Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.7 No.3, 2018, pp.58-61
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajcst-2018.7.3.1897

Rank Based Virtual Community Detection in Mobile User Network

S. S. Sunanna[1] and S. Aji [2]
Department of Computer Science, University of Kerala, Kerala, India
E-Mail: sunanna94@gmail.com

(Received 22 September 2018; Revised 1 October 2018; Accepted 13 October 2018; Available online 20 October 2018)

***Abstract* - A virtual community network is a network of individuals who are sharing something in common such as behavior, thoughts, and ideas. The detection of virtual communities in social networks encourages the researchers in the computational and social science to invest more effort and time to explore the different properties of virtual networks. This paper introduces, mobile user network, another application area of virtual community networks with some proven techniques. The work explains the link prediction methods in which the relation between the different parameters in the dataset is explored. The virtual network hence formed could give more insights to the investors in the mobile communication domain to fulfill the customers' needs and attract more users to their tent.**

***Keywords* - OSN, Virtual community, Call log Network, Social Network Analysis**

I. INTRODUCTION

In this digital era, mobile phones are playing an essential role in the daily life. The developments in 'Online Social Networks' (OSN) and cellular networks make people more interactive even though they are in different geographical areas. Billions of people are actively participating in this kind of interactions and sharing their ideas, thoughts, and emotions without any restriction and hesitation. These interactions create a considerable amount of socially relevant data which can be treated for various analytical purposes because these data consists of core information that spreads through communities in OSNs to the whole world, regardless of their status or location. Since this kind of dataset posses the behaviors and feelings of mass, the corporate companies, organizations, and governments are the potential users of social network analysis. Handling of missing data has been a significant risk for most of the researchers in the field of data analytics and especially in the social network analysis. Several methods have been described from the 1970s, and the common among them are EM algorithm [1], multiple imputation methods [10]. The role of missing data in the efficiency of the algorithms and the performance of simple imputation techniques are scarcely studied in those periods [2]. The primary cause of missing of data in social networks is due to non - response or boundary specification problems or fixed choice designs as described by Kossinets [11]. Zuo and Xiao [3] proposed a method for finding missing data on fuzzy soft set and later on a data filling approach described in [4]. However, this approach cannot be used for predicting data in fuzzy soft sets which have unknown data and incomplete.

DFIS [5] is the best method comparing all other previous developments, and afterward, the ADFIS [6] was introduced which focus more on the reliability of association through parameters in the soft set. These methods give more emphasis on the association of different attributes in the data set than the probability, and the missing links were categorized accordingly. In order to minimise the computational expenses of these methods, researchers give an improvement to ADFIS [6] and object parameter method [7] in 2017. That was treated as one of the efficient algorithms for prediction of missing nodes through an association between parameters as described in [16] and thereby improves the accuracy of ranking algorithms[19].

This paper explains how the missing value prediction methods is applied in the mobile phone network for better analysis. The rapid growth of the mobile users [20] have created a lot of healthy competitions between the cellular service providers, the virtual networks formed because of the communications between the users will give meaningful insights to the people in this domain.

II. MISSING LINK PREDICTION

The methods used for prediction of missing nodes are described in [16]. As the first step, it is essential to represent a network data in the form of Boolean-valued Information System (BIS), adjacency matrix in which 1 denotes the nodes are connected and otherwise 0. An OSN consists of followee nodes and follower or linked nodes. For BIS conversion, a unique set of nodes from both the followees and follower nodes are selected in the OSN and later it is represented in rows and columns. The BIS conversion follows the definitions which are mentioned in [16]

Definition 1: For two nodes x and y, if x is following y, then they are represented by xy wherein x is connected to y; hence xy = 1 and they are called linked nodes.

Definition 2: For two nodes x and y, if x is not following y, then they are represented by x*y wherein x is not connected to y; hence x * y = 0 and they are called unlinked nodes.

Consider the Fig. 1, a small network with five nodes A, B, C, D and E. The links of the nodes are AD, AC, DA, DC, EA, EB, DB, BD, BC, CB and BE. The nodes D, C and E are similar and form one virtual community by connecting to their prime node B. The same community is also be connected to A except C. The BIS representation of this incomplete network dataset is in a data structure called adjacency matrix, where 1 represents the connection between nodes and 0 represents unlinked nodes.

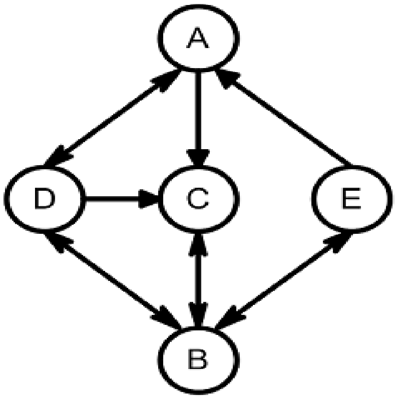

Fig. 1 A small network with five nodes

TABLE I REPRESENTATION OF FIG. 1 IN BIS.

