Chapter 5

Link Prediction in Complex Networks Based on Significance of Higher-Order Path Index (SHOPI)

While network analysis is almost 70 years old, less studies focus on the structure of paths in complex networks. When humans or other entities traverse a complex network, they usually do not take the shortest path, but they also do not move randomly. The structure of these paths is an important research area where few works are available. This chapter [1] studies the path feature of social networks.

5.1 Introduction

Earlier, we have studied about several structure or topology based link predictors in the literature review section. Structural similarity-based methods extract information about the underlying structures or topology range from local to global, including quasi-local

[1] Published in Physica A: Statistical Mechanics and its Applications

(more than local and less than global information). Some of which are the neighborhood-based methods (i.e., Common Neighbors (CN) [56], Jaccard [62], Adamic/Adar (AA) [58], Resource Allocation (RA) [59], Preferential attachment (PA) [57], etc.) that use local information, path-based methods (i.e., Katz index [55], Inverse path distance [17] Average commute time (ACT) [81], PageRank [40], Leicht-Holme-Newman Index [70], Random walk with restart (RWR) [72], etc.), which explore global information of the underlying network. Quasi-local methods employ as much information as global methods and computationally efficient as the local methods. So these approaches are trade-offs between local and global methods. Example of such methods include local path index (LP) [53], local random walk (LRW) [53], superposed random walk (SRW) [53], etc.

From the viewpoint of paths, the existing works can be categorized in the following taxonomy shown in Figure 5.1. Initial categorization includes deterministic and random walk approaches where random walks are used to find the next vertex in the path. The next level hierarchy (categorization) states about the contribution of paths (homogeneous and heterogeneous) to the similarity score computation. Homogeneous methods include the summation of the total number of paths with equal contributions, e.g., CN, LP, Katz, etc. Heterogeneous methods, on the other side, incorporates different contributions of different path based on some priority or another scheme, e.g., AA, RA, Significant Path (SP) [281], Effective Path (EP) [282], etc. Earlier work [282] models the influence between two nodes as the connectivity of paths between them where the connectivity of a path is defined as the product of transfer probability of each link involved in the path. Further, [281] proposed an index based on the heterogeneity of paths where the larger score is assigned to the node pair having lower degree intermediate nodes in the smaller path (they called such paths as significant).

The proposed work viz., SHOPI falls into the heterogeneous category where different common neighbors of the node pair are penalized based on their connections to other nodes so that information leaks through them can be minimized. Moreover, paths of longer lengths are also penalized but lightly compared to Katz and LP.

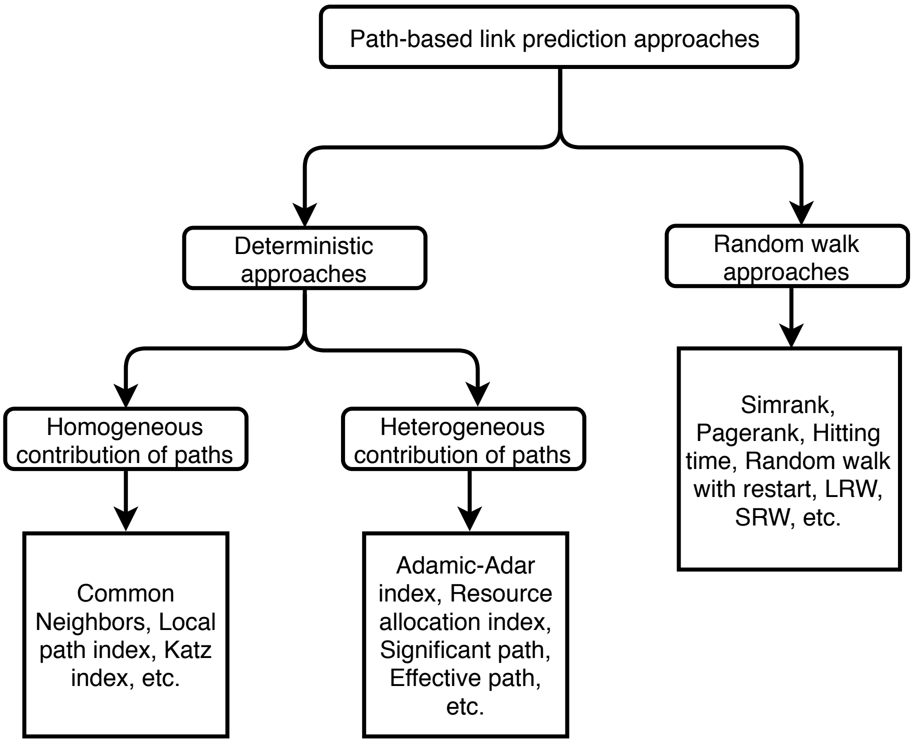

Path-based link prediction approaches
Deterministic approaches
Random walk approaches
Homogeneous contribution of paths
Heterogeneous contribution of paths
Simrank, Pagerank, Hitting time, Random walk with restart, LRW, SRW, etc.
Common Neighbors, Local path index, Katz index, etc.
Adamic-Adar index, Resource allocation index, Significant path, Effective path, etc.

FIGURE 5.1: Path-based approaches to link prediction

The major contribution of this work is as follows

- Motivated by the resource allocation process in networks, this work employs a higher-order path index as a discriminating feature and find missing links in complex networks.
- An iterative algorithm viz., SHOPI (Link prediction based on $\bar{S}$ignificance of $\bar{H}$igher $\bar{O}$rder $\bar{P}$ath $\bar{I}$ndex) and computational complexity are included in the work.
- This work is experimentally evaluated on twelve networks from different areas to show its performance.
- Effects of higher-order paths and different values of the parameter ψ have also experimentally evaluated.

The common neighbor approach can be viewed with another angle called path-based method, i.e., "friends of a friend is also a friend" can be realized using a path[2] of length 2. Several studies are available in the literature that exploits the path as a discriminating feature for link prediction. The path can be of varying length, and several paths may exist between two nodes. These path features can be utilized to compute the similarity score between two nodes or as features in a machine learning framework. For example, negated shortest path (NSP) [17], Katz [55], LP index [53], SR [84], PR [40], etc. Katz index is based on the ensemble of all paths and directly sums these paths to compute the similarity between two nodes. This index incorporates matrix inversion for implementation, so it is quite complex and realized for small networks. LP considers path up to length 3 when degeneracy states need to be resolved in the calculation. As this index does not consider inverse matrix computation, it is somehow simpler to compute. Negated shortest path or inverse path distance incorporates the shortest paths between two nodes to compute similarity and takes $O(|E|logn)$ time using Dijkstra's algorithm. These methods, based on ensemble of paths, are deterministic. Random walk based methods that collect paths information viz., SR, PR, hitting time, etc., are also available [17] in the literature. Recent works [54, 88, 89] argue that features extracted from paths of length 3 are more relevant than 2 length paths. Kovàcs et al. [54] show relevancy of 3 length path features in protein-protein interaction (PPI) networks extensively. They proposed a degree normalized similarity method based on 3 path length, namely the path of length 3 (L3), and showed a significant accuracy improvement in PPI. Pech et al. [88] proposed a theory showing the number of 3-hop path to be a simple and degenerated index induced from a more complex linear optimization. Meanwhile, Muscoloni et al. [89] introduced Cannistraci-Hebb (CH) network automata and local community paradigm (LCP) [68] in a framework and proposed a novel method called CH2-L3. CH2-L3 is based on paths of length 3 that maximize the internal community links and minimizing external links. They showed a significant improvement of prediction

[2] A path is a sequence of links connecting a sequence of nodes in the network. The path length is the number of links in the path.

accuracy on resource allocation network automata both in L2 (path length 2) and L3 (path length 3) in several networks.

