HM-EIICT: Fairness-aware link prediction in complex networks using community information

Akrati Saxena¹ · George Fletcher¹ · Mykola Pechenizkiy¹

Accepted: 19 July 2021 / Published online: 27 August 2021
© The Author(s) 2021

Abstract

The evolution of online social networks is highly dependent on the recommended links. Most of the existing works focus on predicting intra-community links efficiently. However, it is equally important to predict inter-community links with high accuracy for diversifying a network. In this work, we propose a link prediction method, called HM-EIICT, that considers both the similarity of nodes and their community information to predict both kinds of links, intra-community links as well as inter-community links, with higher accuracy. The proposed framework is built on the concept that the connection likelihood between two given nodes differs for inter-community and intra-community node-pairs. The performance of the proposed methods is evaluated using link prediction accuracy and network modularity reduction. The results are studied on real-world networks and show the effectiveness of the proposed method as compared to the baselines. The experiments suggest that the inter-community links can be predicted with a higher accuracy using community information extracted from the network topology, and the proposed framework outperforms several measures especially proposed for community-based link prediction. The paper is concluded with open research directions.

Keywords Link prediction · Link analysis · Similarity-based indices · Social networks

✉ Akrati Saxena
a.saxena@tue.nl

George Fletcher
g.h.l.fletcher@tue.nl

Mykola Pechenizkiy
m.pechenizkiy@tue.nl

¹ Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

1 Introduction

In a complex network, the nodes represent the objects or entities, and the edges or links between nodes denote the relationship between the respective objects. Complex networks have been highly used to study a variety of complex systems, such as biological, technological or, social systems (Onnela et al. 2007; Almaas et al. 2007; Caldarelli 2007). Researchers have been interested in studying the evolution of these complex networks, their structural properties, and several dynamic phenomena taking place over these networks, such as information diffusion and opinion formation. The methods for link prediction in complex networks also have been of great interest due to their wide applications. The methods for predicting links that might be formed in the future can be used to recommend friends in social networks, identify collaboration opportunities among researchers, identify probable missing interactions in protein networks, and so on. The link recommendation methods require a good understanding of network evolution and influence the further evolution of the network (Saxena 2020).

In real-world networks, most of the time, it is infeasible to get the meta-information of the nodes due to the privacy policies or the time-taking process of data collection (Praveena and Smys 2016). For example, in a friendship network, it is infeasible to get the characteristics of people so that their similarity can be computed that can be further used to predict future links. Therefore, the link prediction methods based on network structure information have received much attention. These heuristic methods compute the similarity score of a node-pair using the node proximity information.

Some of the recent works in link prediction also considered community information for improving the accuracy of the proposed methods (Cannistraci et al. 2013; Yan and Gregory 2012; Valverde-Rebaza and de Andrade 2012). The community structure is a well-known mesoscale structure in real-world networks (Clauset et al. 2004). The nodes belonging to the same community are densely connected with each other, and the nodes belonging to different communities have a very small number of connections (Saxena and Iyengar 2016). The intra-community links are driven by homophily, i.e., the nodes having similar interest are more probable to be connected (McPherson et al. 2001). The evolution of inter-community links is still not well-studied in the literature. However, the concepts that can explain the formation of inter-community links are triad formation (Holme and Kim 2002), bridge or weak-ties (Granovetter 1973), and the random formation of the link between two nodes due to some activity or event (Watts and Strogatz 1998). Therefore, efficiently predicting inter-community links has been an open question.

Most of the existing heuristic methods have focused on intra-community link prediction to improve the overall accuracy of their methods (Yan and Gregory 2012; Cannistraci et al. 2013). They have not explored the community structure information and network properties in-depth to improve inter-community link prediction accuracy. The inter-community links are essential to increase the diversity in the network. For example, in a social network, the opinion of a person is highly influenced by the opinion of its neighbors (Saxena et al. 2020a, b). If a person is well connected with other people of her own community having similar opinions, the person will receive the information that confirms her own belief. However, if the person is also well connected with the people having a different opinion, the person will receive the

information from different perspectives. The psychological experiments have shown that a person is more likely to believe in the correct information if the person has seen it from different perspectives (Park et al. 2009). There are several other real-life applications of increased diversity (Hofstra et al. 2017; Garimella et al. 2017; Matakos et al. 2020).

In this work, we propose a link prediction framework, called **HM-EIICT** (**H**euristic **M**ethod-**E**xtended using **I**ntra and **I**nter **C**ommunity **T**hresholds), that extends similarity-score based heuristic link prediction methods to predict both intra-community as well as inter-community links with a high accuracy. We observe that the similarity scores based on any traditional heuristic link prediction method highly vary for intra-community and inter-community links. The intra-community links have a higher similarity score for existing links than the inter-community links. Based on our observation, we propose an HM-EIICT framework that considers different threshold values of similarity score for predicting inter and intra-community links using a given heuristic method. The proposed framework is verified on real-world networks, and the results show that the proposed method highly improves the inter-community link prediction accuracy. The accuracy for intra-community link prediction is either improved or remained intact. Therefore, The HM-EIICT method shows a huge improvement in the overall accuracy. The proposed link recommendation method improves the diversity in a network by reducing the network modularity and can be used for evolving a diverse social network.

The paper is structured as follows. In Sects. 2 and 3, we discuss related work and preliminaries, respectively. In Sect. 4, we discuss the proposed framework. In Sect. 5, we study the performance of the proposed framework, including the details of the datasets and evaluation metrics. The paper is concluded with future directions in Sect. 6.

