Enhance Ambiguous Community Structure via Multi-strategy Community Related Link Prediction Method with Evolutionary Process

Qiming Yang, Wei Wei, Ruizhi Zhang, Bowen Pang, and Xiangnan Feng

Abstract—Most real-world networks suffer from incompleteness or incorrectness, which is an inherent attribute to real-world datasets. As a consequence, those downstream machine learning tasks in complex network like community detection methods may yield less satisfactory results, i.e., a proper preprocessing measure is required here. To address this issue, in this paper, we design a new community attribute based link prediction strategy HAP and propose a two-step community enhancement algorithm with automatic evolution process based on HAP. This paper aims at providing a community enhancement measure through adding links to clarify ambiguous community structures. The HAP method takes the neighbourhood uncertainty and Shannon entropy to identify boundary nodes, and establishes links by considering the nodes' community attributes and community size at the same time. The experimental results on twelve real-world datasets with ground truth community indicate that the proposed link prediction method outperforms other baseline methods and the enhancement of community follows the expected evolution process.

Index Terms—Complex network, community enhancement, link prediction, community detection, Shannon entropy.

1 Introduction

Owing to the fact that a great deal of real world data could be expressed in complex network fashion [1], complex network analysis attracts more and more attention in many scientific disciplines. There are various approaches to unveil the underlying information behind networks. Among these studies, community detection has been considered as one of the most vital topics [2], [3], [4].

The network community is defined as a group of nodes which are densely connected to each other, while sparsely connected to the rest nodes [5]. In many real-world cases community structure appears frequently. Social networks are paradigmatic examples of graphs with communities, since people have the tendency to form groups with similar interest or ideology [6]. In biological bodies, community structure in protein interaction networks could represent a group of proteins with similar functions [7].

Many community detection methods have been proposed to approach the problem. According to the work by Fortunato et al. [8], community detection methods can be roughly divided into five categories. The first category contains the traditional graph partitioning methods like Kernighan-Lin algorithm [9] and it's extended version by Suaris et al. [10]. The second category adopts the hierarchical clustering, represented by the GN algorithm [6] and the FN algorithm [11]. The third category is the modularity based methods, which convert the clustering process into optimising modularity measures [5], [12], [13]. The fourth one contains the spectral clustering algorithms since the eigenvectors of network Laplacian matrices have several desirable properties related to community structure [14], [15], [16]. The last category is the dynamic methods, in which the spin models [17] and random walks [18] are often used.

However, most real-world datasets are severely incomplete [8], e.g., in online social networks like Facebook (currently renamed as Meta), Twitter and Sina, only part of global social information could be collected; in gene interaction networks, links among genes are usually measured by costly experiments. The prevalent imperfections on real-world network datasets might lead to incorrect community detection outcomes, so community enhancement methods are critical.

Researches have been conducted to fix impaired network systems by predicting which node pairs are more likely to establish links, also known as the link prediction methods [19], [20], [21], [22], [23], [24], [25]. However, to the best of our knowledge, researches on the topic of community enhancement by link prediction have been rarely discussed. Link predcition could be regarded as a data preprocssing procedure, which has played a key role in the practice of machine learning models [26], [27]. Therefore, it is natural to assume that the link prediction methods could have significant function and play a major role in community enhancement task.

Moreover, there are three main drawbacks in existing

- *Q. Yang, R. Zhang and B. Pang are with the School of Mathematical Sciences, Beihang University, Beijing, China, and also with Key Laboratory of Mathematics Informatics Behavioral Semantics, Ministry of Education, China. E-mail: {asdyqm, ruizhiz, pangbw}@buaa.edu.cn*
- *W. Wei is with the School of Mathematical Sciences, Beihang University, Beijing, China, also with Key Laboratory of Mathematics Informatics Behavioral Semantics, Ministry of Education, China, also with Institute of Artificial Intelligence, Beihang University, Beijing, China, and also with Peng Cheng Laboratory, Shenzhen, Guangdong, China. E-mail: weiw@buaa.edu.cn*
- *X. Feng is with the Center for Humans and Machines, Max Planck Institute for Human Development, Lentzeallee 94, Berlin, Germany. Email: fengxiangnan@gmail.com*
- *Corresponding author: Wei Wei*

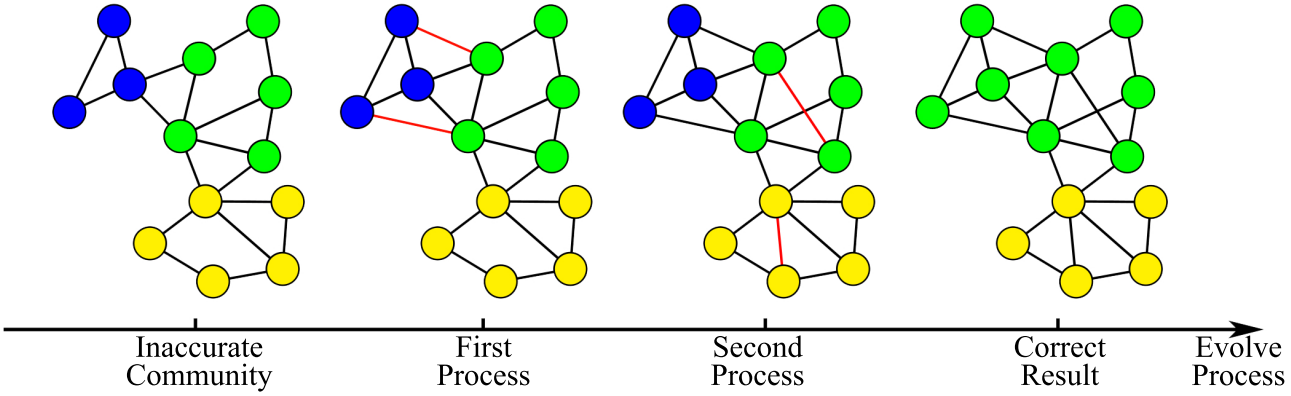


Fig. 1. Diagram of proposed community enhancement measure. The first subfigure indicates the biased community detection result by algorithms and the ground truth community is illustrated at the last subfigure. In the first process the community enhancement measure adjusts biased result by adding inter-cluster links between inaccurately separated subcommunities. On the contrary, it enhances the correct result by adding intra-cluster edges in the second process.

methods:

- Most methods do not take the evolution of network structure into consideration. Most experiments are conducted on single iteration or snapshot, such as the EdgeBoost [28] algorithm, in which the long term numerical stability is not discussed and only two-hop neighbours are concerned.
- Most methods treat the community detection output as the ground truth result. However, as aforementioned, due to the imperfection of real-world data, clustering results are usually flawed, which means adjustments are required in link prediction algorithms to fix the results.
- Most methods only consider single strategy. However, applying different strategies simultaneously might gain notable improvement. For example, In graph representation learning, the IDGNN by Leskovec et al. [29] successfully achieved expressive power over 1-WL test [30] by applying two activation functions at the same time.

In this paper, we design a harmony-based aggregation preferred (HAP) link prediction method and propose a community enhancement algorithm based on this strategy. Our method comprises two distinct procedures. At the first stage it will suture fractured communities into a complete one, while the repaired communities will get enhanced at stage two. The diagram of such procedure is presented in figure 1. Experimental results on real-world datasets with ground truth community information show that our method achieves better performance than baseline methods in most cases. The main contributions of our work are:

- We design a new link prediction strategy HAP for community enhancement task. Our HAP method does not require any prior knowledge about the ground truth community such as the number of communities.
- The HAP method could be applied as a plug-in module in preprocessing procedure. It could be complementary to arbitrary community detection algorithms easily.
- Using Shannon Entropy, the definitions of consistency and harmony perform practical potential in community related problems or situations where nodes have attributes and POIs.

The remainder of this paper is organized as follows. Sect 2 presents related research work. Next, Sect 3 provides formal definition of community detection problem and illustrates the details of our method including the inductive biases. Furthermore, Sect 4 shows the experimental results and comparison with other baseline models. Finally Sect 5 gives the conclusion and future work.

2 Related Work

Both link prediction and community detection are of great significance in network analysis since both of them provide network topology information from various perspectives.

