Clique Comparison and Homophily Detection in Telecom Social Networks

Preliminary Communication

Marin Mandić
University of Mostar
Faculty of Mechanical Engineering, Computing and Electrical Engineering
Mostar, Bosnia and Herzegovina
marin.mandic.sb@gmail.com

Davor Škobić
University of Mostar
Faculty of Mechanical Engineering, Computing and Electrical Engineering
Mostar, Bosnia and Herzegovina
davor.skobic@gmail.com

Goran Martinović
J.J. Strossmayer University of Osijek
Faculty of Electrical Engineering, Computer Science and Information Technology
Osijek, Croatia
goran.martinovic@ferit.hr

Abstract *– Social Network Analysis (SNA) is based on graph theory and is used for identification of the structure, behavioral patterns and social connectivity of entities. In this paper, SNA is used in the telecom industry in terms of a call detail record referring to phone call data separated into two groups, i.e., domicile network and virtual operator network data. Emphasis was placed on community detection. Comparison was made among communities detected in domicile and virtual operator networks. Results show that in contrast to domicile network, the number of cliques in the virtual operator network is larger. Also, homophily was detected between domicile network and virtual operator network users.*

Keywords *– cliques, community detection, homophily, prepaid users, Social network analysis (SNA)*

1. INTRODUCTION

Telecom users make connections and build communities among themselves. Those connections can be detected by using a call and text detail record in a telecom system. Telecom operators detail data of any call (call detail records) and text messages exchanged between person A and person B. Those detailed data are very important because it is possible to extract knowledge and patterns which can be used for business improvement. One of the ways for extracting knowledge from detailed data is to create and analyze social networks. The usage of SNA in the telecom industry mostly refers to churn [1], fraud detection, a tariff model recommendation system [2], the identification of central nodes in the network [3], and the identification of communities and groups [4]. In this paper, emphasis is placed on community detection in social networks. Every user can get influenced by people from his/her or other communities he/she is connected with. Users also influence other users, and consequently they can influence future behavior of other users. This kind of influence between users is very important for telecom companies for the purpose of recognizing whether a user will stop using their services and change the telecom operator. It is very important to identify user departures (i.e., churn) in order to prevent them from taking place by using an appropriate marketing campaign. Marketing campaigns are activities directed towards users aimed at increasing user satisfaction. This enables telecom companies to keep or increase income and profit [5]. Churn detection was usually done by means of traditional methods, i.e., by using machine learning algorithms on different user attributes. By building social networks it is possible to increase the probability of churn prediction. In addition to churn, the influence of other telecom users is also important for a user decision to buy extra services or change a tariff model. Telecoms need to identify key subscribers that have a large number of connections with other subscribers. Key subscribers can influence the opinion and decisions of other subscribers [3]. Keeping a subscriber with a great

influence on other subscribers could, on a large scale, decrease the churn rate.

A telecom provider can offer a friend and family rate plan, in which a telecom user gets cheaper calls with their friends and becomes more satisfied with telecom services. With this offer, a telecom provider increases this user's loyalty and accordingly reduces the churn probability. The prevention of churn is essential to the preservation of revenue for any telecom company. Churners have influence on their social circles, so preventing churn reduces the churn probability of their social neighbors. It is easy to make a friends and family offer to postpaid users, because a telecom company knows their addresses and their family names, so it is easy to draw a conclusion who belongs to a certain family. Detecting a family based on demographic data is not possible for prepaid users because of their anonymity. SNA and community detection can help a telecom provider to make a friends and family offer to prepaid users because SNA helps us to detect who belongs to which community. In addition, SNA can help us to detect friends and family members although members of the family are interchangeably prepaid and postpaid users.

Key contributions of this paper are:

- Community creation comparison in different telecom networks.
- Homophily detection between domicile network and virtual operator networks.
- Calculation of connection weight in which the importance of calls and text messages is leveled.

This paper is organized as follows: Section 2 describes previous research in the field of social network analysis in the telecom industry, Section 3 demonstrates steps for community and homophily detection in telecom social networks, while Section 4 presents the conclusion and directions for future research.

