2012 IEEE 12th International Conference on Data Mining

Predicting Links in Multi-Relational and Heterogeneous Networks

Yang Yang and Nitesh Chawla
Department of Computer Science
University of Notre Dame
Notre Dame, Indiana, US
Email: {yyang1,nchawla}@nd.edu

Yizhou Sun and Jiawei Hanı
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois, US
Email: {sun22, hanj}@illinois.edu

***Abstract*—Link prediction is an important task in network analysis, benefiting researchers and organizations in a variety of fields. Many networks in the real world, for example social networks, are heterogeneous, having multiple types of links and complex dependency structures. Link prediction in such networks must model the influence propagating between heterogeneous relationships to achieve better link prediction performance than in homogeneous networks. In this paper, we introduce Multi-Relational Influence Propagation (MRIP), a novel probabilistic method for heterogeneous networks. We demonstrate that MRIP is useful for predicting links in sparse networks, which present a significant challenge due to the severe disproportion of the number of potential links to the number of real formed links. We also explore some factors that can inform the task of classification yet remain unexplored, such as temporal information. In this paper we make use of the temporal-related features by carefully investigating the issues of feasibility and generality. In accordance with our work in unsupervised learning, we further design an appropriate supervised approach in heterogeneous networks. Our experiments on co-authorship prediction demonstrate the effectiveness of our approach.**

I. Introduction

Link prediction, that is, predicting the formation of links in a network in the future or predicting the missing links in a network, has become a hot topic in recent years. Most of the recent link prediction methods [6] [7] [8] [1] are designed for homogeneous networks, where only one type of link exists in the networks. However, many important real-world networks, such as DBLP bibliographic networks and human disease-gene networks, are complicated and modeled as heterogeneous interactions. For example, the DBLP network contains conferences, papers, and authors as nodes, with links from types of co-author, author-write-paper, paper-published in-conference, and so on. A few studies have worked on the link prediction in heterogeneous networks, from the early work of [13] to the recent work of [20].

The complexity of structural dependency and heterogeneity of links produces obstacles for link prediction in heterogeneous networks. Well-known topological features designed for homogeneous networks are difficult to apply in such complex scenarios. There are two typical ways of handling the link prediction problem in heterogeneous networks: 1) treating all types of link equally; 2) studying each type of link independently and ignoring its correlation with other link types

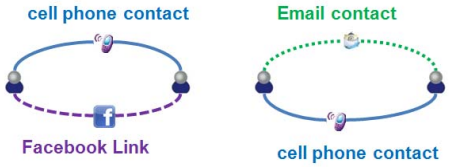

Fig. 1. Correlation between Different Types of Links

[2]. However, both of these methods lead to a loss of information. In particular for the second case, topological properties may differ in each homogeneous projection of heterogeneous networks, but different types of links are correlated with each other and thus influence each other. For example, two people connected in a cellphone network have a high probability to be friends on Facebook. Likewise, two people who send emails to each other frequently may also call each other (Figure 1). The problem is: *how to quantitatively capture the correlation between different types of links and employ this information to design an effective, general method for the link prediction in heterogeneous networks?*

To that end, we develop both unsupervised and supervised learning methods in heterogeneous networks, which are based on the topological structures and timestamps of links (if available). We first introduce our method called the Multi-Relational Influence Propagation (MRIP) for heterogeneous networks. It is motivated from the work of Kempe et al. [4] that which was designed for maximizing the spread of influence in homogeneous networks. Experimental results on the disease-gene network [2] and the DBLP network [15] are presented for comparisons. Then we introduce our approach of temporal link predictors which are time-involved variants of classical link predictors, such as *Common Neighbors* [14] and *Adamic/Adar* [6]. Most of these extended link predictors outperform their original incarnations by more than 9% in terms of AUROC (area under the receiver operating curve). With careful extraction of features and in accordance with our work in unsupervised learning, we design effective supervised learning approaches for link prediction in heterogeneous networks. To summarize, the **contributions** of this paper are as follows:

(a) We study the link prediction problem in heterogeneous networks, where multiple types of link that are correlated

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.144

with each other exist in the network.

(b) We propose a new topological feature called Multi-Relational Influence Propagation (MRIP) that can capture the correlation between different types of links for the link prediction problem.

(c) We further propose temporal features in heterogeneous networks to achieve even better link prediction accuracy.

(d) Experiments on real datasets have demonstrated the effectiveness of our approaches compared with typical solutions and most recently published solutions.

The remainder of the paper is organized as follows. We introduce the preliminary concepts of heterogeneous networks and define the problem in Section III. Section IV explains standard unsupervised approaches and introduces our new method MRIP. Extended temporal link predictors are described in Section V. We show our methods of supervised learning in heterogeneous networks and analyze the experimental results in Section VI, and conclude the study in Section VII.

II. Related Work

A few studies have worked on the link prediction in heterogeneous networks, from the early work of [13] to the recent work of [2] [20]. However in the work of [13] the attribute values of nodes are usually difficult to obtain in real world dataset, thus in this paper our method will not be compared with the method described in the work of [13].

In the work of [2] Davis et al. propsed to explore triad information in heterogeneous network to assist the link prediction task. Their method MRLP is a probabilistically weighted extension of the *Adamic/Adar* measure for heterogenous information networks. The MRLP was proved to be successful in predicting links in heterogeneous networks when comparing with traditional link predictors. In the work of [20] Litchenwalter et al. proposed the concept of a vertex collocation profile (VCP) for the purpose of topological link analysis and prediction. In their definition a vertex collocation profile (VCP), $VCP_{v,u}^{n,r}$ is a vector describing the relationship between two vertices, u and v , in terms of their common membership in all possible subgraphs of n vertices over r relations [20]. In their paper VCP3U is designed to work in undirected homogeneous/heterogeneous networks by describing node pair u and v relationship in all possible subgraphs of 3 vertices, correspondingly VCP4U is employed to capture the information of all possible subgraphs of 4 vertices for any node pair u and v. In this paper our method will be compared with these two most recent work in heterogeneous network.

