Press,** p. 221–41.
- Wellman, B. & Marin, A., Jun 2009. **An introduction. In:** ***Social Network Analysis.*** s.l.:s.n.
- Whitler, K. A., 2014. ***Why Word Of Mouth Marketing Is The Most Important Social Media.*** [Online] Available at: http://www.forbes.com/sites/kimberlywhitler/2014/07/17/why-word-of-mouth-marketing-is-the-most-important-social-media/#356d13cd7a77 [Accessed 10 February 2017].

- Barry, W., Berkowitz, S. & eds., 1988. ***Social Structures: A Network Approach..*** **Cambridge, Cambridge University Press.**
- Cheliotis, G., 2010. *Social-network-analysis.* [Online] Available at: http://www.slideshare.net/gcheliotis/social-network-analysis-3273045 [Accessed 5 June 2016].

- Dasgupta, K. et al., 2008. ***Social ties and their relevance to churn in mobile telecom networks,*** **India: University of Maryland Baltimore County.**
- DataFormats, 2006. ***"DataFormats".*** [Online] Available at: http://netwiki.amath.unc.edu/DataFormats/Formats [Accessed 2 October 2015].
- De-Shuang Huang, Vitoantonio Bevilacqua, Juan Carlos Figueroa, Prashan Premaratne, 2013. ***Intelligent Computing Theories.*** Nanning, China, ICIC.
- Devi, P., Gupta, A. & Dixit, A., 2014. Comparative Study of HITS and PageRank Link. ***International Journal of Advanced Research in Computer and Communication Engineering,*** III(2), p. 6.
- Executive Information Systems, 2016. ***SAS GSA Price List.*** [Online] Available at:
- http://www.execinfosys.com/SAS%20GSA%20PriceList.pdf [Accessed 28 Februrary 2016].
- Feng, Y., 2007. ***Application of Data Mining in CRM System.*** s.l.:s.n.
- Fowler, H. & Christakis, A., 2008. **Dynamic spread of happiness in a large social network: longitudinal analysis over** 20. 5 December.
- Freeman, L., 2007. ***the Development of Social Network Analysis,*** **Vancouver:** s.n.
- Hansen, W. B. & Reese, E. L., 2009. ***Network Genie User Manual.*** [Online] Available at:

- https://secure.networkgenie.com/admin/documentation/Network_Genie_Manual.pdf [Accessed 20 September 2015].
- Harding, F., 2002. ***Cross-Selling Success.*** **s.l.:Adams Media.**
- Hogan, B., 2009. ***The Networked Individual: A Profile of Barry Wellman..*** [Online] Available at: http:/ /www.semioticon.com/ semiotix/semiotix14/sem-14-05.html [Accessed 25 Aug 2015].
- IBM Corporation , 2012. ***Describing networks.*** [Online] Available at: http://www.ibm.com/support/knowledgecenter/api/content/nl/en-us/SS3RA7_15.0.0/com.ibm.spss.sna.doc/sna_overview_statistics.htm [Accessed 05 April 2016].
- IBM Corporation , 2012. ***Network density.*** [Online] Available at:http://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.sna.doc/sna_overview_statistics_density.htm[Accessed 04 April 2016].
- IBM Corporation, 2012. ***About social network analysis.*** [Online] Available at:http://www.ibm.com/support /knowledgecenter/ api/content /nl/en-us/SS3RA7_15.0.0/com.ibm.spss.sna.doc/sna_overview.htm [Accessed 05 April 2016].
- IBM Corporation, 2012. ***Describing groups and group members.*** [Online] Available at:http://www.ibm.com/support/knowledgecenter/api/content/nl/enus/SS3RA7_15.0.0/com.ibm.spss.sna.doc/group_analysis_describe.htm[Accessed 05 April 2016].
- IBM Corporation, 2012. ***Nodal degree.*** [Online] Available at: http://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm

.spss.sna.doc/sna_overview_statistics_degree.htm[Accessed 04 April 2016].

- IBM Corporation, 2012. ***Partitioning into groups.*** [Online] Available at: http://www.ibm.com/support/knowledgecenter/api/content/nl/en-us/SS3RA7_15.0.0/com.ibm.spss.sna.doc/group_analysis_partition.htm [Accessed 04 April 2016].
- IBM, 2016. ***IBM SPSS Modeler - Premium.*** [Online] Available at: https://www.ibm.com/marketplace/cloud/XConfigureProductView?catalogId=12301&langId=-1&storeId=18251&partNumber=SW-Modeler+Client+Premium&krypto=EBRbTD2Ug%2BR7enzq2srw6UzwVKsMQWNkpii1RK0LEhsJFyMS%2FGrpZXEt49fl0SjyOxicvqqogKRQu%2Bo9C35jpeR2nYaDxe6J3SoRFZJ3d[Accessed 19 April 2016].
- Ingen, E. V., 2013. ***Social Network Analysis.*** [Online] Available at: http://erikvaningen.nl/?page_id=47 [Accessed 01 March 2016].
- International Telecommunication Union, 2010. ***World Telecommunication/ICT Development Report,*** **Geneva Switzerland: Place des Nations.**
- Jawwal, 2015. ***Jawwal Ltd..*** [Online] Available at: http://www.jawwal.ps/index.php?page=section&pid=1214&section_parent=1208&catid=3&langid=2[Accessed 01 09 2015].
- Krebs, V., 2013. ***Social Network Analysis: An Introduction.*** [Online] Available at: http://www.orgnet.com/sna.html [Accessed 15 February 2016]

