A Novel Framework for Mobile Telecom Network Analysis using Big Data Platform

M. M. Abo Khedra[1]
Dept. of Computer Science
Faculty of Graduate Studies for Statistical Research
Cairo, Egypt

A. A. Abd EL-Aziz[2]
College of Computer and Information Sciences
Jouf University, Al jouf, KSA
Faculty of Graduate Studies for Statistical Research
Cairo University, Cairo, Egypt

Hedi HAMDI[3]
College of Computer and Information Sciences
Jouf University, Al jouf, KSA
University of Manouba, Tunisia

Hesham A. Hefny[4]
Vice-Dean of Faculty of Graduate Studies for Statistical Research
Cairo, Egypt

***Abstract*—Social Network Analysis measures the interconnection between humans, entities or communities and the streaming of messages between them. This kind of Analysis studies the relationship between different people in a very deep way; it shows how one node (subscriber) in the network can affect the others. This research studies the connections between the customers in many different ways to help any telecom operator increase the cross and up-selling of its products and services as follows: detect communities of subscribers which are a group of nodes collected together to form a community, identify the connection types and label the links between the customers as (business, friends , family and others), as well as identifying the top influencers in the network who can spread positive or negative messages about products and services provided by the company through communities in the network and determine off-net customers who can be acquired to be targeted by specific marketing campaigns. A real cell phone dataset of 116 Million call detailed records of SMS and Voice Calls of an Egyptian Communication Service Provider (CSP) is used.**

***Keywords*—SNA; influencer; acquisition; community detection; link prediction; call detailed record; on-net node; off-net node**

I. Introduction

A Social network is established from the connection of nodes with each other. Much more nodes and links mean much more types of links as friends, colleges, dislike, likes, same perspective or financial exchange. Social Network Analysis (SNA) is a way of discovering social relation by utilizing the network, the structure of the network consists of two main features: vertices (accounts, subscribers inside the network) and linkage (relationships or interconnection) that link them. It analyses the graph of the social network to detect relationship types between subscribers or customers. Social Network Analysis has some metrics to define the importance and figure out the behavior of each node in the graph and structure of the whole graph. The SNA metrics include density, centrality, degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and clique. Data mining based techniques can be utilized to apply the concept of SNA as it is proved a significant impact in this area. SNA can be applied in any industry that can avail some sort of data includes a connection between entities (subscribers, accounts, etc.).

- **Health**

The authors in [1] and [2] tracked and prevented diseases spread, target individuals that are at high-risk positions in the network, track changes in the health status of the population over time, all of this accomplished by using the SNA.

- **Payment fraud detection**

For the financial industry, SNA has a key part to play in detecting fraud as to a large extent it is being committed by organized networks. What the analysis does is putting some context around the content of the data. In payments and banking, the social networks consist of accounts, card numbers and so on. Connections (ties) between these various accounts is the transactions of transfer to figure out the relations between customers in the social graph and to be analyzed to determine patterns that could indicate fraudulent behaviour.

- **Security**

The author in [4] utilized the SNA metrics to detect terrorist networks by tracing communication they made on the events of September 11, 2001, through slicing connections among the hijackers from news statements.

- **Telecommunications**

SNA can be applied easily in the telecom industry due to the availability of data any telecom operator had. A social graph established from using Call Detailed Records (CDRs) and subscribers demographic extracted from the data warehouse or any big data tool like Apache Hadoop. SNA can use this data to discover the hidden relationships between subscribers through representing subscribers as vertices and interactions (calls) as edges between those vertices (nodes). In [3], the author presented the business pains telecom operators had without using SNA. In addition, described how to utilize

the availability of huge good data to increase the Average Revenue Per User (ARPU). Moreover, used to detect influencers, enhancing operations of cross and up-sell campaigns. Telecom operators are facing some main problems when deciding to spend on offers and campaigns that much. Traditional targeting techniques have the drawback of treating every customer as an individual apart from his social interactions and relationships. These social relationships have a great effect on individuals' opinions and usage of products and services. They do not know the relationship between customers (caller and callee) so they cannot be able to make specific offers to customers who are related with each other as a family, friends, co-workers or others. Telecom operators cannot detect the top influencer customers who can attract others to be in the same community. Moreover, the most important off-net customers to be acquired. Who is going to churn from the company, what else they can sell them? Thus, affecting their marketing costs and campaigns ROI. Our first and foremost task is implementing new proposed scientific techniques of Social Network Analysis on large telecom social network to create end-to-end generic SNA solution. To be able to use in the telecom industry to detect communities of customers, Identify relationships between the subscribers of the network, Identify influencer and key players in the network, identify the Ideal members that have the highest probability for acquisition.