2 Related work

Link prediction methods can be mainly categorized as similarity-score based heuristic methods and machine learning based methods. The heuristic methods compute the similarity score of the given node-pair, and node-pairs having a higher similarity score are more likely to have a missing connection or build a connection in the future. The similarity score can also be computed using nodes' characteristics, and two nodes are considered more similar if they have more common properties. However, the characteristics of the nodes are not available due to privacy-related issues, and therefore, most of the similarity based methods consider the structural similarity of the nodes based on network structure.

The similarity based heuristic indices can be further categorized as (i) local indices, (ii) semi-local indices, and (iii) global indices. The local indices consider neighborhood information of the nodes, such as Jaccard coefficient (Liben-Nowell and Kleinberg 2007), Adamic Adar index (Adamic and Adar 2003), resource allocation index (Zhou et al. 2009), CCLP index (Wu et al. 2016), and Leicht-Holme-Newman Index (Leicht et al. 2006). The global similarity indices are mainly based on the shortest distance or number of paths between the given nodes (Fouss et al. 2007; Tong et al. 2006). In

semi-local similarity indices, the local paths or local information gathered using local random walk is used to compute the similarity. The well known semi-local similarity indices include Local Path Index (Lü et al. 2009), Local Random Walk (Liu and Lü 2010), Superposed Random Walk (Liu and Lü 2010), Neighbor Set Information index (Zhu and Xia 2015), and Extended resource allocation index (Liu et al. 2017).

In real-world networks, nodes are organized into communities, and there have been proposed some similarity indices that also consider the community information of the nodes to improve the link prediction accuracy. However, most of the community based indices have focused on improving the accuracy of intra-community links to improve the overall accuracy. Cannistraci et al. (Cannistraci et al. 2013) proposed the CAR index that considers both the common neighbors and local community links to compute the similarity. The WIC index computes the similarity score using within-community (W) and inter-community (IC) information of the shared neighbors where within-community neighbors contribute positively, and inter-community (IC) neighbors contribute negatively in the final score (Valverde-Rebaza and de Andrade 2012). Yan and Gregory (Yan and Gregory 2012) proposed a method based on the concept that the intra-community links are more likely to be connected than the inter-community links. Therefore, the authors precede intra-community node pairs from inter-community node pairs while computing the final ranking based on the similarity score. The proposed method is unfair for inter-community pairs and will end up reducing the diversity from the network.

Biswas and Biswas considered edge-centrality (EC) measures and community-based edge-weight (CEW) to define the importance of existing links (Biswas and Biswas 2017). The proposed method improves the intra-community link prediction by assigning positive weight to intra-community links while computing the CEW. Gao et al. (Gao et al. 2017) proposed a Community Bridge Boosting Prediction Model (CBBPM) that predicts links differently for bridge nodes by boosting their similarity score based on their structural position. Ding et al. (Ding et al. 2016) defined a method to compute the similarity between different communities and used this information to predict missing links. However, this method will assign the same likelihood value to two different intra-community pairs of nodes even if they have a diverse common neighborhood. Li et al. (Li et al. 2019) proposed a link prediction framework that computes the Community Relationship Strength (CRS) and then uses it with similarity-based local indices to compute the final likelihood for a node-pair. Some other community-based link prediction methods include (Wang et al. 2019; Singh et al. 2020; Wu et al. 2017; Jeon and Kim 2017a); however, none of them has focused on improving the inter-community link prediction accuracy.

The machine learning based methods train a model based on the properties of the nodes or edges for the existing links and use this learned model to predict the likelihood of the link for a given node-pair. These methods can be further categorized as classification-based methods (Pecli et al. 2018), probabilistic and statistical methods (Yu and Chu 2007), and matrix factorization methods (Gao et al. 2011). Another approach of link prediction methods is based on network embedding that aims to predict missing links using low dimensional feature representation of the nodes (Grover and Leskovec 2016; Saxena et al. 2021).

Recently, researchers have focused on fairness while designing network science based solutions (Rahman et al. 2019; Li et al. 2021; Spinelli et al. 2021). Masrour et al. (Masrour et al. 2020) proposed a fairness-aware method for recommending links between people belonging to the same and different genders. The proposed method used the adversarial approach to learn a low-dimensional network embedding. As per the best of our knowledge, there has not been proposed any link prediction method that considers fairness for each community and has shown results for both intra-community as well as inter-community link prediction. In this work, we propose a simple and fast heuristic method to improve the intra-community as well as inter-community link prediction accuracy.

3 Preliminaries

3.1 Notations

In Table 1, we explain the notations used in this work.

3.2 Baseline heuristic methods for link prediction

In Table 3, we discuss the formulation of similarity-score based heuristic measures that we consider in our study. The JC, AA, and RA methods only consider the proximity information of the nodes, and CACN, CARA, CRS-RA, CMS-RA, and ICRA methods consider both the node-proximity and community information for computing the similarity score of the given node-pair.