For future references, three well-known community detection algorithms are introduced here:

- Louvain [13]: It is a heuristic method based on modularity optimization. This algorithm first assigns different community labels to all nodes then optimizes the modularity by aggregating those separate communities.
- Infomap [31]: This method uses the probability flow of random walks on a network as a proxy for information flows in the real system. It is an information based approach capable in revealing community structure in weighted and directed networks.
- Label Propagation (LPA) [18]: The major advantage of this algorithm is that it has a near linear time complexity. LPA method solely uses the network structure as its information with each node adopting the label that most of its neighbours currently have at every iteration step.

Besides, several community detection algorithms were proposed based on the altering of network topology structure. Zhang et al. [32] designed an enhanced semi-supervised learning framework for community detection, which required prior knowledge about nodes. Yang et al. [33] considered which prior information is critical for performance improvement and proposed an active link selection framework. Su et al. [34] proposed CSE method based on central and boundary node identification for community

enhancement, which successfully removed the limitation of prior knowledge about nodes. Zhou et al. [35] proposed genetic algorithm and similarity ensemble based community enhancement methods to explore the robustness under adversarial attack.

In link prediction oriented problem, the community detecting results could be regarded as a global attribute to provide extra information for link prediction algorithms. Soundarajan and Hopcroft [36] rewrote the classic CN index and RA index with community information and the experimental results showed improvement. Rebaza and Lopes [37] took intra-cluster and inter-cluster into consideration and proposed WIC measure, which can be extended on directed and asymmetric large-scale networks [38]. Ai et al. [39] presented a link prediction method for personalized recommendation circumstance based on complex network modelling and community detection results.

Meanwhile, link prediction methods can be used to enhance ambiguous community structure. Yang et al. [40] proposed a conditional model for link prediction and a discriminative model for content analysis. Chen et al. [41] tested three traditional link prediction methods for enhancing community structure. Bacco et al. [42] proposed a generative model for multilayer network with interdependence among its layers. Jiang et al. [43] designed a strategy based on node centralities to establish clear boundaries among communities. Burgess et al. [28] proposed EdgeBoost structure and explored the improvement of community detection performance of three link prediction algorithms with six community detection methods.

3 METHODS

In this section, we explain the proposed community enhancement algorithm in detail. The key component of the proposed enhancement algorithm is the HAP link prediction method, based on which the inductive biases and the intuitions will be illustrated. Firstly we will formally define the problem and all the symbols used in this paper.

3.1 Problem and Definitions

Every complex network system can be presented as an ordered tuple $G = (V, E)$, where V is the set of nodes and $E \subseteq V \times V$ represents the edge set of network G [44]. In this paper we mainly concern about undirected graph, i.e., $\forall v_i, v_j \in V \; and \; (v_i, v_j) \in E \Rightarrow (v_j, v_i) \in E$.

In addition, for network with community attributes, each node has its clustering label from ground truth knowledge C_G or by applying community detection algorithms C_A. In a network with K clusters, $C = \{C_1, C_2, \cdots, C_K\}$ denotes the set of clustering labels. We consider both C_G and C_A as a function that satisfying $C_G, C_A : V \mapsto C$, namely $\forall v \in V$, it has its ground truth community attribute $C_G(v) \in C$ and algorithm result $C_A(v) \in C$. Furthermore, $C_G(v)$ does not necessarily equal to $C_A(v)$.

N_{C_i} stands for the number of nodes with the community attribute C_i. CM is the connection matrix satisfying $CM \in \mathbb{Z}^{K \times K}$. We let $CM(i, j)$ be the number of edges between cluster C_i and C_j.

For future reference, the definition of *revising edges* is illustrated here. Given the ground truth community C_G and algorithm's output C_A, edge $e = (v_i, v_j)$ is a *revising edges* if $C_G(v_i) = C_G(v_j)$ and $C_A(v_i) \neq C_A(v_j)$. To clarify, revising edges might not exist in some networks. At the same time we provide the definition of *reinforcing edges*: edge $e = (v_i, v_j)$ is a *reinforcing edge* if and only if $C_G(v_i) = C_G(v_j)$ and $C_A(v_i) = C_A(v_j)$.

3.2 Inductive Biases

It is worth noticing that in most cases, community detection methods will yield a larger number of clusters on both real world and synthesized networks.

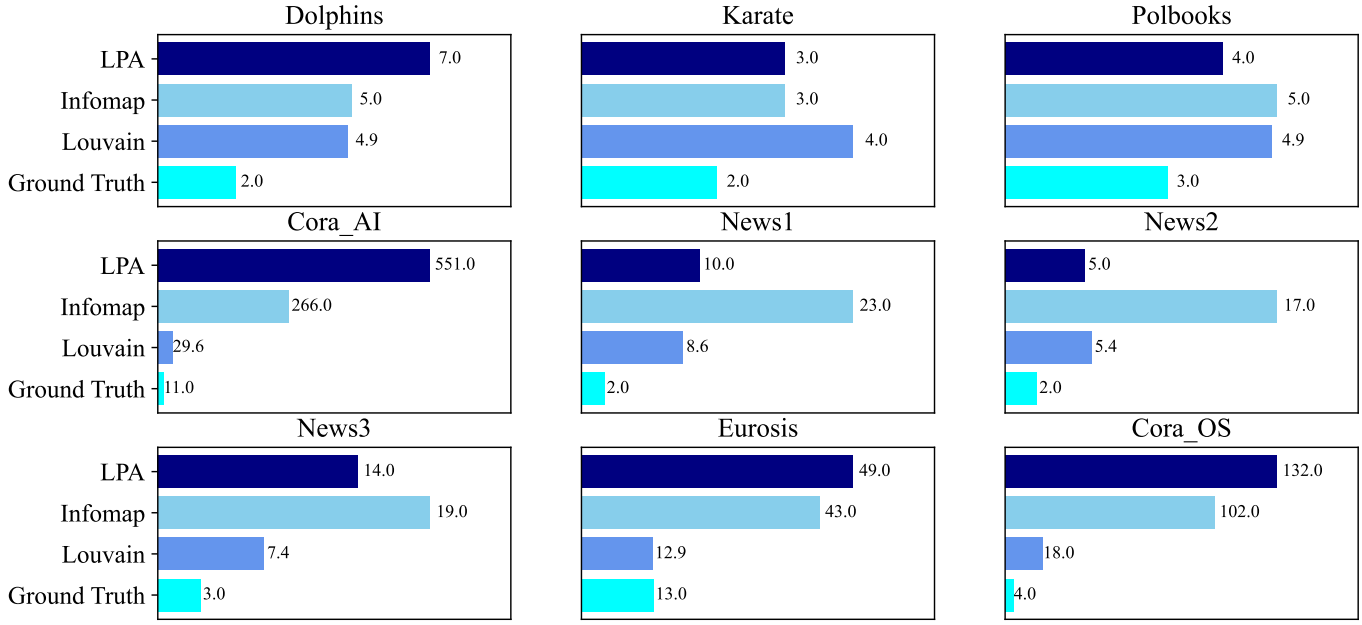
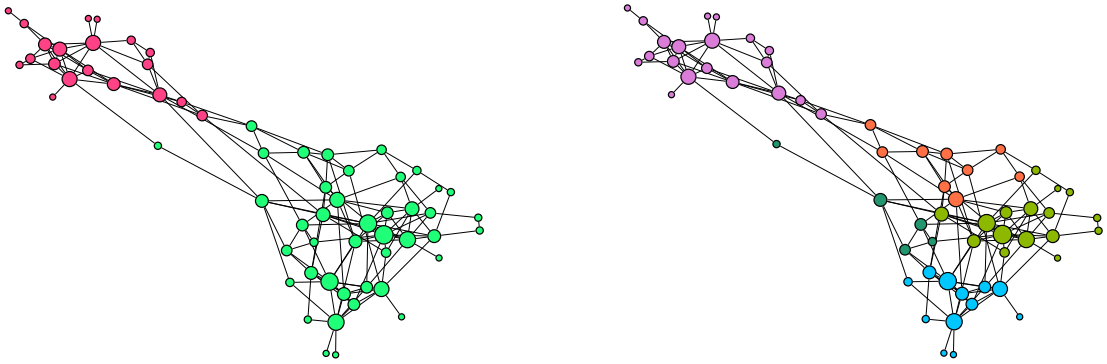


Fig. 2. Average number of clusters in 10 independent iterations. Comparing with the ground truth, LPA and Infomap community detection methods clearly yields higher number of communities. On the other hand, Louvain algorithm tends to detect fewer communities than those two aforementioned methods but is still biased in most cases except Eurosis dataset.