5.2 Proposed work

Evidence suggests that most real networks especially, social networks, show three consistent topological properties, namely, small-world phenomenon [28, 29], clustering [11, 12] and scale-free properties [9]. Their corresponding features are the path, clustering coefficient, and degree distribution. In this work, we exploit the path features of different lengths to compute the similarity score between two nodes of the underlying network. This work is based on the resource allocation process [283] in networks, i.e., a sender (first node) sends a resource or information to a destination (second node) either through a direct connection or through common neighbors. In our link prediction framework, the information flows through the common neighbors. The amount of resources received by the receiver represents the similarity between two nodes (sender and receiver). We employ this concept and compute two path length score.

Our work is based on the idea of the resource allocation process wherein the amount of information received by the destination node derives the similarity between them. We try to maximize the information received at the destination node by restricting the information leaks through their common neighbors and hence maximizing the similarity between these two nodes. Figure 5.2 shows the spreading and receiving of resources (information) through common neighbors. The chance of leaking information through the common neighbors of (p,q) is higher than that of (x,y) because more connections are emerging from the common neighbors of (p,q) compared to that of (x,y). In other words, the amount of information (I_2+I_4) received by the node y is higher than the information (I_6+I_8) received by the node q. This results in the node pair (x,y) to have more similarity score than the node pair (p,q).

Significance of the path index of length 2 The above concept can be encoded by the social theory called degree of the common neighbors. Mathematically, the similarity score having 2 path length between any node pair (x,y) can be expressed as

$$S^{'}(x,y) = \sum_{z \in \Gamma x \cap \Gamma y} \frac{1}{k_z}, \tag{5.1}$$

where k_z is degree of the node z.

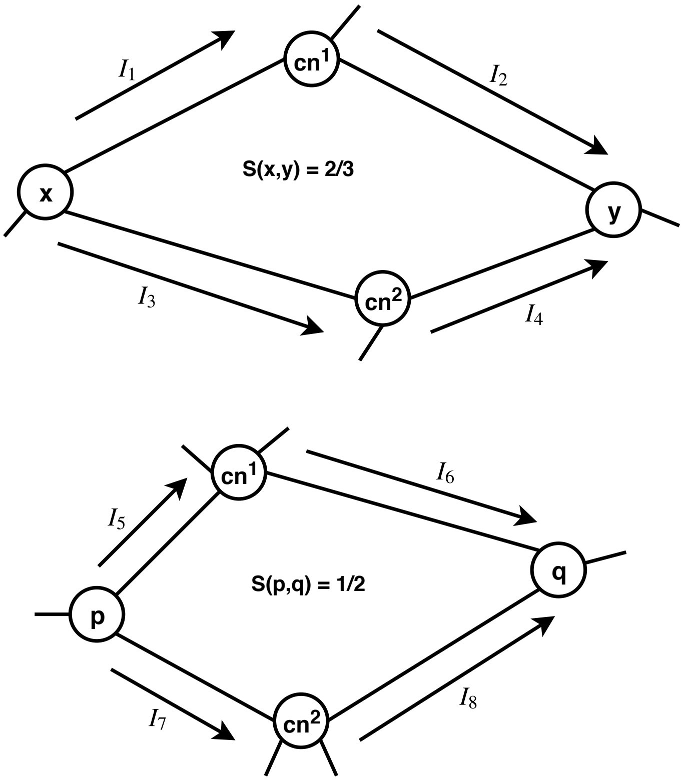


FIGURE 5.2: Path length-2 score calculation: the score between x and y, $S(x,y) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$, and the score between p and q, $S(p,q) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. Clearly, $S(x,y) > S(p,q)$ because the common neighbors of the node pair (x,y) leaks less information compared to that of the node pair (p,q).

Significance of path index of higher order [3] Every path of length $l > 2$ can be decomposed into path of length $(l-1)$ and an edge connected to it. For example, all possible paths of length 3 (i.e., $x - i_1 - cn_1 - y$, $x - cn_1 - i_1 - y$, and $x - i_1 - i_2 - y$)

[3] In this paper, higher and longer are used interchangeably

between the two nodes x and y is shown in the Figure 5.3, where i_1 and i_2 are the two intermediate nodes other than common neighbors in the paths and cn_1 is a common neighbor between these two nodes. Now, the likelihood score of two nodes being connected is the product of the likelihood score of path $(l-1)$ and the significance score of the constituent edge. Mathematically, longer path score can be expressed as

$$S(x,y) = \sum_{l=3}^{l_{max}} f_1 \times f_2 \times \psi^{(l-2)}, \tag{5.2}$$

where f_1 is the significance score of the constituent edge and f_2 is the score of previous iteration. Here, l_{max} is a constant equal to 6 (due to small-world behavior of most social networks) and ψ is penalization parameter that penalizes longer paths. The significant score is computed from the 2 path length score only for the first iteration (i.e. for 3 path length score computation). In other words, The equation 5.2 has been used to calculate the score of the paths of length greater than 2 with the help of 2 length path score. For example, 3 length path can be decomposed into two parts: first an edge and second a path of length 2. Though, we can compute score of 2 length path using equation 5.1. Next, we have to compute the edge score. $f1$ in the equation 5.2 captures edge score calculation and $f2$ captures the 2 path length score computed from previous iteration.

