2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)

SAM: a Similarity Measure for Link Prediction in Social Network

Abdul Samad[1], Mamoona Qadir[2], Ishrat Nawaz[3]
[1,2,3] Department of Computer Science
[1] Capital University of Science and Technology
Islamabad, Pakistan
[2] Khwaja Fareed University of Engineering and Information Technology
Rahim Yar Khan, Pakistan
[3] The Islamia University of Bahawalpur
Bahawalpur, Pakistan
writetosamadalvi@gmail.com[1], Mamoonaqadir17@gmail.com[2], writetoishratnawaz@gmail.com[3]

***Abstract*— Research in the field of social network analysis attracting majority of the researchers nowadays. Out of many social network analysis problems, link prediction gaining high attention due to a growing number of social network users. Link prediction is a task to predict which new interaction is going to be occurring in the future. Traditional link prediction techniques considered pair of node as one unit and make decisions based on the commonality between them. We argued that both nodes in a pair have their own similarity to each other. It may be that one person is 100% similar to another, but the other person is not the same as the first. Moreover, we have proposed a similarity measure SAM for link prediction in the social network. We have compared SAM similarity with four other state-of-the-art link prediction techniques (i.e., Jaccard, Salton Index, Salton Cosine and Resource Allocation). The experiments in this paper are performed on five different datasets (i.e., Astro, CondMat, GrQc, HepPh and HepTh). Our results show that SAM performs better than rest of the link prediction techniques on all datasets.**

Keywords—Link Prediction; Social Network Analysis; node-Base Similarity; Topological-Based Similarity; Co-author network;

I. INTRODUCTION

The growing trend towards social networking is changing our lives and global business on a daily basis [13], which has been addressed in recent research. The social network is a place where two or more than two people having some relationship come and share their information, exchange views, make discussion and follows each other. Social networks can be off-line through face-to-face contacts [14] in schools, universities, conferences and other public places but it can be on-line by using Twitter [15], Face book [16], Google+ and LinkedIn. Social network is a social graph where people are represented as nodes and their contact is the edge between them. The edge in the social graph means they communicate or interact with each other even if they are geographically distant from each other. From the past few years, the attractiveness of people towards social networks such as Face book and Twitter have created many oopportunities for researchers to study and analyze characteristics of social network as well as various aspects of human behaviourr throughout the social network. Analysis of social network provides help to people in identifying individuals with mutual interest and their respective communities [17]. Moreover, we can find how communities build and how people interact to each other in social network.

Researchers, in the field of social network analysis, are facing too many problems [12]. Link prediction is one of them that mean predicting the unobserved or missing links in social network. The dynamic nature of social network makes this challenge more interesting. Consider a social network about 3 persons in Fig 1, in which solid link shows that "Sobia" is friend of "Ali" and "Abid" at time T. Another thing, there is no friendship between "Ali" and "Abid" at time T. It becomes interesting to think up the probability that there may be link between "Ali" and "Abid" at T_{1+1}. The goal of link prediction is here to predict the newly added friendship between persons.

A Link prediction has been attracting the attention of different domain's researchers. Researchers are researching in various domains, such as recommender systems which recommends new friends of common interest [1], bioinformatics in protein-protein interaction networks [2] and citations in citation network [3]. Additionally, link prediction have been used to predict future links in the networks (i.e., Face book), predict missing links [4], items for sale at Amazon [5], criminals in criminal networks [6]. Therefore, prediction of future links is very important for the analysis as well as completion of the network.

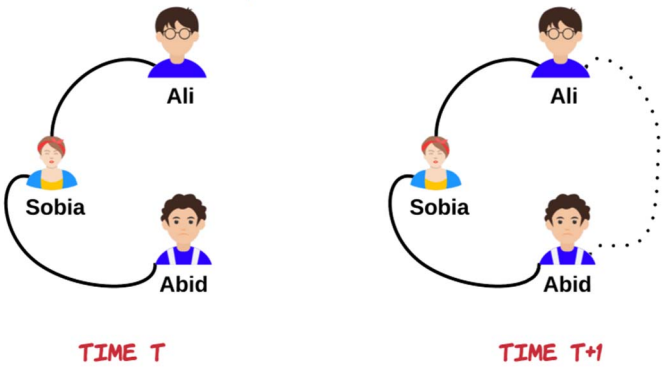

Fig. 1. Link Prediction in Social Network.

978-1-7281-4956-1/19/$31.00 ©2019 IEEE

There are many link prediction techniques which use information of nodes and topology in the social graph to compute the similarity between pair of nodes to predict the link. The computed similarity is represented as a score, which is assign to a pair of node (x, y). A high score indicates more chances that x and y will be linked in future. On the other hand, a low score represents high probability that x and y will not be linked in future. In the literature, state-of-the-art link prediction approaches [3,8,9,10,11,27,28], treat both nodes equally to compute the similarity score between pair of nodes, which are sometimes unable to predict future links.

In this paper, we have proposed a similarity measure SAM, which treats both nodes individually to compute similarity.

The remainder of this paper is as follows: Section 2 represents related work. The proposed similarity measure SAM is explained in section 3. The results of experiments are discussed in section 4. In the end, section 5 concludes the paper and presents some future directions.

II. LITERATURE REVIEW

A familiar task in prediction of link between two nodes x and y is to measure the similarity between them. The more similar nodes are, the more probability there exist a link between these nodes. The less similar nodes are, the less probability there exist a link between these nodes. To give accurate prediction of links between nodes in social network, it is preferable to use similarity measures to find out the most potential links. Using this approach, majority of researches have been done as shown below.