(a) Ground Truth (b) Infomap Result

Fig. 3. An example of the Dolphin network indicating that community detection algorithms turn complete community into fragments. Compared with the ground truth information, the larger community in dolphin network gets fractured by the Infomap community detection method.

Here we present the results of three community detection algorithms on nine real-world network systems in figure 2. Notice the difference in cluster numbers between ground truth knowledge and clustering results. Except Louvain method on Eurosis network, community detection methods output larger number of communities than ground truth. Especially in Cora_OS dataset, the results of LPA and Infomap are 25 times larger than the actual instance number. The experimental results of those community detection methods on LFR benchmark graphs yields higher number of community is referred to [28]. All real-world datasets mentioned here will be formally introduced in section 4.

Furthermore, in most cases the emergence of extra clusters comes from the fracture of complete ground truth communities. For example, demonstrated in figure 3, the larger ground truth community in dolphins network gets fragmented into four smaller communities by the Infomap method. If the link prediction method could make connections among different parts of fractured subcommunities, community detection methods would achieve better performance by recognising and merging those subcommunities into a complete one.

If we could adjust the network topology structure by connecting nonexistent revising edges, the community detection algorithms would get a higher likelihood to recognize the ground truth community of network.

Moreover, since the proposed method is designed to make connections among fractured subcommunities, it has the tendency to form links between small community and large community. That is to say, if the size of community in the complex network has great difference, the link prediction method is risky to dissipate the small clusters.

In conclusion, the inductive biases in this paper are listed as follow

- The outputs of community detection algorithms are highly likely to be incorrect and contain more clusters than actual case because detection methods tend to split large community into smaller ones.
- The connecting of revising edges enhances the ambiguous community structure, which could help the downstream community detection algorithms preform better.
- The sizes (number of nodes) of ground truth communities in a complex network are in similar scales.

3.3 Algorithm Skeleton

It is commonly accepted that ambiguous community structure is challenging for community detection studies due to the subtle difference between inter-edges and intra-edges [34]. The leading thought of the HAP link prediction method is to add links among fractured components of a complete community, thus turning misunderstood inter-edges into affirmative intra-edges.

As a community attribute related unsupervised link prediction method, the HAP method requires a community detection algorithm to trigger the community enhancement procedure since no prior knowledge about the ground truth community is available.

To be specific, the proposed community structure enhancement method contains three main iterating processes: 1) community detection; 2) central and boundary nodes recognition; 3) adding links. The community detection process will not be discussed here since it is given by the users.

3.4 Central Nodes Recognition

In order to achieve the goal of community enhancement, the HAP link prediction method will recognise central and boundary nodes while simultaneously adds links to the network. The proposed community enhancement method has an evolutionary process which will be explained in short notice.

When detecting the central and boundary nodes, most methods use centrality measure to define whether nodes are on the edge of community or not [34] [35]. Such centrality measure can be defined through average distance between intra-community nodes, which can be regarded as geometric distance centrality. Other methods, such as calculating the fraction of neighbour nodes that have the same community attributes, could be regarded as the probability of one-step random walk ending within the community.

These two measures are both successful in identifying boundary nodes. However, distance based centrality measure will bring unwanted calculation complexity. Meanwhile, nodes with large degree have a higher likelihood to

yield smaller average distance. Also the definition of boundary node usually does not consider the node's neighbours with different community attributes.

The fraction of connection based method has less computation complexity and takes the neighbour nodes' community attributes into consideration. However, the linear expression in fraction of connection might not utilize the neighbourhood information fully since communities are usually treated without difference.

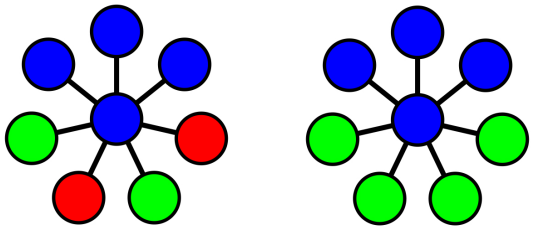


Fig. 4. Diagram of node with its neighbourhood information. The diversity of nodes' colour indicates they belong to different community

Take figure 4a as an example. In [34], the centrality score of a node is defined as:

$$CS_u = \frac{|N(u) \cap N_{LS}|}{|N(u)|}, \tag{1}$$

where the numerator stands for the number of nodes with the same community attribute while the denominator is the degree of node u. So we can easily calculate that for the central node in figure 4a its centrality score equals to $3/7$.

If we alter the neighbour nodes' community attribute, the shortage of such centrality measure will be exposed. As shown in figure 4b. It can be seen that in figure 4b, the centrality score of the central blue node remains $3/7$.

In other word, this centrality function fails to express the difference between figure 4a and 4b. Thus we should take the neighbour nodes' community attributes into deeper consideration.

There is no doubt that the most useful and easily accessible information for the centrality score is the neighbourhood of nodes. Furthermore, for node x, all its neighbour nodes $\Gamma(x)$ should contain the community attribute information. Here we define the neighbourhood community enumeration (NCE) of node x, given the community mapping function C_G:

$$NCE_G(x) =< C_G(i) | i \in \Gamma(x) > . \tag{2}$$

Such as in figure 4a, the central node's NCE is $< 1, 1, 1, 2, 2, 3, 3 >$ ($1, 2$ and 3 stand for blue, green and red community respectively). In figure 4b, the central node's NCE is $< 1, 1, 1, 2, 2, 2, 2 >$. We apply entropy [45] to assess the mapping distribution in NCE. If a node has intricate neighbourhood information, namely a large entropy value, then it can be treated as the boundary node of a community. Here we use the Shannon Entropy to quantify the uncertainty of the NCE. For a node x, given the community mapping function C_G, its boundary score $BS(x)$ can be calculated as follow:

$$BS(x) = \frac{ShannonEntropy(NCE_G(x))}{log(|\Gamma(x)|)}. \tag{3}$$

If $|\Gamma(x)| = 1$, its boundary score is set as 0. It can be easily proven that the maximum value of the numerator is $log(|\Gamma(x)|)$, thus the BS value is always among 0 and 1. Higher BS score of a node indicates its neighbourhood community attributes are more various.

For example, the central node in figure 4a has a boundary score 0.5545 while in figure 4b the boundary score equals to 0.3509. It is more precise since in figure 4a, the central node is on the overlapping section of three communities while in figure 4b it stands between two communities.

Algorithm 1: Boundary Score Calculation

Input: Graph $G = (V, E)$, Community Mapping Function C_M and node x.
Output: Boundary Score $BS(x)$.

```
NCE ⟵ ∅
Counter ⟵ ∅
for v ∈ V do
    if (x, v) ∈ E then
        NCE ⟵ NCE ∪ C_M(v)
        Counter[C_M(v)] += 1
Counter ⟵ Normalize(Counter)
BS(x) ⟵ 0
for key ∈ Counter do
    p ⟵ Counter[key]
    BS(x) ⟵ BS(x) − p ∗ log_2(p)
return BS(x)/log_2(len(NCE))
```

With the help of Boundary Score, the definition of *consistency score (CS)* is defined as:

$$CS(x) = 1 - BS(x). \tag{4}$$

The maximum value of CS of a node will be achieved when there is only one type of community in its neighbourhood. Larger CS value of a node suggests its neighbours' community attributes perform less uncertainty. This consistency measure can be viewed as a node centrality index evaluating the uncertainty of its neighbourhood clustering information.

3.5 Link Connection

3.5.1 Harmony Similarity Measure

As discussed before, due to the imperfection of community detection algorithms, there is no reason that link prediction methods should take the clustering results as the ground truth. Experimental results in section 4 will prove that the usage of community attributes with insufficient consideration might damage the community enhancement.

Consider the RA index [46]. The intuition behind this method is the resource allocation process on networks. The similarity function is formulated as:

$$s(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{|\Gamma(z)|} \tag{5}$$

It can be interpreted as a way to quantify the information transmission efficiency, which is under the assumption that intermediate nodes between x and y will equally spread information towards their own neighbours. So equation 5

can be regarded as a measure of amount of information flow coming out of node x and received by node y.

