EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Structural Importance-based Link Prediction Techniques in Social Network

Abdul Samad [1,*], Muhammad Azam[2], Mamoona Qadir[3]

[1]Capital University of Science and Technology, Islamabad Pakistan
[2]The University of Agriculture Faisalabad
[3]Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan Pakistan

Abstract

Link prediction in social network gaining high attention of researchers nowadays due to the rush of users towards social network. Link prediction is known as the prediction of missing or unobserved link, i.e., new interaction is going to be occurring in a near future. State-of-the-art link prediction techniques(e.g., Jaccard Index, Resource Allocation, SAM Similarity, Sorensen Index, Salton Cosine, Hub Depressed Index and Parameter-Dependent) considers only similarity of the pair of node in order to find the link. However, we argued that nodes having same status of centralization along with high similarity can connect to each other in a future. In this paper, we have proposed structural importance-based state-of-the-art link prediction techniques and compared. We have compared structural importance-based link prediction techniques with state-of-the-art techniques. The experiments are performed on four different datasets (i.e., Astro, CondMat, HepPh and HepTh). Our results show that structural importance-based link prediction techniques outperformed than state-of-the-art link prediction techniques by getting 95% at threshold 0.1 and 68% at threshold 0.7.

Received on 04 December 2020; accepted on 19 December 2020; published on 07 January 2021

Keywords: Link Prediction, Social Network Analysis, Similarity Measure, Structural Importance, Centralization

doi:10.4108/eai.7-1-2021.167840

1. Introduction

Social Network is a place where a set of people participates to-gather in the form of societies and communicates with each other or creates relationships [27]. Interaction of students in collages and universities or gathering of people on public places makes off-line social networks. On the other hand, interaction of people on online social networking sites represents online social networks such as Facebook, Twitter, Instagram, and YouTube [22]. Use of social networking sites enables people to make new friends, or stay in touch with old college friends, celebrate functions with your families half around the world, meet people whose likes and dislikes are similar to each other, join groups of related interest, or can disappear and leave that particular group afterwards [28]. Social interaction can be defined as a social graph where people or participants correspond to the nodes and their relationship represents with edge between these nodes.

Analysis of Social network is the process to mine the structure of the social network using graph theory. It specifies network structure in terms of nodes (people) and links to connect them. The analysis of social networks provides help to people that how they can meet other people of similar interest, what communities they should join and find the products of their interest. There are also many uses of analysis of network social such as you can target the peoples for the product [25], you can detect how people makes their neighbor?[13] and how people participates in the communities?[11]

Researcher are facing many problems in the field of social network analysis [24], one of them is link prediction which means that which new interaction between nodes appearing or disappearing near in the future. The dynamic nature of social network makes this challenge more interesting. Moreover we can also use incomplete information for prediction of missing links in a network. Consider a social graph G (as shown in Figure 1) about five persons (i.e., U, V, W, Y and Z).

*Corresponding author. Email: writetosamadalvi@gmail.com

Some of the nodes have no link with other such as (U, V, Y, Z) at time T. After some while, at time T+1, we can think about link prediction that person (U, V, W) can make a new relationship during all this time as all three nodes have a common friend W. This task is known as link prediction.

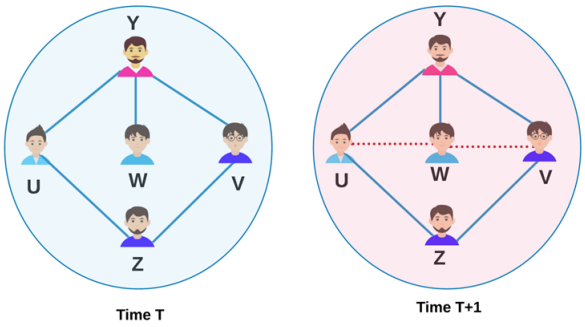

Figure 1. Example of Link Prediction

Predicting a link for social network has many dominant uses. First, it can be used in recommender systems to help people for finding friends [15][19]. Second, it can be used to find or recommend luxury hotels for tourist and travelers [12]. Third, in the field of research, professional findings and co-authors interactions in any research area [9]. Certainly, the link prediction method can be used in biology and biomathematics [6], for example, in gene expression network and health care a, specialists predicting about outmost probability which reversals near future possibilities and managing them in proportion by communication with related people through online and off-line social network.

The link prediction method can be categories in four parts from high to low [29]: First-one is the knowledge of social network for prediction of link. These categories have information of social theory or nodes. Category two consists of path, common neighbor and walks in random, these are part of topology. Category three adds that most famous method for the prediction of link is consisted on paths, nodes, random-walks and common neighbors. Last category may have meaningful information on link prediction as it is about the size, attribute and storage of nodes and edges. All categories are important but third play active role to predict link on the base of social network knowledge.

In the past studies, most of the link prediction techniques have been observed which uses the topological information of nodes within the social graph to find the similar pair of nodes in order to predict the link. where, the estimated similarity is considered as a score, that is assigns to the pair of nodes (X, Y). Where, a high similarity score creates more chances that X will be connected to Y in the future. Moreover, a low similarity score represents high probability that there will be no link between X and Y in the future. Our hypothesis is that nodes having same status of centralization along with high similarity can connect to each other in a future. Here, in this paper, we have proposed structural importance-based link prediction techniques. Where, we have considered importance of nodes within their neighborhood for the similarity score computation. It makes this study different from the rest of studies that are available in the literature.

The reminder of the paper is as follows: Section 2 represents the literature review. State-of-the-art link prediction techniques and structural importance-based link prediction techniques are presented in Section 3. Results are discussed in Section 4. In the end, Section 5 concludes the this study and presents future direction.

2. Literature Review

Similarity could be measured for different nodes with the help of link prediction method for this suppose that there are two nodes X and Y, link between them on the basis of similarity could be measured with the help of probability. Least similar node has less probability that is how chances of link occurrence decrease in future. More the probability between the nodes of (x, y) makes them more similar for link between them. Many of researches have been using this approach; exact prediction of links among nodes could lead to the most active links in network.

Wang et al [2] worked that probability used for local graphical structure and graphs to measure border scale to change the link among the nodes. . Tylenda et al.[1] presented that probability in local structure could be measure current similarity of nodes using approach called Adamic Adar (AA) Root and Page Rank .In Adamic Adar the applied weight on edges in common neighbor nodes and at algorithm are different in the time prediction links. Three measurement scales are mostly used in this rooted Page Rank, Katz and escape prospect. Munasinghe et al. [16] Discussed that Time Score (TS) are used to measure for prediction link among the confined pair of nodes makes strong bound connection with time and destiny for better communication. Sores et al. [7] worked that topological structure find similarity pairs of links. Zhang and Phili. [31] Presented two hope similarity model where measuring the similarity between similar edges takes place. Ibrahim and Chen. [8] Worked that integrated model, node neutrality used to predict temporary information of model in social network. Han et al. [30] (worked that community similarity .they discussed in community similarity degree (CSD).This method are used to measure the similarity in multiple community.Murata et al. [18] worked that weight score and weighted namely score method are used to measure

similarity in common neighbor [3] and AdamicAdar [20] respectively.

