Collective Prediction of Multiple Types of Links in Heterogeneous Information Networks

Bokai Cao
Department of Computer Science
University of Illinois at Chicago
Chicago, IL, USA
caobokai@uic.edu

Xiangnan Kong
Department of Computer Science
Worcester Polytechnic Institute
Worcester, MA, USA
xkong@wpi.edu

Philip S. Yu
Department of Computer Science
University of Illinois at Chicago
Chicago, IL, USA
psyu@cs.uic.edu

***Abstract*—Link prediction has become an important and active research topic in recent years, which is prevalent in many real-world applications. Current research on link prediction focuses on predicting one single type of links, such as friendship links in social networks, or predicting multiple types of links independently. However, many real-world networks involve more than one type of links, and different types of links are not independent, but related with complex dependencies among them. In such networks, the prediction tasks for different types of links are also correlated and the links of different types should be predicted collectively. In this paper, we study the problem of collective prediction of multiple types of links in heterogeneous information networks. To address this problem, we introduce the *linkage homophily principle* and design a relatedness measure, called RM, between different types of objects to compute the existence probability of a link. We also extend conventional proximity measures to heterogeneous links. Furthermore, we propose an iterative framework for heterogeneous collective link prediction, called HCLP, to predict multiple types of links collectively by exploiting diverse and complex linkage information in heterogeneous information networks. Empirical studies on real-world tasks demonstrate that the proposed collective link prediction approach can effectively boost link prediction performances in heterogeneous information networks.**

***Index Terms*—collective link prediction; heterogeneous information networks; meta path.**

I. Introduction

Network analysis, especially link prediction, is prevalent in a wide range of application domains. Examples include recommender systems, gene analysis, *etc.* In these applications, the potential existence of unknown links representing a particular relationship need to be predicted. For example, in gene-disease association prediction, different gene sequences can lead to certain diseases. Researchers would like to predict the association relationships between genes and diseases. In drug-target binding prediction, different chemical compounds can bind with certain gene targets. In order to discover new drugs for diseases, researchers are interested in predicting the binding relationships between genes and chemical compounds.

Link prediction has been extensively studied in the literature [19], [8], [20]. Conventional research on link prediction mainly focuses on homogeneous information networks which are composed of one single type of objects and links. Recent studies extend to study the link prediction problem in heterogeneous information networks [25], [6], [16], [7], where multiple types of objects are interconnected through multiple types of links. Most of existing studies either predict one single type of links or independently predict multiple types of links. However, in many real-world applications, we need to predict multiple types of links within a heterogeneous information network, and the prediction tasks for multiple types of links are correlated.

For example, in drug discovery networks, we usually need to predict multiple types of links within a heterogeneous information network, such as gene-disease association links and drug-target binding links. Moreover, different types of links are correlated, and should be predicted collectively instead of independently. Gene-disease association relationship is correlated with drug-target binding relationship, since there are diverse and complex relationships connecting gene, disease, chemical compounds, and related objects. The tasks of predicting drug-target binding links and gene-disease association links are closely related, because any discovery of a certain association between gene sequences and a particular disease can provide important clues about developing the corresponding drugs for such disease, vice versa. Therefore, in order to capture such inter-link dependencies, multiple types of links need to be predicted collectively.

In this paper, we study the problem of heterogeneous collective link prediction, which corresponds to collectively predicting the potential existence of multiple types of links in a heterogeneous information network. If we consider collective link prediction and heterogeneous information networks as a whole, the major research challenges of this paper can be summarized as the following aspects.

First, how can we compute the existence possibility of a link between different types of objects? Conventional approaches for link prediction mainly consider one single type of relationships in the data. However, heterogeneous information networks can encode diverse and complex relationships among objects, involving multiple types of source node correlations and target node correlations. For example, two chemical compounds can be considered similar due to (1) sharing of common substructure, (2) treating the same disease, or (3) binding to the same gene, *etc.* Each of these types of similarities or correlations corresponds to a different linkage or path relationship in Figure 2. That is to say, similar

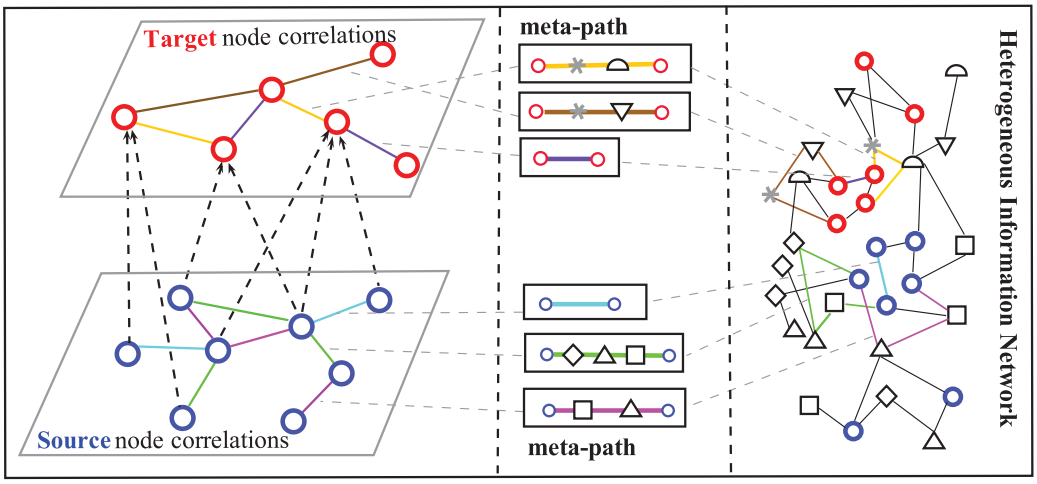


Fig. 1. Collective link prediction in heterogeneous information networks. Different types of nodes are drawn as different shapes. The relationships from blue circles (the source nodes) to red circles (the target nodes) are the links to be predicted. Each line in different colors represents a particular correlation among the source nodes or the target nodes, extracted from the heterogeneous information network using meta-paths.

node sets representing such correlations can be discovered through meta-path based connections [25]. In order to investigate the potential existence of a link, we introduce the *linkage homophily principle* by referring to existing linkage information between the similar nodes of the source node and those of the target node. Different from previous work where proximity measures for link prediction are defined on one single type of nodes (*e.g.*, common neighbors, Jaccard's coefficient, preferential attachment) [19], [25], we design a relatedness measure between a pair of nodes which can be of different types. We also extend some representative proximity measures to heterogeneous links [26].

Second, how can we simultaneously predict multiple types of links in heterogeneous information networks to capture the inter-link dependencies? It is usually assumed that the prediction tasks for multiple types of links are independent which may not hold in some real-world applications. To tackle this problem, inspired by co-training [2], we try to combine labeled and unlabeled data in heterogeneous information networks, where the label concepts are different types of links. Instead of directly exchanging their most confident predictions on unlabeled data, we consider adding the most confidently predicted links of each type into the network schema. Therefore, an iterative framework is proposed to effectively address the problem of collective prediction of multiple types of links by leveraging the complementary prediction information from different types of links.