Follower/ Followee	A	B	C	D	E
A	0	0	1	1	0
B	0	0	1	1	1
C	0	1	0	0	0
D	1	1	1	0	0
E	1	1	0	0	0

In Table I, some links are found missing because of network-imposed restrictions which makes the data or network incomplete. The cells of the BIS with the value equal to 0 denote that they are unlinked nodes. However, there is a probability, a mandatory part for link prediction, for those nodes to be linked. Link prediction tries to infer new links that are likely to exist and the meantime it avoids self-loops. By taking the unknown nodes of the first column as unknown, we can predict the unknown links through association, if $x_{ij} = 0, i \neq j$, then consider the cell x with index i and j as unknown. It can be represented as * for better understanding.

The consistency is of the first column with all other columns are calculated in the next stage of the algorithm.

$$CN_{1k} = \{x | F(x_{i1}) = F(x_{ik})\} \tag{1}$$

where $F(x_{i1})$ is the cell values in the first column of the BIS, $F(x_{ik})$ denotes all the cell values in all the other columns of BIS, and CN_{1k} is the set of cells in the column-1 that are consistent with the respective cells in the other k columns.

The consistency degree is,

$$CD_{1k} = CN_{1k}/U_1 \tag{2}$$

where CD_{1k} is the consistency degree of the first row and k column and U_1 is the number of known values, 1's, in the first column.

$$U_1 = |\{x | F(x) = 1\}| \tag{3}$$

The maximum consistency ratio among CD_{1k} is,

$$CD_{max} = max(CD_{ik}) \tag{4}$$

A predefined filter $\lambda, 1 \geq \lambda > 0$ is used to select strong associations.

If $CD_{1k} \geq \lambda$, then the unknown values in the column-1 are calculated as the corresponding values of column k. The Prime nodes are nodes that represent two or more consistent columns and the collection of followers in the prime nodes gives the Virtual community in a network. The algorithm used for prediction of missing nodes through an association between prime nodes is described in [17] and it is used for prediction.

III. RANKING ALGORITHMS FOR OSNs

A. Page Rank Algorithm

Sergey Brin and Larry Page developed the 'random surfer model' [8, 12, 13], the PageRank Algorithm, for ranking the web pages. A stochastic Markov process, browse through the web pages links, is explained through the Markov chain matrix. The Page Ranking algorithm is used to organize the web pages according to its significance by finding the total number and quality of links to a page and hence it gives an estimate that implies the importance of a website.

B. K- Core Rank Algorithm

Another most efficient and effective ranking algorithm is k-core. In 'Social Network Analysis', the grouping of individuals according to the firm or weak interaction is a significant concern [14, 15]. The presence of frequent, intense, direct, positive ties among a subset of users, then they are called cohesive groups [22]. Vladimir Batagelj and Matjaz Zaversnik proposed the k-core ranking algorithm [9]. Identifying top spreaders is one of the main challenges faced by researchers in understanding and controlling spreading processes on complex networks. The K- core ranking algorithm is used to identify the top spreaders [18] in a social network.

IV. EXPERIMENT AND RESULTS

The evaluation of this algorithm is done based on the imprecision function after and before prediction as described in [16].If the value of imprecision function (E) is low then the accuracy of prediction will be more, because those nodes were also contributed the most to the information diffusion. The call log dataset used for experiments are taken from [21]. It consists of two parameters calls and SMS. The calls dataset consists of four fields such as who $(userid)$ called which hashed phone numbers ($destphonehash$,which uniquely identifies a phone number in the caller phone), when $(timestamp)$, duration, and $userid$ of the recipient $(destuseridknown)$ if the recipient is also a subject in the experiment. The data related to calls and SMS are used to find the friendships between the users in the data set. The SMS data consists of the $userid$ of the sender, the number (hashed format) and user id of the receiver, time, and duration.

A community can be formed either as a physical community or as a virtual community. Physical community relation is established in society by either living in the same geographical area, sharing the same workplace, or being the members of the same institution. The well-known members in the region such as teachers, organizational leaders are marked as prime nodes. The second is the virtual community relationship in which the users have no such real community relationship in real life, but their preferences are correlated. This virtual or social association can be occurred based on similar choices or a shared world view, their relationship with prime nodes can be that of a particular product, intellectual, ideology, or any other possible relation. Some of them may be followers of prime nodes, and they may interact with the posts of the influencer by sharing, liking, or commenting. Some of them become followees. Not only the related prominent nodes but also the predicted links is also used to form a community. New link prediction is sensible when the followers share the same physical community relationship with their followees. The algorithm used here for link prediction uses consistent association, and it is given in the equation (1), which selects only consistent associations. For inconsistent association, the relation should be changed, and the equal sign $(=)$ should be replaced by not equal to sign $(\neq)$. Link prediction by finding the similarity between nodes is also used when there is a chance to link those node each other even though the nodes are not linked directly. This kind of growth of network through the relationship between the nodes is consistent and helpful to bring forward the essential links and hence to identify the top spreaders of a network.