4 The proposed method: HM-EIICT

In real-world networks, nodes are organized into communities. The connections are denser among the nodes belonging to the same community and sparser between the nodes belonging to different communities (Saxena and Iyengar 2016). We first analyze the characteristics of intra-community and inter-community links on the datasets mentioned in Table 2 (refer Sect. 5.1 for further details of the used datasets). The results are shown in Table 4 for eight similarity-score based heuristic methods mentioned in

Table 1 Notations

Notation	Explanation
$G(V, E)$	A graph G where V is the set of nodes and E is the set of edges
u, v, w	Denote nodes in graph G , and $u \in V$, $v \in V$, and $w \in V$
$\Gamma(u)$	Denotes set of neighbors of node u
$\deg(u)$	Degree of node u , and $\deg(u) = \Gamma(u) $
C_u	Community label of node u

Table 2 Datasets

Networks	#Nodes	#Edges	#Communities	References
Eu-Email	986	16064	7	Leskovec et al. (2007)
Facebook	4039	88234	16	Mcauley and Leskovec (2012)
GrQc	4158	13422	42	Leskovec et al. (2007)
Hep-th	8638	24806	50	Leskovec et al. (2007)
Astro-ph	17903	196972	37	Leskovec et al. (2007)

Table 3, where we show mean, standard deviation, minimum, and maximum value of similarity scores for both intra-community and inter-community links, separately. We observe that the similarity scores of inter-community links are lower than the similarity score of intra-community links for all heuristics methods. The results clearly show that the mean similarity score has a huge difference for different kinds of links.

Based on our observation, we propose that the heuristics methods should consider different threshold values for similarity scores while predicting intra-community and inter-community links. The threshold value should be higher for intra-community links than inter-community links for all considered heuristic methods. We propose a link prediction framework that extends the baseline heuristic methods by using different threshold values for different types of links. The proposed method is referred to as **HM-EIICT (Heuristic Method-Extended using Intra and Inter Community Thresholds)**. The EIICT extension for the Jaccard Coefficient method is referred to as JC-EIICT, and it can be computed as,

$$JC - EIICT(u, v) = \begin{cases} 1, & \text{if } (C_u = C_v \text{ \& } JC(u, v) > \theta_1) \text{ or } (C_u \neq C_v \text{ \& } JC(u, v) > \theta_2) \\ 0, & \text{otherwise} \end{cases}$$

where θ_1 is the threshold value for intra-community links, and θ_2 is the threshold value for inter-community links. The other heuristic methods can similarly be extended for their EIICT version.

The value of θ_1 and θ_2 is decided based on the structural properties of the network. The simplest way is (i) compute the similarity score for existing intra-community and inter-community links, and (ii) then decide the intra-community and inter-community thresholds such that some f fraction of intra-community and inter-community links have similarity scores higher than that, respectively. f might be different for computing intra-community and inter-community thresholds.

Complexity The complexity of the proposed framework depends on two factors, (i) identifying community labels, and (ii) computing threshold values (θ_1 and θ_2). If the ground-truth community information is not available, the communities are identified using the Louvain community detection method Blondel et al. (2008) that has $O(n \cdot \log n)$ complexity. To compute the thresholds' value, a small fraction (x and $x \ll m$) of intra and inter community edges are uniformly sampled and their similarity score is used to decide the threshold value as described above. Once the communities are identified, the complexity to compute the similarity score for JC, AA, RA, CACN, CARA, CMS-RA, and ICRA method is $O(deg_{avg}^2)$, where deg_{avg} is the average

Table 3 Baseline similarity-score based heuristic methods that we have considered for the analysis

Method	Description	Formulation	Ref
JC	Jaccard Coefficient	$JC(u, v) = \frac{ F(u) \cap F(v) }{ F(u) \cup F(v) }$	(Liben-Nowell and Kleinberg, 2007)
AA	Adamic Adar Index	$AA(u, v) = \sum_{w \in CN(u, v)} \frac{1}{\log F(w)}$	(Adamic and Adar, 2003)
RA	Resource Allocation Index	$RA(u, v) = \sum_{w \in CN(u, v)} \frac{1}{F(w)}$	(Zhou et al., 2009)
CACN	Community-adaptive CN	$CACN(u, v) = \begin{cases} \lambda \cdot A(u, v) , & C_u = C_v \\ (1 - \lambda) \cdot A(u, v) , & C_u \neq C_v \end{cases}$	(Jeon and Kim, 2017b)
CARA	Community-adaptive RA method	$CARA(u, v) = \begin{cases} \lambda \cdot RA(u, v), & C_u = C_v \\ (1 - \lambda) \cdot RA(u, v), & C_u \neq C_v \end{cases}$	(Jeon and Kim, 2017b)
GRS-RA	Community Relationship Strength based RA method	$GRS_RA(u, v) = RA(u, v) \cdot CRS(C_u, C_v)$ $CRS(i, j) = \frac{ (A(i) \cup F(i)) \cap (A(j) \cup F(j)) }{\min((A(i) \cup F(i)) , (A(j) \cup F(j)))}$ where $A(i) = \{u C_u = i\}$ and $F(i) = \{v C_u = i, C_v \neq i, (u, v) \in E\}$	(Li et al., 2019)
CMS-RA	Community Membership based Similarity Score by extending RA index	$CMS - RA(u, v) = \frac{1}{\sum \Omega_{uv}^1 \frac{1}{deg(w)} + \lambda \sum \Omega_{uv}^2 \frac{1}{deg(w)}}, C_u = C_v$ $= \begin{cases} \lambda \left(\sum \Omega_{uv}^3 \frac{1}{deg(w)} + \sum \Omega_{uv}^4 \frac{1}{deg(w)} \right), & C_u \neq C_v \\ \Omega_{uv}^1 = \{w w \in A(u, v), C_u = C_v, C_w = C_u\} \\ \Omega_{uv}^2 = \{w w \in A(u, v), C_u = C_v, C_w \neq C_u\} \\ \Omega_{uv}^3 = \{w w \in A(u, v), C_u \neq C_v, C_w \in \{C_u, C_v\}\} \\ \Omega_{uv}^4 = \{w w \in A(u, v), C_u \neq C_v, C_w \notin \{C_u, C_v\}\} \end{cases}$	(Bai et al., 2019)
ICRA	Intra-Com based RA	$ICRA(u, v) = \frac{ A^{IN}(u, v) }{ A(u, v) } + \sum_{w \in A(u, v)} \frac{\max\{deg(u), deg(v)\}}{deg(w)}$ where $A^{IN}(u, v) = \{w C_w = C_u = C_v\}$	(Wang et al., 2017)

