VOL. 13, NO. 7, APRIL 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

MISSING LINK PREDICTION IN COLLABORATIVE RESEARCH BASED ON TRIADIC ANALYSIS

Akila Venkatesan and Govindasamy Vaiyapuri
Pondicherry Engineering College, Pondichery, India
E-Mail: akila@pec.edu

ABSTRACT

Researchers, academicians collaborate in academic institutions to perform research. Most of the time, the outcome of this collaboration is in the form of Co-authored paper publication. The Co-authoring of research papers can be modeled as a network. In this context, many works have been done about the link prediction in Co-authorship networks. To predict whether a link will appear or not, each pair of nodes within the network is assigned a score called similarity or proximity. The links having higher similarity score are supposed to be of higher existence likelihoods. So far, majority of previous works in link prediction focus on the application of similarity measures such as Common Neighbors, Jaccard Coefficient and Simrank, in order to predict new connections in the future. To analyze and understand any evolving network, it is necessary to study not only the link formation but also the link dissolution. Link dissolution is a relatively new problem. This paper addresses Unlink Prediction based on (1) Social Embededdness with Network Overlapping and (2) Social Embededdness with Triadic Analysis. The experiments are conducted using the bibliographic dataset from DBLP-ACM. The results shows that the proposed work performs better in terms of Precision and Recall.

Keywords: co-authorship networks unlink Prediction, social embededdness, triadic analysis.

INTRODUCTION

A social network is a social structure made up of a set of social actors (such as individuals or organizations), and the social interactions between actors. The social network perspective provides a set of methods for analysing the structure of whole social entities as well as a variety of theories explaining the patterns observed in these structures [1]. A subset of social network is the Co-authorship network. Collaborations among authors often form a network of connections which defines the Co-authorship network.

The link prediction problem is relevant to a number of interesting current applications of social networks. It is believed that the informal social network that prevails in an organization is beneficial more than the official hierarchy imposed by the organization itself [13]. This concept can be extended to Co-authorship network also. It is proved that collaborations between authors/researchers often yield more fruitful results than individual researchers' performance [4] [5]. Predicting the evolution of Co-authorship network thus plays a significant role in (1) analysing the trend of the structure of scientific collaborations; (2) detecting potential research communities as well as their evolutions; (3) assessing the future influence of scientists; and (4) recommending companions, assistants, or colleagues for individual researchers. Further, the Co-Authorship network is an evolving network. It is also important to analyse what links may disappear in future. This information will be helpful in allocation of funds to research groups etc. So, Unlink Prediction pertains to predicting links that will be removed given only the structure of a network. The problem of predicting whether a link will be removed can be viewed as the inverse problem of predicting the creation of links. To this end, this paper has proposed methods based on Social Embeddedness to predict links that may be unlinked.

RELATED WORK

A brief survey on papers predicting link dissolution or Unlink Prediction is given here. In [6] the Unlink Prediction uses probabilistic ontology using the probabilistic description logic CRALC. Unlink Prediction for German politicians are done in [7]. A computational social science approach is used in the paper. The paper reports that past link information is useful in Unlink Prediction. Paper [8] is based on the following and unfollowing behavior of German politicians. It is reported that there are differences in factors that influence formation and dissolution of ties. DecLiNe - a model to [predict decay of links in social networks is presented in [9]. It is based on structural analysis of graphs. Novel metrics are introduced in the paper which can predict link decay irrespective of the network type. The system has been evaluated with Wikipedia dataset. From the survey, it is evident that there is only a brief amount of work available in Unlink Prediction. In this context, the proposed system of Unlink Prediction based on Social embeddedness with triadic analysis is a novel work which is presented in the next section.

Unlink prediction using social embeddedness

In Co-authorship networks, Unlink Prediction predicts missing links in current networks. The objective of our approach is to validate how far unlinks can be predicted. Relationships are unlikely to resolve when the partner's networks are well connected. This indicates that the social support of the neighbors of the two users is important to determine if they will connect or disconnect. The Unlink Prediction is based on Social Embeddedness [10] [11] [12]. The Social Embeddedness is computed based on two methods i) Network Overlap and ii) Structural Balance.