The dataset consists of 57413 nodes and hence constructed a $57413X57413$ matrix and applied the algorithm for prediction of missing nodes.

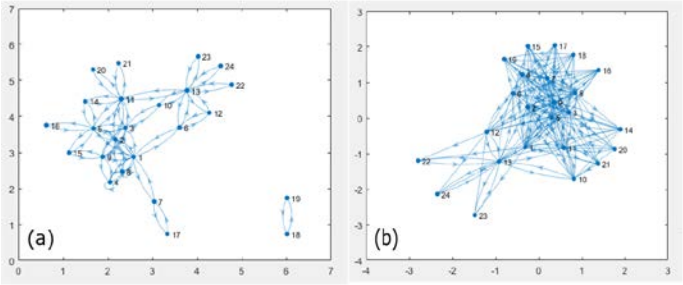

Fig. 2 Graph representation of (a) incomplete call log network (b) call log network after prediction of missing nodes.

The Fig. 2 (a) shows the graph representation of an incomplete call log network and Fig. 3 (b) shows the graph after prediction of missing nodes.

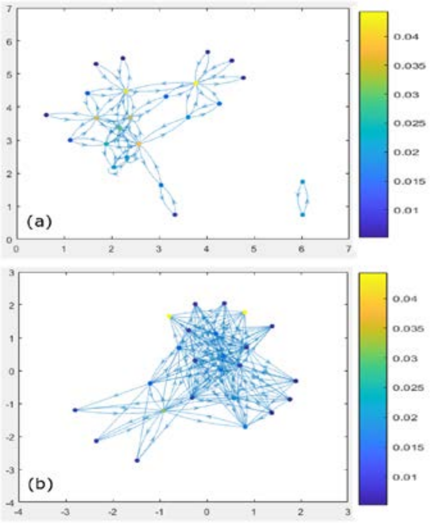

Fig. 3 Graph representation of call network based on PageRank score (a) before and (b) after prediction.

The graph on Fig. 3 (a) and (b) shows the call network with nodes whose PageRank score is higher than 0.005 based on this we can find out the active users in a particular area. The color scale on the right of the graph shows the Page Rank score from high to low. The yellow color denotes the node with high Page Rank, and it becomes dark blue when the nodes have low PageRank.

The Fig. 4 shows the statistics of PageRank and Fig. 5 that of k core, before and after prediction of missing nodes. Here, Eap represents the imprecision function after prediction and Ebp represents the imprecision function before prediction. The values in the x-axis means the imprecision function values for the top $1\%, 10\%, 20\%, 30\%, 40\% and 50\%$ top users from the network.

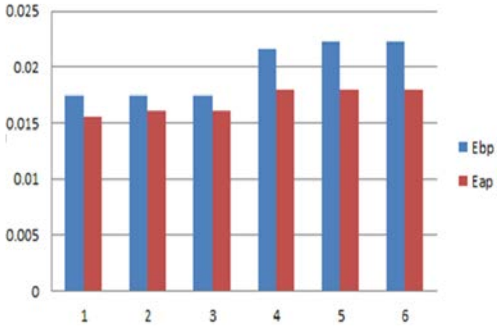

Fig. 4 PageRank of the call log network before and after prediction.

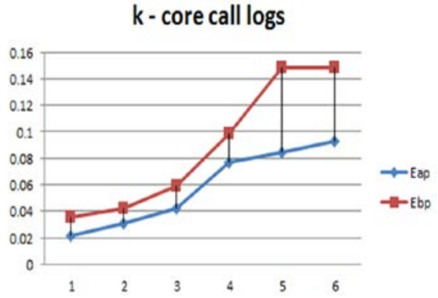

Fig. 5 k - core of the call log network before and after prediction.

Table II shows the statistics of the call log data set before and after finding the missing nodes. In this call network domain, more than 5 percent of new links are predicted.

