Social Network Analysis Framework in Telecom

Mona Saleh Amer
Acting ERP Department Head, Information Technology Institute (ITI)

***Abstract*—**this paper is based on a project prototype that uses Social Network Analysis (SNA) in the telecommunication industry to detect the communities of subscribers (business, friends and family communities), identify the most influential customers in the network who can spread positive or negative messages through the network and recommend best customer acquisitions to be targeted by marketing campaigns. The project is based on a real dataset of 100,000 customers of an Egyptian Communication Service Provider (CSP). It is implemented by two Technical professionals working for an Egyptian CSP, seventeen students and two technical professionals working for Information Technology Institute (ITI). The objective of this paper is to present a prototype that can be used by telecom operators to increase cross-selling and up-selling products and reduce marketing costs using SNA with very low cost compared to available SNA market products. The proposed business value of this prototype is believed to be:

- Optimized marketing campaigns by targeting the segment of customers that is most likely to be interested in a product/service instead of targeting the whole bases.
- Increase cross-selling and up-selling of products and services by utilizing existing customers' word of mouth using viral marketing.
- Better customer value management and increased customer satisfaction.

***Keywords*—** Acquisition, Big Data, Influencer, Social Network, Social Network Analysis, SNA, Telecom

I. INTRODUCTION

SNA is a methodology developed by sociologists and researchers in social psychology and further developed in collaboration with mathematics, statistics, and computing [1].

SNA is based on an assumption of the importance and effectiveness of relationships among interacting nodes (people, products …etc.). It is used to analyze these relationships which are important to discover the structure and interdependencies of individuals, products or organizations. SNA also tests the strength and effectiveness of a relationship. It allows managers to visualize and understand the great number of relationships that can either facilitate or impede knowledge creation and transfer. SNA can utilize any data source that includes interactions between entities, such as interactions between human or bank account.

In telecommunication industry, SNA uses the call detail records (CDRs) of the CSP in conjunction with other data sources like customers' profiles to analyze the social relations of their customers. Linking this information together with SNA leads to better insights and values that affect significantly revenue and customer satisfaction.

This paper has the following sections: Section 2, in this section we present a literature review on the usage of SNA in different domains and list some of the commercial SNA solutions that is used in telecommunication industry specifically. Section 3 represents the work methodology. Section 4 describes the techniques and algorithms used in this paper and introduces an overview of the solution proposed to get the communities, influencers and customer acquisitions. Section 4 also discusses the results of applying these algorithms with some explanatory figures. It also includes a function that we manually developed to calculate the acquisition scores. Section 5 describes the conclusion derived from this paper and Section 6 contains some of the suggested areas of future research.

II. PREVIOUS WORK

SNA has proved its importance by its high penetration rate in most of industry verticals such as Retail, Online Gaming and Finance and Banking Sectors [4]. "Since 1999 there seems to be a rapid increase in the number of empirical studies employing network analysis, thus looking at the detailed structuring of the relationships between organisations/individuals and at the impact of network structure on performance" [7]. It is also an important tool for use in preparing the prosecution of criminal cases [5] and can be used to analyze networks of organizations associated with terrorist activity [6] and even more to analyze the networks of individuals in an organization itself which is called Organizational Network Analysis (ONA) [1] [2].

Since SNA plays great role in characterizing network structure and relationships among these networks, it has become of high importance for CSPs for its great impact in the Return of Investment (ROI) of these operators. Therefore many SNA products were developed to serve the telecommunication industry specifically and other industries generally. Here we list some of the leading SNA products that serve the telecommunication industry.

1) Social-3

This product offers business solutions in several industry sectors, including Telecommunications, Retail and Insurance [8]. The product features the following options:

- Prediction of Churn and Customer Actions
- Influencer Marketing
- Prediction of upcoming Tweet Crisis
- Fraud Scoring

2) iDiro Analytics

Idiro provides predictive analytics for companies with large consumer networks. The platform is a scalable solution capable of analyzing billions. iDiro provides features for:

- Customer retention & acquisitions
- identifying families and households
- identifying rotational churn

3) IBM SPSS

IBM SPSS is predictive analytics software that is part of the IBM Business Analytics portfolio. Along with the many standard nodes delivered, new nodes are added in IBM SPSS Modeler 15 which are Social Network Analysis nodes that include the results of social network analysis in the analysis streams [10]. These nodes are:

- Group Analysis: it imports CDR data from a fixed-field text file, identifies groups of nodes within the network defined by the records, and generates key performance indicators for the groups and individuals in the network.
- Diffusion Analysis: it imports call detail record data from a fixed-field text file, propagates an effect across the network defined by the records, and generates key performance indicators summarizing the results of the effect on individual nodes.