The computational procedure for longer paths can be well understood with the help of the Figure 5.3. Let i_1 and i_2 be two intermediate nodes other than common neighbors (cn_1, cn_2, cn_3, and cn_4) of the node pair (x,y). The longer path equation (i.e. equation 5.2) contains three parts; f_1, f_2, and penalization function ψ. Three possible paths of length 3 are shown in the figure. f_1 is the significance score (computed from the two length path score only for the first time) of the edges $x-i_1$ of the leftmost and the rightmost figure and $x-cn_1$ of the middle figure. The second part f_2 is the score of 2 path length of previous iteration as shown by paths i_1-cn_1-y (the leftmost), cn_1-i_2-y (the middle one), and i_1-i_2-y in the Figure 5.3. The cumulative effects $(f_1 \times f_2)$ of the significance score f_1 and two path length score f_2 results in the three path length score. Further, more higher length score can be computed in the similar way iteratively, as shown in the Algorithm

2. The algorithm 2 shows an iterative procedure to compute missing links in the complex networks.

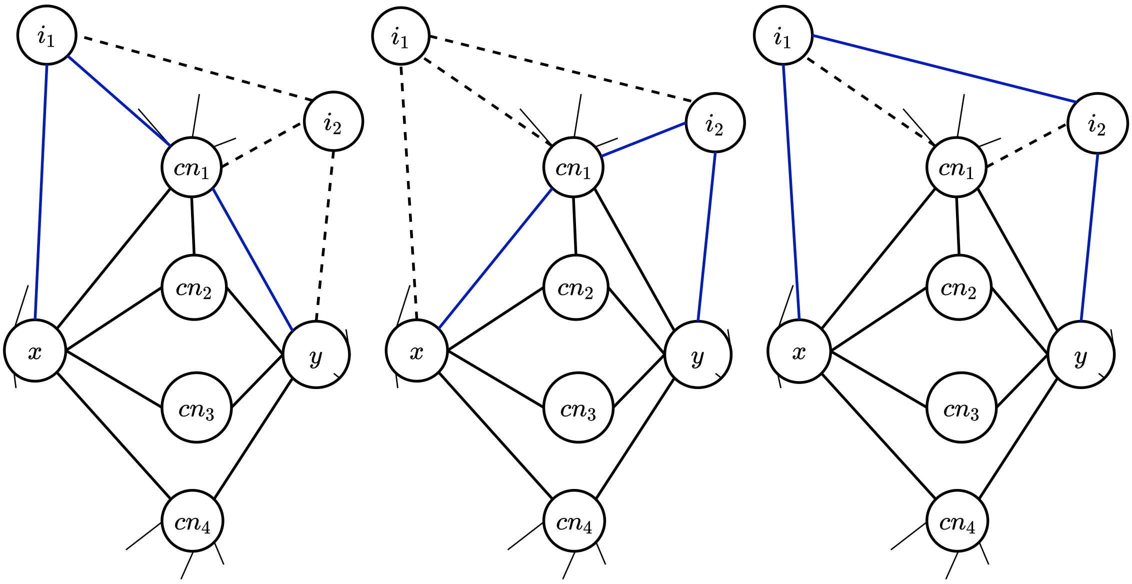


FIGURE 5.3: Three possible paths (blue colored edges) between the node pair (x, y) during the computation of path length-3 score.

Algorithm description The input to the algorithm is a graph (or network), and output is the matrix with the score of all node pairs. This algorithm mainly consists of the following phase; Initialization, computation, and updation. The initialization phase assigns the two matrices *score* and *prev* with the two path length score between every node pair of the network (Line $1-6$). The computation phase of the algorithm iteratively computes scores for higher path length based on the two length path score matrices (Line $7-14$). Finally, the updation phase iteratively updates these two matrices based on the score computed in the previous phase (Line $15-18$).

Algorithm 2: SHOPI

```
Input: Graph $G(V,E)$
Output: Score matrix $score_{n\times n}$ of size $n \times n$
Computation of 2 path length score
$prev_{n\times n} = score_{n\times n} \leftarrow 0$                ▷ Initialization phase
foreach Node pair $(i,j) \in G$ do
    $z \leftarrow CN_{i,j}$                                  ▷ $z$ is the common neighbor of $(i,j)$
    $score_{i,j} \leftarrow \frac{1}{k_z}$                          ▷ $k_z$ is the degree of the node $z$
    $prev_{i,j} \leftarrow score_{i,j}$
Computation of higher path length score                    ▷ Computation phase
for Path length $l \in (l_{max} - 2)$ do
    $temp_{n\times n} \leftarrow 0$                                ▷ Initialization of temp matrix
    foreach Node pair $(i,j) \in G$ do
        foreach Node neighbour $N_i$ of $i$ do
            $f_1 \leftarrow score_{i,N_i}$
            $f_2 \leftarrow prev_{N_i,j}$
            $temp_{i,j} += f_1 \times f_2 \times \psi^{(l-2)}$          ▷ $\psi$ is the penalization parameter
    foreach Node pair $(i,j) \in G$ do
        $score_{i,j} = score_{i,j} + temp_{i,j}$
                                                           ▷ Updation phase
        $prev_{i,j} = temp_{i,j}$
return The Score matrix; $score_{n\times n}$
```

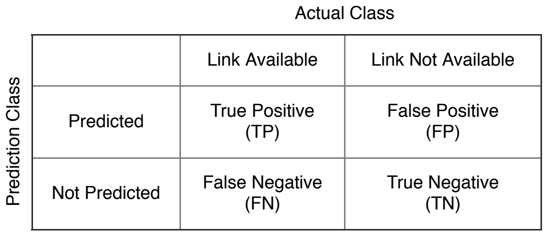
5.3 Experimental study

5.3.1 Evaluation metrics

The link prediction problem is treated as a binary classification task [63], so most of the evaluation metrics of any binary classification task can be used in link prediction evaluation. The evaluation of a binary classification task having two classes can be represented as a confusion matrix [180], as given in Figure 5.4.

In the confusion matrix,

- True Positive (TP): The positive data item (Link Available) predicted as positive (Predicted).



		Actual Class	
Prediction Class		Link Available	Link Not Available
	Predicted	True Positive (TP)	False Positive (FP)
	Not Predicted	False Negative (FN)	True Negative (TN)

FIGURE 5.4: Confusion Matrix

- True Negative (TN): The negative data item (Link Not Available) predicted as negative (Not Predicted).
- False Positive (FP): The negative data item (Link Not Available) predicted as positive (Predicted).
- False Negative (FN): The positive data item (Link Available) predicted as negative (Not Predicted).

Based on the confusion matrix, several metrics can be derived as follows [180].

True Positive Rate (TPR)/Recall/Sensitivity

$$TPR = \frac{\#TP}{\#TP + \#FN}. \tag{5.3}$$

False Positive Rate (FPR)

$$FPR = \frac{\#FP}{\#FP + \#TN}. \tag{5.4}$$

True Negative Rate (TNR)/Specificity

$$TNR = \frac{\#TN}{\#TN + \#FP}. \tag{5.5}$$

$$Precision = \frac{\#TP}{\#TP + \#FP}. \tag{5.6}$$

In the above equations, # represents “the number of”.

