Al-Molhem *et al. J Big Data* (2019) 6:99
https://doi.org/10.1186/s40537-019-0264-6

Journal of Big Data

RESEARCH Open Access

Social network analysis in Telecom data

Nour Raeef Al-Molhem*, Yasser Rahal and Mustapha Dakkak

*Correspondence:
nour.almolhem@hiast.edu.sy
Faculty of Information Technology, Higher Institute for Applied Sciences and Technology, Damascus, Syria

Abstract

Many systems can be represented as networks or graph collections of nodes joined by edges. The social structures in these networks can be investigated using graph theory through a process called social network analysis (SNA). In this paper, networks and SNA concepts were applied using Telecom data such as call detail records (CDRs) and customers data to model our social network and to construct a weighed graph in which each relation carries a different weight, representing how close two subscribers are to each other. In addition, SNA is used to explore the Telecom network and calculate the centrality measures, which help determine the node importance in the network. Depending on centrality measures as well as influence capability of node measure, the influencers in network were detected and targeted by marketing campaigns resulting in 30% raise in growth rate of mobile traffic compared with traditional ways. Finding Multi-SIM subscribers within the same operator or across different operators presents another important concern to Telecom companies because it allows to improve campaigns and churn prediction models. Social network similarity measures and social behavioral measures between nodes were calculated in the Telecom network to detect these Multi-SIM subscribes and 85% accuracy result was achieved for subscribes from different operators and 92% for subscribes from the same operator. The paper is based on a real dataset of 3 months CDRs and customer data provided by a local Telecom operator. This dataset is used to build a network with more than 16 million nodes and more than 300 million edges on a big data platform.

Keywords: Influencers, Similarity, Telecom social network analysis, Multi-SIM, CDR, Big data

Introduction

The process of defining social communication based on both the network and the graph theory is known as social network analysis (SNA) [1]. SNA is used to analyze relationships among interacting nodes (customers, products ...etc.) which is important to discover the structure and reliance of individuals or organizations. This approach has now become a powerful tool to study networks in various areas like banking, Telecom, web applications, physics and social science. Relationships in a network can either be directional or non-directional. In a directional relationship one person is the initiator (or source of the relationship) while the other is the receiver (or destination of the relationship), a weight indicating the strength of the relationship can be added. In Telecom domain, SNA uses the CDRs of the operator [2, 3], which is a data record produced by a telephone exchange. This data record documents the details of any Telecom transaction (calls, SMS , internet ...) that passes through the

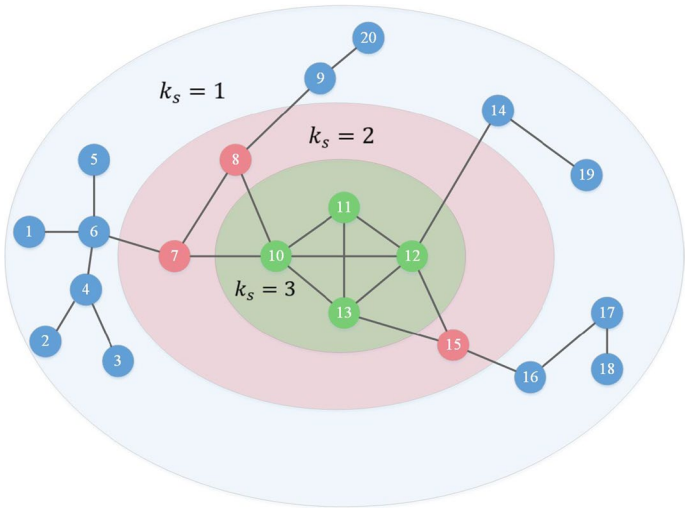

Fig. 1 Sample of decomposition by *k-shell* method

mobile devices. In addition to CDRs, SNA uses other data sources like customers data to analyze the social relations of subscribes. Linking this information together with SNA gives better insights and values that affect the revenue and the customer satisfaction [4]. CDRs and customers data gathered for 3 months were used in our work to implement the data ETL (Extracting, Transforming and Loading) and the data summary to build a social network where nodes represent GSM numbers of subscribers and edges represent interactions between subscribers (calls in our case).

An influencer in network is defined as a node which is well connected. It is capable of propagating information to lots of people [5]. In Telecom domain influence subscribers are usually considered to be well-connected to other subscribers in network. This good connection guarantees low risk for churn but high potential for diffusion of products and services [6]. Finding the actual influencers in Telecom field is a challenging topic and it depends largely on intuition, personal experiences and how the influencer is interpreted. In our work eigenvector centrality [7] is used to measure the importance of subscribers [8]. This centrality indicates that one node's importance is determined by its neighbors' importance as well as number of nodes linked to this node. In addition to eigenvector centrality, *k-shell* decomposition method which is a fast node ranking method for large-scale networks [9] is also used in detecting influencers, to identify the most influential nodes located in the core of the graph. It partitions a graph into sub-structures that are directly linked to centrality; Fig. 1 represents a sample of decomposition by *k-shell* method.

This method assigns an index k_s to each node. This index represents the location of the node in the graph and depends on the node degree [10] which is the number of edges the node has. First, all nodes with degree k = 1 are removed recursively from the network and a $k_s = 1$ value assigned to them. This procedure is repeated iteratively until only nodes with degree $k > 1$ are left in the network. All previous steps are repeated with changing of values of k and k_s until all nodes in the network

Table 1 The iteration process of k-shell decomposition

Node degree	Deleted nodes	Iteration order	k_s
1	1 , 2 , 3 , 5 , 18 , 19 , 20	1	1
1	4 , 9 , 14 , 17	2	1
1	6 , 16	3	1
2	7 , 8 , 15	1	2
3	10 , 11 , 12 , 13	1	3

are assigned to one of the *k-shell* values. Table 1 shows the iteration process and the assigned k_s values of each node of the network in Fig. 1.