Wang et al. [19] presented local probabilistic graphical models which can extent to large scale graphs to measure the chance of link existence between two nodes. Tylenda et al. [18] escalated the local probability model by latest similarity measures namely Root PageRank and Adamic Adar (AA) which is denoted by time-aware link prediction. Three frequently used measures namely rooted PageRank, Katz and escape probability are used for link prediction by Song et al. [20]. Munasinghe et al. [21] presented the Time Score (TS) measure for link prediction between the pair of nodes that confide on nearby interaction time and the node number to search out the power of the relationship between strong bonds and engagement timestamp connected to the future. Soares et al. [22] presented work that computes and gives similarity score to every pair of detached nodes through topological similarity measures. Zhang and Philip [23] considered 2-hop for similarity computation where they proposed a edge-based similarity search method that computes the similarity between two nodes as the 2-hop similarity. Ibrahim and Chen [24] incorporated structure of community, centrality of node and temporal information in predicting possible links in social networks. Han et al. [25] worked on similarity between communities where they proposed Community Similarity Degree (CSD) metric to guess the degree of interest similarity between multiple users in a community. Murata et al. [26] proposed two weighted similarity scores namely WAA and WCN escalated from Common Neighbors [27] and Adamic Adar [28] respectively. Based on Murata's work [26] , Ismail et al. [29] also proposed Weighted Jaccard Coefficient (WJC). Some more research on this road can be seen here in [30, 31, 32, 33, 34].

To find new or missing contact links between humans in a social network is considered as a challenging problem. Wang et al. [35] give promising results for Graph inference problem. Using homophily theory with which missing edges in a graph can be predicted with limited knowledge of people's social profile as nodes. Also using offline data they have predicted people contacts through their mobility patterns and by giving weight between edges representing the strength between contact and hence predict the missing parts of graph. Samad et al. [11] analyzed social profiles of humans and profiles' impact on link prediction and prove that all social features of humans have different importance. In a graph, between two nodes on the edges weight represent the bonding of different contacts. Samad et al. [11] says on the basis of most common and dominated features of different profiles the weight on the edges must be different. In Similarity measures the best accuracy on Nationality feature scores 0.92. In simple words, humans prefer to contact with other humans belonging to identical language and nationality. Junuthula et al. [36] focus on Link prediction for online social network (OSN's) using predictive power of combining friendship and interaction networks. Using friendship networks they improved predictions of future edges in interaction networks. In a friendship network there is a 'follow' link exist but in interaction network two nodes have conversation on a particular day. For predicting future interaction networks predictors are split into three categories, predictors that do not use friendships, predictors that use only current friendships and predictors that use predicted friendships. Results are incorporating current friendships does indeed result in a significantly better link predictor for interaction networks. Zhou et al. [37] worked on different algorithms for the problem of attacking similarity-based link prediction for deleted link nodes in the network. For this approach they define two broad classes, one is local information to make certain optimal attacks for CND metrics about target links in a group, and another is global network information for well-motivated special cases which uses NP-Hard metrics . To track these missing or deleted links they use polynomial-time algorithms. Lim et al. [38] worked on hidden links prediction using criminal network. They have explored that supervised machine learning metrics required large datasets for training and testing. Therefore, predicting hidden links they have used the application of deep reinforcement learning (DRL) for reconstructing criminal networks. In their experiments, they have concluded that link prediction through DRL presents better performance than supervised machine learning. Moreover, Lim et al. [39] compared performance of DRL with supervised machine learning in terms of predictive accuracy and computing power.

III. RESEARCH METHODOLOGY

Five different co-author social networks, represented by undirected graph $G= (V, E)$, are discussed in this paper. In the social graph, E denotes the set of links between authors and V is set of nodes that represent the authors. Our work is to give a score, $S(x; y)$, to each pair of nodes $(x; y)$. The possibility of a link between nodes is indicated by the given score. The

higher *S(x; y)* represents the higher possibility that there is a link between node *x* and *y* for a pair of nodes (*x; y)*.

A. Dataset Description

In this paper, we have used 5 different datasets [7] (i.e., *AstroPh, CondMat, GrQc, HepPh and HepTh*) containing co-author network. These datasets covers research participation between authors and derived from *e-print arXiv*. If an author *x* co-authored a paper with author *y*, the graph contains an undirected edge between *x* and *y*. if there are *z* co-authors on a paper, this generates completely (sub) graph on *z* nodes. Moreover, detailed statistics of datasets are given in Table 1.

TABLE I. DATASETS STATISTICS

Dataset	Properties		
	Nodes	***Edges***	***Triangles***
AstroPh	18772	198110	1351441
CondMat	23133	93497	173361
GrQc	5242	14496	48260
HepPh	12008	118521	3358499
HepTh	9877	25998	28339

B. Jaccard Coefficient

Jaccard Coefficient [11] considered both total and common number of neighbors for similarity computation between pair of nodes. Jaccard Coefficient between two nodes u and v is computed as follows using Equation 1.

$$Jc(u,v) = \frac{|\tau(u) \cap \tau(v)|}{|\tau(u) \cup \tau(v)|} \qquad (1)$$

Here in this Equation 1, τ(u) means set of nodes that are adjacent to node u and $|\tau(u) \cap \tau(v)|$ is number of common neighbors between pair of nodes u and v

C. Sorensen Index

Sorensen similarity [9] measures are defined as Equation 2. It's considered the size of the common neighbors and also points out that lower degree of nodes would have higher link likelihood.

$$SI(u,v) = \frac{|\tau(u) \cap \tau(v)|}{|\tau(u)| + |\tau(v)|} \qquad (2)$$

D. Sorensen Cosine

Salton cosine [9] is common cosine metric that is used to compute similarity between pair of nodes. This metric is defined as Equation 3.

$$Sc(u,v) = \frac{|\tau(u) \cap \tau(v)|}{\sqrt{|\tau(u)| \cdot |\tau(v)|}} \qquad (3)$$

E. Resource Allocation

Resource allocation is proposed by Zhou et al. [8], and punishes the higher degree nodes more heavily. This similarity measure performs better for the network which has high average degrees. Another most interesting fact about resource allocation is that it uses not only neighbors but also neighbor of neighbors. Similarity between pair of nodes using resource allocation is computed as follows by Equation 4.

$$RA\ (u,v) = \sum_{z \epsilon |\tau(u) \cap \tau(v)|} \frac{1}{|\tau(z)|} \quad (4)$$

F. Sam Similarity

In this paper, we have proposed Sam similarity measure, which considered similarity of *X* towards *Y* as well as similarity of Y towards *X*. It divides the task of similarity computation into two parts. First, it gets know that how much u is similar with v using Equation 5. Second, it computes similarity from v towards u using Equation 6. Finally, uses both results from Equation 5 and 6 using Equation 7. Here we punish higher degree nodes in order to find the links between higher degree nodes and lower degree nodes. In the literature, most of the techniques ignore link between higher degree and lower degree nodes. Our technique punishes heavily to higher degree nodes and creates more chances for lower degree nodes to link with higher degree nodes. Moreover, state-of-the-art traditional link prediction techniques considered pair of node as one unit and make decisions based on commonality between them. We argued that both nodes in a pair have their own similarity to each others. It may be that one person is 100% similar to another, but the other person is not the same as the first.