Consider the information spreading with community attributes. Instead of spreading to neighbours uniformly, the information flow might be measured inaccurately by the uncertainty of intermediate nodes' neighbourhood information. That is to say, in this case, the information flow is not point-to-point but in the community-to-community pattern.

In other words, creating links between nodes that have higher consistent common neighbours is equivalent to creating links that are more likely to have the same community attribute. Here we define the $harmony(HM)$ value to evaluate this consistency between node x, y, namely:

$$HM(x,y) = \frac{1}{|\Gamma(x) \cap \Gamma(y)|} \sum_{z \in \Gamma(x) \cap \Gamma(y)} CS(z) \qquad (6)$$

where CS is the *consistency* measure. Through equation 6, it is clearly seen that the HM index is a second order similarity measure via entropy.

The HM score of two nodes is higher if their common neighbours have larger average CS value. Following the definition of consistency, a node with higher consistency value indicates that it is more likely to be a central node of a community. Then nodes x, y will yield a large harmony value if they share a neighbourhood of central nodes, suggesting that x and y are more likely to share the same community.

Algorithm 2: Harmony Score Calculation

Input: Graph $G = (V, E)$, Community Mapping Function C_M and nodes x, y.
Output: Harmony Score $HS(x, y)$.
$CommonNeighbour(CN) \longleftarrow \emptyset$
$Counter \longleftarrow 0$
for $v \in V$ **do**
  if $(x, v) \in E$ **and** $(y, v) \in E$ **then**
    $CN \longleftarrow CN \cup v$
    $Counter \mathrel{+}= 1$
$HS(x, y) \longleftarrow 0$
for $v \in CN$ **do**
  $HS(x, y) \mathrel{+}= (1 - BoundaryScore(v))$
return $HS(x, y)/Counter$

3.5.2 Evolution Transformation

According to equation 6, nodes in the same community are expected to have higher HM score. It is a desirable characteristic, but it might also accelerate the enhancement of fractured community at the very beginning if solely based on the harmony index. Therefore this enhancement ability of link prediction is still insufficient. We need to control the evolutionary process by community size to determine whether it is time to adding reinforcing edges or not.

Firstly we explain the two stages of HAP method in detail. As illustrated in figure 1, the enhancement procedure can be separated into two processes. The first process is to add links between fractured communities, which adds *revising edges* to mend broken communities. After that, the enhancement measure tends to strengthen those relatively large-scale communities by adding *reinforcing edges* to reinforce the outputs of community detection algorithms.

In order to automatically transform the enhancement procedure, the variables in such process need to be carefully designed. Following the demonstration of figure 1, the minimum community size escalates. It can be seen that in figure 3, comparing with the connective frequency between different community in the ground truth result, the connection among fractured components from Infomap are much more frequent. The merging of small communities should have higher priority at the beginning. For two nodes x, y with different community labels, HAP method tends to add link between them if the communities they belong to are relatively frequent to connect with each other.

Here we define Community Size Attribute (CSA):

$$CSA(x,y) = \begin{cases} \frac{CM(C_x,C_y)}{\min\{CM(C_x,C_x),CM(C_y,C_y)\}}, & C_x \neq C_y \\ \frac{\sqrt{N_{C_x}}}{\max_i N_{C_i}}, & C_x = C_y \end{cases} \qquad (7)$$

CM is the connection matrix. Equation 7 calculates the ratio between inter edges and the intra edges for smaller communities when $C_x \neq C_y$.

Algorithm 3: Community Size Attribute

Input: Graph $G = (V, E)$, Community Mapping Function C_M and node x, y.
Output: CSA value $CSA(x, y)$.
$C_x \longleftarrow C_M(x)$
$C_y \longleftarrow C_M(y)$
if $C_x = C_y$ **then**
  $CM \leftarrow$ ZeroMatrix
  for $e = (i, j) \in E$ **do**
    $CM[C_M(i), C_M(j)] \mathrel{+}= 1$
  if $CM[C_x, C_x] \leq CM[C_y, C_y]$ **then**
    $CSA = CM[C_x, C_y]/CM[C_x, C_x]$
  if $CM[C_x, C_x] > CM[C_y, C_y]$ **then**
    $CSA = CM[C_x, C_y]/CM[C_y, C_y]$
if $C_x \neq C_y$ **then**
  $SizeOfCommunity(SoC) \longleftarrow$ ZeroArray
  for $v \in V$ **do**
    $SoC[C_M(v)] \mathrel{+}= 1$
  $Scale \longleftarrow$ LargestValue(SoC)
  $CSA = \sqrt{Soc[C_x]}/Scale$
return CSA

As aforementioned, when $C_x = C_y$ we do not want to enhance the community when it is relatively small, since reinforcing small communities too early might cause improper community detection result. Empirically we set $CSA = \sqrt{N_{C_x}}/\max_i\{N_{C_i}\}$ when $C_x = C_y$, which performs satisfactorily in experiments. The CSA value will yield higher value when the size of community gets larger during the revising process.

Generally, the CSA index can be re-garded as the indicator of connection possibility between communities since it utilizes the community attribute of nodes. On the other hand, the HM index explicitly points out the detail about

which pair of nodes should build connection. Combing those two indexes together could achieve the community enhancement measure with desired automatic transform of procedure.

Following this, combine equation6 and 7, the similarity function of HAP method is listed below:

$$HAP(x,y) = CSA(x,y) \cdot HM(x,y)$$

Algorithm 4: HAP Method

Input: Graph $G = (V, E)$, Community Mapping Function C_M and Number of Adding L.
Output: New Graph $G = (V, E')$.
$NonExistEdges(NEE) \longleftarrow \emptyset$
for $i \in V$ **do**
  for $j \in V$ **do**
    if $(i,j) \notin E$ **then**
      $CEE \longleftarrow CEE \cup (i,j)$
$HAP\ Scores \longleftarrow \emptyset$
for $e = (i,j) \in NEE$ **do**
  $HS \longleftarrow$ Harmony Score Calculation(G, C_M, i, j)
  $CSA \longleftarrow$ Community Size Attribute(G, C_M, i, j)
  $HAP \leftarrow HS \cdot CSA$
  $HAP\ Scores \longleftarrow HAP\ Scores \cup (HAP,\ e)$
$HAP\ Scores \longleftarrow$ Descend Order$(HAP\ Scores)$
$Counter \longleftarrow 0$
$E' \longleftarrow E \cup$ Top L Edges In$(HAP\ Scores)$
return $G = (V, E')$

With all the preparation work, the community enhancement process is demonstrated in algorithm 5. In the experimental section, we will replace the HAP method with other link prediction methods to verify its validity.

Algorithm 5: Community Enhancement Method

Input: Graph $G = (V, E)$, Community Detection Method CDM, Number of Adding L, Number of Interation N.
Output: Community Mapping Function C_M.
$G_{old} \longleftarrow G$
$C_{old} \longleftarrow CDM(G_{old})$
for $n \in 1, ..., N$ **do**
  $G_{new} \longleftarrow$HAP Method(G_{old}, C_{old}, L)
  $C_{new} \longleftarrow CDM(G_{new})$
  $G_{old} \longleftarrow G_{new}$
  $C_{old} \longleftarrow C_{new}$
return C_{new}

4 Experiments Results and Analysis

4.1 Datasets

Twelve networks are tested in experimental process, including a network consisting of 62 dolphins in a community living off Doubtful Sound, New Zealand (Dolphins, for short) [47], network of friendships among 34 members of Zachary's karate club (Karate, for short) [48], books about US politics sold by the online bookseller Amazon.com in 2004 (Polbooks, for short) [49], a mapping interactions between Science in Society actors on the Web of 12 European countries (Eurosis, for short), an online hyperlinks network between weblogs on US politics (Polblogs, for short) [50], a network of the relationship between publication and the corresponding word from a dictionary (Cora, for short) [51]. A series of citation network on different sub domains, listed as Cora Artificial Intelligence(Cora_AI, for short), Cora Human Computer Interaction (Cora_HCI, for short) and Cora Operating Systems (Cora_OS, for short) [52], three subsets of the 20 newsgroups dataset which comprise around 18000 newsgroups posts on 20 topics (News_1, News_2, News_3, for short) [53].[1]

Details of the twelve datasets are listed in table 1. The column of Transitivity indicates the fraction of close triangles in the network system, which is an indicator of connectivity. N_C stands for the ground truth number of clusters in the network. To demonstrate the community structure, the index of *Intra* is defined as the ratio of edges within communities. *L* is the hyperparameter of edge increment for each iteration, which is approximately proportional to the number of edges in the network system.