To mention that in the work of [25] Rossetti et al. proposed multidimensional versions of the Common Neighbors and Adamic/Adar, and derived predictors that aimed at capturing the multidimensional and edge level temporal information. However their methodology of approaches are significantly different from ours. We employ the network alignment technology to capture interrelations between link types, while they use connectivity measure for multidimensional networks to guide their design. In temporal methods design, we are gathering nodal historical data and try to capture the preference

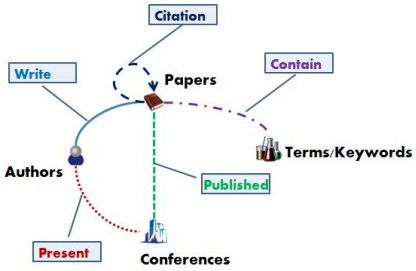

Fig. 2. DBLP Bibliographic Network

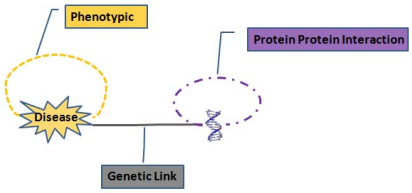

Fig. 3. Disease-Gene Network

of topological features when two nodes are associated by new link; while they are interested in edge level communication data. The work of [25] did not include a comparison with any competitive approaches in recent publications, such as [1] and [3].

III. Concepts and Preliminaries

A. Heterogeneous Network

Given a heterogeneous network, it can be modeled as $G = (V_1 \cup V_2 ... \cup V_M, E_1 \cup E_2 ... \cup E_N)$, where V_u $(u \in N)$ represents the set of nodes of the same type u and E_j $(j \in M)$ represents the set of links with the type j. In the real world, many important networks are heterogeneous. For example in the DBLP bibliographic network there are several types of nodes and links (Figure 2). Another example is the human disease-gene network, which has two kinds of nodes (disease and gene) and three types of links among these nodes (disease-disease phenotypic link, gene-gene PPI link and disease-gene genetic link) (Figure 3). Links, or interactions, in these networks can occur between nodes of the same type, such as co-authorship links and phenotypic links; or occur between nodes of different types, such as links from the author-write-paper relation and genetic links between diseases and genes.

B. The Link Prediction Problem Definition

Given a heterogeneous network $G = (V_1 \cup V_2 ... \cup V_M, E_1 \cup E_2 ... \cup E_N)$, the link prediction task in such networks is to predict whether there is or will be a link of type i $(i = 1, 2, ..., N)$ between a pair of nodes u and v, where $u \in V_x$ and $v \in V_y$. In the unsupervised link prediction scenario, the problem is to assign a score $s(u, v, i)$ that indicates the possibility of a link between nodes u and v, where i is the link type. While in the supervised learning scenario, the target is to answer whether a link of type i will form between two given nodes u and v.

C. Preliminaries

We first briefly introduce the work of Kempe et al. [4] for influence maximization in homogeneous network. The *influence maximization* problem was originally proposed in 2003 by Kempe, Kleinberg and Tardos [4] as finding the k most influential nodes in a social network under some stochastic cascade models, such as the *independent cascade* (IC) model or the *weighted cascade* (WC) model. In *weighted cascade* model node v is assumed to activate its neighbor u with the probability

$$p_{v,u} = 1 - (1 - \frac{1}{degree(u)})^{weight(u,v)}$$

To find out k most influential nodes in *linear threshold* model for the network, many researchers employ breadth-first search procedure to propagate the probability of activating from the source node v to any reachable node u, the score $s_{v,v}$ is initially assigned probability 1. Throughout this procedure, the influence probability between each pair of nodes u and v will be recorded to support the mining of the top k most influential nodes. In the evolution of the network, influence and link formation are closely interrelated. The link between two nodes forms due to the reason that the mutual influence between them is strong enough, while in turn the formation of link enhances the influence between these two nodes.

This inspires us to use such influence score $s_{v,u}$ as an estimation of the likelihood for new links (Equation 1). There is minor change in the equation 1, we use $\frac{weight(v,u)}{degree(v)}$ as the probability of activating for the simplicity of computation.

$$flow\,(v,u) = score\,(v) \cdot \frac{weight\,(v,u)}{degree\,(v)} \tag{1}$$

However this methodology is designed for the homogeneous network, much work need to be considered to make it feasible in the heterogeneous network. While MRIP could employ the similar method of propagating probability, it also brings up a fundamental question: *how to propagate influence/probability in the heterogeneous network?* This question has not yet been answered in typical *influence maximization* research area. Our solution to this problem will be discussed in Section IV.

IV. The MRIP Method

In this section, we introduce our MRIP method for unsupervised link prediction in detail, and show its effectiveness using real-world examples and compare it with several baselines [10] and most recent work of [2].

A. Baseline Predictors

Most approaches to link prediction are based on measures that analyze the proximity of nodes in the network. Feature-based link prediction methods can be categorized as: 1) methods based on node neighbors; 2) methods based on ensemble of all paths. In category 1 there are several baseline predictors, such as *Common Neighbors* [14], *Jaccard Coefficient*, *Adamic/Adar* [6] and *Preferential Attachment* [7]. In category 2 a number of methods refine the notion of shortest-path distance by implicitly considering the ensemble of all paths between two nodes.

B. MRIP

In last section we bring up a fundamental question of our method design: *how to propagate influence/probability in the heterogeneous network?*. To solve this problem we need to know the relations between any given pairs of edge types i and j. However, it leads to additional questions:

(a) How do we represent the relationship between any given pair of edge types i and j quantitatively?
(b) Is the relationship between two edge types i and j symmetric or asymmetric?

We propose the following solutions to the aforementioned questions, which will be detailed in Section IV.