- Manning, C. D., Raghavan, P. & Schütze, H., 2008. *Hubs and **Authorities.*** [Online] Available at: http://nlp.stanford.edu/IR-book/html/htmledition/hubs-and-authorities-1.html [Accessed 17 Jun 2016].
- Manning, C. D., Raghavan, P. & Schütze, H., 2008. Link analysis. In: ***Introduction to Information Retrieval.*** **Cambridge: Cambridge University Press,** pp. 461-481.
- Maryland, U. o., 2009. ***Lec10-modularity.*** [Online] Available at: https://www.cs.umd.edu/class/fall2009/cmsc858l/lecs/Lec10-modularity.pdf [Accessed 12 August 2016].
- Moreno, J. L., 1934. **Who Shall Survive?. In:** ***Foundations of Sociometry, Group Psychotherapy, and Sociodrama.*** **DC: Nervous and Mental Disease Publishing Co.**
- Newman, M. E. J., 2004. ***Detecting community structure in networks,*** **Michigan: University of Michigan.**
- Oracle, 2016. ***Database SQL Reference.*** [Online] Available at: https://docs.oracle.com/cd/B19306_01/server.102/b14200/statements_4010.htm [Accessed 19 March 2016].

- Oracle, 2016. ***ROWNUM Pseudocolumn.*** [Online] Available at: https://docs.oracle.com/cd/B19306_01/server.102/b14200/pseudocolumns009.htm [Accessed 6 March 2016].
- Otte, E. & Rousseau, R., 2002. **a powerful strategy, also for the information sciences. In:** ***Social network analysis.*** s.l.:s.n., pp. 441-453.
- Passmore, D. L., 2011. ***Social network analysis: Theory and applications.*** s.l.:s.n.
- Penheiro, C. A. R., 2011. ***Social Network Analysis in Telecommunications.*** **Hoboken,** New Jersey: John Wiley & Sons Inc..
- Richter, Y., Yom-Tov, E. & Slonim, N., 2010. ***Predicting customer churn in mobile networks through the analysis of social groups.*** s.l., s.n., p. 732–741.
- Rosenblatt, G., 2014. ***NetworkDensity.*** [Online] Available at: http://www.the-vital-edge.com/what-is-network-density/ [Accessed 04 April 2016].
- Rouse, M., 2005. ***macro.*** [Online] Available at: (http://whatis.techtarget.com/definition/macro) [Accessed 10 08 2015].
- Rouse, M., 2014. ***Business Intelligence (BI).*** [Online] Available at: http://searchdatamanagement.techtarget.com/definition/business-intelligence[Accessed 10 February 2017].

- SAS, 2015. ***History of SAS.*** [Online] Available at: https://www.sas.com/en_ae/company-information.html#history [Accessed 2 March 2016].
- Scott, J., 1991. ***Social Network Analysis.*** London: Sage.
- Sobek, M., 2003. ***The PageRank Algorithm.*** [Online] Available at: http://pr.efactory.de/e-pagerank-algorithm.shtml[Accessed 30 July 2016].
- Sonamine, 2010. ***Whitepapers.*** [Online] Available at: http://www.sonamine.com/home/index.php?option=com_phocadownload&view=category&id=2:whitepapers&download=5:social-network analysis-for-telecommunications-marketing&Itemid=1
- Sonamine, 2010. ***Whitepapers.*** [Online] Available at: http://goo.gl/5nHexg
- SPSS Inc., 2000. *CRISP-DM 1.0.* s.l.:SPSS.
- SPSS, I., 2013. ***IBM SPSS Modeller.*** [Online] Available at: http://www-01.ibm.com/software/analytics/spss/products/modeler/[Accessed 2 October 2015].
- Stanford, 2008. ***Hubs and Authorities.*** [Online] Available at: http://nlp.stanford.edu/IR-book/html/htmledition/hubs-and authorities-1.html[Accessed 24 Jun 2016].
- Tong, H., Faloutsos, C. & Pan, J.-y., 2006. ***Random Walk with Restart and Its Applications.*** Pittsburgh, Carnegie Mellon University.

- TopPredictiveSoftware, 2014. ***top-predictive-analytics-software.*** [Online] Available at: http://www.predictiveanalyticstoday.com/top-predictive-analytics-software/[Accessed 2 October 2015].
- Vodafone, 2013. ***introduce-a-friend.*** [Online] Available at: http://www.vodafone.co.uk/terms-and conditions/consumer/mobile/other/previous-terms-and conditions/vodafone-introduce-a-friend/ [Accessed 27 February 2016].
- Wasserman, S. & Faust, K., 1994. ***Social Network Analysis: Methods and Applications.*** **Cambridge , Cambridge University Press.**
- Wellman, B., Chen, W. & Weizhen, D., 2002. ***Networking Guanxi.*** **Cambridge, Cambridge University Press**, p. 221–41.
- Wellman, B. & Marin, A., Jun 2009. ***An introduction.*** **In: Social Network Analysis.** s.l.:s.n.
- Whitler, K. A., 2014. ***Why Word Of Mouth Marketing Is The Most Important Social Media.*** [Online] Available at: http://www.forbes.com/sites/kimberlywhitler/2014/07/17/why-word-of-mouth-marketing-is-the-most-important-social-media/#356d13cd7a77 [Accessed 10 February 2017].