Challenge is to detect missing or new links between nodes in a social communication. Wang et al. [4] presented the efficient result of inferential graph problem. Study of homophile used link prediction finding missed edges in social graph collect knowledge of people in social profile. Social graph have offline data predict social contact of people strength of content and predict the missing part of graph. Samad et al. [23] discussed the different social features and check that impact of link prediction in a graph structure by giving weight on the edge they represent the boundary of different connection. Samad et al. [23] worked on different features to enhance and justify the most important feature which plays important role in social profile. Also, discuss similarity for social contacting people prefer to participate other belonging to ideal language and nationality.

Junuthula et al. [14] presented that predicts link for online social network that interacts between different network and combine friendship group having mutual contacts. They proved that friendship networks improved with prediction of link and interaction on network discussed on a particular day. Zhou et al. [5] enplaned the problem attacking similarity-based on link prediction discussed in a different algorithm for deleted node in the network. On the base of algorithm, the network is divided in two classes such as global and local similarity. Global similarity worked on special cases use for NP-Hard metrics. They track deleted and missing link using an algorithm from any graph. Local similarity used in optimal attack and focus on CND target link in a group. Lime et al. [17] worked on hidden link in criminal data and analyze the method in supervised machine learning for big data training and testing. Hidden links explored the application of deep reinforcements learning (DRL) use in the criminal network. The DRL experiment performs better in supervise machine learning. Moreover, Lime et al. [17] measured that working with supervises machine learning in DRL is a predictive accuracy and power of computing.

3. Methodology

For the experiments, we have used four different co-author social network dataset in this paper. These dataset represented by a social graph G (V, E), where V denotes the vertices (i.e., authors) in the graph and E corresponds to edges (i.e., co-author relationship between authors). Our study is to find the best way to predict links based on similarity between nodes. For the link prediction, we give a similarity score S(a,b) to every pair of node (a,b). The possibility of predicted link between pair of node is then estimated by the given similarity score. A higher similarity score S(a,b), close to 1, indicates high chances that link between a and b will be occur in future, while, the lower similarity score S(a,b), close to 0, shows high chances there will be no link between a and b in future.

3.1. Dataset Description

In this research, we have used four different co-author social network datasets (i.e., HepTh, HepPh, CondMat and Astro). These datasets are taken from *e-print arXiv* and shows the research collaboration of authors. The graph represents an edge that denotes the co-author relationship between x and y if they co-authored a paper. Likewise, completely subgraph of k nodes represents the k co-authors on a single paper. More detailed statistics about datasets are shown in Table 1 and notations used in this paper are presented in Table 2.

Table 1. Statistics of used datasets

Dataset	Nodes	Edges	Triangles
AstroPh	18772	198110	1351441
CondMat	23133	93497	173361
HepPh	12008	118521	3358499
HepTh	9877	25998	28339

Here, in the Table 1, nodes corresponds to the authors, edges represents the co-author relation and triangles shows the co-author relation between 3 authors. For example, author i published article with author j and k. On the other hand, author j also published article with author k then there will be a triangle between i, j and k.

3.2. State-of-the-art Techniques

Jaccard Index. Another name of Jaccard Index is Jaccard coefficient [10], which is famous as a normalize form of common neighbors and treats the neighbors as two different groups i.e., intersection of neighbors and union of neighbors. Further it takes in account both groups to compute the similarity of two nodes u and v. In the literature, it has been observed by Liben et al. that common neighbors outperforms Jaccard Index. It is estimated as Equation 1

$$JC(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u) \bigcup \Gamma(v)|} \quad (1)$$

Resource Allocation. In operating system, resource allocation is known as the process of utilizing available resources for various uses. This measure [32] behaves same as one in operating system. Most of the researchers considered it same as Adamic Adar, however, it more penalized to high degree common

Table 2. Symbols and notations used in this paper

Notation	Definition
$\|\Gamma(u)\|$	Represents the number of neighbors of node u
$\|\Gamma(u)\|$	Represents the number of neighbors of node u
$\|\Gamma(u) \cap \Gamma(v)$	Represents the number of common neighbors between node u and v
$deg(u)$	number of direct neighbors of node u
$deg(v)$	number of direct neighbors of node v
$S - Jac$	Structural Importance-based jaccard index
$S - SAM$	Structural Importance-based SAM similarity
$S - SI$	Structural Importance-based sorenson index
$S - SC$	Structural Importance-based salton cosine similarity
$S - HD$	Structural Importance-based hub depressed similarity
$S - PD$	Structural Importance-based parameter-dependent similarity

neighbors. That is the reason behind closer results of both Resource Allocation and Adamic Adar. The foremost characteristic of Resource Allocation is that it looks at direct neighbors as well as neighbors of direct neighbors. It is estimated as in Equation 2

$$RA(u,v) = \sum_{z \in \Gamma(u) \bigcap \Gamma(v)} \frac{1}{|\Gamma(z)|} \tag{2}$$

Sørensen Index. Another name of Sørensen Index is Sørensen coefficient, and statistically used to compute the similarity between two nodes u and v. It was published in 1948 by famous researcher Thorvald Sørensen [26] to test on the ecological community to find similarity portion between data samples. Its foremost feature is to pull up the lower degree nodes to find their interactions. It is computed as in Equation 3.

$$SI(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u) + \Gamma(v)|} \tag{3}$$

SAM Simialrity. In the literature, almost all the similarity measures emphasized on the statement that commonality between nodes two nodes to find their similarity i.e., node u and v one similarity to each other such that if u and 70% similar to v and then v is also 70% similar to u. Samad et al. [23] published a new similarity measure by stated that two nodes u and v have their own similarity in their own perspective, i.e., it is possible that u is 100% similar to v, but at the same time v is not similar as u. SAM similarity is estimated as Equation 4

$$SAM(u,v) = \frac{\frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(u)|} + \frac{|\Gamma(u) \bigcap \Gamma(v)|}{|\Gamma(v)|}}{2} \tag{4}$$

Salton Cosine. Another name of Slaton Cosine [21] is cosine similarity, which is similar as Jaccard coefficient and Sørensen coefficient. In the literature, it has been observed through some studies that Salton Cosine produces the similarity twice the Jaccard Index, however, few results have been found against the observation. Salton Cosine is computed as in Equation 5

$$SC(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{\sqrt{|\Gamma(u) \cdot \Gamma(v)|}} \tag{5}$$

Hub Depressed Index. In the literature, it has been observed that Hub Depressed Index [33] works same as Hum Promoted Index, however, it assigned high score to links connected with hub (i.e., nodes with higher degree called hub). The reason behind the score assigning to links with higher degree nodes is that the denominator is determine by the lower degree only. Similarity can be defined as in Equation 6.

$$HDI(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{Max(|\Gamma(u)|, |\Gamma(v)|)} \tag{6}$$

Parameter-Dependent. This measure [34] improves the accuracy of link prediction for both unpopular and popular. Here, λ have many goodness that, in case $\lambda = 0$, this measure debased to Common Neighbors. Besides, if $\lambda = 1$ and $\lambda = 0.5$, it debased to Salton Cosine and Leicht-Holme-Nerman, respectively. Formula is

shown in Equation 7.

$$PD(u,v) = \frac{|\Gamma(u) \bigcap \Gamma(v)|}{(|\Gamma(u)| \cdot |\Gamma(v)|)^{\lambda}} \quad (7)$$

