Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks*

Yizhou Sun† Rick Barber† Manish Gupta† Charu C. Aggarwal‡ Jiawei Han†
† University of Illinois at Urbana-Champaign, Urbana, IL
‡ IBM T. J. Watson Research Center, Hawthorne, NY
†{sun22, barber5, gupta58, hanj}@illinois.edu ‡ charu@us.ibm.com

***Abstract*—The problem of predicting links or interactions between objects in a network, is an important task in network analysis. Along this line, link prediction between co-authors in a co-author network is a frequently studied problem. In most of these studies, authors are considered in a homogeneous network, *i.e.*, only one type of objects (author type) and one type of links (co-authorship) exist in the network. However, in a real bibliographic network, there are multiple types of objects (*e.g.*, venues, topics, papers) and multiple types of links among these objects. In this paper, we study the problem of co-author relationship prediction in the heterogeneous bibliographic network, and a new methodology called *PathPredict*, *i.e.*, meta path-based relationship prediction model, is proposed to solve this problem. First, meta path-based topological features are systematically extracted from the network. Then, a supervised model is used to learn the best weights associated with different topological features in deciding the co-author relationships. We present experiments on a real bibliographic network, the DBLP network, which show that meta path-based heterogeneous topological features can generate more accurate prediction results as compared to homogeneous topological features. In addition, the level of significance of each topological feature can be learned from the model, which is helpful in understanding the mechanism behind the relationship building.**

I. Introduction

Link prediction in networks has been an important topic since the emergence of online social networks. Most of the existing link prediction studies ([7], [4], [15], [8], [6]) are designed for homogeneous networks, in which only one type of objects exists in the network. Examples of such networks include friendship and co-author networks. Recent research [14] has also studied the problem of link prediction in networks containing different kinds of attribute values associated with objects. However, most of the networks in real world are heterogeneous, and attribute values of objects are often difficult to fully obtain. Therefore, the use of topological features between objects in a heterogeneous network is critical in predicting links in a holistic way.

Table I
Top-5 Predicted Co-Authors for Jian Pei in 2003-2009

Rank	Hybrid heterogeneous features	# Shared authors
1	**Philip S. Yu**	**Philip S. Yu**
2	**Raymond T. Ng**	Ming-Syan Chen
3	Osmar R. Zaïane	Divesh Srivastava
4	**Ling Feng**	Kotagiri Ramamohanarao
5	**David Wai-Lok Cheung**	**Jeffrey Xu Yu**

* Authors in bold format are the true new co-authors of Jian in the time period 2003-2009.

In this paper, we study the problem of predicting future co-author relationships between existing authors in a heterogeneous bibliographic network, using heterogeneous topological features. Different from the traditional co-author network setting, a heterogeneous bibliographic network is considered, which contains multiple types of objects, such as authors, venues, topics and papers, as well as multiple types of links denoting different relations among these objects, such as "write" and "written by" relations between authors and paper, "cite" and "cited by" relations between papers, and so on. In link prediction tasks, paths between two objects play a very important role in generating topological features in homogeneous networks. For example, the number of common neighbors used in [7] is the number of length-2 paths between the two objects; and the $Katz_\beta$ measure used in [5] is a weighted sum of counts of paths with different lengths. However, in heterogeneous networks, different paths between the same pair of authors in the network may represent different relations and denote different semantic meanings. For example, a path between two authors "Jim" and "Mike" could be "Jim-P_5-SIGMOD-P_6-Mike" (Fig. 3), that is Jim and Mike are linked together as they both published papers (P_5 and P_6) in the conference "SIGMOD". They can also be connected through a path denoting they have common co-authors, *e.g.*, "Jim-P_1-Ann-P_3-Mike", and so on. We can see that the type information associated with objects and links makes the topological structure in heterogeneous networks more complex and with richer semantics than that in homogeneous networks.

We then propose a new methodology called *PathPredict*, *i.e.*, meta path-based relationship prediction model,

*The work was supported in part by U.S. National Science Foundation grants IIS-09-05215, the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), and MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

978-0-7695-4375-8/11 $26.00 © 2011 IEEE
DOI 10.1109/ASONAM.2011.112

to solve the problem. Instead of treating objects and links of different types equally or extracting homogeneous sub-networks from the original network, we propose a meta path-based topological feature framework in the heterogeneous bibliographic network. The goal is to systematically define the relations between authors encoded in different paths using the meta-structure of these paths, *i.e.*, the meta paths [12]. For example, the meta path for "Jim-P_5-SIGMOD-P_6-Mike" is "author-paper-venue-paper-author". Further, several measures are proposed to quantify these meta path-based relations, each of which quantifies the relation in a different way. We then use a supervised learning framework to learn the best weights associated with each topological feature. Experiments show that by considering the rich semantics of heterogeneous topological features, the accuracy of link prediction can be improved significantly. For example, Table I shows the top-5 predicted co-authors for Jian Pei in the time period of 2003 to 2009 in DBLP network, using hybrid heterogeneous topological features and the number of common neighbors in the extracted co-author sub-network from year 1996 to 2002 respectively. We can see that, the results generated by heterogeneous features has a higher accuracy compared with the homogeneous one. Furthermore, from the model we can tell which topological feature plays a more important role in deciding their future collaboration, which is helpful for us to understand the *mechanism* of future relationship construction. The contributions of this paper include:

- We study the problem of co-author relationship prediction in **heterogeneous** bibliographic networks;
- A new methodology called *PathPredict*, *i.e.*, meta path-based relationship prediction model, is developed to solve this problem;
- Experiments on the real DBLP bibliographic network show that by considering the heterogeneous types of objects and links in the network collectively, the co-author relationship prediction accuracy can be significantly improved.

II. Problem Definition

In this section, we introduce the definition of a heterogeneous bibliographic network and the co-author relationship prediction task in this network setting.

A. Heterogeneous Bibliographic Network

In this paper, we use the DBLP bibliographic network as an example of heterogeneous bibliographic networks. The DBLP bibliographic dataset with citation information provided by [13] consists of rich information for publications, such as the authors, venues, titles and so on. We further extract frequent phrases from titles as topics using the sequential pattern mining algorithm *PrefixSpan* [10]. The network then contains 4 types of objects, namely papers, authors, topics, and venues (conferences or journals).

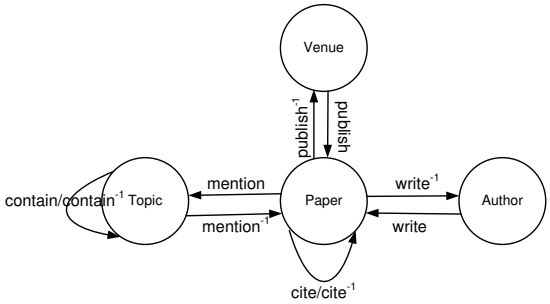


Figure 1. Schema for DBLP Bibliographic Network

As an abbreviation, we use the initial capital letters to denote these object types, namely P for papers, A for authors, T for topics, and V for venues. Links exist between authors and papers by the relations "write" and "written by" (denoted as $write^{-1}$), between papers and topics by "mention" and " mentioned by" (denoted as $mention^{-1}$), between venues and papers by "publish" and "published by" (denoted as $publish^{-1}$), between papers by "cite" and "cited by" (denoted as $cite^{-1}$), and between topics by "contain" and "contained in" (denoted as $contain^{-1}$) if one topic is contained in the other topic. We can see that the DBLP bibliographic network is a directed graph with type information on objects and links. Further, we use a meta structure called **network schema** to summarize the network, which is shown in Fig. 1. In the network schema, the nodes are the types of objects, and the edges are relations between types.

B. The Co-Author Relationship Prediction Task

Given a heterogeneous network, the link prediction task is then generalized to **relationship** building prediction, which is to predict whether two objects will build a relationship following a certain **target relation**. Notice that relationships between objects are instances of the target relation. In our case, we say Jim and Mike have built a co-author relationship, if they follow a co-author relation. Unlike homogeneous co-author networks, the co-author relation is not defined in our DBLP network schema directly. Nevertheless, it can be defined through the composition of two relations "$write$" and "$write^{-1}$", that is, two authors a_i and a_j are co-authors, if and only if a_i has written a paper p that is written by a_j.

Formally, following the work [12], we use the concept of **meta path** defined over the network schema to describe the relations that can be derived from the network. A meta path is a path defined on the network schema, where nodes are object types and edges are relations between object types. For example in the DBLP network, the co-author relation can be described using the meta path $A \xrightarrow{write} P \xrightarrow{write^{-1}} A$, and in abbreviation as $A-P-A$, if there is no ambiguity

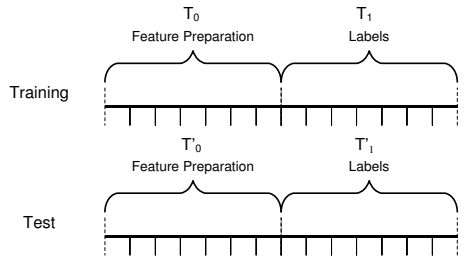


Figure 2. Supervised Framework for Relationship Prediction

in either the meaning or the order of the relation. Another example is $A - P \rightarrow P - A$, which is short for $A \xrightarrow{write} P \xrightarrow{cite} P \xrightarrow{write^{-1}} A$. This describes the citation relation between authors. Notice that the network schema provides a meta structure description for the network, and a meta path provides a meta structure description for paths between objects in the network.