The main contribution is proposing a generic, fully integrated, scalable, end-to-end SNA solution built for organizations to gain more insights on their customers from a social perspective. The introduced solution architecture starts with Community detection to identify the one big community for each consumer, which includes all his/her, interactions with other consumers. Then proposed a new Business based Rules for the Community Labelling, Through the extraction of new derived features from the data set to determine the type of relationship between the subscribers in the network such as Family, business, friends or others according to set of business behaviors of customers. Furthermore, identification of influencers in each community by implementing two algorithms used for various types of influence. Finally The creation of a new mathematical equation to acquire off-net customers. The business value behind that is as the following: Offer the customers suitable packages according to the communities they belong to whether it is family, friends or workmates and increasing the company market share by bringing in new subscribers.

The rest of the paper is organized as follows: Section II listed the related work for SNA, Section III explained the full solution architecture and studied the different techniques for implementing SNA, Section IV discussed the results of the techniques used in SNA and finally, Section V presented the conclusion.

II. Related Work

In this section, a literature survey is done for social network analysis, community detection, community labelling and influencer detection.

A. Social Network Analysis

Due to the big size of social network data, there is a need for a big data architecture to host the data and make the analysis on top of it. The author in [5] discussed this issue as they proposed a new prototype platform of SNA upon the architecture of big data using Spark GraphX framework in combination with JavaScript. The author in [7] proposed a study for handling the centrality analysis by using uncertainty methods, they integrated multiple centrality measurements by using fuzzy set and MYCIN theory. The author in [8] used mixed integer linear programming models to a dataset extracted from the social network. This research used two MILP clustering models M1 and M2. M1 model minimizes the maximum diameter between the cluster diameters; on the other hand, M2 reduces the total distance among objects related to the same cluster. M1 has computation time lower than M2.

B. Community Detection

Community detection is one of the key concepts in SNA. Nodes that tied together are grouped to build a cluster or a community. The author in [19] proposed a study for handling the centrality analysis by using uncertainty methods, they integrated multiple centrality measurements by using fuzzy set and MYCIN theory. The author in [6] utilized the clustering technique and centrality measure to detect link intensity among Brazilian Scientific researchers, Researches act as edges and researchers act as nodes. The author in [12] used the Girvan Newman algorithm to detect communities by progressively removing edges from the original network. The author in [13] used K-cliques algorithm for finding overlapping communities that is the combination of two cliques that share a common area that contains (k-1) node. The author in [14] used Fast Unfolding algorithm that unveils hierarchies of communities and allows zooming within communities to discover sub-communities, sub-sub-communities.

C. Community Labelling

The author in [18] predicted links in social networks using deep learning. Proposing computationally efficient and network-size-independent feature vector with deep learning that is fit for the real-time application. The author in [16] used a supervised algorithm (SVM) for relationship identification which gives a low accuracy (55%) and they admit that a specifically designed model for relationship identification will work better. The author in [11] listed some rules to characterize a group of subscribers into one of three different categories (Youth, Corporate, and Homebound). For example, if groups of subscribers have a high usage of messaging service with long duration calls at night and less frequent calls are labelled as Youth. Nevertheless, he did not mention anything related to the below point, which leads to an increase in the accuracy for the community labelling:

- Age difference (between the caller and the receiver, divide it by ranges, this would help to identify friends from families – ex-parent and child have a difference greater than 15 years).

- Home and work location (while taking into consideration that the other user maybe off-net and we would not know either age or home location).

D. Influencer Detection

The detection of influencers is a recent line of research in SNA. An Influencer is the most connected node (customer) in the network. The author in [9] used the HITS (Hyper-link Induced Topic Search) to detect the most important pages by analyzing the links and rates the Web pages. The author in [10] used the IP algorithm on twitter dataset to detect the influencers. The author in [15] used the Limited Recursive Algorithm to calculate social network influence for users on the social network Twitter and Facebook. The author in [17] focused on visualizing and measuring the different metrics of social network and based on those values, it can detect the most influential node.