3.3. Structural Importance-Based Link Prediction

Depending on the place of a node in network graph under consideration, some of the network structures can play a important role in link prediction between similar nodes. In the past studies, it is observed that two kinds of social network structures control the literature: (1) network density; (2) network centralization. The roles of different nodes within a network are often understood through centrality analysis, which aims to quantify the capacity of a node to influence, or be influenced by, other nodes via its connection topology. Here, our hypothesis is that nodes having same status of centralization along with high similarity can connect to each other in a future. In order of the analyze the centralization status, we have used degree centrality which is defined as follows 8.

$$C_d(v) = deg(v) \quad (8)$$

Where, $deg(v)$ returns the number of adjacent nodes of node v. In order to use degree with similarity, we have defined the following formula 9.

$$S(u,v) = Sim(u,v) + (\frac{\Gamma(u) \bigcap \Gamma(v)}{deg(u)} + \frac{\Gamma(u) \bigcap \Gamma(v)}{deg(v)}) \quad (9)$$

Here, $S(u,v)$ represent the similarity score between node u and v. While, Sim represent the state-of-the-art similarity measures (as discussed in Section 3.2). After merging the centralization saltus with similarity score, we have proposed the following Structural Importance based state-of-the-art similarity measures (as shown in Table 3).

3.4. Generating Edge List

We have considered four different dataset for our experiment (i.e., HepTh, HepPh, CondMat and Astro). In order to predict the link, three different set of edges (i.e., Edge5, Edge10 and Edge15) are selected for prediction from each dataset. Where, the first set contains 5% edges, second contains 10% edges, while, third set consist of 15% edges. In addition, to generate the edge lists, the following steps are taken.

- First, five percent edges are randomly taken from CondMat dataset and formed an edge list called CondMat-Edge5.
- Secondly, in order to make second edge list, again ten percent edges are randomly take for prediction from CondMat dataset and called it CondMat-Edge10.
- In order to make the third edge list, 15 percent edges are randomly picked and formed another edge list known as CondMat-Edge15.

In this way, we have collected 30% edges from CondMat dataset. Moreover, we have repeated the above three steps for each remaining dataset and found 12 another edge lists i.e., Astro-Edge5, Astro-Edge10, Astro-Edge15, HepPh-Edge5, HepPh-Edge10, HepPh-Edge15, HepTh-Edge5, HepTh-Edge10 and HepPh-Edge15. Moreover, more statistics about edge lists are shown in Table 1.

4. Results and Experiments

For the experiment, we have used four co-author datasets (i.e., Astro, HepPh, CondMat and HepTh). In the start, from the datasets contained a co-author social network along with different number of edges and nodes (as shown in Table 2), we have generated social graphs. In the first step, we have randomly generate three edge lists for each social graph. In order to predict the links, these 12 edge lists are further used in our experiment. in the second step, these picked edges are removed from each social graph and applied similarity measures on the social graph. In order to predict the removed edges, similarity score is assigned to every removed edge. After that, four different threshold (i.e., 0.1, 0.3, 0.5 and 0.7) are applied on every social graph and created 48 predicted graphs, 12 for each dataset using 3 edge lists. In the end, results of predicted graphs are evaluated using accuracy measures.

4.1. Evaluation on Astro Dataset

For the link prediction, 30% edges were used from the Astro dataset and divided into three edge lists (i.e., Astro-Edge5, Astro-Edge10 and Astro-Edge15). Where, edge list Astro-Edge5 contributed with 9902 edges, Astro-Edge10 contained 19804 edges and Astro-Edge15 represented 28292 edges. The results from Astro dataset are shown in Figures 2, 3 and 4. Where, Figure 2, representing the results of edge list Astro-Edge5. In Figure 2, X-axis represents the similarity techniques, while, Y-axis shows the prediction accuracy. In addition, the bars represents the threshold. The same pattern is designed in all the figures. The results showed that with the use of structural importance along with state-of-the-art similarity measures we succeed in getting highest accuracy. Resultant threshold showed that on all thresholds, structural importance of nodes played an important role in achieving high accuracy. At threshold 0.1, S-SAM obtained high accuracy as compared to SAM, Where, S-SAM succeed in getting accuracy 1 and SAM achieved 0.93. Likewise, other state-of-the-art similarity approaches also performed better along with structural importance as compared

Table 3. Structural Importance based State-of-the-art Similarity Measures

No.	State-of-the-art	Structural Importance-Based	Similarity Measure
1	$Jac(u,v) = \frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{\lvert\Gamma(u)\bigcup\Gamma(v)\rvert}$	$S-Jac(u,v) = \frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{\lvert\Gamma(u)\bigcup\Gamma(v)\rvert} + (\frac{\Gamma(u)\bigcap\Gamma(v)}{deg(u)} + \frac{\Gamma(u)\bigcap\Gamma(v)}{deg(v)})$	Jaccard Similarity
2	$RA(u,v) = \sum_{z\in\Gamma(u)\bigcap\Gamma(v)} \frac{1}{\lvert\Gamma(z)\rvert}$	$S-RA(u,v) = \sum_{z\in\Gamma(u)\bigcap\Gamma(v)} \frac{1}{\lvert\Gamma(z)\rvert} + (\frac{\Gamma(u)\bigcap\Gamma(v)}{deg(u)} + \frac{\Gamma(u)\bigcap\Gamma(v)}{deg(v)})$	Resource Allocation
3	$SI(u,v) = \frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{\lvert\Gamma(u)+\Gamma(v)\rvert}$	$S-SI(u,v) = \frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{\lvert\Gamma(u)+\Gamma(v)\rvert} + (\frac{\Gamma(u)\bigcap\Gamma(v)}{deg(u)} + \frac{\Gamma(u)\bigcap\Gamma(v)}{deg(v)})$	Sørensen Index
4	$SAM(u,v) = \frac{\frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{\lvert\Gamma(u)\rvert} + \frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{\lvert\Gamma(v)\rvert}}{2}$	$S-SAM(u,v) = \frac{\frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{\lvert\Gamma(u)\rvert} + \frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{\lvert\Gamma(v)\rvert}}{2} + (\frac{\Gamma(u)\bigcap\Gamma(v)}{deg(u)} + \frac{\Gamma(u)\bigcap\Gamma(v)}{deg(v)})$	SAM Similarity
5	$SC(u,v) = \frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{\sqrt{\lvert\Gamma(u)\cdot\Gamma(v)\rvert}}$	$S-SC(u,v) = \frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{\sqrt{\lvert\Gamma(u)\cdot\Gamma(v)\rvert}} + (\frac{\Gamma(u)\bigcap\Gamma(v)}{deg(u)} + \frac{\Gamma(u)\bigcap\Gamma(v)}{deg(v)})$	Salton Cosine
6	$HDI(u,v) = \frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{Max(\lvert\Gamma(u)\rvert,\lvert\Gamma(v)\rvert)}$	$S-HD(u,v) = \frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{Max(\lvert\Gamma(u)\rvert,\lvert\Gamma(v)\rvert)} + (\frac{\Gamma(u)\bigcap\Gamma(v)}{deg(u)} + \frac{\Gamma(u)\bigcap\Gamma(v)}{deg(v)})$	Hub Depressed
7	$PD(u,v) = \frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{(\lvert\Gamma(u)\rvert\cdot\lvert\Gamma(v)\rvert)^{\lambda}}$	$S-PD(u,v) = \frac{\lvert\Gamma(u)\bigcap\Gamma(v)\rvert}{(\lvert\Gamma(u)\rvert\cdot\lvert\Gamma(v)\rvert)^{\lambda}} + (\frac{\Gamma(u)\bigcap\Gamma(v)}{deg(u)} + \frac{\Gamma(u)\bigcap\Gamma(v)}{deg(v)})$	Parameter-Dependent

