A Quasi-local Approach for Link Prediction in Graphs Based on Node Affinity Tree

Mohammad Nowsin Amin Sheikh, Md Kamrul Islam, *Member, IAENG,* Sagar Chakraborty, I-Shyan Hwang, Wissem Inoubli, and Md Abdul Kader

***Abstract*—Similarity-based approaches, commonly used measures for link prediction in graphs, rely on the assumption that two nodes with high similarity are likely to form a link. Among the three categories of similarity-based approaches, local approaches are simpler and faster, while global approaches are more complex and slower but often yield better performers in real-world graphs. To achieve a good compromise between computational cost and prediction performance, quasi-local approaches utilize neighborhood information up to a certain hop distance. However, most of them are still designed based on global information in graphs. Moreover, they consider all neighborhood information, even though some may be noisy and affect prediction performance. Besides, most of these approaches define similarity functions based solely on structural features, ignoring node attributes that are available in many real-world graphs. This paper presents a novel quasi-local approach, LPST (Link Prediction based on Similarity between affinity Trees), for link prediction utilizing affinity among neighbors and node attributes. LPST starts with generating affinity trees with neighbors based on a branch and bound strategy for each end node. Then, a similarity function is defined based on the available attributes and structural features that are extracted from the affinity trees. LPST is evaluated against both non-attributed and attributed homogeneous graphs. The performance of the proposed approach is then compared to state-of-the-art similarity-based link prediction approaches. Experimental results show the effectiveness of the proposed LPST approach in predicting links in graphs.**

***Index Terms*—link prediction, homogeneous graph, similarity-based approach, affinity tree.**

I. Introduction

LINK prediction is one of the most compelling yet challenging problems in the area of graph mining. It predicts the probability of a link between two unlinked nodes by leveraging the available information in a graph [1], [2]. Link prediction approaches have been widely used in many real-world applications, such as friend suggestions in social networks [3], product recommendations in e-commerce [1], inferring interactions in biological networks [4], [5], [6], [7], hotel recommendations in hotel sales networks [8].

Similarity-based link prediction approaches, which assume that two similar nodes are likely to be connected, have been extensively studied in the literature because of their explainability. These approaches are categorized into three groups: local approaches, global approaches, and quasi-local approaches [9]. Local approaches define the similarity function based on the local neighborhood information only. Local approaches, such as common neighbour (CN) [10], Adamic-Adar(AA) [11], Resource Allocation (RA) [12], Jaccard (JA) [13], CAR [14], LLHN [15], CCLP [16], NLC [17], utilize neighbourhood information. Though local approaches are simple, highly parallelable and applicable to large graphs, they are less effective in predicting links beyond the second hop neighbourhood. Global approaches, such as Katz(KZ) [18], Global Leicht-Holme-Newman (GLHN) [15], Flow Propagation(FP) [19], Random Forest Kernel(RFK) [9], Random Walk(RW) [20], Random Walk with Restart (RWR) [21], overcome the limitation of local approaches by utilizing the global adjacency matrix to compute the link probability beyond second-hop neighbourhood. However, the use of global information generates high computational time and difficulties in parallel implementation for the global approaches, and they are infeasible to be applied to large real-world graphs. Quasi-local approaches make a good trade-off between computational time and prediction performance; they are almost as computationally efficient as local approaches, but not limited to neighbors of neighbors. Local path index (LPI) [22], local random walk (LRW) [23], and friend link (FL) [24] are examples of quasi-local approaches that consider the neighbours up to a predefined hop. Quasi-local approaches have gained much popularity in recent years. In this paper, we address two important issues in state-of-the-art quasi-local approaches. Firstly, although the possibility of spurious or noisy links in a real-world graph cannot be ignored and can have an impact on the prediction results, the issue of spurious links in graphs is neglected in nearly all state-of-the-art similarity-based approaches [25], [26]. Secondly, most of the state-of-the-art quasi-local approaches ignore the presence of node attribute information, which could improve prediction performance. In this paper, we present a new noble quasi-local link prediction approach, namely LPST (Link Prediction based on Similarity between affinity Trees), by maximizing the use of available structural features and node attributes and pruning out the suspectable spurious (we call weak) links neighbors from

Manuscript received July 24, 2024; revised April 21, 2025.

This work was supported in part by the University Grants Commission, Bangladesh and Jashore University of Science and Technology, Bangladesh under a research grant (Grant no. 23-FOET-07) and NSTC 113-2221-E-155-055, Taiwan

Mohammad Nowsin Amin Sheikh is an assistant professor of Computer Science and Engineering department, Jashore University of Science and Technology, Jashore 7408, Bangladesh. (e-mail: n.amin@just.edu.bd).

Md. Kamrul Islam is an associate professor of Computer Science and Engineering department, Jashore University of Science and Technology, Jashore 7408, Bangladesh. (corresponding author to provide phone: +8801729885196, e-mail: mk.islam@just.edu.bd).

Sagar Chakraborty is a doctoral candidate in the Jashore University of Science and Technology, Jashore 7408, Bangladesh. (e-mail: sagar_cse_ku@yahoo.com)

I-Shyan Hwang is a professor of Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 32003, Taiwan. (corresponding author to provide e-mail: ishwang@saturn.yzu.edu.tw)

Wissem Inoubli is an associate professor of Computer Science, University of Artois and a member of the Centre de Recherche en Informatique de Lens (CRIL), 62030 Arras Cedex, France. (e-mail: wissem.inoubli@univ-artois.fr)

Md. Abdul Kader is a senior lecturer of Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan-Pahang, Malaysia. (e-mail: abdkader@umpsa.edu.my)

the link prediction function. The precise contributions are: (i) generating an affinity tree for each end node based on the branch and bound strategy by discarding weak links, (ii) utilizing available node attributes along with the structural features to define the link prediction function , and (iii) evaluating LPST on both attributed and non-attributed graphs.

II. The Proposed LPST approach

In the following sections II-A—II-B, we present the proposed LPST approach. Overall, the approach starts with the crucial step of generating an affinity tree for each end node of a link based on the branch and bound strategy. Then, a noble node similarity function is defined based on the extracted structural and available attribute features from the two trees.

A. Affinity tree generation

A **graph**, $G(V, E)$ formally represents a network, where V denotes the set of vertices (nodes), and E denotes the set of edges (links). The link between two nodes *x* and *y* is noted as $e_{x,y}$. A **path** is a sequence of links between a pair of nodes *x, y* that connects multiple nodes in a graph. If there is a link $e_{x,y}$ such that $e_{x,y} \in E$, then *y* is called a neighbour of the node, *x*. For an undirected graph, the node, *x* is also called the neighbour of the node, *y*. The neighbours of a node form a neighbour set for the node, *x* and denoted by Γx.

The first step of the proposed LPST approach is to generate an affinity tree for each node of a link. For a link $e_{x,y}$, two trees will be generated based on the branch and bound strategy for the nodes *x* and *y*. The breadth first search (BFS) is followed for the branching operation. The branching operation is guided by a probability value that indicates the probability of reaching the other end node (*y/x*) from the current node (*x/y*). Assume a x-rooted tree will be generated for the node *x*, assuming the probability of reaching node *y* from node *x* is $P_x = 1.0$. When branching, the cost of the link $e_{x,t}$ for each child *t* of *x* is computed from their structural and attribute affinities of the nodes pair, *(x,t)* where the structural affinity is defined as the Jaccard coefficient in (1).

$$JA_{xt} = \frac{|\Gamma x \cap \Gamma t|}{|\Gamma x \cup \Gamma t|} \tag{1}$$

For the attribute affinity, we consider both numerical and categorical attributes of nodes. Overlapping and cosine similarity functions are used to compute the affinity of categorical and numerical attributes, respectively. Let, $CT^x = \{ct_1^x, ct_2^x, ..., ct_C^x\}$ and $CT^t = \{ct_1^t, ct_2^t, ..., ct_C^t\}$ are sets of C categorical attributes of *x* and *t*, respectively. The categorical attribute affinity of the nodes *x* and *t* is defined in (2).

$$S_{xt}^{ct} = \frac{\sum_{i=1}^{C} S(ct_i^x, ct_i^t)}{C} \tag{2}$$

where, $S(ct_i^x, ct_i^t) = 1$ if $ct_i^x = ct_i^t$, and 0 otherwise.
Now let, $NM^x = \{nm_1^x, nm_2^x, ..., nm_N^x\}$ and $NM^t = \{nm_1^t, nm_2^t, ..., nm_N^t\}$ are the sets of N numerical attributes of *x* and *t*, respectively. The affinity of *x* and *t* with respect to their numerical attributes is defined in (3).

$$S_{xt}^{nm} = \frac{\sum_{i=1}^{N} nm_i^x \times nm_i^t}{\sqrt{\sum_{i=1}^{N}(nm_i^x)^2} \times \sqrt{\sum_{i=1}^{N}(nm_i^t)^2}} \tag{3}$$

The accumulated attribute affinity of the node, *x* and its child, *t* is computed from their numerical (2) and categorical (3) attributes affinities using (4).

$$S_{xt} = \frac{(C \times S_{xt}^{ct}) + (N \times S_{xt}^{nm})}{C + N} \tag{4}$$

The overall affinity between nodes *x* and *t* is defined based on their attributes (4) and structural (1) affinities in (5).

$$S(x, t) = \gamma \times S_{xt} + (1 - \gamma) JA_{xt} \tag{5}$$

The user-defined parameter $0 \leq \gamma \leq 1$ gives the flexibility to emphasize more on the structural or attribute information in defining the affinity function (5). For a graph without node attributes, we set $\gamma = 0$.