In this paper, we define the problem of collective link prediction in heterogeneous information networks. We observe that meta-path is defined as a type of path containing a certain sequence of link types [27], which can be a good tool to extract diverse and complex relationships among objects in heterogeneous information networks, since different meta-paths usually represent different correlations among connected objects with different semantic meanings, as shown in Figure 1. Moreover, we introduce the *linkage homophily principle*, based on which we design a relatedness measure, called RM, that can characterize the existence probability of a link between different types of nodes. We also extend conventional proximity measures to heterogeneous links. Furthermore, we propose a general iterative framework for heterogeneous collective link prediction, called HCLP, to predict multiple types of links collectively by capturing the diverse and complex relationships among different types of links and leveraging the complementary prediction information. It is demonstrated through experiments on real-world tasks that our proposed method can significantly improve the performance of collective link prediction in heterogeneous information networks.

For the rest of the paper, we first state the problem of heterogeneous collective link prediction and introduce data collections and related notations in section II. Then we introduce the relatedness measure in section III-B and an iterative framework in section III-C. Experimental results are discussed in section IV. In section VI, we conclude the paper.

II. Problem Definition

In this section, we first introduce related concepts and notations, then define the problem, and introduce the dataset used in this paper.

A. Heterogeneous Information Network

A heterogeneous information network is a type of information network with multiple types of nodes and multiple types of links [28], [26]. It can be represented as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ denotes the set of nodes, which involves t types of nodes: $\mathcal{V}^1 = \{v_1^1, \cdots, v_{n_1}^1\}, \cdots, \mathcal{V}^t = \{v_1^t, \cdots, v_{n_t}^t\}$, where v_j^i represents the j-th node of type i. $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of links between the nodes in $\mathcal{V}$, which involves m types of links.

Each type of links starting from a source node of type i and ending at a target node of type j corresponds to a binary relation R^{ij}, where R_{pq}^{ij} holds if v_p^i and v_q^j are linked by

a link of type R^{ij}. For example, in Figure 2, the link type "hasTissue" is a relation between *genes* and *tissues*, where R^{ij}_{pq} holds if the p-th *gene* node has a link of type "hasTissue" to the q-th *tissue* node in the network. We can write this link type as "gene$\xrightarrow{hasTissue}$tissue" or "$\mathcal{V}^i \xrightarrow{R^{ij}} \mathcal{V}^j$".

In addition, a relation R^{ij} is mathematically described by a $|\mathcal{V}_i|\times|\mathcal{V}_j|$ weighted adjacency matrix $\mathcal{W}^{ij}$, where $\mathcal{W}^{ij}_{pq} \in [0,1]$ is the possibility that there exists a link of type R^{ij} between v^i_p and v^j_q. Particularly, $\mathcal{W}^{ij}_{pq} = 1$, if there is an existing link between v^i_p and v^j_q. Otherwise, $\mathcal{W}^{ij}_{pq}$ is set to 0 in initialization for all the unknown links.

B. Collective Link Prediction over Multiple Link Types

The task of collective link prediction is to predict the potential existence of multiple types of links, which are correlated. In this paper, we propose to utilize heterogeneous information networks to facilitate the process of collective link prediction.

Given a heterogeneous information network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $\mathcal{E}$ is the set of existing links involving m types of links, then all the remaining unknown links are denoted as $\mathcal{U} = \mathcal{V} \times \mathcal{V} - \mathcal{E}$, from which a set of test links are randomly sampled, $\mathcal{T} \subseteq \mathcal{U}$. The task of collective link prediction is to find a predictive function $f : (\mathcal{V}, \mathcal{E}, \mathcal{T}) \rightarrow \mathcal{Y}$, where $\mathcal{Y} = \{y_1, y_2, \cdots, y_{|\mathcal{T}|}\}, y_i \in \{0,1\}$ is a set of inferred results for whether the test links can exist.

To build a classifier, for each link, a feature vector $\mathbf{x}_i$ should be derived from the network schema or linkage structures. Then, our task reduces to build a model to estimate the probability $P(y_i|\mathbf{x}_i)$. $\mathbf{x}_i$ can be normally designed as to reflect the degree of similarity or closeness of two adjacent nodes of the i-th link. However, computing $\mathbf{x}_i$ is a challenging task in the context of heterogeneous information networks, since conventional proximity measures [19] can not be directly applied here. In section III-B, a novel relatedness measure is proposed to capture diverse and complex relationships between multiple types of nodes, in different semantics.

C. Data Collections

In this paper, we study a heterogeneous information network from bioinformatics. SLAP dataset[1] is a heterogeneous information network composed of over $290K$ nodes and $720K$ edges. It integrates data involving 10 types of nodes, such as *genes*, *diseases*, *chemical compounds*, *etc.*, which are connected through 11 types of links, such as "gene $\xrightarrow{PPI}$ gene", "gene $\xrightarrow{hasTissue}$ tissue", *etc.* Its network schema is shown in Figure 2. There are extensive applications of predicting these types of links in bioinformatics, and several examples are listed in Table I.

[1]SLAP dataset [4]: this dataset is an information network that integrates many datasets into a single framework using Semantic Web technologies for drug discovery. It includes public datasets related to systems chemical biology: such as PubChem, DrugBank, PPI, SIDER, CTD diseases, KEGG Pathways, *etc.*

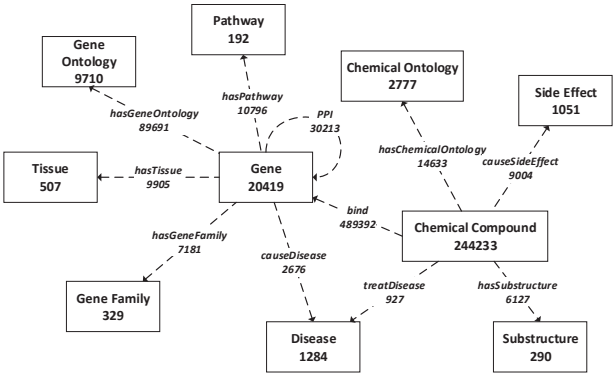


Fig. 2. An example of heterogeneous information network schema (the data schema of SLAP dataset). Each box represents one type of nodes in the network, and each dashed line represents one type of links. The numbers in the figure represent the numbers of nodes/links of different types. [14]

III. Proposed Method

To address the first challenge of how to compute the existence possibility of a link between different types of objects, we first introduce the notion of meta-path in section III-A, based on which we design a novel relatedness measure between a pair of nodes in section III-B. The second challenge of how to simultaneously predict multiple types of links is tackled by our proposed iterative framework for heterogeneous collective link prediction in section III-C.

