2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Link Prediction in Multi-relational Collaboration Networks

Xi Wang
Department of EECS
University of Central Florida
Orlando, Florida USA
Email: xiwang@eecs.ucf.edu

Gita Sukthankar
Department of EECS
University of Central Florida
Orlando, Florida USA
Email: gitars@eecs.ucf.edu

***Abstract*—Traditional link prediction techniques primarily focus on the effect of potential linkages on the local network neighborhood or the paths between nodes. In this paper, we study the problem of link prediction in networks where instances can simultaneously belong to multiple communities, engendering different types of collaborations. Links in these networks arise from heterogeneous causes, limiting the performance of predictors that treat all links homogeneously. To solve this problem, we introduce a new link prediction framework, Link Prediction using Social Features (*LPSF*), which weights the network using a similarity function based on features extracted from patterns of prominent interactions across the network.**

I. Introduction

In many social media tools, link prediction is used to detect the existence of unacknowledged linkages in order to relieve the users of the onerous chore of populating their personal networks. The problem can be broadly formulated as follows: given a disjoint node pair (x, y), predict if the node pair has a relationship, or in the case of dynamic interactions, will form one in the near future [7]. One weakness with network-based link prediction techniques is that the links are often treated as having a homogeneous semantic meaning, when in reality the underlying relationship represented by a given link could have been engendered by different causal factors. In some cases, these causal factors are easily deduced using user-supplied meta-information such as tags or circles, but in other cases the provenance of the link is not readily apparent. In particular, the meaning of links created from overlapping communities are difficult to interpret, necessitating the development of heterogeneous link prediction techniques.

When a person's true affiliations are unknown, our proposed method, *LPSF*, models link heterogeneity by adding weights to the links to express the similarities between node pairs based on their social features. These social features are calculated from the network topology using Edge Clustering [5] and implicitly encode the diversity of the nodes' involvements in potential affiliations. The weights calculated from the social features provide valuable information about the true closeness of connected people, and can also be leveraged to predict the existence of the unobserved connections. In this paper, different similarity-based prediction metrics were adapted for use on a weighted network, and the corresponding prediction scores are used as attributes for training a set of supervised link prediction classifiers. Experiments on a real-world scientific collaboration dataset (DBLP) demonstrate that *LPSF* is able to outperform homogeneous predictors in the unweighted network.

II. Method

Most of previous work in link prediction focuses on node-similarity metrics computed for unweighted networks, where the strength of relationships is not taken into account. However, proximities between nodes can be estimated better by using both graph proximity measures and the weights of existing links [1], [4]. Most of this prior work uses the number of encounters between users as the link weights. However, as the structure of the network can be highly informative, social dimensions provide an effective way of differentiating the nodes in collaborative networks [5], [6]. In this paper, the weights of the link are evaluated based on the user's social features extracted from the network topology under different similarity measures.

Our proposed link prediction framework (*LPSF*) consists of the following steps:

- Extract every node's social features (SF) using the *EdgeClustering* method.
- Calculate the similarity between node pairs based on their SFs using the *Histogram Intersection Kernel*.
- Reweight the network based on the similarities between connected node pairs.
- Apply supervised learning models for predicting links, where the prediction scores from unsupervised link prediction metrics are used as features.

We construct the node's social feature space using the scalable edge clustering method proposed in [5]. In this feature space, edges that share a common node are more similar than edges that do not. Based on the features of each edge, K-means clustering is used to separate the edges into groups using this similarity measure. Each edge cluster represents a potential affiliation, and a node will be considered involved in one affiliation as long as any of its connections are assigned to that affiliation.

In order to investigate the impact of link weights for link prediction in collaboration networks, we compare the performances of eight benchmark unsupervised metrics for unweighted networks and their extensions for weighted networks: Common Neighbors, Jaccard's Coefficient, Preferential Attachment, Adamic/Adar Coefficient, Resource Allocation

ASONAM'13, August 25-29, 2013, Niagara, Ontario, CAN

Index, Inverse Path Distance, PropFlow, and PageRank. The prediction scores from these unsupervised metrics can further be used as the attributes for learning supervised prediction models.

As mentioned in [4], unsupervised link prediction methods exhibit several drawbacks. First, they can only perform well if the network link topology conforms to the scoring function *a priori*. In other words, the assumption is both the links in the existing network and the predicted links score highly on the given measure. Second, the ranking of node pairs is performed using only a single metric, and hence the strategy may completely explore different structural patterns contained in the network. By contrast, supervised link prediction schemes can integrate information from multiple measures and can usually better model real-world networks. Most importantly, unlike in other domains where supervised algorithms require access to appropriate quantities of labeled data, in link prediction we can use the existing links in the network as the source of supervision.