The cost of the link $e_{x,t}$ is computed from its affinity score, $S(x, t)$ as (6).

$$Cost(x, t) = 1 - S(x, t) \tag{6}$$

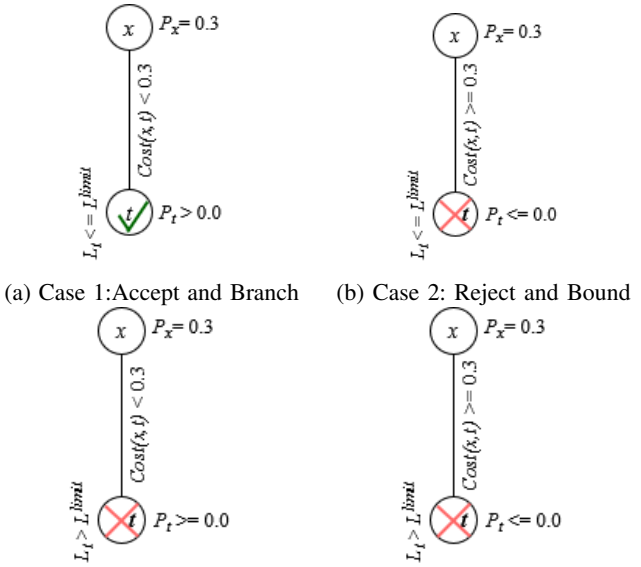
Hence, the probability of reaching the node *y* from root node *x* through node *t* is computed using (7).

$$Probability, P_t = P_x - Cost(x, t) \tag{7}$$

The inclusion of the link $e_{x,t}$ in the x-rooted affinity tree is decided based on its current probability (p_t) and a user defined depth limit parameter, L^{limit}. The link $e_{x,t}$ is included in the x-rooted tree only in the case its probability is above zero (*i.e.* $P_t > 0$) and its depth (L_t) from the root is below or equal to the depth limit (*i.e.* $L_t <= L^{limit}$). The four cases of accepting or rejecting the link $e_{x,t}$ in the x-rooted tree are illustrated in Fig. 1. The branching and bound process is repeated for each child and grandchild of the root *x* until every node is visited or all children meet the bounding conditions to generate the x-rooted tree. For each node, *t* of the tree, a data structure is maintained that stores its $parent$, a probability score(P) to reach the destination node, current depth (L) from its root, a popularity score as the number of visit (Pop) to the node, its visiting status (generated/ visited) and degree (D) of the node. When a node t is branched from its parent x, its data structure is initialized with $parent_t = x$, $Pop_t = Pop_t + 1$, depth to $L_t = L_x + 1$, $D_t = |\Gamma t|$ and $Status_t =' generated'$. During branching, a node could be branched from several parents in a graph. Each time a branched node is explored, its data structure is updated based on its current probability score, P_t and its proposed probability score, $P_x - Cost(x, t)$. Two cases are observed here:

1) $P_t < [P_x - Cost(x, t)]$: In this case, the proposed probability is lower than the current probability. As a result, the connection request from the current node (*x*) to the child (*t*) is rejected with setting $Pop_t = Pop_t + 1$ and $D_t = |\Gamma t|$.
2) $P_t \geq [P_x - Cost(x, t)]$: In this case, the connection request is accepted as the proposed probability is greater than or equal to the current probability. The information of the node, *t* is updated as $P_t = P_x - Cost(x, t)$, $Pop_t = Pop_t + 1$, $L_t = L_x + 1$ and $D_t = |\Gamma t|$.

The overall procedure of affinity tree generation is described in Algorithm 1.



(a) Case 1:Accept and Branch (b) Case 2: Reject and Bound

(c) Case 3: Reject and Bound (d) Case 4: Reject and Bound

Fig. 1: Branch and bound decisions (x=parent, t=child, $Cost(x,t)$=branching cost, P_x=parent probability score, L_t=child's depth, L^{limit}=depth limit): (a) The branching is accepted as the branching cost is less than the parent's probability and the child's depth is less than the depth limit; (b) The branching is rejected as the branching cost is greater than the parent's probability; (c) The branching is rejected as the child's depth is greater than the depth limit; (d) The branching is rejected as the branching cost is greater than the parent's probability and the child's depth is greater than the depth limit.

Algorithm 1: x-rooted affinity tree generation

Input: Root node, x; graph, $G(V,E)$; depth limit, L^{limit}; Node Queue, Q ;
Output: x-rooted tree;
Initialization: Set, $x.Status$="generated", $x.P = 1, x.Pop = 1, x.D = 0, x.Parent = \emptyset, front = 1, rear = 1$;
Insert x into the queue, $Q[rear] = \{x\}$;
while $rear \geq front$ **do**
  $t = Q[front]$;
  Set, $t.Status$="visited";
  if *t.P > 0.0 & t.L* $< L^{limit}$ **then**
    Extract neighbour set, $\Gamma t = \{g_1, g_2, g_3, ...\}$;
    foreach $g \in \Gamma t$ **do**
      Compute cost, $Cost(t,g)$ using (6) ;
      Find the index of g in Q, $ind = Q.Indexof(g)$;
      if $ind < 0$ **then**
        if $node.Pro > Cost(node, g)$ **then**
          Set, *g.Status*='generated', *g.Pop=1, g.L=t.L+1, g.P=t.P*-Cost(*t,g*), *g.D=* $|\Gamma g|$, *g.Parent=t* ;
          Increment rear and insert g into queue, *rear=rear+1, Q[rear]=g;*
        end
      else
        Set, *g.Pop=g.Pop+1*;
        if $g.Status$*="generated"* **AND** $g.P < (t.P - Cost(t,g))$ **then**
          Set, *g.P=t.P-Cost(t,g), g.L=t.L+1, g.Parent=t* ;
        end
      end
    end
  end
  Increment front, *front=front+1*;
end

B. Similarity score computation

The affinity tree generation step generates two affinity trees for two end nodes of the link $e_{x,y}$. Suppose, Γx and Γy are the sets of all nodes in the x-rooted and y-rooted trees, respectively. The first structural similarity between two trees is defined based on their common neighbours in (8).

$$S^1_{xy} = \frac{|\Gamma x \cap \Gamma y|}{|\Gamma x \cup \Gamma y|} \tag{8}$$

The second structural similarity score is generated as ratio of total nodes in two trees and total nodes in the original graph as (9).

$$S^2_{xy} = \frac{|\Gamma x| + |\Gamma y|}{2 \times |V|} \tag{9}$$

The similarity of two trees with respect to popularity and depth is computed based on the common neighbour set, $\Gamma x \cap \Gamma y$. Suppose that Pop^x_z, L^x_z and Pop^y_z, L^y_z are the popularity and depth of the common neighbour, $z \in \Gamma x \cap \Gamma y$ in x-rooted and y-rooted tree respectively. The popularity of a common neighbour, z is normalized by its degree, and the similarity score with respect to its popularity is defined in Equation 10. The depth of the neighbour, z is normalized by the depth limit (L^{limit}) and the similarity function is defined as (11).

$$S^3_{xy} = \frac{1}{|\Gamma x \cap \Gamma y|} \sum_{z \in \Gamma x \cap \Gamma y} \frac{Pop^x_z + Pop^y_z}{2 \times D_z} \tag{10}$$

$$S^4_{xy} = \frac{1}{|\Gamma x \cap \Gamma y|} \sum_{z \in \Gamma x \cap \Gamma y} \frac{L^x_z + L^y_z}{2 \times L^{limit}} \tag{11}$$

The structural similarity between x-rooted and y-rooted trees is then defined based on similarity scores in ((8)-(11)) as the following (12).

$$S^{ST}_{xy} = \sum_{i=1}^{4} \alpha_i \times S^i_{xy} \tag{12}$$

The parameter α_i describes the contribution of each structural information S^i_{xy} on the link formation between nodes x and y where $1 \leq \alpha_i \leq 0$ and $\sum_i^4 \alpha_i = 1$.

On the other hand, the attribute similarity, S^{AT}_{xy} between node x and y is computed using (4). Thus, the similarity score between nodes x and y is defined based on the structural and attribute similarity as follows.

$$S(x,y) = \beta \times S^{AT}_{xy} + (1-\beta) S^{ST}_{xy} \tag{13}$$

The parameter $0 \leq \beta \leq 1$ gives the flexibility to set more contribution of structural or attribute information in the final similarity score, $S(x,y)$. For the data set with no node attributes, we set $\beta = 0$.