TABLE 1
Information of Twelve Real-world Networks

Network	#Nodes	#Edges	Transitivity	N_C	Intra	L
Dolphins	62	159	0.30878	2	0.96226	10
Karate	34	78	0.25568	2	0.87179	10
Polbooks	105	441	0.34840	3	0.84127	20
Eurosis	1272	6454	0.23478	13	0.82290	100
Polblogs	1222	16717	0.22596	2	0.90578	200
Cora	2458	5069	0.09003	7	0.80410	100
Cora_AI	4633	12985	0.15621	11	0.82636	150
Cora_HCI	1053	2350	0.17730	5	0.96213	50
Cora_OS	2068	8645	0.13664	4	0.82852	120
News_1	398	3347	0.42667	2	0.93188	50
News_2	598	5041	0.36420	3	0.80401	100
News_3	595	4557	0.35152	3	0.85561	100

4.2 Baseline Methods and Evaluation Measures

4.2.1 Baseline Methods

The kernel of HAP method is the similarity paradigm link prediction algorithm. In order to verify the performance of proposed HAP method, it is compared with six other link prediction methods, including JA [54], PA [55], CN [20], CN1 [36], RA [46] and RA1 [36].

- PA(Preferential Attachment): $s(x,y) = |\Gamma(x)| \times |\Gamma(y)|$. It is solely based on degrees under the assumption that higher degree nodes tend to connect each other and independent with network's topology information.
- CN1(Adjusted Common Neighbour): Different from CN, CN1 takes clustering information into consideration. It gets a bonus point if the common neighbours of nodes x and y belong to the same community with them.
- RA1(Adjusted Resource Allocation): Same as CN1, the numerator term in RA gets additional point for

1. All datasets are available on https://github.com/vlivashkin/community-graphs

common neighbours being in the same community with nodes x and y.

The community detection algorithms concerned here are Infomap, Louvain and Label-Propagation, all of which have been discussed previously. All seven link prediction methods are aggregated with those three community detection algorithms to get twenty-one combination, which are tested on twelve real-world datasets. Further details will be provided in the following subsection.

4.2.2 Evaluation Measures

In order to quantify the quality of community detection result, as well as measuring the improvement of graph enhancement, the widely used Normalized Mutual Information (NMI) evaluation methods will be introduced here:

$$NMI(X,Y) = \frac{MI(X,Y)}{\sqrt{H(X) \times H(Y)}}$$

where MI [56] stands for Mutual Information score and H is the Shannon Entropy function. If we take X as the algorithm output and Y is the ground truth community structure, value 1 means perfect correlation while 0 stands for no mutual information. That is to say, higher NMI values mean better community detection results.

In addition, the correction ability of link prediction methods is another desirable characteristic. To quantify this ability, the fraction of revising edges will be discussed here. Furthermore, to express the automatic transition between two stages, the dynamic about fraction of revising and reinforcing edges will be demonstrated. Last but not least, to test the generalization ability and numerical stability, the difference between the original NMI value and the final output NMI value will be compared.

Meanwhile, we illustrate the average fraction of revising edges of all link prediction algorithms among three different community detection methods. Since the damping factor $\sqrt{N_{C_x}}$ in equation 7 is clearly not suitable for all situations, only part of real-world datasets about the dynamic between two stages will be demonstrated.

4.3 Evaluation of NMI

In this section, we demonstrate the numerical experimental results on twelve networks. Since the ground truth community structure behind those networks are all available, we compare the algorithm output with the ground truth by applying the NMI measure. The result of best NMI performance of all link prediction methods among three different community detection algorithms are demonstrated in figure 5.

For each network, we independently apply the different twenty-one combinations of link prediction methods with community detection algorithms and repeat the link

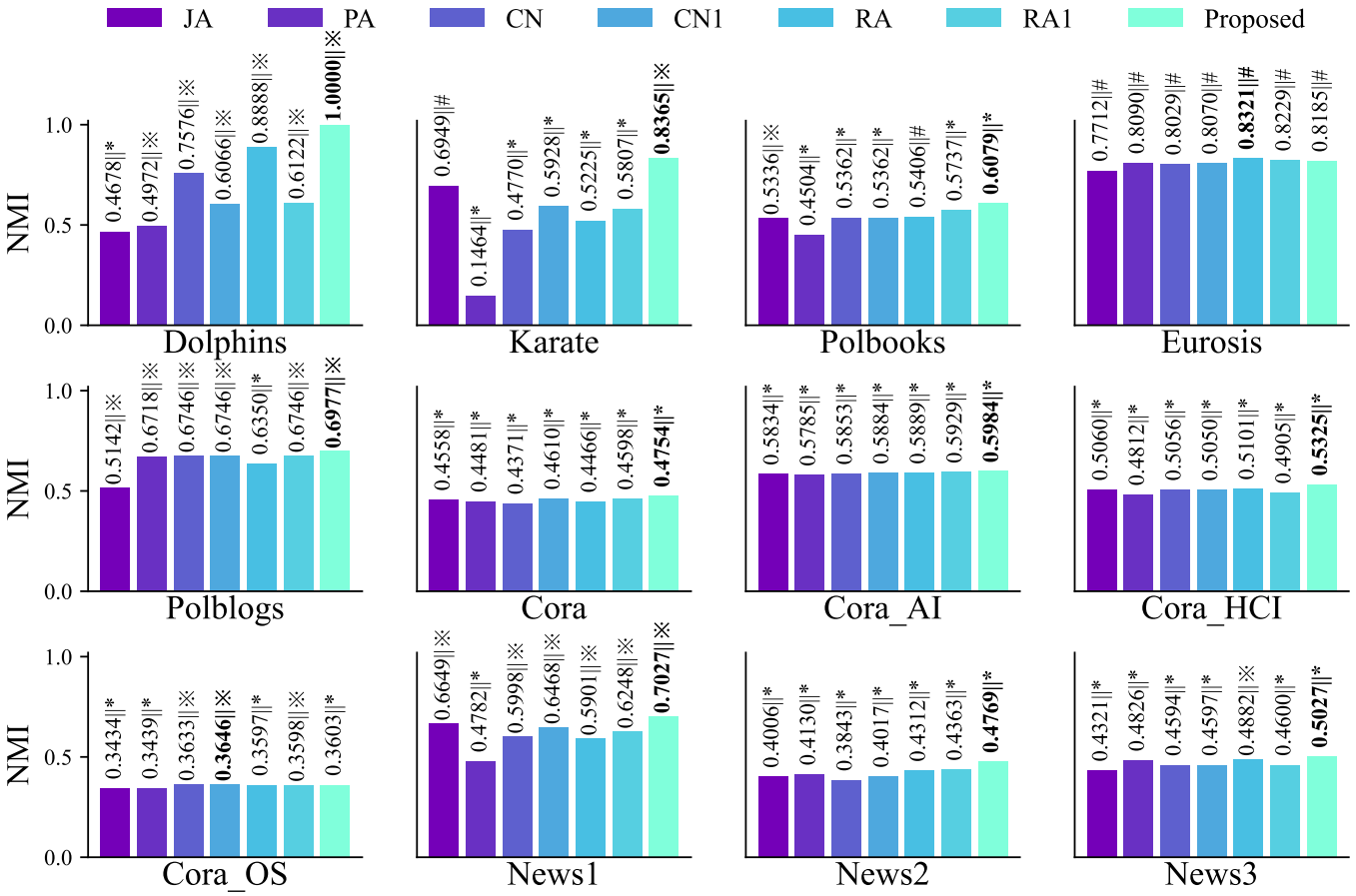


Fig. 5. Best NMI performance of all link prediction methods, labels after || stand for best performance achieved by Louvain(∗), Infomap(#) and LPA(※). Furthermore, we do not implement Louvain algorithm on Eurosis dataset because the experimental results in figure 2 already demonstrated that the Lovain method achieved satisfying result and the community enhancement task wouldn't be significant.