2. REVIEW OF CURRENT RESEARCH

Current research in the field of social network analysis in the telecom industry is described in this section with a special overview of community detection. In their community detection related research authors have focused on different sizes of communities [6], the definition of a weakening clique [4], or creating new algorithms for community detection [7], while this paper deals with community comparison in different telecom networks since none of the referenced papers deals with such comparison. That is one of the key contributions of this paper.

2.1. SOCIAL NETWORK ANALYSIS

Social network analysis (SNA) is a set of analytic methods used to show and measure connectivity and interaction between people, groups, organizations, computers and other connectable entities. Social network does not necessarily have to be a social network (like Facebook or LinkedIn), but, as in this paper, it can be a network of users interacting with each other through mobile services. Interactions between users shown by using graphs enable visualization of connections that could not be seen otherwise, with an insight into the structure, enabling different measurements and test connections. Social networks are created by means of directed acyclic graphs, where nodes represent social entities. In this example, those are telecom users, and connections represent relations between those entities. In [8], the authors stated that social network analysis in the telecom industry can help the company to recognize customer behavior, predict the relationship strength and the influence on the events between them.

2.2. BASIC SOCIAL NETWORK ELEMENTS

A social network is composed of nodes and one or more kinds of connections between nodes [1]. Nodes can represent people, organizations, computers, web sites, and, as in this paper, telecom users. Nodes can be enriched with additional demographic attributes, such as region, date of birth, gender, etc. Connections show different kinds of relationships between nodes. In this paper, connections represent phone calls and text messages exchanged between telecom users. Connections between nodes can be directional or unidirectional. In directional connections, connection direction is visible, e.g., person A calls person B. Connections between nodes can be described by connection weight and type.

2.3. COMMUNITIES AND MEMBER GROUPING

One of the characteristics of social networks in a telecom environment is that they contain communities (subgroups) of nodes which are interconnected. This is a natural phenomenon that mimics real life, in which a person connects with a family, friends, colleagues, etc. A clique is a subscriber group, in which each subscriber is directly associated with any other subscriber within the clique. Since this definition of the clique is very restrictive, several ways of weakening this definition are created, such as n-cliques, n-clans, k-plexus and k-cores [4].

In an N-clique, each node is connected to each other with a distance (N) greater than one. Usually, a two-step distance is used. The N-clan provides an additional requirement on the n-clique, i.e., the link between the nodes must be over the other nodes inside the clique. A k-plexus weakens the clique so that the node can be a member of the clique if it is linked to all the nodes, with the exception of the K node number. For a k-core, a node belongs to a group if it is associated with the k number of nodes inside the kernel, regardless of the number of nodes it is not connected with. This definition makes a k-core even more relaxed than a k-plexus.

In this paper, communities are identified, where each member is directly related to each other member of the clique.

A group of authors [16] divided community detection based on the type of telco communication. They created an SMS community, a call community and a multidimensional community. They concluded that a multidimensional approach provides a richer level of insight into a people community and that the created community is different from the community created based on a single type of communication.

In this paper, we created a community based on SMS and voice communication and leveled the importance of voice and SMS messages.

Future research can expand a comparison of other types of clique to domicile and virtual network operators.

3. RESEARCH METHODOLOGY AND EXPERIMENT RESULTS

Figure 1 shows steps in the identification of community telecom social networking processes.

Furthermore, these steps for the identification of communities and homophily within a social network are described in the following subsections.