A. Meta-path

Before exploiting the correlations among multiple types of links, we first consider potential relationships among multiple types of nodes derived from a heterogeneous information network. We briefly review the notion of meta-path following the previous work [26], [17], [15], [27].

In general, a meta-path $\mathcal{P}$ corresponds to a type of path within the network schema, containing a certain sequence of link types. For example, in Figure 2, a meta-path "disease $\xrightarrow{causeDisease^{-1}}$ gene $\xrightarrow{causeDisease}$ disease" denotes a composite relationship between *diseases*, and the semantic meaning of this meta-path is that two *diseases* are caused by a common *gene*. The link type "causeDisease^{-1}" represents the inverted relation of "causeDisease". We consider that *diseases* connected through the meta-path can be regarded as more similar than those without such correlations.

Different meta-paths usually represent different relationships among linked nodes with different semantic meanings. For example, the meta-path "gene $\xrightarrow{PPI}$ gene" denotes two *genes* are connected through "PPI" links, while the meta-path "gene $\xrightarrow{hasTissue}$ tissue $\xrightarrow{hasTissue^{-1}}$ gene" corresponds to the semantic meaning that two *genes* share common *tissues* [14]. In this way, similarity between a type of nodes can be described by different meta-paths with different semantics.

$\mathcal{P}^{ij}$ is defined as a type of meta-paths starting from source nodes of type i and ending at target nodes of type j. Particularly, $\mathcal{P}^{ii}$ (or $\mathcal{P}^{jj}$) is a type of meta-paths starting from and

TABLE I
VARIOUS APPLICATIONS OF LINK PREDICTION IN BIOINFORMATICS.

Predicted Link Type	Application	Reference
Gene $\xrightarrow{PPI}$ Gene	Protein-protein Interaction Prediction	[12]
Gene $\xrightarrow{causeDisease}$ Disease	Prioritization of Candidate Disease Genes	[22]
Gene $\xrightarrow{hasGeneOntology}$ Gene Ontology	Automated Gene Ontology Annotation	[29]
Chemical Compound $\xrightarrow{bind}$ Gene	Drug-Target Interaction Prediction or Drug Repositioning	[5]
Chemical Compound $\xrightarrow{treatDisease}$ Disease	Drug Discovery	[9]
Chemical Compound $\xrightarrow{causeSideEffect}$ Side Effect	Drug Side Effect Profiling	[23]
Chemical Compound $\xrightarrow{hasChemicalOntology}$ Chemical Ontology	Automatic Annotation of Chemical Ontology	[11]

ending at nodes of the same type i (or j). The s-th (or the t-th) meta-path of $\mathcal{P}^{ii}$ (or $\mathcal{P}^{jj}$) is denoted as $\mathcal{P}^{iis}$ (or $\mathcal{P}^{jjt}$). For example, $\mathcal{P}^{iis}$ can be the green line in Figure 1, representing a meta-path connecting the source nodes. Similarly, $\mathcal{P}^{jjt}$ can be the yellow line representing a meta-path connecting the target nodes. In Figure 1, R^{ij} is indicated as the arrows linking from the source nodes to the target nodes. Concatenating $\mathcal{P}^{iis}$, R^{ij} and $\mathcal{P}^{jjt}$ in sequence can compose a meta-path of $\mathcal{P}^{ij}$ going from the source nodes to the target nodes, denoted as $\mathcal{P}^{ijst}$. It can be written as a certain sequence of relations: $R^{k_0k_1}, R^{k_1k_2}, \cdots, R^{k_{n-1}k_n}$, where $k_0 = i$, $k_n = j$ and n is the length of $\mathcal{P}^{ijst}$.

B. Relatedness Measure

Similarity measures for analyzing the proximity of nodes in a network are introduced in previous work [19], [25], which however are defined on one single type of nodes. In real-world applications of heterogeneous information networks, there are multiple types of nodes. Thus, we design a novel relatedness measure between a pair of nodes, which can be of different types.

Let's consider a link type "$\mathcal{V}^i \xrightarrow{R^{ij}} \mathcal{V}^j$". An effective measure function should be designed to reflect the possibility that a link can exist between the source node v_p^i and the target node v_q^j. An intuition is to find a set of nodes of type i similar to v_p^i, and a set of nodes of type j similar to v_q^j, then investigate the existing linkage information across the two sets of nodes. We consider that the more existing links across the similar node set of v_p^i and the similar node set of v_q^j, the more likely that there exists a link between v_p^i and v_q^j, which can be referred to as the following principle.

PRINCIPLE 1. *(Linkage Homophily Principle) Two nodes are more likely to be directly linked if most of their respective similar nodes are linked.*

Normally, similar nodes are defined within neighbors. In heterogeneous information networks, however, similar nodes of the same type may be connected through composite paths instead of direct links. Thus, meta-path based connections can be used to capture such generalized neighbor dependencies, upon which similarity can be defined. Following this idea, we need to select the meta-paths that define the similarity of source nodes and that of target nodes. Assume the source node is v_p^i (*i.e.*, the p-th node of type i) and the selected meta-path is $\mathcal{P}^{iis}$ (*i.e.*, the s-th meta-path of $\mathcal{P}^{ii}$). The set of nodes of type i that are connected to v_p^i via the meta-path $\mathcal{P}^{iis}$ are regarded as its generalized neighbors and denoted as N_p^i. Similarly, assume the target node is v_q^j (*i.e.*, the q-th node of type j) and the selected meta-path is $\mathcal{P}^{jjt}$ (*i.e.*, the t-th meta-path of $\mathcal{P}^{jj}$). The set of nodes of type j that are connected to v_q^j via the meta-path $\mathcal{P}^{jjq}$ are regarded as its generalized neighbors and denoted as N_q^j. The idea is to use the connectivity between the two generalized neighbor sets, N_p^i and N_q^j, to predict the likelihood of linkage between nodes v_p^i and v_q^j.

This concept can be further refined by taking into consideration the similarity of generalized neighbor nodes as a weight. Weighted by the similarity between N_p^i and v_p^i and that between N_q^j and v_q^j, the relatedness measure can be defined as the expected number of links between similar source nodes and similar target nodes, divided by the maximum number of potential links between them.

Let EPC^{iis} denote the similarity matrix of i-th node type along the meta-path $\mathcal{P}^{iis}$, then $EPC^{iis}_{pp'}$ is the similarity between nodes v_p^i and $v_{p'}^i$. As in [26], the similarity can be measured based on path counts. In general, expected path count, EPC^{ijst}, can be computed by the product of weighted adjacency matrices $\mathcal{W}^{k_{p-1}k_p}$ corresponding to each link type $R^{k_{p-1}k_p}$ along the meta-path $\mathcal{P}^{ijst} = \{R^{k_0k_1}, R^{k_1k_2}, \cdots, R^{k_{n-1}k_n}\}$ as follows:

$$EPC^{ijst} = \prod_{p=1}^{n} \mathcal{W}^{k_{p-1}k_p} = EPC^{iis} \times \mathcal{W}^{ij} \times EPC^{jjt}$$

where $\mathcal{P}^{ijst}$ is a composition of $\mathcal{P}^{iis}$, R^{ij} and $\mathcal{P}^{jjt}$. Here, EPC^{ijst} is a $|\mathcal{V}_i| \times |\mathcal{V}_j|$ matrix and EPC^{ijst}_{pq} is the expected number of path instances between v_p^i and v_q^j. EPC^{iis} (or EPC^{jjt}) can be similarly computed and regarded as the similarity measures between nodes of type i (or j).