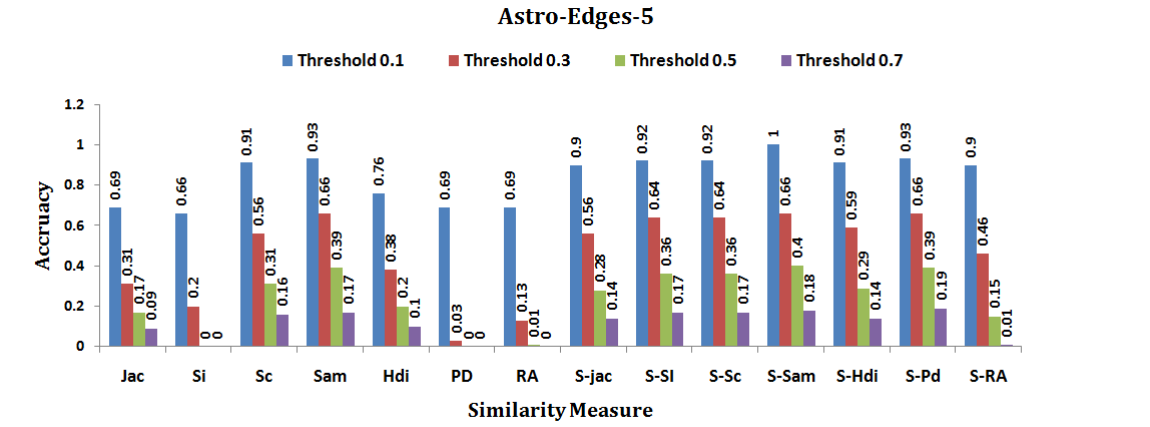

Figure 2. Prediction Results using Astro Dataset with 5% Edges

to without structural importance. At thresholds 0.5 and 0.7, SI and PD could not produce reasonable results. In detail, at threshold 0.3, S-SAM and S-PD obtained 0.66, S-JAC succeed in getting 0.56, S-SI and S-SC produced 0.64, S-HDI obtained 0.59 and S-RA achieved 0.46 accuracy. On the other hand, at threshold 0.3, State-of-the-art techniques JAC obtained 0.31, SI achieved 0.2, SC produced 0.56, SAM obtained 0.66, HDI achieved 0.38, RA obtained 0.17 and PD succeed in getting 0.03 accuracy. Overall on all thresholds, S-SAM obtained maximum accuracy by 1 and both SI and PD achieved minimum accuracy by 0.

Similarly, Figures 3 and 4 addressing the prediction results of edge list Astro-Edge10 and Astro-Edge15.

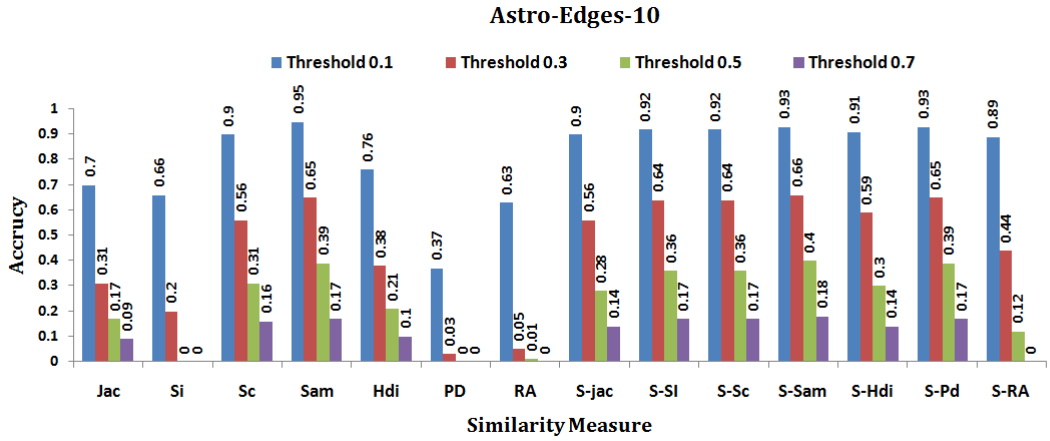

Figure 3. Prediction Results using Astro Dataset with 10% Edges

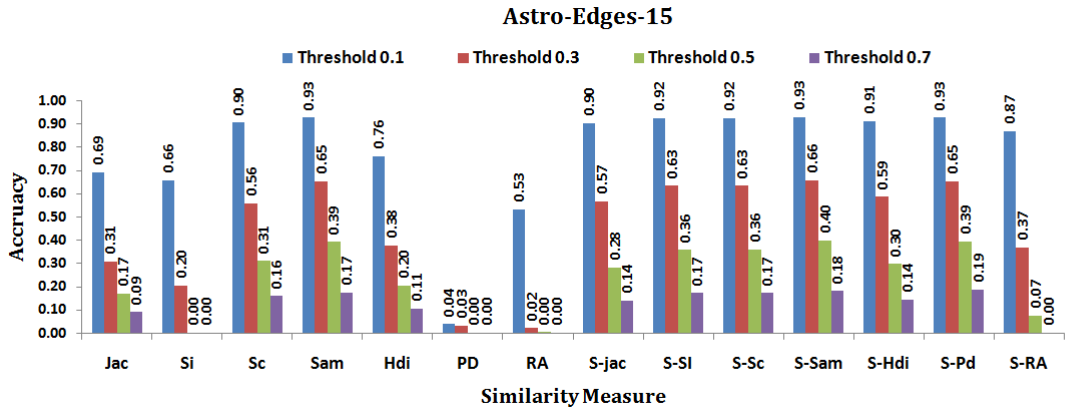

Figure 4. Prediction Results using Astro Dataset with 15% Edges

On all the thresholds, in both Figures, State-of-the-art link prediction techniques with structural importance performed better as compared to link prediction techniques without structural importance. In Figure 3, At threshold 0.3, S-SAM produced maximum accuracy by 0.66 and PD achieved minimum accuracy by 0.03. Likewise, at threshold 0.5, S-SAM obtained maximum accuracy by 0.40 and both SI and PD achieved minimum accuracy by 0.In Figure 4, at threshold 0.7, S-SAM and S-PD obtained maximum accuracy by 0.19, while, SI, PD and RA achieved minimum accuracy by 0. In Figure 4, State-of-the-art techniques SI, PD and RA could not produced better results. Overall, S-SAM and S-PD performed best on all thresholds.