TABLE 2
NMI ranking of seven link prediction methods

	Dolphins	Karate	Polbooks	Eurosis	Polblogs	Cora	Cora_AI	Cora_HCI	Cora_OS	News_1	News_2	News_3	Mean
JA	7	2	6	7	7	4	6	3	7	2	6	7	5.33
PA	6	7	7	4	5	5	7	7	6	7	4	3	5.67
CN	3	6	4	6	2	7	5	4	2	5	7	6	4.75
CN1	5	3	4	5	2	2	4	5	**1**	3	5	5	3.67
RA	2	5	3	**1**	6	6	3	6	5	6	3	2	3.67
RA1	4	4	2	2	2	3	2	2	4	4	2	4	3.25
HAP	**1**	**1**	**1**	3	**1**	**1**	**1**	**1**	3	**1**	**1**	**1**	1.33

adding and clustering for ten iterations as the community enhancement measure. The results in figure 5 are the best performance for each link prediction methods among three different community detection algorithms. The corresponding ranking is listed in table 2. As we can see in table 2, our proposed HAP link prediction method has leading performance in 10 out of 12 networks, especially in the dolphins network where it reaches the maximum value of NMI.

Not limited to the best performance of NMI values, the improvement on each real-world datasets are still significant. Since there remains uncertainty in the community detection section, several link prediction methods might get leading performance since they acquire better community detection result at the beginning of community enhancement process (demonstrated in algorithm 5). Here we provide such information in figure 6, where red part indicates the final result is decreased while green part shows the improvement between the initial partition result and the terminal output after the final round of graph enhancement.

From figure 6 we could find out that our proposed methods have top-2 performance in most cases (11 out of 12). In addition, our proposed HAP method achieves improvement in performance on all datasets. The corresponding ranking is demonstrated in table 3. Furthermore, comparing table 2 with table 3, we can verify our hypothesis such as the main contribution for RA1 measure's higher NMI values originates in better initial partition result rather than its community enhancing ability.

Furthermore, observing the average ranks of all baseline

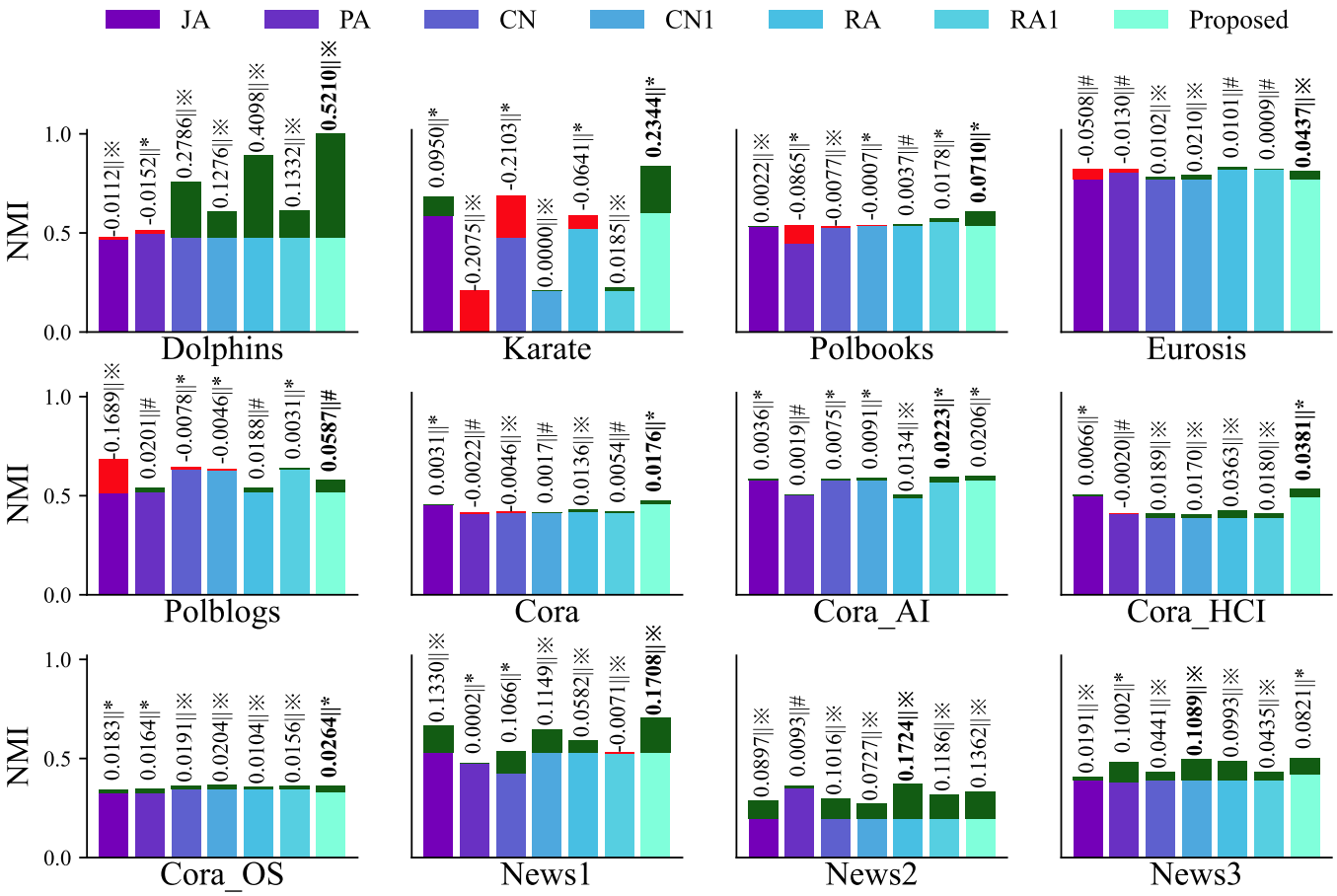


Fig. 6. Largest improvement for each combination, labels after || stand for best performance achieved by Louvain(∗), Infomap(#) and LPA(⋇). The Louvain community detection on Eurosis dataset is also excluded.

TABLE 3
Best NMI improvement ranking of seven link prediction methods

	Dolphins	Karate	Polbooks	Eurosis	Polblogs	Cora	Cora_AI	Cora_HCI	Cora_OS	News_1	News_2	News_3	Mean
JA	6	2	4	7	7	4	6	6	4	2	5	7	5.00
PA	7	6	7	6	2	6	7	7	5	6	7	2	5.67
CN	3	7	6	3	6	7	5	3	3	4	4	5	4.67
CN1	5	4	5	2	5	5	4	5	2	3	6	**1**	3.92
RA	2	5	3	4	3	2	3	2	7	5	**1**	3	3.33
RA1	4	3	2	5	4	3	**1**	4	6	7	3	6	4.00
HAP	**1**	**1**	**1**	**1**	**1**	**1**	2	**1**	**1**	**1**	2	4	1.42

methods in table 2 and 3, we find that the RA family (RA, RA1) generally performs better than the CN family (CN, CN1). Considering the characteristics and intuitions behind those link prediction methods, this finding might indicate that the information flow paradigm (RA) might be more appropriate than the neighbourhood overlapping index (CN) in the scene of community enhancement in real-world data.

4.4 Evaluation of Revising Edges

Taking a step further to explore the reason of such leading performance by HAP method, we need to consider the role of revising edges in the community enhancement task. According to inductive biases, it is believed that the connection of revising edges will deeply affect the community detection methods and lead those algorithms to better NMI performance. As shown in table 4, each value indicates the average fraction of revising edges among three community detection methods for the corresponding link prediction method. For example, if the accumulating fraction of CN index in complex network system on three community detection methods are $6, 9$ and 12, the corresponding value in this table is $(6 + 9 + 12)/3 = 9$. The best value on each network is highlighted.

As can be seen from table 4, the proposed algorithm can achieve the best performance in all datasets with overwhelmingly highest fraction of revising edges. Additionally, comparing the CN and RA index with their community attributes version CN1 and RA1, the experimental results on revising fraction elucidate that improper consideration of community attributes will bring damage upon the correction ability of enhancement measure.

Not only the revising edges, the reinforcing edges also play an important role in mending fractured communities.