We used *k-shell* values to calculate the influence score for each node in the network. The most influential nodes are the highest in score.

A subscriber in a Telecom operator might possess more than one line within the same operator or across different operators. Multi-SIM subscribers with different operators have a higher potentiality to churn than ordinary subscribers. They cost the operator to lose the chance of initiating and receiving all their traffic. Detecting the Multi-SIM subscribers across different operators allows for even more usage profiling that will help create more tuned campaigns and enrich churn prediction models. Using the social dimension manifested in the relationships among customers provided by SNA, we built a strong Multi-SIM detection model that guarantees more stability against seasonal changes and traffic usage fluctuations, and ensures linking the other dials belonging to the single customer within the same operator and across different operators.

The Telecom operators have massive data resources such as CDRs, customers profile data, customers location data and so on. How to effectively store, parse and analyze this amount of data represent the main challenge that faces Telecom operators especially when new techniques such as SNA and machine learning algorithms, which require huge amount of memory and distributed processing solutions, are applied. Here comes the need of a big data platform that can effectively processes data of varying sizes and complexities, facilities the calculations and reduces the processing time [11]. A big data platform which contains various tools and capabilities was chosen to handle the previous challenges in Telecom data.

One of our primary motivations in this paper is to present a new way to build a large-scale social network from Telecom data using the big data platform. This platform is designed as a solution to the challenges that we usually face in extracting the SNA measures in large-scale networks. Another motivation of our research is to provide models based on real data that can help decision makers in Telecom companies to plan for the right offers and targeting the most influence customers in the network, which increase cross-selling and up-selling products and reduce the marketing costs.

The rest of the paper is structured as follows: In "Related work" section, we present related works on using SNA in Telecom domain; "Methods" section describes the data set we used in this research as well as how we built Telecom network and feature extraction methods to determine Influencers and Multi-SIM subscribers. In "Results and discussion" section, we describe our results and evaluate our approach. Finally, we conclude our work and describe future work in "Conclusion" section.

Related work

The majority of approaches depending on CDRs to build the Telecom social network in related work, where nodes (represented by customers) are connected by links (represented by calls/SMS's or interactions). Onnela et al. [12] analyzed a weighted call graph of mobile phone call records by examining its degree, strength, weight distributions, clustering and weighted clustering, together with correlations between these quantities. Nanavati et al. [13] analyzed the graph properties such degree distribution and neighborhood distribution over time of calls and SMS networks. Mona et al. [14] proposed a prototype that uses SNA to detect the communities of subscribers using two phases. The first phase is community labeling by K-means algorithm [15] and the second phase is community detection as a membership vector that identifies the community ID to which every node belongs. They identified the most influential customers who can spread positive or negative messages through the network using PageRank algorithm [16] and recommended the best customer acquisitions to be targeted by marketing campaigns. Nattapon et al. [17] proposed a data cleansing process for CDR in order to filter the anomaly numbers. Moreover, they invented a measure to capture influencers based on calling behaviors; their experiment was conducted on CDR's of a Telecom operator in Thailand. Wang et al. [18] proposed *k-shell* iteration factor which is a novel node ranking measure to quantify the influence capability of nodes. This factor utilizes the iteration information of *k-shell* decomposition to distinguish the influence capability of nodes with the same *k-shell* value, which can help to discriminate the influence capability of nodes more accurately and provide a more reasonable ranking list than other measures. Ahmed et al. [19] proposed a model based on SNA, which represented the data warehouse that is continuously fed by switches and charging/billing systems. Centrality measures and behavioral attributes were calculated to facilitate equivalence analysis between each pair of nodes. They also analyzed the way each node in the pair interacts with other nodes in network basing on link attributes. In addition, they came up with a score that was used in clustering the customers into Multi-SIM probability clusters for the marketers to target. Zhan et al. [5] proposed a new way of calculating the most influential top-K communities in large networks using Katz centrality [20], which measures the relative influence of each node in network by taking into account the node's immediate neighbors. They calculated the average of Katz centrality for all communities instead of using the traditional centrality measures.

Customer churn prediction models aim to detect customers with a high propensity to leave the company [21], these churn prediction models have been widely used in the Telecom companies to identify customers who are likely to churn and provide suitable intervention to encourage them to stay [22]. Also, SNA features were used to enhance the results of churn prediction models in Telecom domain after representing CDRs data as a graph. Dasgupta et al. [23] used the SNA to study the evolution of churners in the network over a period of time and explored the propensity of a subscriber to churn depending on the number of friends who have already churned, Ahmad et al. [24] also used SNA features to enhance the results of predicting the churn.

Sorić et al. [25] presented a prototype platform for SNA analysis in big data environment and gave an overview of the architectural integration in combination with multiple technologies, frameworks and techniques through big data architecture. Moreover, they

used a platform of a combination of Spark GraphX[1] framework and JavaScript[2] for an efficient social network analysis with different types of powerful and interactive visualizations. Brdar et al. [26] provided an overview of all steps in discovering knowledge from raw telecom data in the context of different applications, they also presented a discussion about approaches that are analyzing mobile operators' data sets via graph theory and machine learning.