Appendices

Appendix 1: SQL scripts to extract CDRs and Profile Data for the selected region

1. **Extract the aggregated CDRs data from the data warehouse:**

```
CREATE TABLE sna_sample_25000_aggr_cdrs PARALLEL AS
SELECT *
FROM  (
        SELECT  a_number,
            b_number,
            call_type,
            Avg(call_duration) Avg_Du,
            Avg(calls_count) Avg_Calls,
            Avg(call_duration / calls_count) ,          from    dwh_subscribers_fac
t
        where   call_date BETWEEN
            '01-Jan-  2016' AND '31-Mar-2016'
        AND     a_number_cell_id BETWEEN 80 AND 93          GROUP BY a_
number,
            b_number,
            call_type
        ORDER BY dbms_random.value)
WHERE  ROWNUM <= 25000;
```

dbms_random.value is an oracle built in package that provides a built in random number generator. Rownum is a pseudocolumn returns a number indicating the order in

which Oracle selects the row from a table or set of joined rows. The first row selected has a ROWNUM of 1, the second has 2, and so on. (Oracle, 2016)

2. **Extract the profile data for the previously extracted subscribers:**

```
CREATE TABLE sna_sample_25000_profile
PARALLEL AS
 SELECT a.a_number,
      a.tenure,
      a.connection_type,
      a.region /*for the selected 14 cells id*/
 FROM   dwh_subscribers_active_profile a
 WHERE  a.a_number IN (SELECT a_number
                  FROM   sna_sample_25000_aggr_cdrs
                    GROUP  BY a_number)
 GROUP  BY a.a_number,
      a.tenure,
      a.connection_type,
      a.region;
```

The statement in the ‘where’ clause is classified as nested statement. This statement adds another layer of filtration of the data to insure that the output is same as previously generated sample.

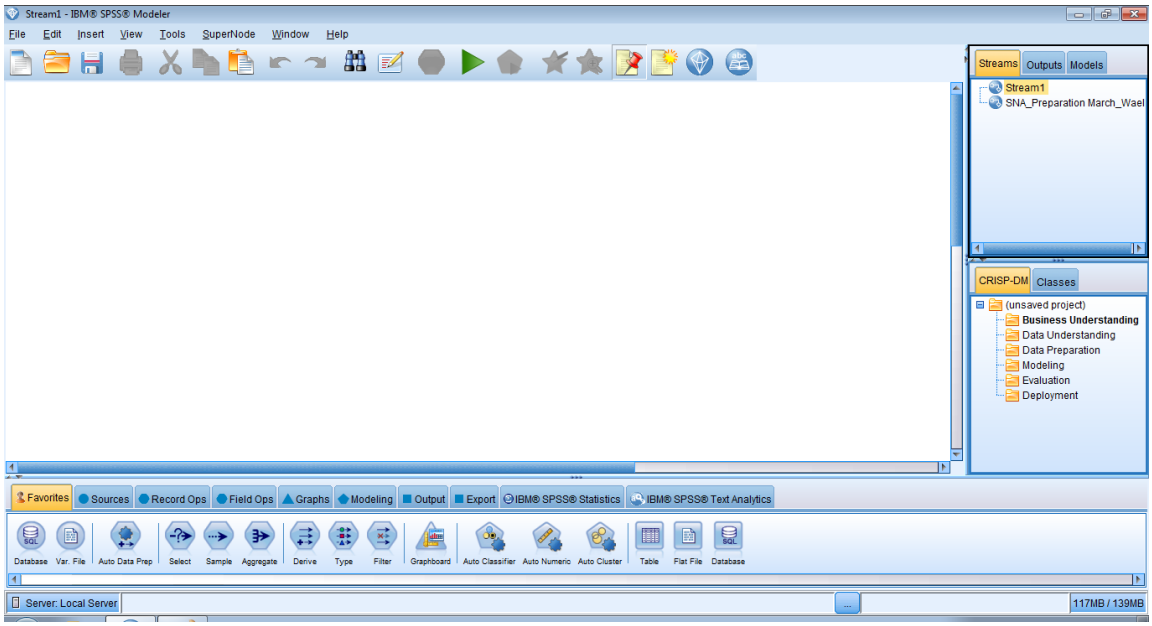
3. **SNA Base Duration Percentile:**

```
SELECT Avg(avg_du)            "Mean",
   Percentile_cont(0.01)
    within GROUP (ORDER BY avg_du) over (
```

