CoupledLP: Link Prediction in Coupled Networks

Yuxiao Dong[†], Jing Zhang[‡], Jie Tang[‡], Nitesh V. Chawla[†], Bai Wang[♮]

[†]Department of Computer Science and Engineering, University of Notre Dame
[†]Interdisciplinary Center for Network Science and Applications (iCeNSA), University of Notre Dame
[‡]Department of Computer Science and Technology, Tsinghua University
[♮]Department of Computer Science and Technology, Beijing University of Posts and Telecommunications

{ydong1, nchawla}@nd.edu, {zhangjing12, jietang}@tsinghua.edu.cn, wangbai@bupt.edu.cn

ABSTRACT

We study the problem of link prediction in *coupled networks*, where we have the structure information of one (source) network and the interactions between this network and another (target) network. The goal is to predict the missing links in the target network. The problem is extremely challenging as we do not have any information of the target network. Moreover, the source and target networks are usually heterogeneous and have different types of nodes and links. How to utilize the structure information in the source network for predicting links in the target network? How to leverage the heterogeneous interactions between the two networks for the prediction task?

We propose a unified framework, CoupledLP, to solve the problem. Given two coupled networks, we first leverage atomic propagation rules to automatically construct implicit links in the target network for addressing the challenge of target network incompleteness, and then propose a Coupled Factor Graph Model to incorporate the meta-paths extracted from the coupled part of the two networks for transferring heterogeneous knowledge. We evaluate the proposed framework on two different genres of datasets: disease-gene (DG) and mobile social networks. In the DG networks, we aim to use the disease network to predict the associations between genes. In the mobile networks, we aim to use the mobile communication network of one mobile operator to infer the network structure of its competitors. On both datasets, the proposed CoupledLP framework outperforms several alternative methods. The proposed problem of coupled link prediction and the corresponding framework demonstrate both the scientific and business applications in biology and social networks.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications; J.4 [**Social and Behavioral Sciences**]: Sociology

Keywords

Link Prediction; Coupled Networks; Social Networks; Healthcare; Mobile Communication Networks

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
KDD'15, August 10-13, 2015, Sydney, NSW, Australia.
© 2015 ACM. ISBN 978-1-4503-3664-2/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2783258.2783329.

1. INTRODUCTION

Link prediction is a fundamental problem in social networks, attracting considerable interest from different research fields, e.g., computer science [27, 38, 4], network science [7, 42], and biology [30, 5, 36, 8, 14]. Typically, the link prediction problem is formalized as: given a snapshot of a network at time t, predict which links will be created in the following time t+1. The problem can be addressed by using unsupervised methods such as Adamic/Adar [1] and random walk with restart [37], or supervised learning models such as supervised random walk [3] and random forest [28] by defining a set of features.

In this paper, we study the link prediction problem in an interesting new setting: *coupled networks*, where we have two networks: one source network G^S and one target network G^T. Basically, we have structure information of the source network G^S and interactions G^C between the two networks, but do not have any structure information for the target network. The objective of link prediction here is to predict the existence of links in the target network G^T.

The problem exists in many data mining applications. As the example illustrated in Figure 1, the disease-gene coupled networks are decomposed as a disease network (Fig. 1(b)), a gene network (Fig. 1(d)), and a cross network that links source and target networks together (Fig. 1(c)). Link prediction in coupled networks is then formalized as a problem of using the disease network and associations between diseases and genes to predict the relationships that exist between two genes (Fig. 1(b) + Fig. 1(c) $\rightarrow$ Fig. 1(d)). Solving the problem automatically is quite useful, because otherwise arduous and expensive medical experiments on a huge selection by biologists and geneticists are required to figure out the links in the gene network [28]. In other domains such as social networks, the problem is also very important. In mobile social networks, an operator such as AT&T is motivated to infer the link structure among users of its competitors (such as Verizon and T-Mobile); Or in online social networks, it would be very useful for Google+ to acquire new users by having Facebook connections among GMail users who are registered Facebook users.

Coupled network link prediction is different from the *classical link prediction* problem [27, 28, 42, 23], which generally aims at predicting the future links in the next time period. Meanwhile, the proposed problem differs from *link prediction in heterogeneous network* [43, 45, 38, 22, 2], in which partial multi-typed links are given to predict the remaining single- or multi-typed links. Our problem is also different from the problem of *transfer link prediction* [6, 11, 39], which focuses on leveraging the estimated parameters in one network to improve the prediction performance of the other network based on the common features between the two networks. Finally, our problem is different from the problem of *cross-domain link prediction* [40, 20], whereas it aims to predict the links

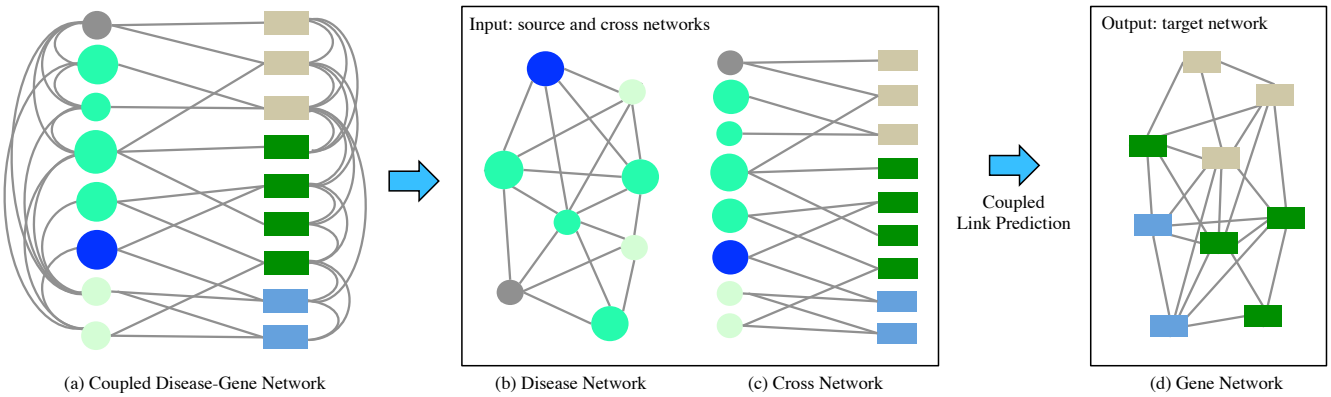


Figure 1: Illustrative example of link prediction in coupled disease-gene networks. (a) Coupled disease-gene network; (b) Disease network; (c) Cross network; (d) Gene network. Taking disease network as the source and gene network as the target, the problem of coupled link prediction aims to predict the links in gene network (d) by leveraging both the disease network (b) and cross network (c).

Table 1: Differences of link prediction problems.We take the disease and gene networks in Figure 1 as an example.

Problem	Transfer link prediction [6, 11]	Cross-domain link prediction [40, 20]	Heterogeneous link prediction [45, 38]	Coupled link prediction
Input	disease network + (part of gene network)	disease network + gene network + (part of cross network)	part of coupled networks	disease network + cross network + (part of gene network)
Output	remaining links in gene network	remaining links in cross network	remaining links in coupled networks	links in gene network

in the cross network (Fig. 1(c)) between two networks. We summarize the differences of different link prediction problems in Table 1. The significant advantage of the proposed problem lies in that it can be applied to real applications such as inferring the links in a competitor's or enemy's network to better understand it.