Our approach is evaluated on two metrics viz., area under the ROC curve (AUROC) [52, 182] and average precision [180].

Area under the Receiver Operating Characteristics Curve (AUROC). An roc curve is a plot between the true positive rate (sensitivity) on the Y-axis and the false positive rate (1-specificity) on the X-axis. The true positive rate and false positive rate can be evaluated using equation 5.3 and 5.4 respectively. The area under the roc curve [182] is a single point summary statistics between 0 and 1 that can be computed using the trapezoidal rule which sums all the trapezoids under the curve. The value of the AUROC of a predictor should be greater than 0.5, which is the value of a random predictor, i.e., higher the value of AUROC, better the performance of the predictor.

Average Precision (AP). This metric is also a single point summary value computed based on varying threshold[4] values of recall. The average precision value is equal to the precision averaged over all values of recall between 0 and 1, i.e.,

$$AP = \int_{r=0}^{1} p(r)dr,$$

where p is the precision at different threshold value of recall r.

Practically, integral is approximated to sum over the precision at each threshold value, multiplied by the change in recall i.e.,

$$AP = \sum_{k=1}^{R} p(k)\triangle r(k), \tag{5.7}$$

where $\triangle r(k)$ is the change in recall on the set R of different threshold values.

[4] https://sanchom.wordpress.com/tag/average-precision/

5.3.2 Datasets description

This work used twelve network datasets from four different fields (viz., collaboration networks, social networks, citation networks, and biological networks) to study the performance of our approach. Jazz, Netscience, Ca-GrQc, and Ca-HepTh are collaboration networks, Dolphins, Political blogs, Facebook, and Twitter are social networks. Smagri, Cora belong to citation networks and Protein-protein interaction (PPI), Celegansneural belongs to biological networks.

Jazz[5] [178] is the collaboration network of jazz musicians where each musician corresponds to a node, and an edge between two nodes shows that the two musicians have played music together in a band. Netscience[6] [179] is a coauthorship network of scientists working on network theory and experiment compiled by Newman in 2006. Ca-GrQc[7] and Ca-HepTh[4] are collaboration networks of arXiv General Relativity and High Energy Physics Theory, respectively. Dolphins[3] [175] is an undirected social network of 62 bottlenose dolphins in a community living off Doubtful Sound, New Zealand. Political blogs[3] [273] is a directed network of hyperlinks in Political blogs leaning towards the conservatives and the democrats preceding US election 2004. Facebook[4] [284] is social network of 4039 user profiles and network data consisting of 193 circles (i.e. friend-lists) extracted from 10 ego-networks. Twitter[2] is a directed network where each node corresponds to a twitter user, and each directed edge from user A to user B represents that the user A mentioned the user B in a tweet using the "@username". SmaGri[8] [285] is a citation network from Garfield collection produced by HistCite software. The network is the result of searches in Web of Science. Cora[2] [286] is also a directed citation network where a node represents scientific paper, and a directed edge from A to B represents that the paper A cites the paper B. Protein-protein-interaction[9] (PPI) [274] is a biological network of proteins in a cell

[5] http://konect.uni-koblenz.de/networks/

[6] http://www-personal.umich.edu/ mejn/netdata/

[7] https://snap.stanford.edu/data/

[8] http://vlado.fmf.uni-lj.si/pub/networks/data/

[9] https://icon.colorado.edu/#!/networks

TABLE 5.1: Topological information of real-world network datasets

Class	Network	$\|V\|$	$\|E\|$	$\langle D\rangle$	$\langle K\rangle$	$\langle C\rangle$	r	H
Collaboration networks	Jazz	198	2742	2.235	27.697	0.6201	0.020	1.395
	Netscience	1589	2742	5.823	3.451	0.878	0.461	2.010
	Ca-GrQc	5242	14496	6.049	5.531	0.687	0.659	3.051
	Ca-HepTh	8361	15751	7.025	3.768	0.636	0.293	2.305
Social networks	Dolphins	62	159	3.302	5.129	0.258	-0.043	1.326
	Political blogs	1490	16718	2.738	22.44	0.361	-0.221	3.621
	Facebook	4039	88234	3.693	43.691	0.617	0.063	2.439
	Twitter	5182	84851	3.028	32.748	0.295	-0.192	4.083
Citation networks	SmaGri	1059	4917	2.981	9.286	0.349	-0.192	4.083
	Cora	2708	5278	6.310	3.898	0.293	-0.065	2.798
Biological networks	PPI	2375	11693	5.096	9.847	0.388	0.453	3.475
	Celegansneural	297	2148	2.447	14.465	0.308	-0.163	1.800

where a node represents a protein and edge denotes the interaction between two proteins. Celegansneural[3] [11]: A neural network of C. Elegans compiled by D. Watts and S. Strogatz in which each node refers a neuron and, an edge joins two neurons if they are connected by either a synapse or a gap junction.

The basic structural properties of 12 networks are tabulated in the Table 5.1 wherein $|V|$, $|E|$, $\langle D\rangle$, and $\langle K\rangle$ are number of vertices, number of edges, average distance, and average degree of the network respectively. $\langle C\rangle$, r, and H are average clustering coefficient, coefficient of assortativity, and degree of heterogeneity, respectively.

5.3.3 Results analysis

We have performed a comprehensive experiment of the proposed method viz., SHOPI, with the baseline methods on a different class of networks. We evaluate our method against two evaluation metrics mentioned in section 5.3.1. As the proposed method focus on different paths and parameters, so its sensitivity to parameters is also represented.

AUROC Table 5.2 shows the AUROC results of the proposed and baseline methods on 12 network datasets. The best result against each dataset is shown in bold-face. We observe that the SHOPI best performs mostly on collaboration (i.e., Jazz, Netscience, and Ca-HepTh) and social networks (i.e., Political blogs, Facebook, and Twitter). For biological networks, the LP is the best on the PPI network, and SHOPI is best on the Celegansneural network. On the PPI network, SHOPI shows better results than neighborhood-based methods and comparable to the embedding method (i.e., Node2vec). Jaccard performs best on the Ca-GrQc dataset, and AA is best on the Dolphin dataset. On citation networks, LP and Node2vec perform individually best for SmaGri and Cora networks, respectively. Our method performs equivalent to the Katz method on the SmaGri dataset. One thing to note that, on citation networks, the CAR method shows lower performance even worst than a random predictor. The same behavior is observed on Ca-HepTh and Dolphins datasets. The reason might be the absence or least number of local community structures in the citation datasets.