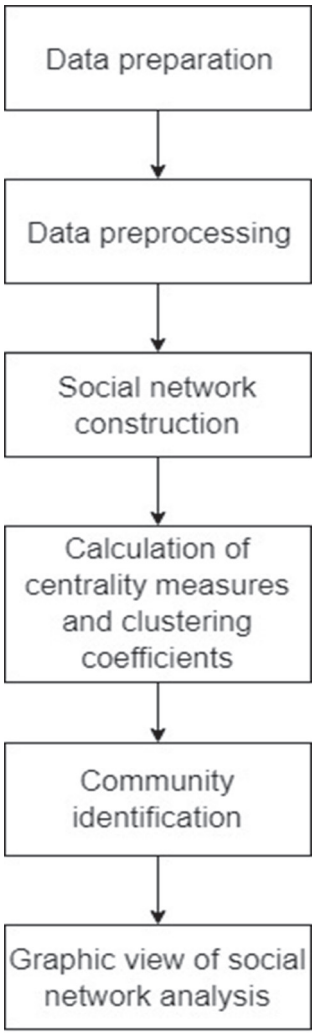

Fig. 1. Diagram of the process of identifying community telecom social networks

3.1. DATA PREPARATION

Telecom operators are high-tech companies that have a lot of information about their customers. Telecoms thus have demographic information about their users such as gender, age, geographic data, etc. Telecoms also have detailed call data that contain the A-number, i.e., the number that initiates the call, the B-number, i.e., the number that receives the call, call duration, call cost, call origination, call termination, etc. Out of the detailed data calls and text messages to other users since October 2017, prepaid user call data were used in this paper. The telecom user is a node, and the connection is created if a call exists or an SMS is exchanged between two users. The data used belong to large data sets. Large sets of data are characterized by the 5Vs of Big Data: Volume, Velocity, Variety, Value and Veracity [9]. It consists of large volumes of data, high velocity, a large variety of data, veracity and data values. Records of calls and sent messages are very large. They have a high rate of occurrence due to a large number of users and the intensive use of telecom services. Telecom has a variety of information about its users and these data need to undergo quality control in order to verify their authenticity. Customer data are extremely valuable to telecoms; hence telecom operators invest significant resources in the systems that handle and store such data. Data worth for telecoms is manifested in data exploitation for user segmentation, successful marketing campaigns creation, churn prevention, etc.

3.2. DATA PREPROCESSING

The data on the number of users used in the experiment were anonymized. All calls made to postpaid numbers, numbers of other mobile operators and all fixed operator numbers were dismissed as well. That led to nodes composed of pre-paid users. Due to a large amount of data, i.e., detailed calls, A-number and B-number groups were made and the data on call duration and the number of sent text messages were summarized. In this paper, connection weight was made with respect to call duration and the number of sent text messages. Call duration expressed in seconds was divided by 60 to obtain the value in minutes and then added to the number of sent text messages s. In this way, the importance of sent text messages and one minute of the call was made equal, and the measure of the connection strength between callers was obtained. The method of calculating the weight of the link differs in referenced papers and in this paper. Varun and Ravikumar [10] preprocess data by calculating the Euclidean distance of call duration, the number of calls, SMS messages sent and the buyer's age. A group of authors [11] calculate connection weight by summing up duration of calls between the telecom users. Since mathematical calculations needed to analyze social networks are very demanding, the data for one municipality in Bosnia and Herzegovina was taken, with approximately the same number of domicile network users and sub-

scribers who are members of the virtual network. The number of virtual operator network nodes is 10,753 compared to the number of nodes in the domicile network of the home operator, which is 6,099.

3.3. SOCIAL NETWORK CONSTRUCTION

A social network can be created by using data located in the adjacency matrix or the edge list. References [4] and [10] create a social network by using such matrix. In our paper, a social network was constructed using data from the pre-processing phase, i.e., by using the created edge list. The edge list is a table with two columns in which column members are nodes. A pair of nodes in the same line indicates that these nodes are connected. Additional columns in the edge list describing the connection strength (the number of sent txt messages and call duration in minutes) and the operator (domicile and virtual) are entered. Table 1 shows part of the data from the edge list from which the social network was created in this paper.

Table 1. Edge list

Number A	Number B	Network of number A	Network of number B	Strength
159207	384338	Virtual	Virtual	63.30
308494	385417	Domicile	Domicile	8.03
164105	386803	Virtual	Domicile	3.60
113642	387721	Virtual	Virtual	7.33
381835	387789	Virtual	Virtual	26.83
384995	388886	Virtual	Virtual	24.97
181668	391315	Domicile	Domicile	2.50
1737064	394016	Domicile	Domicile	2.63
104532	394373	Virtual	Virtual	76.87

When the network was created, link direction was not demonstrated, so it can be said that a non-direct social network was created. The aim of this paper is to compare the clique and detect homophily in telecom social networks, so link direction is irrelevant. The created network has a total of 16,852 nodes and 33,668 connections between nodes.