Table 4 Similarity-scores computed using heuristic baseline methods for Intra-community and Inter-community links

	Eu-Email		Facebook		GrQc		Hep-th		Astro-ph	
	Intra	Inter	Intra	Inter	Intra	Inter	Intra	Inter	Intra	Inter
JC	Mean	0.199	0.110	0.095	0.402	0.060	0.207	0.058	0.332	0.101
	Std.	0.116	0.072	0.085	0.325	0.053	0.223	0.049	0.258	0.079
	Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Max	0.923	0.439	0.926	0.576	0.957	0.444	0.939	0.984	0.721
AA	Mean	4.780	4.473	11.749	7.770	3.715	1.582	0.687	5.951	3.235
	Std.	3.645	3.846	9.055	8.852	3.730	1.757	0.591	4.335	2.765
	Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Max	43.681	43.913	55.458	61.318	18.530	14.916	6.516	86.339	37.236
RA	Mean	0.372	0.265	0.561	0.500	0.538	0.183	0.346	0.498	0.302
	Std.	0.292	0.248	0.306	0.473	0.354	0.199	0.306	0.161	0.225
	Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Max	3.832	3.899	7.115	4.173	3.103	1.484	3.567	7.141	3.061
CACN	Mean	11.755	7.943	33.383	13.725	7.073	2.282	0.642	14.392	5.095
	Std.	9.223	6.769	28.360	17.255	8.772	3.374	0.545	11.617	4.849
	Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Max	103.800	68.800	151.800	117.200	36.600	20.400	7.200	210.000	63.200
CARA	Mean	0.223	0.106	0.336	0.200	0.323	0.208	0.062	0.299	0.121
	Std.	0.175	0.099	0.184	0.189	0.212	0.080	0.064	0.181	0.090
	Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Max	2.299	1.559	4.269	1.669	1.862	2.140	0.693	4.285	1.224
CRS-RA	Mean	0.372	0.222	0.561	0.430	0.538	0.026	0.346	0.498	0.141
	Std.	0.292	0.209	0.306	0.460	0.354	0.040	0.306	0.301	0.113
	Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Max	3.832	3.190	7.115	3.233	3.103	0.299	3.567	7.141	1.743
CMS-RA	Mean	0.352	0.159	0.560	0.300	0.536	0.110	0.342	0.487	0.181
	Std.	0.269	0.149	0.305	0.284	0.354	0.119	0.305	0.298	0.135
	Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Max	3.116	2.339	7.115	2.504	3.103	0.890	3.427	6.740	1.837
ICRA	Mean	37.623	43.641	87.746	327.604	15.870	5.214	7.234	44.273	36.478
	Std.	50.390	64.879	137.252	469.548	17.924	9.350	8.726	60.547	47.745
	Min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Max	889.386	1345.021	4355.164	4360.996	209.767	108.748	156.849	3600.122	1307.093

degree of the network Wang et al. (2015). The complexity to compute thresholds θ_1 and θ_2 is $O(x \cdot \log x)$ as the values will be sorted, and then a value will be chosen such that f fraction of sampled edges have value higher than this. Therefore, the complexity for these methods is $O(n \cdot \log n + x \cdot \deg_{avg}^2 + x \cdot \log x)$. In real-life applications, if $x < n$, the overall complexity is $O(n \cdot \log n + x \cdot \deg_{avg}^2)$. In CRS-RA method, the complexity to compute the community relationship strength is $O(n^2)$ and the complexity to compute the similarity score is $O(\deg_{avg}^2)$, and therefore, the complexity of the proposed framework for CRS-RA method is $O(n \cdot \log n + n^2 + x \cdot \deg_{avg}^2 + x \cdot \log x)$; if $x < n$, then the complexity is $O(n \cdot \log n + n^2 + x \cdot \deg_{avg}^2)$.

5 Experimental analysis

In this section, we discuss datasets, evaluation metrics, and the performance analysis of the proposed method.

5.1 Datasets

The experiments have been performed on different kinds of real-world networks, including friendship networks, collaboration networks, and communication networks. The details of the datasets are mentioned in Table 2. Eu-Email is an email communication network extracted from a European research institution. Facebook is a snapshot of the network extracted from the Facebook social networking website. The GrQc, Hep-th, and Astro-ph are collaboration networks extracted from Arxiv papers for general relativity, high-energy physics theory, and astrophysics scientific research areas, respectively.

5.1.1 Community detection

In most real-world networks, the ground truth community information is not available. The scientific community has defined several methods to identify communities using network structure if the ground truth information is not known. In our work, we apply the most used community detection method, known as the *Louvain community detection method* to identify communities in a network (Blondel et al. 2008). The Louvain method uses two-step greedy optimization to optimize the modularity of a community partition of the network. First, the method optimizes the modularity locally to find small communities. In the second step, it merges all nodes belonging to the same community and creates an aggregated network where each node represents a community. These steps are repeated iteratively until the maximum modularity is achieved and the obtained communities are returned.

In all the networks, the communities are detected using the Louvain Method, and a community label is assigned to each node based on which community it belongs to. A node pair is referred to as intra-community node pair if both nodes belong to the same community, otherwise, it will be referred to as inter-community node pair.