Social embeddness using network overlap

Network overlap is defined as the degree of linkage between two peoples' networks. It is measured by ratio of the number of common neighbors and the number of possible common neighbor. To quantify the network overlap of two users i and j, the following measures i) Common Neighbors ii) Jaccard Coefficient and iii) SimRank are used. Common Neighbors is defined as the number of common neighbors shared by two nodes. Jaccard Coefficient is the ratio of common neighbors out of all neighbors, and can be used for comparing the similarity and diversity of neighbor set. SimRank score can be interpreted as the time before two random walkers meet on the network if they start at nodes and randomly walk the network. For all two nodes in the Co-authorship network, the three measures are computed. The Social Embeddeness based on Network Overlap is computed based on the following algorithm given in Figure-1.

Procedure **NetworkOverlap**
Boolean linkloss
N: Set of nodes
Link: Set of links
//CN- Common Neighbours //JC-Jaccard Coefficient
Begin
For all n_i , n_j Є {N}and n_i != n_j
$CN(n_i,n_j)$
$JC(n_i,n_j)$
$SimRank(n_i,n_j)$
endfor
If ($CN(n_i,n_j)$ > $\overline{CN(n_i,n_j)}$) && ($JC(n_i,n_j)$ > $\overline{JC(n_i,n_j)}$) &&
($SimRank(n_i,n_j)$ > $\overline{SimRank(n_i,n_j)}$
$Link(n_i , n_j)$=T
else
$Link(n_i , n_j)$=F
endif

Figure-1. Social embeddness using network overlap.

The Similarity measure of Common Neighbors Jaccard Coefficient and SimRank are computed for the two nodes ni and nj.It is hypothised that if the computed similarity measures are greater than the mean values for all the three measures, then the link will be retained otherwise the link will dissolve.

Social embededdness with triadic analysis

A triad is a subgraph of size 3. Triadic analysis refers to the analysis of the properties of all triads that can be formed in a network. The triadic analysis is performed to compute the social embeddedness of any two nodes. The first step is to convert the Co-authorship network to a signed network. To perform this exploratory analysis was performed with the three similarity measures for computing the cross density between actors. The results are shown in the Figure-2.

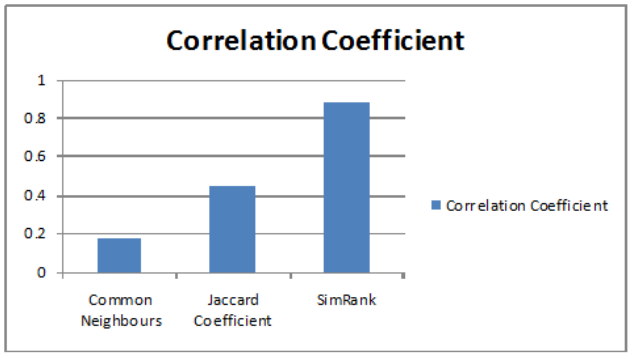

Figure 2. Correlation coefficient of the similarity measures.

Based on the experiments performed, SimRank measure gives the best Correlation Coefficient for cross density computation. So SimRank measure is used to convert the Co-Authorship network to a signed network. If the SimRank measure > 0.4 then a +ve sign is assigned to the link or a -ve sign is assigned in the Co-authorship network. In the second step, all the triads in the network are extracted. Every triad is checked for structural balance. The unbalanced triads are extracted for further examination. The unbalanced triads may have three negatives or one -ve sign.

Case 1: The triads with three -ve signs indicate that all the three authors are weakly bound so all the three links may dissolve.

Case 2: For the triads with one -ve link indicate that if the first author has a strong link with the third author and the second author has a weak link with the third then there is dissonance in the triad and the triad may dissolve.The balanced triads may have three +ve or one +ve sign in them.

Case 3: The triads with three +ve sign indicate that all the three bonds are strong so all the three links will stay.

Case 4: The triads with one +ve link indicate that the first author as well as the second author has a weak relation with the third author while the first and second author share a strong bond.