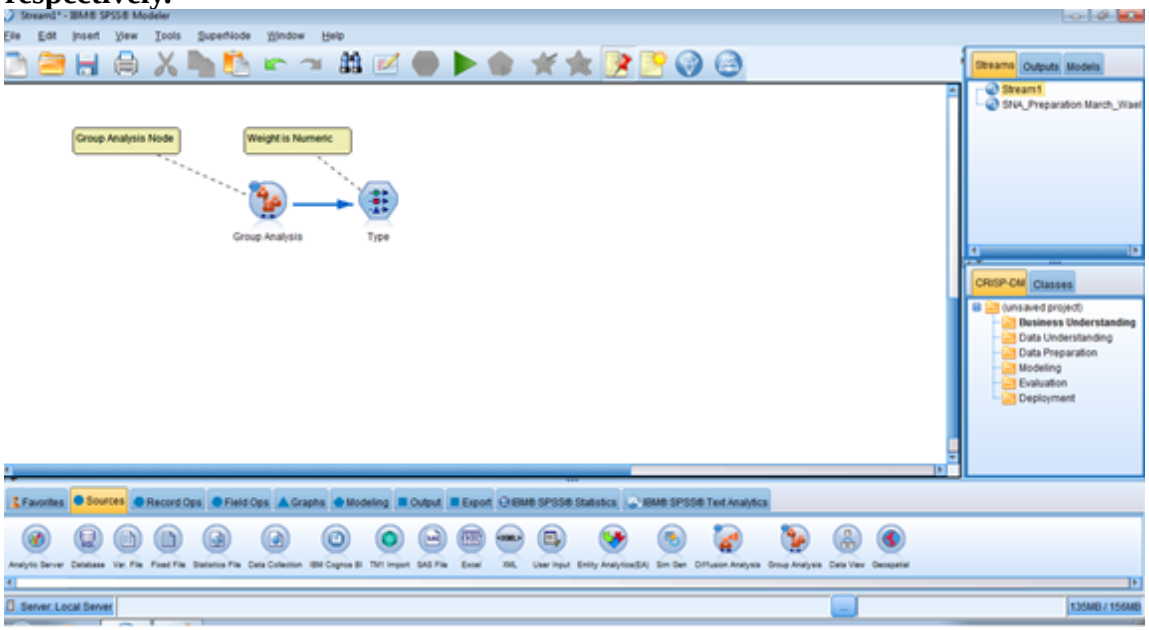
```
      PARTITION BY a_number) "1%",
    Percentile_cont(0.10)
     within GROUP (ORDER BY avg_du) over (
      PARTITION BY a_number) "10%",
    Percentile_cont(0.25)
     within GROUP (ORDER BY avg_du) over (
      PARTITION BY a_number) "25%",
    Percentile_cont(0.50)
     within GROUP (ORDER BY avg_du) over (
      PARTITION BY a_number) "50%",
    Percentile_cont(0.75)
     within GROUP (ORDER BY avg_du) over (
      PARTITION BY a_number) "75%",
    Percentile_cont(0.90)
     within GROUP (ORDER BY avg_du) over (
      PARTITION BY a_number) "90%",
    Percentile_cont(0.95)
     within GROUP (ORDER BY avg_du) over (
      PARTITION BY a_number) "95%",
    Percentile_cont(0.99)
     within GROUP (ORDER BY avg_du) over (
      PARTITION BY a_number) "99%"
FROM   sna_sample_25000_aggr_cdrs;
```

Appendix 2: IBM SPSS Modeler Tool

1. IBM Modeler environment 'New Stream'.

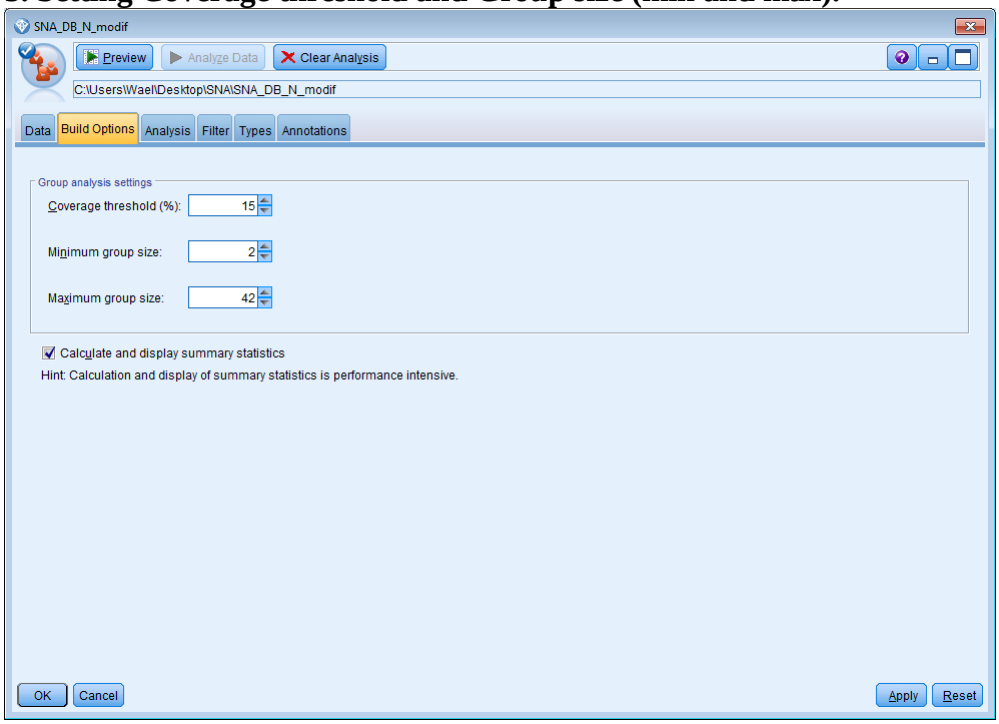


2. Importing SNA model nodes 'Group Analysis & Type Nodes' and establishing connections (arrow connections). Sections are 'Source' and 'Field Ops' respectively.