All previous related works are based on the calculation of centrality measures (Degree, Closeness [10]: indicates how close a node is to all other nodes in the network, PageRank, Katz centrality, ...) to detect influencers in small or medium scale networks, but in large-scale networks some of these measures, such as Closeness and Katz centrality, become insufficient because they are complex and difficult to calculate. Also, there are some measures that incompatible with the nature of Telecom data such as PageRank which is created for ranking web pages. Our contribution in this paper comes through presenting new measures for detecting influencers that can be applied in large-scale networks and compatible with Telecom data. In addition, we have tried to improve the methods of detecting Multi-SIM customers by applying social network concepts which considered a new way in this area. Previous studies in this domain are rare and limited in detecting customers from the same operator using customers data only, while with the new proposed method we can find Multi-SIM customers within the same operator and across different operators by using the social network and customers data.

Our research is considered as the first of its kind that uses a real and big Telecom dataset in Syria. The paper proposes a novel approach to detect influence subscribers in the Telecom social network, this approach is depending on calculating eigenvector and influence capability of node for each subscriber. The new approach is more accurate and efficient than traditional methods that using only centrality measures as we'll see in "Results and discussion" section. Another new approach is presented in this paper: Multi-SIM subscriber detection model, which is based on the idea of similarity between nodes in the graph plus the mutual behavioral characteristics between customers.

Methods

This section describes the data sets used in this work, how we built social network and the feature extraction methods.

Solution architecture

We have chosen Hortonworks Data Platform (HDP)[3] as a big data platform to install and use in the study. HDP is a free open-source framework under the Apache 2.0 License[4] designed to deal with data from different sources and formats. It has a variety of open source systems and tools such as:

[1] https://spark.apache.org/graphx/.

[2] https://www.javascript.com.

[3] https://hortonworks.com/.

[4] https://www.apache.org/licenses/LICENSE-2.0.

- Hadoop Distributed File System (HDFS)[5]: it is a Java-based, file system for storing large volumes of data; it provides scalable and reliable data storage.
- Apache YARN[6]: it represents the processing layer for managing distributed applications that run on multiple machines in a network, it allows using various data processing engines for batch, interactive and real-time stream processing of data stored in HDFS, so YARN provides resource management while HDFS provides storage.
- Apache Spark[7]: A distributed, in-memory data processing engine designed for large-scale data processing and analysis.
- Apache Zeppelin[8]: A web-based notebook which supports interactive data exploration, visualization and collaboration.

We stored Data in HDFS as a spark DataFrame[9] format which is a Dataset organized into named columns; it is similar to data frame in R/Python or a table in a relational database, but with more optimizations. We used Spark tools for processing data, building the telecom social network and calculating SNA features.

Data description and preparation

Four data sources were selected for our work:

1. Customer data: The customer data has been collected from CRM system, it contains customer contract information (subscriber GSMs, subscription type, age, gender, location ...). In addition to the above, the customer data also contains all services, offers, packages that were subscribed by the customer.
2. Mobile IMEI information: It contains all information about the customer mobile devise such model, brand and if the device is dual-SIM or not.
3. Call details records (CDRs): It contains all transactions and actions that were taken by the customer; we selected for our work only data of calls. This data source is generated as text files.
4. Cells and towers information: it contains the information of actions location like longitude and latitude coordinates, sub-area, area and city.

We have exerted great effort in the process of collecting and clearing previous sources of information due to the large volume of data and the variety of its sources. In addition to collecting and clearing data, we had to understand and link all types of data so that they can be used for our research. In the end, 3 months of data sets were prepared which contained more than 10 million customers with their data and about one billion records of calls between customers.

An ETL on CDRs were performed in a duration of 3 months and prepared using big data platform HDP. Before the storage operation, the data were cleansed by eliminating the irrelevant numbers from CDRs. These numbers were classified in three types: first

[5] https://hadoop.apache.org/docs/r3.1.1/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html.

[6] https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html.

[7] https://spark.apache.org/.

[8] https://zeppelin.apache.org/.

[9] https://spark.apache.org/docs/latest/sql-programming-guide.html.

Table 2 Sample of detailed data

Calling number	Called number	Call duration (s)	Start call time
963-9******08	963-9******92	330	21/11/2018 08:30:47 AM
963-9******43	963-9******17	138	21/11/2018 11:23:52 AM
963-9******21	963-9******49	212	21/11/2018 04:12:31 PM
963-9******76	963-9******19	99	21/11/2018 10:36:11 PM

Table 3 Sample of summarized data

Calling number	Called number	Calls duration (s)	Calls count
963-9******14	963-9******22	4500	22
963-9******31	963-9******62	3257	18
963-9******94	963-9******11	7368	38
963-9******34	963-9******78	5327	29

type is call center numbers which are numbers that receive a lot of calls from massive amount of other numbers but do not themselves make any calls. Second type is Telesales numbers (the opposite behavior of call center numbers) which are numbers make a lot of calls to a lot of numbers and don't receive many calls back. The last type is the wrong calls numbers which are numbers that received one call with short duration and didn't make any calls at all.

After cleansing operation, we stored two types data in HDFS:

- Detailed data: contains all calls in each day with time of the call and the duration. This type of data is used to extract Multi-SIM subscribers model features. Table 2 shows a sample of detailed data.
- Summarized data: contains the aggregation of detailed data over 3 months considering direction of calls, each record contains calling and called part with number of all their calls and total calls duration in the whole period. This type of data is used to build the social network. Table 3 shows sample of summarized data.

Building social network

Summarized data was used as mentioned before to build the social network for 3 months where nodes represent GSM numbers of subscribers and edges represent any interactions between subscribers (calls in our case). The result was a direct graph which contains about 16 million nodes and about 300 million edges. Figure 2 visualizes a sample of our social network where size and color of nodes express ranking degrees and lines between the nodes express ranking weights. We used Fruchterman-Reingold algorithm [27] which is a force-directed layout algorithm for drawing the graph.