$$Sam(u_v) = \frac{|\tau(u) \cap \tau(v)|}{|\tau(u|} \quad (5)$$

$$Sam(v_u) = \frac{|\tau(u) \cap \tau(v)|}{|\tau(v)|} \quad (6)$$

$$Sam(u,v) = {(Sam(u_v) + Sam(u_v))}/{2} \quad (7)$$

For example, consider a social network shown in Fig 2; "Abid" is husband of "Sobia" which are friends to each other. On the other hand, "Sobia" is mother of "Ali" and connected by edge to each other. If we predict the link between "Ali" and "Abid", state-of-the-art techniques gives low similarity score as they did not support the real world scenarios most of the time. After applying Jaccard we gets 0.14 score which shows that "Abid" and "Ali" are 14% similar to each other. On the other hand, after applying Sam Equation 5 and 6, "Abid" is 100% similar to "Ali" and "Ali" 14% similar to "Abid". Using Equation 7, we can say that both are 57% similar.

G. Generating Lists of Edges

We have performed experiments over 5 social networks of co-authors. In order to predict links from existing social network we have randomly picked 3 different sets (i.e., Edges-1, Edges-5 and Edges-10) of edges for each dataset.

- First, we randomly picked 1000 edges from Astro dataset and formed edge list called *Astro-Edge-1*.
- For the second edge list, 5000 edges are randomly picked for prediction and formed another edge list *Astro-Edge-5*.
- For the third edge list, we have picked randomly 10000 edges in order to predict them. We called this list as *Astro-Edge-10*.

After applying above 3 steps for remaining datasets, we have obtained more 12 edges lists (i.e., *CondMat-Edge-1, CondMat-Edge-5, CondMat-Edge-10, GrQc-Edge-1, GrQc-Edge-5, GrQc-Edge-10, HepPh-Edge-1, HepPh-Edge-5, HepPh-Edge-10, HepTh-Edge-1, HepTh-Edge-5, HepTh-Edge-10*) for the prediction.

H. Evluation

In order to compare our technique with state-of-the-art link prediction techniques, we have checked the percentage of predicted social links. Moreover, Accuracy measure (ratio of correct predictions to the total number of input edges) is used to evaluate similarity measures.

$$Accuracy = 1 - \frac{E(G\gamma) + E(G\rho) - 2E(G\gamma \cap G\rho)}{Max(E(G\gamma), E(G\rho))} \qquad (8)$$

Here in Equation 8:

- E represents the Edges of social graph,
- Gγ represents the original social graph,
- Gρ represents the predicted social graph,
- Max function will return the maximum number of edges from the original and predicted graph.

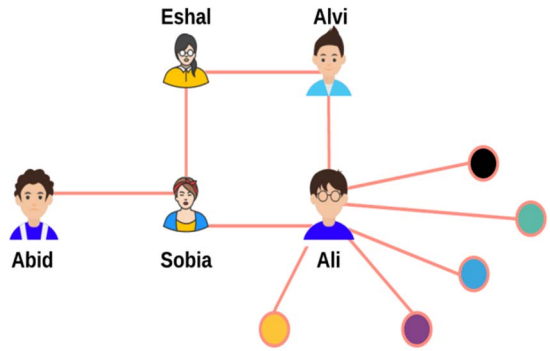

Fig. 2. Example of Social Network.

IV. Experiments and Results

Five co-author datasets (i.e., *Astro, CondMat, GrQc, HepPh, and HepTh*) are used in experiments. Initially, these datasets contains a social graph of authors with different number of nodes and edges as stats are shown in Table 1. First, we extracted 3 edges lists for each dataset. The experiments are performed on these 15 edge lists. Secondly, we remove these edges from every social graph and applied similarity measures on every pair of node in each social graph. After applying similarity measures, we have applied some thresholds on similarity between each pair of node and created 100 predicted social graphs, 20 for each similarity measure and dataset on four different thresholds. Finally, we have checked the accuracy of every similarity measure.

A. Comparisons using Astro Dataset

For computing similarity, 16 thousand edges are used for prediction from *Astro* dataset. These edges are divided in three edge lists, where first edge list (i.e., *Astro-Edge-1*) contains 1000 edges, second (i.e., *Astro-Edge-5*) contained 5000 and last edge list (i.e., *Astro-Edge-10*) contained 10000 edges. Figures 3, 4 and 5 are showing the results from *Astro* dataset. Results from *Astro-Edge-1* are shown in Fig 3, where threshold is shown on X-axis and percentage of prediction in the form of accuracy is shown on Y-axis. The same pattern is followed in all figures. Resultant thresholds showed that at threshold 0.1, Sam achieved the highest results with 93% predicted links. In detail, the minimum accuracy achieved by Sam is 18% and maximum is 93%. Moreover, Sam, Salton Cosine and Resource Allocation achieved good results on all thresholds. On the other hand, Jaccard and Salton Index could not obtain reasonable results. Overall, the minimum accuracy achieved by Jaccard, Salton Index, Salton Cosine and Resource Allocation is 9%, 6%, 16% and 16%.

Similarly, Fig 4 addresses the prediction results of 5000 edges (i.e., *Astro-Edge-5*) and Fig 5 showing 10,000 edges (i.e., Astro-Edge-10) from *Astro* dataset. At threshold 0.1, in both Fig 4 and Fig 5, Sam obtained maximum 93% predicted links. Similarly, at threshold 0.1, the maximum result obtained by Salton Cosine and Resource Allocation is 90% and 90%. On the other hand, Jaccard and Salton Index predicted maximum 70% and 66% links correctly. Overall, using *Astro* dataset, the accuracy of Sam is better than remaining link prediction techniques.