3.4. CALCULATION OF CENTRALITY MEASURES AND CLUSTERING COEFFICIENTS

The calculation of centrality measures was made for the domicile and for the virtual operator separately, and for the clustering coefficient, all presented in Table 2. Measure calculations were made according to the formulas that follow. The average degree measure shows an average number of direct links the node has towards other nodes. The calculation of the degree measure of node k is shown in [12] and it is calculated by formula (1), where k_i is a degree of each node i.

$$\bar{k} = \frac{1}{n}\sum_{i=1}^{n} k_i \tag{1}$$

The average geodesic distance is the average length of the shortest path between all nodes. The average distance measure is calculated by formula (2), where $d(i,j)$ is the shortest distance between nodes i and j, and The average geodesic distance is the average length of the shortest path between all nodes. The average distance measure is calculated by formula (2), where d(i,j) is the shortest distance between nodes i and j, and ½ n (n-1) is the number of possible ties of n nodes in the network.

$$l = \frac{1}{\frac{1}{2}n(n-1)}\sum_{i\geq j} d(i,j) \tag{2}$$

Measurements carried out in the created social network demonstrate that virtual operators have a greater average degree of connected nodes and a smaller average distance than the domicile provider. This has demonstrated that the virtual operator social network is significantly more related.

Diameter was measured, which is in fact the highest value of the shortest path between two nodes. This measure demonstrates how large a network is, i.e., how many steps must be taken to get from one end of the network to the other. The diameter is calculated by formula (3), where d (u, v) is the distance from node u to node v.

$$max_{u,v \in V}\ d(u,v) \tag{3}$$

In Table 2, it can be seen that there are no significant differences in diameter values between the domicile and the virtual operator, which means that the networks with the same distribution of nodes do not differ significantly in terms of this parameter.

A clustering coefficient is a measure that quantifies transitivity as a natural feature of a social network. Social networking indicates that friends of a particular node are likely to be mutual friends [2]. The clustering coefficient can be measured globally, i.e., for the entire network, or locally, i.e., for a particular node. The local clustering coefficient tells us how close neighbors of a particular cluster are to form a clique.

The global clustering coefficient can be calculated in several ways. In this paper, the computational method was applied, as suggested by Watt and Strogatz [13]. The global clustering coefficient c is calculated for the entire social network by computing the average of all local values ($i = 1, \ldots, n$) and it is shown in formula (4).

$$c = \frac{1}{n}\sum_i c_i \tag{4}$$

The local clustering coefficient is calculated according to formula (5) shown in Oliveira and Gama [2], where node N_i is adjacent to v_i . e_{jk} represents the link connecting node v_j to node v_k. k_i is the degree of node v_i.

$$c_i = \frac{2|e_{jk}|}{k_i(k_i-1)} \; ; \; v_j, \; v_k \in N_i; \;\; e_{jk} \in E \quad (5)$$

In [14], Hanneman and Riddle have shown that a higher clustering coefficient defines a greater likelihood of creating cliques. The measurements from this paper show that the virtual social network is more prone to clique creation.

Table 2. Demonstration of the centrality measure and the clustering coefficient of the domicile and virtual operator

Measurement	Domicile operator	Virtual operator
Average node stage	2.36	3.40
Average shortest distance	9.69	9.14
Diameter	28.00	27.00
Global clustering coefficient	0.05	0.16
Local clustering coefficient	0.12	0.26

3.5. COMMUNITY IDENTIFICATION

In this paper, each member of the clique is directly linked to other member of the clique. The number and clique size data showed that there is a larger number of cliques on the telecom social network with a small number of members, and a smaller number of cliques with a larger number of members. These data confirmed the conclusion given in [7]. It is detected that there is a significantly higher number of cliques in the virtual operator than in the domicile one. The main reason for this is the tariff model of the virtual operator, where free calls to other users in the same network are allowed, with a specific fee for the established call. Table 3 shows the number of cliques per cluster member divided by the type of a network operator.