Weights were added to our graph where each edge carries a different weight. The calculated weights depend on the count and duration of all calls between each side of edge. In order to unify weight over all the graph we normalized the duration of calls for each edge

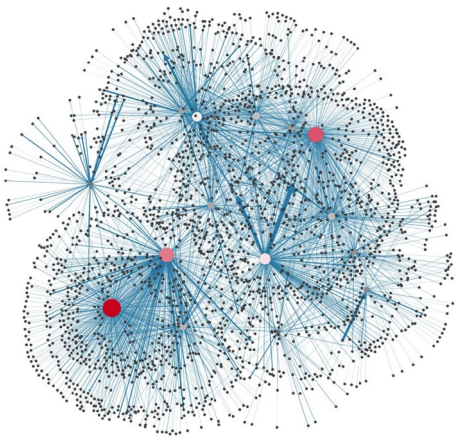

Fig. 2 Visualization for a sample of the Telecom social network

Table 4 Sample of Telecom social network data

Source	Destination	Weight
963-9******14	963-9******22	0.0425
963-9******31	963-9******62	0.0496
963-9******94	963-9******11	0.0272
963-9******34	963-9******78	0.0335

by dividing to the max value of duration in the network and we did the same thing for the count of calls. Finally, edges weight was calculated using the following equation:

$$W = \alpha \,.\, D_{Norm} + (1 - \alpha) \,.\, N_{Norm} \tag{1}$$

where W represents edge weight, D_{Norm} represents normalized duration of calls, N_{Norm} represents normalized number of calls and α represents the importance factor where $0 < \alpha < 1$. By choosing $\alpha > 0.5$ we give duration of calls more importance than number of calls and vice versa when $\alpha < 0.5$. In our network, calls duration was considered a little bit more important than calls number so we selected $\alpha = 0.6$ and the calculated weight must be $0 < W < 1$. Table 4 shows sample of Telecom social network data.

Social network analysis features

With the help of social network analysis, Telecom companies can recognize the customer's behavior and predict the strength of relations among customers. We analyzed our social network and calculated centrality measures. First, degree centrality (In-Degree, Out-Degree, Degree) was calculated for each node. Neighborhood degree (ND) [28] which is the node degree plus the sum degrees of node neighbors was also calculated, ND is given by the equation:

$$ND(v) = d(v) + \sum_{u \in N(v)} d(u) \tag{2}$$

Table 5 Sample of SNA centrality measures

Id (GSM)	In-degree	Out-degree	Degree	ND	LCCF
963-9******14	0.039	0.024	0.042	0.02867	0.241417
963-9******31	0.134	0.165	0.1993	0.124	0.022958
963-9******94	0.108	0.085	0.1286	0.092	0.037237
963-9******34	0.021	0.011	0.0213	0.018	0.074074
963-9******89	0.050	0.051	0.0673	0.0453	0.120723

where $d(v)$ is the degree of the node v and $N(v)$ represents neighbors of node v. Another calculated centrality measure is local clustering coefficient (*LCC*) [29], which indicates how close the node's neighbors are to be a clique (complete graph). This measure is given by the equation:

$$LCC(v) = \sum_{u \in N(v)} \frac{|N(v) \cap N(u)|}{|N(v)| * (|N(v)| - 1)} \tag{3}$$

All calculated measures are normalized by dividing to the max value of each measure over all the graph. Table 5 shows a sample of calculated SNA centrality measures.

The calculated SNA features (in-degree, out-degree, degree, *ND*, *LCC*) were used to enhance the churn prediction models that used in the Telecom company by adding social network features on top of the traditional churn predictors.

Detecting influencers

Previous measures can help find out the most influenced customers in the network but they are not enough, so more measures were calculated to come up with an influence score which expresses the importance of each node in the graph. This importance is based on the strength of links with other nodes presented by eigenvector centrality (*EV*), and the global location of the node within the graph presented by influence capability of node (*IC*).

Eigenvector centrality measures a node's importance while considering the importance of its neighbors; the main idea is that links from important nodes (as measured by previous centrality measures) are more valuable than links from unimportant nodes. All nodes start with equal *EV* value, but as the computation progresses, nodes with more degree start gaining importance. This importance propagates out to the nodes to which they are connected. After a number of computing iterations, the *EV* values stabilize and give the final values for eigenvector centrality for all nodes. *EV* centrality is calculated by using the equation:

$$EV(v) = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} v_j \tag{4}$$

Where λ is a constant scalar value and A is the adjacency matrix [30] which represents the network mathematically and has values:

$$A_{ij} = \begin{cases} 1 & \text{If there is an edge between vertices } v_i \text{ and } v_j \\ 0 & \textit{Otherwise} \end{cases} \tag{5}$$

Table 6 Sample of EV and IC measures

Id	EV	IC
963-9******14	0.030832	0.241417
963-9******31	0.094961	0.022958
963-9******94	0.051983	0.037237
963-9******34	0.135877	0.074074
963-9******89	0.046235	0.120723

Edges weight was included in the calculations of *EV* by replacing the one values in adjacency matrix A with the edge weight.