B. Comparisons using CondMat Dataset

For the purpose of similarity computation, 16 thousand edges out of 93497 are used for prediction from *CondMat* dataset. We have divided edges in three edge lists (i.e., *CondMat-Edge-1, CondMat-Edge-5 and CondMat-Edge-10*). Where first edge list (i.e., *CondMat-Edge-1*) contains 1000 edges, second (i.e., *CondMat-Edge-5*) contained 5000 and last edge list (i.e., *CondMat-Edge-10*) contained 10000 edges. Figures 6, 7 and 8 are addressing the results from *CondMat* dataset.

Fig 6 addresses the results of *CondMat-Edge-1*, where threshold is shown on X-axis and percentage of prediction in form of accuracy is shown on Y-axis. At threshold 0.1, Sam achieved the highest results with 90% predicted links. Moreover, the minimum accuracy achieved by Sam is 18%. At thresholds 0.1 and 0.3, Sam, Salton Cosine and Resource Allocation achieved good results. On the other hand, Jaccard and Salton Index again could not obtain suitable results. Overall, the minimum accuracy achieved by Jaccard, Salton Index, Salton Cosine and Resource Allocation is 6%, 3%, 12% and 14%. In the case of thresholds 0.5 and 0.7, although all similarity measures unable to produced better results but we can see better results from Sam than others.

Likewise, Fig 7 showing the prediction results of 5000 edges (i.e., *CondMat-Edge-5*) and Fig 8 showing 10,000 edges (i.e., *CondMat-Edge-10*) from *CondMat* dataset. In the case of threshold 0.1, in both Fig 4 and Fig 5, Sam obtained maximum 91% predicted links. In the same way, at threshold 0.1, the maximum result obtained by Salton Cosine and Resource Allocation is 89% and 87%. On the other hand, Jaccard and Salton Index predicted maximum of 67% and 64% links

correctly. Overall, using *CondMat* dataset, the accuracy of Sam performed better than other link prediction techniques.

TABLE II. RESULTS OF ALL SIMILARITY MEASURES FOR ALL DATSETS.

Dataset	No of Edges to Predict	*Jaccard*		*Salton Index*		*Salton Cosine*		*Resource Allocation*		*Sam*	
		Max	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*	*Min*	*Max*	*Min*
Astro	1,000	71%	9%	68%	**6%**	91%	16%	92%	16%	**93%**	18%
	5,000	68%	9%	65%	**4%**	90%	16%	90%	17%	**93%**	19%
	10,000	70%	9%	66%	**5%**	90%	16%	90%	18%	**93%**	20%
CondMat	1,000	68%	6%	65%	**3%**	88%	12%	87%	14%	**90%**	18%
	5,000	67%	5%	64%	**2%**	89%	11%	87%	12%	**91%**	16%
	10,000	67%	**2%**	64%	**2%**	89%	6%	86%	6%	**91%**	9%
GrQc	1,000	74%	23%	73%	**15%**	87%	32%	81%	36%	**88%**	39%
	5,000	72%	21%	70%	**17%**	86%	31%	82%	34%	**87%**	39%
	10,000	72%	21%	70%	**14%**	86%	31%	82%	34%	**87%**	37%
HepPh	1,000	82%	30%	80%	**22%**	92%	45%	92%	47%	**95%**	50%
	5,000	82%	30%	80%	**18%**	92%	47%	92%	50%	**95%**	49%
	10,000	83%	31%	81%	**23%**	92%	48%	92%	51%	**94%**	56%
HepTh	1,000	55%	5%	52%	**4%**	78%	7%	71%	9%	**79%**	12%
	5,000	55%	5%	51%	**4%**	78%	7%	73%	10%	**80%**	7%
	10,000	54%	5%	51%	**4%**	78%	7%	72%	10%	**79%**	14%

C. Comparisons using GrQc Dataset

For link prediction, three different sets of edges are picked from *GrQc* dataset. Furthermore, these edges are divided in three edge lists, where first edge list (i.e., *GrQc-Edge-1*) contains 1000 edges, second (i.e., *GrQc-Edge-5*) contained 5000 and last edge list (i.e., *GrQc-Edge-10*) contained 10000 edges. Figures 9, 10 and 11 are showing the results from *GrQc* dataset.

Results from *GrQc-Edge-1* are shown in Fig 9. Resultant thresholds showed that at threshold 0.1, Sam achieved the highest results with 88% predicted links. In detail, the minimum accuracy achieved by Sam is 39% and maximum is 88%. Moreover, Sam, Salton Cosine and Resource Allocation achieved good results on all thresholds. On the other hand, Jaccard and Salton Index obtained a maximum of 74% and 73% which are a little bit good than obtained in the previous two datasets (i.e., *Astro* and *CondMat*). The same behavior is followed by Salton Cosine and Resource Allocation as they obtained maximum 87% and 81% predicted links. Overall, the minimum accuracy achieved by Jaccard, Salton Index, Salton Cosine and Resource Allocation is 23%, 15%, 32% and 36% which is better as compared with previous two datasets (i.e., *Astro* and *CondMat*).

Similarly, Fig 10 addresses the prediction results of 5000 edges (i.e., GrQc-Edge-5) and Fig 11 showing 10,000 edges (i.e., GrQc-Edge-10) from *GrQc* dataset. Again Sam outperformed than other similarity measures on all thresholds. At threshold, 0.1, in both Fig 4 and Fig 5, Sam succeed in getting maximum 87% predicted links. Similarly, at threshold 0.1, the maximum results obtained by Salton Cosine and Resource Allocation are 86% and 82%. On the other hand, Jaccard and Salton Index predicted maximum 72% and 70% links correctly. Overall, using *GrQc* dataset, Sam predicted links more than rest of the techniques.

D. Comparisons using HepPh Dataset

For the purpose of similarity computation, 16 thousand edges out of 118521 are used for prediction from *HepPh* dataset. We have divided edges in three edge lists (i.e., *HepPh-Edge-1, HepPh-Edge-5 and HepPh-Edge-10*). Where first edge list (i.e., *HepPh-Edge-1*) contains 1000 edges, second (i.e., *HepPh-Edge-5*) contained 5000 and last edge list (i.e., *HepPh-Edge-10*) contained 10000 edges. Figures 12, 13 and 14 are showing the results from *HepPh* dataset.

