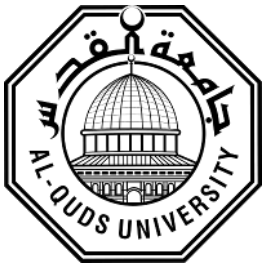
Deanship of Graduate Studies

Al-Quads University



Customize Social Network Analysis for Telecommunications Companies

Aktham Faheem Aqel Sawan

M. Sc. Thesis

Jerusalem-Palestine

1438/2017

Customize Social Network Analysis for Telecommunication Companies

Prepared by:

Aktham Faheem Aqel Sawan

M.Sc.: Computer Science-Al-Quds University-Palestine

Supervisor: Dr. Rashid Jayousi

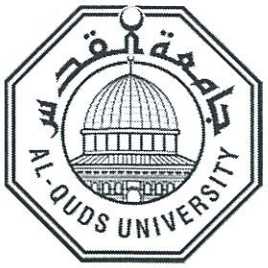
A thesis Submitted in Palestine Fulfillment of the Requirements for the Degree of Master of Computer / Department of Computer Science/ Faculty of Graduate Studies at Al-Quds University

1438/2017

Al-Quads University

Deanship of Graduate Studies

Computer Science Department



Thesis Approval

Customize Social Network Analysis for Telecommunications Companies

Prepared by: Aktham Faheem Aqel Sawan

Registration No: 21411486

Supervisor: Dr. Rashid Jayousi

Master thesis submitted and accepted. Date 05/07/2017

The names of signatures of the examining committee members as follow:

1-Head of Committee: Dr. Rashid Jayousi signature:.....

2-Internal Examiner: Dr. Nidal Kafri signature:.....

3-External Examiner: Dr.Yousef Abu Zir signature:.....

Jerusalem-Palestine

1438/2017

Dedication

This work is dedicated…
To my parents, for their love, support and encouragement…
To be beloved wife, without her caring support it would not have been possible…
To my lovely kids …Leen and Yahya…
To my brothers, sister, friends and colleagues…
To all of you I say a big “thanks” for your support.

Declaration

I certify that this thesis submitted for the degree of Master, is the result of my own research, except where otherwise acknowledged, and that this study (or any part of the same) has not been submitted for a higher degree to any other university or Institution.

Signed ……

Aktham Fahem Aqel Sawan

Data : 05 / 07 /2017

Acknowledgment

First praise is to the Almighty Allah, Lord of all creatures, the most gracious, most merciful, for his graces and blessings throughout all my life. Without him, everything is nothing.

My sincerer thanks for supervisor Dr.Rashid Jayousi, for his sincere efforts, interest and time he have kindly spent to guide my research.
I am very grateful to all professors at Al-Quds University in computer science department, for the time they have spent to teach me.

My study was funded by my employer Jawwal, special thanks to Jawwal represented by it GM Mr. Abdulmajeed Melhem and all the employees in Jawwal.

Finally, and most importantly, I would like to thank my wife. Her support and encouragement

Abstract

Social Network Analysis (SNA) is created to analyze social network data. Therefore, the main companies in the data mining filed (such as IBM, SAS, R and python) have created their own SNA algorithms. The aim of this research is to create customized SNA algorithm for telecom companies because the current algorithms were not designed just for the telecom networks, in addition when current algorithms were used for telecom many high value customer not include in final result plus results coverage just 55% from input customers, so in the new algorithm relation strength and extenders were used to enhance final results.

300 million records that belong to around 4 million customers in the last three were collected from (Jawwal Telecommunications Company) as case study. The current algorithms and the new algorithm were used the same data. In this research six experiments were applied based on call duration, call count and ratio between call duration and call count, in addition two groups size were used (15 and 20), Oracle Sql-PL/SQL was used to implement algorithm.

The results that approved by Jawwal were based on parameters that used in experiment number six (ratio between calls count and call duration with group size till 20 customer), it has increased the coverage of NW to be 75.9% instead of around 55% for current algorithms, in addition all high valued customers has included in results for the new algorithm, moreover algorithm have applied in Mobily in Saudi Arabia and the same positive results have been found same as Jawwal.

New novelty ideas have created in this research such as, extenders this type of customers used for customer who is influencer in one group and follower in the other group. Also relation strength used to create groups and assign followers to their most related influencer; furthermore, Super Group used as new layer to connect related groups in one group and find super influencer.

Table of Contents

List of figures

List of tables

Chapter one
Introduction

Social networks have become progressively the most important channel to connect people with each other due to what social media companies offers from services which has made the whole world connected such a small piece.

Recently big companies work to analyze social network that includes nodes(individuals or organizations) and who they exchange data between them (voice, data, SMS, video values, ideas, visions, financial exchange, kinship, friendship, dislike, trade or conflict, etc.,) [7, 15].

The general definition of SNA (social network analysis) is how to create groups for nodes based on their links and communications or activities. Thus, it draws the relationships between these nodes [8, 22].

There are several domains for social network analysis (SNA) such as (computer science, sociology, mathematics and physics) this leads to different methodological ways and many tools. That's why there are so many programs that were created to study and manipulate such networks [23].

SNA can be used in marketing in fields such as: Retention, attrition, prevention/ Churn, Segmentation, Acquisition new customers Fraud detection and up sell and Crosse ling much product adoption [22].

To recognize social communities depends on two aspects; firstly on the social relationships between nodes (customers); and secondly nodes measure that are based on influencer and disseminator. By using SNA we can target customers depending on a social communications and status changes within communities (e.g. a community influencer Churn you can target his or her followers to save them from churn) [22, 23, 24].

However, although the satisfied results from analyze social network data by using current SNA algorithms, these algorithms are general and has unsatisfied result for telecommunication company's data.

Below point explain some of SNA advantages:

1) Identifying which individuals are playing significant roles (knowledge brokers leaders, information managers ,etc.,).
2) Distinguishing information bottlenecks وbreakdowns وstructural holes, as well as isolated units, individuals, and teams.
3) Creating opportunities to fast-track knowledge flows through organizational and functional boundaries.
4) Strengthening the effectiveness and efficiency existing channels.
5) Raising alertness of important networks and groups to find ways to improve their performance.
6) Enhancing strategies.[26,27]

The present thesis aims to create a new SNA algorithm customized for telecommunication companies. The main idea for this algorithm is using relation strength as a main factor for creating groups to enhance the final results coverage and quality. The new algorithms has increased NW coverage comparing with current algorithm, Moreover high valued customers (who are classified such that based on their revenue average that from high section of customers revenue in telecom companies) include in final result for the new algorithm which do not exist in the final results for current algorithms.

Motivation

Recently the SNA model from IBM SPSS was applied in Jawwal NW data, but there was a problem in the final results of this model, because results covered only 55% of input Network, In addition, many high value customers Do not exist in SNA model results. These results were less than expected because IBM SNA is not designed for telecom only, for that reason decision has been taken to analyze the results of IBM SPSS, and try to customize their work for telecom. And also, to increase the model coverage and type of customers that should be included. For that reason in the new SNA model telecom business concepts are involved as a layer to build models, besides, this research approaches a new idea by adding network topologies (hierarchical NW design) that will be used to create NW clusters.

Research question

- What is the effect of applying new topologies on telecom customer Networks?

Objectives

- Create a new SNA model for telecom based on IBM SNA skeleton with Applying telecom business concepts as a guide for it.
- Create a new layer in SNA (super group) for SNA communities to enhance SNA results.
- Increase model coverage for the new SNA model.

Chapter two
Background

IBM, SAS, R and Python companies, created SNA models to analyze networks even social NW or telecom companies NW, these models focus on groups not on individual users. It creates groups consisting of individuals, who have relations with each other based on their calls or SMS transactions in telecom.

3.1 SNA form IBM

They have on each group influence and dissemination users for incoming and outgoing calls, which are the most important users in each group, and through the marketing campaigns target these users were tried for example[6].

Below the entities that should be used as input for the SNA model.

- Users: - there are direct or indirect relation between users (source number, target number).
- Value: - weight of relationship between users like the following figure 3.1.1 (value of relation in telecom for example call duration plus number of SMS)[24]

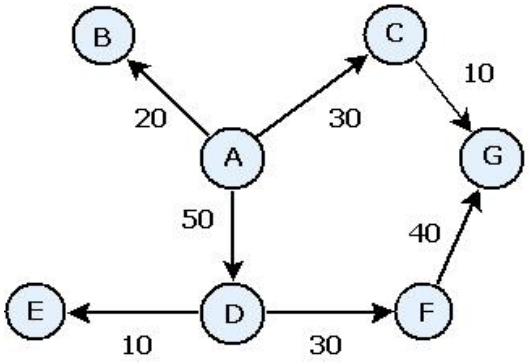
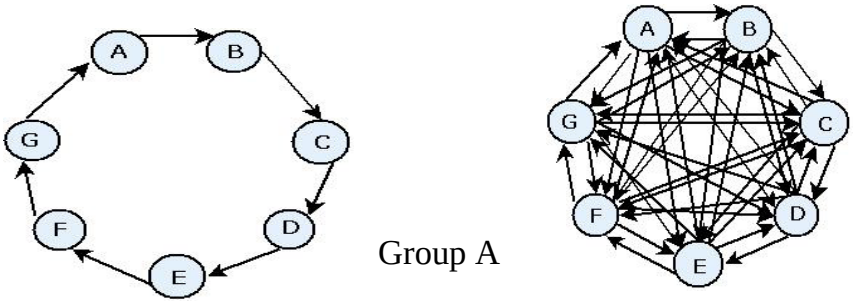


Figure 3.1.1: Sample of related nodes [6]

- Can be seen in figure 3.1.1that A calls B, C and D in different weights, based on these weights many measures can be determined by using the SNA model [25].
- When the SNA model creates group based model inputs, each group has two main characteristics (density and degree)
  - Density: - equal number of relations that exists in a group divided into all possible relations in a group.

For example:-as appears in figure 3.1.2 the density in group A equals 7/ (7*6) =0.17 and the density in group B equals 42/ (7*6) =1.0, the mean relation between individuals in group B being stronger than in group A



Group A

Group B

Figure 3.1.2: Sample of groups density [6]

- Degree: - each node in a group has in degree and out degree, Table 3.1.1 represents in degree and out degree for each node in as can be seen in Figure 3.1.1 for example.

Table 3.1.1: Sample of in and out degree for one group [6]

Node	Degree	In-degree	Out-degree
A	3	0	3
B	1	1	0
C	2	1	1
D	3	1	2
E	1	1	0
F	2	1	1
G	2	2	0

In Table 3.1.1 as can be found that node A has a greater number of out degree, which means it's the central node. In addition can be found that node G has a greater number of in degree nodes which means it's the prestige node [6].

When the SNA model create groups it uses the steps below.