III. Methodology

A. Proposed Framework

The framework based on Big Data tools, which contains four layers as shown in Fig. 1:

- Storage layer: a real data came from one of the communication service providers (CSPs) in Egypt to be loaded by the extract transform and load (ETL) process into a relational database management system (RDBMS), and then transferred to the Hadoop HDFS.
- Processing (analysis) layer: Apache Spark is a well-designed engine used to enhance graphs execution. A group of tools built on top of apache spark that contains a tool to handle structured data processing and SQL queries called Spark SQL, a tool for the analysis and mining models called PySpark that used with Python programming language, a tool for graph processing called GraphX.
- Post-Processing Storage layer: The output from the processing layer stored in the PostgreSQL and Neo4J for further analysis and visualization.
- Visualization layer: This layer uses Business Intelligence (BI) software to provide insights on the results of the analysis using reports and dashboards.

B. Data Collection

A cell phone dataset is used containing aggregated Call Data Record (CDR) for three months per the hours of the call for weekdays from (Sunday to Thursday) and the weekend includes (Friday and Saturday) per direction (incoming or outgoing) for 116 Million records CDRS of SMS and Voice Calls. The dataset also contains 12 Million records for customers' profiles that map to the aggregated CDRS with 26 columns like (age, location of the home, work, etc.). The caller and callee represent a network of nodes linked by Calls or SMS. The network based on the centrality measures and adjacency matrix.

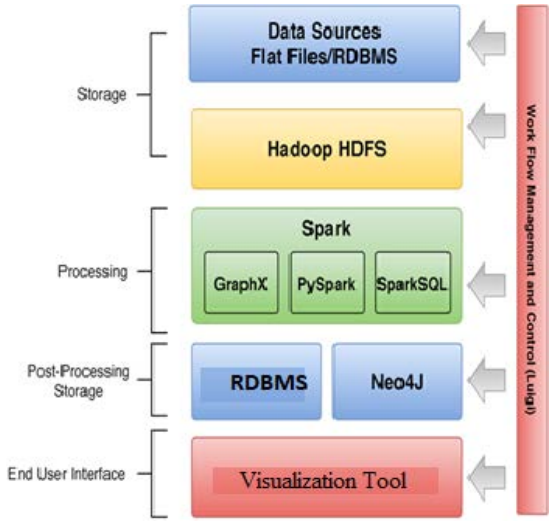

Fig. 1. Solution Architecture for SNA.

C. Social Network Analysis Techniques

Many techniques settled under the umbrella of SNA, the following sub-sections illustrate the flow of SNA methodologies and the approaches used to detect communities, community labelling, define key influencers and acquiring a new customer as shown in Fig. 2.

1) Community detection: Communities or clusters are a number of nodes collectively having a great chance of being linked to each other than to nodes of other groups.

There is no unique definition of a community so far which is widely accepted in social networks. A variety of definitions for a community has been proposed according to different aspects, which can be mainly classified into three main categories: intuitive definition, functional definition and definition from the process of an algorithm.

Community detection can support in predicting churners in a community, i.e. if the rate of churners increases in this community so most probably much more members will churn in a short time, so the operators have to stop that.

The following two algorithms are the most common methods to find out the communities from a given huge telecom network:

- Fast unfolding (Louvain Method) algorithm
- Label propagation (LPA) algorithm.

Both of them represents the heuristic method for greedy modularity optimization. In order to assess the performance and outcome of each technique, both algorithms ran on a sample of nodes and compared the output (nodes divided into clusters) as shown in Table I.

According to the NMI value that measures the accuracy as shown below in Fig. 3 and results in Table I, The Label Propagation algorithm used as a factor of initial splitting to discover communities and built on top of the community labelling.

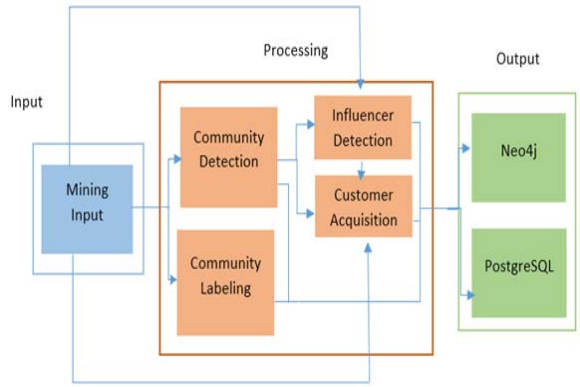

Fig. 2. The Flow of SNA Techniques.

2) Community labelling: The interactions between every two nodes "edge" indicate some distinct behaviour between them. Based on the interaction weight and time, and by using some defined rules, the edges between nodes could be classified into multiple categories or labels e.g. Friends, family, business or others. These labels could be used further to guide business strategies and plans to maximize the profit e.g. Promos and offers. By applying this technique in the telecom industry, it found that the structure of customer communication networks provides a natural way to understand customers" relationships and to a certain extent the behaviour of highly connected customers. The intention is to identify the type of call whether it is a family, friend, or business. The business value behind community labelling could be by offering the customers suitable packages according to the communities they belong to whether it is family, friends or workmates.