Wang et al. [18] proposed an Influence Capability measure based on *k-shell* values and the iteration information in the decomposition process to distinguish nodes with the same k_s values. As we see in Table 1 many nodes have the same k_s values but different locations in graph so their influence capability differs and nodes with higher iteration values are closer to the core nodes. First, they proposed a *k-shell* iteration factor:

$$\delta_v = k_s \,.\, \left(1 + \frac{n}{m}\right) \tag{6}$$

where, k_s is the *k-shell* value for node v, m is the total iteration number and v is the removed node in the *n-th* iteration of the *k-degree* process. Then they proposed an influence capability factor defined as follows:

$$IC_v = \delta_v \,.\, D_v + \sum_{u \,\epsilon\, N(v)} \delta_u \,.\, D_u \tag{7}$$

where IC_v represents the influence capability of node v, δ_v is the *k-shell* iteration factor of node v, D_v is the degree of node v and $N(v)$represents neighbors of node v. The Influence Capability factor *IC* contains the degree which is a local measure and the *k-shell* iteration factor which is a global measure. Therefor, *IC* takes into consideration the local and global influence capabilities of the node which help to distinguish between nodes more accurately. We calculated *EV* and *IC* for all nodes in network and normalized values. Table 6 shows a sample of calculated *EV* and *IC* measures.

Multi-SIM subscribers model

There are only two major Telecom companies in our country, we have named the first operator to which the data set belongs to as "Original operator", the other operator has been named as "Second operator". We used the social network and detailed data to build Multi-SIM subscribers' model. First, we benefited from graph proprieties to find nodes (subscribers) that share mutual nodes (neighbors) because the subscriber often calls the same people from his different SIMs. Finding nodes with mutual neighbors is very complex and expensive, so we simplified it by distributing calculations over periods of 10 days throughout the 3 months duration, and removing node pairs that share less than three neighbors in each period. After finding node pairs that share neighbors, two filters were added to the resulted data. The first filter removes pairs that have edges between them because if the subscriber has two SIMs, it is rare that he calls from one of his SIMs to the other SIM (calling himself). The second filter removes pairs that have a number of

Table 7 Sample of calculated similarity SNA measures

Id_1	Id_2	Sim_{Jacc}	Sim_{Cos}	Sim_{Score}
963-9******14	963-9******72	0.241417	0.253801	0.247609
963-9******31	963-9******16	0.124975	0.149874	0.137425
963-9******94	963-9******11	0.291038	0.314238	0.302638
963-9******34	963-9******54	0.184251	0.196213	0.190232
963-9******89	963-9******37	0.312462	0.337125	0.324794

shared nodes less than the selected threshold. We selected 10 shared nodes as a threshold. The next step in our Multi-SIM subscribers' model was calculating two types of measures: SNA similarity measures and SNA behavioral measures for each pair of nodes after filtration step. SNA similarity [31] measures include the following:

- Jaccard measure: in this measure we normalize the number of shared neighbors between two nodes based on the size of union of its two neighborhoods'. This measure is given by the equation:

$$Sim_{Jacc}(v\ ,\ u) = \frac{|N(v)\ \cap N(u)|}{|N(v)\ \cup N(u)|} \tag{8}$$

  where $N(v)$represents neighbors of node v.
- Cosine measure: it is the cosine of the angle between the characteristic vectors of the neighborhoods of two nodes. This measure is given by the equation:

$$Sim_{Cos}(v\ ,\ u) = \frac{|N(v)\ \cap N(u)|}{\sqrt{|N(v)\ | * |N(u)\ |}} \tag{9}$$

We calculated a similarity score, which is the average between the two previous similarity measures. The similarity score plays a main role to detect pairs that have high probability to be similar and exclude ones with low probability by filtering on a threshold (we selected 0.1 threshold). Table 7 shows a sample of calculated similarity SNA measures.

SNA behavioral measures focus more on the behavior of each node of the pair with the shared neighbors. We used detailed data to extract these measures and use them in addition to similarity score to increase the probability. SNA behavioral measures include the following:

- Common duration range: it represents the most common duration range of voice calls with each node of the shared neighbors. Duration ranges are clustered into 6 groups:
  - 0 to 4 min
  - 5 to10 min
  - 11 to 20 min
  - 21 to 40 min
  - 41 to 60 min
  - More than 60 min

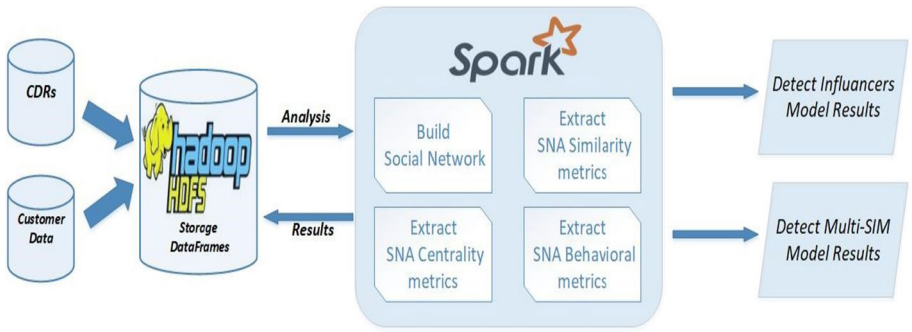

Fig. 3 Solution architecture cycle

- Common time of day: it represents the most common period of the day where a voice call occurs with each node of the shared neighbors. The day is divided into 4 periods; and each period spans 6 h:
  - Period 1: 00:00 a.m. to 06:00 a.m.
  - Period 2: 06:00 a.m. to 12:00 p.m.
  - Period 3: 12:00 p.m. to 06:00 p.m.
  - Period 4: 06:00 p.m. to 00:00 a.m.
- Common day of week: it represents the most common day of week where a voice call occurs with each node of the shared neighbors.

A friend is considered as a mutual behavioral friend when he has at least two of the behavioral measures with the same values with each side of the pair.