1- Determine similarity for the nodes to include it in one group

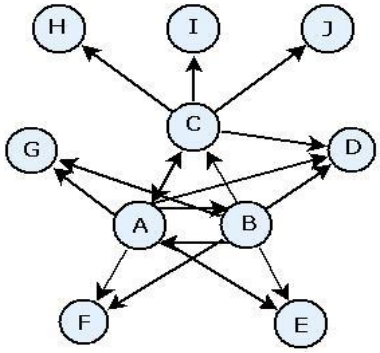


Figure 3.1.3: Sample of relations between nodes [6]

- As appears in the Figure 3.1.3 can be found that node A has relations with (B, C, D, E, F, G), and node B has relations with (A, C, D, E, F, G) and node C has relations with (A, D, H, I, J).
- From that can be found the node A and B calling the same 5 nodes(C, D, E, F, G), in addition can be found that node A and C call the same one node (D). So node A and B are more similar than node A and C and may have same group [24].

2- Partitioning node into groups

There are three criteria to partition nodes into groups.

- First the minimum group size to determine the minimum number of subscribers for each group.
- Second the maximum group size to determine the maximum number of subscribers for each group.
- Third coverage threshold, in this criteria SNA set high relationship weights for subscribers in the same group based on similarity. For example, if coverage was set to 40% that means stronger 40% of relation between subscribers were needed to be in the same group, but here may be faced some groups that contain more than the maximum group size. In such cases, SNA divides Groups into sub groups suitable with the maximum group size [25].

3- Describing groups and members in these groups

- Each group has density and in degree and out degree if group in Figure 3.1.1 was taken

- Can be found that density equals 14/42=33%, in degree equals 7/42 =17%, out degree equals 7/42=17%. also for members assign authority and dissemination users , high authorities will be assigned too high in degree in group in our example is (G) , high dissemination will be assigned to high out degree in group in our example is (A)[6]

SNA model output

In Table 3.1.2 analysis results for groups can be found

Table 3.1.2: IBM SNA output measures for network [24]

Statistic	Statistic Description
Total Nodes in Groups	Number of nodes included in the identified groups
Total Links in Groups	Number of links included in the identified groups
Total Number of Groups	Number of groups identified in the network
Mean Group Size	Average number of nodes in a group
Mean Group Density	Average fraction of direct connections between nodes in a group.
Mean Fraction of Core Members	Average fraction of nodes in a group that are core nodes for the group.
Mean Density of Core Group	Average fraction of direct connections between core nodes in a group
Mean InDegree	Average number of incoming links
Mean OutDegree	Average number of outgoing links.

In Table 3.1.3 for each group performance indicators can be found.

Table 3.1.3: IBM SNA output measures for groups [24]

Field	Description
GAG_Group	Number Unique identifier for a group
GAG_Size	Number of individuals in a group
GAG_Density	Fraction of direct connections between individuals in a group.
GAG_KernelDensity	Fraction of direct connections between core individuals in a group
GAG_CoreNodesFraction	Fraction of individuals in a group that are core individuals for the group.
GAG_MaxRankType1	Maximum authority score of any group member.
GAG_MinRankType1	Minimum authority score of any group member
GAG_MaxMinRankRatioType1	Ratio of the largest authority score to the smallest. This value reflects the authority strength of the group leader.
GAG_MaxRankType2	Maximum dissemination score of any group member.
GAG_MinRankType2	Minimum dissemination score of any group member
GAG_MaxMinRankRatioType2	Ratio of the largest dissemination score to the smallest. This value reflects the dissemination strength of the group leader.

In Table 3.1.4 for each individual performance indicators can be found.

Table 3.1.4: IBM SNA output measures for nodes [24]

Field	Description
GAI_NodeNumber	Unique identifier for an individual
GAI_CoreNode	Indicator of whether the individual is a core individual for a group or not.
GAI_RankType1	Authority score for the individual.
GAI_RankOrderType1	Rank order in the group based on the authority scores
GAI_RankType2	Dissemination score for the individual.
GAI_RankOrderType2	Rank order in the group based on the dissemination scores
GAI_InDegree	Number of relationships in which the individual is the target of the
GAI_OutDegree	Number of relationships in which the individual is the source of the
GAI_GroupLeaderType1	Whether the node is an authority leader, whose leadership score is derived from incoming links.
GAI_GroupLeaderConfidenceType1	The confidence that the node is an authority leader.
GAI_GroupLeaderType2	Whether the node is a dissemination leader, whose leadership score is derived from outgoing links.
GAI_GroupLeaderConfidenceType2	The confidence that the node is a dissemination leader.

3.2 SNA from R

SNA for R are used to categorize Main actors within group. For identifying actors, there are several centrality metrics in NW. these matrices are the following:-

- Degree :- to calculate a number of connections for each node (in degree, out degree)
- Betweenness :- to calculate total of shortest paths between nodes
- Closeness :- to calculate distance to all other related nodes
- Eigenvector::- centrality with weighted degree (to count more incoming links from highly central nodes).[16,17]
- Density :- to calculate a number of real connections in NW that is divided into the number of possible connections in NW
- Reciprocity :- the amount of nodes that are symmetric
- Mutuality :- to calculate number of complete transactions
- Transitivity :- to summarize the number of transitive trios
- Graph mean :-to calculate average of graph based on to the graphs density
- Centralization: - based on degree, closeness and Betweenness [17].

When using R tools there are diversity for analysis SNA based on graphs and it can be used for: bipartite networks and random graphs, also it is easy to calculate a degree, to weight networks and Built-in visualization tools. Moreover there are advantage of R's for

tools that used in built-in graphics, that connect immediately to extra statistical Analysis Achieve network and SNA based econometrics in the same roof [16].

3.3 SNA from SAS

There's a wide diversity of capable tools. SAS/Internet built reports integrating the group utilizing an applet MP connect are one of SAS's many Influential SNA tools. Also identified as Link Analysis, is a graphical and mathematical analysis focusing on the linking among "persons of interest". This approach of analysis has vast practical prominence in fields like fraud and epidemiology analysis [18].

Main measure of SNA from SAS.

- In Degree: - in this measurement numbers of incoming links for each subscriber are calculated.
- Out Degree: - in this measurement number of outgoing links for each subscriber are calculated.
- Centrality :- collected of In and out degree
- Density: - in this measure the density of groups is calculated, based on the numbers of links in group divided into maximum probable links that can exist in each group.
- Prestige :- divided number of culms into row based in size of NW.[19]

3.4 SNA from python

Can be found that a NW is a set of nodes connected via links Nodes. It can be anything based on the context for SNA they used many measures and packages to analyze the NW and relation between its nodes. Below you can find some of the main measurement that were used on SNA and created by python [21]

- Degree (in, out): calculated incoming/outgoing edges for each.
- Centrality :- collected of In and out edged
- Density: - the density of a group is calculated by the number of links in groups that were divides into maximum probable links that can exist in each group.
- Closeness :- to find distance to all other related nodes
- Betweenness: - to calculate a total of shortest paths between nodes.
- Eigenvector: - centrality with weighted degree (count more incoming links from highly central nodes).
- Page Rank: like Eigenvector but not firmly on a centrality measurement [20].

Chapter three
Literature review

Many studies have been done on social network analysis, these studies used SNA in different domains and tried to enhance their results and processes.

In study [1] the researcher states that when you work with social network data it means you are working with huge big data. For that the researcher aims to improve the performance of algorithm that used to assign communities in SN from processing side. By comparing implementing sequential processing, Dynamic Parallel (DP) and a Hybrid CPU-GPU (HCG), then finds the more efficient processing to speed-up Clustering-Based Community Detection in Social Networks.

The results in [1] discovered that hybrid and dynamic parallel processing are faster than sequential implementation. At the same time hybrid implementation is faster than dynamic parallel.

Can be found that the online SN is sharing similar topological features as real world SN. Recently there are a lot of studies that have been discussing how to analyze SN but there are difficulties in linking SN software design with human interests.

The author in [2] proposed a new methodology to involve human interactions and analyze it by Designing online SN software. The approach in [2] proposed a novel use of SNA techniques for eliciting requirements to design a good online SN-based software. To validate the new methodology the author in [2] involved a questionnaire to measure a NW design construction by collecting data from real-world .The main idea is to examine NW for helping in the identification of interests and behaviors of people for better elicit into software (SW) requirements.

The results in [2] proved that the proposed novel methodology was effective for the designing of online social network (SN) software. Also new methodology improves comparing to traditional methods for SW design, where it involves SN modeling by analyzing and eliciting the behavior of requirements to create stronger online SN sites.

The huge amount of data in social networks can cause plenty of elements in visualization of data, which will be challenging when discovering the information. To raise the level of abstraction the author in [3] represented Chord diagrams.

The author in [3] adapted and introduced a new technique to visualize social network (SN) in process mining. This approach is categorized as visualization technique, called Chord diagram.

Technique for interactions between nodes and nodes is declared formally in [3] by using chord diagram and there is a plug-in to support the SN visualization in BPM area that was developed in Prom framework. In study [3], the plug-in is working to visualize SN in a real scenario, by using traditional graph and chord diagram visualization technique.

Results for [3] show that the new product can prove new insights into social network, like association direction, resource involvement, association contributions, resource abstraction and special nodes. It also uncovered that a node cannot be isolated from others in this approach.

To improve the classification of different users, the author in study [4] proposed the efficient method OWL DL this method utilized semantic in web technology to describe target rules, Friend-Of-A-Friend (FOAF) and SNA for defining users' domain.

The authors in study [4] proposed method, the field ontology that will be generated from decision tree domain then the domain will be classified to users and web. Pages, after that SNA is used for analyzing the tags that are coming from the profiles of users Friend-Of-A-Friend (FOAF), then to obtain the users' interests.

To verify the efficiency of proposed study in [4] the authors manage an experiment in which they gathered the users' interests for those who voted for the Untied State of America presidential candidates.

The results of study [4] have a high level of accuracy and precision for classifying users, In addition, in the gained interests for growing of users, more consistent as more

web. Pages are additional to the classification. Accuracy and precision of this approach are 91.5, and in classifying the users, individually 93.1.

The authors of study [5] proposed to use SNA instead of current tools that are used for analyzing transportation, because the mentioned tools are time consuming and expensive. In addition it needs rigorous data to have reliable results. However, SNA can deal with connectivity and complexity of NW in a cost-and time effective manner.

In this study [5] SNA is used to analyze transportation NW and therefore corroborate its efficiency as a complementary tool to improve planning of transportation. For that, the author in study [5] accepted four steps For research methodology: (1) Exploring the connection between the concepts and language of Social Network Analysis and those of systems of transportation;(2) working for the transportation context in different Social Network Analysis centrality measures;(3) working on SNA for two cases in Mississippi and studies; and (4) studying the results of drawing and the case studies.