TABLE I. COMPARISON OF LABEL PROPAGATION AND FAST UNFOLDING ALGORITHM

Measure	Label propagation	Fast unfolding
Average degree per clusters	1.5363	1.5485
Average density per cluster	0.2286	0.1827
Average in degree per cluster	0.7693	0.7746
Average out degree per cluster	0.7692	0.7746
Average closeness centrality	0.2550	0.2331
Average betweenness centrality	0.0598	0.0080
NMI Value	0.912	0.435

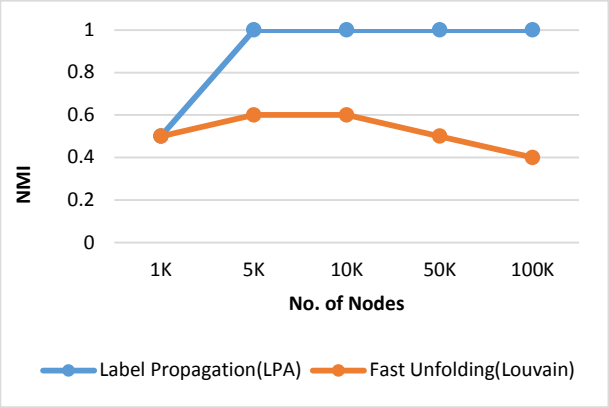

Fig. 3. Comparison of LPA and Louvain Algorithm.

A new business based rules proposed with taking into consideration age difference, location and some customers might be both a work colleague and a friend.

The day is categorized into four-time slots with different usage pattern:

- 6 am – 8 am (Early morning).
- 8 am – 5 pm (Working hours).
- 5 pm – 12 am (Early evening).
- 12 am - 6 am (late evening).

The Proposed Business based Rules for Community Labelling are as follows:

Family

- v_low_EM_WD =1.
- v_low_EE_WD =1 or v_medium_EE_WD =1.
- v_low_WE =1 or v_medium_WE =1.
- SMS_low_WD=1.
- SMS_medium_WE=1.
- low_age_difference =1 or high_age_difference=1.
- same home location = 1.

Friends

- v_low_EM_WD =1 or v_medium_EM_WD=1.
- v_medium_EE_WD=1 or v_high_EE_WD=1.
- v_medium_LE_WD=1 or v_high_LE_WD=1.
- v_medium_WE=1 or v_high_WE=1.
- SMS_medium_WD =1.
- SMS_high_WE=1.
- low_age_difference=1.

Business

- v_low_EM_WD=1.
- v_medium_WH_WD=1 or v_high_WH_WD=1.
- low_age_difference =1 or medium_age_difference=1.
- SMS_low_WE=1.
- SMS_high_WD=1.
- same work location =1.

3) Influencer detection: Influencer detection module has the ability to identify the most Influential persons in a telecom's operator network. Telecom operators pay attention to identify the influencers in each community to offer them a suitable promotion. Influencers can use viral marketing to talk about the advantages gained from telecom operators they belong. Several companies have produced tools, which supports marketing departments demand to detect leading

users who make some influence on the followers and who will spread more effectively information about the products. This is necessary for the recent trend to invest in the so-called Word of Mouth Marketing, where companies need to discover the influencers who will spread more effectively information about the new products. Which lead to increase the revenue In case of detecting and ranking influential Customers, the following algorithm is the most common method to detect the influencers when coming to the telecom industry and have a distributed implementation on Hadoop unlike the other algorithms, we can implement it on map reduce:

- Page rank: mathematical method

The PageRank algorithm is based on using the hyperlinks as an indicator of a page's importance. Every unique page is assigned a page rank. If a lot of pages' vote for a page by linking to it then the page that is being pointed to will be considered important.

4) Customer acquisition: The telecommunication operators networks could have some off-net customers who can be acquired to further increase the profit and the market share. Using the existing customer base to spread products and ideas beyond the current network borders. The foremost task is to identify top potential off-net consumers to join the network if they were targeted with the right campaign. The acquiring criteria depending on the acquisition score of the off-net customers, more score, more probability to be acquired.