This coupled link prediction problem presents several unique challenges. First, *incompleteness*, we do not have structure information between two users in the target network—that is, there is a visibility of links that go from the source network to the target network but not beyond that. Second, *heterogeneity*, the source and target networks with multi-typed objects are twisted and coupled with one another. This makes it difficult to directly use a supervised learning approach due to the different types of links in source and target networks. Third, *asymmetry*, following the heterogeneity, the two coupled networks usually present different network properties—such as the average degree k or clustering coefficient cc as shown in Table 2.

In light of these differences and challenges, we present a unified two-phase framework CoupledLP to predict links in coupled networks. At the first phase, we leverage atomic propagation rules to propagate the implicit knowledge from the source network to the target network and construct "complete" coupled networks. At the second phase, we first extract features from the "complete" coupled networks, and then generate informative meta-paths from the coupled part between the source and target networks. We then propose a supervised Coupled Factor Graph Model to incorporate the meta-paths as structural correlation factors. Our contributions can be summarized as follows:

- We formally define a novel problem of coupled link prediction in networks and propose a unified CoupledLP framework to solve it.
- We propose a Coupled Factor Graph Model to utilize the implicit knowledge for predicting links in the target network. To incorporate the heterogeneous information between two networks, we define the meta-paths extracted from the coupled networks as structural correlation features.
- Finally, our experimental results on two different types of coupled networks demonstrate the effectiveness of the proposed CoupledLP framework. CoupledLP significantly outperforms several state-of-the-art link prediction algorithms on different networks.

The datasets used in the paper are three sets of large-scale real-world coupled networks, in which the first are the networks with diseases and genes coupled together as shown in Figure 1(a), the second are the mobile communication networks from three operators coupled together with 712 million call records in a European country, and the last are also the mobile networks from two operators with 42 million calling records in an Asian city. The experimental results on the large-scale real networks demonstrate that 1) CoupledLP offers a greater than 84% potential predictability for determining the existence of phenotypic links between disease pairs and 2) a mobile operator—such as AT&T—can achieve an accuracy of 80% for predicting the top links of its competitor's network—such as Verizon.

Organization. Section 2 formalizes the problem of link prediction in coupled networks; Section 3 presents the proposed framework to solve the problem; Section 4 explains the experimental results; Section 5 discusses related work and Section 6 concludes the work.

2. PROBLEM DEFINITION

Generally, we use $G = \{V, E\}$ to denote a network, where $V = \{v_i\}$ is the set of nodes, and $E \subseteq V \times V$ is the set of links between nodes, with each link denoted as $e_{ij} = (v_i, v_j) \in E$.

Definition 1. **Coupled networks and cross network:** Given a source network $G^S = (V^S, E^S)$ and a target network $G^T = (V^T, E^T)$, they compose coupled networks if there exists a cross link e_{ij} with one node $v_i \in V^S$ and the other node $v_j \in V^T$. The

cross network $G^C = (V^C, E^C)$ is a bipartite network containing all the cross links in the coupled networks.

Figure 1(a) shows a typical example of coupled networks with a disease network as the source network G^S and a gene network as the target network G^T. The links between diseases and genes represent the genetic association links between them, which, with their linked nodes in G^S and G^T, constitute the cross network G^C.

Problem 1. **Coupled network link prediction:** Given the source network G^S and the cross network G^C in coupled networks $G = (G^S, G^T, G^C)$, the task is to find a predictive function:

$$f : (G^S, G^C) \rightarrow Y^T$$

where Y^T is the set of labels for the potential links in the target network G^T, with $y_{ij} = 1$ indicating a link exists between v_i and v_j, and $y_{ij} = 0$ indicating no link exists between them. Henceforth we use y_e to denote the label of a link e.

For the nodes in the target network without any links associated in the cross network, it is intractable to predict the links among them. In this sense, the objective actually can be reduced to predicting the links in the target network among the nodes contained in both the target network and the cross network, i.e., $\{v \in V^T \cap V^C\}$. However, in this work we still abbreviate the objective as predicting links in the target network G^T.

The coupled network link prediction problem is general for both directed and undirected networks. In this work, both kinds of networks are investigated, including directed mobile networks and undirected disease-gene networks. The details of the datasets are introduced in Section 4.1.

The fundamental challenge of this problem is how to capture the link formation patterns in the target network with little prior knowledge about its network structure and a few heterogeneous information about the cross network. In the traditional link prediction problem, we usually can observe most of the network structure, thus the patterns used to predict links can be easily obtained. However, in our problem, the link structure between two users in the target network is totally opaque, which makes the problem non-trivial. Therefore, how to effectively capture the features of the links in the target network without any inside information becomes a significant challenge.

3. COUPLED NETWORK LINK PREDICTION FRAMEWORK

In this section, we first introduce the framework to solve the proposed link prediction problem in coupled networks, and then explain the two main phases in the framework respectively.

3.1 CoupledLP Framework

To solve the challenges of coupled link prediction, we propose a framework, CoupledLP, to first enrich the target network structure, from which we then extract features and predict the links in the target network.

At the first phase, we infer the possible links [24] with the highest potential in the target network based on several heuristic rules. We name the inferred network as an implicit target network. The motivation is with the highly probable links inferred at first, we can extract more informative structural features for the links to be predicted. Otherwise, the target network is difficult to be leveraged for extracting features.

At the second phase, based on the implicit target network, we propose a Coupled Factor Graph Model (CoupledFG) to predict the links in the target network. The idea is to use both features extracted from the implicit target network and the structural meta-paths extracted from coupled networks.

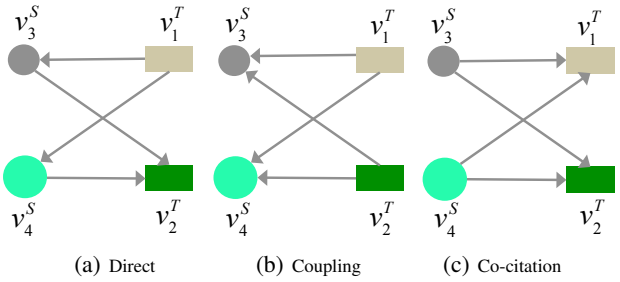


Figure 2: Atomic propagations in coupled networks.

3.2 Implicit Target Network Construction

To enrich the structural information in the target network, we construct an implicit target network based on three atomic propagation rules proposed in [15, 25], which include direct propagation, coupling propagation, and co-citation propagation. The three rules explain several real life phenomena such as information propagation in social networks, and traffic flow in railway or flight networks. Figure 2 illustrates how to infer the implicit links in the target network by using the three atomic propagation rules. In the figure, the links between v_1^T and v_2^T are to be predicted in the target network, the links denoted by the solid line are observable in the cross network, the objective is to infer the unknown links with high probability based on the observed links.

Direct Propagation. Figure 2(a) illustrates direct propagation. Suppose we observe many indirect links pointing from v_1^T to v_2^T, such as $v_1^T \rightarrow v_3^S \rightarrow v_2^T$ and $v_1^T \rightarrow v_4^S \rightarrow v_2^T$ in the cross network. This indicates the potential direct propagation between the two nodes v_1^T and v_2^T in the target network. For example, in flight networks, when passengers always fly from Chicago to Boston through transferring at New York, or fly from Chicago to Boston through transferring at Detroit and so on, the airlines will probably feel the requirements and add the direct flights from Chicago to Boston.

Coupling Propagation. Figure 2(b) illustrates coupling propagation. Suppose we observe v_1^T and v_2^T always link to one common node, e.g., v_3^S or v_4^S, it then indicates that the two nodes v_1^T and v_2^T also link to each other. A real case about Facebook wall posts may well explain the phenomenon. Generally, it is highly probable that the users who post comments on the same message know each other. Moreover, the more comments there are posted on a common message, the more likely they are friends.