In addition to previous SNA measures, two other measures were used: subscriber location and IMEI. The operator can only records location and IMEI for its subscribers, as a result we can use these information only for detecting similarities between subscribers in the same operator. Subscribers often switch between their SIMs using the same mobile device, so IMEI can be used as an indicator of similarity between subscribers who have the same IMEI value. Subscriber location can help also as IMEI by identifying the most common places where the subscriber resides (home and job). We extracted these measures for pairs whose their two nodes are from the same operator using customer data, mobile IMEI information and towers information to increase accuracy in detecting similar subscribers.

Results and discussion

Performance

Performance usually plays a key role in the success of analytical models similar to our model, so the solution was designed to deliver high performance and speed, especially with ETL activities and SNA operations. HDP framework was installed and customized with a variety of systems and tools such as Hadoop, Spark, Yarn and Zeppelin. It contained eight nodes, eight Terabyte storage capacity, 32 Gigabyte RAM and eight core processors for each node. Figure 3 represents the solution architecture cycle.

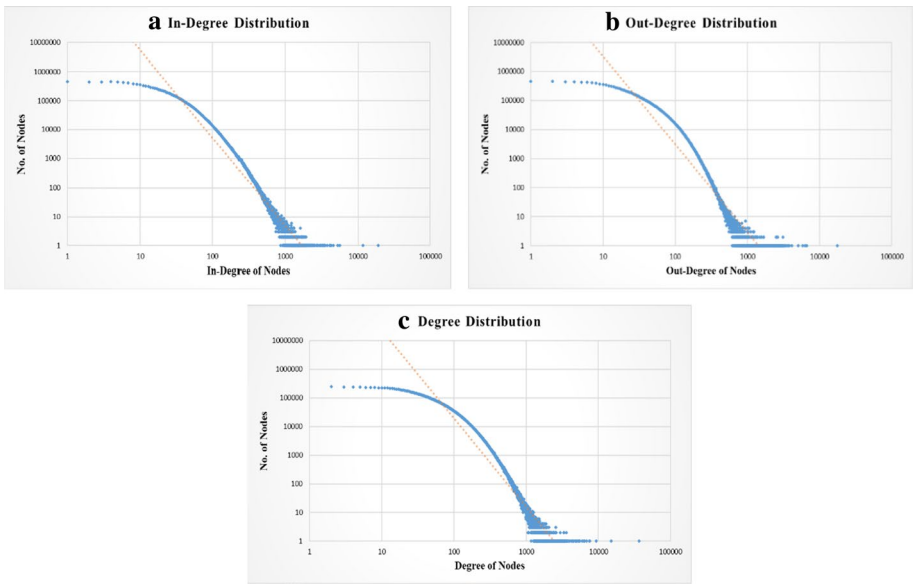

Fig. 4 Frequency distribution of degrees types

Using Spark is very useful for ETL processing and analytics because of its ability to perform calculations in-memory. This allows Spark to be faster than MapReduce and more efficient in batch and interactive processing of SQL queries. In addition, Spark supports several programming languages including Scala, Java, Phyton and R. Spark abilities were used to build the social network of 3 months CDRs provided by the Telecom company and to store network nodes and edges in HDFS as DataFrames that facilitates processing operations in the next steps. A window of 3-month data was chosen to build the network due to the limitation in capabilities of framework hardware and the complexity of SNA operations. The complexity of SNA comes from iterative calculations (*LCCF*, *EV* and *IC*) which need a massive memory to save all previous iterations. To solve this problem, calculation algorithms were modified by storing each iteration output and loading it to the next step with updating the iteration output each time to save the storage. Another performance challenge in SNA was finding node pairs that have mutual neighbors, which consumed memory very fast. To overcome this challenge calculations were distributed on 10-days periods over all 3-month duration as we mentioned before. We considered the max range of each period is 10 days according to the experiments we did. Any period range more than 10 days had failed due to hardware limitations.

Accuracy

Network was built and weighed according to the marketing needs in the operating company which cares more about call duration than calls number. Figure 4 presents the frequency distribution of in-degree, out-degree and degree, where frequency distribution is the fraction of nodes in the network with different types of degrees. All the charts (a), (b) and (c) in Fig. 4 are plotted in log-log scale with types of degrees on the X-axis and the number of nodes on the Y-axis, we chose the log-log scale for its suitability to represent the large range of degree values. The distribution in the figure demonstrates a power law

Table 8 Multi-SIM detection model results

Condition	Pairs operator type	Result (%)
SNA similarity measures	(Original-Operator, Second-Operator)	68
SNA similarity measures + SNA behavioral measures	(Original-Operator, Second-Operator)	76
SNA similarity measures	(Original-Operator, Original-Operator)	74
SNA similarity measures + SNA behavioral measures	(Original-Operator, Original-Operator)	85
SNA similarity measures + SNA behavioral measures + additional measures	(Original-Operator, Original-Operator)	92

behavior with exponents 3.017 for in-degree, 3.02 for out-degree and 3.085 for degree. The figure also shows that the distribution of the types of degrees gives similar results that there are very few nodes which have very high degrees. Therefore, these nodes may be suitable for individual targeting by a Telecom service provider.

A research analysis has been done by using our results together with data of the most common locations of customers. We have selected randomly a sample of 50,000 customers belong to the same city from our social network. A 15% highest-ranked influence customers' group were chosen from the sample depending on the *EV* and *IC* scores, which represent the importance of customer in the network, we called this group as influence group. Another 15% customers' group were chosen from the remaining sample but this time randomly, we called it the random group. The influence group and the random group were targeted by a particular bundle purchasing campaign. Considering the subsequently forward 2 months, we have monitored customers who subscribed in the new bundle from the social network related to each group. The result of applying the campaign was that a 16,245 new customers have purchased the new bundle from the social network related to the random group and 22,255 new customers have purchased the new offer from the social network related to the influence group, which mean a 37% raise in number of customers who subscribed in the new offer included in the campaign. Also, there was a 30% raise in growth rate of mobile traffic of the new customers from the social network related to the influence group comparing to the new customers from the social network related to the random group. According to the past methods, the customer influence score has presented additional gains in terms of business effectiveness and performance.