SNA based score formula used to determine the network value for the off-nets. Each node (off-net consumer) assigned a score indicating its network value. The proposed equation to decide which off-net customers to acquire using Eq. (1).

$$\text{Net Value} = \sum (\text{weight of direct neighbors in the community}) \times \text{community density (Number of Links divided by Number of Vertices)} \quad (1)$$

IV. Results and Discussion

The results presented in this section are obtained from applying the algorithms mentioned in Section III.

A. Community Detection

The output extracted from Label Propagation algorithm used to decide which communities would be targeted by the new campaign. The related customers grouped together to represent one community as shown in Fig. 4.

B. Community Labelling

A social graph created G (V, E) with mobile phone subscribers as the vertices and with edges representing the social links, where (u, v) ∈ E represents the directed link from vertex (U, V) ∈ E to vertex v ∈ V and exists if u initiated a call to v. This graph is built from our set of call detail records as follows. For each region, we collected subscribers after eliminating outliers like automated calling machines and subscribers with low call counts, we pooled subscribers from all the regions and ranked them by their total number of calls and removed the top and bottom 10%. We then uniformly selected a sample set of 100,000 subscribers to our vertex set V. For each sampled subscriber we added their immediate neighbors to the vertex set and created the directed edges using our set of call detail records. Table II shows the final distributions of the relationship types in our directed graph.

C. Influencer Detection

A sample of nodes examined using the page rank algorithm and the output listed the node id and the authority score per node as shown in Table III.

To interpret the result, the output sorted and extracted the top 10 nodes (can be substituted later on by percentile); the selected nodes would be used as our influencers. Page rank algorithm denotes some node ids that are the new influencers where the telecom companies need to find them out to pay attention.

D. Customer Acquisition

Equation (1) that is mentioned in the customer acquisition helps in identifying which off-net customers would be most profitable for us to acquire. It requires a filtered input based on node type (on-net or off-net) as well as output from the influencer detection algorithms used as well as calculating the community density. The output is a list of recommended off-net nodes to acquire along with their company belong to as shown in Table IV.

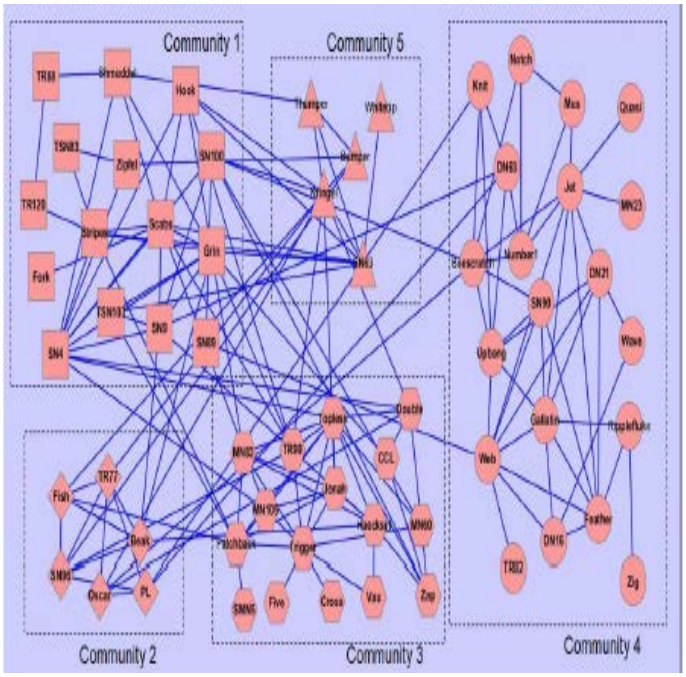

Fig. 4. Output of Community Detection.

TABLE II. Output of Community Labelling

Family Edges	63,350
Business Edges	191,581
Friends Edges	185,214
Others Edges	5267

TABLE III. OUTPUT OF PAGE RANK ALGORITHM

Page rank Top 10 Scores	
Customer Number	**Authority Score**
599000001	0.31637741
599012461	0.066292251
599013241	0.043618953
599012575	0.039831044
599012839	0.025255006
599011440	0.024302746
599011441	0.024302746
599011817	0.022412784
599010494	0.017147495
599010613	0.017147495

TABLE IV. OUTPUT OF CUSTOMER ACQUISITION

Off-Net Customer Number	**Company Belong To**
140254695	Etisalat
157372237	Orange
157372361	Orange
140254731	Etisalat
140254788	Etisalat
50651513	Orange
140254714	Etisalat
157372449	Orange
157373063	Orange

V. CONCLUSION

In this paper, a number of significant challenges pertaining to social network analysis and its implementation in the telecom operators. By focusing on building end-to-end generic SNA solution using big data tools that can be used in the telecom industry. SNA uses the call detail records (CDRs) of the Customer Service Provider (CSP) in conjunction with other data sources like customers' profiles to analyze the social relations of their customers. The paper has a full implementation of the following SNA key points: Community detection to find a community for each subscriber with (90%) accuracy, secondly, to detect the relationship between customers as (friends, family, co-workers, others) to make specific offers, thirdly, to identify the main influencers in the network who can convince other off-net customers to join the company with 99% accuracy and finally created a formula to acquire new customers.