1) Link Interrelation: There are many work in the analysis of multi-relational and heterogeneous networks, and several measures/notions of correlation among dimensions/relations are proposed in the work of [26] [27] [28]. Our solution is motivated by the network alignment work in biological research, which encompasses interactions of different link types to study their interrelations [22] [23]. An important evaluation metric for network alignment algorithms is called *edge correctness*, which measures the percentage of correctly aligned edges.

We use the conditional probability $probability(i|j)$ to represent the correlation between link type i and link type j. The conditional probability $probability(i|j)$ is equivalent to the *edge correctness* measurement for the network alignment, which is a simple and effective method for link types interrelation study. A toy example is given in Figure 5. In Figure 5 the conditional probability is equivalent to the *edge correctness* when we construct an alignment from one network to another. For example when we map facebook network (5 edges) to cellphone network (3 edges), only one out of five edges are correctly aligned, thus $probability(c|f) = 0.2$; when we map cellphone network to facebook network, there are one out of three edges are correctly aligned, such that $probability(f|c) = 0.33$.

The *edge correctness* captures the interrelation between one-hop distant node pairs (edges in the network) in different dimensions, this method can be easily extended to measure the correlation between two-hop distant node pairs or three-hop distant node pairs. By scaling the hop distance between node pairs, for h-hop distant node pairs of link type i and link type j, we can obtain a correlation value $C^h_{i,j}$ (Figure 4), describing h-hop distant node pairs interrelation between link type i and link type j. And then we can construct a correlation vector $\vec{CV}_{i,j}$ (Equation 2) for each pair of link type i and link type j, which describes the relationship between two link types, i and j , in terms of their "*edge correctness*" in all possible hop distance h.

$$\vec{CV}_{i,j} = (C^1_{i,j}, C^2_{i,j}, ..., C^h_{i,j}, ...) \tag{2}$$

This vector can be useful in describing the topological similarity of different dimensions in the heterogeneous network.

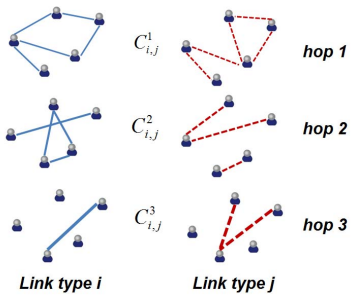

Fig. 4. Examlpes of Correlation Values

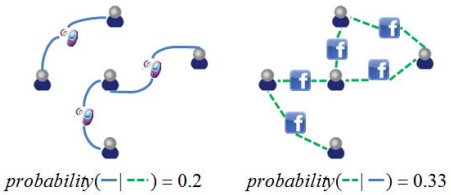

Fig. 5. Toy Example for Correlation between Different Link Types

In this paper we only use the value $C^1_{i,j}$ (denoted as $probability(i|j)$ in following sections) to describe the correlation of link types. In future our method MRIP can be trivially extened to employ the information of $C^h_{i,j}$, $h > 1$.

2) Asymmetric Interrelation: Intuitively for any given pairs of link types i and j, $probability(i|j)$ should be different from $probability(j|i)$. For example, while it may be likely for two friends to call each other, two people who call each other are not necessarily friends.

MRIP method designation is based on the following considerations.

(a) For any given link type i, the influence propagates not only through the links of type i but also propagates through other types of links.
(b) The probabilities that propagate through other link type j depend on the correlation between link type i and link type j.

To that end, we modify Equation 1.

$$flow(v,u,i) = score(v)\cdot\beta\cdot\frac{weight(v,u,i)}{degree(v,i)} + score(v)\cdot\beta\cdot\sum_{j\neq i}^{K}(\sigma(i,j)\cdot\frac{weight(v,u,j)}{degree(v,j)})/(|E(v,u)|-1) \quad (3)$$

where v and u are nodes, $\beta = 0.05$ is the $katz$ [9] factor, $\sigma(i,j)$ is the $probability(i|j)$, and $|E(v,u)|-1$ is the number of link types between node v and u except type i (Figure 6).

MRIP employs breadth-first search procedure to propagate the probability. Therefore $score(v)$ is the probability of a link between the source node (breadth-first search source node) and node v. The $katz$ factor is included into design to penalize the case described in Figure 6 (b). The long distance propagation

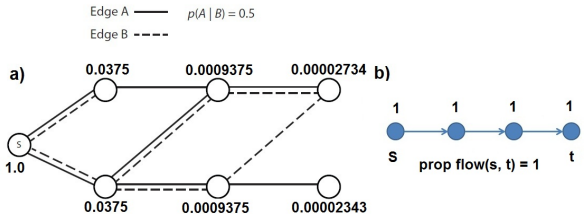

Fig. 6. A conceptual overview of our Multi-relational Influence Propagation algorithm. Flow propagates outward from the source node S.

will also be penalized by the $katz$ factor. Additionally in the equation the top part is the influence score flowing through link type i, while the bottom part is the "hidden" information propagating through other types of links, such as type j. If link type j has a significant correlation with link type i, the "hidden" information flowing through it is also large. The contribution from link type j to link type i depends on its own network structure and its correlation with link type i. If there are multiple types of links related to link i, we take the mean of the scores propagating through them as the 'bonus' part; if there is no other link type ($|E(v,u)| = 1$), there is zero 'bonus' from other link types.

C. Discussion

In our current work we only employ the *weighted cascade* model in our designation, actually the influence score propagating from source node v to target node u is different in different influence spreading models. Our future work will find out which model works best in the real world network. In recent work of [2] Davis et al. employed triad information to capture the interrelations between different link types, however the expense of computation is not so feasible in large-scale network. As for our method we employ the dyad information to study the interrelations, the complexity of computation is reduced. Additionally in unsupervised experiments we can show that the performance of our method is comparable to or better than MRLP [2] in the disease-gene network. The time complexity of calculating conditional probability is $O(|E|)$, and the MRIP algorithm takes $O(|V|.|E|)$ time for all reachable node pairs. The MRIP complexity can be reduced heavily when we restrict the propagation within h hops.