Now we can formulate our relatedness measure between source nodes of type i and target nodes of type j upon the meta-path $\mathcal{P}^{ijst}$ as follows:

$$RM^{ijst} = \frac{EPC^{ijst}}{EPC^{iis} \times \mathbf{1} \times EPC^{jjt}} = \frac{EPC^{iis} \times \mathcal{W}^{ij} \times EPC^{jjt}}{EPC^{iis} \times \mathbf{1} \times EPC^{jjt}}$$

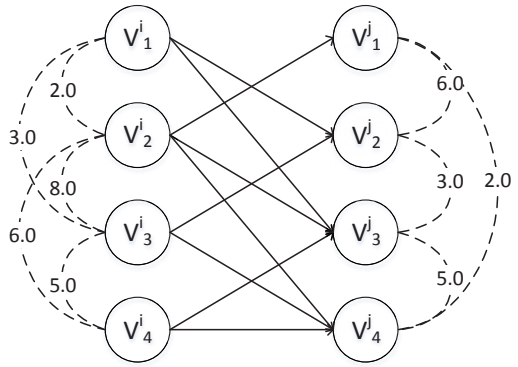


Fig. 3. A simplified example of computing the relatedness measure. The arrows represent existing links between source nodes of type i and target nodes of type j, and the numbers on the dashed lines represent similarity between the same types of nodes (*i.e.*, expected path count derived from meta-paths $\mathcal{P}^{iis}$ and $\mathcal{P}^{jjt}$).

where $\mathbf{1}$ is a $|\mathcal{V}_i| \times |\mathcal{V}_j|$ matrix in which all the elements are 1, and the division means respectively dividing elements in the numerator matrix by elements in the denominator matrix in the corresponding positions. This way, RM^{ijst} is also a $|\mathcal{V}_i| \times |\mathcal{V}_j|$ matrix and RM^{ijst}_{pq} is the relatedness we measure between v^i_p and v^j_q following $\mathcal{P}^{ijst}$.

We give an intuitive interpretation of the formulation that the numerator represents the expected number of links between generalized neighbor sets of a node pair, and the denominator represents the maximum number of potential links between them. Both of them are weighted by the product of similarity between generalized neighbors and the node pair. Here, the maximum number of potential links between two node sets corresponds to the extreme case where all nodes in the two sets are linked in pairs. We can see that the higher RM, the more likely that there can exist a link between the pair of source node and target node.

To be specific, let's have a look at an example shown in Figure 3. We can derive the value of EPC^{iis}, EPC^{jjt} from the figure and compute RM^{ijst} as follows:

$$EPC^{iis} = \begin{pmatrix} 0 & 2 & 3 & 0 \\ 2 & 0 & 8 & 6 \\ 3 & 8 & 0 & 5 \\ 0 & 6 & 5 & 0 \end{pmatrix}, EPC^{jjt} = \begin{pmatrix} 0 & 6 & 0 & 2 \\ 6 & 0 & 3 & 0 \\ 0 & 3 & 0 & 5 \\ 2 & 0 & 5 & 0 \end{pmatrix}$$

$$RM^{ijst} = \begin{pmatrix} 0.70 & 0.40 & 0.85 & 0.40 \\ 0.69 & 0.17 & 0.78 & 0.36 \\ 0.34 & 0.67 & 0.58 & 0.86 \\ 0.59 & 0.55 & 0.80 & 0.55 \end{pmatrix}$$

Considering the existence possibility of the link between v^i_2 and v^j_2, we first find v^i_1, v^i_3 and v^i_4 as the generalized neighbors of v^i_2, and find v^j_1 and v^j_3 as the generalized neighbors of v^j_2. In this case, there are only two existing links across the sets $N^i_2 = \{v^i_1, v^i_3, v^i_4\}$ and $N^j_2 = \{v^j_1, v^j_3\}$. Therefore, the numerator is $2 \times 3 + 6 \times 3 = 24$, the denominator is 144, and $RM^{ijst}_{2,2} = 24/144 = 0.17$. It indicates that v^i_2 and v^j_2 are not likely to be linked.

$\{\mathbf{x}_p\}$ = ComputingFeatures(i,j,$\mathcal{I}$)

1) Construct the meta-path set $\mathcal{P}^{ii}$ and $\mathcal{P}^{jj}$ through breadth-first-search
2) Combine meta-paths in $\mathcal{P}^{ii}$ and $\mathcal{P}^{jj}$ in pairs and compute the *relatedness measure* as the feature vector $\mathbf{x}_p$, $p \in \mathcal{I}$
3) Return $\{\mathbf{x}_p\}$

Fig. 5. The function of computing features using the relatedness measure, parameterized by the type of source node i, the type of target node j, and the index set of instances $\mathcal{I}$.

However, in the case of v^i_3 and v^j_4, four links exist across $N^i_3 = \{v^i_1, v^i_2, v^i_4\}$ and $N^i_4 = \{v^j_1, v^j_3\}$, the generalized neighbor sets of v^i_3 and v^j_4 respectively. Especially, v^i_2 is the most similar node of v^i_3, and it links to both similar nodes of v^j_4 in N^i_4. It is important because v^i_2 counts more in the computation of the relatedness measure, for its largest similarity with v^i_3, $EPC^{iis}_{2,3} = 8$. $RM^{ijst}_{3,4} = 0.86$ indicates a high chance that there can exist a direct link between v^i_3 and v^j_4, according to the meta-path $\mathcal{P}^{ijst}$.

To address the problem of collective link prediction, we propose an iterative framework below, using the relatedness measure to construct the feature vectors of links. Therefore, we can build a learning model for each type of links.

C. Iterative Framework

There are multiple types of links in heterogeneous information networks. For each link type, the links are partially known. Sparsity can be a problem for some link types, where there are far more unknown linkage relationship than existing links [18]. Since the potential existence of multiple types of links can interact with each other, the process of link prediction should be conducted in a collective way. In other words, apart from the existing linkage information, we can leverage predicted existence of other link types to facilitate the prediction of a particular sparse link type, which can also share its enriched linkage information in return with other link types. Thus, it is a collective and complementary process.