The results of the proposed study in [5] approved that SNA is able to quickly and easily determine the greatest critical intersections in the examined transportation NW. It is applied on the Department of Transportation (MDOT), the research also believed that SNA is an innovative and effective tool for the analysis of transportation.

Social Network Analysis has been accepted as a promising method after the 11 September, 2001 attacks as the best method for investigating and ranking terrorists and their communities, as well as to find leaders /players in these communities, but when using SNA alone there are problems on ranking coming from centrality. For that reason, authors in [9] suggest to use Analytical Hierarchy Process (AHP) with SNA as plug in to improve results of centrality measure, for that and as a result of this improvement.

The results in [9] when combining SNA with Analytical Hierarchy Process (AHP), are promising. It can rank and order 19 terrorists and these terrorists involved in the attack and can find the key leaders /players between them. Also sensibility analysis is devoted to deal with modifications in particular judgments.

In paper[10] the authors suggest a new method Social network based engineering education (SNEE) for analyzing the students' relations to Evaluate students learning, also for organizing the education method as a social NW based method, for that they compare the new method that depends on students behavior with traditional models of students' evaluation.

To validate the new method the authors in [10] correlate between SNA measures (closeness centrality, degree centrality, eigenvector centrality, between centrality and tie strength average) and the grades gained by students (grades for quality of work, volume of work, diversity of work, and also final grades).

Main findings in paper [10] create a new model for evaluation of the students depended on their behavior in the courses and create partial automated student evaluation based on applying SNA, also increasing the productivity and objectivity for teachers and giving them a more scalable process in evaluation.

Results of paper [10] interaction that have more different for students, had the degree of centrality and the most frequent the student was among the interaction paths with the other students (centrality of between-ness), the superiority was the worth of the work.

All five SNA measures had a strong and positive correlation to the grade for final grades and volume of work.

All students with better average tie power had the highest grade for variety of work comparing with low ties.

In paper [11] the authors work to create a new algorithm. The main aim of this algorithm is to find better central nodes (influence nodes) that were calculated by the SNA algorithm centrality measure. Since finding influence nodes and most active users in their groups is the most challenging and important task in latest years.

For that the authors in [11] define four factors which affect the relationships' strength. Then they modified the measure for the eigenvector centrality and joint it with a system of fuzzy inference to have more realistic results.

To do that, they used a fuzzy interface system to calculate the strength of each relation, then created a crisp matrix in which the corresponding elements identify the strength of each relation, and using this matrix eigenvector measure calculated the most influential node.

The results of paper [11], where the authors apply the new algorithm, the authors have more realistic influence nodes (central nodes). By considering the strength of connected nodes for all friendships.

In paper [12] the author designed a hybrid model based on ANN and fuzzy techniques for opinion system recommendation, for communities and users respectively.

They propose this model because the necessary date is frequently hidden and distributed on servers of social sites. For that they thought to design a new approach for the analysis and collection of such data in the social web.

The results of paper [12] after applying and testing the hybrid technique of ANN and Fuzzy (HFANN) technique, the author gather data for testing from three main sites for social network analysis, after which they used to test and train the proposed methods, the hybrid approach of combining ANN and Fuzzy both have advantages for results with high accuracy and utilization in classifying social data.

In paper [13] the authors in this research propose a collective model using user based and content-based approach, along with a method for centrality measurement.

Content-based method tends to attention on tweet content and studies of nodes that exists in a NW, but the user-based method focuses into communications between nodes through connection in the NW twitter.

The results of paper [13] showed that the interaction that happens between nodes and the centrality degree, affect the purpose of the maximum significant nodes in a NW twitter

In paper [14] the author develop a new method when using scaling of a multidimensional scale to differentiate between some social NW. Moreover, the others used various factors and methods to compare social NWs, like comparing density of the social NW, degree of node, the number of triads and path length in a NW. The Proposed technique works well if there are more than three social NWs compared.

The result of paper [14] explain, that techniques that used by the author were more successful when it was compared it with some social NWs. It achieved by generating a Metadata table of several social NW variables and using scaling of multidimensional techniques like Hiclusand Prefscal. Also the techniques that were used by the author gave better results when more than three social NWs were compared. So it's clear that these techniques deliver a more systematic and comprehensive process of comparing social NWs

In this research, we've created a new layer for SNA to detect groups based on telecom business concepts, In addition there is a new type of users called extender. This user can be member in one group (extender) and disseminator or influencer in the other group. Also we created a new grouping technique (super group).This technique is based on hierarchical NW tree, it communicates groups through extenders to create a main group (super group) that has more than one basic group. The results of the new SNA model were promising because the coverage of NW increased from 55% in IBM SNA to 75.9% in the new SNA model, furthermore, all high value customers exist in new SNA model.

Below table contain summary of all literature review papers by having brief about each paper problem, tools that used and results.

Table 2.1.1 - A: sample of in and out degree for one group

Papers	Problem	Tools	Results
[1]	The author focused to Improve the performance of algorithm that used to assign communities in SN by using hybrid approach.	Run data into SNA algorithm and measure performance of CPU	From empirical results the authors find that hybrid and dynamic parallel processing faster than sequential implementation. At the same time hybrid implementation is faster than the dynamic parallel.

Table 2.1.1 - B: sample of in and out degree for one group			
Papers	Problem	Tools	Results
[2]	The author proposed a new methodology to involve human interactions and analyzed it by designing online SN software	developed the methodology is depends on a software development design process	Results for this study proved that the proposed novel methodology was effective for designing online SN software. Also the new methodology improved compared to traditional methods for SW design, where it involves SN modeling by analyzing and eliciting the behavior of requirements to create stronger online SN sites.
[3]	The huge amount of data in social networks can cause plenty of elements for the visualization of data, which will be a challenge when discovering the information. To raise the level of abstraction the author represent Chord diagrams.	Prom framework open source, with supports visualization for chord diagram	Results for this study show that the new product provide a new insight for the investigation of social network, like association direction, resource involvement, association contributions, resource abstraction and special nodes. It also uncovers that nodes cannot be isolated from others in this approach.
[4]	How to classify different users, to do that the author proposed the efficient method OWL DL this method is utilize semantic in web technology for describing target rules and FOFA and SNA for define users domain	SNA ,Decision tree, OWL DL and FAFA	The results of the proposed method has high level of accuracy and precision for classifying users, in the gained interests for growing of users, more consistent as more webpages are additional to the classification. Accuracy and precision of this approach are 91.5 and in classifying the users individually 93.1%.

Table 2.1.1 - C: sample of in and out degree for one group			
Papers	Problem	Tools	Results
[5]	Using SNA instead of current tools that is used for analyzing transportation because the mentioned tools are time consuming and expensive. In addition it needs rigorous data to have reliable results. However, SNA can deal with connectivity and complexity of NW in a cost and time effective manner.	Using SNA for transportation	The results of the proposed study approved that SNA is able to quickly and easily determine the greatest critical intersections in the examined transportation NW. It applied on (MDOT) Department of Transportation, The research also believes that SNA is an innovative and effective tool for the analysis of transportation.
[9]	Improving SNA results to discover and rank terrorists and their target lists by using Analytical Hierarchy Process (AHP).	SNA and AHP	Results of combining SNA with AHP are promising, based on rank ordering find 19 it could terrorists which were involved in attacks and could find the key leaders /players among them.
[10]	In this paper the authors suggest a new method. Social network based engineering education (SNEE), for analyzing the students' relations to evaluate students learning, also for Organizing the education method as a social NW based method.	Social network based engineering education SNEE, SNA	Results of this paper interaction that have more different for student, had the degree of centrality and the most frequent the student was among the interaction paths with the other students (centrality of betweenness), the superior was the worth of the work. All SNA five measures had a strong and positive correlation to the grade for final grades and volume of work. All students with better average tie power had the highest grade for variety of work compared to low ties.

Table 2.1.1 - D: sample of in and out degree for one group			
Papers	Problem	Tools	Results
[11]	In this paper authors worked to create a new algorithm. The main aim of this algorithm was to find better central nodes (influence nodes) that were calculated by SNA algorithm centrality measure. They did that by using fuzzy interface system.	SNA ,Fuzzy interface system	Result of this paper when applying the new algorithm, show that authors have more realistic influence nodes (central nodes), with taking into consideration the strength of connected nodes for all friendships.
[12]	In this paper the author designed a hybrid model based on ANN and Fuzzy techniques for opinion system recommendation, for communities and users respectively.	SNA, ANN and Fuzzy	The results of this paper after applying and testing hybrid technique of ANN and Fuzzy (HFANN) provide themselves were accurate and utilized for classifying social data.
[13]	The authors in this research propose a collective model using user based and content-based approach, along with a method for centrality measurement.	SNA	The results of this paper showed that the interaction that happens between nodes and the centrality degree, affect the purpose of the maximum significant nodes in a NW twitter.
[14]	In this paper the author develops a new method when using scaling of a multidimensional scale to differentiate among some social NW. also the others used various factors and methods to compare social NWs.	SNA	The Result of this paper explains that techniques that used by the author were more successful when it was compared with some social NWs. it achieved by generating a Metadata table of several social NW variables and using scaling of multidimensional techniques like Hiclus and Prefscal.
My research	Current SNA models not customize for telecom. For that results of these models coverage around 55% of input NW, moreover there are many high value customers for telecom that do not exists in results.	SNA	Creating new SNA model and customizing it for telecom companies, the results of new model coverage 77% of input NW, also all high value customers are covered in the new SNA model.

Chapter four

Methodology

To achieve the objectives of this research that is customizing SNA for telecom companies, collecting data that needed to be input for model was done. SPSS SNA Skelton has been used in this new model. The work has been processed as following

4.1 Data collection

The data was collected based on Jawwal Network data, all incoming and outgoing calls and SMS for all Jawwal customers for the last three months. Around 300 million records were collected which is belong to around 4 million customers. Data includes sender number and receiver number, but these numbers were encrypted for security purposes. Encrypted ID was used for each number instead of the real number. In addition, duration and count of calls and SMS between sender and receiver were collected, every two SMSs counted as one minute of call. Outliers' data were removed such as numbers that were used from systems.

4.2 Method

The main goal of SNA models is to create groups for customers and find important customers (influencers and disseminator) as can be seen in figure 4.2.1.

Figure 4.2.1: Influencers and Disseminations in Groups

The below steps were used to find them.

As can be seen in figure 4.2.1 the subscriber in red is influencer subscriber and the others subscriber in blue are flowers and who they communication in one group.

1. Calculate the degree of centrality for each node based on the node degree distribution, the duration of calls and the number of calls and SMS (two SMSs counted as one call). Example if customer number599??33?3 C1 has 5 outgoing relations with other customer, calculate degree of centrality of this group was needed based on duration and count of calls for each customer.