III. Experiments

We evaluate and describe performance of the proposed LPST approach on both attributed and non-attributed graphs. In Section III-A, we describe redour experimental graph

datasets and procedures for preparing training and test datasets. We define prediction performance metrics in Section III-B and discuss experiment results in Section III-C.

A. The Datasets

We evaluate the proposed LPST link prediction approach on five non-attributed and five attributed real-world benchmark graphs from different areas. The experimental graphs are homogeneous and undirected. The non-attributed graphs include Celegans [27], Hamster [28], Ecoli [29], PB [30], Power [31], Router [32] and Yeast [33] in which the node attributes are not available. Celegans is a metabolic graph of neurons and synapses in the C. elegans worm. Hamster is a social graph in which users are represented by nodes and user friendships and family ties are represented by links. Ecoli is a biological graph representing biological links between operons in E.Coli bacteria. The PB (Political Blog) graph represents the links between political blog pages within the United States. Power represents an electrical grid connectivity among power transmitters (generators, transformers, substations) in western US. The Router graph represents the links between routers in the internet. All experimental graphs are homogeneous. The attributed graphs include AdjNoun [34], Epinions [35], Football [36] Movie-Tweeting [37], Political-Books [38], UK-Faculty [39] and USAir [2]. AdjNoun is an undirected graph representing co-occurring common nouns and adjectives in the novel *David Copperfield* by Charles Dickens. Each node represents a labeled word (noun/adjective), with edges linking adjacent words in the text. Epinions is a general consumer review graph where nodes represent products and links represent connectives between products which were co-rated by consumers. Each node is attributed with its product's category. The Football graph is a network of American football matches during the Fall 2000 where nodes represent Division IA college teams, and edges connecting teams that competed against each other. The node attribute indicates to the conference to which each team belongs, with 10 different conferences. Movie-Tweetings is a movie rating graph where nodes represent movies and two nodes are connected if they are co-rated by the same user. The genre of each movie is considered as its attribute. Political-Books is another attributed graph where nodes represent books on US politics available on Amazon.com, and links represent co-purchases of books by the same customer. The type of each book (liberal, neutral, or conservative) is considered as the node attribute. UK-Faculty is a friendship network within a UK university faculty, with nodes representing individuals, edges representing friendships, and school affiliations serving as node attributes. The attributed USAir graph represents the transportation network of US airports, where airports are represented as nodes connected by edges representing direct flight between airports. Each airport node is attributed with its location. In addition, we generate two syntactic graphs, SynA and SynNA. SynA is an attributed graph, where nodes have both numerical and categorical attributes whereas SynNA is a simple non-attributed graph. These two syntactic graphs are used only to investigate the effect of graph sparsification on link prediction approaches. Table I summarizes the characteristics of the graphs. For evaluating the prediction performance of LPST, we follow a 10-cross validation protocol. The existent link set E of a graph $G(V, E)$ is split into 10 groups. In each iteration, one group is taken as the positive test set, and the remaining 9 groups form the training set. A set of non-existent links, E', is also sampled randomly such that $\forall_{e_{xy} \in E'} e_{x,y} \notin E$ and $|E| = |E'|$. The set E' is also split into 10 groups and one group is taken as the negative test set. Each test set consists of one positive and one negative test set. As a result, 10 test and 10 training sets are prepared for each graph.

B. Performance Metrics

Precision measures the proportion of missing links that are correctly predicted as positive link ([40], [41], [42]). The computation of precision involves ranking all predicted links from a test set, containing $|E|$ positive and $|E'|$ negative links, in descending order of their similarity scores. In the case where a negative link has the same similarity score as a positive test link, the negative link is ranked lower. If L is an user defined parameter and L_r is the number of positive test links among the L top ranked (also written as top-L) predicted links, then the precision is defined as

$$Precision = \frac{L_r}{L} \tag{14}$$

A higher precision value signifies higher accuracy of the prediction approach.

On the other hand, **AUP (Area Under Precision curve)** measures the area under the precision curve ([14], [17]). The precision curve is generated by evaluating precision at ten incremental values of L (10-100% of existent links), each paired with an equal number of non-existent links from the test dataset. Since AUP takes into account the performance at different values of L, it is considered more robust and reliable.

C. Experimental Results & Discussion

In the following sections III-C1-III-C2, we summarize the experimental results of the proposed LPST approach in non-attributed and attributed graphs, comparing the results with other local, quasi-local and global link prediction approaches. In the experiments, we choose L as the size of positive test links. We also investigate the effect of increasing graph sparsity on the proposed approach in Section 4.3. We give equal emphasis to all structural features used in (12) (i.e. $\alpha 1 = \alpha 2 = \alpha 3 = \alpha 4 = 0.25$). Additionally, we assume that attribute and structural information are eqhave equal importance in defining the similarity function, and so $\beta = 0.5$ for the attributed graphs.

1) Performance in non-attributed graphs: The proposed LPST approach is evaluated on seven real-world non-attributed graphs. We select the non-attributed graphs that are the most dense (Ecoli) and most sparse (Router) to visualize the distribution of similarity scores of the top-L ranked test links computed by the proposed LPST approach in Fig. 2. From the figures, it is clearly seen that the majority of links in the top-L links are positive test links for the dense one (Fig. 2a). In other words, the average rank of positive (or existent) test links is higher than the average rank of negative (or non-existent) test links. On

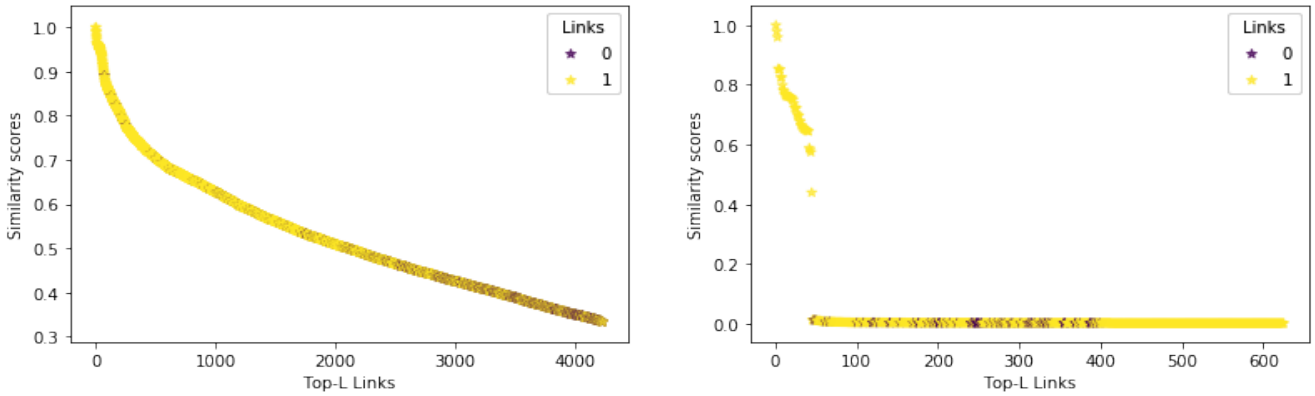
TABLE I: Topological statistics of graph datasets: number of nodes(#Node), links(#Link), average node degree (NDeg), clustering coefficient (C.Coef), average path length (APL), diameter (Diam) and attribute (– denotes no attribute), artifitially generated graphs are marked with the * mark

Graph	#Nodes	#Links	NDeg	C.Coef	APL	Diam	Attribute
Celegans	297	2148	14.465	0.308	2.455	5	–
Ecoli	1805	42325	46.898	0.350	2.714	10	–
Hamster	1576	4031	5.115	0.32	3.217	8	–
PB	1222	14407	23.579	0.239	2.787	8	–
Power	4941	6594	2.669	0.107	18.989	46	–
Router	5022	6258	2.492	0.033	6.449	15	–
Yeast	2375	11693	9.847	0.388	5.096	15	–
AdjNoun	112	425	7.589	0.19	2.53	5	Categorical
Epinions	456	957	4.197	0.927	3.239	7	Categorical
Football	115	613	10.66	0.403	2.508	4	Categorical
MovieTweet	182	509	5.593	0.817	3.997	9	Categorical
PolBooks	105	441	8.4	0.448	3.079	7	Categorical
UKFaculty	81	577	14.247	0.574	2.098	4	Categorical
USAir	332	2126	12.807	0.749	2.738	6	Numerical
SynNA*	300	3000	20.0	0.066	2.17	3	–
SynA*	300	3000	20.0	0.922	1.18	3	Categorical, Numerical

the other hand, the similarity scores of a few positive links are high, but many positive and negative test links in top-L links have very low similarity scores for the sparse one (Fig. 2b). The possibility of generating small affinity trees in sparse graphs may cause the similarity score to be low. Based on the top-L ranked links, the precision scores of the proposed LPST approach are computed for the non-attributed graphs. Table II compares the precision scores of LPST with those of local, global and quasi-local approaches. From the table, precision improvements of 3.2%,3.2%,7.6% and 44.4% are observed in Ecoli, Hamster, PB, and Router graphs, respectively, when comparing LPST with the best performing existing approach for each graph. For other graphs, LPST is also very competitive. The extraction of multiple structural features in the proposed LPST approach likely contributes to these improvements in precision. Unfortunately, the LPST approach shows less precision scores than that of a few other global approaches in some sparse graphs like power; however, it still performs better than all quasi-local and many local approaches. This can be due to the partially reliance on common neighbourhood information (likewise local approaches) for LPST to defining the similarity function. This assumption is further strengthened by the low precision scores for the common neighbour-based local approaches. Carrying rich information by the affinity trees for the end nodes could be the most post possible reason behind showing less variance in standard deviation scores across the experimental graphs. For most attributed graphs, LPST shows significant improvements in precision scores compared to global and quasi-local approaches. LPST is able to utilize both structural and node attribute information in attributed graphs and this could be the reason for success of LPST in attributed graphs. Overall, it is observed that LPST exhibits better precision scores for both highly dense and sparse graphs. The standard deviation shows that precisions of the LPST approach do not vary much across real-world graphs.