Fig 12 is showing the results of *HepPh-Edge-1* edge list, where threshold is shown on X-axis and percentage of prediction in form of accuracy is shown on Y-axis. At threshold 0.1, *Sam* obtained the highest results with 95% predicted links. On the other hand, the minimum accuracy achieved by Sam is 50%. At thresholds 0.1 and 0.3, Sam, Salton Cosine and Resource Allocation achieved good results. On the other hand, Jaccard and Salton Index again could not obtain suitable results. Overall, the minimum accuracy achieved by Jaccard, Salton Index, Salton Cosine and Resource Allocation is 30%, 22%, 45% and 47%. In case of thresholds 0.5 and 0.7, all similarity measures produced better results than previous datasets (i.e., *Astro*, *CondMat* and *GrQc*). Another interesting thing which we can see Sam outperformed that rest of the techniques.

Similarly, Fig 13 addresses the prediction results of 5000 edges (i.e., *HepPh-Edge-5*) and Fig 8 showing 10,000 edges (i.e., *HepPh-Edge-10*) from *HepPh* dataset. In case of threshold 0.1, in both Fig 4 and Fig 5, Sam obtained maximum 95% predicted links. In the same way, at threshold 0.1, the maximum result obtained by Salton Cosine and Resource Allocation is 92% and 92%. On the other hand, Jaccard and Salton Index predicted maximum 83% and 81% links correctly. Overall, using *HepPh* dataset, the accuracy of Sam performed better than other link prediction techniques.

E. Comparisons using HepTh Dataset

For link prediction on *HepTh* dataset, three different sets of edges (i.e., *HepTh-Edge-1, HepTh-Edge-5 and HepTh-Edge-10*) are picked. Where *HepTh-Edge-1* is contains 1000 edges, *HepTh-Edge-5* contains 5000 and *HepTh-Edge-10* contains 10000 edges. Figures 15, 16 and 17 are showing the results from *HepTh* dataset.

Results from *HepTh-Edge-1* are shown in Fig 9. Resultant thresholds showed that at threshold 0.1, our proposed approach Sam achieved 79% predicted links highest than rest of the techniques. In detail, the minimum accuracy achieved by Sam is 12% and maximum is 79%. Moreover, Sam, Salton Cosine and Resource Allocation achieved good results on all thresholds. On the other hand, Jaccard and Salton Index obtained maximum 55% and 52% which are little bit unsatisfactory than obtained in previous datasets (i.e., *Astro*, *CondMat*, *GrQc* and *HepPh*). The same behavior is followed by Salton Cosine and Resource Allocation as they obtained maximum 78% and 71% predicted links. Overall, the minimum accuracy achieved by Jaccard, Salton Index, Salton Cosine and Resource Allocation is 5%, 4%, 7% and 9% which is too lowest as compared with previous datasets (i.e., *Astro*, *CondMat*, *GrQc* and *HepPh*).

Similarly, Fig 16 addresses the prediction results of 5000 edges (i.e., *HepTh-Edge-5*) and Fig 11 showing 10,000 edges (i.e., *HepTh-Edge-10*) from *HepTh* dataset. Again Sam outperformed than other similarity measures on all thresholds. At threshold, 0.1, in both Fig 16 and Fig 17, Sam succeed n getting maximum 80% predicted links. Similarly, at threshold 0.1, the maximum results obtained by Salton Cosine and Resource Allocation are 78% and 73%. On the other hand, Jaccard and Salton Index predicted maximum 55% and 51% links correctly. Overall, using *HepTh* dataset, Sam predicted links more than rest of the techniques.

F. Summary of Comparisons

Table 2 is representing the overall results of our experiments. All the similarity measures performed well on HepPh datasets by getting maximum results. The maximum results achieved by Sam are 95% on dataset *HepPh*. On the other hand, Jaccard obtained maximum 83%, Salton Index achieved 80%, Salton Cosine obtained 92% and Resource Allocation obtained 92%. All similarity measures, except Sam, performed worst on *CondMat* dataset by getting low results. Sam achieved minimum 7% on HepTh dataset, Jaccard obtained minimum 2%, Salton Index achieved 2%, Salton Cosine obtained 6% and Resource Allocation Obtained 6%.

Moreover, in Table 2, second column representing the number of edges used for prediction from every dataset. In each row, maximum predicted results are represented in green color and minimum are in blue. For every edge list, in each dataset, Sam achieved maximum results and Salton Index obtained minimum results. In case of 1000 edges prediction, Sam achieved 95%, Jaccard obtained 82%, Salton Index obtained 80%, Salton Cosine and Resource Allocation achieved 92%. In case of 5000 edges, Sam obtained 95%, Jaccard achieved 82%, Salton Index obtained 80%, Salton Cosine achieved and Resource Allocation obtained 92%. Similarly, in case of 10000 edges, all similarity measures followed the same behavior.

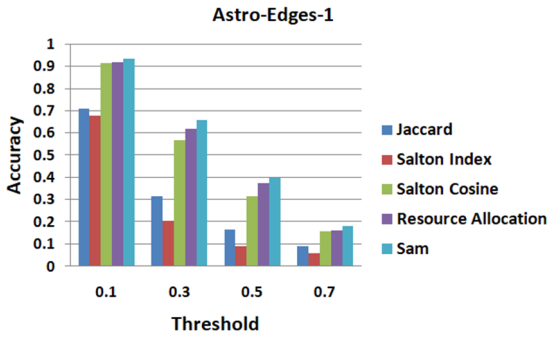

Fig. 3. Results of Astro dataset for the prediction of 1000 edges.

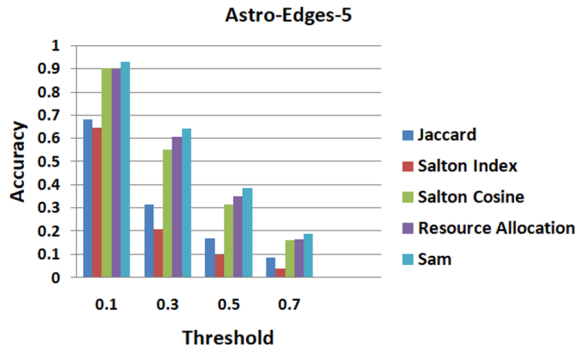

Fig. 4. Results of Astro dataset for the prediction of 5000 edges.