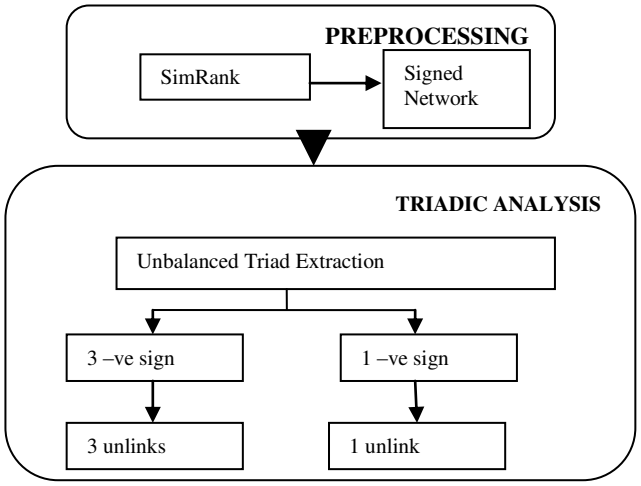

Figure-3. Embededdness with Triadic Analysis.

This indicates because of the resonance of thought between the first and second author the other two weak links may be nurtured in future. The high level architecture diagram is given in Figure-3.

RESULTS AND DISCUSSIONS

The experiments were conducted in a Intel core i3-2330M processor @2.20 GHz with Hard disk - 320GB. The experiments were conducted in NetBeans IDE 8.0 environment with Microsoft Access Tool and Graphviz tool 5.2.2. The experiment to evaluate the Unlink Prediction using Social Embeddeness with Network Overlap and Unlink Prediction using Social Embeddeness with Triadic Analysis were performed using the bibliographic dataset from DBLP-ACM. The data from 2003 to 2007 were analyzed. The experiments were conducted in four test runs. In the first Test case T1, the data from 2003 to 2004, in the second test case T2 the data from 2003 to 2005, in the third test case T3 the data 2003 to 2006 and in the final test case the data from 2003 to 2007 were used. The results are statistically analyzed to verify the significance of the results. The t-test for two paired samples / Two-tailed test were used for the statistical analysis.

Hypothesis with respect to precision

Null Hypothesis H0: There difference between the means of Precision in Social Embeddeness with Network Overlap and Precision in Social Embeddeness with Triadic Analysis is 0.

Alternate Hypothesis Ha: There difference between the means of Precision in Social Embeddeness with Network Overlap and Precision in Social Embeddeness with Triadic Analysis is not 0.

Table-1. Results of the T-Test for two paired samples W.R.T precision.

Difference	**-0.144**
t (Observed value)	-13.146
|t| (Critical value)	4.303
DF	2
p-value (Two-tailed)	0.006
alpha	0.05

The number of degrees of freedom is approximated by the Welch-Satterthwaite formula. As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis H0, and accept the alternative hypothesis Ha. The risk to reject the null hypothesis H0 while it is true is lower than 0.57%. The results of the t-test for two paired samples / Two-tailed test with respect to Precision are shown in Figure-4.

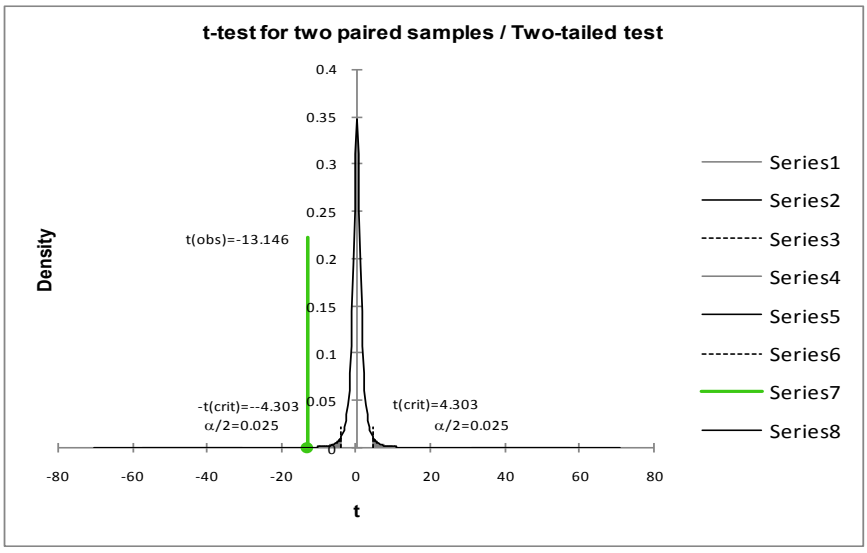

Figure-4. T-test for precision.