The computation and evaluation for all possible links on large networks is infeasible, due to multiple computational reasons including time and storage capacity. The work of [1] and [21] both proposed that the link prediction within a short hop geodesic distance (i.e. 2 hops or 3 hops) provides much greater baseline precision in many networks. Effectively predicting links within this set offers a strong indicator of reasonable deployment performance. In this paper for all methods metioned, we restricted the prediction task within the set of three hops node pairs due to their higher prior probability of formation and computational feasibility. And in this case the computation complexity of our MRIP method is reduced significantly when the hop distance is within three.

D. Dataset

We use two real-world heterogeneous networks to demonstrate the effectiveness of our methods in this paper.

1) DBLP: Based on the DBLP dataset from [15], we attach timestamps for each activity in the data and choose 3,215 authors who published at least 5 papers in conferences relating to four areas (Data Mining, Database, Information Retrieval, and Machine Learning) between 1990 and 2010. There are four types of nodes—authors, papers, conferences, and terms—with the network relation structure described in Figure 2. In this paper we focus on the co-author relation (collaborate on paper), common terms relation (publications have similar terms) and common conference relation (show up in the same conference in the same year, physical proximity). For unsupervised learning, we choose data between 1990 and 2000 as our training set, and data between 2001 and 2005 as testing set. While for supervised learning, data between 1990 and 2000 is employed as feature set, data between 2001 and 2005 is used as label set, and data between year 2006 and 2010 is the testing set.

2) Disease-Gene Network: The disease-gene (DG) network was constructed from three individual datasets from [2]. As the name suggests, this network has two distinct node types, diseases and genes, with three link types connecting them as described in Figure 3. This dataset was only used to evaluate unsupervised learning experiments due to the reason that we only have the same unsupervised learning setting from the authors of [2]. The disease-gene network consists of 703 diseases and 1,132 genes, 10,483 genetic links, 10,483 phenotypic links, and 2,450 PPI interactions exist among these diseases and genes.

E. Experimental Results

In order to show the power of our MRIP method in link prediction for heterogeneous networks, we use *Common Neighbor*(CN), *Jaccard Coefficient*(JC), *Adamic/Adar*(AA), *Preferential Attachment*(PA) and *PropFlow* as baselines. For the disease-gene network, we use a 10-fold cross-validation stratified edge holdout scheme. We chose holdout evaluation since longitudinal data was not relevant for the disease-gene network. Link prediction is evaluated for each link type i separately on all eligible node pairs (u, v).

Link prediction performance is evaluated separately for each link type using AUROC, which are shown in Table I and Table II. Methods in bold face indicate the best overall link predictor for the corresponding link type. First, we notice that there is no universally dominant method, which is an expected result since unsupervised link prediction methods are domain-specific [2]. In the disease-gene network, MRIP outperforms the other methods in predicting genetic and PPI links, while it has comparable performance to the top performer in predicting phenotypic links. And we also can see that our MRIP method is comparable to or better than MRLP in all link types of disease-gene network. In the DBLP network MRIP has better performance in predicting co-authorship between authors and predicting new terms/research shared by authors, while it also has a comparable performance in conference presentation prediction. Notice that, for the sparse link types, such as co-authorship, terms, and PPI, MRIP performs better than *PropFlow*, as that other dimensions of link information are considered by MRIP, which undoubtedly improves the effectiveness of MRIP. Generally speaking, MRIP works well in most of link types, and is also stable (comparable to the best predictors if not the best).

TABLE I
AUROC on Disease-Gene Network for Unsupervised Learning

Disease-Gene	PA	PF	JC	CN	AA	MRIP	MRLP [2]
Genetic	0.903	0.951	0.957	0.951	0.956	**0.975**	0.974
Phenotypic	**0.943**	0.762	0.771	0.909	0.911	0.901	0.938
PPI	0.827	0.888	0.786	0.788	0.789	**0.890**	0.808

TABLE II
AUROC on DBLP Network for Unsupervised Learning

DBLP	PA	PF	JC	CN	AA	MRIP
Collaboration	0.673	0.676	0.590	0.597	0.596	**0.766**
Conference	0.503	**0.704**	0.701	0.698	0.689	0.691
Terms	0.738	0.742	0.545	0.546	0.532	**0.811**

V. Temporal Feature Based Methods

Taskar et al. [13] employed the attributes of objects to support link prediction tasks in heterogeneous networks. However the attribute information is generally difficult to collect in real-world networks, often due to security reasons and privacy issues. Additionally even if some of this information is available, such as user surveys, it is usually incomplete or unreliable. We need information that can potentially expose users' subconscious behavior, and time is the best choice we have. On the other hand, time is important because network evolution is associated with it. When users in the network make a decision, their activities are tagged by timestamps, which can serve as data for analyzing their behavior patterns. Therefore, the combination of network topology and corresponding temporal information can benefit the link prediction task. However, it is a difficult conclusion to draw without careful investigations.

A. A Simple Case for Temporal Network Analysis

We extract the DBLP co-author network from between 1980 and 2010, and for each year we compute the number of new links constructed, creating a time series as shown in Figure 7 (a). According to the figure, we can see that this time series has a significant trend. To analyze a time series, the preliminary step is to determine that whether there is unit root at significant level. Dickey et al. [16] proposed a method called the *augmented Dickey-Fuller* test, which can tell whether there is unit root in the time series. With *augmented Dickey-Fuller* test we find that the unit root *p-value* is significant (0.99). Thus we know that this time series is non-stationary and that a difference operator needs to be applied to the time series. In this way we can determine an Unit-Root Nonstationary model