Also, similar to the seminal work of link prediction in homogeneous network [7], we are interested in predicting new relationships rather than repeated relationships. In other words, we are interested in predicting whether two authors that have never co-authored before will co-author sometime in the future rather than predicting how many times two authors will co-author in the future.

The co-author relationship building between two authors can be affected by many factors, and in this paper we are particularly interested in the impact of topological structures on the relationship building process. In other words, we want to know what kind of connections between two authors are more helpful to lead to future collaboration(s). In order to solve this problem, we first systematically design the topological features in the DBLP network, and then a supervised learning method is proposed to learn the weights associated with each topological feature in determining relationships.

The supervised learning framework is summarized in Fig. 2. Generally, given a past time interval $T_0 = [t_0, t_1]$, we want to use the topological features extracted from the aggregated network in the time period T_0, to predict the relationship building in a future time interval, say $T_1 = [t_1, t_2]$. In the **training stage**, we first sample a set of author pairs that have never co-authored in T_0, collect their associated topological features in T_0, and record whether a relationship is to appear between them in the future interval T_1. A training model is then built to learn the best coefficients associated with each topological feature by maximizing the likelihood of relationship building. In the **test stage**, we apply the learned coefficients to the topological features for the test pairs, and compare the predicted relationship with the ground truth.

III. The *PathPredict* Model

In this section, we introduce the *PathPredict* model in detail, which includes two components: (1) the meta path-based topological feature definition and (2) the logistic regression-based co-authorship prediction model.

A. Topological Features in Heterogeneous Networks

First, we study how to systematically define the topological features in the DBLP network. Topological features are also called structural features, which aim at extracting connectivity properties for pairs of objects. Topological feature-based link prediction aims at inferring the future connectivity by leveraging the current connectivity of the network. There are some frequently used topological features defined in homogeneous networks, such as the number of common neighbors, preferential attachment [2], [9], $katz_\beta$ [5] and so on. We first review several commonly used topological features in homogeneous networks, and then propose a systematic meta path-based methodology to define topological features in heterogeneous networks.

1) Review Existing Topological Features: We now list several well-known and frequently used topological features in homogeneous networks. For more topological features, the readers can refer to [7] .

- *Common neighbors*. Common neighbors is defined as the number of common neighbors shared by two objects a_i and a_j, namely $|\Gamma(a_i) \cap \Gamma(a_j)|$, where $\Gamma(a)$ is the notation for neighbor set of the object a and $|\cdot|$ denotes the size of a set.
- *Jaccard's coefficient*. Jaccard's coefficient is a measure to evaluate the similarity between two neighbor sets, which can be viewed as the normalized number of common neighbors, namely $\frac{|\Gamma(a_i) \cap \Gamma(a_j)|}{|\Gamma(a_i) \cup \Gamma(a_j)|}$.
- $Katz_\beta$. $Katz_\beta$ [5] is a weighted summation of counts of paths between two objects with different lengths, namely $\sum_{l=1}^{\infty} \beta^l |path_{a_i,a_j}^{\langle l \rangle}|$, where β^l is the damping factor for the path with length l.
- *PropFlow*. In a recent work [8], a random walk-based measure PropFlow is proposed to measure the topological feature between two objects. This method assigns the weighs to each path (with fixed length l) using the products of proportions of the flows on the edges.

We can see that, most of the existing topological features in homogeneous networks are based on neighbor sets or paths between two objects. However, as there are multi-typed objects and multi-typed relations in heterogeneous networks, the neighbors of an object could belong to multiple types, and the paths between two objects could follow different meta paths and indicate different relations. Thus, we need to design a more complex strategy to generate topological features in heterogeneous networks to distinguish paths with different meanings.

2) Meta Path-based Topological Features: To design the topological features in the heterogeneous networks, we first define the topology between two objects using meta paths, and then define measures on the specific topology.

Table II
META PATHS UNDER LENGTH 4 BETWEEN AUTHORS IN THE DBLP NETWORK

Meta Path	Semantic Meaning of the Relation
$A-P-A$	a_i and a_j are coauthors (the target relation)
$A-P \rightarrow P-A$	a_i cites a_j
$A-P \leftarrow P-A$	a_i is cited by a_j
$A-P-V-P-A$	a_i and a_j publish in the same venues
$A-P-A-P-A$	a_i and a_j are co-authors of the same authors
$A-P-T-P-A$	a_i and a_j write the same topics
$A-P \rightarrow P \rightarrow P-A$	a_i cites papers that cite a_j
$A-P \leftarrow P \leftarrow P-A$	a_i is cited by papers that are cited by a_j
$A-P \rightarrow P \leftarrow P-A$	a_i and a_j cite the same papers
$A-P \leftarrow P \rightarrow P-A$	a_i and a_j are cited by the same papers