5.1.2 Prepare training-testing dataset

To generate the training and testing data, we follow the same methodology as used in previous works (Epasto and Perozzi 2019; Grover and Leskovec 2016); however, we maintain the ratio of inter and intra-community links that is not considered in previous studies. First, we remove 10% of inter-community and 10% of intra-community edges uniformly at random from E in a complex network and put them in set E_{lp} that will be used for analyzing the HM-EIICT link prediction method. While removing these 10% edges, it is ensured that the network remains connected. For the link prediction task, the same number of inter and intra-community node pairs for non-existent links are chosen uniformly at random, as we have in E_{lp} . These sampled pairs will work as negative cases and are added to set E_{lp} . If a link is formed between a given node pair, then it is referred to as a positive case, and otherwise, it will be referred to as a negative case. To create train and test data, the node pairs in E_{lp} are split into E_{train} and E_{test} , and while splitting, we ensure that the ratio of inter and intra-community node pairs is maintained for both positive and negative cases. The training and testing data ratio is (.5 : .5) if it is not mentioned explicitly. The positive cases of training data are used to compute the threshold values θ_1 and θ_2 . In our experiments, we first compute the similarity score values for all the existing edges in the training dataset and then used it for computing intra-community and inter-community thresholds. For example, suppose we have $f = 0.9$ for intra-community link prediction, then in simple words, an intra-community node pair is predicted positive (recommended to have a link in future) if the similarity score for this pair falls in the range of similarity score of the top 90% pairs in positive intra-community train cases.

5.2 Evaluation metrics

The performance of the proposed method is measured using the following two metrics.

1. **Accuracy:** Accuracy shows the fraction of correctly predicted positive and negative test cases in the testing dataset. It is computed as,

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

where TP is the number of true positive cases, TN is the number of true negative cases, FP is the number of false positive cases, and FN is the number of false negative cases using the confusion matrix.

2. **Modularity Reduction:** The network modularity was originally proposed to identify communities in a network (Newman and Girvan 2004). It compares the link density between the communities with the expected density if the links are distributed uniformly at random in the given network. For a given network, it is defined as,

$$Q = \frac{1}{2m} \sum_{i,j} \sum_{i,j} \left(A_{i,j} - \frac{d_i d_j}{2m} \right) \delta(C_i, C_j) \quad (2)$$

where A is the adjacency matrix representation of the network, m is total number of edges, C_i is the community of node i , d_i is the degree of node i , $\delta(C_i, C_j)$ is the Kronecker delta function. The homophily of a network is higher if a significant portion of the links is between nodes that belong to the same community.

The modularity reduction (modred) method uses modularity to determine whether the proposed link prediction method is unfair to predict more intra-community links than the inter-community links (Masrour et al. 2020). It is defined as,

$$\text{modred} = \frac{Q_{\text{ref}} - Q_{\text{pred}}}{Q_{\text{ref}}} \quad (3)$$

where Q_{ref} is the modularity of the reference network (e.g., the ground truth network when evaluating link prediction algorithms) and Q_{pred} is the modularity of the predicted network that we obtain by adding the edges predicted by the proposed method to the original network. If one method gives a higher modred value than another method, it indicates that the first link prediction method has predicted more inter-community links than the second method.

5.3 Performance study

In our experiments, E_{train} is used for computing θ_1 and θ_2 threshold values. For each heuristic method, the similarity score values are computed for intra-community and inter-community existing links in E_{train} , and the threshold values are chosen using that. The intra-community threshold value is computed using $f = 0.9$ for all datasets and the inter-community threshold is computed using $f = 0.8$ for Eu-Email, $f = 0.9$ for Facebook and Astro-ph, $f = 0.6$ for GrQc, and $f = 0.7$ for Hep-th network. These values are chosen based on the preliminary accuracy analysis that provides good results; more explanation is provided in Sect. 5.4 and refer Fig. 1. The threshold values can be chosen differently for different methods. However, we have used the same value of f for all the methods to maintain consistency in the experiments while comparing different methods. The threshold value for baseline heuristic methods is computed using the same approach, though the only difference is that the similarity-score values are not separated for intra-community and inter-community links. For baseline heuristic methods $f = 0.7$, as the experimental observations showed that this value provides a good accuracy trade-off for both types of links. Each experiment is performed 100 times, and the mean value is reported.

The results for accuracy of different heuristic baseline methods and their EIICT version are shown in Table 5. The results show that the HM-EIICT framework highly improves the accuracy for inter-community link prediction for local (JC, AA, RA) as well as global (CACN, CARA, CRS-RA, CMS-RA, ICRA) heuristic methods, which already considered the community information while computing the similarity score. The accuracy for intra-community link prediction remains intact or improves. Therefore, the HM-EIICT method improves the accuracy of all baseline heuristic methods. The EIICT version of simple heuristic methods, such as RA-EIICT, gives close to the maximum accuracy on GrQc and the maximum accuracy on Eu-Email, Facebook, and

Fig. 1 Impact of varying f for deciding threshold value on Intra-community and Inter-community link prediction accuracy