The main goal is to support the decision-makers in the telecom industry to make the right decisions regarding the campaigns and offers to the right customers by providing them with the gaps in their relationships with customers and giving a three-sixty view about the customers through Business object tool for reports and dashboards.

We like to extend this research work by predicting rotational churn clients who left the company and returns back to the company but with a better offer than he took. In addition to, the churn prediction as one of the SNA key points to detect who is the client that has the intent to leave the company.

REFERENCES

[1] I. J. Saldanha, T. Li, C. Yang, C. Ugarte-Gil, G. W. Rutherford, and K. Dickersin, "Social network analysis identified central outcomes for core outcome sets using systematic reviews of hiv/aids," Journal of clinical epidemiology, vol. 70, pp. 164–175, 2016.

[2] K. E. Poundstone, S. A. Strathdee, and D. D. Celentano, "The social epidemiology of human immunodeficiency virus/acquired immunodeficiency syndrome," Epidemiologic reviews, vol. 26, no. 1, pp. 22–35,2004.

[3] S. Doyle, "Social network analysis in the telco sector marketing applications"Journal of Database Marketing & Customer Strategy Management, vol. 15, no. 2, pp. 130–134, 2008.

[4] V. Krebs, "Uncloaking terrorist networks," First Monday, vol. 7, no.4,2002.

[5] I. Sori´c, D. Dinjar, M. ˇ Stajcer, and D. Oreˇsˇcanin, "Efficient social network analysis in big data architectures," in Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017 40th International Convention on. IEEE, 2017, pp. 1397–1400.

[6] V. Str¨oele, F. Campos, J. M. N. David, R. Braga, A. Abdalla, P. I. Lancellotta, G. Zimbrˇao, and J. Souza, "Data abstraction and centrality measures to scientific social network analysis," in Computer Supported Cooperative Work in Design (CSCWD), 2017 IEEE 21st International Conference on. IEEE, 2017, pp. 281–286.

[7] X. Zuo, B. Yang, and W. Zuo, "Exploring uncertainty methods for centrality analysis in social networks," in Data Mining Workshops (ICDMW), 2017 IEEE International Conference on. IEEE, 2017, pp. 163–169.

[8] H. Pirim, "Mathematical programming for social network analysis," in Big Data (Big Data), 2017 IEEE International Conference on. IEEE, 2017, pp. 2085–2088.

[9] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze (2008). "Introduction to Information Retrieval". Cambridge University Press. Retrieved 2008-11-09.

[10] Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011, September). Influence and passivity in social media. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 18-33). Springer, Berlin, Heidelberg.

[11] SARAVANAN, M., et al. Labeling communities using structural properties. In: 2010 International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2010. p. 217-224.

[12] Girvan M. and Newman M. E. J., Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002).

[13] Fortunato, S. (2010). Community detection in graphs. Physics reports, 486(3-5), pp. 75-174.

[14] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), pp.100-108.

[15] Hajian, B., & White, T. (2011, October). Modelling influence in a social network: Metrics and evaluation. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing (pp. 497-500). IEEE.

[16] Yu, M., Si, W., Song, G., Li, Z., & Yen, J. (2014, August). Who were you talking to-Mining interpersonal relationships from cellphone networkdata. In 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014) (pp. 485-490). IEEE

[17] Farooq, Aftab, Gulraiz Javaid Joyia, Muhammad Uzair, and Usman Akram. "Detection of influential nodes using social networks analysis based on network metrics." In 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1-6. IEEE, 2018

[18] Chiu and J. Zhan, "Deep learning for link prediction in dynamic networks using weak estimators," IEEE Access, vol. 6, pp. 35937-35945, 2018.

[19] Zuo, B. Yang, and W. Zuo, "Exploring uncertainty methods for centrality analysis in social networks," in Data Mining Workshops (ICDMW), IEEE International Conference on. IEEE, pp. 163–169, 2017.