4) SAP

InfiniteInsight is SAP Predictive Analytics software developed by KXEN which was acquired later by SAP. SAP Predictive Analytics has a social component for SNA which can provide predictive insights to [11] [12]:

- Gain end-to-end social network analysis capabilities
- Use powerful visualization capabilities and graph exploration
- Natively integrate social attributes into predictive models
- Identify "influencers" and node segmentation for specific domains
- Detect hidden links in your data and connect individuals using multiple identities

III. Work Methodology

3.1 Solution Architecture

The infrastructure used is a big data cloud based solution which contains three layers (see Fig.1):

- Storage layer: where data is being loaded to a relational database management system (RDBMS) through the extract transform and load (ETL) process
- Analysis layer: this layer uses Apache Spark for the analysis and mining models which are coded using R programming language. The output is stored in the storage layer for further analysis or visualization.
- Visualization layer: this layer uses Business Intelligence (BI) software to provide insights on the results of analysis using reports and dashboards.

3.2 Analysis

The SNA is based on building a network of nodes (represented by customers) which are connected by links (represented by calls/sms or interactions). The network is built using the adjacency matrix and centrality measures. The following sub-sections describes the methodology and models used to identify communities, influencers and pinpointing new customer acquisitions

3.2.1 Community detection

The identification of communities is done in two phases, community labeling and community detection.

3.2.1.1 Community labeling

Community labeling is identifying each interaction (link) between customers (nodes) as a business, family or friend link. This is done using descriptive K-means clustering algorithm shown in Fig. 2. Links- which are read from CDR- between

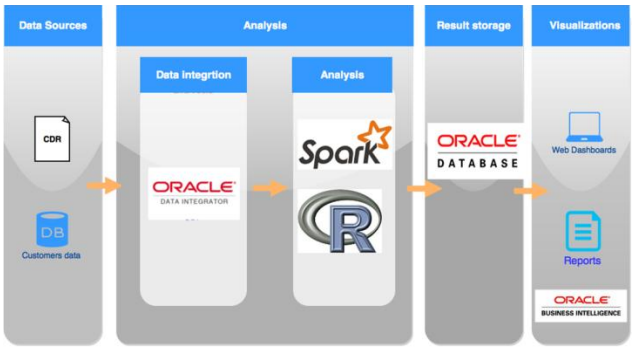

Figure 1- Solution Architecture

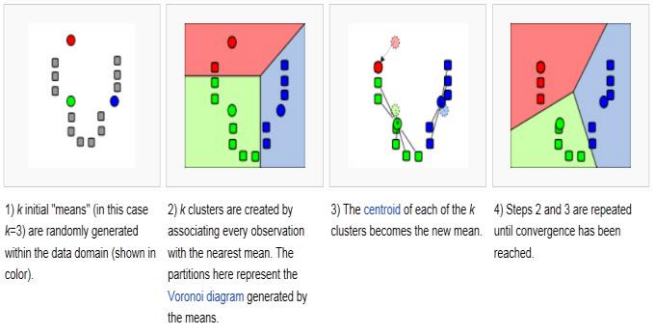

Figure 2- Demonstration of K-means standard algorithm [13]

each two nodes are filtered based on regularity and strength to exclude links that doesn't represent a relation, happened once or by coincidence. The algorithm uses these links along with some other fields from the customer profiles as input to k-means. Links are categorized based on: timing of day (early morning, late morning, peak hours …etc.) and average duration of all calls during a week day and a week end. Each link is then assigned a label (family, friend or business) by setting rules that is based on domain knowledge combining these variables (for example, a business link happens mostly during the working hours of a week day, and so on). Location data is also used to differentiate family and friends links.

The output of this step is to label each link in the network with family, business or friend link.

3.2.1.2 Community detection

In this step, the three communities of each node are detected. This step uses algorithm based on the fast unfolding technique and modularity of the nodes. The algorithm is built on top of the community labeling. The weight of links is calculated and used as input for this algorithm. The output of this algorithm is a membership vector that identifies the community ID to which every node belongs.