Meta Path-based Topology. As introduced in Sec. II, a meta path is a path defined over the network schema, and denotes a composition relation over the heterogeneous networks. By checking the existing topological features defined in homogeneous networks, we can find that both the neighbor set-based features and path-based features can be generalized in heterogeneous information networks, by considering paths following different meta paths. For example, if we treat each type of neighbors separately and extend the immediate neighbors to n-hop neighbors (*i.e.*, the distance between one object and its neighbors are n), the common neighbor feature between two authors is then becoming the count of paths between the two authors following different meta paths. For path-based features, such as $Katz_\beta$, it can be extended as a combination of paths following different meta paths. Hence, each meta path defines a unique topology between objects, representing a special relation.

Meta paths between two object types can be obtained by traversing on the DBLP network schema, by using standard traversal methods such as the BFS (breadth-first search) algorithm. As the network schema is a much smaller graph compared with the original network, this stage is very fast. For co-authorship relation, we extract all the meta paths within a length constraint, say 4, starting and ending with the author type A. The meta paths between authors up to length 4 are summarized in Table II, where the semantic meaning of each relation denoted by each meta path are given in the second column.

Measure Functions on Meta Paths. Once the topologies given by meta paths are determined, the next stage is to propose measures on these meta paths. In this paper, we propose four measures along the lines of topological features in homogeneous networks. These are path count, normalized path count, random walk, and symmetric random walk, which are defined as follows.

- **Path count**. Path count measures the number of path instances between two objects following a given meta path, denoted as PC_R, where R is the relation denoted by the meta path. Path count can be calculated by the products of adjacency matrices associated with each relation in the meta path.
- **Normalized path count**. Normalized path count is to discount the number of paths between two objects in the network by their overall connectivity, and is defined as $NPC_R(a_i, a_j) = \frac{PC_R(a_i,a_j)+PC_{R^{-1}}(a_j,a_i)}{PC_R(a_i,\cdot)+PC_R(\cdot,a_j)}$, where R^{-1} denotes the inverse relation of R, $PC_R(a_i, \cdot)$ denotes the total number of paths following R starting with a_i, and $PC_R(\cdot, a_j)$ denotes the total number of paths following R ending with a_j. $PC_R(a_i, \cdot)$ and $PC_R(\cdot, a_j)$ can be viewed as degrees of a_i and a_j in the network respective to R and R^{-1}.
- **Random walk**. Random walk measure along a meta path is defined as $RW_R(a_i, a_j) = \frac{PC_R(a_i,a_j)}{PC_R(a_i,\cdot)}$, which is a natural generalization of PropFlow [8].
- **Symmetric random walk**. Symmetric random walk considers the random walk from two directions along the meta path, and defined as $SRW_R(a_i, a_j) = RW_R(a_i, a_j) + RW_{R^{-1}}(a_j, a_i)$.

We now use the example in Fig. 3 to show the calculation of these measures. Let R denote the relation represented by meta path $A-P-V-P-A$. It is easy to check it is symmetric, *i.e.*, $R = R^{-1}$. Let J denote Jim, and M denote Mike. We can see that $PC_R(J, M) = 7$, $NPC_R(J, M) = \frac{7+7}{7+9} = 7/8$, $RW_R(J, M) = 1/2$, $RW_R(M, J) = 7/16$, and $SRW_R(J, M) = 15/16$.

For each meta path, we can apply any measure functions on it and obtain a unique topological feature. In the experimental section, we will compare the different topological features for the co-author relationship prediction task.

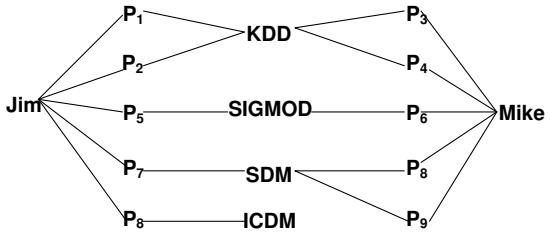


Figure 3. An Example of A-P-V-P-A Paths Between Two Authors

B. The Co-authorship Prediction Model

Second, we introduce the relationship prediction model which models the probability of co-authorship between two authors as a function of topological features between them. Given the training pairs of authors, we first extract the topological features for them, and then build the prediction model to learn the weights associated with these features.