III. Experimental Setup

A. Multi-relational Dataset

Our proposed method is evaluated on two real-world multi-relational collaboration networks extracted from the DBLP dataset. The DBLP dataset provides bibliographic information for millions of computer science references. In this paper we only consider authors who have published papers between 2006 and 2008, and extract their publication history from 2000 to 2008. In the constructed network, authors correspond to nodes, and two authors are linked if they have collaborated at least once. For the weighted variant, the number of coauthored publications is used as the weight on each link. Link heterogeneity is induced by the broad research topic of the collaborative work. Similar DBLP datasets have previously been employed by Kong et al. to evaluate collective classification in multi-relational networks [3]. Here, we aim to predict the missing links (coauthorship) in the future based on the existing connection patterns in the network.

B. Evaluation Framework

The supervised link prediction models are learned from training links (all existing links) in the DBLP dataset extracted between 2000 and 2008, and the performance of the model is evaluated on the testing links, new co-author link generated between 2009 and 2010. Link prediction using supervised learning model can be regarded as a binary classification task, where the class label (0 or 1) represents the link existence of the node pair. When performing the supervised classification, we sample the same number of non-connected node pairs as that of the existing links to use as negative instances for training the supervised classifier.

In our proposed *LPSF* model, the edge clustering method is adopted to construct the initial social dimensions. When conducting the link prediction experiment, we use *cosine* similarity while clustering the links in the training set. The edge-based social dimension in our proposed method, *LPSF*, is constructed based on the edge cluster IDs using the *count* aggregation operator, and varying numbers of edge clusters are tested in order to provide the best performance of *LPSF*. The weighted network is then constructed according to the similarity score of connected nodes' social features under the selected weight measure. We evaluate the performance of four supervised learning models in this paper, which are *Naive Bayes* (NB), *Logistic Regression* (LR), *Neural Network* (NN) and *Random Forest* (RF). All algorithms have been implemented in WEKA [2], and the performance of each classifiers are tested using their default parameter setting.

In DBLP dataset, the number of positive link examples for testing is very small compared to negative ones. In this paper, we sample an equivalent number of non-connected node pairs as links from the 2009 and 2010 period to use as the negative instances in the testing set. The evaluation measures for link prediction performance used in this paper are precision, recall and F-Measure.

IV. Results

Figure 1 and 2 display the comparisons between *LPSF* and the baseline methods on DBLP dataset using a variety of supervised link classification techniques, against both the unweighted and weighted supervised baselines. The same features are used by all methods, with the only difference being the weights on the network links. In this paper, we compare the proposed method *LPSF* with alternate weighting schemes, such as the number of co-authored papers, as suggested in [1]. We see that in both DBLP datasets, *Unweighted*, *Weighted* and *LPSF* perform almost equally under Precision, though *LPSF* performs somewhat worse for some classifiers (*Random Forest* and *Naive Bayes*). When considering the number of collaborations between author pairs, the *Weighted* method slightly improves upon the performance of the *Unweighted* method. The *Weighted* approach receives the most improvements on *Naive Bayes* (3% on Recall and 5% on F-Measure) in the DBLP-A dataset and on *Neural Network* (5% on Recall and 10% on F-Measure).

The proposed reweighting (*LPSF*) offers substantial improvement over both the *Unweighted* and *Weighted* schemes on Recall and F-Measure in both datasets. In the DBLP-A dataset, *LPSF* outperforms the unweighted baseline the most dramatically on *Logistic Regression*, with about 23% improvement and 40% on Recall and F-Measure respectively. In the DBLP-B dataset, *LPSF* shows the best performance using *Neural Network* with accuracy improvements over baselines for 13% on Recall and 30% on F-Measure.

LPSF calculates the closeness between connected nodes according to their social dimensions, which captures the nodes' prominent interaction patterns embedded in the network and better addresses heterogeneity in link formation. By differentiating different types of links, *LPSF* is able to discover the possible link patterns between disconnected node pairs that may not be determined by the *Unweighted* and simple *Weighted* method, and hence exhibits great improvement on Recall and F-Measure. Since *LPSF* can be directly applied on the unweighted network, without considering any additional node information, it is thus broadly applicable to a variety of link prediction domains.