The experimental result for precision is shown in Figure-5.

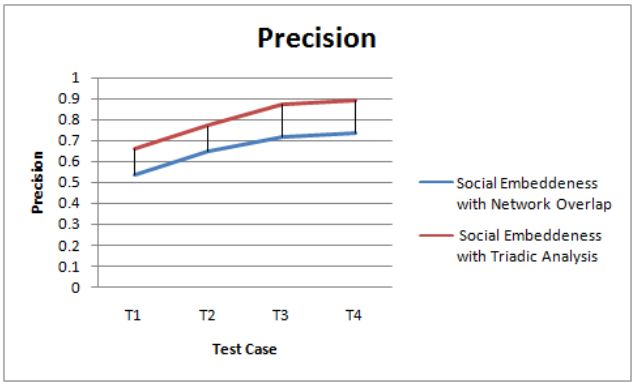

Figure-5. Performance Analysis based on Precision.

Hypothesis with respect to recall

Null Hypothesis H0: There difference between the means of Recall in Social Embeddeness with Network Overlap and Recall in Social Embeddeness with Triadic Analysis is 0.

Alternate Hypothesis Hb: There difference between the means of Recall in Social Embeddeness with Network Overlap and Recall in Social Embeddeness with Triadic Analysis is not 0.

Table-2. Results of T-Test for two paired samples W.R.T recall.

Difference	-0.105
t (Observed value)	-8.934
|t| (Critical value)	4.303
DF	2
p-value (Two-tailed)	0.012
alpha	0.05

The number of degrees of freedom is approximated by the Welch-Satterthwaite formula. As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis H0, and accept the alternative hypothesis Hb. The risk to reject the null hypothesis H0 while it is true is lower than 1.23%. The results of the t-test for two paired samples / Two-tailed test with respect to Recall are shown in Figure-6.

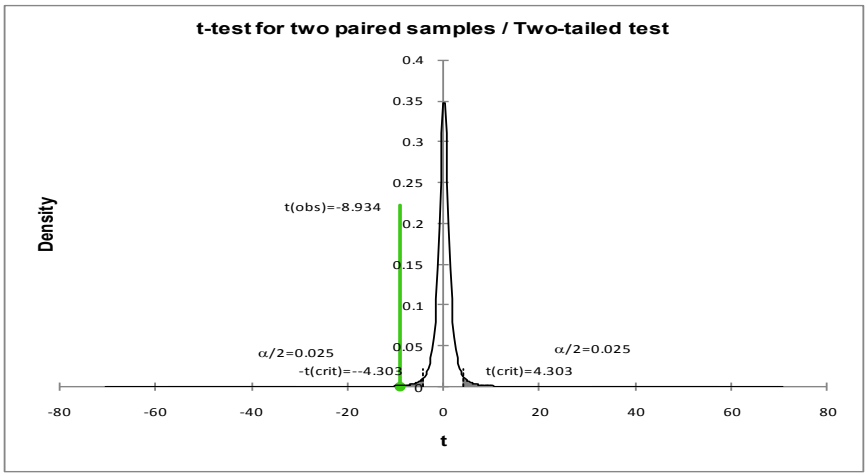

Figure-6. Performance analysis based on precision.

The experimental results for Recall are shown in Figure-7.

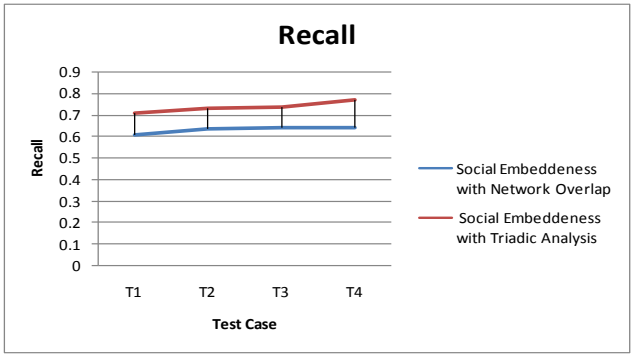

Figure-7. Performance analysis based on recall.