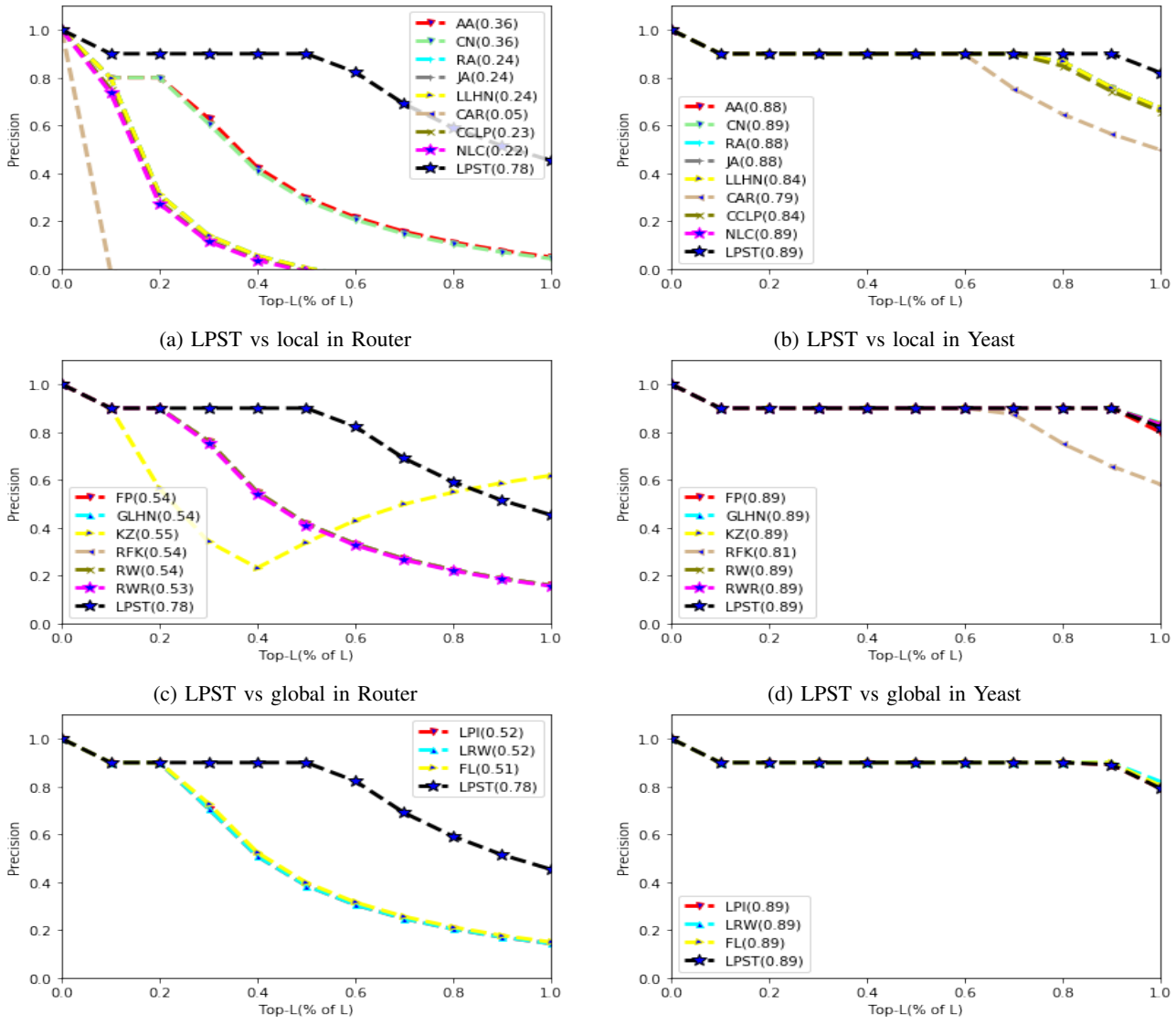
The above precision scores are computed for a single value of top-L. To describe the effect varying top-L on precisions, we plot precision scores for 10 different values of L in Fig. 3 for two randomly selected graphs: Router and Yeast. In both graphs, precision scores decrease with increasing the value of top-L, with a few exceptions. From Fig. 3a, 3c and 3c, it is evident that precisions are very high (very close to 1.0) for small values of top-L and fall with increasing top-L. In Fig. 3a, the precision scores for local approaches decrease significantly as top-L increases as these approaches highly depend on common neighborhood, which are less in sparse graphs. Similar trends are also observed for quasi-local and global approaches in Fig. 3e and 3c, respectively. However, precision scores remain still very high till top-$L = 0.5L$. Similar to other approaches, the precisions for LPST fall for the Router graph. It should be emphasized that the decrease in precisions are less significant in LPST approach when comparing to other approaches. For the Yeast graph, the precision scores for all approaches including LPST are quite high, even for larger values of top-L. Yeast is a comparatively dense graph, and the high number of common neighbors could be the most possible reason behind the success of existing approaches in this graph. Notably, the precision scores for the proposed LPST approach in the Yeast graph are quite stable and consistently higher than those of the local approaches.

The performance metric AUP for the non-attributed graphs is computed from the precision curves. The mean AUP scores with their standard deviations are presented in Table III. The success in precision table (Table II) holds in AUPs. From the table, the highest AUP scores are observed for the LPST approach in most cases. The most remarkable improvement in AUP is found for the Router graph, where LPST improves the AUP by nearly 28% compared to best performing existing approach(LRW). Another important point is that the highest AUP score for the Power graph is found for LPST, though the precision score was much lower than the highest precision in Table II. LPST is able to extract rich information from sparse graphs than many similarity-based approaches and



(a) Ecoli (b) Router

Fig. 2: Similarity score distribution for the most dense (Ecoli) and most sparse (Router) non-attributed graphs where yellow and khaki colored points represent the positive(=1) and negative(=0) test links respectively.



(a) LPST vs local in Router (b) LPST vs local in Yeast

(c) LPST vs global in Router (d) LPST vs global in Yeast

(e) LPST vs quasi-local in Router (f) LPST vs quasi-local in Yeast

Fig. 3: Precision curve for two non-attributed graphs Router and Yeast: Horizontal axis represents the values of top-L as the percentage of links in positive test set (e.g. top-$L = 0.1L, 0.2L, ..., L$ where L=number of positive test links) and vertical axis represents the precision scores for each top-L. Fig. 3a, 3c, 3e show the precision curves for local, global and quasi-local approaches respectively in Router graph for a single fold. And Fig. 3b, 3d, 3f show the precision curves for local, global and quasi-local approaches respectively in Yeast graph for a single fold. The number in () in each cell represents area under the precision curve.

TABLE II: Precision scores for the non-attributed graphs: Graph-wise highest precision scores are marked in bold. The numbers in the parentheses indicate standard deviations (e.g. the cell value 0.89(48) means the precision is 0.89 and the standard deviation is 48×10^{-3}). The superscript L, G and Q stands for local, global and quasi-local approaches respectively.