Co-citation Propagation. Figure 2(c) illustrates co-citation propagation. Suppose we observe v_1^T and v_2^T are always linked by one common node, e.g., v_3^S or v_4^S. This indicates that the two nodes v_1^T and v_2^T also link to each other. An intuitive case is the paper citation network. If two papers are always cited together by the other papers, it will be more likely that they share similar research topics, and that one paper may also cite another one.

Formally, we represent the network structure of the cross network as a matrix M, with each element M_{ij} as the transition probability between node v_i and v_j. In a static network, M_{ij} can be calculated as the normalized link strength between node v_i and v_j. In a dynamic network, the duration between two sequential propaga-

tions can be incorporated into the transition probability [29]. Then the network structure in the target network can be represented as the matrix multiplication, MM, according to the direct propagation rule, MM^T, according to the coupling propagation rule, and M^TM, according to the co-citation rule. We combine the three results together to represent the final network structure of the target network. Specifically, the network structure M^* of the target network is shown as follows: $M^* = MM + MM^T + M^TM$. There are several ways to construct the implicit target network $G^{T'}$ by using M^*. For example, we can map the resulted matrix element M^*_{ij} into a probability space and construct a probabilistic target network [24]. In this work, we select the top $z\%$ links ranked by M^*_{ij} as the inferred links in the implicit target network, where z is threshold parameter to prune the noisy links inferred by the three propagation rules.

Although the atomic propagation rules are explained by the directed networks, when applied in the undirected networks, the three rules can be simply reduced into the same rule denoted by MM because $M = M^T$. There are other methods for inferring the implicit network structure, such as random walk with restart or directly counting common neighbors. However, the proposed propagation method is empirically proved to perform better than several alternative methods in Section 4.

3.3 Coupled Factor Graph Model

This section proposes a Coupled Factor Graph Model (CoupledFG) to predict the label set Y^T for the links in the target network G^T based on the network structure in the source network G^S, cross network G^C, and the implicit target network $G^{T'}$ inferred in the first phase. Our goal is to train a supervised classification model.

Model description. We treat the links in G^T with two nodes contained in $V^T \cap V^C$ as the candidates to predict. Specifically, we first extract features $\mathbf{x}_e$ from $(G^S, G^C, G^{T'})$ for each link e in the candidate set and learn the weights for different features by using our proposed Coupled Factor Graph Model. We then estimate a formation probability $P(y_e|\mathbf{x}_e)$ for each link in the test set.

The objective of our model is to maximize the formation probability of the links in coupled networks given the observed features and the model parameters, i.e., $P(Y|\mathbf{X}, G)$. In factor graph, the "global" probability can be factorized as a product of "local" factor functions [21], which is relatively easy to optimize.

The attribute feature vector $\mathbf{X}_e$ for each link can be defined as different types, such as the number of common friends, the Jaccard distance, the score of random walk with restart between the two nodes of the link. However, due to the asymmetry of the two coupled networks, the features extracted for edges in the source network G^S and edges in the (implicit) target network G^T ($G^{T'}$) may be located in different spaces. We have to use two separate sets of factors to capture the attributes of edges in the source network G^S and the (implicit) target network G^T, respectively, and factorize the conditional probability $P(\mathbf{X}|Y)$ of generating attributes $\mathbf{X}$ for links given their labels Y as:

$$P(\mathbf{X}|Y) \propto \prod_{e\in E^S}\prod_{k=1}^{K} P(x^S_{ek}|y^S_e) \prod_{e\in E^T}\prod_{k=1}^{K} P(x^T_{ek}|y^T_e) \quad (1)$$

where $P(x^S_{ek}|y^S_e)$ is the probability of generating the k^{th} feature x^S_{ek} in the source network given the label y^S_e. Accordingly, $P(x^T_{ek}|y^T_e)$ is the probability of generating the k^{th} feature x^T_{ek} in the target network given the label y^T_e, and K is the number of attribute features.

In addition to the features defined on source and target networks separately, we define the structural factor $P(Y|G)$ in coupled networks to bridge the source and target networks. The meta-path based methods have been demonstrated as effective solutions for solving the heterogeneity of link prediction tasks [46, 38]. The major issue here is how to design informative meta-paths in coupled networks and utilize them in the factor graph model.

We design meta-paths in coupled networks as follows. Following the definitions in the work [38], we define the *schema* of coupled networks to be $S = (Q, R)$, where $Q = \{S_Q, T_Q\}$ with S_Q as the set of node types in the source network and T_Q as the set of node types in the target network, $R = \{S_R, T_R, C_R\}$ with S_R as the set of relation types in the source network, T_R as the set of relation types in the target network, and C_R as the set of relation types in the cross network. For example, in coupled disease-gene networks, we have $Q = \{$*Disease, Gene*$\}$ and $R = \{$*D-D relation, G-G relation, D-G relation*$\}$. We define the following meta-paths, based on their physical mechanisms, with illustrative examples in disease-gene networks.

$S_QT_QS_Q$: Disease $\xrightarrow{express^{-1}}$ Gene $\xrightarrow{express}$ Disease, which means two diseases are expressed by one gene.

$T_QS_QT_Q$: Gene $\xrightarrow{express}$ Disease $\xrightarrow{express^{-1}}$ Gene, which means one disease is expressed by two genes.

$S_Q[T_QT_Q]^rS_Q$: Disease $\xrightarrow{express^{-1}}$ [Gene $\xrightarrow{associate}$ Gene]^r $\xrightarrow{express}$ Disease, which means two diseases are expressed by r associated genes.

$T_Q[S_QS_Q]^rT_Q$: Gene $\xrightarrow{express}$ [Disease $\xrightarrow{family}$ Disease]^r $\xrightarrow{express^{-1}}$ Gene, which means two genes express r diseases belonging to one family.

We now introduce how to model the coupled meta-paths in factor graphs. Again, according to the theory of factor graph [21], the probability $P(Y|G)$ of labels given the structure of the network can be factorized over all meta-paths in networks as following:

$$P(Y|G) \propto \prod_{\pi\in\Pi} P(Y_\pi) \quad (2)$$

where Π denotes the pre-defined meta-path set in coupled networks and π is a group of meta-paths in Π. Y_π denotes the labels of candidate links instantiated by the group of meta-paths π. For example, two candidate links e_1 and e_2 that share the same meta-paths can be defined as $Y_\pi = (y_{e_1}, y_{e_2})$. The modeling of meta-paths as structured heterogeneous correlations makes our method different from previous work, in which meta-paths are usually used as a vector of attributes for machine learning models.