Table Ii Statistics Of The Predicted Call Log Data Set

Data Set	Number of Links Before Link Prediction	Number of Links After Link Prediction	Number of New Predicted Links	Percentage of New Predicted Links
Calllog	10.6945	668.406	561.461	5.25

V. CONCLUSION

This paper discusses the methods for predicting the missing links in Social Networks. The missing nodes are predicted according to the association, the intensity and frequency of relation, between parameters in the dataset. The virtual communities are detected using the top spreaders identified with the help of ranking algorithms. Around five percentages of the total links is predicted in the experiments with call log dataset. There are notable differences between the networks formed before and after the link prediction. The virtual community detected using page ranking and K-core algorithms give meaningful information about the network. This information about the network is valuable for the telecom services providers for exploring better utilities for the customers. The virtual communities reveal the information flow happening in a mobile users network; this kind of knowledge has potential implications in this digital world where more than Eighty percent of the people use mobile phones for their communication.

REFERENCES

[1] J. Schafer, and M. Olsen, "Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective", *Multivariate Behavioral Research*, Vol. 33, No. 4, pp. 545-571, 1998.

[2] M. Huisman, "Imputation of Missing Network Data: Some Simple Procedures", *Journal of Social Structure*, Vol. 10, pp. 1-29, 2009.

[3] Y. Zou and Z. Xiao, "Data analysis approaches of soft sets under incomplete information", *Knowledge-Based Systems*, Vol. 21, no. 8, pp. 941-945, 2008.

[4] Kong, Zhi Zhang, Guodong Wang, Lifu Wu, Zhaoxia Qi, Shiqing Wang and Haifang. "An efficient decision making approach in incomplete soft set", *Applied Mathematical Modelling*, Vol. 38, pp. 2141-2150, 2014.

[5] H. Qin, X. Ma, T. Herawan and J. M. Zain. " DFIS: A novel data filling approach for an incomplete soft set", *International Journal of Applied Mathematics and Computer Science*, Vol. 22, No. 4, pp. 817-828, 2012.

[6] M. Sadiq Khan, M. Al-Garadi, A. Wahab and T. Herawan, "An alternative data filling approach for prediction of missing data in soft sets (ADFIS)", *SpringerPlus*, Vol. 5, No. 1, 2016.

[7] Y. Liu, K. Qin, C. Rao and M. Alhaji Mahamadu, "Object–Parameter Approaches to Predicting Unknown Data in an Incomplete Fuzzy Soft Set", *International Journal of Applied Mathematics and Computer Science*, Vol. 27, No. 1, pp. 157-167, 2017.

[8] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hyper - textual Web search engine", *Computer Networks*, Vol. 56, No. 18, pp.3825-3833, 2012.

[9] V. Batagelj and M. Zaversnik, "An O(m) Algorithm for Cores Decomposition of Networks", *Advances in Data Analysis and Classification*, Vol. 5, No. 2, pp. 129-145, 2011.

[10] J. Schafer and J. Graham, "Missing data: Our view of the state of the art.", *Psychological Methods*, Vol. 7, No. 2, pp. 147-177, 2002.

[11] G. Kossinets, "Effects of missing data in social networks", *Social Networks*, Vol. 28, No. 3, pp. 247-268, 2006.

[12] Krapivin, Mikalai and M. Marchese, "Focused Page Rank in Scientific Papers Ranking", In *International Conference on Asian Digital Libraries*, pp. 144-153, 2008.

[13] L. Page, S. Brin, R. Motwani and T. Winograd, "The Pagerank Citation Ranking: Bringing Order to the Web", *Social Networks*, 1999.

[14] Wasserman S. and Faust K. "Social Network Analysis: Methods and Applications", *Cambridge University Press*, 1994.

[15] S. B. Seidman, " Network structure and minimum degree", *Social Networks*, Vol. 5, pp. 269-287, 1983.

[16] M. Khan, A. Wahab, T. Herawan, G. Mujtaba, S. Danjuma and M. Al-Garadi, "Virtual Community Detection Through the Association between Prime Nodes in Online Social Networks and Its Application to Ranking Algorithms", *IEEE Access*, Vol. 4, pp. 9614-9624, 2016.

[17] Y. Liu, M. Tang, T. Zhou and Y. Do, "Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition", *Scientific Reports*, Vol. 5, No. 1, 2015.

[18] S. Pei, L. Muchnik J. S. Andrade, Z. Zheng and H. A. Makse, "Searching for superspreaders of information in real-world social media", *Scientific Reports*, Vol. 4, No. 1, 2014.

[19] J. Y. Halpern *et al*. "A ranking algorithm for online social network search", *In Proceedings of the 6th ACM India Computing Convention Article No. 17*, 2013

[20] W. Dong, B. Lepri and A. Pentland "Modeling the co-evolution of behaviors and social relationships using mobile phone data", In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia - MUM*, 2011.

[21] P. A. Chirita, J. Diederich, and W. Nejdl, "MailRank: using ranking for spam detection", In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management-CIKM*, 2005

[22] K. Mankirat and S. Sarbjeet, "Analyzing negative ties in social networks: A survey", *Egyptian Informatics Journal*, Vol. 17, No. 1, pp. 21-43, 2016.