In this paper, we choose the standard method, namely, the logistic regression model as the prediction model. For each training pair of authors $\langle a_{i_1}, a_{i_2} \rangle$, let $\mathbf{x}_i$ be the $(d+1)$-dimensional vector including constant 1 and d topological features between them, and y_i be the label of whether they

will be co-authors in the future ($y_i = 1$ if they will be co-authors, and otherwise 0), which follows binomial distribution with probability p_i. The probability p_i is modeled as follows:

$$p_i = \frac{e^{\mathbf{x}_i \boldsymbol{\beta}}}{e^{\mathbf{x}_i \boldsymbol{\beta}} + 1}$$

where $\boldsymbol{\beta}$ is the $d + 1$ coefficient weights associated with the constant and each topological feature. We then use standard MLE (Maximum Likelihood Estimation) to derive $\hat{\boldsymbol{\beta}}$, that maximizes the likelihood of all the training pairs: $L = \prod_i p_i^{y_i}(1 - p_i)^{(1-y_i)}$.

IV. Experiments

In this section, we show that our proposed meta path-based topological features can improve the co-authorship prediction accuracy compared with the baselines that only use homogeneous object and link information.

A. Dataset

The DBLP bibliographic network, which has been introduced in Sec. II-A, is used for experiments. This network contains $1632K$ papers published before 2010, $1037K$ authors, $7.7K$ venues, and $280K$ topics that have appeared more than 5 times in the paper titles.

B. Experiment Setting

We consider three time intervals for the network, according to the publication year associated with each paper: $T_0 = [1989, 1995]$, $T_1 = [1996, 2002]$, and $T_2 = [2003, 2009]$. For the training stage, we use T_0 as the past time interval, and T_1 as the future time interval, which is denoted as $T_0 - T_1$ time framework. For the test stage, we consider the same time framework $T_0 - T_1$ for most of the studies, and consider $T_1 - T_2$ time framework for the model generality test (Sec. IV-D) and the query-based test (Sec. IV-E).

Let an author pair be $\langle a_i, a_j \rangle$, we call a_i the source author, and a_j the target author. Two sets of source authors are considered. The first set is comprised of highly productive authors, who has published no less than 16 papers in the past time interval; and the second set is comprised of less productive authors, with between 5 and 15 publications. We confine the target authors that are relatively close to the source authors, to avoid the excessive computing between authors that are unrelated. The target authors are selected if they are 2-hop co-authors or 3-hop co-authors of the source author. For each source author set under each target author constraint (2-hop or 3-hop co-authors), we first find all the source authors that have new relationships building with existing authors in the future time interval, and use these new relationships as positive training pairs. We also sample an equal sized set of negative pairs. Therefore, in the training dataset, the sizes of positive pairs and negative pairs are balanced. We summarize the training datasets in Table III. It can be noticed that, highly productive authors are more likely to co-author with authors within a small distance than the less productive authors (64.91% of the highly productive authors have new relationships building with 2-hop co-authors, while only 36.58% of the less productive authors build new relationships with their 2-hop co-authors). We will study other behavior differences for the two groups of sources authors in the following parts. In all, we have four labeled datasets: (1) the highly productive source authors with 2-hop target authors (denoted as $HP2hop$); (2) the highly productive source authors with 3-hop target authors (denoted as $HP3hop$); (3) the less productive source authors with 2-hop target authors (denoted as $LP2hop$); and (4) the less productive source authors with 3-hop target authors (denoted as $LP3hop$).

To evaluate the prediction accuracy, two measures are used. The first measure is the classification accuracy rate (accuracy) for binary prediction under the cut-off score as 0.5, and the second one is the area under ROC curve (AUC).

C. Overall Accuracy

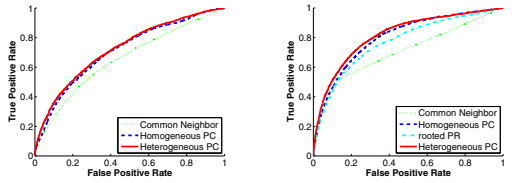
In this section, we evaluate the accuracy of our methodology on the four datasets, using a 10-fold cross validation.

We first compare the heterogeneous topological features with the homogeneous ones. For the heterogeneous topological features, we use path count measure for 9 meta paths (denoted as heterogeneous PC) listed in Table II (not including the target relation itself); for homogeneous topological features, we use (1) the number of common co-authors, (2) the rooted PageRank ([7]) with restart probability $\alpha = 0.2$ for the co-author sub-network, and (3) the number of paths between two authors of length no longer than 4, disregarding their different meta paths (denoted as homogeneous PC). The rooted PageRank measure is only calculated for the $HP3hop$ dataset, due to its inefficiency in calculation for large number of authors. The comparison results are summarized in Fig. 4 and Table IV. We can see that the heterogeneous topological feature beats the homogeneous ones in all the four datasets, which validates the necessity to consider the different meta paths separately in heterogeneous networks. We also notice that, in general the co-authorship for highly productive authors is easier to predict than less productive authors, by looking at the overall prediction accuracy on the two groups of source authors. Finally, we can see that the prediction accuracy is higher when the target authors are 3-hop co-authors, which means the collaboration between closer authors in the network is more affected by information that is not available from network topology.