4.2. Evaluation on CondMat Dataset

For the link prediction, 30% edges were used from the CondMat dataset and divided into three edge lists (i.e., CondMat-Edge5, CondMat-Edge10 and CondMat-Edge15). Where, edge list CondMat-Edge5 contributed with 9902 edges, CondMat-Edge10 contained 19804 edges and CondMat-Edge15 represented 28292 edges. The results from CondMat dataset are shown in Figures 5, 6 and 7. Where, Figure 5, representing the results of edge list CondMat-Edge5. In Figure 5, X-axis represents the similarity techniques, while, Y-axis shows the prediction accuracy. In addition, the bars represents the threshold. The same pattern is followed in all the figures. The results showed that with the use of structural importance along with state-of-the-art similarity measures we succeed in getting highest accuracy. Resultant threshold showed that on all thresholds, structural importance of nodes played an important role in achieving high accuracy. At threshold 0.1, S-HDI and S-PD obtained high accuracy as compared to HDI and PD, Where, S-HDI and S-PD succeed in getting accuracy 0.90 , while, HDI and PD achieved respectively 0.74 and 0.54. Likewise, other state-of-the-art similarity approaches also performed better along with structural importance as compared to the similarity approaches without

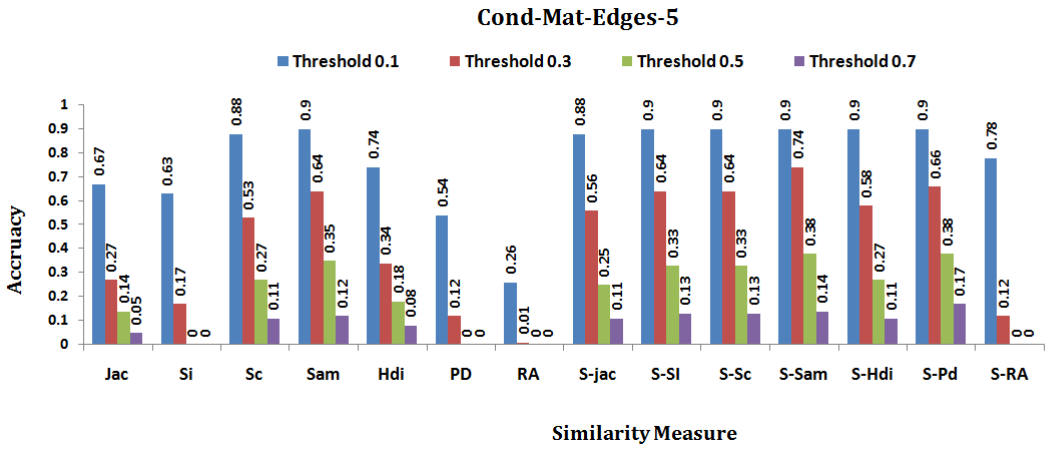

Figure 5. Prediction Results using CondMat Dataset with 5% Edges

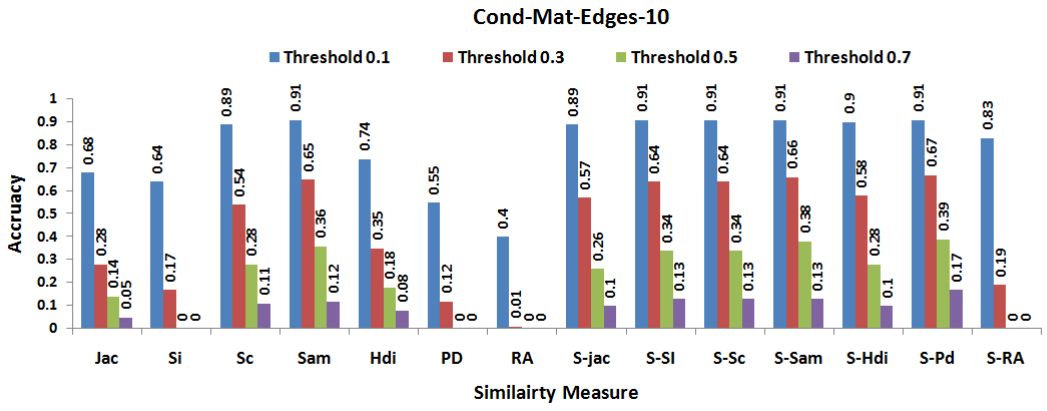

Figure 6. Prediction Results using CondMat Dataset with 10% Edges

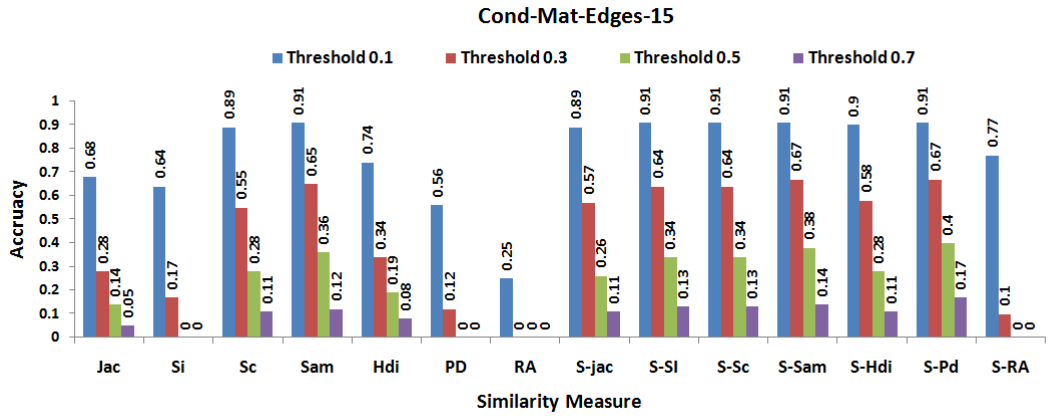

Figure 7. Prediction Results using CondMat Dataset with 15% Edges

structural importance. At thresholds 0.5 and 0.7, SI, PD and RA could not produce reasonable results. In detail, at threshold 0.3, S-SAM obtained 0.74, S-RA succeed in getting 0.78, S-SI and S-SC produced 0.64, S-HDI obtained 0.58, S-JAC obtained 0.56 and S-PD achieved 0.66 accuracy. On the other hand, at threshold 0.3, State-of-the-art techniques JAC obtained 0.27, SI achieved 0.17, SC produced 0.53, SAM obtained 0.64,

HDI achieved 0.34, RA obtained 0.01 and PD succeed in getting 0.12 accuracy. Overall on all thresholds, S-SAM, S-SI, S-SC, S-HDI and S-PD obtained maximum accuracy by 1, while, SI, PD and RA achieved minimum accuracy by 0.

Similarly, Figures 6 and 7 addressing the prediction results of edge list CondMat-Edge10 and CondMat-Edge15. On all the thresholds, in both Figures, State-of-the-art link prediction techniques with structural importance performed better as compared to link prediction techniques without structural importance. In Figure 6, At threshold 0.3, S-PD produced maximum accuracy by 0.67 and PD achieved minimum accuracy by 0.12. Likewise, at threshold 0.5, S-PD obtained maximum accuracy by 0.39, while, SI, PD and RA achieved minimum accuracy by 0.In Figure 7, at threshold 0.7, S-PD obtained maximum accuracy by 0.17, while, SI, PD, RA and S-RA achieved minimum accuracy by 0. In Figure 7, State-of-the-art techniques SI, PD, RA and S-RA could not produced better results. Overall, S-SAM and S-PD performed best on all thresholds.