3.2.2 Influencer detection

Influencers are considered to be the most connected customers in the network. In this step, the influence of each node in the network is represented by score. The most influential nodes are the highest in score. To calculate the score, page rank algorithm is used which is applied on top of community detection phase. The page rank algorithm has two main inputs which are the graph and weights on the links in the graph. The graph is built from the CDR using the adjacency matrix. The output of this step is the influence score of each node which is then saved to the storage layer.

3.2.3 Customer Acquisition

The objective of this this step is to pinpoint all real potential off net customers for joining your network to be later targeted in campaigns. This is done by calculating an acquisition score for each node based on its links and usage with other on net members. It also classifies all off net customers to categories; for example (High potential, ideal, Low potential). A scoring function that calculates the acquisition score is developed based on some parameters and weights like regularity and strength and direction of the link and duration. The output of this step is the acquisition score on two levels:

- Score on the link level between on-net and other off-net node
- Score on the node level aggregated for all scores on all links between on-net all nodes and one off net node.

IV. Proposed Solution and Experimental Results

The project is a prototype for an end-to-end solution that is used by decision makers to identify communities, influencers and new customer acquisitions. The solution includes all processing phases (ETL, Mining/Analysis and Visualization) required to produce the desired output.

The objective of this paper is to present a prototype that can be used by telecom operators to increase cross-selling and up-selling products and reduce marketing costs using SNA with very low cost compared to available SNA market products.

After applying the SNA on a sample of 100000 subscribers' real CDR data of an Egyptian CSP, three main outputs are resulted:

- List of all nodes who belong to each node's business, family or friends community
- List of all nodes and their influence score in the network
- List of all off-net nodes and a score representing their potential for joining the network

3.3 Community detection

After applying the clustering algorithm to label each link between nodes, the output which represents three clusters is then visualized to determine the clusters' characteristics. The output of this phase is a variable assigned to each link that indicates the community type this link belongs to. Visualization output of business cluster variables is shown in Fig. 3.

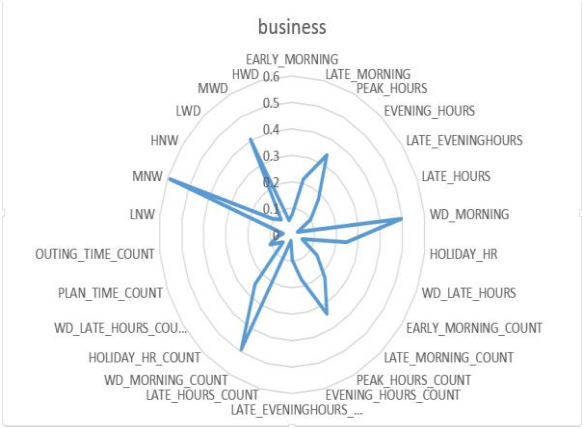

Figure 3- Business Community- the business links are dragged in the direction of early morning, late morning, peak hours and low and medium duration

3.4 Influencers

After community labeling is completed, the output is used as an input to the page rank algorithm which is applied to detect the influence score of each node in the network. The page rank algorithm has two main inputs which are the graph and weights on the links in the graph. Graph is built by constructing the adjacency matrix from the data file. Weights are calculated based on the count and duration of interactions which indicate the regularity and strength of the links. The algorithm output is a score that represents the influence of this node in the network.

3.5 Acquisition

The acquisition score is calculated based on the following technique: All CDR transactions that contain off net numbers are pinpointed. Only transactions that are related to connections with high regularity and strength are filtered. Regularity is a connection that occurs at least in 3 different weeks. A connection is considered strong when the number of interactions is at least three in the duration of 3 months. A function is developed to calculate the acquisition score based on multiple variables is shown in Equation 1.

$$Link\ score = (.05/12) * V1 + (1/12) * V2 + (1/12) * V3 + (.05/12) * V4 + (.025/12) * V5 + (3/12) * V6 + (2/12) * V7 + (1/12) * V8 + (1/12) * V9 + (.05/12) * V10 + (1/12) * V11 + (.025/12) * V12 \quad (1)$$