We conclude that the proposed SHOPI is significantly better than neighborhood methods and other path-based methods (i.e., Katz and LP) on collaboration and social networks. Our method shows lower AUROC results with the neighborhood methods on citation networks. The reason may be the large number of components available in the network. SmaGri and cora networks contain a total of 36 and 78 connected components respectively. The largest component in the SmaGri is 1024 nodes networks while other 35 nodes are isolated. Moreover, Cora contains a largest component with 2485 nodes and other components are with lower degree less that 10. Another 56 out of 78 components contain only two nodes that contribute zero to the proposed index. This results in less number of paths of different lengths and contributing less to the similarity score. This phenomenon also observed in the Table 5.4 against the citation networks where there is a large change in the AUROC values from path of length 3 to path of length 6. Though large number of connected components are also available in the Netscience, Ca-GrQc, Ca-HepTh datasets but the size of their components other than largest ones are not very small or isolated. Moreover, these networks are highly clustered resulting in more three length paths. And hence SHOPI performs better on these datasets.

TABLE 5.2: AUROC Results

	CN	AA	JC	PA	CAR	Katz	LP	Node2vec	SHOPI
Jazz	0.9481	0.9520	0.9504	0.7895	0.9314	0.9357	0.9433	0.8732	**0.9565**
Netscience	0.9412	0.9287	0.9506	0.6390	0.5328	0.9262	0.9403	0.8924	**0.9511**
Ca-GrQc	0.9215	0.9162	**0.9294**	0.7417	0.6055	0.9160	0.9281	0.9089	0.9225
Ca-HepTh	0.8957	0.8949	0.8874	0.7281	0.4388	0.8893	0.9024	0.8750	**0.9086**
Dolphins	0.7454	**0.7765**	0.7697	0.7264	0.3571	0.7562	0.7693	0.7507	0.7454
Political blogs	0.9410	0.9333	0.9079	0.9342	0.7399	0.9466	0.9498	0.8666	**0.9508**
Facebook	0.9918	0.9932	0.9895	0.8328	0.9447	0.6076	0.9912	0.9915	**0.9938**
Twitter	0.9321	0.9338	0.8997	0.9221	0.7385	0.4452	0.9389	0.8522	**0.9426**
SmaGri	0.8588	0.8498	0.7905	0.8368	0.4443	0.8692	**0.8916**	0.7452	0.8685
Cora	0.7057	0.7389	0.7345	0.6712	0.4594	0.8304	0.8209	**0.8779**	0.7863
PPI	0.8883	0.9001	0.8900	0.8154	0.7461	0.9384	**0.9409**	0.9101	0.9129
Celegansneural	0.8154	0.8382	0.7927	0.7351	0.8465	0.8796	0.8709	0.7956	**0.8851**

TABLE 5.3: Average Precision (AP) Results

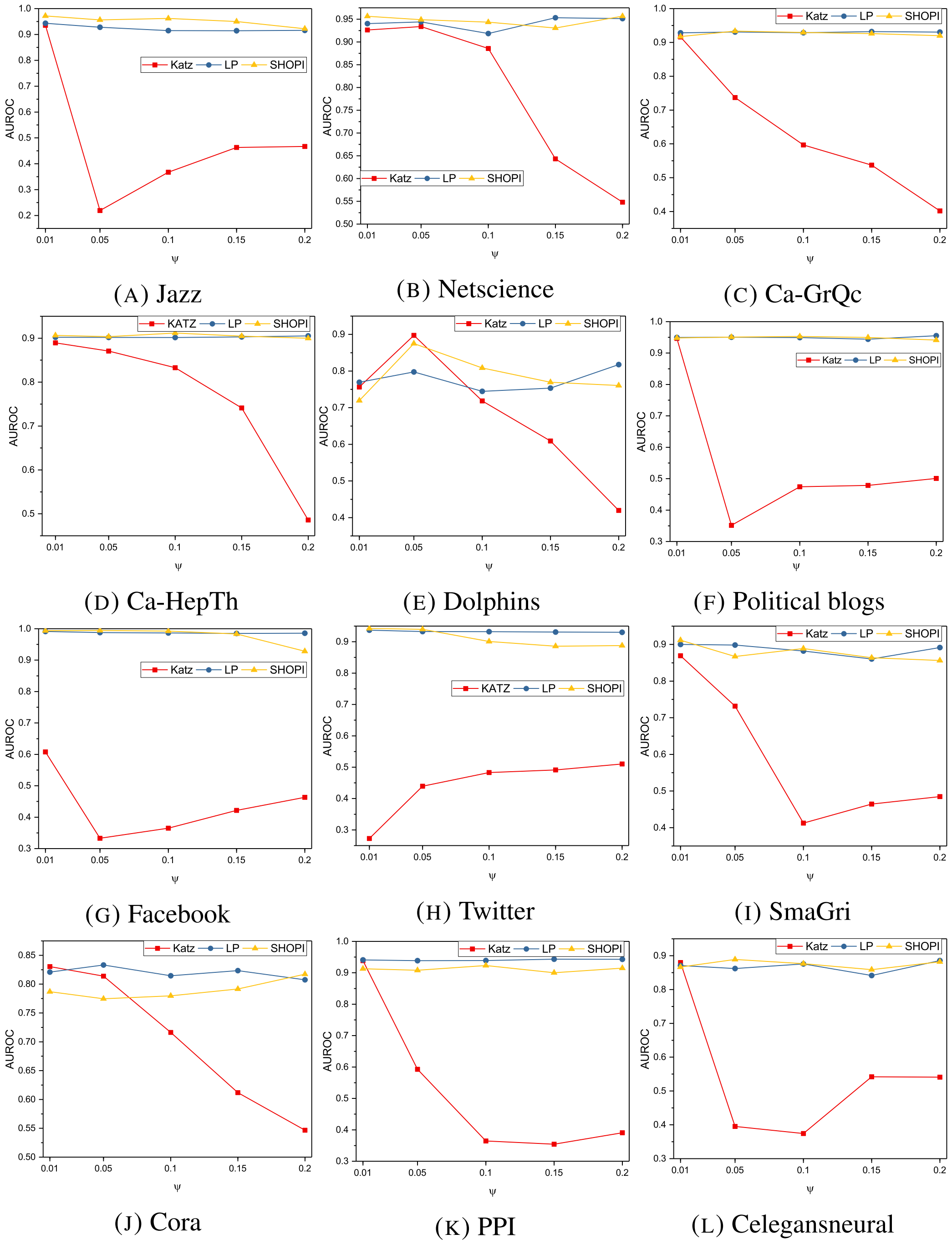
	CN	AA	JC	PA	CAR	Katz	LP	Node2vec	SHOPI
Jazz	0.3302	0.3354	0.2674	0.1064	0.3216	0.3118	0.3157	0.0860	**0.3409**
Netscience	0.1622	0.1714	0.0976	0.0033	0.1087	**0.1845**	0.1473	0.0817	0.1350
Ca-GrQc	0.2236	**0.2300**	0.0632	0.0191	0.2172	0.2142	0.2209	0.0489	0.1471
Ca-HepTh	0.0505	**0.0748**	0.0304	0.0005	0.0287	0.0001	0.0550	0.0304	0.0679
Dolphins	0.0313	**0.0517**	0.0403	0.0290	0.0070	0.0284	0.0333	0.0244	0.0410
Political blogs	0.0766	0.0748	0.0173	0.0338	0.0633	**0.0897**	0.0849	0.0084	0.0693
Facebook	0.2446	0.2619	0.1724	0.0190	0.2353	0.0521	0.2293	0.1252	**0.2940**
Twitter	0.0419	**0.0458**	0.0057	0.0192	0.0347	0.0031	0.0382	0.0057	0.0349
SmaGri	0.0266	0.0281	0.0026	0.0150	0.0132	0.0183	0.0284	0.0036	**0.0312**
Cora	0.0052	0.0128	0.0038	0.0008	0.0002	0.0100	0.0092	0.0069	**0.0132**
PPI	0.0916	0.0909	0.0328	0.0654	0.0912	0.0295	**0.1404**	0.0415	0.0929
Celegansneural	0.0299	0.0383	0.0167	0.0206	0.0117	0.0424	0.0346	0.0189	**0.0447**