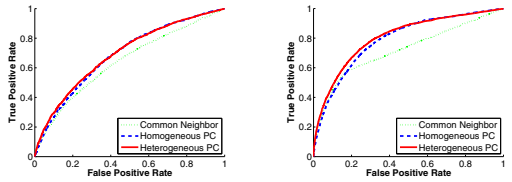
Second, we compare different measures proposed for heterogeneous topological features in Sec. III-A: (1) the path count (PC), (2) the normalized path count (NPC), (3) the random walk (RW), (4) the symmetric random walk (SRW), and (5) the hybrid features of (1)-(4)($hybrid$). The results for the four datasets are shown in Fig. 5. It turns

Table III
FOUR TRAINING DATASETS IN TIME FRAMEWORK $T_0 - T_1$ SUMMARIZATION

Source author type	Constraint	# Source authors	# Source author with new relationships	# New relationships	# Avg. target authors
highly productive	2-hop✓	2538	1548 (64.91%)	4986 (19.43%)	159.01
	3-hop✓	2538	1860 (77.99%)	9215 (35.91%)	930.65
	no	2538	2385 (100%)	25661 (100%)	119246
less productive	2-hop✓	13075	3367 (36.58%)	6189 (12.51%)	47.97
	3-hop✓	13075	4333 (47.08%)	10710 (21.64%)	271.06
	no	13075	9204 (100%)	49483 (100%)	119246

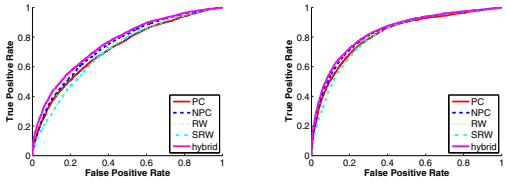


(a) $HP2hop$ (b) $HP3hop$

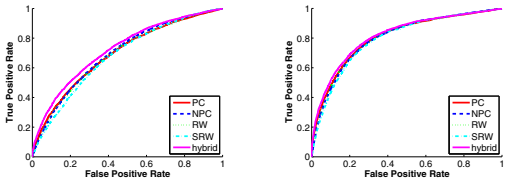


(c) $LP2hop$ (d) $LP3hop$

Figure 4. Homogeneous Features vs. Heterogeneous PC Feature



(a) $HP2hop$ (b) $HP3hop$

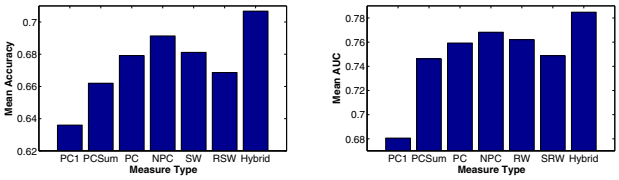


(c) $LP2hop$ (d) $LP3hop$

Figure 5. Comparison of Different Heterogeneous Features

Table IV
HOMOGENEOUS TOPOLOGICAL FEATURES VS. HETEROGENEOUS ONES

Dataset	Topological features	Accuracy	AUC
$HP2hop$	common neighbor	0.6053	0.6537
	homogeneous PC	0.6433	0.7098
	heterogeneous PC	**0.6545**	**0.7230**
$HP3hop$	common neighbor	0.6589	0.7078
	homogeneous PC	0.6990	0.7998
	rooted PageRank	0.6433	0.7098
	heterogeneous PC	**0.7173**	**0.8158**
$LP2hop$	common neighbor	0.5995	0.6415
	homogeneous PC	0.6154	0.6868
	heterogeneous PC	**0.6300**	**0.6935**
$LP3hop$	common neighbor	0.6804	0.7195
	homogeneous PC	0.6901	0.7883
	heterogeneous PC	**0.7147**	**0.8046**



(a) Mean accuracy (b) Mean AUC

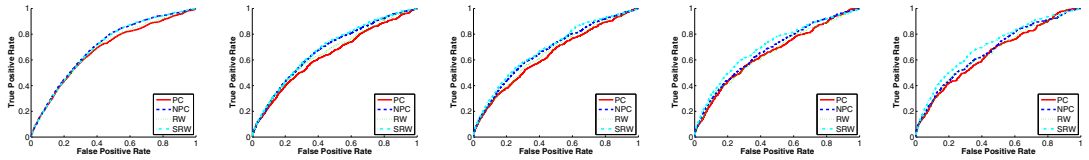
Figure 6. Average Accuracy over 4 Datasets for Different Features

out that in average (see Fig. 6): (1) all the heterogeneous features beat the homogeneous features (common neighbor is denoted as $PC1$, and homogeneous PC is denoted as $PCSum$); (2) the normalized path count beats all the other three individual measures; and (3) the hybrid feature produces the best prediction accuracy.