Inspired by co-training [2], we try to combine labeled and unlabeled data in heterogeneous information networks where existing links can be regarded as limited labeled data while unknown links are unlabeled data in large size. Different from conventional co-training that each example can be partitioned into two distinct views and shares a same set of label concepts, in the setting of heterogeneous information networks, there are different labels for each learning algorithm representing whether the corresponding type of links can exist. However, these different types of links co-exist in a network, which could be leveraged to exploit their correlations. Instead of directly exchanging their most confident predictions on unlabeled data, we consider adding the most confidently predicted links of each type into the network schema. Then meta-paths, computed from the network schema, will be updated for each type of links. In such way, complementary prediction information from different types of links can be exchanged and enriched in an indirect manner.

Input:
$\mathcal{G}$: a heterogeneous information network, A: a base learner (default: SVM),
$\mathcal{Y}_{\mathcal{L}}$: label set of training instances, θ : threshold to update $\mathcal{W}$ (default: 0.9),
$maxlen$: maximum meta-path length (default: 3), $maxiter$: maximum # of iteration (default: 5)
Training Initialization:
For k = 1 to m, learn the local model f^k for the k-th link type, R^{ij}:
1. $\{\mathbf{x}_p^k\}$ = ComputingFeatures(i,j,$\mathcal{E}^k$) and construct training sets $\mathcal{D}^k = \{(\mathbf{x}_p^k, y_p^k)\}$, $p \in \mathcal{E}^k$
2. Let $f^k = A(\mathcal{D}^k)$ be the local model trained on $\mathcal{D}^k$
Iterative Inference:
Repeat until convergence or #iteration $> maxiter$:
For k = 1 to m:
1. Sample a set of unknown links $\mathcal{S}^k$
2. $\{\mathbf{x}_p^k\}$ = ComputingFeatures(i,j,$\mathcal{S}^k$) and obtain the probabilistic output of $f^k(\mathbf{x}_p^k)$, $p \in \mathcal{S}^k$
3. Update the weighted adjacency matrix $\mathcal{W}^{ij}$ where $f^k(\mathbf{x}_p^k)$ is above θ
Output:
For k = 1 to m:
1. $\{\mathbf{x}_p^k\}$ = ComputingFeatures(i,j,$\mathcal{T}^k$), $p \in \mathcal{T}^k$
2. Return the deterministic output of $f^k(\mathbf{x}_p^k)$ as $\hat{\mathcal{Y}}_{\mathcal{T}^k}$

Fig. 4. The HCLP algorithm.

In this paper, we propose an effective algorithm to perform the heterogeneous collective link prediction, called HCLP. We summarize it in Figure 4. The algorithm is essentially a general iterative framework and it calls the function defined in Figure 5 to compute features for each link, where the relatedness measure can be replaced with other possible measure functions. It contains the following key steps:

Training Initialization: Basically, for the k-th link type, say "$\mathcal{V}^i \xrightarrow{R^{ij}} \mathcal{V}^j$", we use breadth-first-search to exploit all the meta-paths of $\mathcal{P}^{ii}$ and $\mathcal{P}^{jj}$ within a maximum path length $maxlen$. It has been shown in [26] that long meta-paths are not quite useful in capturing the linkage structures of heterogeneous information networks. To avoid redundancy, a meta-path is removed if it contains another meta-path that has already been exploited. Then, meta-paths of $\mathcal{P}^{ii}$ and $\mathcal{P}^{jj}$ can be combined in pairs to compute the relatedness measure as the feature vectors for predicting the link type "$\mathcal{V}^i \xrightarrow{R^{ij}} \mathcal{V}^j$". Using such feature vectors, we can train a classifier f^k corresponding to the k-th link type (*i.e.*, R^{ij}).

Iterative Inference: Overall, it is an iterative classification algorithm [21] for the inference step. During each round of iterations, a set of unknown links $\mathcal{S}^k$ is randomly sampled for the k-th link type, and the probability $P(y_p^k|\mathbf{x}_p^k), p \in \mathcal{S}^k$ can be obtained from the corresponding trained classifier f^k. The weighted adjacency matrix $\mathcal{W}^{ij}$ is then updated where the probabilistic output of $f^k(\mathbf{x}_p^k)$ is above a threshold θ.

Based on our proposed HCLP algorithm, especially the step of iterative inference, multiple types of links can be collectively predicted in heterogeneous information networks. In each round of iterations, multiple types of links are predicted in turn, and their weighted adjacency matrices would be updated at the positions where the corresponding predictions are confident enough, thereby enriched network schema can be leveraged as complementary information for prediction of other link types.

IV. EXPERIMENTS

We first introduce the procedure of data processing in section IV-A and compared approaches in section IV-B. Experimental results are summarized in section IV-C. Furthermore, we analyze the robustness of HCLP in section IV-D and conduct parameter sensitivity analysis in section IV-E.

A. Data Processing

We test our algorithm on SLAP dataset [4] to evaluate the performances of collective prediction of multiple types of links in heterogeneous information networks. To balance the numbers of different types of nodes, we extract the top-5000 *chemical compounds* and *genes* respectively, sorted by their degrees to each other. All the other nodes and links remain the same as in the original dataset. The task is to infer the potential existence of 11 types of links collectively. As an evaluation metric, we respectively investigate the prediction accuracy of each link type.

B. Compared Methods

In order to demonstrate the effectiveness of our approach in heterogeneous collective prediction of multiple types of links, we extend the measure functions introduced in [26], so that they can be applied to a pair of nodes of different types like the relatedness measure, since their original definitions as proximity measures are based on one single type of nodes.

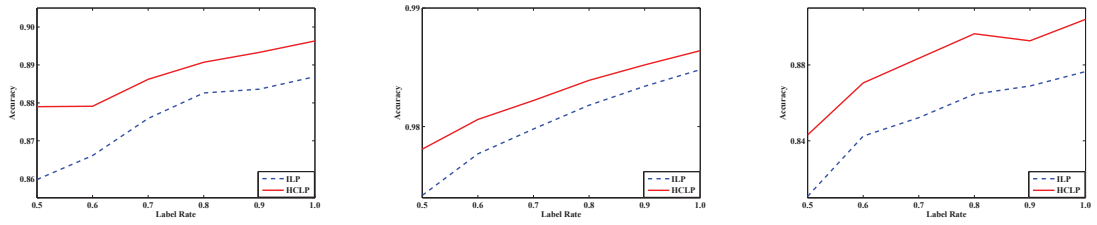
• **NPC (normalized path count)** is to discount the number of paths between a pair of nodes in the network schema by the connectivity of the source node and the target node.

$$NPC_{pq}^{ijst} = \frac{2 \times EPC_{pq}^{ijst}}{EPC_{p,\cdot}^{ijst} + EPC_{\cdot,q}^{ijst}}$$

where $EPC_{p,\cdot}^{ijst}$ denotes the total number of path instances along $\mathcal{P}^{ijst}$ starting from v_p^i and $EPC_{\cdot,q}^{ijst}$ denotes the total number of path instances along $\mathcal{P}^{ijst}$ ending at v_q^j.

TABLE II
LINK PREDICTION PERFORMANCES "AVERAGE ACCURACY (RANK)".