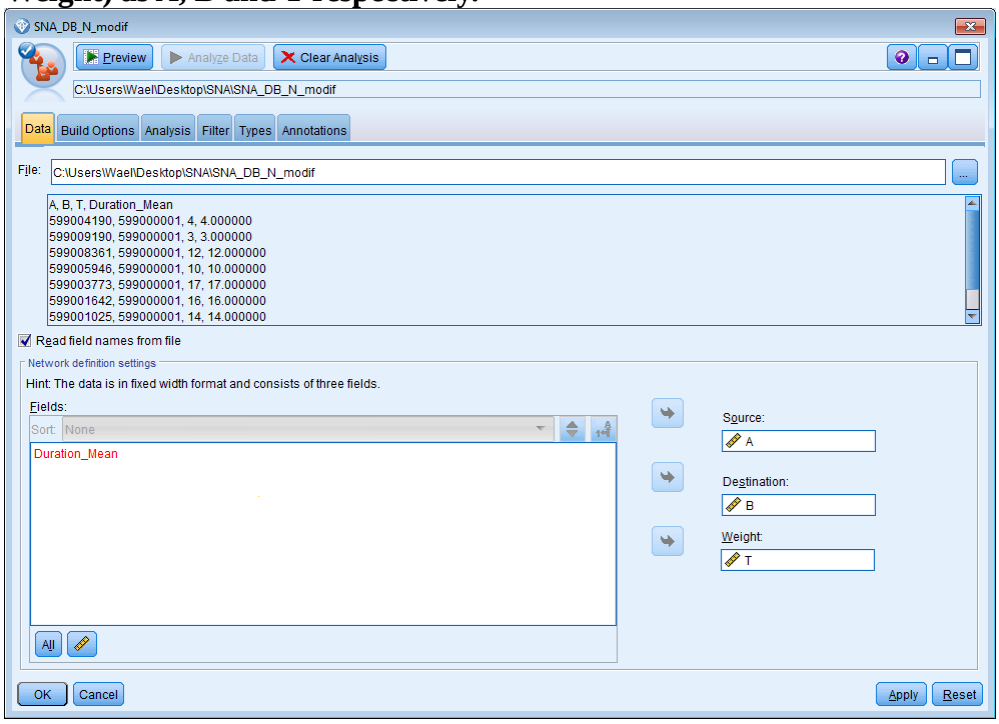


- Dataset import is embedded into Group analysis node.

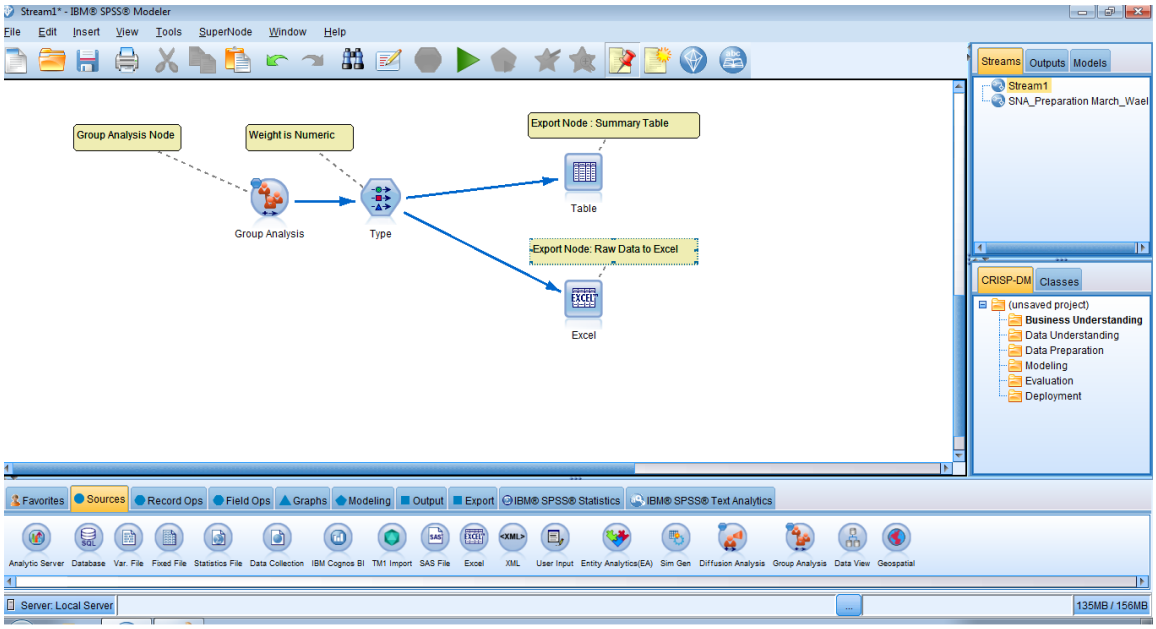
3. Setting Coverage threshold and Group size (min and max).