Table 5 Accuracy results for heuristic methods and their EIICT version

Sr.	Method\datasets		Eu-Email	Facebook	GrQc	Hep-th	Astro-ph
1	JC	Total	0.766	0.826	0.832	0.835	0.846
		Intra	0.821	0.838	0.867	0.875	0.909
		Inter	0.678	0.535	0.545	0.670	0.704
	JC-EIICT	Total	0.787	0.860	0.899	0.909	0.935
		Intra	0.822	0.872	0.917	0.932	0.942
		Inter	0.733	0.564	0.751	0.818	0.918
2	AA	Total	0.807	0.824	0.846	0.843	0.847
		Intra	0.836	0.829	0.870	0.870	0.885
		Inter	0.760	0.699	0.649	0.732	0.763
	AA-EIICT	Total	0.829	0.886	0.916	0.910	0.940
		Intra	0.842	0.889	0.932	0.932	0.943
		Inter	0.808	0.813	0.783	0.819	0.933
3	RA	Total	0.806	0.836	0.847	0.848	0.850
		Intra	0.846	0.840	0.874	0.873	0.892
		Inter	0.743	0.739	0.627	0.744	0.755
	RA-EIICT	Total	0.833	0.905	0.916	0.910	0.945
		Intra	0.852	0.908	0.933	0.933	0.948
		Inter	0.802	0.819	0.778	0.820	0.939
4	CACN	Total	0.789	0.822	0.875	0.886	0.851
		Intra	0.840	0.830	0.903	0.943	0.901
		Inter	0.708	0.638	0.650	0.652	0.738
	CACN-EIICT	Total	0.820	0.877	0.918	0.924	0.938
		Intra	0.842	0.880	0.934	0.943	0.942
		Inter	0.784	0.794	0.786	0.842	0.929
5	CARA	Total	0.773	0.834	0.847	0.841	0.848
		Intra	0.852	0.844	0.884	0.894	0.938
		Inter	0.648	0.600	0.545	0.624	0.647
	CARA-EIICT	Total	0.812	0.903	0.915	0.910	0.941
		Intra	0.856	0.908	0.934	0.933	0.948
		Inter	0.744	0.788	0.758	0.818	0.925
6	CRS-RA	Total	0.801	0.834	0.846	0.837	0.848
		Intra	0.853	0.839	0.888	0.910	0.933
		Inter	0.720	0.711	0.502	0.538	0.657
	CRS-RA-EIICT	Total	0.829	0.903	0.915	0.911	0.940
		Intra	0.856	0.908	0.935	0.933	0.948
		Inter	0.786	0.775	0.758	0.819	0.922
7	CMS-RA	Total	0.797	0.836	0.845	0.843	0.850
		Intra	0.854	0.842	0.876	0.880	0.911
		Inter	0.707	0.682	0.590	0.690	0.711
	CMS-RA-EIICT	Total	0.824	0.904	0.913	0.912	0.944
		Intra	0.855	0.908	0.931	0.934	0.948
		Inter	0.775	0.805	0.763	0.821	0.934
8	ICRA	Total	0.805	0.826	0.845	0.837	0.848
		Intra	0.817	0.828	0.864	0.867	0.863
		Inter	0.786	0.772	0.691	0.715	0.814
	ICRA-EIICT	Total	0.828	0.881	0.900	0.905	0.936
		Intra	0.841	0.884	0.919	0.929	0.940
		Inter	0.807	0.801	0.747	0.810	0.927

The bold values show the best results for each type of method

Table 6 Modularity reduction for heuristic methods and their corresponding EIICT version

Sr.	Method/Datasets	Eu-Email	Facebook	GrQc	Hep-th	Astro-ph
1	JC	-0.017	-0.001	-0.004	-0.005	-0.008
	JC-EIICT	-0.010	0.000	-0.002	-0.003	-0.001
2	AA	-0.008	0.001	-0.003	-0.003	-0.004
	AA-EIICT	-0.006	0.000	-0.002	-0.002	0.000
3	RA	-0.011	0.000	-0.003	-0.003	-0.005
	RA-EIICT	-0.007	0.000	-0.002	-0.002	0.000
4	CACN	-0.012	0.001	-0.003	-0.007	-0.006
	CACN-EIICT	-0.007	0.001	-0.002	-0.002	-0.001
5	CARA	-0.020	-0.001	-0.004	-0.006	-0.011
	CARA-EIICT	-0.009	0.000	-0.002	-0.003	-0.001
6	CRS-RA	-0.013	0.000	-0.005	-0.009	-0.010
	CRS-RA-EIICT	-0.007	0.000	-0.002	-0.003	-0.001
7	CMS-RA	-0.014	0.000	-0.003	-0.004	-0.008
	CMS-RA-EIICT	-0.008	0.000	-0.002	-0.003	0.000
8	ICRA	-0.005	0.001	-0.002	-0.003	-0.002
	ICRA-EIICT	-0.005	0.001	-0.002	-0.002	-0.001

The bold values show the best results for each type of method

Astro-ph network. If we compare the results of JC-EIICT, AA-EIICT, and RA-EIICT, i.e., HM-EIICT for local heuristics, with global heuristic methods, including CACN, CARA, CRS-RA, CMS-RA, and ICRA, for bigger networks (Facebook, GrQc, Hep-th, and Astro-ph), the former methods perform better; that shows the efficiency of the proposed framework compared to global heuristic methods. We further compute the modularity reduction for link prediction to analyze how the diversity is increased. The results in Table 6 show that the HM-EIICT reduces the modularity considerably as compared to baseline heuristic methods, and therefore, improves the diversity. The Facebook network has inter-community links much lesser than intra-community links, and therefore, the modularity reduction is close to 0 for various link prediction methods on Facebook.

We would like to mention that we have used training data to compute the threshold values. However, In real-life applications, the threshold values can also be computed using the similarity score of all existing links (E) in the network. We also performed experiments using this approach and achieved similar accuracy and modularity reduction. In our work, we have shown results only for threshold values computed using the training dataset as it shows the efficiency of the proposed method by only using 5% edges while computing the threshold values. We also observe that different methods give good accuracy on different datasets. The RA-EIICT provides the highest accuracy on Eu-Email, Facebook, and Astro-ph, and CACN-EIICT provides the highest accuracy on GrQc and Hep-th networks.