Average precision (AP) Table 5.3 depicts the average precision results of SHOPI against several datasets. It shows the best results on Jazz and Facebook belong to collaboration and social network respectively. Moreover, It performs best on both datasets of citation networks, and Celegansneural of biological network. AA method achieves best results on Ca-GrQc, Ca-HepTh, Dolphins and Twitter datasets. The precision results of the SHOPI is comparable to the AA and **second-best** performer on these datasets except Ca-GrQc. Katz method shows the best results on Netscience and Political blogs datasets, and the LP is the best performer on PPI dataset which significant compared to other methods. Our method is **second-best** performer on PPI dataset.

Effects of the parameter value ψ and sensitivity analysis This work includes paths of different lengths as features to compute the similarity between two nodes of the network. The equation to compute this score incorporates a parameter ψ. We investigate the effect of this parameter ψ on the accuracy (i.e., AUROC and AP) by applying five different values of it ranging from 0.01 to 0.20.

AUROC sensitivity The effect of ψ on the AUROC values corresponding to the SHOPI and the two existing methods viz., Katz index, and LP index is shown in Figure 5.5. From the table, we observe that the SHOPI is having no significant effects of ψ on collaboration networks [See Figs. 5.5a, 5.5b, 5.5c, 5.5d] and social networks [See Figs. 5.5e, 5.5f, 5.5g, 5.5h] except Dolphins where a significant effect observed. We observed that our method shows higher AUROC values on almost all values of ψ when compared to the LP on collaboration and social network datasets. On citation and biological datasets, the SHOPI and the LP show more effect of ψ. The parameter ψ greatly affects the Katz index on all datasets, as shown in the figure. The Katz index shows high AUROC value when the parameter ψ is very low near to zero and decreased heavily with increasing values of this parameter. The LP method is almost not affected by this parameter.

AP sensitivity Figure 5.6 shows the effects of the parameter ψ on average precision (AP) of the SHOPI, Katz index, and LP method. The X-axis shows the different values of ψ, and the corresponding AP values are shown on the Y-axis. The Katz index here, greatly influenced by increasing values of the parameter on all datasets, as shown in the figure. The SHOPI and the LP show almost the same behavior on different values of ψ. This parameter affects these two methods on citation and biological datasets to some extent. On the Dolphins dataset, they show almost the same fluctuation on different ψ values. The SHOPI is having larger AP compared to the LP method on Jazz and Ca-HepTh, while lower value on Netscience and Ca-GrQc (in collaboration category). The Katz index has the best performance on the Netscience, Political blogs, and PPI datasets when