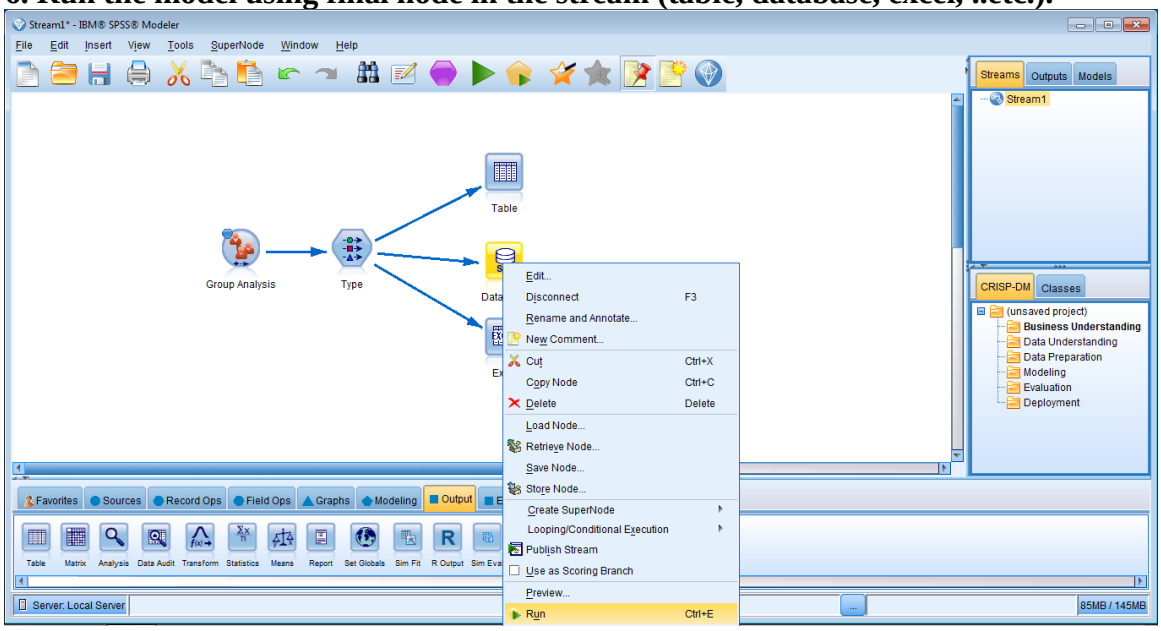
4. Setting the configuration of relationship operators (Originators, Receivers and Weight) as A, B and T respectively.



5. Exporting the final results using 'Export Node' either as a summary table or a raw data. Section is Export.



6. Run the model using final node in the stream (table, database, excel, ..etc.).



Appendix 3: HITS Algorithm Iterations for Authority and Hub Scores

1. Configuring the matrices:

A: Adjacency Matrix (i-->J)

Node No	1	2	3	4
1	0	7	1	0
2	0	0	12	7
3	0	0	0	0
4	0	0	4	0

A[T]

Node No	1	2	3	4
1	0	0	0	0
2	7	0	0	0
3	1	12	0	4
4	0	7	0	0

U: Initiation of Y(0)

Node No	1	2	3	4
Initialization	1	1	1	1

2. Iterations Equations:

Authority (a):	**Hub (h):**
$X_{(1)} = A^T * Y_{(0)}$	$Y_{(1)} = A * X_{(1)}$
$X_{(2)} = A^T * Y_{(1)}$	$Y_{(2)} = A * X_{(2)}$
$X_{(3)} = A^T * Y_{(2)}$	$Y_{(3)} = A * X_{(3)}$
$X_{(4)} = A^T * Y_{(3)}$	$Y_{(4)} = A * X_{(4)}$
..	..
..	..
$X_{(i)} = A^T * Y_{(i-1)}$	$Y_{(j)} = A * X_{(j)}$

3. Trials (1-10):

X(1)	X(1) Bar
0	-
7	0.356
17	0.864
7	0.356

Y(1)	Y(1) Bar
66	0.244
253	0.936
-	-
68	0.252

X(2)	X(2) Bar
-	-
462	0.120
3,374	0.879
1,771	0.461

Y(2)	Y(2) Bar
6,608	0.120
52,885	0.962
-	-
13,496	0.245

X(3)	X(3) Bar
-	-
46,256	0.059
695,212	0.881
370,195	0.469

Y(3)	Y(3) Bar
1,019,004	0.090
10,933,909	0.965
-	-
2,780,848	0.245

X(4)	X(4) Bar
-	-
7,133,028	0.044
143,349,304	0.881
76,537,363	0.471

Y(4)	Y(4) Bar
1,019,004	0.090
10,933,909	0.965
-	-
2,780,848	0.245

X(5)	
-	X(5) Bar
7,133,028	-
143,349,304	0.044
76,537,363	0.881

Y(5)	Y(5) Bar
193,280,500	0.083
2,255,953,189	0.966
-	-
573,397,216	0.245

X(6)	X(6) Bar
-	-
1,352,963,500	0.040
29,558,307,632	0.881
15,791,672,323	0.471

Y(6)	Y(6) Bar
39,029,052,132	0.081
465,241,397,845	0.966
-	-
118,233,230,528	0.245