The results of Multi-SIM detection model contained more than 1.5 million records of pairs with similarity score greater than 0.1 and mutual friends greater than 10. We split the result into two groups: the first group contained pairs from the same operator (Original-Operator, Original-Operator). The second group contained pairs from the two different operators (Original-Operator, Second-Operator). From each group we selected the 25% of highest-ranking subscribers in similarity score before and after applying SNA behavioral measures. The condition of behavioral measures was selecting pairs that have 5 mutual behavioral friends at least. The same process was also done with the first group (Original-Operator, Original-Operator) again but this time the additional factors were added (subscriber location and IMEI). We chose only pairs which have identical location or IMEI. The model can be tested in two ways: first by making direct calls to subscribers in previous groups with a questionnaire about the number of lines that subscriber

possesses and to which operator each line belongs or second by using the customer data. Table 8 represents the results we have obtained. The traditional methods used in detecting Multi-SIM subscribers were able to detect subscribers only in the same operator based on customer data with a success rate between 30 and 40%, but by using our Multi-SIM detection model we have achieved a better results within the same operator and across different operators. Table 8 shows that using SNA similarity measures gave good results in detecting Multi-SIM subscribers and it's getting better when pairs are from the same operator (Original-Operator, Original-Operator). This can be explained by the fact that we have the whole network for Original-Operator subscribers where pairs are (Original-Operator, Original-Operator), but in the other case when pairs are (Original-Operator, Second-Operator) part of the network is missing (Second-Operator to Second-Operator calls).

Also the results show that using SNA behavioral measures increased the accuracy of the model with pairs from both same and different operators, the reason for this is that Multi-SIM subscribers rarely change their behavioral actions with the same people when contacting from different SIMs. We got the max result when we applied SNA similarity and behavioral measures in addition to location and IMEI of subscriber on the pairs from the same operator (Original-Operator, Original-Operator) and that's logical because we add more filtration constrains to our dataset.

In summary, the overall evaluation of our models shows that the social network analysis features gave us more accurate results in detecting influencers and Multi-SIM subscribers than the results obtained by the traditional methods.

The limitation of this work is represented by the window of data we choose for building the network (3 months CDRs) and the type of CDRs we use (calls CDRs only). We were unable to increase the width of the data window more than 3 months also we could not handle other types of CDRs (like SMS or internet CDRs) due to the massive amount of data we have to collect as well as the complexity of SNA operations that consume a large amount of storage and memory.

This work could be improved by getting rid of previously mentioned limitations and adding other data sources like SMS and Internet CDRs, so the data set would be larger and more suitable for building the social network and the models will be more robust and accurate. Finally SNA can play a main role in the analysis of Telecom data from the social point of view.

Conclusion

This paper presents a new prototype for building a social network on a big data platform using Telecom data by weighing and analyzing it. It also presents two models: influencers detection model and Multi-SIM detection model taking advantage of the built network. In influencers detection model we present a new way to measure the importance of subscribers by calculating eigenvector and influence capability of node rather than following the standard approach using centrality measures. This model was tested by targeting the highest ranking influencers. More accuracy and sufficiency were achieved than those obtained in classic ways with a 30% raise in growth rate of mobile traffic. In Multi-SIM detection model we benefited form the idea of mutual friends to calculate the similarity score and the behavioral score and use them to find pairs of SIMs (in the

same or different operators) that are probably similar and came to 85% accuracy result. The location and IMEI of subscriber were also used to increase the accuracy of results for pairs from the same operator and achieved 92% accuracy result. Finally, we see that SNA can help in Telecom industry for planning the right offers and studying the behavior of customers. There are also a lot of Telecom topics such as customers communities, fraud detection and advertising campaigns are worth pursuing and analyzing using SNA techniques.

Abbreviations

SNA: social network analysis; CDR: call detail record; SMS: short message service; SIM: subscriber identification module; ETL: Extracting,Transforming and Loading Data; GSM: Global System for Mobile communications; HDP: Hortonworks Data Platform; HDFS: Hadoop Distributed File System; k_s: location index of the node in the graph; W: edge weight; D_{Norm}: normalized duration of calls; N_{Norm}: normalized number of calls; α: importance factor; ND: neighborhood degree; d: degree of node v; $N(v)$: neighbors of node v; LCC: local clustering coefficient; EV: eigenvector centrality; A: adjacency matrix; λ : constant scalar value; δ_v: *k-shell* iteration factor; IC: influence capability of node; Sim_{Jacc}: Jaccard similarity score; Sim_{Cos}: cosine similarity score; IMEI: International Mobile Equipment Identity; RAM: random access memory.

Acknowledgements

Many thanks to Syriatel and Mr. Adham Troudi for support and motivation. Thanks for Mr. Mhd Assaf, Mr.william Soulaiman, Mr. Ammar Asaad, and Miss. Marwa Hanhoun for their co-operation and help.

Authors' contributions

NRA-M took the main role of performing the literature review, built the Telecom social network and implemented the proposed models. He conducted the experiments and wrote the manuscript. YR and MD took on a supervisory role and oversaw the completion of the work. All authors read and approved the final manuscript.

Funding

The authors declare that they have no funding.