   Let C1 has the following outgoing data with other customers

   C2 (5 duration, 2 count of calls),C3(10 duration, 2 count of calls),C4(5 duration, 2 count of calls),C5(8 duration, 4 count of calls),C6(7 duration, 2 count of calls).

   Degree of centrality equal summation of all durations of outgoing duration of customer multiply by 0.7 and summation of outgoing calls counts multiply by 0.3 =(35*0.7) +(12*0.3) =24.5+4 =28.5

   By using mathematic notations

   Let's have two array for C1 first for outgoing calls duration and second for outgoing calls count

   Duration Array values are DARY{5,10,5,8,7}

   Count Array values are CARY{2,2,2,4,2}

   //*to calculate degree of centrality for C1*

   Deg_of_Cent= $\sum_{i=1}^{n-1} DARY[i] * 0.7 + \sum_{i=1}^{n-1} CARY[i] * 0.3$
2. Rank all nodes in networks for 4 million customers based on degree of centrality by using adjacency matrix for directed graph.
3. Calculate relation strength between every two nodes.

   Example if total outgoing calls for A equal 100 minutes and A call 5 numbers as following (B 15 minutes, C 10 minutes, D 15 minutes, F 20minutes and E 40 minutes). Relation strength between A and B =15/100 =15% and maximum relation strength between A and E =40/100=40%

   By using mathematic notations

   Let's have array for A customer and his outgoing calls

   Array values are ARY{15,10,15,20,40}

   X= $\sum_{i=1}^{n-1} ARY[i]$ //*to calculate outgoing calls for A*

   //to calculate relation strength

Res←0

For i←0 to n-1

Res←ARY[i]/X

Return Res

4. Find the most important customer for each customer.
5. Created group for most important customer who have maximum number of followers in networks (10 top percentile), groups created based on parameter of maximum size for groups (15, 20) the assign followers to influencer customer in each group based on relation strength.
6. For each group calculate closeness centrality to determine disseminator customers who have the highest number of outgoing relations with the other customers in the same group and influencer customer who have the highest number of incoming relations with the other customers in the same group as can be seen in figure 4.2.2.
7. Rank customers for each group based on incoming and outgoing relations and assign extenders.
8. Calculate measurement about each group (such as size, density, max rank for dissemination, min rank for dissemination , group max rank for influencer ,group min rank for influencer ,group in degree and group out degree).

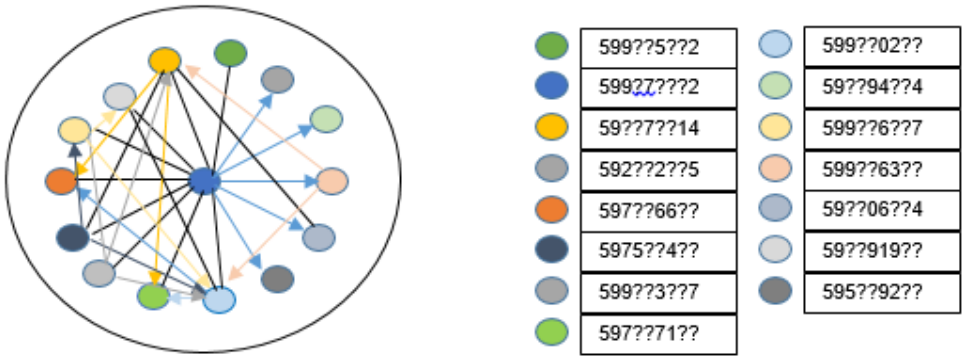


Figure 4.2.2: SNA sample real group

Also a new type of customers were determined in this research, that's called Extender. This customer is disseminator or influencer in a group in the meantime is follower in other group as can be seen in figure 4.2.3.

Based on Extenders in this research a new type of groups were identified called super groups. To calculate super groups the following steps were used based on groups (normal groups) that was generated.

1. Calculate the number of extenders in each normal group.
2. Calculate the degree of centrality of normal groups

   Example if group number1 G1 has 5 customers, calculate degree of centrality of this group was needed based on duration and count of calls for each customer

   Let G1 has the following customer's data

   C1(10 duration, 5 count of calls),C2(20 duration, 10 count of calls),C3(15 duration, 10 count of calls),C4(25 duration, 5 count of calls),C5(35 duration, 15 count of calls).

   Degree of centrality equal summation of all durations of customers multiply by 0.7 and summation of all calls counts multiply by 0.3 =(105*0.7) +(45*0.3) =73.5+13.5=87

   By using mathematic notations

   Let's have two array for G1 first for calls duration and second for calls count

   Duration Array values are DARY{10,20,15,25,35}

   Count Array values are CARY{5,10,10,5,15}

   //*to calculate degree of centrality for G1*

   Deg_of_Cent= $\sum_{i=1}^{n-1} DARY[i] * 0.7 + \sum_{i=1}^{n-1} CARY[i] * 0.3$
3. Rank the whole normal groups that were calculated based on number of extender in each group then degree of centrality of group.
4. Start from the highest rank of normal group (that have maximum number of extenders) and used hierarchical NW design mixed with extenders as it appears in chart 4.2.3.

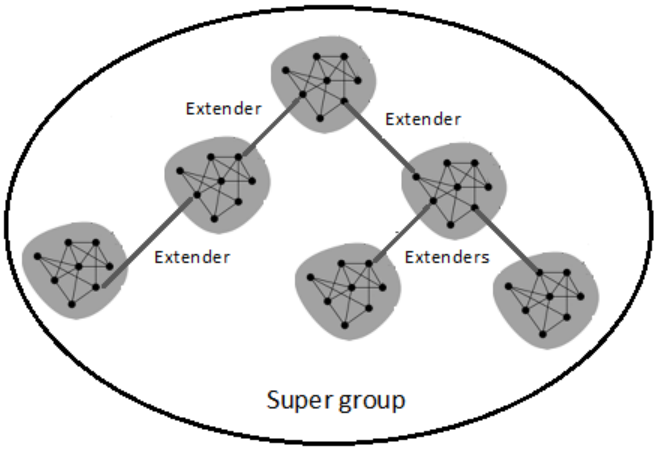


Figure 4.2.3: Super group

5. Calculated closeness centrality for each super group to determine disseminator customers who have the highest number of outgoing relations with the other customer

in a super group and influencer customer who have the highest number of incoming relations with the other customers in a super group.

6. Rank customers based on incoming and outgoing relations to find influencers and disseminator as can be seen in figure 4.2.4.
7. Calculate measurement about each super group(such as size, density, max rank for dissemination, min rank for dissemination , group max rank for influencer ,group min rank for influencer ,group in degree and group out degree).

Figure 4.2.4: Influencers and Disseminations in super groups

Below table contains a sample group from SNA groups

Table 4.2.1:Sample results from SNA groups measure

Proposed model	
GAG_Super_GroupNumber	**10**
GAG_GroupNumber	**15680**
GAG_Size	**15**
GAG_Density	**0.21**
GAG_MaxRankType1	**0.20**
GAG_MinRankType1	**0.02**
GAG_MaxMinRankRatioType1	**10.00**
GAG_MaxRankType2	**0.32**
GAG_MinRankType2	**0.01**
GAG_MaxMinRankRatioType2	**33**

In table 4.2.1 as can be seen the measures that were calculated for each groups.

- GAG_ Super_Group Number: - in this measure each super group has a unique number.
- GAG _Group Number: - in this measure each group has a unique number.
- GAG _size:- in this measure the numbers of subscribers in each group are calculated.
- GAG _density:- in this measure the density of a group was calculated based on the number of links in a group divided by the maximum probable links that can exist in each group.
- GAG_MaxRankType1:-in this measure the maximum ratio for dissemination user was calculated based on his links in a group divided by the total of links in the group.
- GAG_MinRankType1:- in this measure the minimum ratio for dissemination user was calculated based on his links in a group divided into the total of links in the group.
- GAG_MaxMinRankRatioType1:-in this measure the maximum ratio of dissemination user divide into minimum ratio of dissemination user was calculated.
- GAG_MaxRankType2:- in this measure the maximum ratio for influencer user was calculated based on his links in the group divided into the total of links in the group.
- GAG_MinRankType2:- in this measure the minimum ratio for influencer user was calculated based on his links in a group divided into the total of links in the group.
- GAG_MaxMinRankRatioType2:-in this measure the maximum ratio of influencer user divide into the minimum ratio of influencer user.

Table 4.2.2.1:Sample results from SNA subscribers measure

GAI_Node Number	GAI_ InDegree	GAI_ OutDegree	GAI_Rank OrderType1	GAI_Rank OrderType 2	GAI_ RankType 1
599??5??2	1	1	12	9	0.02
599?7???2	9	14	1	1	0.20
59???7?14	5	5	3	3	0.11
592??2??5	1	0	15	15	0.02
597??66??	3	1	6	10	0.07
5975??4??	2	4	9	6	0.05
599??3??7	2	4	8	5	0.05
597??71??	3	1	7	12	0.07
599??02??	6	4	2	4	0.14
59??94??4	1	0	13	13	0.02
599??6??7	3	5	5	2	0.07
599??63??	1	2	11	7	0.02
59??06??4	2	1	10	11	0.05
59??919??	4	2	4	8	0.09
595??92??	1	0	14	14	0.02

- GAI_ Node Number: - Mobile number subscriber exist in this filed.
- GAI_In Degree: - in this measure the number of incoming links for each subscriber was calculated.
- GAI _Out Degree: - in this measure the number of outgoing links for each subscriber was calculated.
- GAI_RankOrderType1:- in this measure the order of subscriber based on incoming links were calculated.
- GAI_RankOrderType2:-in this measure the order of subscriber based on outgoing links were calculated.

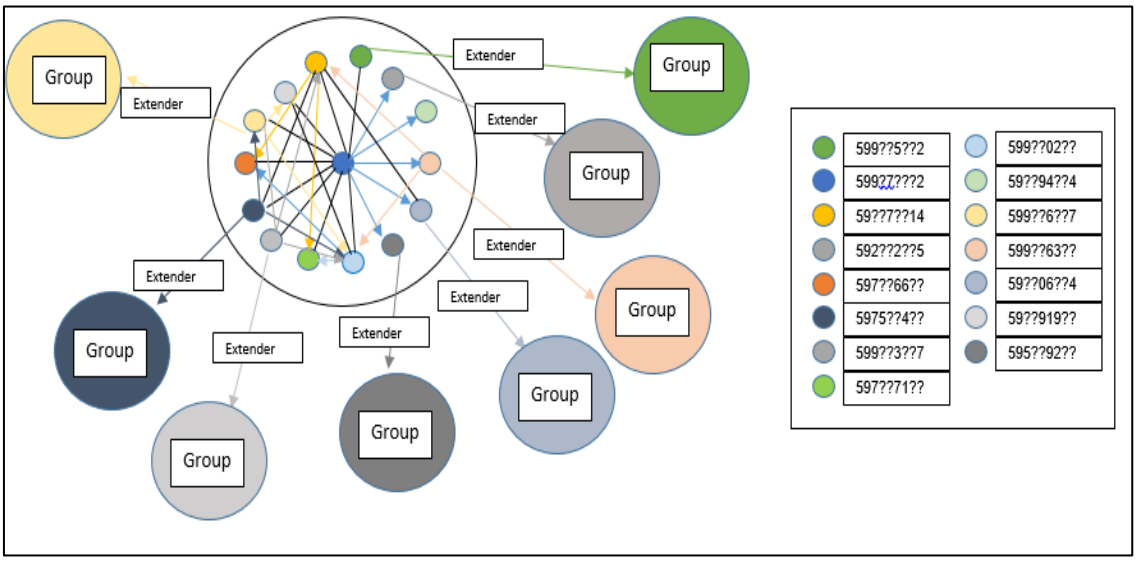


Figure 4.2.5: Sample super group with real results

- GAI_RankType1:-in this measure the number of incoming links divided into the total of incoming links for each subscriber were calculated.