X(7)	X(7) Bar
-	-
273,203,364,924	0.040
6,094,858,748,384	0.881
3,256,689,784,915	0.471

Y(7)	Y(7) Bar
8,007,282,302,852	0.081
95,935,133,475,013	0.966
-	-
24,379,434,993,536	0.245

X(8)	X(8) Bar
-	-
56,050,976,119,964	0.039
1,256,746,623,977,150	0.881
671,545,934,325,091	0.471

Y(8)	Y(8) Bar
1,649,103,456,816,900	0.081
19,781,781,028,001,500	0.966
-	-
5,026,986,495,908,610	0.245

X(9)	X(9) Bar
-	-

Y(9)	Y(9) Bar
339,944,491,160,497,000	0.081

11,543,724,197,718,300	0.039	4,078,968,331,689,700,000	0.966
259,138,421,776,469,000	0.881	-	-
138,472,467,196,010,000	0.471	1,036,553,687,105,880,000	0.245

X(10)	X(10) Bar	Y(10)	Y(10) Bar
-	-	70,091,059,286,724,700,000	0.081
2,379,611,438,123,480,000	0.039	841,074,798,891,120,000,000	0.966
53,433,779,219,860,400,000	0.881	-	-
28,552,778,321,827,900,000	0.471	213,735,116,879,441,000,000	0.245

Appendix 4: Matlab Script to apply Monte-Carlo Simulation

1. Without incorporating SNA

```
%% No SNA analysis
for i = 1:10000
% number of members
x = randi(41,1,1) + 1  ; % group size from 2 to 42
daily_revenue = (randi(18,1,x) + 2 ) ./ 10 ; %daily revenue from $0.3 to $2.0
d_v = (rand(1,x)) ;   % dummy variable
cost = (d_v <= 0.3) .* 0.05 + 0.05 ; % offer cost is 70% 0.05 or and 30% 0.1
take_offer = rand(1,x) <= 0.1 ;  % percentage of taking the offer
uplift=take_offer.*
    ((randi(11,1,x)+14).*0.01).*daily_revenue.*(randi(12,1,x)+0).*(randi(10,1,1)+0)
    ;
% the equation takes into account if a customer took the offer then it
% calculate how much the company profitted from the offer between 15 and
% 25%
total_uplift = sum(uplift) ; % to caluclate the total ROI for the group we
% need to sum the uplift values
total_cost = sum(cost);  % we also calculate the cost, then we use the ROI formula
ROI12(:,i) = (total_uplift-total_cost)./total_cost *100;
all_uplift_1(:,i)= total_uplift;
all_cost_1(:,i) = total_cost;
end
```

```
ROI13=mean(ROI12)

hist(ROI12)

meanROI15=mean(ROI12)

 stdevROI15=std(ROI12)
```

2. With SNA

```
%% with SNA analysis

for i = 1:10000

% number of members

x = randi(41,1,1)+1  ;% group size from 2 to 42

daily_revenue = (randi(18,1,x)+2) ./ 10 ; % daily revenue from $0.3 to $2.0

d_v = (rand(1,x)) ;   % dummy variable

cost = (d_v <= 0.3) .* 0.05+ 0.05 ; % offer cost is 70% 0.2 or and 30% 0.3

cost(1,[3:end])=0;

take_offer = rand(1,2) < 0.3 ;

for j = 3:x

if take_offer(1,1) == 1 || take_offer(1,2)==1

   take_offer(1,j) = rand(1,1)<0.3;

else take_offer(1,j) = rand(1,1)<0.1 ;

end

end

% if the one of the group leaders (the highest authority or dissemination

% score) opted in, the remaining of the group would opt in, otherwise, the

% probability of opting in is as without SNA which is 7%

uplift = take_offer.*
```

```
((randi(11,1,x)+14).*0.01).*daily_revenue.*(randi(10,1,1)+0) ;
% the equation takes into account if a customer took the offer then it
% calculate how much the company profitted from the offer between 15 and
% 25%
total_uplift = sum(uplift) ; % to caluclate the total ROI for the group we
% need to sum the uplift values
total_cost = sum(cost);  % we also calculate the cost, then we use the ROI formula
ROI14(:,i) = (total_uplift-total_cost)./total_cost *100;
all_uplift_2(:,i)= total_uplift;
all_cost_2(:,i) = total_cost;
end
hist(ROI14)
meanROI15=mean(ROI14)
 stdevROI15=std(ROI14)
```

جامعة النجاح الوطنية
كلية الدراسات العليا

تحليل أثر الشبكة الإجتماعية في قطاع الإتصالات في فلسطين – شركة جوال

إعداد الطالب
وائل خليل أبورزق

اشراف
د .رامز عساف

قدمت هذه الأطروحة استكمالاً لمتطلبات الحصول على درجة الماجستير في الإدارة الهندسية بكلية الدراسات العليا في جامعة النجاح الوطنية في نابلس، فلسطين.