Table 3. Demonstration of clique quantity by clique size

Clique number of members	Domicile operator	Virtual operator
3	319	1,180
4	12	225
5	0	47
6	0	3

3.6. GRAPHIC VIEW OF SOCIAL NETWORK ANALYSIS

The analysis of social networks can be visually observed if a graph is created. Figure 2 shows a 6-node clique represented by the following numbers: 131,488, 250,055, 250,926, 332,163, 454,326, and 1,102,805. The nodes that belong to a clique are marked red and they are members of the virtual operator network. Green nodes are nodes from the domicile operator and it can be seen in the graph that it is relatively weakly associated with nodes from the virtual operator. A weak link between the nodes of the domicile operator node with virtual operator nodes points to homophily, which will be described in detail in the next subsection of the paper. Green nodes do not belong to the clique and are represented by the following numbers: 112,726, 218,079, 257,654, and 269,067. The nodes in the virtual operator network that are not part of the clique are also shown in the graph. They are labeled with the following numbers: 174,331, 201,076, 287,257, and 378,606. Pink nodes have a lot of links to clique nodes, but do not belong to the clique because they are not associated with all nodes in the clique. Red and pink nodes are the nodes of a virtual operator that have a very large number of mutual connections. A large number of connections between virtual operator nodes indicate a high likelihood of creating a clique.

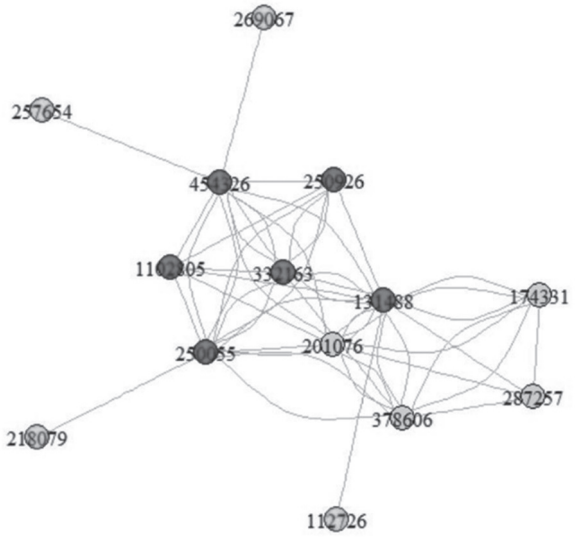

Fig. 2. Graphic representation of the telecom social network

3.7. Homophily

Homophily is a tendency for users to socialize and connect with users similar to them [15]. Homophily can be created with different attributes such as gender, race, nationality, social status, etc. In this paper, homophily was determined by the operator. Homophily can be recognized in the graph given in Figure 2, where it is apparent that green nodes, i.e., the domicile operator nodes, are poorly associated with nodes in the

virtual network. Table 4 shows the number of connections between different network users. For the preparation of Table 4, data from the pre-processing step were used. The table was created by summing up the number of connections between users grouped by the networks users belong to. Homophily can be proven if the connection strength between operators from Table 4 is taken into consideration. The table shows that most connections amounting to 83.6% are established within the same operator. Only 16.39% of connections are established towards users from other networks. Table 4 data confirm the conclusion that can be seen in Figure 2, i.e., that users are connected much better with users within the same network operator.

Table 4. The number of connections between different network users

A number network	B number network	Connection number	Per cent
Virtual operator	Virtual operator	8,757	51.96%
Domicile operator	Domicile operator	767	4.55%
Virtual operator	Domicile operator	1,996	11.84%
Domicile operator	Domicile operator	5,333	31.64%

4. CONCLUSION

This paper presents an analysis of social networks by using data on telecommunication service users. Special emphasis is given to community detection. This paper compares the clique and SNA measurements between the domicile and the virtual operator. By comparing the obtained data, it can be concluded that there is a significantly higher number of cliques in the virtual operator than in the domicile operator. The reason for creating a larger number of cliques in a virtual operator are free calls between virtual operator users. Nodes and links among subscribers are represented in a graph, and a greater number of connections between the users of the same network is observed. The graph also shows a significantly smaller number of connections between members of different networks, thus proving homophily. In addition, a comparison of the number of connections between subscribers of the same operator to the other operator is shown. It can be seen in Table 4 that most connections, i.e., 83.6%, were established within the same operator and thus homophily is confirmed.

Practical application of SNA created communities would be to attract users from other telecom operators who are community members with domicile network users. Such users are naturally under the influence of members of their community and are therefore more likely to switch to a domicile provider. Marketing campaigns should focus on such users. Membership in the community, respecting the number of connections to their own and to other users of the network, are the data that telecom companies can successfully utilize to create customer segmentation. In addition, a telecom provider can offer a friend and family rate plan based on SNA analysis.