Third, we compare the accuracy of our model under different strengths of the relationship definition. In the previous cases, we say two authors have a co-authorship if they have co-authored *one* paper. Here, we study the relationships defined by different collaboration frequency. From Fig. 7, we can see that, the measure symmetric random walk is more important in deciding high frequency co-author relationships. In other words, two authors who can be reached with high probability mutually in the network will be more likely to build strong collaboration relationships.

D. Model Generalization

We now test the model generalization over different time periods. In reality, we may need to train the model using $T_0 - T_1$ time framework, but apply the model to a different time framework with a shift ΔT. In our case, we consider the time shift as 7 years, namely the $T_1 - T_2$ framework. In other words, we want to see whether the model trained 7 years ago can still produce reasonable prediction results according to the new topological features. We can see that, the accuracy of the prediction for using last time framework as training is comparable with results using the same term training. Notice that, the accuracy rate using a cut-off of 0.5



(a) $freq > 1$ (b) $freq > 2$ (c) $freq > 3$ (d) $freq > 4$ (e) $freq > 5$

Figure 7. Impacts of Collaboration Frequency on Different Measures

Table V
MODEL GENERALIZATION TEST OVER TIME EVOLVING

Training framework	Test framework	Prediction Accuracy	
		Accuracy	AUC
$T_0 - T_1$	$T_0 - T_1$	0.7368	0.8211
$T_0 - T_1$	$T_1 - T_2$	0.7123	0.8325
$T_1 - T_2$	$T_1 - T_2$	0.7442	0.8313

Table VI
SIGNIFICANCE OF META PATHS WITH NORMALIZED PATH COUNT MEASURE FOR $HP3hop$ DATASET

Meta Path	p-value	significance level[1]
$A - P \rightarrow P - A$	0.0378	**
$A - P \leftarrow P - A$	0.0077	***
$A - P - V - P - A$	1.2974e-174	****
$A - P - A - P - A$	1.1484e-126	****
$A - P - T - P - A$	3.4867e-51	****
$A - P \rightarrow P \rightarrow P - A$	0.7459	
$A - P \leftarrow P \leftarrow P - A$	0.0647	*
$A - P \rightarrow P \leftarrow P - A$	9.7641e-11	****
$A - P \leftarrow P \rightarrow P - A$	0.0966	*

[1] *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$, ****: $p < 0.001$

is underestimated as predicted scores have a shift due to the growth of the network; while the measure of AUC is more trustable as it considers all possible cut-offs.

E. Case Study

For the case study, we first show the learned importance for each topological feature in deciding the relationship building in DBLP, and then show the predicted co-author relationships for several source authors in a query mode.

First, we show the learned importance for all the 9 meta paths with NPC measure, as NPC is the best measure for co-author relationship prediction overall. We show the p-value for the feature associated with each meta path under Wald test and their significance level in Table VI. From the results, we can see that for the $HP3hop$ dataset, the shared co-authors, shared venues, shared topics and co-cited papers for two authors all play very significant roles in determining their future collaboration(s). For the asymmetric meta paths that represent the asymmetric relations, such as citing and cited relations between authors, they have different impacts in determining the relationship building. For example, for a highly productive source author, the target authors citing her frequently are more likely to be her future co-authors than the target authors being cited by her frequently.

Table VII
TOP-6 SIGNIFICANT TOPOLOGICAL FEATURES IN HYBRID MEASURE SPACE FOR $HP3hop$ DATASET

top-k	Meta Path + Measure	p-value
1	$A - P - V - P - A + NPC$	3.12e-38
2	$A - P - A - P - A + SRW$	2.14e-27
3	$A - P - T - P - A + NPC$	1.54e-13
4	$A - P - A - P - A + RW$	2.14e-06
5	$A - P - V - P - A + SRW$	0.0001
6	$A - P \leftarrow P \rightarrow P - A + PC$	0.0008

Table VIII
QUERY AUTHOR SUMMARIZATION

Query author	# Candidates	# True relationships
Jiawei Han	11934	36
Christos Faloutsos	12945	45
Charu Aggarwal	5166	12
Jian Pei	4809	42
Xifeng Yan	1617	8

For the case of using hybrid features, we list the top-6 featured denoted as the combination of meta paths and measures for $HP3hop$ dataset in Table VII.