Table 1- Scoring Function Inputs Description

Scoring Function Inputs Description	
Variable	Description
(V1)	All the outgoing calls that the subscriber make from total number of calls
(V2)	All the incoming calls that the subscriber make from total number of calls
(V3)	{(count total high duration calls)/(total outgoing duration)} * (sum of high outgoing duration)
(V4)	{(count total medium duration calls)/(total outgoing duration)} *(sum of medium outgoing duration)
(V5)	{(count total short duration calls)/(total outgoing duration)} *(sum of low outgoing duration)
(V6)	{(count total long duration calls)/(total incoming duration)} *(sum of high incoming duration)
(V7)	{(count total medium duration calls)/(total incoming duration)} *(sum of medium incoming duration)
(V8)	{(count total short duration calls/(total incoming duration)} * (sum of low incoming duration)
(V9)	The percentage of outgoing messages from all messages
(V10)	The percentage of incoming messages from all messages
(V11)	The percentage of calls from all events (messages and calls)
(V12)	The percentage of SMS from all events (SMS and calls)

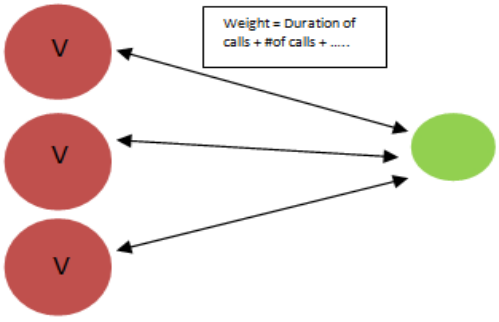

Figure 4- Scoring function weight on node and link level

The output will be acquisition score on two levels as shown in Fig. 4:

- Score on the link level will be between on-net and other off-net node.
- Score on the node level will be aggregated for all scores on all links level that associated between all on-net nodes and one off net node.

V. Conclusion

This paper presents a prototype that can be used by telecom operators to increase cross-selling and up-selling products and reduce marketing costs using SNA with very low cost compared to available SNA market products. The main usage of these outputs is to help decision makers in telecom industry to study behavior of their customers to plan for the right offers and utilize network usage to guarantee the best service for their current users using different dashboards and BI reports that clarify mining and statistics results.

VI. Future Work

As mentioned, this project prototype used RDBMS as storage layer, Spark and R programming as computational and analysis layer and a commercial BI tool as a visualization layer. A future improvement will be implementing this on a NoSQL (Not only SQL) database that can accommodate the huge amount of data generated in the CSP and thus suits for faster analysis and retrieval. Furthermore, a churn prediction in a network can utilize the already built network and be added.

Acknowledgement

The author has great acknowledgement to the team who participated and was the major factor of the success of the project which this paper is based on. This acknowledgement goes for Vodafone VIS professionals Mohamed Nayer and Yara Maher. Also goes for ITI, the teaching assistant Rana Salah and the great team who implemented the project; the graduates of the Data Warehousing and Business Analytics track at ITI Intake 35.

References

[1] Scott, 2000, Social Network Analysis, A Handbook, Sage.

[2] http://www.gartner.com/it-glossary/social-network-analysis-sna Gartner Research, last visited on 16th Oct 2015

[3] Charlotte Patrick, Competitive Landscape: Vendors Providing CSP Social Network Analysis, Gartner, 2011.

[4] http://www.frost.com/prod/servlet/press-release.pag?docid=257217740 Frost & Sullivan Research, last visited on 16th Oct 2015

[5] Natarajan, M. (2006). Understanding the Structure of a Large Heroin Distribution Network: A Quantitative Analysis of Qualitative Data. Journal of Quantitative Criminology, 22(2), 171-192

[6] Basu, A. (2005, June). Social Network Analysis of Terrorist Organizations in India. Paper accepted for presentation at the North American Association for Computational, Organizational and Social Science (NAACSOS 2005), Notre Dame, June 2005

[7] Fabrice Coulon, (2005, Jan). The use of Social Network Analysis in Innovation Research: A literature review

[8] http://www.social-3.com/ Social-3 social analytics company, last visited on 16th October 2015

[9] http://idiro.com/services/#section-sna Idiro analytics, last visited on 16th October 2015

[10] ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/SNA_UserGuide.pdf IBM SPSS Modeler Documentation last visited on 16th October 2015

[11] http://www.sap.com/pc/analytics/predictive-analytics/software/predictive-analysis/index.htmlhttp:.html last visited on 16th October 2015

[12] https://en.wikipedia.org/wiki/KXEN_Inc. last visited on 16th October 2015

[13] https://en.wikipedia.org/wiki/K-means_clustering last visited on 16th October 2015