Given the probability $P(\mathbf{X}|Y)$ of generating the attribute features $\mathbf{X}$ and $P(Y|G)$ of generating meta-path based structural features, the conditional distribution over the coupled networks is factorized as:

$$\begin{aligned} P(Y|\mathbf{X}, G) &\propto P(\mathbf{X}|Y) \cdot P(Y|G) \\ &\propto \prod_{e\in E^S}\prod_{k=1}^{K} P(x^S_{ek}|y^S_e) \prod_{e\in E^T}\prod_{k=1}^{K} P(x^T_{ek}|y^T_e) \prod_{\pi\in\Pi} P(Y_\pi) \end{aligned} \quad (3)$$

Now the problem becomes how to instantiate the probabilities $P(x_{ek}|y_e)$ and $P(Y_\pi)$ in Eq. (3). In principle, they can be instantiated in different ways. In this work, we model them by Markov random fields based on Hammersley-Clifford theorem [17], which states that a probability distribution that has a positive density satisfies one of the Markov properties with respect to an undirected graph if and only if it is a Gibbs random field. And the density of the probability distribution can be factorized over the cliques of the

graph. Thus the three probabilities in Eq. (3) can be initialized as:

$$P(x_{ek}^S|y_e^S) = \frac{1}{Z_\alpha}\exp\{\alpha_k f_k(x_{ek}^S, y_e^S)\} \quad (4)$$

$$P(x_{ek}^T|y_e^T) = \frac{1}{Z_\beta}\exp\{\beta_k g_k(x_{ek}^T, y_e^T)\} \quad (5)$$

$$P(Y_\pi) = \frac{1}{Z_\gamma}\exp\{\gamma_\pi h_\pi(Y_\pi)\} \quad (6)$$

where α_k, β_k, and γ_π are the corresponding weights of the factor functions, respectively representing the influence degree of the k^{th} factor function $f(\cdot)$ and $g(\cdot)$, and π^{th} factor function $h(\cdot)$. $f(\cdot)$ and $g(\cdot)$ are defined as a vector of feature functions in source and target networks, respectively. Similarly, $h(\cdot)$ is defined as a vector of indicator functions. Z_α, Z_β, and Z_γ are the normalization factors.

Parameter estimation. The parameters to be estimated are $\theta = \{\{\alpha\}, \{\beta\}, \{\gamma\}\}$. We learn the parameters through maximizing the logarithm of the likelihood function $P(Y|\mathbf{X}, G, \theta)$. Based on Eqs. (3), (4), (5), and (6), the log-likelihood objective function can be written as:

$$\mathcal{O}(\theta) = \sum_{e\in E^S}\big(\sum_{k=1}^{K}\alpha_k f_k(x_{ek}^S, y_e^S)\big) + \sum_{e\in E^T}\big(\sum_{k=1}^{K}\beta_k g_k(x_{ek}^T, y_e^T)\big) + \sum_{\pi\in\Pi}\gamma_\pi h_\pi(Y_\pi) - \log Z \quad (7)$$

where $Z = Z_\alpha Z_\beta Z_\gamma$ is a normalization factor. In this objective function, the first term and second term respectively define the likelihood over the source network and the target network separately, and the third term defines the likelihood over the meta-paths through the cross network that connects the source and target networks. Such a definition implies that while the source and target networks are optimized with different parameters $\{\alpha\}$ and $\{\beta\}$, the coupled networks are bridged by the parameters $\{\gamma\}$ of the meta-path factors. The idea here is inspired by the model proposed in [39], wherein Tang et al. presented a transfer-based method for inferring social tie across heterogeneous networks. The difference lies in that [39] uses common factors—defined based on social theories—extracted from source and target networks separately to bridge them, while in this work we use factors defined on meta-paths that naturally connect two coupled networks.

To solve the log-likelihood function, we adopt a gradient descent algorithm (or Newton-Raphson algorithm). Specifically, we derive the gradients of each parameter with regards to Eq. (7). For example, the gradients for each α_k and γ_π are derived as:

$$\frac{\mathcal{O}(\theta)}{\alpha_k} = \mathbb{E}[f_k(x_{ek}^S, y_e^S)] - \mathbb{E}_{P_{\alpha_k}(y_e^S|\mathbf{X})}[f_k(x_{ek}^S, y_e^S)] \quad (8)$$

$$\frac{\mathcal{O}(\theta)}{\gamma_\pi} = \mathbb{E}[h_\pi(Y_\pi)] - \mathbb{E}_{P_{\gamma_\pi}(Y_\pi|\mathbf{X},G)}[h_\pi(Y_\pi)] \quad (9)$$

where $\mathbb{E}[t_\theta(\cdot)]$ ($t_\theta(\cdot)$ represents $f_k(x_{ek}^S, y_e^S)$ or $h_\pi(Y_\pi)$) is the expectation of factor function $t_\theta(\cdot)$ given the data distribution of the input network, and $\mathbb{E}_{P_\theta(Y|\mathbf{X},G)}[t_\theta(\cdot)]$ represents the expectation of factor function under the distribution $P_\theta(Y|\mathbf{X}, G)$ learned by the model.

Usually, it is intractable to estimate the marginal probability in the second term of Eq. (8) and (9) as the graphical structure can be arbitrary and may contain cycles. In this work, we use loopy belief propagation (LBP) [31] to approximate the gradients. The learning algorithm contains two main parts: First, perform LBP algorithm to calculate corresponding marginal distributions; Second, update each parameter to maximize the objective function. It is worth noting that the learning process perform the LBP algorithm twice in each iteration, one is for estimating the marginal distribution of unknown variables and the other for marginal distribution over all cliques. In this way, the algorithm utilizes the unlabeled information in the learning process. Finally, each parameter is updated with the learning rate η:

$$\theta_{new} = \theta_{old} + \eta \cdot \frac{\mathcal{O}(\theta)}{\theta} \quad (10)$$

The time complexity of the learning algorithm at each iteration is $O(|E_{cand}| \cdot |y| + |E_{pair}| \cdot |y|^2)$ if considering the candidate link pairs connected by meta-paths as structural factors, where $|E_{cand}|$ is the number of candidate links, $|y|$ is the number of labels, and $|E_{pair}|$ is the number of pairs of candidate links that are connected by meta-paths.

Link prediction. After we obtain the learned parameters $\theta = \{\{\alpha\}, \{\beta\}, \{\gamma\}\}$, we estimate the link labels Y^T in the test set. All the links in the test set are assigned with labels that can maximize the marginal probabilities with the estimated parameters:

$$Y^{T\star} = \arg\max \mathcal{O}(Y^L|\mathbf{X}, G, \theta).$$

where the LBP algorithm is used to solve this problem.

Due to the computing complexity, it is intractable to enumerate all meta-paths in given coupled networks. There are several ways to determine the choice of meta-paths. We use the physical meaning in real world as a principle to choose meta-paths [38], and limit the length of meta-paths into three and four. Taking disease-gene networks as an example, we choose four types of meta-paths, including the following cases: 1) Disease−Gene−Disease, which means two diseases are expressed by one gene; 2) Disease−Gene−Gene−Disease, which means two diseases are expressed by two associated genes; 3) Gene−Disease−Gene, which means one disease is expressed by two genes; 4) Gene−Disease−Disease−Gene, which means two genes express two diseases who belong to one family. The first two types of meta-paths are used for predicting from gene network to disease network, and reversely, the last two are used for predicting links in gene network from disease network.

3.4 Summary

In conclusion, targeting at the three challenges in coupled network link prediction problem, we propose a unified framework CoupledLP that contains two phases. At the first phase, we construct an implicit target network to solve the incompleteness of the target network. At the second phase, we model attribute features by using separate sets of parameters to address the issue of the asymmetry of the source and target networks. We then incorporate informative meta-paths as the structural factors into factor graphs to untangle the heterogeneity of coupled networks.

4. EXPERIMENTS

In this section, we evaluate our proposed framework, CoupledLP, for predicting links in coupled networks on three networks and demonstrate its effectiveness. The disease-gene dataset and code are publicly available[1].