4.3. Evaluation on HepPh Dataset

In order to predict the links, 30% edges were used from the HepPh dataset and divided into three edge lists (i.e., HepPh-Edge5, HepPh-Edge10 and HepPh-Edge15). Where, edge list HepPh-Edge5 contributed with 9902 edges, HepPh-Edge10 contained 19804 edges and HepPh-Edge15 represented 28292 edges. The results from HepPh dataset are shown in Figures 8, 9 and 10. Where, Figure 8, representing the results of edge list HepPh-Edge5. In Figure 8, X-axis represents the similarity techniques, while, Y-axis shows the prediction accuracy. In addition, the bars represents the threshold. The same pattern is designed in all the figures. The results showed that with the use of structural importance along with state-of-the-art similarity measures we succeed in getting highest accuracy. Resultant threshold showed that on all thresholds, structural importance of nodes played an important role in achieving high accuracy. At threshold 0.1, S-SAM and S-PD obtained high accuracy as compared to SAM and PD, Where, S-SAM and S-PD succeed in getting accuracy 0.95, while, SAM and PD achieved respectively 0.95 and 0.45. Likewise, other state-of-the-art similarity approaches also performed better along with structural importance as compared to link prediction approaches without structural importance. At thresholds 0.5 and 0.7, SI, PD and RA could not produce reasonable results. In detail, at threshold 0.3, S-SAM and S-PD obtained 0.83, S-JAC succeed in getting 0.77, S-SI and S-SC produced 0.81, S-HDI obtained 0.78 and S-RA achieved 0.66 accuracy. On the other hand, at threshold 0.3, State-of-the-art techniques JAC obtained 0.61, SI achieved 0.53, SC produced 0.77, SAM obtained 0.83, HDI achieved 0.64, RA obtained 0.03 and PD succeed in getting 0.03 accuracy. Overall on all thresholds, S-SAM, S-PD and SAM obtained maximum accuracy by 0.95 and both SI, PD and Ra achieved minimum accuracy by 0.

Similarly, Figures 9 and 10 addressing the prediction results of edge list HepPh-Edge10 and HepPh-Edge15. On all the thresholds, in both Figures, State-of-the-art link prediction techniques with structural importance performed better as compared to link prediction techniques without structural importance. In Figure 9, At threshold 0.3, S-PD and SAM produced maximum accuracy by 0.83, while, PD and RA achieved minimum accuracy by 0.03. Likewise, at threshold 0.5, S-SAM, S-PD and SAM obtained maximum accuracy by 0.68, while, both SI, PD and RA achieved minimum accuracy by 0. In Figure 10, at threshold 0.7, S-SI, S-SC, S-SAM, S-PD and SAM obtained maximum accuracy by 0.49, while, SI, PD and RA achieved minimum accuracy by 0. In Figure 10, State-of-the-art techniques SI, PD and RA could not produced better results. Overall, S-SAM, S-PD and SAM performed best on all thresholds.

4.4. Evaluation on HepTh Dataset

For the link prediction, we have picked 30% edges from the HepTh dataset and divided into three edge lists (i.e., HepTh-Edge5, HepTh-Edge10 and HepTh-Edge15). Where, edge list HepTh-Edge5 contributed with 9902 edges, HepTh-Edge10 contained 19804 edges and HepTh-Edge15 represented 28292 edges. The results from HepTh dataset are shown in Figures 11, 12 and 13. Where, Figure 11, representing the results of edge list HepTh-Edge5. In Figure 11, X-axis represents the similarity techniques, while, Y-axis shows the prediction accuracy. In addition, the bars represents the threshold. The same pattern is designed in all the remaining figures. The results showed that with the use of structural importance along with state-of-the-art similarity measures we succeed in getting highest accuracy. Resultant threshold showed that on all thresholds, structural importance of nodes played an important role in achieving high accuracy. At threshold 0.1, S-SI, S-SC and S-PD obtained high accuracy as compared to SI, SC and PD, Where, S-SI, S-SC and S-PD succeed in getting accuracy 0.81, while, SI, SC and PD achieved respectively 0.52, 0.79 and 0.49. Likewise, other state-of-the-art similarity approaches also performed better along with structural importance as compared to without structural importance. At thresholds 0.5 and 0.7, SI, PD and RA could not produce reasonable results. In detail, at threshold 0.3, S-SAM obtained 0.49, S-PD achieved 0.51, S-JAC succeed in getting 0.39, S-SI produced 0.45, S-SC obtained 0.46,