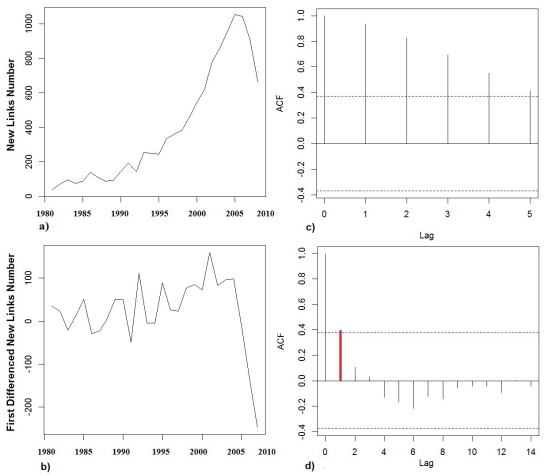

Fig. 7. a) New Links Time Series for DBLP Co-author Network b) First Differenced Time Series c) ACF of New Links Time Series d) ACF of First Differenced New Links Time Series

is suitable for this time series. To analyze such time series, we first take the difference from the original data and then analyze its lag order. For example, if we know that the last step had a decreasing trend, then we may forecast that this trend will continue in the next step. However we don't know how many step-back-look should take without lag test. With the first differenced time series, if we employ *MLE* (maximum likelihood estimation) method to verify its lag order, we find that lag order 1 is the best we have. This is also confirmed by its Autocorrelation Function (ACF) plot in Figure 7 (d).

This observation demonstrates that link formation in networks is strongly associated with time, which serves as a guideline for our method designation. In conclusion: 1) the evolution of network of current step depends on the network in last step at significant level (ACF); 2) link formation is significantly correlated with time and can be modeled with the associated time information.

Besides the global trend, the link formation is also influenced by individual behaviors. An obstacle, however, to analyzing individual behavior is the lack of data. DBLP degree distribution follows power law, which means most nodes have low degrees and thus provide little information for statistical analysis. Bootstrapping technology is also an option to solve this issue. Notice, though, that link formation occurs between two nodes; likewise, if there is enough information for one of the nodes, we can use it to guide our analysis. In the DBLP network, high degree nodes definitely have enough temporal information for analysis, however we need to know what percentage of new links are associated with them. Based on the degree information, we rank the nodes in descending order, and we statistically analyze how many new links in the future are associated with the top $K\%$ of them. In Figure 8 we first rank nodes in the network observed between 1980 and 2000, and compute how many new links between 2001 and 2005 are related to the top $K\%$ of them; we repeat this experiment for the network between 1980 and 2005 and new links within 2006 and 2010. We can see that about 60% of new links are

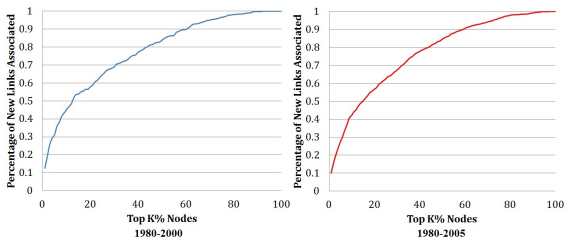

Fig. 8. Top K Percentage Author's and New Links Associated with Them

constructed from the top 20% high degree nodes, which means that in the link formation two end nodes do not play the same role; rather, one of them dominates the formation of the link. This finding guides our design of temporal features in the following sections.

B. Temporal Features and Generalized Temporal Methods

We now introduce the temporal features used in our link prediction solution, and some generalizations of baseline predictors.

1) Recency and Activeness: Recency is proposed by Potgieter et al. [17] and *activeness* is proposed by Huang and Lin [18]. Originally these two features are used to predict the recurrence of links in the future. *Recency* is the length of time elapsed since a node made its last communication, and *activeness* is the number of communications made in last time step. We alter the definition to fit into a new link prediction scenario, where *recency* is the time elapsed since a node made its last new link and *activeness* is the number of new links made in last time step.

2) Degree Preferential Likelihood: This feature is designed to capture the personalized behavior of a node in the network when it's trying to choose another node to form a new link. *Preferential Attachment* suggests that high degree nodes have a high probability of developing a new link. This is true when the network has an outside intermediate to propagate influences between nodes. For example, in an academic co-authorship network, the influence not only propagates through the network, but nontrivial percentage of influence is spread through other media, such as magazines, TV, or newspaper. When such outside media does not exist, such as in a cellphone network, the hypothesis of Preferential Attachment could lead to a poor prediction quality [1]. Thus, in our paper we decompose *Preferential Attachment*. The benefit of decomposition is demonstrated in Figure 11. In traditional *Common Neighbor* method we can not differentiate the likelihood of two node pairs sharing the same number of common neigbors, however by decomposition we can capture the preference of individual nodes and then distinguish that one node pair has larger probability to occur than the other.

Definition 1. *The degree preferential vector of a node v is a sequence of historical data that describes the degree of node v's neighbor u when v and u form a link.*

For each node in the network we record a *degree preferential vector* as shown in Figure 10. Then given two nodes u, v

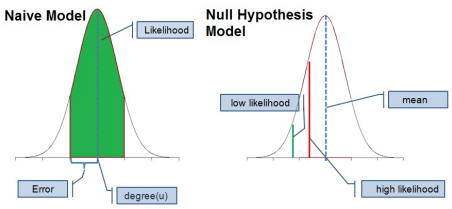

Fig. 9. Two Models

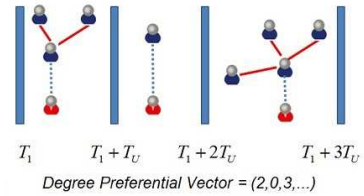

Fig. 10. Degree Preferential Vector

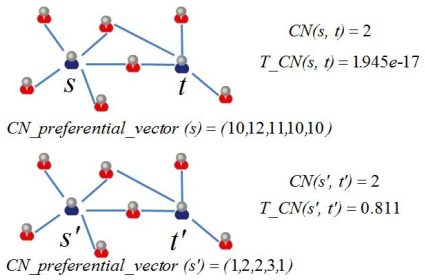

Fig. 11. Decomposition Benefit

and their *degree preferential vectors*, we can compute the $prob(u|vector(v))$ and $prob(v|vector(u))$. In this paper we employ two methods to calculate the $prob(u|vector(v))$:

1) Naive Model:

$$prob(u|vector(v)) = \frac{|x|, x \in [\alpha, \beta], x \in vector(v)}{|vector(v)|}$$
$$\alpha = degree(u) - std(vector(v))$$
$$\beta = degree(u) + std(vector(v)) \tag{4}$$

2) Null Hypothesis Model:

$$prob(u|vector(v)) = p\_value(t\_ratio)$$
$$t\_ratio = \frac{degree(u) - mean(vector(v))}{std(vector(v))} \tag{5}$$

The obstacle to compute $prob(u|vector(v))$ is that the probability of a single value is zero for a given continuous PDF function. For model 1, the hypothesis is, if node u has a high probability of being selected by node v for link formation, then the scope (α, β) should cover a large portion of values in the *degree preferential vector*. This is the most naive way to estimate the probability that node u selects node v. In model 2, we have a null hypothesis H_0 : the mean value of $vector(v)$ is $degree(u)$. The p-value can reveal the significant level of this null hypothesis H_0. Our heuristic, the probability of link

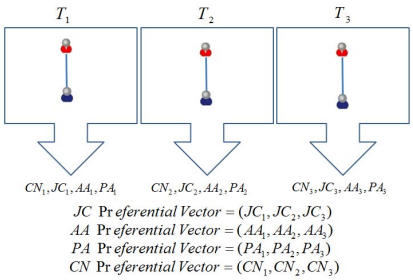

Fig. 12. Temporal Methods

formation between nodes u and v, depends on the closeness between $degree(u)$ and $mean(vector(v))$.

Based on our findings in Section V-A, one node will dominate the formation of a link. Thus we define the likelihood of link between nodes u and v as:

$$degree\_preferential(u, v) =$$
$$\begin{cases} prob(u|vector(v)) \cdot prob(v|vector(u)) \text{ if } d(u), d(v) \geqslant \alpha \\ prob(u|vector(v)) \text{ if } d(u) < \alpha, d(v) \geqslant \alpha \\ prob(v|vector(u)) \text{ if } d(v) < \alpha, d(u) \geqslant \alpha \\ 0 \text{ if } d(v) < \alpha, d(u) < \alpha \end{cases} \tag{6}$$

α is a threshold that determines whether a node has enough historical data for analysis, if $degree(u) < \alpha$ we consider the other node v dominates the preference, while if both of them do not have enough historical data the likelihood is assigned probability zero.

Both models assume that if properties of two nodes u and v fit well then there is a high chance that a link will form. However in real world networks node u degree may be very close to the mean of $vector(v)$, in which case it is still possible that they fail to link each other. If we can statistically collect this information and develop a more decent model, temporal methods discussed above may achieve better performance.

3) Temporal Methods: To generalize static baseline predictors into methods including time information, eligible methods should meet some requirements:

1) Simplicity: The original baseline predictors should be simple in complexity, such as *Common Neighbor* and *Jaccard Coefficient*. The paths involving methods are too complex for computation when considering time, i.e., *Katz*.
2) Generality: These original baseline predictors should reveal the generality of nodes behavior, such as *Preferential Attachment*, which describes the general behavior pattern when people are choosing collaborators.

Based on above requirements we select *Common Neighbor*(CN), *Jaccard Coefficient*(JC), *Adamic Adar*(AA) and *Preferential Attachment*(PA) for generalization. Similar to the degree preferential vector (Figure 10) we can collect the *common neighbor preferential vector* for each node, and then compute the common neighbor perferential likelihood score for each pair of nodes u, v. We can generalize JC, AA and

PA in the same way. (Figure 12)

4) Heterogeneous Network Discussion: We can also develop some temporal features by using other types of link information. For example, we can collect the common conferences number when two authors construct a new link, or gather the information about how similar their publications terms are when a link forms. In this way we create a feature called *temporal Conference* to measure the likelihood of link formation between two given nodes u, v. Which link type is selected to construct such a temporal feature depends on its correlation with our predicting link type. A similar method like meta-path [3] can be employed to evaluate the significance of the correlation.

C. Unsupervied Experimental Results

In this section we perform the experiments for temporal methods introduced above, and give their performance measured by AUROC in Table III. From the table we can see that the temporal method generally yields better results than their static counterparts. For example *temporal Common Neighbor* outperforms *Common Neighbor* by 9% in terms of AUROC. And *temporal Conference* performance (0.681) is even better than *PropFlow* (0.676). The *temporal degree pereferential likelihood* does not outperform *Preferential Attachment*, but is still comparable. Additionally, we would like to note that all of these temporal methods have very low complexity of computation, however they achieve better or comparable performance compared with *PropFlow*.

TABLE III
AUROC Comparison on DBLP Co-author Network

Temporal Predictor	T_PA	T_CN	T_JC	T_AA	T_Conf
AUROC	0.670	**0.651**	**0.648**	**0.650**	**0.681**
Static Predictor	PA	CN	JC	AA	PF
AUROC	**0.673**	0.597	0.590	0.596	0.676

In these tables **T_PA** states for *Temporal Preferential Attachment* method, **T_CN** is the *Temporal Common Neighbor* method, **T_JC** is the *Temporal Jaccard Coefficient* method, **T_AA** represents the *Temporal Adamic Adar* method, and **T_Conf** is the *Temporal Conference* method.

In conclusion if temporal information is considered, then we can achieve better predictive performance by generalizing static baseline predictors, both on homogenous and heterogeneous networks.