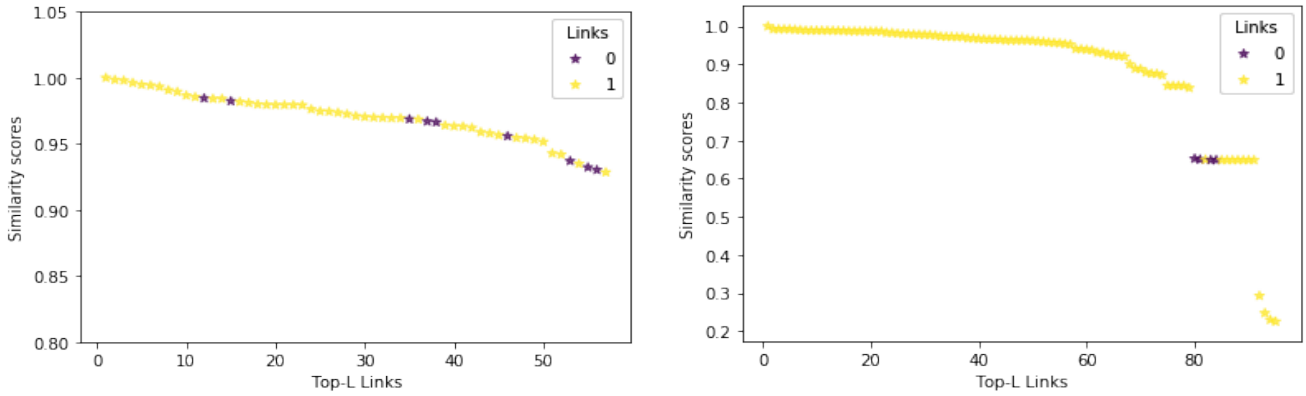
Approach	Celegan	Ecoli	PB	Power	Router	Hamster	Yeast
AA^L	**0.89**(48)	0.94(02)	0.92(07)	0.18(13)	0.12(13)	0.77(13)	0.79(17)
CAR^L	0.41(25)	0.86(04)	0.77(10)	0.01(03)	0.02(06)	0.34(18)	0.61(16)
$CCLP^L$	**0.89**(51)	0.91(18)	0.92(06)	0.09(07)	0.11(12)	0.76(14)	0.78(07)
CN^L	0.88(66)	0.94(13)	0.91(09)	0.18(13)	0.12(13)	0.77(13)	0.79(17)
JA^L	0.87(73)	0.93(20)	0.90(52)	0.18(13)	0.12(13)	0.77(13)	0.79(17)
$LLHN^L$	0.86(84)	0.93(02)	0.91(23)	0.18(13)	0.12(13)	0.77(13)	0.79(17)
NLC^L	**0.89**(45)	0.92(20)	0.91(07)	0.08(07)	0.11(12)	0.75(13)	0.78(20)
RA^L	0.89(48)	0.93(08)	0.92(06)	0.18(13)	0.12(13)	0.77(13)	0.79(17)
FP^G	0.82(84)	0.73(109)	0.88(78)	0.61(24)	0.26(20)	0.89(89)	0.91(39)
$GLHN^G$	0.79(103)	0.74(114)	0.88(64)	0.59(31)	0.26(38)	0.93(13)	0.91(15)
KZ^G	0.64(103)	0.68(126)	0.71(159)	0.46(27)	0.36(23)	0.93(14)	**0.92**(13)
RFK^G	0.83(86)	0.75(127)	0.89(53)	**0.66**(21)	0.26(18)	0.79(92)	0.75(138)
RW^G	0.81(93)	0.73(115)	0.86(87)	0.61(28)	0.26(31)	0.93(16)	**0.92**(14)
RWR^G	0.77(146)	0.72(118)	0.89(73)	0.35(35)	0.25(28)	0.92(29)	0.91(16)
LPI^Q	0.80(124)	0.74(100)	0.87(80)	0.28(12)	0.28(16)	0.91(12)	0.90(14)
LRW^Q	0.86(82)	0.73(99)	0.87(84)	0.28(12)	0.28(16)	0.91(11)	0.90(13)
FL^Q	0.76(25)	0.79(15)	0.88(16)	0.26(10)	0.19(18)	0.92(17)	**0.92**(15)
$LPST$	**0.89**(51)	**0.97**(21)	**0.99**(09)	0.30(23)	**0.52**(27)	**0.96**(19)	0.89(42)

this could contribute behind the success of showing better AUP scores for LPST in sparse graphs. For the remaining graphs, LPST remain competitive when compared to other approaches. Overall, LPST performs better in terms of AUP scores on either highly dense or highly sparse graphs. The standard deviations for LPST are lower than those of the competing approaches in all the graphs except for the Power graph. These indicate the stability of LPST approach in prediction performance in terms of AUP.

2) Performance in attributed graphs: LPST effectively utilizes available node attribute information to improve prediction performance. The approach is evaluated against seven real-world attributed graphs. Similar to the experiments with non-attributed graphs, the similarity scores of test links are computed and ranked according to their similarity scores in decreasing order. A negative link has a higher ranking when it has the same similarity score as a positive test link. We select the attributed graphs that are the most dense (UK-Faculty) and most sparse (Epinions) to visualize the distribution of similarity scores of the top-L ranked test links computed by the proposed LPST approach in Fig. 4. From the figures, it is clearly seen that the majority of links in the top-L links are positive test links for both dense and sparse graphs. In other words, the average rank of positive (or existent) test links is higher than the average rank of negative (or non-existent) test links. The possibility of generating big affinity trees in sparse graphs may cause the similarity score to be low. The average node degree difference between the most dense and most sparse non-attributed graphs is 44.406, compared to just 10.050 for attributed graphs. Hence, a nearly similar visualization is expected for the most dense (UK-Faculty) and most sparse (Epinions) attributed graphs, unlike the more distinguishable difference observed in non-attributed graphs. Based on the top-L ranked test links, the precision scores of the LPST approach for seven attributed graphs are tabulated in Table IV along with those of other local, global and quasi-local approaches. From the table, local approaches shows very high precision in the USAir and UK-Faculty graphs. LPST further improves the precision slightly, by 2-3%, utilizing node attributes. A notable improvement of about 9% in precision is observed for LPST in the Epinions graph. LPST outperforms in the MovieTweet graph by 4.5%, UK-Faculty by 2.1%, and USAir by 3.2% compared to the best existing approach. For other graphs, no single approach wins always. Overall, the proposed LPST approach shows significant improvements in precision compared to global and quasi-local approaches for most of the attributed graphs in the experiment. In contrast to non-attributed graphs, LPST shows high precision scores in most attributed graphs by utilizing the node attributes. To visualize the robustness of the approaches in terms of precision, we plot the precision curves for two randomly selected attributed graphs (AdjNoun, USAir), similar to Fig. 3. For the AdjNoun graph, the precision scores begin to decrease in the second half. The overall AUP score of LPST in this case is higher than some local approaches and comparable to some quasi-local and global approaches. In case of USAir graph, the curve is nearly flat and the overall AUP score in this case is higher than that of existing approaches. This is due to the fact that most of the positive test links have higher similarity scores than the negative test links. We observe an exceptional behaviour in the RFK global approach, where the precision scores normally fall with increasing top-L but rise again for larger values of top-L. We compute the AUPs for the attributed graphs from their precision curves. We record the mean and standard deviation of AUPs for all graphs in Table V. From the table, it is seen that the AUP scores of the proposed LPST approach outperform those of the existing approaches in most attributed graphs. A significant improvement in the AUP scores of LPST is seen for the Epinion and USAir graphs. For the remaining graphs, the AUPs for LPST remain competitive when compared to other approaches.

TABLE III: AUPs for the non-attributed graphs: Graph-wise highest AUPs are marked in bold. The AUP scores are interpreted in the similar way to Table II.The superscript L, G and Q stands for local, global and quasi-local approaches respectively.

Approach	Celegans	Ecoli	PB	Power	Router	Hamster	Yeast
AA^L	**0.88**(20)	0.87(03)	0.90(07)	0.38(25)	0.28(24)	0.87(04)	0.88(04)
CAR^L	0.67(24)	0.86(01)	0.88(05)	0.02(06)	0.06(15)	0.60(19)	0.81(08)
$CCLP^L$	**0.88**(20)	0.87(01)	0.89(07)	0.22(17)	0.26(23)	0.87(04)	0.86(05)
CN^L	0.87(27)	0.87(06)	0.89(08)	0.38(25)	0.28(24)	0.87(04)	0.88(04)
JA^L	0.87(38)	0.87(02)	0.89(10)	0.38(25)	0.28(24)	0.87(05)	0.88(04)
$LLHN^L$	0.84(81)	0.85(12)	0.90(02)	0.38(25)	0.28(24)	0.87(08)	0.85(05)
NLC^L	**0.88**(14)	0.87(01)	0.89(08)	0.18(16)	0.26(24)	0.87(04)	0.87(06)
RA^L	**0.88**(16)	0.87(03)	0.90(05)	0.38(25)	0.28(24)	0.87(04)	0.88(05)
FP^G	0.83(67)	0.84(37)	0.89(12)	0.41(20)	0.51(23)	0.87(37)	0.89(15)
$GLHN^L$	0.83(76)	0.84(48)	0.89(08)	0.40(17)	0.51(19)	**0.90**(02)	**0.90**(02)
KZ^G	0.75(107)	0.76(104)	0.78(105)	0.45(27)	0.52(13)	**0.90**(03)	0.89(03)
RFK^G	0.84(66)	0.85(37)	0.89(06)	0.42(18)	0.51(32)	0.81(36)	0.85(42)
RW^G	0.83(71)	0.83(45)	0.88(28)	0.41(31)	0.51(21)	**0.90**(03)	**0.90**(02)
RWR^G	0.83(64)	0.80(74)	0.88(25)	0.31(26)	0.50(18)	0.89(10)	0.89(06)
LPI^Q	0.83(68)	0.82(61)	0.89(16)	**0.51**(19)	0.54(21)	0.89(03)	0.89(02)
LRW^Q	0.85(68)	0.82(48)	0.88(24)	**0.54**(16)	0.55(25)	0.89(02)	0.89(01)
FL^Q	0.78(21)	0.82(42)	0.80(08)	0.53(11)	0.45(24)	0.89(03)	0.89(02)
LPST	**0.88**(15)	**0.90**(01)	**0.95**(06)	**0.54**(26)	**0.76**(17)	**0.90**(05)	0.87(17)