Table 4.2.2.2:Sample results from SNA subscribers measure

GAI_RankType2	GAI_GroupLeaderType1	GAI_GroupLeaderConfidenceType1	GAI_GroupLeaderType2	GAI_GroupLeaderConfidenceType2	IS Extender	Is core	Relation strength
0.02	0	0	0	0	1	1	0.15
0.32	1	10	1	33	0	1	0.28
0.11	0	0	0	0	0	1	0.63
0.00	0	0	0	0	1	0	0.52
0.02	0	0	0	0	0	1	0.23
0.09	0	0	0	0	1	1	0.05
0.09	0	0	0	0	1	1	0.61
0.02	0	0	0	0	0	0	1
0.09	0	0	0	0	0	0	0.08
0.00	0	0	0	0	0	0	0.72
0.11	0	0	0	0	1	1	0.32
0.05	0	0	0	0	1	1	0.09
0.02	0	0	0	0	1	0	0.21
0.05	0	0	0	0	0	0	0.07
0.01	0	0	0	0	1	0	0.67

- GAI_RankType2:-in this measure the number of outgoing links divided into the total of outgoing links for each subscriber was calculated.
- GAI_GroupLeaderType1:-in this measure the disseminator user assigned by flag 1 and follower users were assigned by flag 0.
- GAI_GroupLeaderConfidenceType1:-in this measure the disseminator user assigned by his confidence (GAG_MaxMinRankRatioType1) and normal users were assigned by flag 0.
- GAI_GroupLeaderType2:-in this measure the influencer user assigned by flag 1 and normal users were assigned by flag 0.
- GAI_GroupLeaderConfidenceType2:-in this measure the influencer user assigned by his confidence (GAG_MaxMinRankRatioType2) and normal users were assigned by flag 0.
- Is extender: -in this measure the extender customer assigned by flag 1 and normal customer was assigned by flag 0.
- Is core: - in this measure the core user that directly related to the influencer assigned by flag 1 and not related user to the influencer user assigned by flag 0.

- Relation strength: - in this measure percentage of relation between user and influencer assigned, based on both duration and count of transaction that happened between them from user side.

4.3 Validation

To validate the new SNA model in creating groups, three different parameters for that degree of centrality (duration , count and mix duration with count) have been tested, in addition data based on two maximum groups size (15 and 20) have been tested because the average community size for each customer in Jawwal was 18.2 customers.

To validate the quality of results for the proposing model in real environment, campaigns have been created to find out how many indirect customers will be opted-in into offers when target their influencers without target them. The same exercise will apply on IBM SPSS communities.

Chapter five

Experiment and Results

This research experiments were built to test if association, clustering and classification algorithms can be used to buildup groups for networks. To have optimal results different parameters were used for the degree of centrality (duration, count and mix duration with count) with two maximum groups (size 15 and 20) because to mean community size for each customer in Jawwal is 18.2 customers.

5.1 SNA with association results

This research experiments were built to test if association algorithms can be used to create SNA grouping (like Apriori and CARMA) , in this experiment sender customer was provided as ID for association algorithm and the receiver customer was provided as target. The results of this experiment were 5 thousand groups, each group has two to four customers, that means these groups just covered around 15 thousand customers. It's just 0.03% from the base, for that reason the association algorithm can't be used to create SNA groups.

5.2 SNA with clustering algorithms results

This research experiments were built to test if clustering algorithms can be used to create SNA grouping (like K-mean, Two Steps and Kohonen), these algorithms are used to create groups for customers that have the same Pattern. In addition, it works just with sender and some measures about him. In this research groups should be built based on communications between senders and receivers such as incoming and outgoing calls and SMS . For that reason the clustering algorithm can't be used to create SNA groups.

5.3 SNA with classifications algorithms results

This research experiments were built to test if classifications algorithms can be used to create SNA grouping (like ANN, Regression and Decision Tree), these algorithms are used to predict future action for customer based on significant measures provided to

classification algorithm and the result should be between 0 and 1. In this research groups should be built based on communications between senders and receivers such as incoming and outgoing calls and SMS .For that reason the classification algorithm can't be used to create SNA groups.

5.4.1 Proposed algorithm VS SPSS SNA centrality bases on duration and max group size 20

In this experiment, both proposed algorithms and SPSS SNA model were compared. By using different parameters as the following.

- Degree of centrality calculated based on groups sizes that should be minimum 2 and maximum 20 and data should be based on duration.
- Closeness centrality calculated based on summation of duration for incoming and outgoing calls and SMS that happened between nodes.

Table5.4.1.1: Results for group size between 2 and 20 and centrality based on duration

Measure	IBM SPSS	Proposed model
Nodes in groups	2,109,681	3,009,346
Links in groups	4,688,195	6,594,848
Number of groups	183,410	274,527
Mean group size	11.5	11.91
Groups density	0.23	0.24
Mean of In Out degree	1.96	2.02

In this experiment, as can be seen in table 5.4.1.1, as can be found in the main points below

- By using proposed model the coverage of NW increased to be around 75.2 %, furthermore SPSS SNA coverage 52.7%.
- Links between nodes in proposed model are around 6.6 M, but in SPSS SNA 4.68 that means coverage of links increases by 41%.
- The number of groups in proposed model are around 275K but in SNA SPSS 183Kthe number of groups increased by 50%.

- The same figures in groups (density, mean in and out) can be found, though the coverage of NW increased.

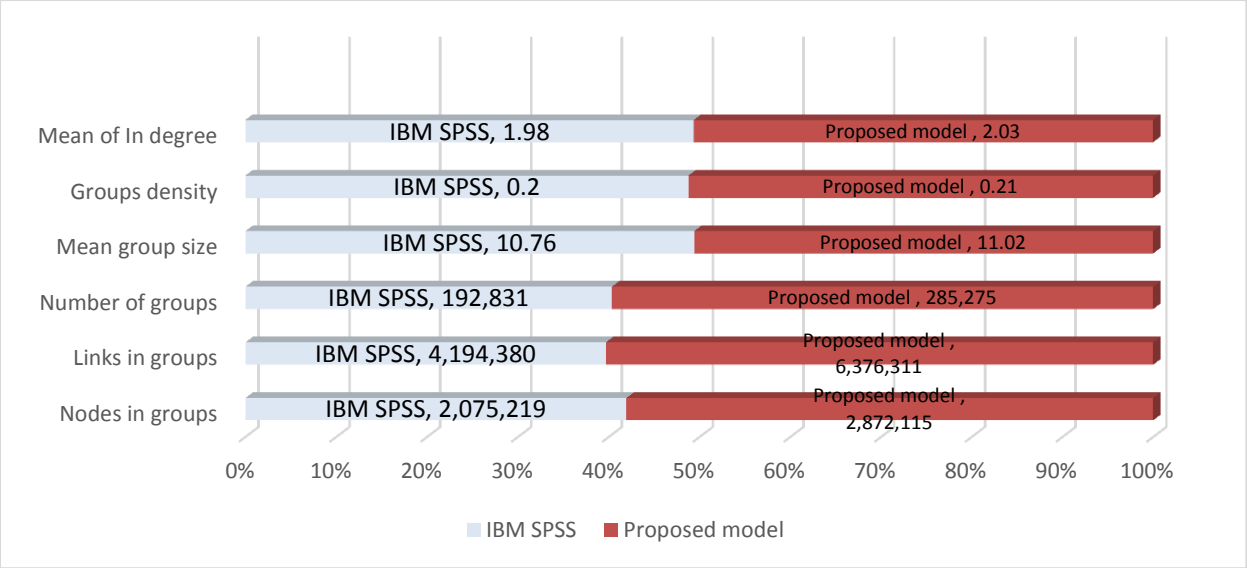


Figure5.4.1.1: (Stacked bar) Results for group size between 2 and 20 and centrality based on duration

- As can be seen in figure 5.4.1.1 these measures compare between SPSS SNA and proposed model based on ratio for each one.

Table 5.4.1.2: super groups results for group size between 2 and 20 and centrality based on duration

Measure	Proposed model
Nodes in groups	2,676,274
Links in groups	6,743,396
Number of Extenders	263,214
Mean super group size	134.48
Mean Influencers in super group	13.2
Mean of In degree	2.51

As can be seen in table 5.4.1.2 about super groups results for group size between 2 and 20 and centrality based on duration, as can be found in the below points

- New type of customers extenders was 263 thousand customers
- Mean super group size was134 customers.

- Each super group has around 13.2 groups and influencers
- Coverage of NW that have super groups was 67% .
- Links between nodes that have super groups were 6.74 M.
- Note that there are main influencers and disseminators in each super

5.4.2 Proposed algorithm VS SPSS SNA centrality bases on duration and max group size 15

In this experiment, both proposed algorithms and SPSS SNA model were compared. By using different parameters as the following.

- Degree of centrality calculated based on groups sizes that should be minimum 2 and maximum 15 and data should be based on duration.
- Closeness centrality calculated based on summation of duration for incoming and outgoing calls and SMS that happened between nodes.

Table 5.4.2.1 : Results for group size between 2 and 15 and centrality based on duration

Measure	IBM SPSS	Proposed model
Nodes in groups	1,991,490	2,841,413
Links in groups	4,120,028	6,232,455
Number of groups	190,630	283,350
Mean group size	10.39	10.97
Groups density	0.21	0.22
Mean of In Out degree	1.95	2.01

In this experiment as can be seen in table 5.4.2.1 results for group sizes between 2 and 15 and centrality based on duration, as can be found in the main points below

- Coverage of NW by using our research model increased to be around 71 %, moreover SPSS SNA coverage 49.7%.
- Links between nodes in our research are around 6.2 M, but in SPSS SNA 4.1 that means coverage of links increases by 51%.
- Number of groups in our research are around 283K but in SNA SPSS 191K the number of groups increases by 49%.