Type of Links	Methods							
	ILP(NPC)	ILP(RW)	ILP(SRW)	ILP(RM)	HCLP(NPC)	HCLP(RW)	HCLP(SRW)	HCLP(RM)
Gene $\xrightarrow{PPI}$ Gene	0.889 (5)	0.878 (8)	0.893 (3)	0.888 (6)	0.892 (4)	0.884 (7)	0.898 (2)	0.899 (1)
Gene $\xrightarrow{cause}$ Disease	0.690 (8)	0.696 (6)	0.701 (3)	0.697 (5)	0.696 (6)	0.701 (3)	0.708 (1)	0.704 (2)
Gene $\xrightarrow{has}$ Pathway	0.888 (7)	0.899 (5)	0.909 (2)	0.909 (2)	0.881 (8)	0.894 (6)	0.906 (4)	0.913 (1)
Gene $\xrightarrow{has}$ GO term	0.862 (5)	0.851 (8)	0.867 (4)	0.862 (6)	0.870 (3)	0.857 (7)	0.875 (1)	0.874 (2)
Gene $\xrightarrow{has}$ Gene family	0.669 (7)	0.674 (3)	0.673 (5)	0.678 (1)	0.669 (7)	0.673 (4)	0.677 (2)	0.672 (6)
Gene $\xrightarrow{has}$ Tissue	0.813 (8)	0.828 (6)	0.846 (3)	0.841 (4)	0.818 (7)	0.830 (5)	0.849 (2)	0.851 (1)
Chemical $\xrightarrow{bind}$ Gene	0.984 (4)	0.978 (7)	0.984 (4)	0.984 (4)	0.985 (2)	0.978 (7)	0.985 (2)	0.986 (1)
Chemical $\xrightarrow{treat}$ Disease	0.887 (4)	0.867 (8)	0.886 (5)	0.873 (7)	0.893 (3)	0.876 (6)	0.894 (2)	0.901 (1)
Chemical $\xrightarrow{cause}$ Side effect	0.901 (6)	0.886 (8)	0.913 (4)	0.931 (2)	0.913 (4)	0.890 (7)	0.927 (3)	0.939 (1)
Chemical $\xrightarrow{has}$ Substructure	0.934 (7)	0.938 (5)	0.943 (3)	0.939 (4)	0.935 (6)	0.932 (8)	0.951 (1)	0.947 (2)
Chemical $\xrightarrow{has}$ CO term	0.917 (6)	0.907 (8)	0.919 (5)	0.922 (3)	0.921 (4)	0.909 (7)	0.925 (2)	0.928 (1)
Average Rank	(6.1)	(6.5)	(3.7)	(4)	(4.9)	(6.1)	(2)	(1.7)



(a) Gene $\xrightarrow{PPI}$ Gene (b) Chemical $\xrightarrow{bind}$ Gene (c) Chemical $\xrightarrow{treat}$ Disease

Fig. 6. Link prediction accuracies with different label rates.

• **RW (random walk)** is to discount the number of paths between a pair of nodes by the connectivity of the source node.

$$RW_{pq}^{ijst} = \frac{EPC_{pq}^{ijst}}{EPC_{p,\cdot}^{ijst}}$$

• **SRW (symmetric random walk)** incorporates random walks from both directions following a meta-path.

$$SRW_{pq}^{ijst} = \frac{EPC_{pq}^{ijst}}{EPC_{p,\cdot}^{ijst}} + \frac{EPC_{pq}^{ijst}}{EPC_{\cdot,q}^{ijst}}$$

• **RM (relatedness measure)** is our designed measure function described in section III-B.

$$RM_{pq}^{ijst} = \frac{EPC_{pq}^{ijst}}{\sum_{p'=1}^{|\mathcal{V}_i|}\sum_{q'=1}^{|\mathcal{V}_j|} EPC_{pp'}^{iis} \times EPC_{qq'}^{jjt}}$$

$$= \frac{\sum_{p'=1}^{|\mathcal{V}_i|}\sum_{q'=1}^{|\mathcal{V}_j|} EPC_{pp'}^{iis} \times \mathcal{W}^{ij}(v_{p'}^i, v_{q'}^j) \times EPC_{qq'}^{jjt}}{\sum_{p'=1}^{|\mathcal{V}_i|}\sum_{q'=1}^{|\mathcal{V}_j|} EPC_{pp'}^{iis} \times EPC_{qq'}^{jjt}}$$

We denote the HCLP algorithm implementing RM as HCLP(RM). By replacing RM in Figure 5 with other measure functions introduced above, we can have HCLP(NPC), HCLP(RW), and HCLP(SRW).

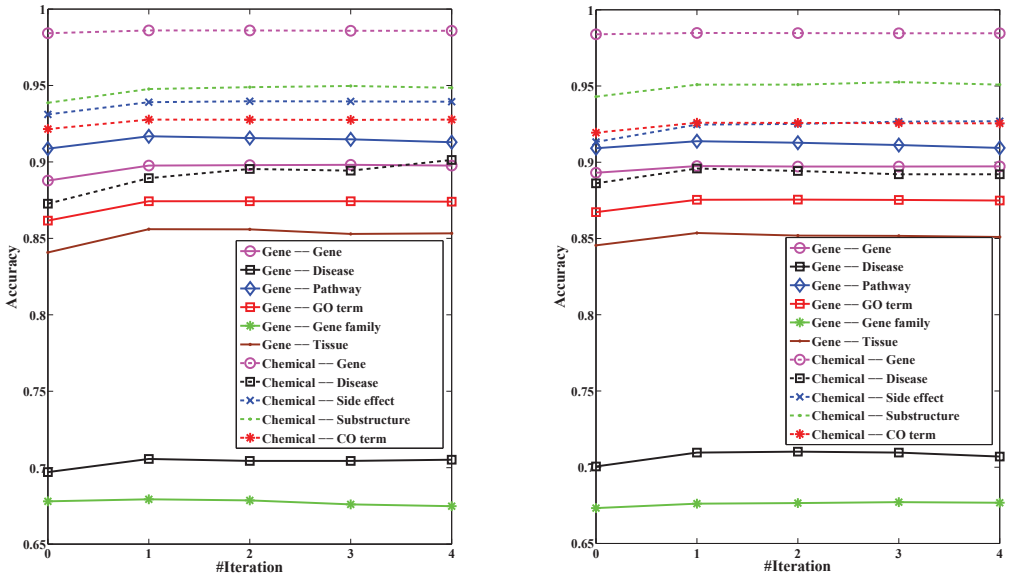
In comparison with our proposed HCLP algorithm, we also employ these measure functions in a conventional supervised setting ILP where the prediction tasks for multiple types of links are assumed to be independent. Such methods complete training and prediction in one round, and they are denoted as ILP(NPC), ILP(RW), ILP(SRW), and ILP(RM), respectively.