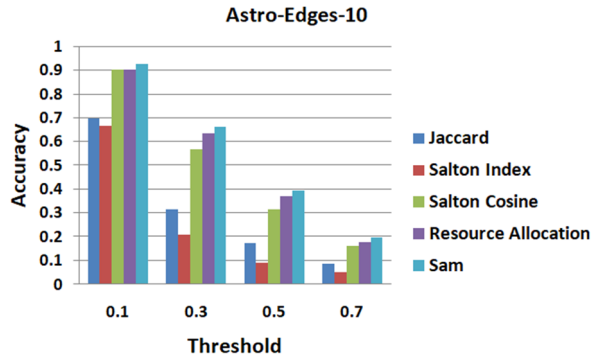

Fig. 5. Results of Astro dataset for the prediction of 10,000 edges.

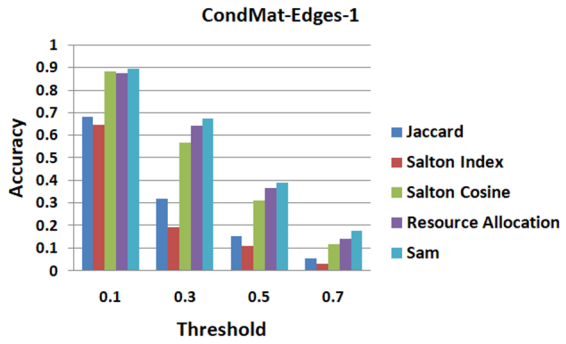

Fig. 6. Results of CondMat dataset for the prediction of 1000 edges.

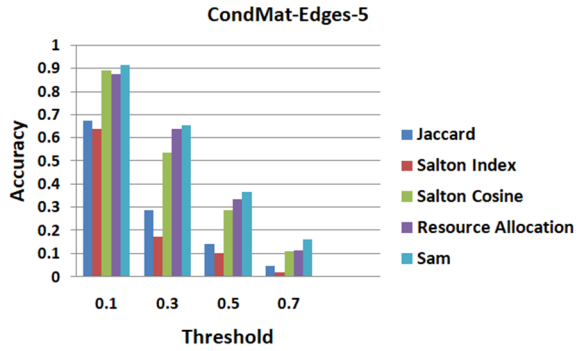

Fig. 7. Results of CondMat dataset for the prediction of 5000 edges.

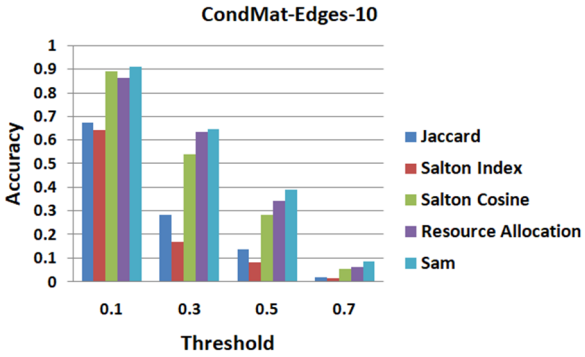

Fig. 8. Results of CondMat dataset for the prediction of 10,000 edges.

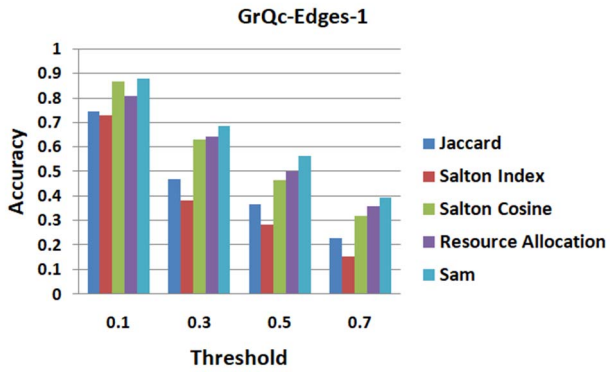

Fig. 9. Results of GrQc dataset for the prediction of 1000 edges.

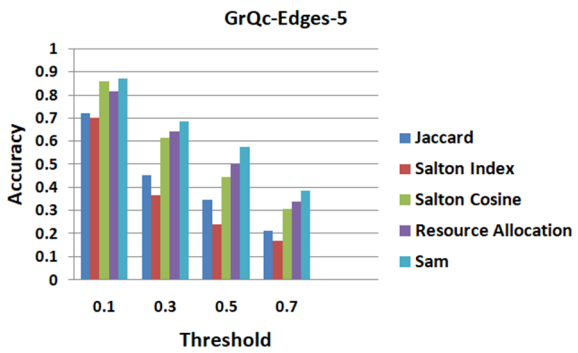

Fig. 10. Results of GrQc dataset for the prediction of 5000 edges.

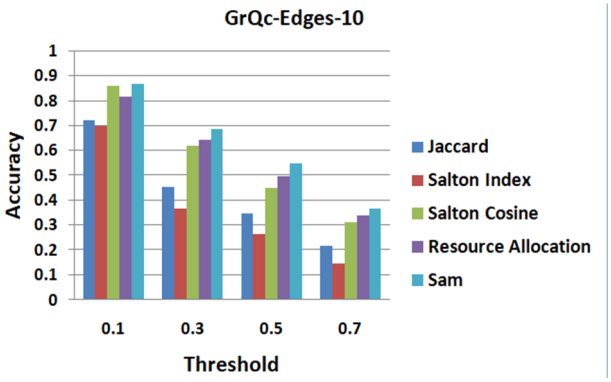

Fig. 11. Results of GrQc dataset for the prediction of 10,000 edges.

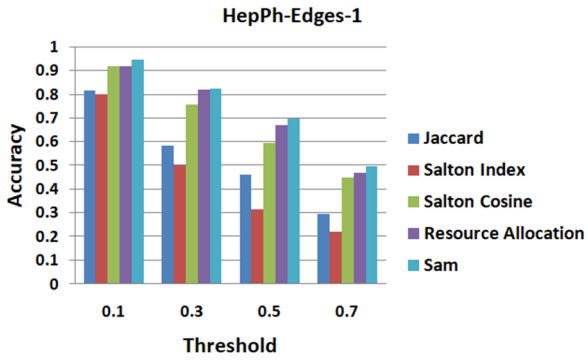

Fig. 12. Results of HepPh dataset for the prediction of 1000 edges.