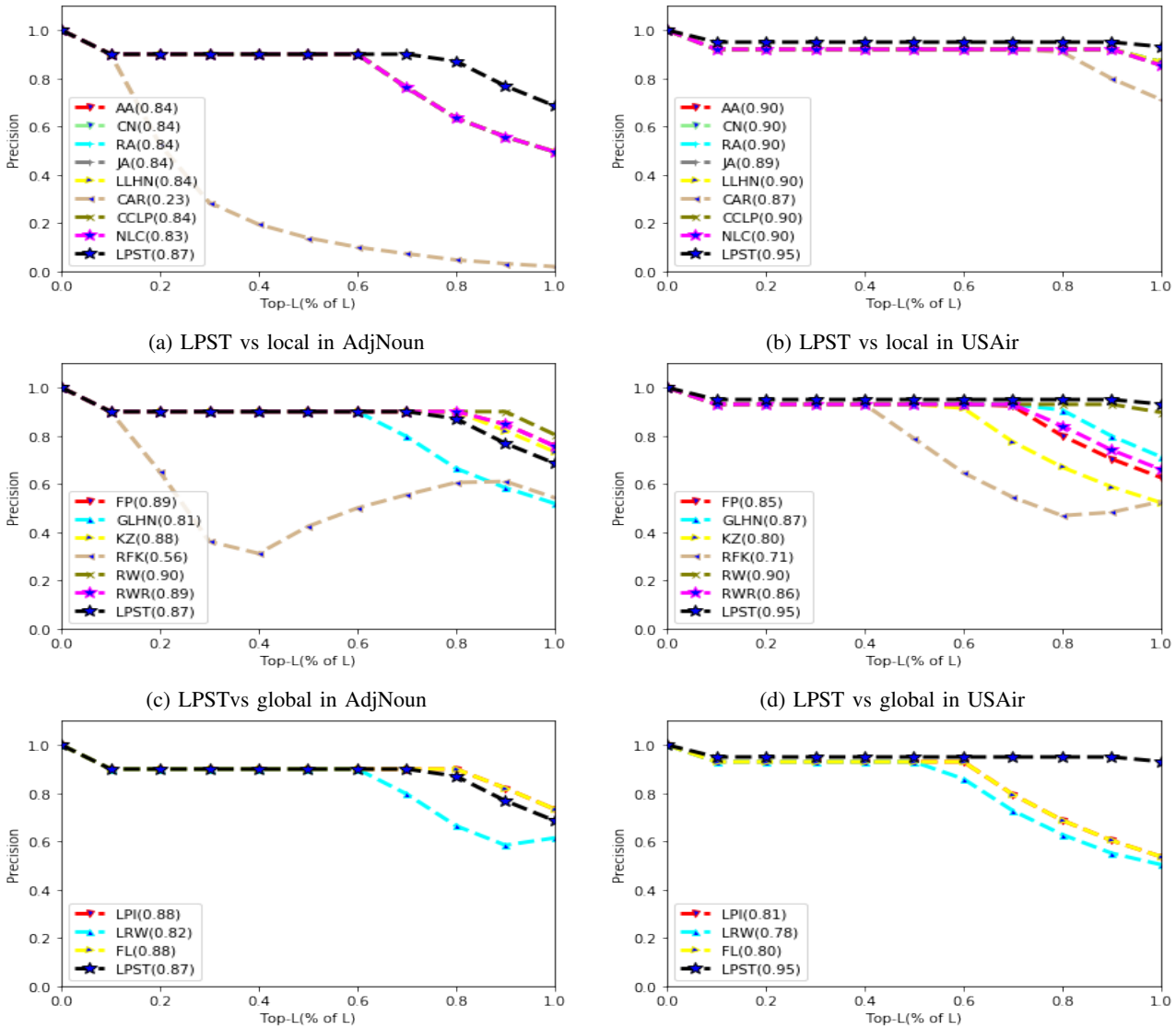
(a) UK-Faculty

(b) Epinions

Fig. 4: Similarity score distribution for the most dense (UK-Faculty) and most sparse (Epinions) attributed graphs where yellow and khaki colored points represent the positive(=1) and negative(=0) test links respectively.

To demonstrate the stability of prediction performance, we analyze the standard deviations across all real-world experimental graphs. We compute the mean and range of standard deviation for all 14 graphs. The mean and range values for each approach are plotted in Fig. 6. For Fig. 6, it is evident that local approaches generally exhibit lower mean deviations in precision scores than all global and most quasi-local approaches across the graphs. It is important to note that LPST also shows lower mean deviation, similar to local approaches. Furthermore, the ranges of deviations for LPST are lower than those of all global and quasi-local approaches and are competitive with those of local approaches. Thus, LPST can be considered more stable than other quasi-local and global approaches with respect to prediction performance. Along with extracting rich information, LPST is able to filter noise when generating affinity tress and consequently minimizing noise in simialrity score computation. This could contribute to provide stability in prediction performance of LPST in real-world graphs.

3) Effect of graph sparsification: We evaluate LPST for different levels of graph sparsity to test prediction robustness. We prepare nine additional graphs from the original SynA with different sparsity by randomly removing 10%-90% with an interval 10% of existent links. For each graph, we apply 10-fold validation to compute the mean AUP, as described earlier. Similarly, we prepare nine additional graphs from SynNA and compute the mean AUP for different sparsity levels. We plot the AUPs of the approaches for different node degrees in Fig. 7. Overall, all approaches show a decline in performance as the average node degree decreases. From Fig. 7a and 7b, it can be seen that the LPST approach is less affected than local approaches when increasing sparsity in both attributed SynA and non-attributed SynNA graphs. This robustness may be attributed to LPST's ability to extract multiple features from the generated affinity trees, which contributes to better AUPs than local approaches. When compared to quasi-local approaches, the effect of increasing the sparsity on LPST is nearly similar. On the other hand,



(a) LPST vs local in AdjNoun

(b) LPST vs local in USAir

(c) LPSTvs global in AdjNoun

(d) LPST vs global in USAir

(e) LPST vs quasi-local in AdjNoun

(f) LPST vs quasi-local in USAir

Fig. 5: Precision curve for two randomly sampled attributed graphs AdjNoun and USAir: Horizontal axis represents the values of top-L as the percentage of links in positive test set (e.g. top-L=10%-100% with an interval of 10% of total test links) and vertical axis represents the precision. Fig. 5a, 5e, 5c show the precision curves for local, quasi-local and global approaches respectively in AdjNoun graph for a single fold. And Fig. 5b, 5f, 5d show the precision curves for local, quasi-local and global approaches respectively in USAir graph for a single fold. The number in () in each approach represents area under precision (AUP) curve.

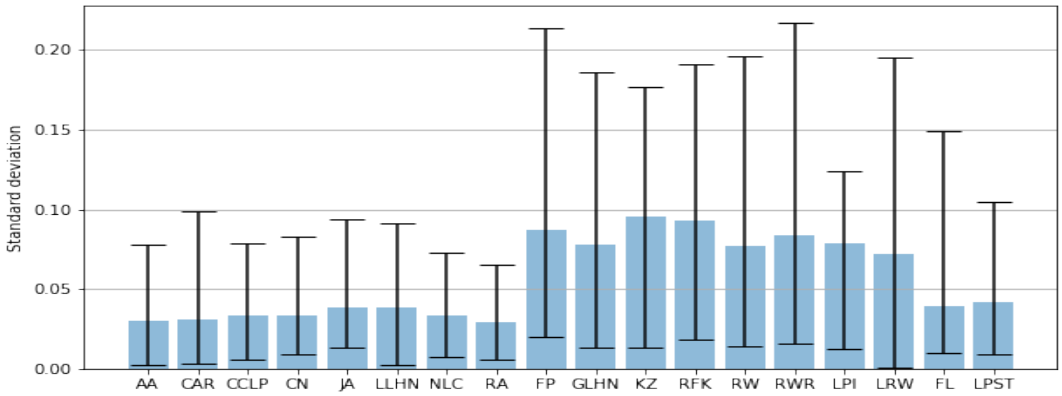


Fig. 6: Precision stability: Thick bar (sky) shows the mean deviation, and thin bar (black) shows the range of the deviations for each approach over the real-world 14 graphs.

TABLE IV: Precision for the attributed graphs: Graph-wise highest precision scores are marked in bold. This table is interpreted in the similar way to Table II. The superscript L, G and Q stands for local, global and quasi-local approaches, respectively.