- The same figures in groups (density, mean in and out) can be found, though the coverage of NW increased.

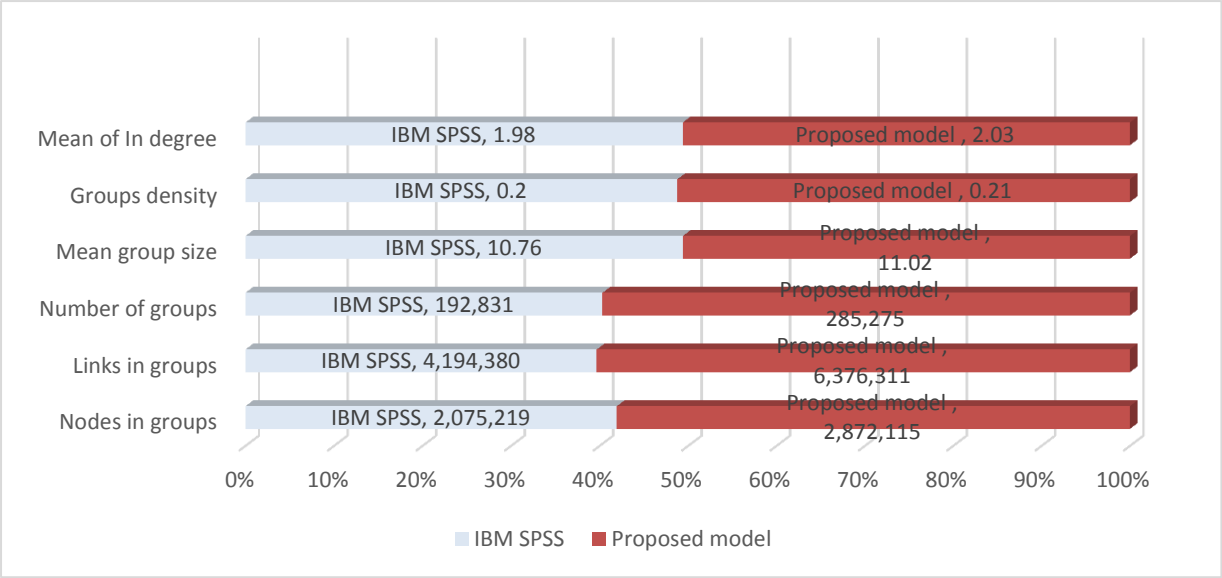


Figure 5.4.2.1: (Stacked bar) Results for group size between 2 and 15 and centrality based on duration

- As can be seen in figure 5.4.2.1 these measures compare between SPSS SNA and proposed model based on ratio for each one.

Table 5.4.2.2 : Super groups results for group size between 2 and 15 and centrality based on duration

Measure	Proposed model
Nodes in groups	2,385,857
Links in groups	6,201,320
Number of Extenders	267,264
Mean super group size	119.91
Mean Influencers in super group	13.4
Mean of In degree	2.59

As can be seen in table 5.4.2.2 about super groups results for group size between 2 and 15 and centrality based on duration, as can be found that in the below points

- New type of customers extenders was 267 thousand customers
- Mean super group size was 119 customers.
- Each super group has around 13.4 groups and influencers
- Coverage of NW that have super groups was 60 %.
- Links between nodes that have super groups were 6.2 M.
- Note that there are main influencers and disseminators in each super group.

5.5.1 Proposed algorithm VS SPSS SNA centrality bases on number of connections (count) and max group size 20

In this experiment, both proposed algorithms and SPSS SNA model were compared. By using different parameters as the following.

- Degree of centrality calculated based on groups sizes that should be minimum 2 and maximum 20 and data should be based on number of transactions (count).
- Closeness centrality calculated based on summation of number of transactions (count) for incoming and outgoing calls and SMS that happened between nodes.

Table 5.5.1.1 : Results for group size between 2 and 20 and centrality based on Count

Measure	IBM SPSS	Proposed model
Nodes in groups	2,248,710	3,076,787
Links in groups	4,598,683	6,520,113
Number of groups	190,274	280,113
Mean group size	11.81	11.96
Groups density	0.2	0.21
Mean of In Out degree	1.91	1.95

In this experiment, as can be seen in table 5.5.1.1 results for group sizes between 2 and 20 and centrality based on count of transactions, as can be found in the below main points

- Coverage of NW by using our research model increased to be around 76.9 %, in addition SPSS SNA coverage 56.2%.

- Links between nodes in our research are around 6.5 M, but in SPSS SNA 4.6 that means coverage of links increases by 42%.
- Number of groups in our research are around 280K but in SNA SPSS 190K the number of groups increased by 47%.
- The same figures in groups (density, mean in and out) can be found, though the coverage of NW increased.

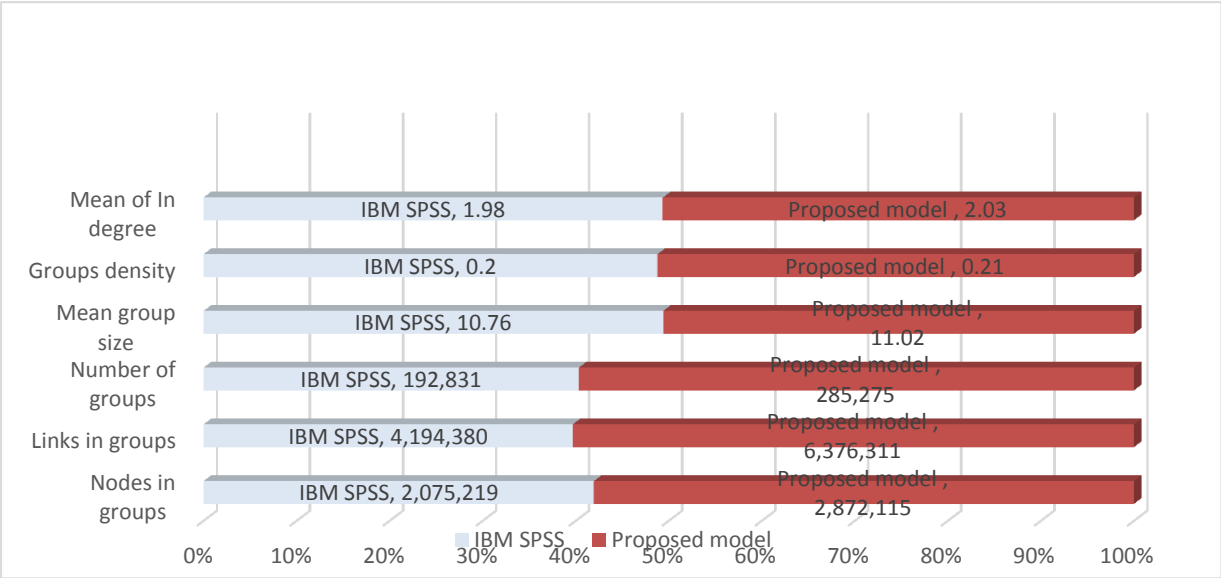


Figure 5.5.1.1 :(Stacked bar)Results for group size between 2 and 20 and centrality based on Count

- As can be seen in figure 5.5.1.1 these measures compare between SPSS SNA and proposed model based on ratio for each one.

Table 5.5.1.2: super groups results for group size between 2 and 20 and centrality based on Count

Measure	Proposed model
Nodes in groups	2,947,609
Links in groups	7,040,243
Number of Extenders	272,987
Mean super group size	148.1
Mean Influencers in super group	13.6
Mean of In degree	2.38

As can be seen in table 5.5.1.2 about super groups results for group size between 2 and 20 and centrality based on count , as can be found that in the below points

- New type of customers extenders was 272 thousand customers
- Mean super group size was148 customers.
- Each super group has around 13.6 groups and influencers
- Coverage of NW that have super groups was73 % .
- Links between nodes that have super groups were 7 M.
- Note that there are main influencers and disseminators in each super group.

5.5.2 Proposed algorithm VS SPSS SNA centrality bases on number of connections (count) and max group size 15

In this experiment, both proposed algorithms and SPSS SNA model were compared. By using different parameters as the following.

- Degree of centrality calculated based on groups sizes that should be minimum 2 and maximum 15 and data should be based on number of transactions (count).
- Closeness centrality calculated based on summation of number of transactions (count) for incoming and outgoing calls and SMS that happened between nodes.

Table 5.5.2.1 : Results for group size between 2 and 15 and centrality based on Count

Measure	IBM SPSS	Proposed model
Nodes in groups	2,134,459	2,918,357
Links in groups	4,155,680	6,252,264
Number of groups	195,274	288,494
Mean group size	10.93	11.08
Groups density	0.2	0.21
Mean of In Out degree	1.92	1.96

In this experiment, as can see in table 5.5.2.1 results for group sizes between 2 and 15 and centrality based on count of transactions, as can be found in the below main points

- Coverage of NW by using our research model increased to be around 72.9 %, in addition SPSS SNA coverage 53.3%.

- Links between nodes in our research are around 6.3 M, but in SPSS SNA 4.2 that means coverage of links increases by 50%.
- Number of groups in our research are around 288K but in SNA SPSS 195K the number of groups increased by 48%.
- The same figures in groups (density, mean in and out) can be found, though the coverage of NW increased.

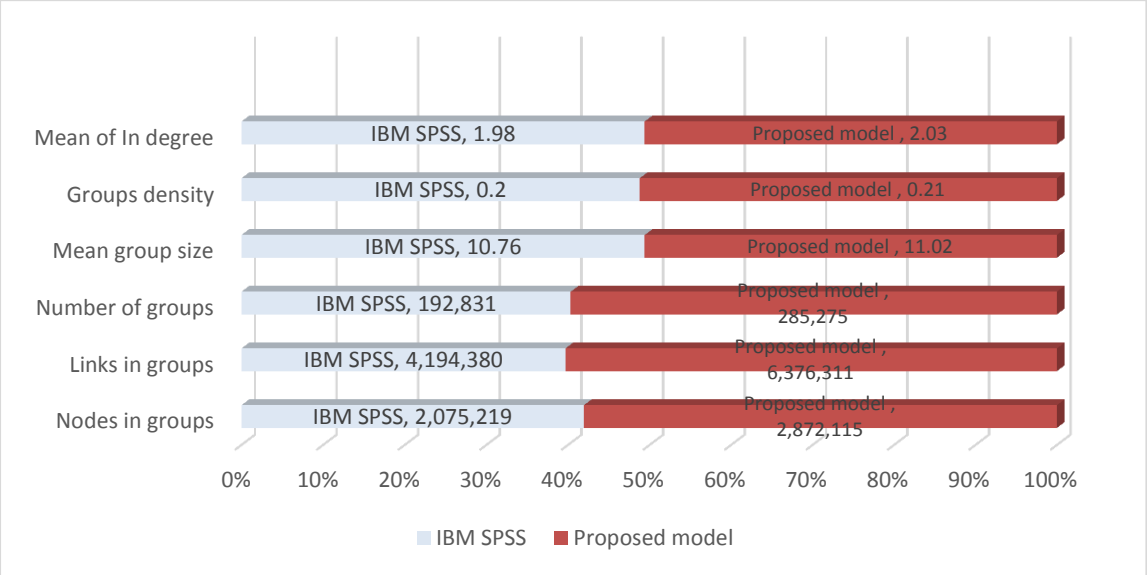


Figure 5.5.2.1 : (Stacked bar)Results for group size between 2 and 15 and centrality based on Count

- As can be seen in figure 5.5.2.1 these measures compare between SPSS SNA and proposed model based on ratio for each one.