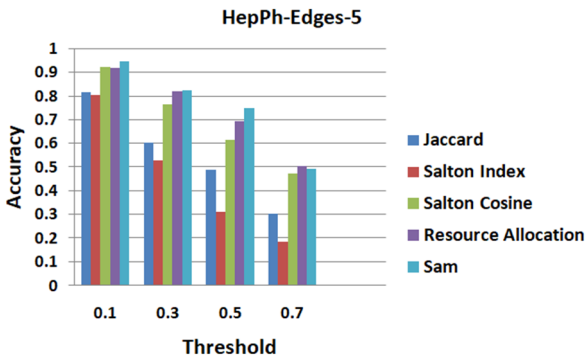

Fig. 13. Results of HepPh dataset for the prediction of 5000 edges.

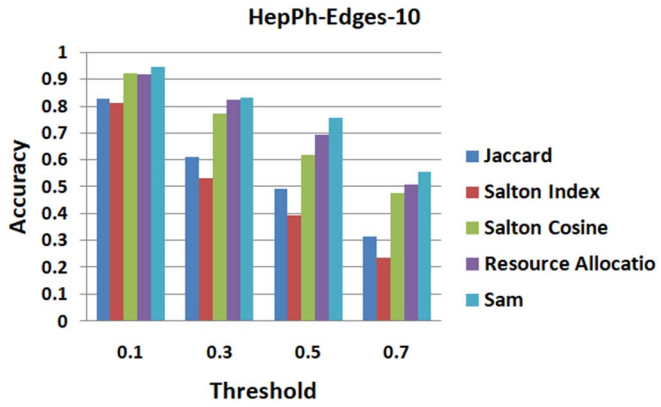

Fig. 14. Results of HepPh dataset for the prediction of 10,000 edges.

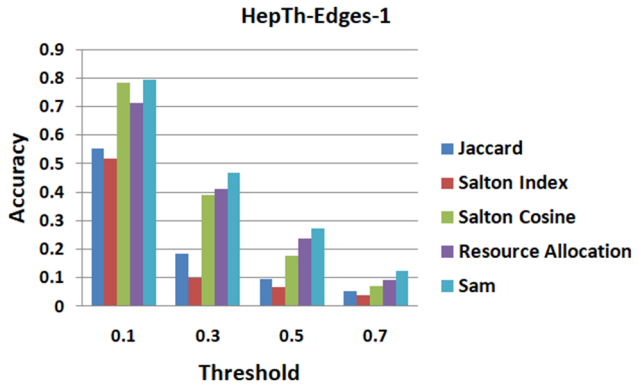

Fig. 15. Results of HepTh dataset for the prediction of 1000 edges.

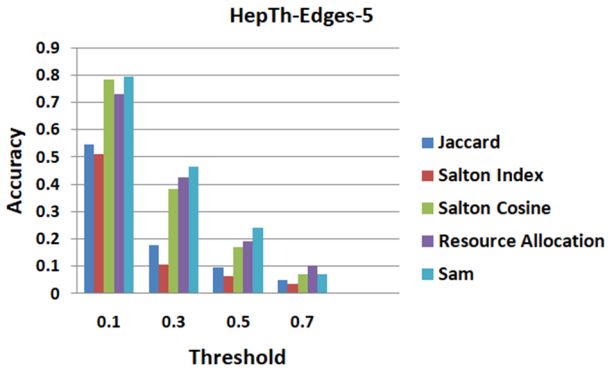

Fig. 16. Results of HepTh dataset for the prediction of 5000 edges.

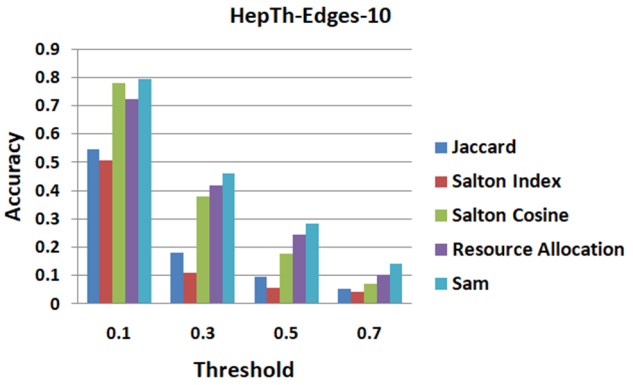

Fig. 17. Results of HepTh dataset for the prediction of 10,000 edges.

V. Conclusion

In this paper, we have experimented on 5 different co-authors datasets. Moreover, we have proposed Sam similarity measure and compared with four other state-of-the-art link prediction techniques. Our proposed method considers that nodes in pair have their own similarity to each other. Our results show that Sam outperformed than rest of the link prediction techniques on all datasets. In future, we will give some weights to similarity of every node in pair.

Acknowledgment

The authors would thankful to Dr. Muhammad Arshad Islam for his valuable discussion on the current topic.

References

[1] Z. Huang, X. Li, H. Chen, Link prediction approach to collaborative filtering, in: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, JCDL'05, ACM, 2005, pp. 141–142.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] C. Lei, J. Ruan, A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity, Bioinformatics 29 (3) (2013) 355–364.K. Elissa, "Title of paper if known," unpublished.

[3] Samad, Abdul. Evaluation of Textual and Topological Similarity Measures for Citation Recommendation. Diss. CAPITAL UNIVERSITY, 2019.

[4] W. Peng et al., "Link Prediction in Social Networks: the State-of-the-Art," Sci China Inf Sci, vol. 58, no. 58, pp. 11101–38, 2015.

[5] B. Cao, N. N. Liu, and Q. Yang, "Transfer Learning for Collective Link Prediction in Multiple Heterogenous Domains," Int. Conf. Mach. Learn., pp. 180–186, 2010.

[6] G. Berlusconi, F. Calderoni, N. Parolini, M. Verani, and C. Piccardi, "Link prediction in criminal networks: A tool for criminal intelligence analysis," PLoS One, vol. 11, no. 4, p. e0154244, 2016.

[7] Leskovec, J., Kleinberg, J., & Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 2.

[8] Zhou, T., Lü, L., & Zhang, Y. C. (2009). Predicting missing links via local information. The European Physical Journal B, 71(4), 623-630.