For a fair comparison, we use LibSVM [3] with linear kernel as the base classifier for all the compared methods. In the experiments, 10-fold cross validations are performed on each type of links.

C. Performances of Heterogeneous Collective Link Prediction

It can be observed from Table II that, in comparison with ILP, performances can be improved by implementing link prediction collectively in heterogeneous information networks, no matter what measure functions are employed. For example, HCLP(RM) generally outperforms ILP(RM), while HCLP(SRW) does better than ILP(SRW). It demonstrates that our proposed HCLP iterative framework can effectively work on different measure functions to improve the performances of predicting multiple types of links. Therefore, in heterogeneous information networks, different types of links should be taken care of simultaneously.

Then we study the effectiveness of different measure functions on prediction of multiple types of links in heterogeneous information networks. In the collective setting HCLP, RM



(a) HCLP(RM) (b) HCLP(SRW)

Fig. 7. Link prediction accuracies with different number of iterations.

outperforms NPC and RW and displays a slight advantage over SRW. However, no one rules all. We can safely conclude that different measure functions can be best fitted into different types of links by capturing its underlying complex physical meanings. In most cases, RM and SRW can be good alternatives.

We also notice that the proposed HCLP iterative framework can be perfectly applied to relatedness measure RM. As can be seen from Table II, HCLP(RM) outperforms all the other compared methods on predicting a majority of link types. Basically, HCLP(RM) effectively boosts the performances of ILP(RM) through an iterative and collective process. For prediction of the link type "chemical compound $\xrightarrow{treatDisease}$ disease", HCLP(RM) improves ILP(RM) in accuracy by nearly three percent.

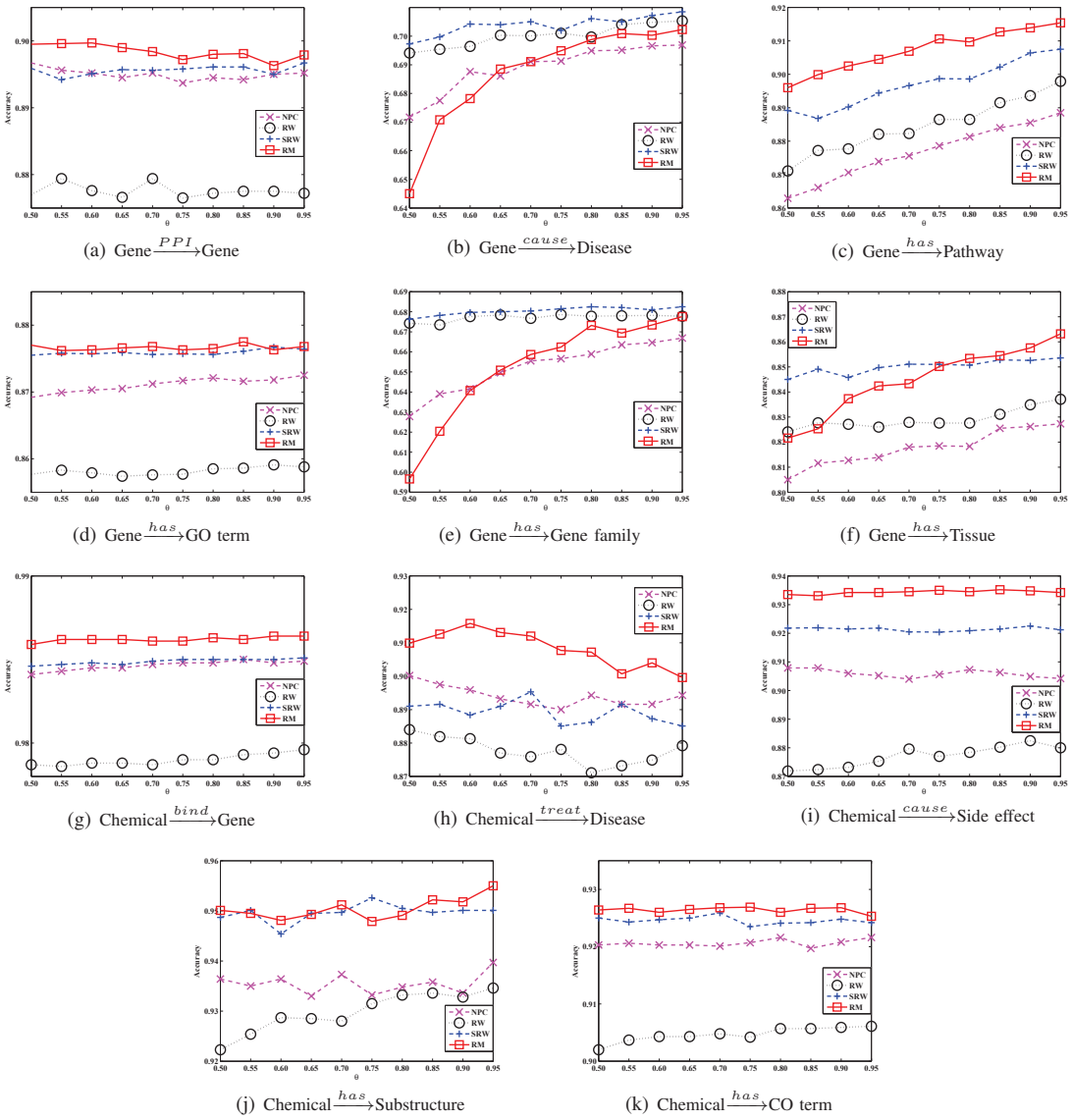
D. Robustness

In many real-world applications, sparsity can always be a problem. To investigate the practical applicability of the HCLP algorithm, we conduct experiments where a proportion of training data of a particular link type is removed and only the remaining training data can be used in the network schema, without any changes in other types of links. Figure 6 shows the experimental results of the methods ILP and HCLP, employing RM, for predicting three important types of links "gene $\xrightarrow{PPI}$ gene", "chemical compound $\xrightarrow{bind}$ gene", and "chemical compound $\xrightarrow{treatDisease}$ disease". It demonstrates that HCLP outperforms ILP at different rates of labeled links. Generally, HCLP does a lot better at lower rate of labeled data. Therefore, the HCLP algorithm can effectively leverage complementary information from other link types, and improve the performances of predicting a particular link type even with its limited training data through a process of heterogeneous collective link prediction.

E. Parameter Sensitivity Analysis

There are two issues in the sensitivity analysis experiments of the HCLP algorithm. Figure 7 illustrates the sensitivity of the methods HCLP(RM) and HCLP(SRW) to the parameter $maxiter$ denoting the maximum number of iterations. As can be seen, the first few rounds of iterations can effectively boost the performances of predicting multiple types of links, and prediction accuracies stabilize as the number of iterations continues to increase. It means our proposed HCLP algorithm can quickly converge within a few rounds of iterations.