Table 5.5.2.2 : Super groups results for group size between 2 and 15 and centrality based on Count

Measure	Proposed model
Nodes in groups	2,641,537
Links in groups	6,541,647
Number of Extenders	277,504
Mean super group size	132.72
Mean Influencers in super group	13.9
Mean of In degree	2.46

As can be seen in table 5.5.2.2 about super groups results for group size between 2 and 15 and centrality based on count , as can be found that the below points

- New type of customers extenders was 277 thousand customers
- Mean super group size was 132 customers.
- Each super group has around 13.9 groups and influencers
- Coverage of NW that have super groups was 66 % .
- Links between nodes that have super groups were 6.5 M.
- Note that there are main influencers and disseminators in each super group.

5.6.1 Proposed algorithm VS SPSS SNA centrality bases on ratio between duration and count and max group size 20

In this experiment, both proposed algorithms and SPSS SNA model were compared. By using different parameters as the following.

- Degree of centrality calculated based on groups sizes that should be minimum 2 and maximum 20 and data should be based on ratio for duration and number of transactions (count).
- Closeness centrality calculated based on summation of ratio for duration and number of transactions (count) for incoming and outgoing calls and SMS that happened between nodes.
- The ratio between duration and count is summation of 0.7* duration +0.3* number of transactions (count).

Table 5.6.1.1 : Results for group size between 2 and 20 and centrality based on ratio between duration and Count

Measure	IBM SPSS	Proposed model
Nodes in groups	2,198,340	3,039,095
Links in groups	4,713,664	6,715,705
Number of groups	186,652	276,430
Mean group size	11.77	11.95
Groups density	0.21	0.22
Mean of In Out degree	1.99	2.03

In this experiment as can see in table 5.6.1.1 results for group size between 2 and 20 and centrality based on ratio between duration and count of transactions, as can be found in the below main points

- Coverage of NW by using our research model increased to be around 75.9 %, moreover SPSS SNA coverage 54.9%.
- Links between nodes in our research are around 6.7 M, but in SPSS SNA 4.7 that means coverage of links increases by 42%.
- Number of groups in our research are around 276K but in SNA SPSS 186K the number of groups increased by 48%.
- The same figures in groups (density, mean in and out) can be found, though the coverage of NW increased.

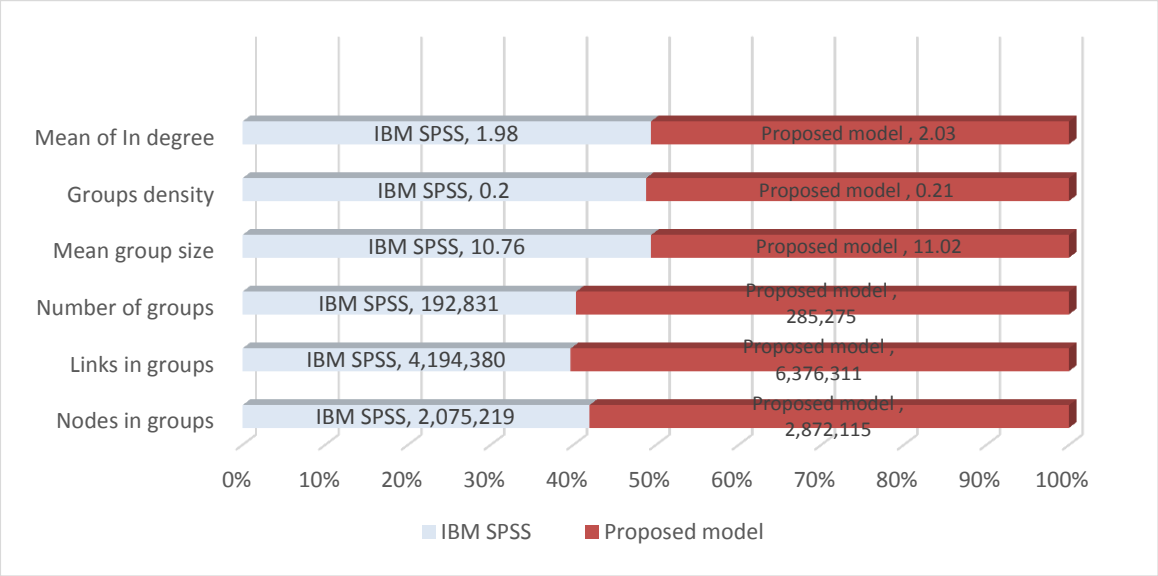


Figure 5.6.1.1 :(Stacked bar) Results for group size between 2 and 20 and centrality based on ratio between duration and Count

- As can be seen in figure 5.6.1.1 these measures compare between SPSS SNA and proposed model based on ratio for each one.

Table 5.6.1.2 : Super groups results for group size between 2 and 20 and centrality based on ratio between duration and Count

Measure	Proposed model
Nodes in groups	2,793,238
Links in groups	7,067,607
Number of Extenders	266,105
Mean super group size	140.35
Mean Influencers in super group	13.3
Mean of In degree	2.52

As can be seen in table 5.6.1.2 about super groups results for group size between 2 and 20 and centrality based on ratio between duration and count , as can be found that in the below points

- New type of customers extenders was 266 thousand customers
- Mean super group size was 140 customers.
- Each super group has around 13.3 groups and influencers
- Coverage of NW that have super groups was 70 %.
- Links between nodes that have super groups were 7 M.
- Note that there are main influencers and disseminators in each super group.

5.6.2 Proposed algorithm VS SPSS SNA centrality bases on ratio between duration and count and max group size 15

In this experiment, both proposed algorithms and SPSS SNA model were compared. By using different parameters as the following.

- Degree of centrality calculated based on groups sizes that should be minimum 2 and maximum 15 and data should be based on ratio for duration and number of transactions (count).
- Closeness centrality calculated based on summation of ratio for duration and number of transactions (count) for incoming and outgoing calls and SMS that happened between nodes.
- The ratio between duration and count is summation of 0.7* duration +0.3* number of transactions (count).

Table 5.6.2.1: Results for group size between 2 and 15 and centrality based on ratio between duration and Count

Measure	IBM SPSS	Proposed model
Nodes in groups	2,075,219	2,872,115
Links in groups	4,194,380	6,376,311
Number of groups	192,831	285,275
Mean group size	10.76	11.02
Groups density	0.2	0.21
Mean of In Out degree	1.98	2.03

In this experiment as can see in table 5.6.2.1 results for group sizes between 2 and 15 and centrality based on ratio between duration and count of transactions, as can be found in the below main points

- Coverage of NW by using our research model increased to be around 71.8 %, furthermore SPSS SNA coverage 51.8%.
- Links between nodes in our research are around 6.4 M, but in SPSS SNA 4.2 that means coverage of links increases by 52%.
- Number of groups in our research around are 285K but in SNA SPSS 192K the number of groups increased by 48%.
- The same figures in groups (density, mean in and out) can be found, though the coverage of NW increased.

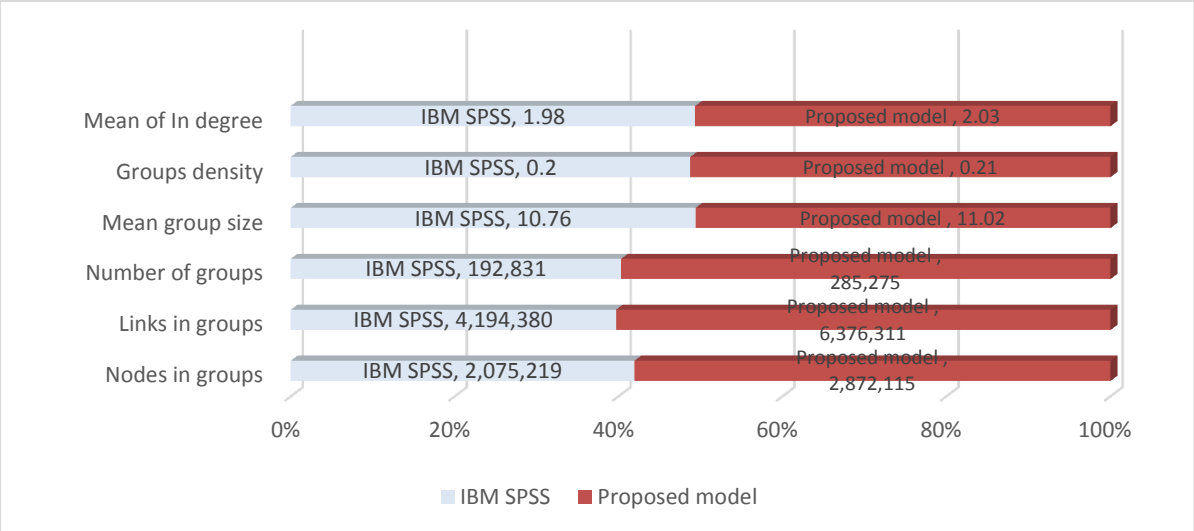


Figure 5.6.2.1: (Stacked bar) Results for group size between 2 and 15 and centrality based on ratio between duration and Count

- As can be seen in figure 5.6.2.1 these measures compare between SPSS SNA and proposed model based on ratio for each one.

Table 5.6.2.2 Super groups results for group size between 2 and 15 and centrality based on ratio between duration and Count

Measure	Proposed model
Nodes in groups	2,473,913
Links in groups	6,490,351
Number of Extenders	271,043
Mean super group size	124.33
Mean Influencers in super group	13.6
Mean of In degree	2.61

As can be seen in table 5.6.2.2 about super groups results for group size between 2 and 15 and centrality based on ratio between duration and count , as can be found that in the below points

- New type of customers extenders was 271 thousand customers
- Mean super group size was 124 customers.
- Each super group has around 13.6 groups and influencers
- Coverage of NW that have super groups was 71 % .
- Links between nodes that have super groups were 6.5 M.
- Note that there are main influencers and disseminators in each super group.