5.4 Sensitivity analysis

First, we study how the accuracy changes as we vary f from 0.1 to 0.9. The results are shown in Fig. 1, where the accuracy is the mean value for 100 random iterations, and the error bars show the standard deviation. The accuracy for Intra-community link prediction shows that $f = 0.9$ gives good results for all the datasets. The accuracy for Inter-community link prediction increases with f and further decreases. The highest inter-community link prediction accuracy is achieved when f ranges from 0.6 to 0.9; it is high when $f \sim 0.8, 0.9, 0.6, 0.7$, and 0.9 for Eu-Email, Facebook, GrQc, Hep-

Fig. 2 Impact of varying training size on link prediction accuracy

th, and Astro-ph datasets, respectively. In GrQc and Hep-th datasets, the accuracy is 0.5 (that is the same as for random prediction) for $f \sim 0.8$ as it gives $\theta_2 = 0$, and therefore, all the links will be predicted positive.

Next, we study how the link prediction accuracy changes with training size and the results are shown in Fig. 2. The results show that for the Hep-th dataset, good accuracy is achieved when the training size is greater than 0.2. For the Astro-ph network (that is the largest considered network), the highest accuracy is achieved when the training size is equal to or greater than 0.1. This shows that even a small fraction of edges to compute the threshold values will provide good link prediction accuracy.

The proposed link recommendation framework is straightforward and fast to compute and will help in evolving a diverse network. The efficiency of the proposed approach can be further improved by choosing optimal values of θ_1 and θ_2 that increase the accuracy for both intra-community as well as inter-community link prediction in a given network, respectively. However, the method to choose optimal threshold values using network structure is still an open research question.

6 Conclusion

In this work, we first studied the structural properties of intra-community and inter-community links using node-pair similarity indices. A node-pair similarity method assigns a similarity score to each pair of nodes based on their neighborhood network structure, and if required using other meta information, such as community labels. We observed that inter-community node pairs have lower node-proximity based similarity than intra-community links, which was expected due to the homophilic structure of real-world networks. Next, based on our observations, we proposed a family of indices, called HM-EIICT (Heuristic Method-Extended using Intra and Inter Community Threshold), to predict both intra-community as well as inter-community links with higher accuracy. The proposed method is evaluated using the accuracy and modularity reduction function. The results showed a huge improvement in inter-community link prediction and also in overall accuracy. The proposed method is fast and easy to compute, and therefore, will be useful in increasing the diversity in the network. The

computation of the optimal value of the threshold for both intra-community as well as inter-community node pairs is an open question that should be looked further.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adamic LA, Adar E (2003) Friends and neighbors on the web. *Soc Netw* 25(3):211–230
- Almaas E, Vázquez A, Barabási AL (2007) Scale-free networks in biology. *Biol Netw* 3(1)
- Bai S, Fang S, Li L, Liu R, Chen X (2019) Enhancing link prediction by exploring community membership of nodes. *Int J Mod Phys B* 33(31):1950382
- Biswas A, Biswas B (2017) Community-based link prediction. *Multimed Tools Appl* 76(18):18619–18639
- Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech: Theory Exp* 2008(10):P10008
- Caldarelli G (2007) Scale-free networks: complex webs in nature and technology. OUP Catalogue, Oxford
- Cannistraci CV, Alanis-Lobato G, Ravasi T (2013) From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Sci Rep* 3(1):1–14
- Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks. *Phys Rev E* 70(6):066111
- Ding J, Jiao L, Wu J, Liu F (2016) Prediction of missing links based on community relevance and ruler inference. *Knowl Based Syst* 98:200–215
- Epasto A, Perozzi B (2019) Is a single embedding enough? learning node representations that capture multiple social contexts. In: *The world wide web conference*, pp 394–404
- Fouss F, Pirotte A, Renders JM, Saeens M (2007) Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Trans Knowl Data Eng* 19(3):355–369
- Gao S, Denoyer L, Gallinari P (2011) Temporal link prediction by integrating content and structure information. In: *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp 1169–1174
- Gao F, Musial K, Gabrys B (2017) A community bridge boosting social network link prediction model. In: *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining* 2017, pp. 683–689
- Garimella K, Gionis A, Parotsidis N, Tatti N (2017) Balancing information exposure in social networks. *arXiv preprint arXiv:1709.01491*
- Granovetter MS (1973) The strength of weak ties. *Am J Sociol* 78(6):1360–1380
- Grover A, Leskovec J (2016) node2vec: Scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp 855–864
- Hofstra B, Corten R, Van Tubergen F, Ellison NB (2017) Sources of segregation in social networks: a novel approach using facebook. *Am Sociol Rev* 82(3):625–656
- Holme P, Kim BJ (2002) Growing scale-free networks with tunable clustering. *Phys Rev E* 65(2):026107
- Jeon H, Kim T (2017a). Community-adaptive link prediction. In: *Proceedings of the 2017 international conference on data mining, communications and information technology*
- Jeon H, Kim T (2017b) Community-adaptive link prediction. In: *Proceedings of the 2017 international conference on data mining, communications and information technology*, pp 1–5
- Leicht EA, Holme P, Newman ME (2006) Vertex similarity in networks. *Phys Rev E* 73(2):026120
- Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. *ACM Trans Knowl Discov Data (TKDD)* 1(1):2-es