Availability of data and materials

The data is not available to public because of Telecom company restriction applied on it, since the data were used under the license for the current study.

Ethics approval and consent to participate

All authors give ethics approval and consent to participate in submission and review process.

Consent for publication

The authors consent for publication.

Competing interests

The authors declare that they have no competing interests.

Received: 17 July 2019 Accepted: 4 November 2019
Published online: 15 November 2019

References

1. Borgatti SP, Everett MG, Johnson JC. Analyzing social networks. Thousands Oaks: Sage; 2018.
2. Mishra S, Mishra BK, Tripathy HK, Mishra M, Panda B. Use of social network analysis in telecommunication domain. In: Modern technologies for big data classification and clustering. IGI Global; 2018. p. 152–178.
3. Gurukar S, Ravindran B. Temporal analysis of telecom call graphs. In: 2014 Sixth international conference on communication systems and networks (COMSNETS). New York: IEEE; 2014.
4. Pinheiro CAR. Social network analysis in telecommunications, vol. 37. Hoboken: Wiley; 2011.
5. Zhan J, Guidibande V, Parsa SPK. Identification of top-k influential communities in big networks. J Big Data. 2016;3(1):16.
6. Phadke C, Uzunalioglu H, Mendiratta VB, Kushnir D, Doran D. Prediction of subscriber churn using social network analysis. Bell Labs Tech J. 2013;17(4):63–76.
7. Mizruchi MS, Mariolis P, Schwartz M, Mintz B. Techniques for disaggregating centrality scores in social networks. Sociol Methodol. 1986;16:26–48.
8. Newman ME. Mathematics of networks. The new Palgrave dictionary of economics. 2016; p. 1–8.
9. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA. Identification of influential spreaders in complex networks. Nat Phys. 2010;6(11):888.
10. Sabidussi G. The centrality index of a graph. Psychometrika. 1966;31(4):581–603.
11. Landset S, Khoshgoftaar TM, Richter AN, Hasanin T. A survey of open source tools for machine learning with big data in the hadoop ecosystem. J Big Data. 2015;2(1):24.
12. Onnela J-P, Saramäki J, Hyvönen J, Szabó G, De Menezes MA, Kaski K, Barabási A-L, Kertész J. Analysis of a large-scale weighted network of one-to-one human communication. New J Phys. 2007;9(6):179.

13. Nanavati AA, Singh R, Chakraborty D, Dasgupta K, Mukherjea S, Das G, Gurumurthy S, Joshi A. Analyzing the structure and evolution of massive telecom graphs. IEEE Trans Knowl Data Eng. 2008;20(5):703–18.
14. Amer MS. Social network analysis framework in Telecom. Int J Syst Appl Eng Dev. 2015;9(1):201–5.
15. MacQueen J, et al. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, Oakland, CA, USA; 1967. , p. 281–297.
16. Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: bringing order to the web. Technical report, Stanford InfoLab; 1999.
17. Werayawarangura N, Pungchaichan T, Vateekul P. Social network analysis of calling data records for identifying influencers and communities. In: 2016 13th international joint conference on computer science and software engineering (JCSSE), New York: IEEE; 2016. p. 1–6.
18. Wang Z, Zhao Y, Xi J, Du C. Fast ranking influential nodes in complex networks using a k-shell iteration factor. Phys A Stat Mech Appl. 2016;461:171–81.
19. Maher A, Ghoneim A. Multiple subscriber-identity-module detection using social network analysis techniques. In: 2014 IEEE international conference on data mining workshop. New York: IEEE; 2014. p. 804–809.
20. Katz L. A new status index derived from sociometric analysis. Psychometrika. 1953;18(1):39–43.
21. Verbeke W, Martens D, Mues C, Baesens B. Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Syst Appl. 2011;38(3):2354–64.
22. Huang B, Kechadi MT, Buckley B. Customer churn prediction in telecommunications. Expert Syst Appl. 2012;39(1):1414–25.
23. Dasgupta K, Singh R, Viswanathan B, Chakraborty D, Mukherjea S, Nanavati AA, Joshi A. Social ties and their relevance to churn in mobile telecom networks. In: Proceedings of the 11th international conference on extending database technology: advances in database technology. New York: ACM; 2008. p. 668–677.
24. Ahmad AK, Jafar A, Aljoumaa K. Customer churn prediction in telecom using machine learning in big data platform. J Big Data. 2019;6(1):28.
25. Sorić I, Dinjar D, Štajcer M, Oreščanin D. Efficient social network analysis in big data architectures. In: 2017 40th international convention on information and communication technology, electronics and microelectronics (MIPRO), New York: IEEE; 2017. p. 1397–1400.
26. Brdar S, Novović O, Grujić N, González–Vélez H, Truică C-O, Benkner S, Bajrovic E, Papadopoulos A. Big data processing, analysis and applications in mobile cellular networks. In: High-performance modelling and simulation for big data applications. Springer, Cham; 2019. p. 163–185.
27. Fruchterman TM, Reingold EM. Graph drawing by force-directed placement. Softw: Pract Exp. 1991;21(11):1129–64.
28. Ma T, Yue M, Qu J, Tian Y, Al-Dhelaan A, Al-Rodhaan M. Psplpa: Probability and similarity based parallel label propagation algorithm on spark. Phys A Stat Mech Appl. 2018;503:366–78.
29. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998;393(6684):440.
30. Biggs N, Biggs NL, Norman B. Algebraic graph theory, vol. 67. Cambridge: Cambridge University Press; 1993.
31. Li Y, Luo P, Wu C. A new network node similarity measure method and its applications. 2014. arXiv preprint arXiv :1403.4303.

Publisher's Note