Approach	AdjNoun	Epinion	Football	MovieTweet	Pol.Book	UKFaculty	USAir
AA^L	0.65(78)	0.89(33)	0.81(40)	0.89(49)	0.91(34)	0.95(55)	0.95(15)
CAR^L	0.10(36)	0.70(40)	0.57(44)	0.65(99)	0.46(48)	0.83(66)	0.82(18)
$CCLP^L$	0.67(79)	0.81(42)	0.81(51)	0.84(67)	0.91(37)	0.94(55)	0.94(18)
CN^L	0.66(83)	0.89(33)	0.82(44)	0.89(49)	0.92(32)	0.93(50)	0.94(26)
JA^L	0.66(94)	0.89(33)	0.82(41)	0.89(49)	0.92(33)	0.95(48)	0.94(33)
$LLHN^L$	0.65(91)	0.89(33)	0.82(44)	0.89(49)	0.92(33)	0.94(73)	0.94(43)
NLC^L	0.66(73)	0.79(40)	0.80(46)	0.82(68)	0.90(43)	0.95(56)	0.94(17)
RA^L	0.65(65)	0.89(33)	0.81(40)	0.89(49)	0.91(34)	0.95(54)	0.95(15)
FP^G	**0.79**(143)	0.47(49)	0.67(91)	0.55(66)	0.84(87)	0.76(214)	0.76(133)
$GLHN^G$	0.72(149)	0.47(49)	0.82(45)	0.54(75)	0.89(63)	0.79(186)	0.75(147)
KZ^G	0.75(135)	0.67(153)	0.83(36)	0.72(112)	0.73(177)	0.64(95)	0.75(162)
RFK^G	0.72(151)	0.47(49)	0.76(88)	0.55(71)	0.83(77)	0.80(191)	0.82(141)
RW^G	0.76(162)	0.47(49)	0.83(34)	0.55(66)	0.85(80)	0.77(196)	0.80(106)
RWR^G	0.65(134)	0.46(48)	0.84(41)	0.37(72)	0.86(82)	0.75(217)	0.77(141)
LPI^Q	0.73(143)	0.46(51)	0.88(57)	0.25(71)	0.84(69)	0.74(211)	0.73(144)
LRW^Q	0.75(134)	0.46(51)	**0.89**(55)	0.25(71)	0.86(68)	0.75(195)	0.75(120)
FL^Q	0.76(149)	0.79(33)	**0.89**(60)	0.80(42)	0.87(52)	0.84(67)	0.85(20)
LPST^Q	0.65(105)	**0.97**(26)	0.80(51)	**0.93**(82)	**0.93**(76)	**0.97**(35)	**0.98**(18)

TABLE V: AUPs for the attributed graphs: Graph-wise highest AUP scores are marked in bold fonts. The table is interpreted in the similar way to Table II. The superscript L, G and Q stands for local, global and quasi-local approaches, respectively.

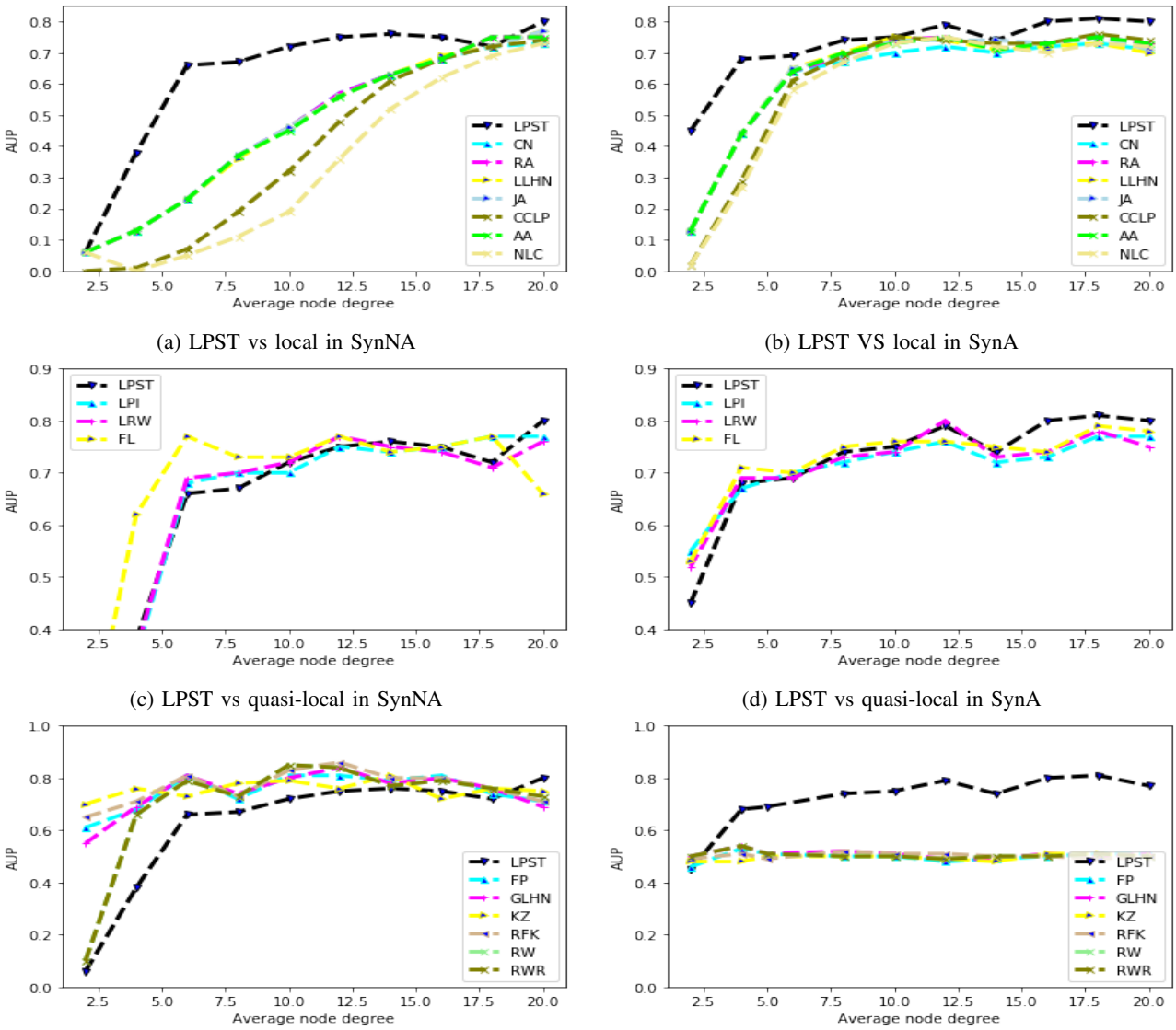
Approach	AdjNoun	Epinion	Football	MovieTweet	Pol.Book	UKFaculty	USAir
AA^L	0.78(72)	0.89(04)	0.88(11)	0.89(05)	0.89(05)	0.89(11)	0.90(03)
CAR^L	0.23(77)	0.85(14)	0.79(26)	0.83(37)	0.71(37)	0.88(17)	0.88(04)
$CCLP^L$	0.80(49)	0.88(09)	0.87(12)	0.88(11)	0.89(05)	0.89(12)	0.90(04)
CN^L	0.75(91)	0.89(04)	0.88(08)	0.89(04)	0.89(06)	0.89(15)	0.90(03)
JA^L	0.77(71)	0.89(04)	0.88(09)	0.89(05)	0.90(03)	0.89(11)	0.89(09)
$LLHN^L$	0.75(99)	0.89(04)	0.88(09)	0.89(05)	0.90(03)	0.88(38)	0.87(64)
NLC^L	0.82(31)	0.88(09)	0.85(19)	0.88(12)	0.89(06)	0.89(13)	0.90(04)
RA^L	0.78(72)	0.89(04)	0.88(11)	0.89(05)	0.89(07)	0.89(16)	0.90(03)
FP^G	0.82(105)	0.72(40)	0.76(89)	0.75(71)	0.86(48)	0.77(159)	0.83(73)
$GLHN^G$	0.82(65)	0.72(40)	0.88(16)	0.75(78)	0.87(37)	0.80(160)	0.82(83)
KZ^G	0.83(66)	0.51(182)	**0.89**(11)	0.70(123)	0.75(114)	0.77(98)	0.73(121)
RFK^G	0.73(181)	0.72(40)	0.79(91)	0.76(59)	0.86(44)	0.80(154)	0.84(76)
RW^G	**0.84**(56)	0.72(40)	**0.89**(09)	0.75(68)	0.86(41)	0.77(153)	0.82(90)
RWR^G	0.77(95)	0.72(40)	**0.89**(10)	0.62(73)	0.86(49)	0.80(115)	0.83(70)
LPI^Q	0.82(67)	0.71(43)	**0.89**(09)	0.47(85)	0.86(46)	0.80(147)	0.82(07)
LRW^Q	0.85(42)	0.71(43)	**0.89**(09)	0.47(85)	0.86(53)	0.79(135)	0.80(09)
FL^Q	0.83(68)	0.80(33)	**0.89**(09)	0.79(05)	0.80(05)	0.79(16)	0.80(03)
LPST^Q	0.79(97)	**0.95**(01)	0.87(17)	**0.90**(08)	**0.92**(11)	**0.90**(06)	**0.94**(04)

different behaviours are observed for non-attributed and attributed graphs in Fig. 7e and 7d. For SynNA, LPST is more affected by sparsity than global approaches, whereas the trend reverses for the attributed graph, SynA. Overall, LPST approach performs better with respect to AUPs for the attributed graph, SynA, than for the non-attributed graph, SynNA. The utilization of node attribute information along with the structural information could be a potential reason behind these trends.