[1] http://aminer.org/CoupledLP

Table 2: Statistics of the three sets of coupled networks. k: average degree; cc: clustering coefficient; ac: associative coefficient. *The asymmetry of network properties between source and target networks is revealed.*

	D	G	$D \leftrightarrow G$	A_a	A_b	$A_a \leftrightarrow A_b$	E_a	E_b	E_c	$E_a \leftrightarrow E_b$	$E_a \leftrightarrow E_c$	$E_b \leftrightarrow E_c$
#Nodes	703	1132	1835	348,640	63,687	235,715	2,531,187	655,755	354,166	1,912,933	1,255,046	625,379
#Links	74523	2450	10483	613,614	96,325	306,213	3,355,197	649,322	311,432	1,844,342	1,131,593	507,894
k	212.01	4.33	11.43	3.52	3.02	2.59	2.65	1.98	1.75	1.92	1.80	1.62
cc	0.2639	0.0377	0	0.0237	0.0225	0	0.0457	0.0366	0.0317	0	0	0
ac	-0.0256	0.1761	-0.2556	0.2011	0.1671	0.0654	0.2848	0.2693	0.2806	0.0231	-0.0305	0.1113

Table 3: Statistics of the candidate and positive links in constructed coupled networks.

Statistics	D to G	G to D	A_a to A_b	A_b to A_a	E_a to E_b	E_b to E_a	E_a to E_c	E_c to E_a	E_b to E_c	E_c to E_b
#Candidate links	243,393	19,014	376,416	1,280,959	972,808	2,594,169	424,793	1,655,878	252,471	372,421
#Positive links	1,582	11,015	25,694	57,138	179,265	373,511	83,657	232,814	46,954	63,544
%Positive links	0.65%	57.93%	6.83%	4.46%	18.43%	14.40%	19.69%	14.06%	18.60%	17.06%

4.1 Datasets

We use two types of coupled networks to evaluate our proposed framework, including Disease-Gene networks, Asian mobile networks with two operators, and European mobile networks with three operators. Table 2 summaries the statistics of the constructed coupled networks from the three datasets. Clearly, we can see the network properties are asymmetric between source and target networks.

Disease-Gene networks (DG). The dataset contains a disease network (D), a gene network (G) and the connections between diseases and genes [8]. Disease pairs are connected by a phenotypic link if there exist significant co-morbidities in real patients (Figure 1(b)). Gene pairs are connected by protein-protein interaction links in accordance with combined physical interaction data collected from Human Protein Reference Database (HPRD) and the Online Predicted Human Interaction Database [8] (Figure 1(d)). Genetic association links exist between diseases and genes in a bipartite graph and represent known disease-gene associations extracted from the Online Mendelian Inheritance in Man database, SwissProt, and HPRD [8] (Figure 1(c)).

Asian mobile networks (A). This is a dataset used in [13], containing about 42 million call records in an Asian city from October 2005 to March 2006. We construct directed networks from the call records by treating each user as a node and creating a link between two users if there exists at least one call record between them. A weight is assigned to a link to represent the number of call records. To conduct our task of link prediction in coupled networks, we construct different coupled networks by treating one operator network as source network, another operator network as target network, and the connections between the two networks as the cross network. Two operators are involved in this dataset, which are denoted as A_a and A_b with the size $|A_a| > |A_b|$. We construct two different coupled networks by using A_a and A_b.

European mobile networks (E). This is a dataset used in [28, 12], containing more than 712 million call records in a European country within two months, i.e., August and September, in 2011. There are three major operators in the mobile networks. We denote the communication network of each of the three operators as E_a, E_b and E_c, respectively with the size $|E_a| > |E_b| > |E_c|$. The European mobile networks are constructed in the same way as the Asian mobile networks. In total, we construct six coupled networks by using this dataset.

4.2 Feature Definition

Basically, for network analysis or graph mining, a series of topological features are usually defined to solve the problem. In this paper, we define the similar features as those defined in HPLP+ [28]. We take an weighted undirected network as example to explain features defined for one potential link e_{ij} between two nodes v_i and v_j. The features include common-neighbor based and path based features.

Common-neighbor based features include the number of common neighbors (CN), Adamic/Adar (AA), and Jaccard Coefficient (JC), and Preferential Attachment (PA). CN simply counts the number of common neighbors between two nodes v_i and v_j. AA also counts the number of common neighbors, but weights each common neighbor by rarity, which is the reciprocal of a node's degree [1]. JC calculates the ratio of common neighbors among the union set of v_i's and v_j's neighbors. PA calculates the similarity between v_i and v_j by the product of their degrees.

Among several alternative path based features, we choose PropFlow (PF) [28] rather than Katz and Random Walk with Restart (RWR), due to their high computing complexity. PF calculates the similarity by conducting random walks from v_i to v_j with a restricted number of steps. We set the number of maximal steps as 4. However, Katz and RWR need to sum over all possible paths between v_i and v_j.

4.3 Experimental Setup

Candidate generation. We explain how to generate candidate links for the prediction tasks. The total number of the potential links equals to $|V| \times |V|$, which is usually very large. Thus, we follow the general method [42, 3] to treat candidate links as those with two nodes at most 2-hops away from each other in the network. Table 3 summarizes the statistics of the candidate and positive links in the coupled networks. From the table, we can see that the percentage of the positive links varies from 4% to 19% for the mobile networks, however, this ratio is only 0.65% for the coupled D to G networks constructed from the disease and gene networks, which is extremely imbalanced. Due to the incompleteness of target network in coupled network link prediction, only 1% of target candidate links are treated as training set and the remaining as test links.

Comparison methods. We compare CoupledLP with both the unsupervised and supervised methods for the task of link prediction in coupled networks. For the unsupervised methods, we directly use the above defined features, CN, AA, JC, PA, and PF, to rank links due to their competitive predictive power in both our experiments

Table 4: The performance of AUPR on different methods.

Method	D to G	G to D	A_a to A_b	A_b to A_a	E_a to E_b	E_b to E_a	E_a to E_c	E_c to E_a	E_b to E_c	E_c to E_b
CN	0.0155	0.6011	0.3017	0.1348	0.3598	0.2319	0.3817	0.2079	0.3145	0.2654
AA	0.0167	0.5912	0.3344	0.1596	0.4541	0.2800	0.4838	0.2562	0.3802	0.3180
JC	0.0803	0.4812	0.0835	0.0903	0.3848	0.3082	0.4140	0.3429	0.3628	0.3579
PA	0.0083	0.7566	0.0820	0.0599	0.1446	0.1287	0.1525	0.1250	0.1560	0.1471
PF	0.0233	0.5501	0.1455	0.0989	0.3504	0.2248	0.3722	0.2138	0.2833	0.2446
IT	0.0155	0.6011	0.3715	0.2059	0.4344	0.3157	0.4568	0.2940	0.4008	0.3559
LRC-IT	0.0140	0.7830	0.3610	0.1880	0.4580	0.3140	0.5240	0.2870	0.4230	0.3500
LRC	0.0190	0.7930	0.3820	0.2030	0.4920	0.3160	0.5190	0.2910	0.4270	0.3590
DT-IT	0.0070	0.6270	0.2760	0.1050	0.3440	0.1620	0.3810	0.1550	0.2900	0.2260
DT	0.0080	0.6310	0.2530	0.1030	0.3580	0.1640	0.3470	0.1557	0.3060	0.2420
CoupledLP-IT	0.0303	0.8249	0.4291	0.2483	0.5088	0.3484	0.5257	0.3240	0.4537	0.3855
CoupledLP	0.0249	0.8432	0.4305	0.2776	0.5481	0.3591	0.5420	0.3399	0.4692	0.4133

Table 5: The performance of AUROC on different methods.