Sensitivity of the HCLP algorithm to the threshold θ is shown in Figure 8 where it plots the experimental results of HCLP implementing different measure functions as θ changes. The parameter θ is a minimum value that the probabilistic prediction of a sampled link should reach when we update its corresponding weighted adjacency matrix in the step of iterative inference. Generally, the performances of HCLP, employing different measure functions, are quite stable regardless of the value of θ for predicting some types of links. In other cases, the performances are improved as θ increases, which implies that a low value of θ can bring in noise into the network schema and mislead the final predictions. Thus, a slightly high value of θ, as 0.90, is preferred in the HCLP algorithm.



(a) Gene $\xrightarrow{PPI}$ Gene (b) Gene $\xrightarrow{cause}$ Disease (c) Gene $\xrightarrow{has}$ Pathway

(d) Gene $\xrightarrow{has}$ GO term (e) Gene $\xrightarrow{has}$ Gene family (f) Gene $\xrightarrow{has}$ Tissue

(g) Chemical $\xrightarrow{bind}$ Gene (h) Chemical $\xrightarrow{treat}$ Disease (i) Chemical $\xrightarrow{cause}$ Side effect

(j) Chemical $\xrightarrow{has}$ Substructure (k) Chemical $\xrightarrow{has}$ CO term

Fig. 8. Link prediction accuracies with different threshold θ.

V. Related Work

Link prediction, mainly in homogeneous information networks [20], has been extensively studied in recent years. Previous work can be generally categorized as follows: (1) unsupervised approaches which focus on proposing different similarity measures, either based upon graph topology or node attributes [1], [19]; (2) supervised approaches which exploit different features from the training set and use supervised learning models to predict the potential existence of links [10], [30], [20]. For a detailed survey on link prediction, please refer to [8].

Heterogeneous information networks have attracted much attention in recent years. Sun et al. studied the clustering problem and the top-k similarity problem in heterogeneous information networks [28], [26]. Ming et al. studied a specialized classification problem in heterogeneous information networks where different types of nodes share a same set of label concepts [13].

Besides, there are related work on link prediction in heterogeneous information networks. The notion of meta-path was applied to redefine the graph proximity measures in heterogeneous information networks [25]. However, it was confined to predicting links between one single type of nodes. Instead of meta-path, the notion of triad census was used to determine the weights of different combinations of edge types [6]. Shi et al. studied the relevance search problem in heterogeneous networks to measure the relatedness of heterogeneous objects [24]. Moreover, the heterogeneous link prediction problem was studied in the unsupervised setting and one single type of edges were predicted independently [16], [7].

VI. Conclusion

In this paper, we studied the problem of collective link prediction in heterogeneous information networks. Conventional approaches for link prediction mainly focus on homogeneous information networks. However, in many real-world applications, there are multiple types of links in heterogeneous information networks and the potential existence of different types of links can interact with each other. Therefore, we introduce the *linkage homophily principle* and design a relatedness measure, called RM, between different types of objects to compute the existence probability of a link. We also extend conventional proximity measures to heterogeneous links. Furthermore, we propose an iterative framework, called HCLP, to predict multiple types of links collectively by exploiting diverse and complex linkage information in heterogeneous information networks. Empirical studies on real-world tasks demonstrate that our proposed method for heterogeneous collective link prediction can effectively facilitate the process of collectively predicting multiple types of links in heterogeneous information networks.

VII. Acknowledgements

This work is supported in part by NSF through grants CNS-1115234, OISE-1129076, and US Department of Army through grant W911NF-12-1-0066.

References

[1] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2001.

[2] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the ACM Conference on Computational Learning Theory*, pages 92–100, 1998.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] B. Chen, Y. Ding, and D. Wild. Assessing drug target association using semantic linked data. *PLOS Computational Biology*, 8(7), 2012.

[5] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, and Y. Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLOS Computational Biology*, 8, 2012.

[6] D. Davis, R. Lichtenwalter, and N. Chawla. Multi-relational link prediction in heterogeneous information networks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2011.

[7] L. Eronen and H. Toivonen. Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, 13(1):119, 2012.

[8] L. Getoor and C. Diehl. Link mining: a survey. *SIGKDD Explorations*, 7:3–12, 2005.

[9] A. Gottlieb, G. Stein, E. Ruppin, and R. Sharan. Predict: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, 7, 2011.

[10] M. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM Workshop on Link Analysis, Counterterrorism and Security*, 2006.

[11] J. Hastings, P. Matos, M. Ennis, and C. Steinbeck. Towards automatic classification within the chebi ontology. *Nature Precedings*, 2009.

[12] S. Jaeger, S. Gaudan, U. Leser, and D. R. Schuhmann. Integrating protein-protein interactions and text mining for protein function prediction. *BMC Bioinformatics*, 9, 2008.

[13] M. Ji, J. Han, and M. Danilevsky. Ranking-based classification of heterogeneous information networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1298–1306, 2011.

[14] X. Kong, B. Cao, and P. S. Yu. Multi-label classification by mining label and instance correlations from heterogeneous information networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2013.

[15] X. Kong, P. S. Yu, Y. Ding, and D. J. Wild. Meta path-based collective classification in heterogeneous information networks. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2012.

[16] T. Kuo, R. Yan, Y. Huang, P. Kung, and S. Lin. Unsupervised link prediction using aggregative statistics on heterogeneous social networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 775–783, 2013.

[17] N. Lao and W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine Learning*, 81(2):53–67, 2010.

[18] V. Leroy, B. Cambazoglu, and F. Bonchi. Cold start link prediction. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 393–402, 2010.

[19] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 556–559, 2003.

[20] R. Lichtenwlter, J. Lussier, and N. Chawla. New perspectives and methods in link prediction. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 243–252, 2010.

[21] Q. Lu and L. Getoor. Link-based classification. In *Proceedings of the International Conference on Machine Learning*, pages 496–503, 2003.

[22] Y. Moreau and L. Tranchevent. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13:523–536, 2012.

[23] E. Pauwels, V. Stoven, and Y. Yamanishi. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*, 12, 2011.

[24] C. Shi, X. Kong, P. S. Yu, S. Xie, and B. Wu. Relevance search in heterogeneous networks. In *Proceedings of the International Conference on Extending Database Technology*, pages 180–191. ACM, 2012.

[25] Y. Sun, R. Barber, M. Gupta, C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 2011.

[26] Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu. PathSim: Meta path-based top-k similarity search in heterogeneous information networks. In *Proceedings of the International Conference on Very Large Data Bases*, 2011.

[27] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1348–1356, 2012.

[28] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2009.

[29] A. Vinayagam, C. Val, F. Schubert, R. Eils, K. Glatting, S. Suhai, and R. Konig. GOPET: A tool for automated predictions of gene ontology terms. *BMC Bioinformatics*, 7, 2006.

[30] C. Wang, V. Satuluri, and S. Parthasarathy. Local probabilistic models for link prediction. In *Proceedings of the IEEE International Conference on Data Mining*, pages 322–331, 2007.