2017

تحليل أثر الشبكة الإجتماعية في قطاع الإتصالات في فلسطين – شركة جوال

إعداد الطالب

وائل خليل أبورزق

اشراف

د .رامز عساف

الملخص

يرتكز هذا البحث على "تحليل واستكشاف المجموعات داخل الشبكة" وذلك عن طريق استنباط الأشخاص ذوي التأثير الكبير في كل مجموعة وتحديدهم؛ للمساهمة في استهدافهم في عروض وحملات أو المحافظة عليهم نظرا لتأثيرهم الكبير على عناصر المجموعة الأخرى، هذه المجموعة قد تكون العائلة، أصدقاء العمل، أصدقاء الدراسة وغيرهم.

وبما أن شركة الاتصالات الخلوية الفلسطينية –جوال هي شركة رائدة في سوق الاتصالات الخلوية في فلسطين فإنه من الضروري أن تمتلك نموذجا علميا لدراسة مشتركيها وتحديد أولئك الذين يملكون السلطة والسيطرة على أفراد مجموعتهم المغلقة.

ويهدف البحث بشكل اساسي إلى الإجابة على ثلاث أسئلة رئيسية وهي:

1. تحديد الأشخاص أصحاب القرار والتأثير على أفراد المجموعة الآخرين من حيث إقناعهم في البقاء في الشبكة أو تقبل المنتج الجديد .
2. تحديد المشتركين الذين يتأثرون بخروج من حولهم من المجموعة أو الشبكة مما يؤثر على ولائهم وايراداتهم للشبكة.
3. حساب المعدل العائد للاستثمار في حملات وعروض الشركة للمشتركين.

وبما ان استخدام المعادلات الحسابية وعمليات تحليل البيانات يعتبر عامل اساسي لحل المشاكل التقنية من جهة والمشاكل المتعلقة في قطاع الأعمال من جهة أخرى، اضافة الى الانتقال من التحليل الفردي إلى تحليل المجموعة وكيفية اتخاذ قرارات شمولية، يقدم هذا البحث نموذج مبني على تحليل درجات التشابه (السلطة والالتقاء) بين عناصر المجموعة المستخدمة في تحليل الشبكة الاجتماعية (SNA)؛ بحيث يتم استخراج قائد المجموعة والشخصية الأكثر تأثيرا فيها.

إن بناء هذا النظام قد اعتمد بشكل أساسي على مستودع البيانات الموجودة في الشركة وكيفية استغلاله واستنباط الفرضيات من هذه البيانات بعد مقارنتها وتدقيقها والتأكد من خلوها من القيم الشاذة مما يضمن كفاءة النظام واستمراريته.

واعتمدت الدراسة على استخدام الفرضيات العلمية والإحصائية التي ساهمت بشكل كبير في بناء أسس وروابط بين البيانات ، اضافة الى استخراج المعلومات وتحليلها مما يضمن ناتجا مبني على أسس علمية دقيقة، وإن تطبيق هذا النموذج من المتوقع أن يساعد الشركات في فهم أنماط وصفات مشتركيها إضافة إلى طرق التواصل فيما بينهم (استقبالا أو إرسالا) باستخدام برنامج IBM SPSS Modeler لتحليل البيانات و Oracle لاستخراج البيانات اضافة الى Matlab لمحاكاة العائد على الاستثمار عند تطبيق النموذج والمقارنة بين عدم تطبيقه باستخدام مونتي-كارلو.

واستنتجت هذه الدراسة أن الاعتماد على "تحليل الشبكة" هو أحد الطرق الفعالة لتغيير طريقة التفكير من الطرق التقليدية في استهداف جميع مشتركي المجموعة بعروض فردية لأحدث الطرق الفعالة في استهداف أشخاص محددين لديهم القدرة على إقناع غيرهم بالاشتراك او البقاء ضمن مجموعتهم وضمن الشركة، وتركز هذه الدراسة على أهمية تطبيق نموذج تحليل المشتركين للوصول إلى الأشخاص المؤثرين في المجموعة بحيث يتم استهدافهم والسعي نحو الحصول على أقصى درجات الرضى لهم مما يضمن بقاءهم لفترة أطول، وقد تم تقسيم الأشخاص المؤثرين إلى قسمين رئيسيين:

الأول: وهو المرجع الذي يقوم أفراد المجموعة بالرجوع إليه لمساعدتهم على اتخاذ القرارات

الثاني: وهو الشخص الذي يقوم بنشر الكلمة بين أفراد المجموعة بأقصى سرعة ووقت ممكن وهذا يعني أيضا أرباحا أكثر بل ويمكن أن يساهم في استدراج مشتركين آخرين من شبكات الإتصالات المنافسة عند تطبيق نظرية "كلمة الفم التسويقية" التي أثبتت فعاليتها على مدار الأعوام الماضية.