Table 5.6.2.3 Summary of coverage percentage for all experiments

Number Of Transaction To Calculate The Degree Of Centrality				**Duration To Calculate The Degree Of Centrality**				**Ration between duration and number of transactions to Calculate The Degree Of Centrality**			
Group size											
2-15		**2-20**		**2-15**		**2-20**		**2-15**		**2-20**	
Proposed	IBM	Proposed	IBM	Proposed	IBM	Proposed	IBM	Proposed	IBM	Proposed	IBM
72.90%	53.30%	76.90%	56.20%	71.00%	49.70%	75.20%	52.70%	71.80%	51.80%	75.90%	54.90%
Super group(Group size)											
2-15		**2-20**		**2-15**		**2-20**		**2-15**		**2-20**	
66.00%		73.30%		60%		67.00%		62.00%		70%	

As can be seen in table 5.6.2.3 that when the number of transactions were used to calculate the degree of centrality, the highest coverage of the network was retrieved, in the meantime when the duration was used to calculate the degree of centrality, the lowest coverage of network was retrieved. But based on the logical building of groups the duration was more representative of Telecom groups because it gave more valuable groups a higher rank than the communicated groups. Thus, create a balanced ratio between duration and number of transactions was aimed (0.7 for duration, 0.3 for number of transactions).

When the ratio between duration and number of transactions were used, middle coverage for the network coverage with suitable groups structured for the telecom was found and Oracle Sql-PL/SQL was used to implement algorithm. Customized algorithm have applied in Mobily in Saudi Arabia and the same positive results have been found same as Jawwal.

Chapter six

Conclusion

Prior work in SNA algorithm has focused to create a general algorithm to be used for many fields such as (social media, telecom and transportation) , but when these algorithm were used for telecom, the final results were coverage around 55% from input network. However these algorithms have created for general use not specific for telecom companies and Oracle Sql-PL/SQL was used to implement it.

These findings motivate us to customize SNA algorithm to be suitable for the telecommunication companies by using business ideas that coming from this domain. In order to test the new algorithm, Jawwal was used as case study. The main objective of this research was to increase the coverage of NW and to add all high valued customers through the network in the final results. To validate the new SNA model, three different parameters were tested for their degree of centrality(duration, count and mix duration with count), also data was tested based on two maximum groups (size15 and 20) because the average community size for each customer in Jawwal was 18.2 customers.

The result of the experiments when the number of transactions were used to calculate the degree of centrality, the highest coverage of the network was retrieved, in the meantime when the duration was used to calculate the degree of centrality, the lowest coverage of network was retrieved. But based on the logical building of groups the duration was more representative of telecom groups because it gave more valuable groups a higher rank than the communicated groups. Thus, balanced ratio between duration and number of transactions (0.7 for duration, 0.3 for number of transactions) was used. Because the results of it were middle coverage for the network with suitable groups structured for the telecom.
The results of this research increased coverage of NW to be 75.9% instead of 55% and included all high valued customers. These results were approved and the objectives were achieved. Moreover, new novelty ideas have created in this research such as, extenders this type of customers used for customer who is influencer in one group and follower in the other group. Also relation strength used to create groups and assign followers to their most related influencer; furthermore, Super Group used as new layer to connect related groups in one group and find super influencer.

Customized algorithm have applied in Mobily in Saudi Arabia and the same positive results have been found same as Jawwal.

As future work, based on this research results when new parameters were added the results be more efficient and NW coverage was increased; therefore, maybe when added another parameters in the future work, the NW coverage and groups distributions will enhance.

References

[1] M. Alandoli, M. Al-Ayyoub, M. Al-Smadi, Y. Jararweh, and E. Benkhelifa, "Using dynamic parallelism to speed up clustering-based community detection in social networks," Future Internet of Things and Cloud Workshops (FiCloudW), IEEE International Conference on, pp. 240– 245, 2016.. Doi:10.1109/W-FiCloud.2016.57.

[2] F. Ghafoor and M. A. Niaz, "Using social network analysis of human aspects for online social network software," 2016.

[3] A. Jalali, "Supporting social network analysis using chord diagram in process mining," Perspectives in Business Informatics Research, pp. 16–32, 2016.

[4] Jeungmin Lee, Hansaem Park, Kyunglag Kwon, Yunwan Jeon, Sungwoo Jung And In-Jeong Chung(v). An Approach for User Interests Extraction sing Decision Tree and Social Network Analysis. Advanced Multimedia and Ubiquitous Engineering. (pp551-560). Doi: 10.1007/978-981-10-1536-6_72.

[5] I. H. E. adaway Ibrahim S. Abotaleb Eric Vechan, "Social network analysis approach for improved transportation planning," Journal of Infrastructure Systems, 2016.

[6] IBM, "Ibm spss modeler social network analysis 18.0 user guide," 2015. [Online]. Available: www- 01. ibm.com/support/docview.wss?uid=swg27046871

[7] A. P. Rob Cross and S. P. Borgatti, "A bird's eye view: Using social network analysis to improve knowledge creation and sharing," 2010.

[8] W. K. J.C.Thomas and T. Erickson, "The knowledge management puzzle: Human and social factors in knowledge management," IBM Systems Journal, vol. 40, no. 4, 2001.

[9] P. Choudhary, "Ranking terrorist nodes of 9/11 network using analytical hierarchy process with social network analysis," .International Symposium of the Analytic Hierarchy, 2016.

[10] M. L. R. V. Goran Putnik, Ctia Filipa Veiga Alves and V. Shah, "Analyzing the correlation between social network analysis measures and performance of students in social network based engineering education," International Journal of Technology and Design Education, vol. 26, no. 3, p. 413437, 2016.

[11] M. G. Fereshte-Azadi Parand and H. Rahimi, "Combining fuzzy logic and eigenvector centrality measure in social network analysis," International Journal of Technology and Design Education, vol. 459, no. 3, p.2431, 2016.

[12] N. K. Dubey, "Fuzzy and ann based mining approach testing for social network analysis," Second International Conference on Information and Communication Technology for Competitive Strategies, vol. 6, no. 3, pp.48–52, 2016.

[13] W. Maharani and A. A. Gozali, “Collaborative social network analysis and content-based approach to improve the marketing strategy of smes in indonesia,” International Conference on Computer Science and Computational Intelligence, vol. 59, pp. 373–381, 2015.

[14] J. Stevens, “Comparing social networks: Comparing multiple social networks using multi dimensional scaling,” DOI: 10.4256/mio.2010.0012. 2010.

[15] J. Roberts. (2015) A brief introduction to social network analysis. [Online]. Available: http://web.mit.edu/vdb/www/6.977/l-jenn.pdf

[16] D. Conway, “Social network analysis in r.new York university -department of politics,” 2009. [Online]. Available:http://files.meetup.com/1406240/sna_in_R.pdf.

[17] G. Zhang, “Social network analysis with r sna package,” 2014. [Online]. Available: http://files.meetup.com/1406240/sna_in_R.pdf

[18] S. Hornibrook and Charlotte, “Degrees. of separation: Social network analysis using the sas system,” 2007.

[19] E. Darom, “Social network analyses with sas/iml.barllan university,” 2012. [Online]. Available:http://www.sascommunity.org/seugi/SEUGI1997/DAROMDATAMIN.PDF

[20] D. S. A. Hale, “Network analysis, oxford internet institute,” 2016.[Online]. Available: http://blogs.it.ox.ac.uk/engage/files/2016/01/Hale-Engage-Impact.pdf

[21] M. Ponce, “A basic introduction to complex networks, using python,” 2015. [Online].Available:https://wiki.scinet.utoronto.ca/wiki/images/f/fb/PDlecture7-ComplexNetworkswPython.pdf

[22] D. McKenzie, “Social network analysis,” 2010. [Online]. Available:https://www.sas.com/content/dam/SAS/en_ca/User%20Group%20Presentations/Toronto-Data-Mining-Forum/McKenzie-SNA-Spring2010.pdf.

[23] E. E.-Z. M. G. David Combe, Christine Largeron, “A comparative study of social network analysis tools,” Laboratoire Hubert Curien, vol. 551, 2010.

[24] IBM, “Ibm spss modeler social network analysis 15 user guide,” 2012. [Online].Available:ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/SNA_UserGuide.pdf

[25] O. Serrat, “Social network analysis,” 2009. [Online]. Available: http://betterevaluation.org/sites/default/files/social-network-analysis.pdf

[26] S.Ponce, “The overseas development institute’s research and policy indevelopment,”2011.[Online].Available:http://www.fao.org/elearning/course/FK/en/pdf/trainerresources/PG_SNA.pdf

تخصيص تحليل الشبكات الاجتماعية لشركات الاتصالات

اعداد : أكثم فهيم عقل صوان
اشراف : د. رشيد الجيوسي

ملخص:

يتم إنشاء تحليل الشبكة الاجتماعية (SNA) لتحليل بيانات الشبكة الاجتماعية؛ وبالتالي، فإن الشركات الرئيسية في مجال استخراج البيانات (مثل IBM, SAS, R and python) خلقت نماذج (SNA) خاصة بها. الهدف من هذا البحث، انشاء خوارزمية مخصصه لشركات الاتصالات لان الخوارزميات الحالية ليست مخصصه فقط لشركات الاتصالات, لذلك استخدم في الخوارزمية الجديدة قوة العلاقة والموسع لتحسين النتائج النهائية.

300 مليون سجل التي تنتمي إلى حوالي 4 ملايين مشترك للاشهر الثلاث الاخيرة جمعت من (شركة جوال للاتصالات) كدراسة حالة. الخوارزمية الجديدة والخوارزميات الحالية استخدمت نفس المعلومات. في هذا البحث ست تجارب قد اجريت بناءً على مدة المكالمة ,عدد المكالمات ونسبة بين مدة المكالمة وعدد المكالمات, بالاضافة الى حجم المجموعات الذي استخدم هو (15 و 20). النتائج التي اعتمدت من شركة جوال كانت بناءً على المحددات التي استخدمت في التجربة السادسة (نسبة بين مدة المكالمة وعدد المكالمات مع حجم مجموعة يصل الى 20 كاعلى تقدير, النتائج زادت من تغطية (NW) لتصبح بنسبة 75.9٪ بدلا من 55٪ في الخوارزميات الحالية. بالاضافة ان نتائج الخوارزمية الحديثة تشمل جميع العملاء ذوي القيمة العالية.

افكار ابداعية جديدة تم استخدامها في هذا البحث مثل المشترك المتوسع وهو المشترك الذي يكون تابع في مجموعة و مؤثر في مجموعة اخرى.ايضاً وقوة العلاقة بين العملاء تستخدم لخلق المجموعات ووضع المشترك التابع مع اكثر مشترك مؤثر له علاقة معه.بالاضافة الى المجموعة الاقوى هي طبقة جديدة لربط المجموعات ذات الصلة في مجموعة واحدة وايجاد المؤثر الاقوى.