TABLE 4
Fraction of revising edges for all methods on 12 real world networks

Methods	Dolphins	Karate	Polbooks	Eurosis	Polblogs	Cora	Cora_AI	Cora_HCI	Cora_OS	News_1	News_2	News_3
JA	6.33	9.00	5.67	3.60	1.58	5.87	4.47	6.00	7.43	1.67	4.10	6.47
PA	15.33	6.33	6.83	0.70	0.95	5.47	7.93	43.73	20.90	23.93	14.27	14.07
CN	18.67	7.33	0.00	0.35	0.05	3.47	1.56	4.60	9.30	2.93	5.43	13.43
CN1	4.33	4.00	0.00	0.00	0.00	0.13	0.07	0.20	1.10	0.00	0.00	0.33
RA	32.33	12.00	5.67	5.60	1.00	23.73	19.64	21.95	23.90	11.80	12.30	14.93
RA1	12.67	8.00	0.50	0.40	0.13	2.90	1.29	2.07	2.93	0.00	0.07	0.17
HAP	**82.67**	**21.00**	**59.33**	**53.20**	**54.25**	**78.33**	**74.20**	**90.00**	**88.47**	**87.67**	**57.67**	**48.57**

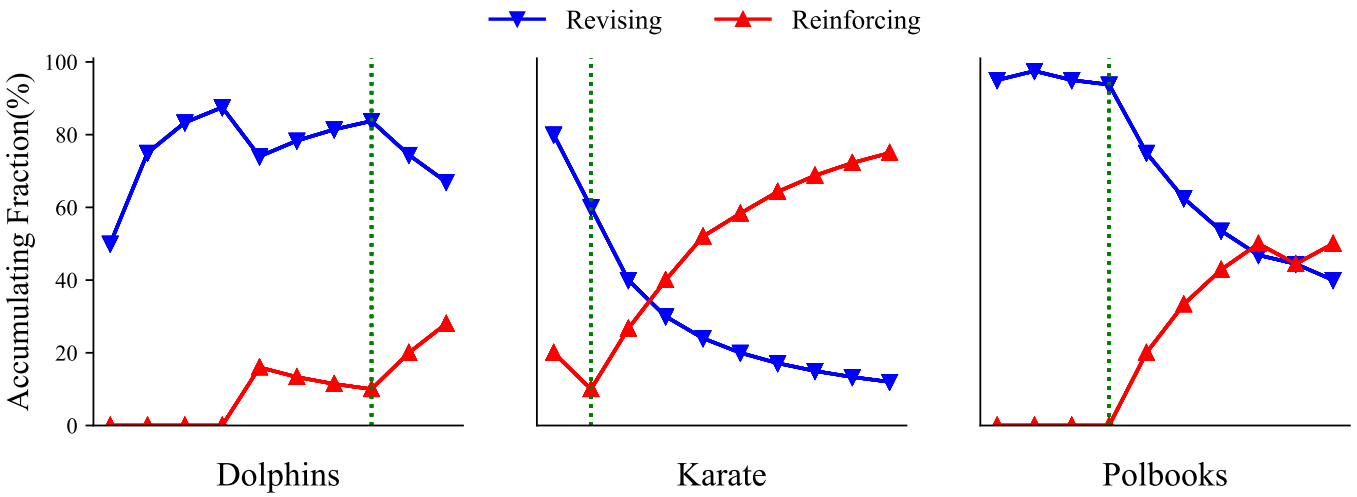


Fig. 7. The accumulating fraction of revising edges and reinforcing edges, the difference between two consecutive points indicates the incremental of corresponding type of edges, the green vertical dash line unveils the auto transformation between two processes.

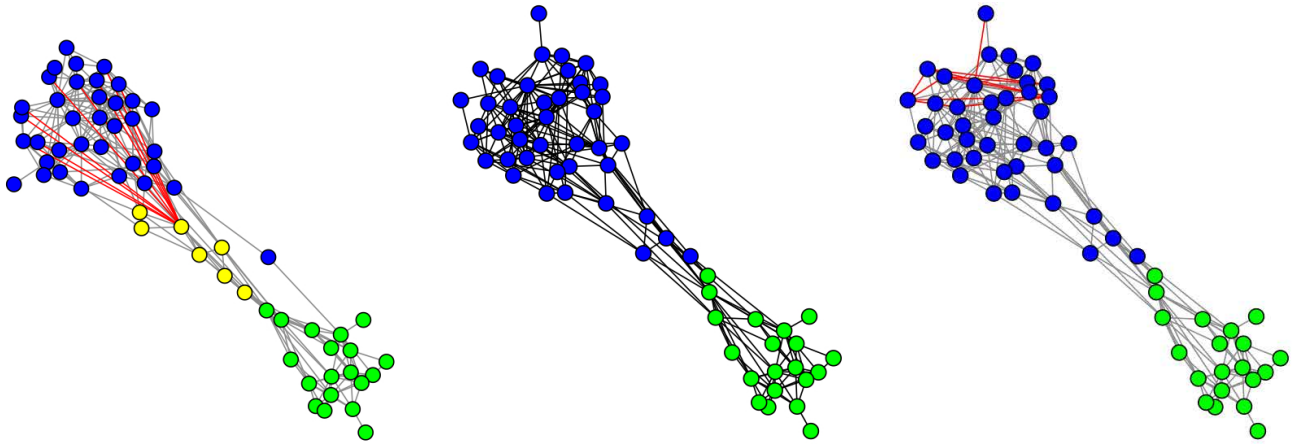


Fig. 8. Three snapshots of the experimental procedure of Dolphin network, red line indicates the links that appended in corresponding round. Subfigure 8a to subfigure 8b represents the first stage of graph enhancement, which makes connection between fractured components and turns them into a complete community. While from subfigure 8b to subfigure 8c shows the reinforcing process to enhance the correct community detection result.

Thus we design our HAP method with evolution process to transform from adding revising edges to reinforcing edges. The transformation has been successfully captured in several datasets. Demonstrated in figure 7, it can be seen that the link prediction method has the ability to automatically transform from revising to reinforcing process. The illustration of transformation about HAP method on Dolphin network is presented in figure 8, where the evolutionary process between two stages is clearly captured.

Last but not least, table 5 shows the sign of difference between the original NMI value and the final output NMI value. In this table, the improvement is labelled with the sign + in red. It can be seen that our HAP method has the greatest application stability and suits for 17 out of 20 cases.

5 Conclusion and Future Work

In this paper we propose the HAP method to fill the void of link prediction based network community enhancement method. Combing the HM index and the CS index, it can be treated as a iterative method which first determines the connection in global perspective (community level) and then focuses into the neighbourhood local information (node level) at each iteration. It has desirable portability and simplicity with low computation cost. Compared with other baseline methods on real-world datasets, we find that our proposed novel index has better performance in most cases. Furthermore, the HAP method does not require any prior knowledge about the distribution or number of communities. Finally, thanks to the iteration paradigm, all local link prediction approaches are no longer bounded by the two hop distance.

There still remain some works that need to be further studied. In this paper we only consider adding edges to the network system, but the removal of existing edges also needs to be concerned. And the square root term in the similarity function of proposed method is an empirical modification term which cannot suit all situations. In future work, we will further explore the preprocessing ability of HAP method in the circumstances where nodes have additional attributes like the node classification tasks or the graph classification tasks.

Acknowledgments

This work was supported by the Research and Development Program of China (Grant No. 2018AAA0101100), the National Natural Science Foundation of China (Grant Nos. 62141605, 62050132), the Beijing Natural Science Foundation (Grant Nos. 1192012, Z180005).

References

[1] S. H. Strogatz, "Exploring complex networks," *nature*, vol. 410, no. 6825, pp. 268–276, 2001.
[2] S. Deng, L. Huang, J. Taheri, J. Yin, M. Zhou, and A. Y. Zomaya, "Mobility-aware service composition in mobile communities," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 3, pp. 555–568, 2016.
[3] Z. Wang, D. Zhang, X. Zhou, D. Yang, Z. Yu, and Z. Yu, "Discovering and profiling overlapping communities in location-based social networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 4, pp. 499–509, 2013.
[4] S. Qiao, N. Han, Y. Gao, R.-H. Li, J. Huang, H. Sun, and X. Wu, "Dynamic community evolution analysis framework for large-scale complex networks based on strong and weak events," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 10, pp. 6229–6243, 2020.
[5] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical review E*, vol. 69, no. 6, p. 066133, 2004.
[6] 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.
[7] A. C. Lewis, N. S. Jones, M. A. Porter, and C. M. Deane, "The function of communities in protein interaction networks at multiple scales," *BMC systems biology*, vol. 4, no. 1, pp. 1–14, 2010.
[8] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.
[9] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *The Bell system technical journal*, vol. 49, no. 2, pp. 291–307, 1970.
[10] P. R. Suaris and G. Kedem, "An algorithm for quadrisection and its application to standard cell placement," *IEEE Transactions on Circuits and Systems*, vol. 35, no. 3, pp. 294–303, 1988.