(A) Jazz

(B) Netscience

(C) Ca-GrQc

(D) Ca-HepTh

(E) Dolphins

(F) Political blogs

(G) Facebook

(H) Twitter

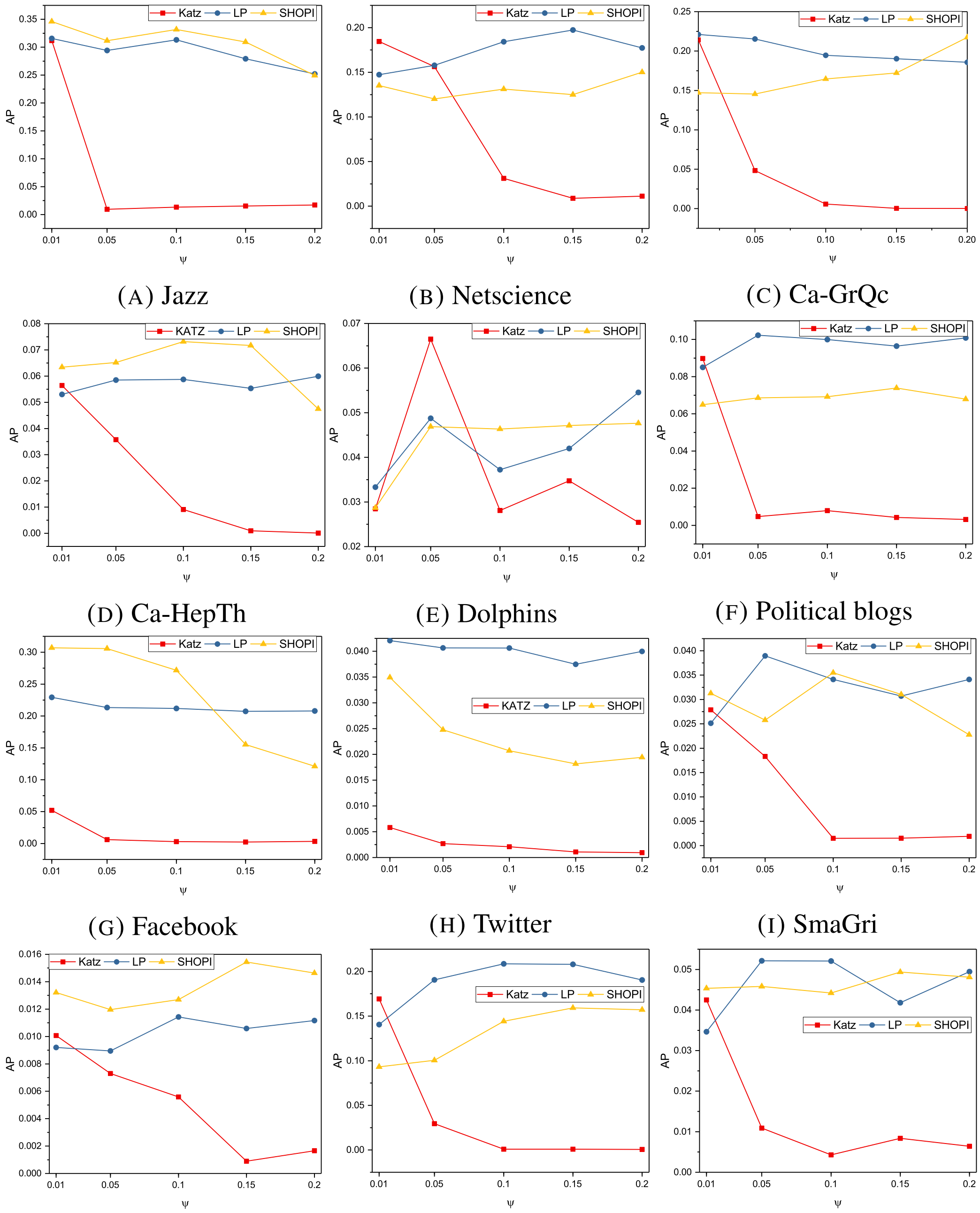
(I) SmaGri

(J) Cora

(K) PPI

(L) Celegansneural

FIGURE 5.5: AUROC results sensitivity corresponding to different parameter values of ψ



(A) Jazz (B) Netscience (C) Ca-GrQc

(D) Ca-HepTh (E) Dolphins (F) Political blogs

(G) Facebook (H) Twitter (I) SmaGri

(J) Cora (K) PPI (L) Celegansneural

FIGURE 5.6: AP results sensitivity corresponding to different parameter values of ψ

the parameter value $\psi = 0.01$ [See Figs. 5.6b, 5.6f, 5.6k] and Dolphins dataset when $\psi = 0.05$ [See Fig. 5.6e].

TABLE 5.4: Effects of considering longer path lengths on the accuracy of link prediction

	AUROC				AP			
	l=3	l=3,4	l=3,4,5	l=3,4,5,6	l=3	l=3,4	l=3,4,5	l=3,4,5,6
Jazz	0.960343	0.963813	0.963826	0.955036	0.312052	0.331052	0.329758	0.313891
Netscience	0.939394	0.937596	0.930313	0.926952	0.152573	0.152268	0.151059	0.123492
Ca-GrQc	0.923291	0.928715	0.922134	0.913700	0.170711	0.175578	0.177503	0.131930
Ca-HepTh	0.879348	0.885693	0.884419	0.908637	0.067320	0.071834	0.068330	0.067944
Dolphins	0.758650	0.724387	0.752349	0.761566	0.040543	0.029157	0.031164	0.033114
Political blogs	0.941831	0.945521	0.936282	0.946530	0.066493	0.074494	0.070124	0.064945
Facebook	0.995228	0.994226	0.994176	0.995033	0.303534	0.308186	0.307047	0.316250
Twitter	0.937118	0.937104	0.934142	0.939811	0.036918	0.038936	0.037355	0.027084
SmaGri	0.828404	0.824603	0.852049	0.868572	0.025766	0.023815	0.028900	0.031848
Cora	0.749547	0.724683	0.743721	0.786383	0.013809	0.011471	0.011824	0.013730
PPI	0.897994	0.904383	0.895074	0.911787	0.125360	0.119760	0.109083	0.092934
Celegansneural	0.856932	0.868140	0.856421	0.885100	0.042001	0.040550	0.043622	0.044754

Significance of higher order paths This work considers path as a discriminating feature. It considers path up to length six as the average length of the path between any two nodes is six (also called six degrees of separation) in most social networks. We also observed that the network datasets considered for our experiments have average path lengths less or equal to six except Ca-GrQc ($\langle D \rangle = 6.049$), Ca-HepTh ($\langle D \rangle = 7.025$), and Cora ($\langle D \rangle = 6.310$). The influence of considering the longer paths on the accuracy (AUROC and AP) of the proposed approach is shown in Table 5.4. We observe that there is no significant difference in the AUROC values with path length $l = 3$ and path length $l = 6$ on social network category (See rows corresponding to the Dolphins, Political blogs, Facebook, Twitter). Moreover, the significant difference observed on Ca-HepTh, Citation networks (See rows corresponding to the SmaGri and Cora datasets), and biological networks (See rows corresponding to the PPI and Celegansneural datasets). This value even decreases in most collaboration networks when longer paths are considered. The effects of considering longer paths diminish the AP values on most networks except Jazz, Ca-HepTh, Facebook, SmaGri, and Celegansneural datasets with an insignificant increase in AP. We conclude that there is no significant increment in the accuracy when compared to path lengths $l = 3$ and $l = 6$.

Complexity analysis. The time complexity here will be discussed based on the assumption that most complex networks are sparse, and hence, the average number of edges (links) for each node is $\langle K\rangle$ (average degree of the network). l_{max} is the maximum length of the path between any pair of node and path longer than l_{max} is considered to have zero influence to the edge likelihood probability. Some optimization has been applied whenever possible.

The main crux of our algorithm is the computation of the 2 and higher path length score. In step $3-5$, for loop of step 3 iterates to $O(n^2)$ times. Line 4 costs $O(\langle K\rangle + \langle K\rangle)$ to compute CNs for a given node pair when the lists of neighbors are hashable and $O(nlgn)$ when adjacency list are used, resulting in total $O(n^2\langle K\rangle))$ time for 2 length path score computation. For higher order paths, step $9-13$ consisting of two loops and take $O(n^2\langle K\rangle)$ time and line $14-16$ take $O(n^2)$. This results in total $((l_{max}-2)[O(n^2\langle K\rangle)+O(n^2)])$ time for higher paths, where $l_{max}=6$ is a constant. Thus, the total computational complexity of the proposed algorithm is $((l_{max} - 2)[O(n^2\langle K\rangle) + O(n^2)]) + O(n^2\langle K\rangle)$ which is equal to $(l_{max}-1)O(n^2\langle K\rangle)+(l_{max}-2)O(n^2)$. Finally, it is approximated to $(l_{max}-1)O(n^2\langle K\rangle)$ time.

The computational complexities of baseline methods are presented in [45]. The CAR costs $O(nK^4)$, which is more complex as it computes time-consuming local community links (LCL). Other methods like CN, AA, JC estimates $O(n\langle K\rangle^3)$ while PA costs $O(n\langle K\rangle^2)$. The computational complexity of the Katz index is $O(n\langle K\rangle + n^3 + n)$ where matrix subtraction takes $O(n\langle K\rangle)$ time, matrix inversion takes $O(n^3)$ time and $O(n)$ required by the subtraction of the diagonal elements in the identity matrix as shown in the equation 3.17. The local path index avoids matrix inversion by exploiting the adjacency matrix power of the previous summation term and costs $O(ln^2\langle K\rangle)$ time. The Node2vec [125] method is based on a random walk sampling, which is efficient compared to pure BFS/DFS. The effective time complexity of the Node2vec is $O(\frac{l}{k'(l-k')})$ per sample where l is walk length and k' is neighborhood size.