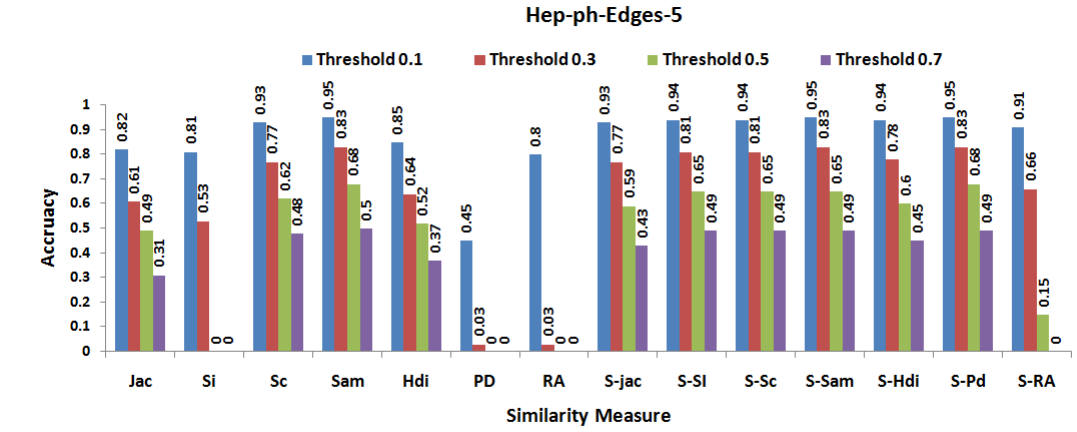

Figure 8. Prediction Results using HepPh Dataset with 5% Edges

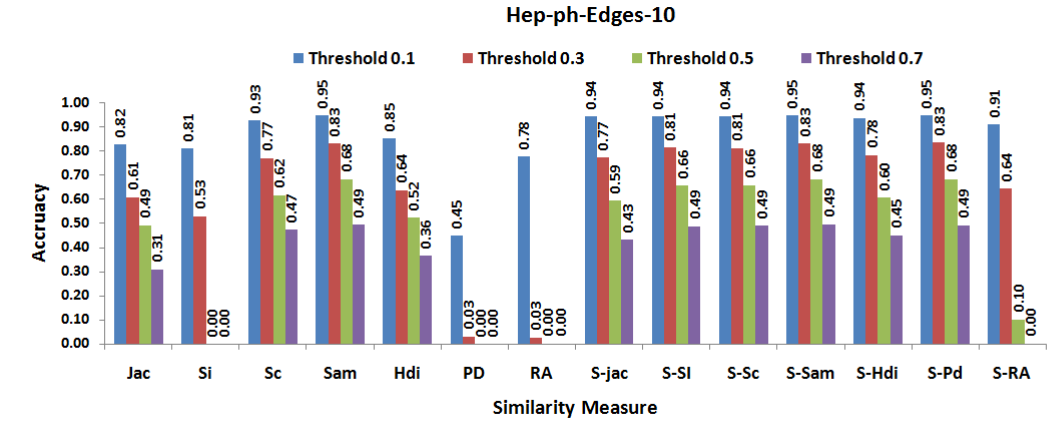

Figure 9. Prediction Results using HepPh Dataset with 10% Edges

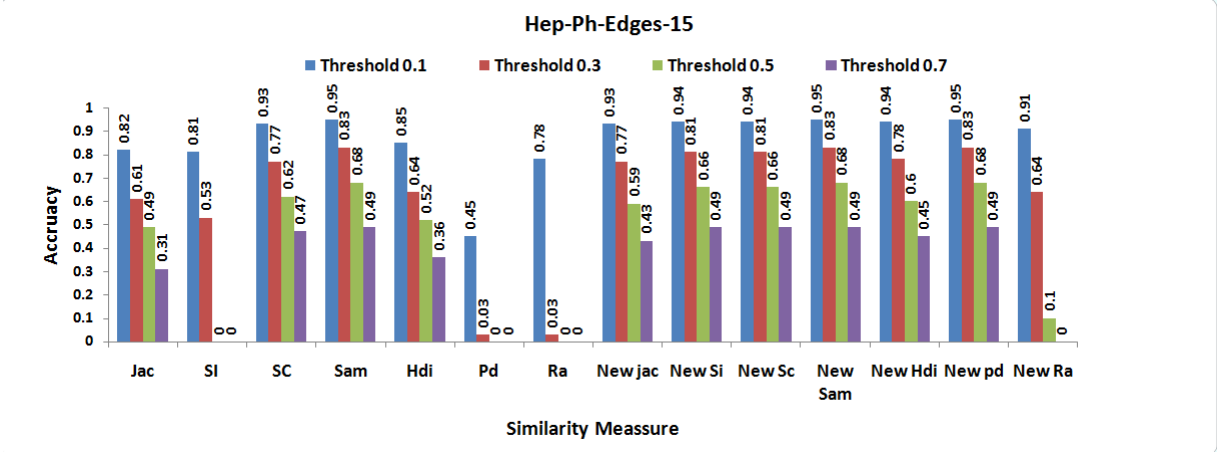

Figure 10. Prediction Results using HepPh Dataset with 15% Edges

S-HDI obtained 0.41 and S-RA achieved 0.28 accuracy. On the other hand, at threshold 0.3, State-of-the-art techniques JAC obtained 0.20, SI achieved 0.11, SC produced 0.39, SAM obtained 0.47, HDI achieved 0.25, RA obtained 0.10 and PD succeed in getting 0.10 accuracy. Overall on all thresholds, S-SI, S-SC, S-SAM

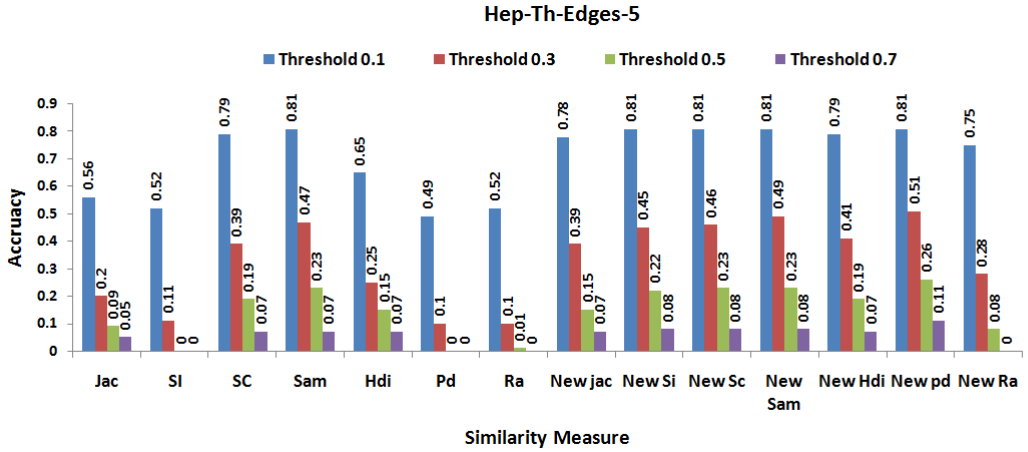

Figure 11. Prediction Results using HepTh Dataset with 5% Edges

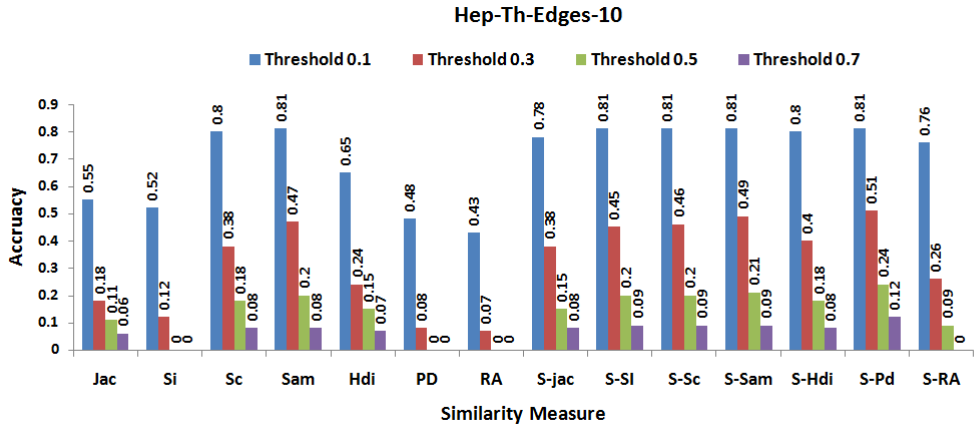

Figure 12. Prediction Results using HepTh Dataset with 10% Edges

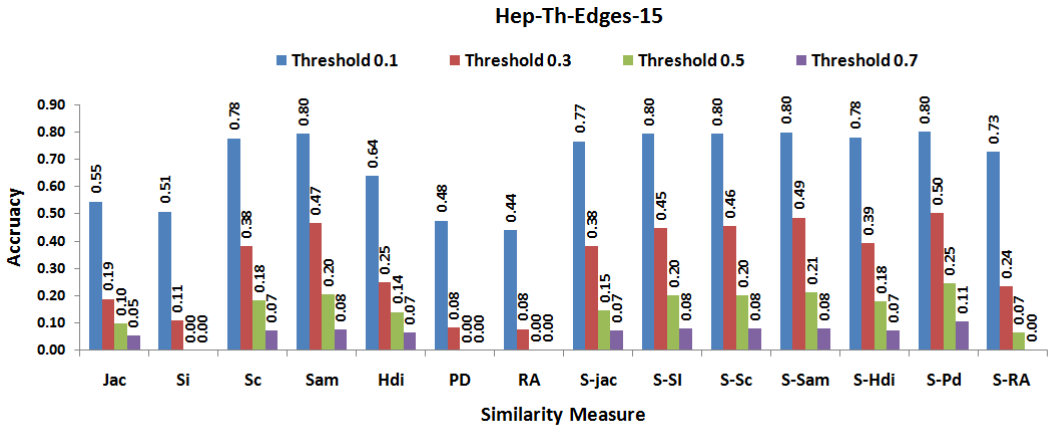

Figure 13. Prediction Results using HepTh Dataset with 15% Edges

and S-PD obtained maximum accuracy by 0.81 and both SI, PD and RA achieved minimum accuracy by 0.