Future research could focus on the calculation of the influence of SNA attributes on the improvement of classical models of churn prediction. It is also possible to make a clique-change analysis through time. For such analysis, call detail records are required from several successive periods. Future research can weaken the strict clique definition mentioned in this paper and compare such communities between the domicile and the virtual operator network.

5. REFERENCES

[1] S. Pushpa, "An Efficient Method of Building the Telecom Social Network for Churn Prediction", International Journal of Data Mining & Knowledge Management Process, Vol. 2, No. 3, 2012, pp. 31-39.

[2] M. Oliveira, J. Gama, "An overview of social network analysis", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 2, No. 2, 2012, pp. 99-115.

[3] Y. Ouyang, M.M. Hu, A. Huet, X. Sun, "Mining of leaders in mobile telecom social networks", Proceedings of the 2016 Wireless Telecommunications Symposium, London, UK, 18-20 April 2016, pp. 4–7.

[4] E. Varun, P. Ravikumar, "Telecommunication Community Detection by Decomposing Network into n-Cliques", Proceedings of the 2016 International Conference on Emerging Technological Trends, Kollam, India, 21-22 October 2016, pp. 1-5.

[5] M. Mandic, G. Kraljevic, I. Boban, "Performance comparison of Machine Learning methods for customer churn prediction in Telecom", International Journal of Electrical Engineering and Computing, Vol. 2, No. 1, pp. 29-36, 2018.

[6] L. Zhou, K. Lu, "Detecting communities with diferent sizes for social network analysis", The Computer Journal, Vol. 58, No. 9, 2014, pp. 1894-1908.

[7] N. Modani, K. Dey, S. Mukherjea, A. A. Nanavati, "Discovery and analysis of tightly knit communities in telecom social networks", IBM Journal of Research and Development, Vol. 54, No. 6, 2010 pp. 7:1-7:13.

[8] C.A.R. Pinheiro, M. Helfert, "Mixing scores from artificial neural network and social network analysis to improve the customer loyalty", Proceedings of the 2009 International Conference on Advanced Information Networking and Applications Workshops, Bradford, UK, 26-29 May 2009, pp. 954-959.

[9] Y. Demchenko, C. de Laat, P. Membrey, "Defining architecture components of the Big Data Ecosystem", Proceedings of the 2014 International Conference on Collaboration Technologies and Systems, Minneapolis, Minnesota, USA, 19-23 May 2014, pp. 104-112.

[10] E. Varun, P. Ravikumar, "Community Mining in Multi-relational and Heterogeneous Telecom Network", Proceedings of the 2016 IEEE 6th International Conference on Advanced Computing, Bhimavaram, India, 27-28 February 2016, pp. 25-30.

[11] A. Backiel, Y. Verbinnen, B. Baesens, G. Claeskens, "Combining local and social network classifiers to improve churn prediction", Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Paris, France, 25-28 August 2015, pp. 651-658.

[12] L.F. Costa, O.N. Oliveira Jr., G. Travieso, F.A. Rodrigues, P.R. Villas Boas, L. Antiqueira, M.P. Viana, L. E. Correa Rocha, "Analyzing and modeling real-world phenomena with complex networks: a survey of applications", Advances in Physics, Vol. 60, No. 3, 2011, pp. 329-412.

[13] D.J. Watts, S.H. Strogatz, "Collective Dynamics of Small-World Networks", Nature, Vol. 393, 1998, pp. 440-442.

[14] R.A. Hanneman, M. Riddle, "Concepts and measures for basic network analysis", The Sage handbook of social network analysis, 2011, pp. 340-369.

[15] F.J. Flynn, R.E. Reagans, L. Guillory, "Do You Two Know Each Other? Transitivity, Homophily, and the Need for (Network) Closure", Journal of Personality and Social Psychology, Vol. 99, No. 5, 2010, pp. 855-869.

[16] M. Zignani, C. Quadri, S. Bernardinello, S. Gaito, G.P. Rossi, "Calling and Texting: Social Interactions in a Multidimensional Telecom Graph", Proceedings of the 10th International Conference on Signal-Image Technology and Internet-Based Systems, Marrakech, Morocco, 23-27 November 2014, pp. 408-415.