TABLE 5
Experimental result of 12 real-world datasets on twenty-one combinations of link prediction methods and community detection methods. Sings in this table show the difference between original and final NMI values.

LP	CD	Dolphins	Karate	Polbooks	Eurosis	Polblogs	Cora	Cora_AI	Cora_HCI	Cora_OS	News_1	News_2	News_3
JA	Louvain	-	+	-	\	-	+	+	+	+	-	-	-
	Infomap	-	-	-	-	-	-	-	-	+	-	+	-
	LPA	-	+	-	-	-	-	-	-	-	+	+	+
RA	Louvain	-	-	-	\	+	-	-	-	+	+	-	+
	Infomap	-	-	-	-	+	-	+	-	+	-	+	-
	LPA	-	-	-	-	-	-	-	-	-	-	-	-
CN	Louvain	+	-	-	\	-	-	+	-	-	+	-	+
	Infomap	+	-	-	-	-	-	-	+	-	-	+	-
	LPA	+	-	-	+	-	-	+	+	+	+	+	+
CN1	Louvain	+	-	-	\	-	+	+	+	-	+	-	+
	Infomap	-	-	-	-	-	+	-	+	+	-	+	-
	LPA	+	=	-	+	-	-	+	+	+	+	+	+
RA	Louvain	-	-	-	\	-	+	+	+	+	-	+	-
	Infomap	-	-	+	+	+	+	+	+	+	+	-	+
	LPA	+	-	-	+	-	+	+	+	+	+	+	+
RA1	Louvain	-	-	+	\	+	-	+	-	-	+	-	-
	Infomap	+	-	+	+	+	+	+	+	+	-	+	-
	LPA	+	+	-	+	-	+	+	+	+	+	+	+
HAP	Louvain	+	+	+	\	+	+	+	+	+	+	+	+
	Infomap	+	+	-	-	+	+	+	+	+	+	+	+
	LPA	+	+	+	+	+	+	-	+	+	+	+	+

[11] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[12] C.-H. Mu, J. Xie, Y. Liu, F. Chen, Y. Liu, and L.-C. Jiao, "Memetic algorithm with simulated annealing strategy and tightness greedy optimization for community detection in networks," *Applied Soft Computing*, vol. 34, pp. 485–501, 2015.

[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[14] W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs," in *Selected Papers Of Alan J Hoffman: With Commentary*. World Scientific, 2003, pp. 437–442.

[15] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[16] A. Mahmood and M. Small, "Subspace based network community detection using sparse linear coding," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 801–812, 2015.

[17] J. Reichardt and S. Bornholdt, "Detecting fuzzy community structures in complex networks with a potts model," *Physical review letters*, vol. 93, no. 21, p. 218701, 2004.

[18] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical review E*, vol. 76, no. 3, p. 036106, 2007.

[19] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: statistical mechanics and its applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[20] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[21] 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.

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

[23] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

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

[25] R.-H. Li, J. X. Yu, and J. Liu, "Link prediction: the power of maximal entropy random walk," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 1147–1156.

[26] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Springer, 2015, vol. 72.

[27] J. Lu, T. Zhang, F. Hu, and Q. Hao, "Preprocessing design in pyroelectric infrared sensor-based human-tracking system: On sensor selection and calibration," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 2, pp. 263–275, 2016.

[28] M. Burgess, E. Adar, and M. Cafarella, "Link-prediction enhanced consensus clustering for complex networks," *PloS one*, vol. 11, no. 5, p. e0153384, 2016.

[29] J. You, J. Gomes-Selman, R. Ying, and J. Leskovec, "Identity-aware graph neural networks," *arXiv preprint arXiv:2101.10320*, 2021.

[30] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *arXiv preprint arXiv:1810.00826*, 2018.

[31] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the national academy of sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[32] Z.-Y. Zhang, K.-D. Sun, and S.-Q. Wang, "Enhanced community structure detection in complex networks with partial background information," *Scientific reports*, vol. 3, no. 1, pp. 1–7, 2013.

[33] L. Yang, D. Jin, X. Wang, and X. Cao, "Active link selection for efficient semi-supervised community detection," *Scientific reports*, vol. 5, no. 1, pp. 1–12, 2015.

[34] Y. Su, C. Liu, Y. Niu, F. Cheng, and X. Zhang, "A community structure enhancement-based community detection algorithm for complex networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 5, pp. 2833–2846, 2019.

[35] J. Zhou, Z. Chen, M. Du, L. Chen, S. Yu, G. Chen, and Q. Xuan, "Robustecd: Enhancement of network structure for robust community detection," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[36] S. Soundarajan and J. Hopcroft, "Using community information to improve the precision of link prediction methods," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 607–608.

[37] J. C. Valverde-Rebaza and A. d. Andrade Lopes, "Link prediction in complex networks based on cluster information," in *Brazilian Symposium on Artificial Intelligence*. Springer, 2012, pp. 92–101.
[38] J. Valverde-Rebaza and A. de Andrade Lopes, "Structural link prediction using community information on twitter," in *2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN)*. IEEE, 2012, pp. 132–137.
[39] J. Ai, Y. Liu, Z. Su, H. Zhang, and F. Zhao, "Link prediction in recommender systems based on multi-factor network modeling and community detection," *EPL (Europhysics Letters)*, vol. 126, no. 3, p. 38003, 2019.
[40] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: a discriminative approach," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 927–936.
[41] M. Chen, A. Bahulkar, K. Kuzmin, and B. K. Szymanski, "Improving network community structure with link prediction ranking," in *Complex Networks VII*. Springer, 2016, pp. 145–158.
[42] C. De Bacco, E. A. Power, D. B. Larremore, and C. Moore, "Community detection, link prediction, and layer interdependence in multilayer networks," *Physical Review E*, vol. 95, no. 4, p. 042317, 2017.
[43] H. Jiang, Z. Liu, C. Liu, Y. Su, and X. Zhang, "Community detection in complex networks with an ambiguous structure using central node based link prediction," *Knowledge-Based Systems*, vol. 195, p. 105626, 2020.
[44] N. Biggs, E. K. Lloyd, and R. J. Wilson, *Graph Theory, 1736-1936*. Oxford University Press, 1986.
[45] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
[46] L. Lü, 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.
[47] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, 2003.
[48] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977.
[49] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.
[50] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*, 2005, pp. 36–43.
[51] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI magazine*, vol. 29, no. 3, pp. 93–93, 2008.
[52] L. Šubelj and M. Bajec, "Model of complex networks based on citation dynamics," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 527–530.
[53] L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens, "Graph nodes clustering based on the commute-time kernel," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2007, pp. 1037–1045.
[54] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.
[55] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.
[56] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

Qiming Yang received the B.Sc. degree from Beihang University, Beijing, China, in 2021. He is currently pursuing the master's degree with the School of Mathematical Sciences, Beihang University, Beijing, China. His current research interests include complex networks, link prediction and graph representation learning.

Wei Wei received the Ph.D. degree in mathematics from the School of Mathematical Sciences, Peking University, Beijing, China, in 2009. He is currently an Associate Professor with the School of Mathematical Sciences, Beihang University, Beijing, China. His research interests include dynamical system and complexity, complex networks, and artificial intelligence.

Ruizhi Zhang is currently pursuing the Ph.D. degree with the School of Mathematical Sciences, Beihang University, Beijing, China. Her current research interests include complex networks, link prediction and graph representation learning.

Bowen Pang received the B.Sc. degree from Beihang University, Beijing, China, in 2020. He is currently pursuing the master's degree with the School of Mathematical Sciences, Beihang University, Beijing, China. His current research interests include complex networks, neural networks, deep learning and time series analysis.

Xiangnan Feng received the Ph.D. degree in mathematics from the School of Mathematical Sciences, Beihang University, Beijing, China, in 2021. He is currently a Postdoctoral Fellow with the Max Planck Institute for Human Development, Berlin, Germany. His research interests include complex networks, computing social science, and artificial intelligence.