Second, we study the predicted co-authors for several source authors as queries. Notice that, predicting co-authors for a given author is an extremely difficult task, as we have too many candidate target authors (3-hop candidates are used), while the number of real new relationships are usually quite small. The statistics for the query authors in $T_1 - T_2$ framework and the recall at position 50 for the predicted results using training in $T_0 - T_1$ framework are summarized in Table VIII and Table IX. We can see that compared with random prediction and using the homogeneous feature of shared common authors, the model using our hybrid heterogeneous topological features gives the best overall performance. Table X shows the top-10 predicted co-authors in time interval T_2 (2003-2009) using the $T_0 - T_1$ training framework, for both the proposed hybrid topological features and the shared co-author feature. It turns out that the previous feature predicts more real relationships by considering multiple factors.

V. RELATED WORKS

The link prediction problem has been studied on homogeneous networks extensively. The earliest works mainly study unsupervised methods [1], [7], in which different similarity

Table IX
$Recall@50$ COMPARISON

Query author	Hybrid Features	Random	# Shared authors
Jiawei Han	0.1111	0.0042	0.0833
Christos Faloutsos	0.0889	0.0039	0.1111
Charu Aggarwal	0.4167	0.0097	0.3333
Jian Pei	0.2619	0.0104	0.2619
Xifeng Yan	0.875	0.0309	0.5
Avg.	**0.3507**	0.0118	0.2579

Table X
TOP-10 PREDICTED CO-AUTHORS FOR JIAWEI HAN

Rank	Hybrid features	# Shared authors
1	**Hans-Peter Kriegel**	Elisa Bertino
2	Christos Faloutsos	Sushil Jajodia
3	Divesh Srivastava	Hector Garcia-Molina
4	H. V. Jagadish	**Hans-Peter Kriegel**
5	Bing Liu[1]	Christos Faloutsos
6	Johannes Gehrke	Divyakant Agrawal
7	George Karypis	Elke A. Rundensteiner
8	**Charu C. Aggarwal**	Amr El Abbadi
9	Mohammed Javeed Zaki	Krithi Ramamritham
10	Wynne Hsu	Stefano Ceri

[1] Although not included in the time interval T_2, Bing Liu co-authored with Jiawei in Year 2010.

measures are constructed from topological structure of the networks or from the object attributes, and are compared to see whether they are consistent with the future link appearance. Subsequently, supervised methods were proposed which combine different features with different coefficients via training data sets [4], [15], [8]. Some recent work [6] has discussed the link prediction problem when the network is not fully observed and thus is modeled in a probabilistic way. A good survey on link prediction may be found in [3]. In this paper, we extend the link prediction problem to more general heterogeneous networks by exploring the topological features in such scenarios.

Another line similar to our problem is the link prediction task in relational data [11], [14], as relational data also involves different types of objects and complex relationships between objects. However, these studies have a different focus compared with our paper. As in [11], they study feature selection in a relational environment using relational languages, and feed these features into supervised link prediction models. In [14], the authors focus on modeling the relational data via a probabilistic model, which relies on the attributes of the objects, and the links are used to capture the dependency relation among different variables. In our paper, we aim at designing a model for relationship building by systematically exploring the topological features in the heterogeneous networks.

VI. CONCLUSIONS

In this paper, we study the problem of predicting co-author relationship among authors in heterogeneous bibliographic networks. In comparison with traditional homogeneous networks, heterogeneous networks contain multiple types of objects and links. We propose the *PathPredict* model to address this problem, which first defines meta path-based topological features in such networks, and then builds logistic regression-based co-authorship prediction model. Experiments on the DBLP bibliographic network show that by considering heterogeneous topological features, the relationship prediction accuracy can be significantly improved, and the model using hybrid features that have combined different meta paths and different measures gives the best overall performance. Furthermore, the learned significance for each topological feature can provide better understanding of the relationship building mechanism in such networks.

REFERENCES

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

[2] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311:590–614, 2002.

[3] L. Getoor and C. P. Diehl. Link mining: a survey. *SIGKDD Explor. Newsl.*, 7:3–12, December 2005.

[4] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM '06*, 2006.

[5] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39 – 43, 1953.

[6] V. Leroy, B. B. Cambazoglu, and F. Bonchi. Cold start link prediction. In *KDD '10*, 2010.

[7] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03*, 2003.

[8] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *KDD '10*, 2010.

[9] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review Letters E*, 64, 2001.

[10] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. Prefixspan: Mining sequential patterns by prefix-projected growth. In *ICDE '01*, 2001.

[11] A. Popescul, R. Popescul, and L. H. Ungar. Statistical relational learning for link prediction. In *Proc. of the Workshop on Learning Statistical Models from Relational Data at IJCAI-2003.*, 2003.

[12] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *VLDB' 11*, 2011.

[13] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *KDD '08.*

[14] B. Taskar, M. fai Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *NIPS '03*, 2003.

[15] C. Wang, V. Satuluri, and S. Parthasarathy. Local probabilistic models for link prediction. In *ICDM '07*, 2007.