# CoupledLP: Link Prediction in Coupled Networks

**Yuxiao Dong**
University of Notre Dame
Notre Dame, IN 46556

**Jing Zhang**
Tsinghua University
Beijing, 100084

**Jie Tang**
Tsinghua University
Beijing, 100084

**Nitesh V. Chawla***
University of Notre Dame
Notre Dame, IN 46556

**Bai Wang**
Beijing U. of Posts and Telecoms
Beijing, 100876

## 1 Introduction

Link prediction is a fundamental problem in social networks, attracting considerable interest from different research fields, e.g., computer science [15, 20, 4], network science [7, 23], and biology [17, 5, 18, 8, 11]. 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 [19], or supervised learning models such as supervised random walk [3] and random forest [16] 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 [16]. 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 [15, 16, 23, 14], which generally aims at predicting the future links in the next time period. Meanwhile, the proposed problem differs from *link prediction in heterogeneous network* [24, 25, 20, 13, 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, 9, 21], 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* [22, 12], whereas it aims to predict the links in the cross network (Fig. 1(c)) between two networks. The significant advantage of the proposed problem lies

*Contact author: nchawla@nd.edu. This work was originally published at ACM SIGKDD 2015 [10]. This abstract is largely extracted from the publication.

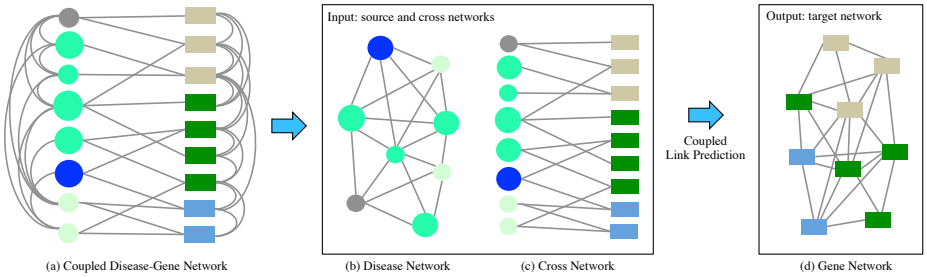


Figure 1: Illustrative example of link prediction in coupled disease-gene networks. (a) Coupled disease-gene network; (b) Disease network; (c) Cross network; (d) Gene network. Taking disease network as the source and gene network as the target, the problem of coupled link prediction aims to predict the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

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 $cc$.

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.

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.

**Acknowledgments.** The work is supported by the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Projects Agency (DARPA) grant #FA9550-12-1-0405, the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, 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.

## 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):720725, 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] 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.

[10] Y. Dong, J. Zhang, J. Tang, N. V. Chawla, and B. Wang. Coupledlp: Link prediction in coupled networks. In *KDD'15*, pages –. ACM, 2015.

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

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

[13] 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.

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

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

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

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

[18] 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.

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

[20] 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.

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

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

[23] 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.

[24] 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.

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