- Li L, Fang S, Bai S, Xu S, Cheng J, Chen X (2019) Effective link prediction based on community relationship strength. *IEEE Access* 7:43233–43248
- Li P, Wang Y, Zhao H, Hong P, Liu H (2021) On dyadic fairness: exploring and mitigating bias in graph connections. In: *Proceedings of international conference on learning representations*
- Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. *J Am Soc Inform Sci Technol* 58(7):1019–1031
- Liu S, Ji X, Liu C, Bai Y (2017) Extended resource allocation index for link prediction of complex network. *Physica A* 479:174–183
- Liu W, Lü L (2010) Link prediction based on local random walk. *EPL (Europhys Lett)* 89(5):58007
- Lü L, Jin CH, Zhou T (2009) Similarity index based on local paths for link prediction of complex networks. *Phys Rev E* 80(4):046122
- Masrour F, Wilson T, Yan H, Tan PN, Esfahanian A (2020) Bursting the filter bubble: Fairness-aware network link prediction. In: *Proceedings of the AAAI conference on artificial intelligence*, vol 34. pp 841–848
- Matakos A, Tu S, Gionis A (2020) Tell me something my friends do not know: diversity maximization in social networks. *Knowl Inf Syst* 62(9):3697–3726
- Mcauley JJ, Leskovec J (2012) Learning to discover social circles in ego networks. In: *NIPS*, vol 2012. pp 548–556
- McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. *Ann Rev Sociol* 27(1):415–444
- Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E* 69(2):026113
- Onnela JP, Saramäki J, Hyvönen J, Szabó G, De Menezes MA, Kaski K, Barabási AL, Kertész J (2007) Analysis of a large-scale weighted network of one-to-one human communication. *New J Phys* 9(6):179
- Park S, Kang S, Chung S, Song J (2009) Newscube: delivering multiple aspects of news to mitigate media bias. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, pp 443–452
- Peceli A, Cavalcanti MC, Goldschmidt R (2018) Automatic feature selection for supervised learning in link prediction applications: a comparative study. *Knowl Inf Syst* 56(1):85–121
- Praveena A, Smyth S (2016) Anonymization in social networks: a survey on the issues of data privacy in social network sites. *J Int J Eng Comput Sci* 5(3):15912–15918
- Rahman TA, Surma B, Backes M, Zhang Y (2019) Fairwalk: towards fair graph embedding. In: *IJCAI*, pp 3289–3295
- Saxena A (2020) A survey of evolving models for weighted complex networks based on their dynamics and evolution. *arXiv preprint* [arXiv:2012.08166](https://arxiv.org/abs/2012.08166)
- Saxena A, Iyengar S (2016) Evolving models for meso-scale structures. In: *2016 8th international conference on communication systems and networks (COMSNETS)*. IEEE, pp 1–8
- Saxena A, Hsu W, Lee ML, Leong Chieu H, Ng L, Teow LN (2020a) Mitigating misinformation in online social network with top-k debunkers and evolving user opinions. In: *Companion proceedings of the web conference 2020*, pp 363–370
- Saxena A, Saxena H, Gera R (2020b) k-truthscore: Fake news mitigation in the presence of strong user bias. In: *International conference on computational data and social networks*, Springer, pp 113–126
- Saxena A, Fletcher G, Pechenizkiy M (2021) Nodesim: Node similarity based network embedding for diverse link prediction. *arXiv preprint* [arXiv:2102.00785](https://arxiv.org/abs/2102.00785)
- Singh SS, Mishra S, Kumar A, Biswas B (2020) Clp-id: Community-based link prediction using information diffusion. *Inf Sci* 514:402–433
- Spinelli I, Scardapane S, Hussain A, Uncini A (2021) Biased edge dropout for enhancing fairness in graph representation learning. *arXiv preprint* [arXiv:2104.14210](https://arxiv.org/abs/2104.14210)
- Tong H, Faloutsos C, Pan JY (2006) Fast random walk with restart and its applications. In: *Sixth international conference on data mining (ICDM'06)*, IEEE, pp 613–622
- Valverde-Rebaza JC, de Andrade Lopes A (2012) Link prediction in complex networks based on cluster information. In: *Brazilian symposium on artificial intelligence*, Springer, pp 92–101
- Wang P, Xu B, Wu Y, Zhou X (2015) Link prediction in social networks: the state-of-the-art. *Sci China Inf Sci* 58(1):1–38
- Wang J, Ma Y, Liu M, Yuan H, Shen W, Li L (2017) A vertex similarity index using community information to improve link prediction accuracy. In: *2017 IEEE international conference on systems, man, and cybernetics (SMC)*, pp 158–163

- Wang J, Ma Y, Liu M, Shen W (2019) Link prediction based on community information and its parallelization. *IEEE Access* 7:62633–62645
- Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. *Nature* 393(6684):440–442
- Wu J, Zhang G, Ren Y (2017) A balanced modularity maximization link prediction model in social networks. *Inf Process Manag* 53(1):295–307
- Wu Z, Lin Y, Wang J, Gregory S (2016) Link prediction with node clustering coefficient. *Physica A* 452:1–8
- Yan B, Gregory S (2012) Finding missing edges in networks based on their community structure. *Phys Rev E* 85(5):056112
- Yu K, Chu W (2007) Gaussian process models for link analysis and transfer learning. In: *NIPS*, pp 1657–1664
- Zhou T, Lü L, Zhang YC (2009) Predicting missing links via local information. *Eur Phys J B* 71(4):623–630
- Zhu B, Xia Y (2015) An information-theoretic model for link prediction in complex networks. *Sci Rep* 5:13707

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.