VI. Integrating MRIP and Temporal Features in a Unified Supervised Model

We study the performance of supervised classification in several contexts: first, the performance of link prediction, if we involve multi-relational features, such as MRIP; second, the performance of link prediction, when temporal features are included; third, the performance of predictors that combine both. In order to show the power of using temporal features and MRIP feature in link prediction, we use bagging with logistic regression (*WEKA*, default parameters) [12] for our frameworks - *Temporal Model* and *MRIP Model*, that is frequently used in binary link prediction tasks as the baseline. And we also include the bagging with *WEKA* random forests (10 trees, default parameters) for HPLP (High Performance Link Prediction framework) [1] that incorporates powerful homogeneous features listed in Table IV, in our paper we denote it as *Homo Model*. Note that for VCP3U and VCP4U we use bagging with random subspace [24] as described in the paper of [20]. VCP3U and VCP4U are the most recent supervised learning models that can work in the heterogeneous network. We undersample training set to 30% positive class prevalence in training. We do not change the size or distribution of the testing data.

A. Temporal Model, MRIP Model and Homo Model

In this section we compare three supervised learning models. In the *Temporal Model* we only include temporal features as feature vectors, for the *MRIP Model* we include MRIP feature, and the *Homo Model* is the best supervised framework for the homogeneous networks proposed in the work of [1]. The features list of these three models are presented in Table IV. Here we would like to note that these temporal features are all computed under Null Hypothesis Model (Section V-B2). Additionally we include the comparisons of our method with VCP3U and VCP4U described in Section II, the detailed features vectors for VCP3U and VCP4U can be found in the work of [20] accordingly.

TABLE IV
Features List

Features	Temporal Model	MRIP Model	Homo Model
Degree		√	√
Volume		√	√
Common Neighbor		√	√
Jaccard Coefficient		√	√
Adamic Adar		√	√
Preferential Attachment		√	√
Max Flow			√
Shortest Path		√	√
MRIP		√	
PropFlow			√
Recency	√		
Activeness	√		
Temporal Common Neighbor	√		
Temporal Jaccard Coefficient	√		
Temporal Adamic Adar	√		
Temporal Preferential Attachment	√		
Temporal Conference	√		

TABLE V
Three Models Comparison

Classifiers	Temporal Model	MRIP Model	Homo Model	VCP3U	VCP4U
AUROC	**0.787**	0.739	0.697	0.718	0.723

The experimental results of these three models, VCP3U and VCP4U are given in Table V. Interestingly we can see that the *Temporal Model* outperforms all the other models. Furthermore, in our observation with multi-relational features (*MRIP Model*) we can achieve better AUROC than only using homogeneous information (*Homo Model*). We conjecture that better performance can be achieved by combining these useful features, especially temporal features.

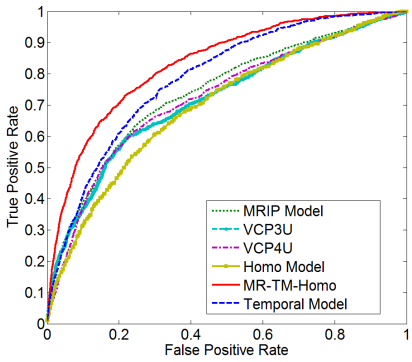

Fig. 13. The ROC curve

B. Temporal Multi-Relational Model

In this section we combine the temporal features with multi-relational features and informative homogeneous features, resulting in a more effective model. For Table VI in this section, MR-TM means that we use features which come from combination of the *MRIP model* and the *Temporal Model*, Homo-TM means the features combination of the *Homo Model* and the *Temporal Model*, and MR-TM-Homo means the combination of all features.

TABLE VI
Combinations of Models

Classifiers	MR-TM	Homo-TM	MR-TM-Homo
Random Forest	0.773	0.773	**0.789**
Logistic	0.825	0.823	**0.832**

In these tables MR-TM model combines features from the *MRIP model* and the *Temporal Model*, Homo-TM model features vector is the union of the *Homo Model* and the *Temporal Model*, and MR-TM-Homo model uses all features listed in Table IV.

Obviously we can conclude *MR-TM-Homo model* outperforms all other solutions, and that we can achieve better AUROC performance with a logistic classifier over a Random Forest classifier.

C. Analysis and Discussion

Because AUROC alone can sometimes be misleading, we also include ROC (receiver operating characteristic curve) curves in Figure 13. From Figure 13 we can see that *MRIP Model* is better than *Homo Model*, *VCP3U* and *VCP4U* in predicting co-authorship, while *Temporal Model* outperforms *VCP3U*, *VCP4U* and *Homo Model*. In our observation with considering temporal information and multi-relational information we can achieve much better performance than most recent competitive work in the link prediction task, i.e. *VCP3U*, *VCP4U* and *Homo Model*. *MR-TM-Homo* outperforms *Homo Model* by almost 20% in terms of AUROC.

A significant challenge of link prediction as a supervised learning problem comes from the sparseness of networks and associated predictors' values. We define a metric called *density ratio* to measure the sparseness of link predictors.

Definition 2. *The density ratio of the predictor p represents how many percentages of node pairs in the link prediction space have corresponding scores assigned by the predictor p.*

The sparseness of the link predictors' values makes it difficult for supervised classifiers to distinguish between positive class and negative class. Multi-relational link predictors and temporal link predictors can achieve better performance than traditional link predictors due to the reason that they overcome the sparseness issue. From Table VII we can see the mulit-relational link predictors and temporal link predictors have much larger *density ratio* than traditional link predictors, i.e. *Common Neighbor*.

TABLE VII
Link Predictors Density Ratio

Homo Predictor	**Density Ratio**	**Heter Predictor**	**Density Ratio**	**Temporal Predictor**	**Density Ratio**
AA	0.005	MRIP	0.31	T_AA	0.139
CN	0.005	VCP3U	1.0	T_CN	0.06
JC	0.005	VCP4U	1.0	T_JC	0.139
PA	1.000			T_PA	0.140
PF	0.110				

We can observe that the sparseness of link predictors is closely correlated with their prediction performance. However the sparseness is not the only factor which determines the performance of link predictors; for example, *PropFlow* values on the DBLP network are much sparser than *Preferential Attachment*, however *PropFlow* has higher AUROC score than *Preferential Attachment* method. The inherent abilitiy (such as heuristic, feasibility and generality) of link predictors is another crucial factor that determines their performance. Intuitively MRIP and temporal features estimate the likelihood of new links based on more abundant information, thus they undoubtedly provide more information for classifier to learn from.