The Unlink Prediction was analyzed using Social Embeddeness with Network Overlap and using Social Embeddeness with Triadic Analysis. The results show that Unlink Prediction using Social Embeddeness with Triadic Analysis outperforms Social Embeddeness with Network Overlap for the parameters - Precision and Recall conclusively.

CONCLUSIONS

Link prediction can be accomplished by implementing both neighbor based methodologies namely Common Neighbors, Jaccard Coefficient and Simrank. The links having higher similarity score are of higher existence likelihood. Our proposed work deals with link dissolution. In order to find out the links which will get missed in future, we have implemented methods for

Unlink Prediction using Social Embeddedness with Network Overlap and Unlink Prediction using Social Embeddeness with Triadic Analysis. Unlink Prediction using Social Embeddeness with Triadic Analysis performs better than the existing methods.

ACKNOWLEDGMENT

The authors acknowledge the support of Sarannya. K, Poovayar Priya. M and Dhanalakshmi. I of Department of Computer Science and Engineering, Pondicherry Engineering College in their assistance of this work. This work was supported by the UGC Minor Research Project with Proposal Number: 2477(F.NO:4-4/2015-16(MRP/UGC-SERO).

REFERENCES

[1] Aggarwal Charu and Karthik Subbian. 2014. Evolutionary network analysis: A survey. ACM Computing Surveys (CSUR). 47(1): 1: 1-1: 10. https://doi.org/10.1145/2601412.

[2] Chen Hung-Hsuan, Liang Gou, Xiaolong Luke Zhang and C. Lee Giles. 2012. Discovering missing links in networks using vertex similarity measures. In Proceedings of the 27th Annual ACM Symposium on Applied Computing. pp. 138-143. ACM.

[3] Sharma Upasana and Bhawna Minocha. 2016. Link Prediction in Social Networks: A Similarity score based Neural Network Approach. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, p. 90. ACM.

[4] Popp József, Sándor Kovács, Péter Balogh and Attila Jámbor. 2016. Co-Authorship and Co-Citation Networks in the Agricultural Economics Literature: The Case of Central and Eastern Europe. Eastern European Economics. 54(2): 153-170. http://dx.doi.org/10.1080/00128775.2015.1135065.

[5] Leydesdorff Loet and Han Woo Park. 2016. Full and Fractional Counting in Bibliometric Networks. arXiv preprint arXiv:1611.06943. https://arxiv.org/abs/1611.06943.

[6] de Oliveira, Marcius Armada, Kate Cerqueira Revoredo, and José Eduardo Ochoa Luna. 2014. Semantic Unlink Prediction in Evolving Social Networks through Probabilistic Description Logic. In Intelligent Systems (BRACIS), 2014 Brazilian Conference on, pp. 372-377. IEEE.

[7] Julia Perl, Claudia Wagner, Jérôme Kunegis and Steffen Staab. 2014. A Theory-Driven Approach for Link and Unlink Predictions in Directed Social Networks, Computational Social Science Winter Symposium.

[8] Perl Julia, Claudia Wagner, Jerome Kunegis and Steffen Staab. 2015. Twitter as a Political Network: Predicting the Following and Unfollowing Behavior of German Politicians. In Proceedings of the ACM Web Science Conference, p. 51. ACM. https://doi.org/10.1145/2786451.2786506.

[9] Preusse Julia, Jérôme Kunegis, Matthias Thimm, and Sergej Sizov. 2014. DecLiNe--Models for Decay of Links in Networks. arXiv preprint arXiv: 1403.4415.

[10] Moody James and Douglas R. White. 2003. Structural cohesion and embeddedness: A hierarchical concept of social groups. American Sociological Review. 68(1): 103-127.

[11] Ferru, Marie, Michel Grossetti, and Marie-Pierre Bès. 2011. Measuring social embeddedness: how to identify social networks in science-industry partnerships.

[12] Des Promotionsausschusses, Vorsitz, and Karin Harbusch. 2014. On Structural Aspects of Unconnectedness in Knowledge and Social Networks. Ph.D thesis, Institute for Web Science and Technologies, University of Koblenz-Landau.