Statistical test We perform a statistical test [277] to show the significant difference of the proposed method with the baseline methods. We employ the Friedman test [278, 279] to analyze whether there is a significant difference among multiple methods. It is a non-parametric counterpart of the repeated measures ANOVA. The Friedman test results for both area under the ROC curve (AUROC) and average precision (AP) are shown in Table 5.5. The observed test values of the Friedman test for both AUROC and AP are 51.178 and 59.336, which are greater than the corresponding χ^2 value (i.e., $\chi^2(\alpha_c, D_f)$). With the confidence interval $\alpha = 0.05$ and degree of freedom $D_f = 8$, χ^2 value is 15.51, obtained from the χ^2 table available in the literature. This results in the rejection of the null hypothesis, as shown in the last column of the table. This test confirms that there is a significant difference among the methods for both AUROC and AP.

Once the null hypothesis is rejected, We perform a post hoc test to find the methods by means of which the significant difference occurs. Lots of post hoc tests are available in the literature, we have opted for the Friedman-Convover post hoc test with the SHOPI as the control method and ”Holm” as the adjusting method to control the family-wise error rate (FWER). The posthoc test results are shown in Table 5.6. The table shows the adjusted p-values using the Holm method for both AUROC and AP. From the table, we observe that the SHOPI is significantly different from all the baseline methods except AA and LP for AUROC, and significantly different except CN, AA, and LP for AP.

5.4 Conclusion

The topology and evolution of complex real-world networks are constrained by various organizing principles of topology and dynamics and happens to be a major research area. Researchers have been addressing structural and topological issues of complex networks as well as their dynamics. Some of the important concepts that have evolved in the field of complex networks are small-world and scale-free networks. Features corresponding to these concepts are the average path length and degree distribution, respectively. Lots of

TABLE 5.5: The Friedman test on area under the ROC Curve (AUROC) and average precision (AP)

	Dataset	IS-value									Test value	State Result
		CN	AA	JC	PA	CAR	Katz	LP	Node2vec	SHOPI	F_f	Is $F_f > \chi^2$?
AUROC	Jazz	0.9481	0.9520	0.9504	0.7895	0.9314	0.9357	0.9433	0.8732	0.9565	51.178	Null Hypothesis Rejected
	Netscience	0.9412	0.9287	0.9506	0.6390	0.5328	0.9262	0.9403	0.8924	0.9511		
	Ca-GrQc	0.9215	0.9162	0.9294	0.7417	0.6055	0.9160	0.9281	0.9089	0.9225		
	Ca-HepTh	0.8957	0.8949	0.8874	0.7281	0.4388	0.8893	0.9024	0.8750	0.9086		
	Dolphins	0.7454	0.7765	0.7697	0.7264	0.3571	0.7562	0.7693	0.7507	0.7454		
	Political blogs	0.9410	0.9333	0.9079	0.9342	0.7399	0.9466	0.9498	0.8666	0.9508		
	Facebook	0.9918	0.9932	0.9895	0.8328	0.9447	0.6076	0.9912	0.9915	0.9938		
	Twitter	0.9321	0.9338	0.8997	0.9221	0.7385	0.4452	0.9389	0.8522	0.9426		
	SmaGri	0.8588	0.8498	0.7905	0.8368	0.4443	0.8692	0.8916	0.7452	0.8685		
	Cora	0.7057	0.7389	0.7345	0.6712	0.4594	0.8304	0.8209	0.8779	0.7863		
	PPI	0.8883	0.9001	0.8900	0.8154	0.7461	0.8384	0.9409	0.9101	0.9129		
	Celegansneural	0.8154	0.8382	0.7927	0.7351	0.8465	0.8796	0.8709	0.7956	0.8851		
AP	Jazz	0.3302	0.3354	0.2674	0.1064	0.3216	0.3118	0.3157	0.0860	0.3409	59.336	Null Hypothesis Rejected
	Netscience	0.1622	0.1714	0.0976	0.0033	0.1087	0.1845	0.1473	0.0817	0.1350		
	Ca-GrQc	0.2236	0.2300	0.0632	0.0191	0.2172	0.2142	0.2209	0.0489	0.1471		
	Ca-HepTh	0.0505	0.0748	0.0304	0.0005	0.0287	0.0001	0.0550	0.0304	0.0679		
	Dolphins	0.0313	0.0517	0.0403	0.0290	0.0070	0.0284	0.0333	0.0244	0.0410		
	Political blogs	0.0766	0.0748	0.0173	0.0338	0.0633	0.0897	0.0849	0.0084	0.0693		
	Facebook	0.2446	0.2619	0.1724	0.0190	0.2353	0.0521	0.2293	0.1252	0.2940		
	Twitter	0.0419	0.0458	0.0057	0.0192	0.0347	0.0031	0.0382	0.0057	0.0349		
	SmaGri	0.0266	0.0281	0.0026	0.0150	0.0132	0.0183	0.0284	0.0036	0.0312		
	Cora	0.0052	0.0128	0.0038	0.0008	0.0002	0.0100	0.0092	0.0069	0.0132		
	PPI	0.0916	0.0909	0.0328	0.0654	0.0912	0.0295	0.2068	0.0415	0.0929		
	Celegansneural	0.0299	0.0383	0.0167	0.0206	0.0117	0.0424	0.0346	0.0189	0.0447		

TABLE 5.6: The Posthoc Friedman Conover Test (Control method = SHOPI, Correction method = Holm)

Metric	p-value							
	CN	AA	JC	PA	CAR	Katz	LP	Node2vec
AUROC	0.0226	0.2209	0.0036	1.40E-09	3.10E-13	4.50E-05	0.8772	0.000029
AP	0.71303	0.78796	1.40E-08	9.10E-10	2.00E-06	0.00257	1	6.90E-10

works are available based on these two topological properties. In work entitled SHOPI, we exploit the path as a discriminating feature to predict missing links in the networks. This work targets the resource allocation process and tries to constrain the information (resource) leak through the common neighbors by penalizing them based on their connections. And, hence, it tries to maximize the information flow between the pair of nodes that characterize the similarity score between them. Higher-order paths (based on the six degrees of separation) are also used as discriminating features with one more penalization function applied to them. The comprehensive experimental results on several networks show that the proposed approach viz., SHOPI outperforms the baseline methods. Consideration of higher-order path index affects little bit to the prediction accuracy, though, significantly affects computational complexity. The statistical test performed here shows the significant difference between the proposed approach with the baseline methods.