D. Performance Trend

When we scale the test set size, we get the different AUROC scores, and we find that the performance trend agrees with the global trend that we analyzed in previous Section V-A.

TABLE VIII
Model Performance Comparison: Label = 5 and Test = 1, 2, 3, 4, 5

DBLP	T=1	T=2	T=3	T=4	T=5
MR-TM-RF	0.779	0.773	0.766	0.763	0.773
Hom-Time-RF	0.780	0.787	0.776	0.773	0.773
MR-TM-Hom-RF	0.813	0.799	0.792	0.789	0.789
MR-Time-Log	0.839	0.837	0.829	0.824	0.825
Hom-Time-Log	0.832	0.835	0.827	0.823	0.823
MR-TM-Hom-Log	**0.849**	**0.842**	**0.836**	**0.833**	**0.832**

An interesting observation is that the global new links construction time series between 2006 and 2010 are highly correlated with the AUROC scores in Table VIII, with the same trend over time. This implies that global link construction trend can influence the performance of link prediction as we discussed in Section V-A. This also inspires us to include the global trend of link formation in our future work, as the performance of a link predictor aware of link occurrence time

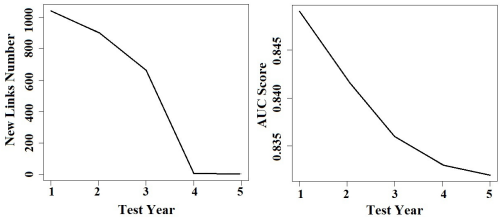

Fig. 14. New Links Time Series Trend and Prediction Performance Trend

should not be impacted significantly by the scale of the test set.

VII. Conclusion

We proposed two types of predictors, which can be used both in unsupervised and supervised models for link prediction in heterogeneous networks. The unsupervised predictor MRIP is demonstrated to have better performance than various competing methods on heterogeneous networks. We also designed several unsupervised temporal link predictors which are extended from classical baseline predictors. They outperform original methods by more than 9% in terms of AUROC. In supervised works of this paper, we employ MRIP and unsupervised temporal predictors as features to construct an effective model for link prediction in heterogeneous networks. The model utilizing temporal features and multi-relational features demonstrates promising performance on DBLP co-authorship prediction.

We also discussed the impact of global link formation trends on the link predictors performance, which suggests that including trend-aware features may benefit the link prediction. Additionally, we show that temporal features can improve the link prediction performance, motivating the next step of our work to efficiently capture the richness of temporal information in the networks. Developing temporal features in the future may bring us a more effective model, which can be used in the heterogeneous networks scenario.

Acknowledgment

Research was sponsored in part by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. We acknowledge financial support from grant #FA9550-12-1-0405 from the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA).

References

[1] R. Lichtenwalter, J. Lussier, and N. Chawla, New perspectives and methods in link prediction, in Proceedings of the 16th ACM SIGKDD, 2010, pp. 243-252.

[2] D. Davis, R. Lichtenwalter, and N. V. Chawla, Multi-Relational Link Prediction in Heterogeneous Information Networks, ASONAM, Kaohsiung, Taiwan, 2011.

[3] Y. Sun, R. Barbery, M. Gupta, C. C. Aggarwal, J. Han, Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks, ASONAM, Kaohsiung, Taiwan, 2011.

[4] D. Kempe, J. M. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 137146, 2003.

[5] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, second edition, 2005.

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

[7] A.-L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaboration, Physica A, 311(3-4):590-614, 2002.

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

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

[10] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology, 58(7):10191031, 2007.

[11] L. Breiman. Random forests. Machine Learning, 45(1):532, 2001.

[12] le Cessie, S. and van Houwelingen, J.C. (1997). Ridge Estimators in Logistic Regression. Applied Statistics, Vol. 41, No. 1, pp. 191-201.

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

[14] M. Mitzenmacher, A brief history of lognormal and power law distributions, In Proceedings of the Allerton Conference on Communication, Control, and Computing, 2001.

[15] H. Deng, J. Han, B. Zhao, Y. Yu and C. Xide Lin. Probabilistic topic models with biased propagation on heterogeneous information networks, KDD 2011, pp.1271-1279.

[16] Dickey, D. A. and Fuller, W.A. Distribution of the estimated for autoregressive time series with a unit root, Journal of the American Statistical Association 74:427-431, 1979.

[17] Potgieter, A.; April, K.; Cooke, R.; and Osunmakinde, I. Temporality in link prediction: Understanding social complexity, Sprouts: Working Papers on Info. Sys. 2007.

[18] Huang, Z., and Lin, D., The time-series link prediction problem with applications in communication surveillance, INFORMS Journal on Computing, 2009.

[19] 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.

[20] R. N. Lichtenwalter and N. V. Chawla. Vertex collocation profiles: subgraph counting for link analysis and prediction in Proceedings of the 21st international conference on World Wide Web, Lyon, France, 2012.

[21] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2011.

[22] R. Singh, J. Xu, and B. Berger. (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection, Proc. Natl. Acad. Sci. USA.

[23] C. Liao, K. Lu, M. Baym, R. Singh, and B. Berger. (2009) IsoRankN: Spectral methods for global alignment of multiple protein networks, Bioinformatics.

[24] L. Breiman. Random forests. Machine Learning, 2001.

[25] G. Rossetti, M. Berlingerio, and F. Giannotti. Scalable link prediction on multidimensional networks via structural analysis. In ICDM Workshop, 2011.

[26] L. Tang and H. Liu. Relational learning via latent social dimensions. In KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, 2009

[27] L. Tang and H. Liu. Uncovering cross-dimension group structures in multi-dimensional networks. In SDM workshop on Analysis of Dynamic Networks, 2009.

[28] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, D. Pedreschi: Foundations of Multidimensional Network Analysis. ASONAM 2011: 485-489