Method	D to G	G to D	A_a to A_b	A_b to A_a	E_a to E_b	E_b to E_a	E_a to E_c	E_c to E_a	E_b to E_c	E_c to E_b
CN	0.6384	0.5330	0.6754	0.5896	0.6090	0.5556	0.6133	0.5418	0.5736	0.5552
AA	0.6544	0.5289	0.7658	0.6933	0.7408	0.6664	0.7486	0.6357	0.6826	0.6543
JC	0.6507	0.3666	0.5974	0.5220	0.7186	0.6116	0.7280	0.5977	0.6652	0.6327
PA	0.4850	0.7073	0.5802	0.5615	0.3835	0.4460	0.3746	0.4462	0.4131	0.4270
PF	0.6426	0.4890	0.7275	0.7006	0.7339	0.6649	0.7389	0.6554	0.6736	0.5552
IT	0.6384	0.5330	0.7735	0.7273	0.6867	0.6435	0.6969	0.6335	0.6756	0.6618
LRC-IT	0.5450	0.7160	0.7590	0.7280	0.7580	0.6930	0.7750	0.6840	0.7200	0.6890
LRC	0.6230	0.7320	0.8210	0.7750	0.7670	0.7070	0.7730	0.6950	0.7290	0.7030
DT-IT	0.5010	0.5830	0.7190	0.6260	0.6690	0.5480	0.6930	0.5410	0.6340	0.5920
DT	0.5140	0.5930	0.7460	0.6530	0.6750	0.5510	0.6730	0.5440	0.6450	0.6040
CoupledLP-IT	0.6825	0.7586	0.8052	0.7424	0.7597	0.7017	0.7664	0.6885	0.7314	0.7004
CoupledLP	0.6790	0.7865	0.8336	0.7807	0.7779	0.7127	0.7769	0.7016	0.7405	0.7157

and the previous work [27]. The propagation method (IT) proposed in Section 3.2 is also used as an unsupervised method.

For the supervised methods, we choose Decision Tree (**DT**) and Logistic Regression Classification (**LRC**) used in [3]. All the methods consider the above defined features. Both DT and LRC simply consider the coupled networks as a homogeneous network.

CoupledLP is our proposed framework that includes both the implicit target network construction at the first phase (z = 10) and the CoupledFG model at the second phase. **DT-IT, LRC-IT, CoupledLP-IT** are the reduced versions of three supervised methods above that exclude the implicit target network construction at the first phase.

Evaluation metrics. For unsupervised methods, we rank links by the similarity scores calculated by these methods. For supervised methods, we rank test links by the probability $P(y_e|x_e, G; \theta)$ generated by them. We evaluate the ranking results by using the area under the Receiver Operating Characteristic curve (AUROC), the area under the Precision-Recall curve (AUPR), and precision at top $k\%$, where we change k from 1 to 10 with 1 as the interval and from 10 to 100 with 10 as the interval in our experiments. AUROC, AUPR and precision at top ranked links are typically used in the evaluation of link prediction tasks [46, 28, 35, 23, 42, 3, 44].

4.4 Experimental Results

In this section, we first present the prediction performance of our proposed CoupledLP framework compared with the baselines; second, verify the effects of implicit target network construction; finally discuss the implications in biology and social networks.

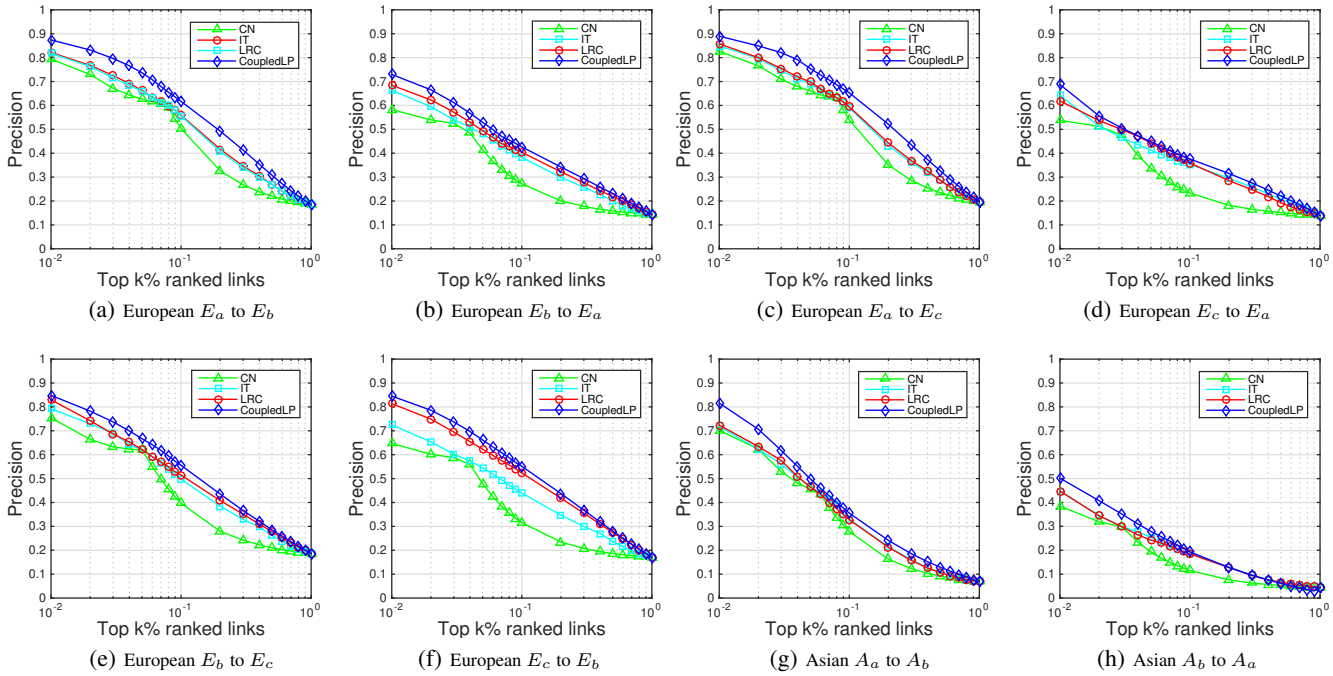
Performance comparison. Tables 4 and 5 show the performance for different methods in terms of AUROC and AUPR on all the constructed 10 coupled networks. Figure 3 presents the curve of the precision at top $k\%$ ranked links. Generally, we can see that our proposed CoupledLP outperforms the other baselines for AUPR, AUROC and precision at top $k\%$ ranked links in most coupled networks. We also conduct t-test for our results, which shows that all the improvements of our proposed CoupledLP over other baselines are statistically significant ($p \ll 0.001$).

IT performs better in most cases than the unsupervised methods in terms of AUPR in Table 4 and precision at top $k\%$ in Figure 3, which demonstrates the effectiveness of our proposed propagation method. That forms the reason that we select it to transfer implicit information from the source to the target network at the first phase in our framework.

The proposed CoupledLP performs better than both the supervised and unsupervised baselines in terms of AUROC. The similar results can be obtained as measured by AUPR in Table 4. We also note that the unsupervised method JC achieves the best AUPR score in D to G case. Figure 3 presents the precision at top $k\%$ ranked links of our method CoupledLP and three baselines CN, IT, and LRC, which obtain the best performance comparing to other baselines. We can see from the figure that the proposed CoupledLP exhibits an improvement of 5%-10% in terms of precision from top 1% to 100% ranked links.

From both tables, we can see that the supervised methods (LRC, DT, and CoupledLP) outperform their reduced versions (LRC-IT, DT-IT, and CoupledLP-IT) that don't use the features extracted from the constructed implicit target network. This demonstrates the effectiveness of the first phase in our CoupledLP framework.