[9] Wang, P., Xu, B., Wu, Y., & Zhou, X. (2015). Link prediction in social networks: the state-of-the-art. Science China Information Sciences, 58(1), 1-38.

[10] Moradabadi, B., & Meybodi, M. R. (2018). Link prediction in weighted social networks using learning automata. Engineering Applications of Artificial Intelligence, 70, 16-24.

[11] Samad, A., Islam, M. A., Iqbal, M. A., Aleem, M., & Arshed, J. U. (2017, December). Evaluation of features for social contact prediction. In 2017 13th International Conference on Emerging Technologies (ICET) (pp. 1-6). IEEE.

[12] Alvi, Abdul & Islam, Arshad & Iqbal, Muhammad & Aleem, Muhammad. (2019). Centrality-Based Paper Citation Recommender System. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems. 6. 159121. 10.4108/eai.13-6-2019.159121.

[13] Mallek, S., Boukhris, I., Elouedi, Z., & Lefèvre, E. (2019). Evidential link prediction in social networks based on structural and social information. Journal of computational science, 30, 98-107.

[14] Pérez-Macías, N., Fernández-Fernández, J. L., & Rua Vieites, A. (2019). Entrepreneurial intentions: trust and network ties in online and face-to-face students. Education+ Training, 61(4), 461-479.

[15] Ahuja, R., Singhal, V., & Banga, A. (2019). Using Hierarchies in Online Social Networks to Determine Link Prediction. In Soft Computing and Signal Processing (pp. 67-76). Springer, Singapore.

[16] Lai, Y. Y., Neville, J., & Goldwasser, D. (2019). TransConv: Relationship Embedding in Social Networks.

[17] Yuan, W., He, K., Guan, D., Zhou, L., & Li, C. (2019). Graph kernel based link prediction for signed social networks. Information Fusion, 46, 1-10.

[18] Tylenda, T., Angelova, R., & Bedathur, S. (2009, June). Towards time-aware link prediction in evolving social networks. In Proceedings of the 3rd workshop on social network mining and analysis (p. 9). ACM.

[19] Wang, C., Satuluri, V., & Parthasarathy, S. (2007, October). Local probabilistic models for link prediction. In Seventh IEEE international conference on data mining (ICDM 2007) (pp. 322-331). IEEE.

[20] Song, H. H., Cho, T. W., Dave, V., Zhang, Y., & Qiu, L. (2009, November). Scalable proximity estimation and link prediction in online social networks. In Proceedings of the 9th ACM SIGCOMM conference on Internet measurement (pp. 322-335). ACM.

[21] Munasinghe, L., & Ichise, R. (2011, August). Time aware index for link prediction in social networks. In International Conference on Data Warehousing and Knowledge Discovery (pp. 342-353). Springer, Berlin, Heidelberg.

[22] da Silva Soares, P. R., & Prudêncio, R. B. C. (2012, June). Time series based link prediction. In The 2012 international joint conference on neural networks (IJCNN) (pp. 1-7). IEEE.

[23] Zhang, J., & Philip, S. Y. (2014). Link prediction across heterogeneous social networks: A survey. Social networks.

[24] Ibrahim, N. M. A., & Chen, L. (2015). Link prediction in dynamic social networks by integrating different types of information. Applied Intelligence, 42(4), 738-750.

[25] Han, X., Wang, L., Farahbakhsh, R., Cuevas, Á., Cuevas, R., Crespi, N., & He, L. (2016). CSD: A multi-user similarity metric for community recommendation in online social networks. Expert Systems with Applications, 53, 14-26.

[26] Murata, T., & Moriyasu, S. (2007, November). Link prediction of social networks based on weighted proximity measures. In Proceedings of the IEEE/WIC/ACM international conference on web intelligence (pp. 85-88). IEEE Computer Society.

[27] Newman, M. E. (2001). Clustering and preferential attachment in growing networks. Physical review E, 64(2), 025102.

[28] Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. Social networks, 25(3), 211-230.

[29] Güneş, İ., Gündüz-Öğüdücü, Ş., & Çataltepe, Z. (2016). Link prediction using time series of neighborhood-based node similarity scores. Data Mining and Knowledge Discovery, 30(1), 147-180.

[30] Sarna, G., & Bhatia, M. P. S. (2017). Content based approach to find the credibility of user in social networks: an application of cyberbullying. International Journal Of Machine Learning and Cybernetics, 8(2), 677-689.

[31] Srilatha, P., & Manjula, R. (2016). Similarity index based link prediction algorithms in social networks: A survey. Journal of Telecommunications and Information Technology.

[32] Nguyen, T. T., Harper, F. M., Terveen, L., & Konstan, J. A. (2018). User personality and user satisfaction with recommender systems. Information Systems Frontiers, 20(6), 1173-1189.

[33] Kaya, B., & Poyraz, M. (2016). Unsupervised link prediction in evolving abnormal medical parameter networks. International Journal of Machine Learning and Cybernetics, 7(1), 145-155.

[34] Moradabadi, B., & Meybodi, M. R. (2016). Link prediction based on temporal similarity metrics using continuous action set learning automata. Physica A: Statistical Mechanics and its Applications, 460, 361-373.

[35] Wang, D., Pedreschi, D., Song, C., Giannotti, F., & Barabasi, A. L. (2011, August). Human mobility, social ties, and link prediction. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1100-1108). Acm.

[36] Junuthula, R. R., Xu, K. S., & Devabhaktuni, V. K. (2018, June). Leveraging friendship networks for dynamic link prediction in social interaction networks. In Twelfth International AAAI Conference on Web and Social Media.

[37] Zhou, K., Michalak, T. P., Waniek, M., Rahwan, T., & Vorobeychik, Y. (2019, May). Attacking Similarity-Based Link Prediction in Social Networks. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (pp. 305-313). International Foundation for Autonomous Agents and Multiagent Systems.

[38] Lim, M., Abdullah, A., Jhanjhi, N. Z., & Supramaniam, M. (2019). Hidden Link Prediction in Criminal Networks Using the Deep Reinforcement Learning Technique. Computers, 8(1), 8.

[39] Lim, M., Abdullah, A., & Jhanjhi, N. Z. (2019). Performance optimization of criminal network hidden link prediction model with deep reinforcement learning. Journal of King Saud University-Computer and Information Sciences.