IV. Conclusion

This paper introduces a novel quasi-local similarity-based approach, LPST, for link prediction in homogeneous graphs. Unlike state-of-the-art similarity-based approaches, LPST leverages both structural and node attribute information in its prediction task. Moreover, LPST distinguishes and discards noisy neighbors when defining the node similarity function. When compared to aligned state-of-the-art similarity-based approaches, LPST demonstrates competitive prediction performance on both attributed and non-attributed graphs. In addition, LPST is less sensitive to graph sparsification than other approaches. Although the computational time of LPST are not directly compared, LPST can be assumed to be slower than the local approaches, but faster than global approaches.

The future extension of this work could be extracting additional useful features from the generated affinity trees to



(a) LPST vs local in SynNA

(b) LPST VS local in SynA

(c) LPST vs quasi-local in SynNA

(d) LPST vs quasi-local in SynA

(e) LPST vs global in SynNA

(f) LPST vs global in SynA

Fig. 7: AUP for different levels of sparsity for the syntactic graphs, SynA and SynNA. Different amount of links are removed randomly from SynA and SynNA to generate graphs with desired average node degree.

further improve prediction performance. In addition, LPST could be applied to very large-scale graphs in a parallel environment to investigate its computational efficiency and prediction performance.

REFERENCES

[1] O. Alshareet and A. Awasthi, "Enhancing e-commerce recommendations with a novel scale-aware spectral graph wavelets framework," *International Journal of Data Science and Analytics*, vol. 19, no. 2, pp. 243–256, 2025.

[2] M. K. Islam, S. Aridhi, and M. Smail-Tabbone, "Appraisal study of similarity-based and embedding-based link prediction methods on graphs," in *Proceedings of the 10th International Conference on Data Mining & Knowledge Management Process*, 2021, pp. 81–92.

[3] X. Li, L. Sun, M. Ling, and Y. Peng, "A survey of graph neural network based recommendation in social networks," *Neurocomputing*, vol. 549, p. 126441, 2023.

[4] M. K. Islam, S. Aridhi, and M. Smail-Tabbone, "From competition to collaboration: Ensembling similarity-based heuristics for supervised link prediction in biological graphs," in *International Conference on Bangabandhu and Digital Bangladesh*. Springer, 2021, pp. 121–135.

[5] N. M. Cong, H.-C. Liu, V. R. Mekala, E. Zaenudin, E. B. Wijaya, and K.-L. Ng, "Identify gene-gene regulatory modules for patients with renal clear cell tumor metastasis," in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering*, vol. 2245, 2023, pp. 31–35.

[6] Y. Wu, Y. Jin, P. Sun, and Z. Ding, "Hbdta: Hierarchical bi-lstm networks for drug-target binding affinity prediction." *Engineering Letters*, vol. 32, no. 2, pp. 284–295, 2024.

[7] J. Luo, D. Yang, and Y. Liu, "Multi-length meta-path sematic fusion in medical heterogeneous graph for disease dignosis." *Engineering Letters*, vol. 32, no. 11, pp. 2063–2072, 2024.

[8] Y. Sevim, G. K. Orman, and O. Kılıçlıoglu, "A link prediction framework for hotel recommendations," in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering*, 2022, pp. 64–69.

[9] V. Martinez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," *ACM Computing Surveys (CSUR)*, vol. 49, no. 4, pp. 1–33, 2016.

[10] F. Lorrain and H. C. White, "Structural equivalence of individuals in social networks," *The Journal of Mathematical Sociology*, vol. 1, no. 1, pp. 49–80, 1971.

[11] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.

[12] T. Zhou, L. Lu, and Y.-C. Zhang, "Predicting missing links via local information," *The European Physical Journal B*, vol. 71, no. 4, pp. 623–630, 2009.

[13] P. Jaccard, "Etude comparative de la distribution florale dans une portion des alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

[14] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi, "From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks," *Scientific Reports*, vol. 3, no. 1, pp. 1–14, 2013.

[15] E. A. Leicht, P. Holme, and M. E. Newman, "Vertex similarity in networks," *Physical Review E*, vol. 73, no. 2, p. 026120, 2006.

[16] Z. Wu, Y. Lin, J. Wang, and S. Gregory, "Link prediction with node clustering coefficient," *Physica A: Statistical Mechanics and its Applications*, vol. 452, pp. 1–8, 2016.

[17] Z. Wu, Y. Lin, H. Wan, and W. Jamil, "Predicting top-l missing links with node and link clustering information in large-scale networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2016, no. 8, p. 083202, 2016.

[18] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[19] O. Vanunu and R. Sharan, "A propagation-based algorithm for inferring gene-disease associations," in *German Conference on Bioinformatics*. Gesellschaft Für Informatik e. V., 2008, pp. 54–63.

[20] K. Pearson, "The problem of the random walk," *Nature*, vol. 72, no. 1867, pp. 342–342, 1905.

[21] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *Sixth International Conference on Data Mining (ICDM'06)*. IEEE, 2006, pp. 613–622.

[22] L. Lu, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," *Physical Review E*, vol. 80, no. 4, p. 046122, 2009.

[23] W. Liu and L. Lu, "Link prediction based on local random walk," *EPL (Europhysics Letters)*, vol. 89, no. 5, p. 58007, 2010.

[24] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos, "Fast and accurate link prediction in social networking systems," *Journal of Systems and Software*, vol. 85, no. 9, pp. 2119–2132, 2012.

[25] E. Nasiri, K. Berahmand, and Y. Li, "Robust graph regularization nonnegative matrix factorization for link prediction in attributed networks," *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3745–3768, 2023.

[26] J. Wu and B. Hooi, "Decor: Degree-corrected social graph refinement for fake news detection," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 2582–2593.

[27] F. Piano, K. C. Gunsalus, D. E. Hill, and M. Vidal, "C. elegans network biology: a beginning," *WormBook*, vol. 2006, pp. 1–20, 2006.

[28] J. Kunegis, "Konect: the koblenz network collection," in *Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 1343–1350.

[29] H. Salgado, A. Santos-Zavaleta, S. Gama-Castro, D. Millan-Zarate, E. Diaz-Peredo, F. Sanchez-Solano, E. Perez-Rueda, C. Bonavides-Martinez, and J. Collado-Vides, "Regulondb (version 3.2): transcriptional regulation and operon organization in escherichia coli k-12," *Nucleic Acids Research*, vol. 29, no. 1, pp. 72–74, 2001.

[30] R. Ackland *et al.*, "Mapping the us political blogosphere: Are conservative bloggers more prominent?" in *BlogTalk Downunder 2005 Conference, Sydney*. BlogTalk Downunder 2005 Conference, Sydney, 2005, pp. 1–12.

[31] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[32] N. Spring, R. Mahajan, and D. Wetherall, "Measuring isp topologies with rocketfuel," *ACM SIGCOMM Computer Communication Review*, vol. 32, no. 4, pp. 133–145, 2002.

[33] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein–protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[34] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 3, p. 036104, 2006.

[35] J. Tang, H. Gao, H. Liu, and A. Das Sarma, "etrust: Understanding trust evolution in an online world," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 253–261.

[36] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[37] S. Dooms, T. De Pessemier, and L. Martens, "Movietweetings: a movie rating dataset collected from twitter," in *Workshop on Crowdsourcing and Human Computation for Recommender Systems, CrowdRec at RecSys*, vol. 2013, 2013, p. 43.

[38] A. Reihanian, M.-R. Feizi-Derakhshi, and H. S. Aghdasi, "Community detection in social networks with node attributes based on multi-objective biogeography based optimization," *Engineering Applications of Artificial Intelligence*, vol. 62, pp. 51–67, 2017.

[39] T. Nepusz, A. Petroczi, L. Negyessy, and F. Bazso, "Fuzzy communities and the concept of bridgeness in complex networks," *Physical Review E*, vol. 77, no. 1, p. 016107, 2008.

[40] L. Pan, T. Zhou, L. Lu, and C.-K. Hu, "Predicting missing links and identifying spurious links via likelihood analysis," *Scientific Reports*, vol. 6, no. 1, pp. 1–10, 2016.

[41] Z. Wu, Y. Lin, Y. Zhao, and H. Yan, "Improving local clustering based top-l link prediction methods via asymmetric link clustering information," *Physica A: Statistical Mechanics and its Applications*, vol. 492, pp. 1859–1874, 2018.

[42] J. Yang and X.-D. Zhang, "Predicting missing links in complex networks based on common neighbors and distance," *Scientific Reports*, vol. 6, p. 38208, 2016.