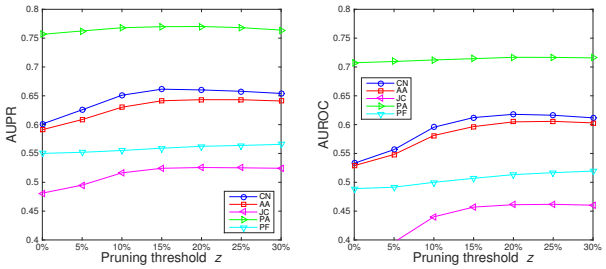
However, we also observe that in the coupled networks D to G, the prediction performance in terms of AUPR is quite limited, while the performance is good on the coupled networks G to D. The reason lies in that in the prediction case from D to G the candidate links



(a) European E_a to E_b (b) European E_b to E_a (c) European E_a to E_c (d) European E_c to E_a

(e) European E_b to E_c (f) European E_c to E_b (g) Asian A_a to A_b (h) Asian A_b to A_a

Figure 3: Precision at top k%. X-axis: k (log scale); Y-axis: precision.



(a) AUPR (b) AUROC

Figure 4: Pruning ratio z% on implicit target network construction from gene to disease. X-axis: z; Y-axis: AUC.

are extremely imbalanced, with only 0.65% positive instances and the remaining over 99% as negative ones.

Effect of implicit target network. We verify the effect of the implicit target network constructed at the first phase of our framework from two aspects.

First, we compare the performance of CoupledLP-IT with CoupledLP in Tables 4 and 5. They clearly show that CoupledLP outperforms CoupledLP-IT in terms of both AUPR and AUROC, which demonstrates the positive effect of the implicit target network constructed at the first phase.

Second, we also verify the effect of the implicit target network directly. Specifically, we first construct the implicit target network and then rank the links based on each unsupervised method executed in the "complete" coupled network with the implicit target network merged in ($z > 0$). We compare the evaluation results with those predicted in the coupled networks without the implicit target network ($z = 0$). The results shown in Figure 4 further demonstrate the effect of the implicit target network.

Convergence and efficiency. The learning process of our Coupled Factor Graph Model model at the second phase can converge within 200 iterations in most cases. The model code is largely developed from previous work [39], which is implemented in C++. The experiments are run on a server with Intel(R) i7 Quad-Core @2.6GHz, 16GB memory, and installed with MAC OS X Mavericks. We test the running time of each supervised method and find that for training the largest coupled network (E_b to E_a) with more than 2.59 million candidate links, LRC, DT, and Coupled Factor Graph Model models cost around 73, 68, and 30 seconds (each iteration), respectively. Note that the proposed CoupledLP also needs to extract meta-paths to construct structural correlations, which is not included in the training time. Although the proposed method needs relatively high computing cost, it is able to handle real large networks in an acceptable duration.

4.5 Applications

Our proposed coupled link prediction problem has potential applications in both scientific research and business intelligence. We discuss the applications of this problem in biology and social networks.

Biology. The identification of new associations between two genes or two diseases has been an important task in biology [30, 36]. A tremendous number of costly biological and genetic experiments have been conducted to explore the existence of genetic associations between two diseases or two genes [14, 8]. Essentially, the biologists and geneticists greedily choose each pair of diseases or genes and then examine whether there is an association between each pair. In this work, we demonstrate that the coupled link prediction problem can be directly applied to reduce the human efforts on the task. The proposed CoupledLP offers a higher 84% (AUPR) potential predictability for determining the existence of phenotypic links between diseases.

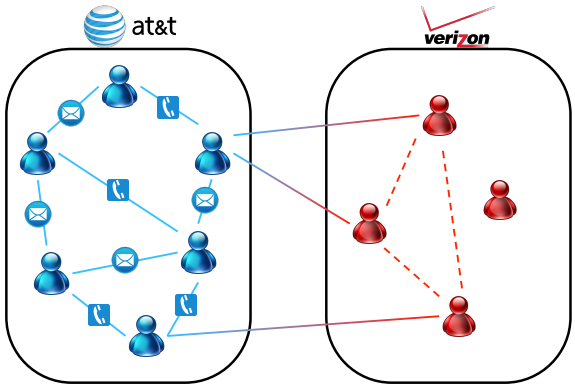


Figure 5: Illustrative example of link prediction in coupled mobile social networks. The source network is AT&T communication network and the target network is Verizon communication network. AT&T has the communication information between its users and users of another operator (Verizon). The objective is to predict the implicit links in the target (Verizon) network.

Social Networks. There are also potential applications in social networks. Figure 5 shows an illustrative example in coupled mobile social networks. In mobile social networks, a mobile operator such as AT&T has the communication network of its users and also the communication information between its users and users of another operator (e.g., Verizon). AT&T is highly motivated to acquire new users from competitors and prevent customer churns by taking the advantages of knowing competitors' network structure and user connections. By applying our coupled link prediction problem and CoupledLP framework into mobile data, a mobile operator such as AT&T can achieve an accuracy of 80% for predicting the links of its competitor's network (e.g., Verizon). Or in online social networks, a significant amount of users may register their Facebook accounts by using their Gmail accounts. It's also useful for Google+ to recommend "people you may know" to its users by knowing the Facebook connections between its Gmail users.

5. RELATED WORK

Link prediction has attracted considerable attentions in various of fields. Generally, the methods can be divided into unsupervised and supervised methods. A survey [27] provides thorough summarization of unsupervised methods. Most of the unsupervised methods are based on similarity measure between two nodes [27], e.g., common neighbors, Adamic/Adar index [1], Jaccard Coefficient, Preferential Attachment, Katz, Random Walk with Restart [37] and so on. More recently, researchers adopt supervised algorithms for link prediction [18]. Backstrom and Leskovec [3] designed a supervised random walk for friend prediction and recommendation in Facebook. Rendle et al. [34] proposed factorization machines and Kim and Leskovec [19] proposed a generative model and used EM algorithm to solve the problem. Spatial and temporal features were also incorporated into the supervised learning framework in several research [35, 42, 23]. Li et al. [26] used deep learning techniques to predict links in dynamic networks. The problem investigated in this paper is totally different from the existing link prediction problems. We propose to predict the links in one network by using the structure information of another network and the interactions between them, which is a novel and non-trivial problem. Besides, Leroy et al. [24] studied the problem of cold start in link prediction by using text and explicit information outside the networks. We propose to construct an implicit target network in our framework to also solve the cold start problem, however, the difference lies in that we only leverage the network structure without any other text information.

Several research has been conducted on predicting links in heterogeneous networks [38, 43, 22, 2, 45]. However, its objective is to predict different types of links, which is different from ours. Extensive research predicts links across multiple social networks or domains [20, 6, 11, 33, 46], while the objectives are still different. One kind focuses on leveraging the estimated parameters in one network to improve the prediction performance of the other network based on the common features between them, named as transfer link prediction. The other kind aims at predicting links in the cross network between two networks.

Our problem is also related to some link analysis tasks in social networks such as relation type prediction [9, 39, 41] and social tie strength prediction [16, 32, 10]. However, they usually target at analyzing the type or strength of a link, while we focus on predicting whether the link exists or not.

6. CONCLUSION

In this paper, we formalize a novel and non-trivial link prediction problem, named as link prediction in coupled networks, which aims at predicting the links in one network by using the pure structure information of another network and the interactions between the two networks. The major challenge of the problem is the missing links of the target network, which makes it difficult to extract features and training instances. We propose a unified framework, CoupledLP, which first propagates the knowledge from the source network to the target network and then use a coupled factor graph model to incorporate the implicit knowledge in the target network for the link prediction. The coupled factor graph model considers both the attribute features in each network and the structural meta-path based features between the two networks. The experiments on two large-scale mobile social networks and one disease-gene network show that our proposed framework outperforms several alternative baseline methods.