Similarly, Figures 12 and 13 addressing the prediction results of edge list HepTh-Edge10 and HepTh-Edge15. On all the thresholds, in both Figures, State-of-the-art link prediction techniques with structural

importance performed better as compared to link prediction techniques without structural importance. In Figure 12, At threshold 0.3, S-PD produced maximum accuracy by 0.51 and RA achieved minimum accuracy by 0.07. Likewise, at threshold 0.5, S-RA obtained maximum accuracy by 0.26 and both SI, PD and RA achieved minimum accuracy by 0. In Figure 13, at threshold 0.7, S-PD obtained maximum accuracy by 0.11, while, SI, PD, RA and S-RA achieved minimum accuracy by 0. In Figure 13, State-of-the-art techniques SI, PD and RA could not produced better results. Overall, S-SAM and S-PD outperformed on all thresholds.

5. Conclusion

In this paper, we have proposed a structural importance based state-of-the-art link prediction approachs. We have experimented on four dataset and compared structural importance based approaches with other state-of-the-art link prediction approaches. Our purposed similarity approach punishes high degree nodes heavily in order to predict the links. The results show that structural importance is useful to find the similar pair of nodes for link prediction. At threshold 0.1, Maximum accuracy was 95% by S-PD and SAM. While, at threshold 0.7, again S-SD and SAM produced 68% accuracy. In the future, we will try to find out the different variants of structural importance.

References

[1] Aggarwal, C.C. (2011) An introduction to social network data analytics. In *Social network data analytics* (Springer), 1–15.

[2] Al Hasan, M. and Zaki, M.J. (2011) A survey of link prediction in social networks. In *Social network data analytics* (Springer), 243–275.

[3] Albert, R. and Barabási, A.L. (2002) Statistical mechanics of complex networks. *Reviews of modern physics* **74**(1): 47.

[4] Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G. *et al.* (2012) Smart cities of the future. *The European Physical Journal Special Topics* **214**(1): 481–518.

[5] Besta, M., Kanakagiri, R., Mustafa, H., Karasikov, M., Rätsch, G., Hoefler, T. and Solomonik, E. (2019) Communication-efficient jaccard similarity for high-performance distributed genome comparisons. *arXiv preprint arXiv:1911.04200* .

[6] Breit, A., Ott, S., Agibetov, A. and Samwald, M. (2020) Openbiolink: A benchmarking framework for large-scale biomedical link prediction. *Bioinformatics* .

[7] Chuan, P.M., Ali, M., Khang, T.D., Dey, N. *et al.* (2018) Link prediction in co-authorship networks based on hybrid content similarity metric. *Applied Intelligence* **48**(8): 2470–2486.

[8] Chuan, P.M., Ali, M., Khang, T.D., Dey, N. *et al.* (2018) Link prediction in co-authorship networks based on hybrid content similarity metric. *Applied Intelligence* **48**(8): 2470–2486.

[9] Ho, T.K.T., Bui, Q.V. and Bui, M. (2019) Co-author relationship prediction in bibliographic network: A new approach using geographic factor and latent topic information. In *Proceedings of the Tenth International Symposium on Information and Communication Technology*: 69–77.

[10] Jaccard, P. (1901) Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat* **37**: 547–579.

[11] Jiang, H., Liu, Z., Liu, C., Su, Y. and Zhang, X. (2020) Community detection in complex networks with an ambiguous structure using central node based link prediction. *Knowledge-Based Systems* : 105626.

[12] Kaya, B. (2020) A hotel recommendation system based on customer location: a link prediction approach. *Multimedia Tools and Applications* **79**(3): 1745–1758.

[13] Koptelov, M., Zimmermann, A., Crémilleux, B. and Soualmia, L. (2020) Link prediction via community detection in bipartite multi-layer graphs. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*: 430–439.

[14] Kumar, S., Zhang, X. and Leskovec, J. (2019) Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*: 1269–1278.

[15] Li, S., Song, X., Lu, H., Zeng, L., Shi, M. and Liu, F. (2020) Friend recommendation for cross marketing in online brand community based on intelligent attention allocation link prediction algorithm. *Expert Systems with Applications* **139**: 112839.

[16] Liao, H., Mariani, M.S., Medo, M., Zhang, Y.C. and Zhou, M.Y. (2017) Ranking in evolving complex networks. *Physics Reports* **689**: 1–54.

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

[18] Lü, L. and Zhou, T. (2011) Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications* **390**(6): 1150–1170.

[19] Luo, J., Sinha, A.P. and Zhao, H. (2020) Location-sensitive friend recommendations in online social networks. In *PACIS*: 155.

[20] Newman, M.E. and Girvan, M. (2004) Finding and evaluating community structure in networks. *Physical review E* **69**(2): 026113.

[21] Salton, G. and McGill, M.J. (1983) *Introduction to modern information retrieval* (mcgraw-hill).

[22] Samad, A., Islam, M.A., Iqbal, M.A., Aleem, M. and Arshed, J.U. (2017) Evaluation of features for social contact prediction. In *2017 13th International Conference on Emerging Technologies (ICET)* (IEEE): 1–6.

[23] Samad, A., Qadir, M. and Nawaz, I. (2019) Sam: a similarity measure for link prediction in social network. In *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)* (IEEE): 1–9.

[24] Samad, A., Qadir, M., Nawaz, I., Islam, M.A. and Aleem, M. (2020) A comprehensive survey of link prediction techniques for social network. *EAI Endorsed Trans. Indust. Netw. & Intellig. Syst.* **7**(23): e3.

[25] Shahmohammadi, A., Khadangi, E. and Bagheri, A. (2016) Presenting new collaborative link prediction methods for activity recommendation in facebook. *Neurocomputing* **210**: 217–226.

[26] Sørensen, T., Sørensen, T., Sørensen, T., SORENSEN, T., Sorensen, T., Sorensen, T. and Biering-Sørensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons .

[27] Vega-Oliveros, D.A., Zhao, L. and Berton, L. (2019) Evaluating link prediction by diffusion processes in dynamic networks. *Scientific reports* **9**(1): 1–14.

[28] Vodă, A.I. and Florea, N. (2019) Impact of personality traits and entrepreneurship education on entrepreneurial intentions of business and engineering students. *Sustainability* **11**(4): 1192.

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

[30] Zhai, Y., Xu, Z. and Liao, H. (2016) Probabilistic linguistic vector-term set and its application in group decision making with multi-granular linguistic information. *Applied Soft Computing* **49**: 801–816.

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

[32] Zhou, T., Lü, L. and Zhang, Y.C. (2009) Predicting missing links via local information. *The European Physical Journal B* **71**(4): 623–630.

[33] Zhou, T., Lü, L. and Zhang, Y.C. (2009) Predicting missing links via local information. *The European Physical Journal B* **71**(4): 623–630.

[34] Zhu, Y.X., Lü, L., Zhang, Q.M. and Zhou, T. (2012) Uncovering missing links with cold ends. *Physica A: Statistical Mechanics and its Applications* **391**(22): 5769–5778.