The problem of link prediction in coupled networks provides a new and practical research direction for link prediction. For future work, in addition to network structure, other information can be also leveraged to help predict the target links. It is also natural to design methods to automatically determine informative meta-paths in coupled networks. Furthermore, it is necessary to propose more efficient methods to integrate the general knowledge between source and target networks and capture the underlying mechanisms that drive link formation in coupled networks. Additionally, it would be interesting to investigate the social behavior of the users in the cross network between two networks. Will it be different from the behavior within one source or target network? To analyze the influence or conformity of those users in the cross network would be an interesting research topic.

Acknowledgments. We thank anonymous reviewers for their useful suggestions. The work is supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA) grant #FA9550-12-1-0405, the National High-tech R&D Program (No. 2014AA015103), National Basic Research Program of China (No. 2014CB340506, 2013CB329603), Natural Science Foundation of China (No. 61222212), National Social Science Foundation of China (No. 13&ZD190), and a research fund supported by Huawei Inc.

7. REFERENCES

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

[2] C. C. Aggarwal, Y. Xie, and P. S. Yu. A framework for dynamic link prediction in heterogeneous networks. *Statistical Analysis and Data Mining*, pages n/a–n/a, 2013.

[3] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *WSDM'11*, pages 635–644, 2011.

[4] N. Barbieri, F. Bonchi, and G. Manco. Who to follow and why: Link prediction with explanations. In *KDD '14*, pages 1266–1275. ACM, 2014.

[5] A.-L. B. Baruch Barzel. Network link prediction by global silencing of indirect correlations. *Nature Biotechnology*, 31(8):720Ű725, 2013.

[6] B. Cao, N. N. Liu, and Q. Yang. Transfer learning for collective link prediction in multiple heterogenous domains. In *ICML'10*, pages 159–166, 2010.

[7] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, May 2008.

[8] D. A. Davis and N. V. Chawla. Exploring and Exploiting Disease Interactions from Multi-Relational Gene and Phenotype Networks. *PLoS ONE*, 6(7):e22670+, July 2011.

[9] C. P. Diehl, G. Namata, and L. Getoor. Relationship identification for social network discovery. In *AAAI'07*, pages 546–552, 2007.

[10] Y. Dong, J. Tang, T. Lou, B. Wu, and N. V. Chawla. How long will she call me? distribution, social theory and duration prediction. In *Machine Learning and Knowledge Discovery in Databases*, pages 16–31. Springer Berlin Heidelberg, 2013.

[11] Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao, and H. Cao. Link prediction and recommendation across heterogeneous social networks. In *ICDM'12*, pages 181–190, 2012.

[12] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla. Inferring user demographics and social strategies in mobile social networks. In *KDD'14*, pages 15–24. ACM, 2014.

[13] N. Du, C. Faloutsos, B. Wang, and L. Akoglu. Large human communication networks: patterns and a utility-driven generator. In *KDD '09*, pages 269–278. ACM, 2009.

[14] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási. The human disease network. *PNAS*, 104(21):8685–8690, 2007.

[15] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW '04*, pages 403–412. ACM, 2004.

[16] M. Gupte and T. Eliassi-Rad. Measuring tie strength in implicit social networks. In *WebSci '12*, pages 109–118, 2012.

[17] J. M. Hammersley and P. Clifford. Markov field on finite graphs and lattices. *Unpublished manuscript*, 1971.

[18] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM'06 workshop on Link Analysis, Counterterrorism and Security*, 2006.

[19] M. Kim and J. Leskovec. The network completion problem: Inferring missing nodes and edges in networks. In *SDM'11*, pages 47–58, 2011.

[20] X. Kong, J. Zhang, and P. S. Yu. Inferring anchor links across multiple heterogeneous social networks. In *CIKM '13*, pages 179–188, 2013.

[21] F. R. Kschischang, B. J. Frey, and H. andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE TOIT*, 47:498–519, 2001.

[22] T.-T. Kuo, R. Yan, Y.-Y. Huang, P.-H. Kung, and S.-D. Lin. Unsupervised link prediction using aggregative statistics on heterogeneous social networks. In *KDD '13*, pages 775–783, 2013.

[23] C. Lee, B. Nick, U. Brandes, and P. Cunningham. Link prediction with social vector clocks. In *KDD '13*, pages 784–792. ACM, 2013.

[24] V. Leroy, B. B. Cambazoglu, and F. Bonchi. Cold start link prediction. In *KDD '10*, pages 393–402. ACM, 2010.

[25] C. W.-k. Leung, E.-P. Lim, D. Lo, and J. Weng. Mining interesting link formation rules in social networks. In *CIKM '10*, pages 209–218, 2010.

[26] X. Li, N. Du, H. Li, K. Li, J. Gao, and A. Zhang. A deep learning approach to link prediction in dynamic networks. In *SDM' 14*, pages 289–297, 2014.

[27] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *ACM CIKM '03*, pages 556–559, 2003.

[28] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *KDD '10*, pages 243–252, 2010.

[29] D. Lo, H. Cheng, and Lucia. Mining closed discriminative dyadic sequential patterns. In *EDBT '11*, pages 21–32, 2011.

[30] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224), 2015.

[31] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *UAI'99*, pages 467–475, 1999.

[32] H. Pham, C. Shahabi, and Y. Liu. Ebm: An entropy-based model to infer social strength from spatiotemporal data. In *SIGMOD '13*, pages 265–276, 2013.

[33] G.-J. Qi, C. C. Aggarwal, and T. S. Huang. Link prediction across networks by biased cross-network sampling. In *ICDE'13*, pages 793–804, 2013.

[34] S. Rendle. Factorization machines with libfm. *ACM Trans. Intell. Syst. Technol.*, 3(3):57:1–57:22, May 2012.

[35] A. Scellato, Salvatore. Noulas and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *KDD '11*, pages 1046–1054, 2011.

[36] U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon, and E. M. Marcotte. Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLoS One*, 8(5):e58977, 2013.

[37] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM '05*, pages 418–425, 2005.

[38] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla. When will it happen?: relationship prediction in heterogeneous information networks. In *WSDM '12*, pages 663–672. ACM, 2012.

[39] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *WSDM'12*, pages 743–752, 2012.

[40] J. Tang, S. Wu, J. Sun, and H. Su. Cross-domain collaboration recommendation. In *KDD '12*, pages 1285–1293, 2012.

[41] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *KDD'10*, pages 203–212, 2010.

[42] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human Mobility, Social Ties, and Link Prediction. In *KDD '11*, pages 1100–1108. ACM, 2011.

[43] Y. Yang, N. V. Chawla, Y. Sun, and J. Han. Predicting links in multi-relational and heterogeneous networks. In *ICDM'12*, pages 755–764, 2012.

[44] Y. Yang, R. Lichtenwalter, and N. V. Chawla. Evaluating link prediction methods. *Knowledge and Information Systems*, pages 1–32, 2014.

[45] J. Zhang, X. Kong, and P. S. Yu. Transferring heterogeneous links across location-based social networks. In *WSDM '14*, pages 179–188, 2014.

[46] J. Zhang, P. S. Yu, and Z.-H. Zhou. Meta-path based multi-network collective link prediction. In *KDD '14*, pages 1286–1295. ACM, 2014.