Figure 1 and 2 also enable us to compare different supervised classifiers for link prediction. We found that the performance of the classifiers varies from datasets. *Logistic*

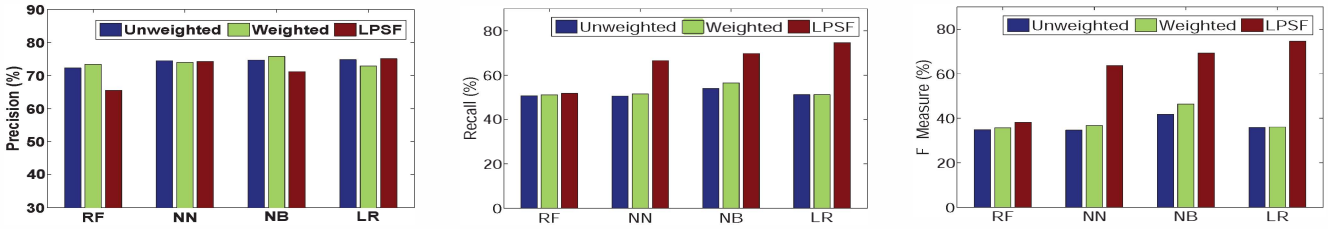

Fig. 1: Comparing the classification performance of supervised link prediction models on unweighted and weighted DBLP-A networks using Precision, Recall and F-Measure. The proposed method (*LPSF*) is implemented using 300 edge clusters and the HIK reweighting scheme. Results show that *LPSF* significantly improves over both unweighted and weighted baselines.

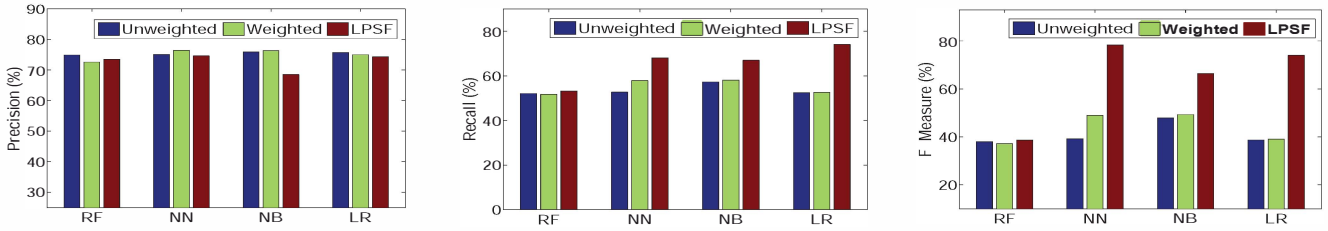

Fig. 2: Comparing the classification performances of supervised link prediction models on unweighted and weighted DBLP-B networks using Precision, Recall and F-Measure. The proposed method (*LPSF*) is implemented using 500 edge clusters and the HIK reweighting scheme. Results show that *LPSF* significantly improves over both unweighted and weighted baselines.

Regression, *Naive Bayes* and *Neural Network* exhibit comparable performance. Somewhat surprisingly, *Random Forest* does not perform well with *LPSF*. We also observe that *LPSF* using *Naive Bayes* will boost the Recall performance over baseline methods at the cost of lower Precision. Therefore *Logistic Regression* and *Neural Network* will be a better choice for *LPSF* in that they improve the Recall performance without decreasing the Precision. Using the traditional weighted features [1] does not help supervised classifiers for link prediction to a great extent. As discussed above, reweighting the unweighted collaboration network using our proposed technique, *LPSF*, performs the best.

V. Conclusion

In this paper, we investigate the link prediction problem in collaboration networks with heterogeneous links. Most commonly-used link prediction methods assume that the network is in unweighted form, and treat each link equally. In this paper, we proposed a new link prediction framework *LPSF* that captures nodes' intrinsic interaction patterns from network topology and embeds the similarities between connected nodes as link weights. The nodes' similarity is calculated based on social features extracted using Edge Clustering to detect overlapping communities in the network. Experiments on the DBLP collaboration network demonstrate that the judicious choice of weight measure in conjunction with supervised link prediction enables us to significantly outperform existing methods. Our proposed method is better able to capture the true proximity between node pairs based on link group information and improve the performance of supervised link prediction methods. The strength of our approach is that it extracts communities in an unsupervised way, and thus can be used to study informal patterns of contact between researchers. These informal patterns, described as the "invisible college" in bibliometric research, can be a powerful but difficult to quantify force behind the process of scientific collaboration [8]. The proposed method is very practical: it can be employed on any unweighted or weighted network in conjunction with any existing link prediction classifier. Moreover, the social features are themselves complementary to node-based approaches.

VI. Acknowledgments

This research was supported in part by DARPA award D13AP00002 and NSF IIS-08451.

References

[1] H. R. de Sá and R. B. C. Prudêncio. Supervised link prediction in weighted networks. In *International Joint Conference on Neural Networks (IJCNN)*, pages 2281–2288, 2011.

[2] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1):10–18, Nov. 2009.

[3] X. Kong, X. Shi, and P. S. Yu. Multi-label collective classification. In *SIAM International Conference on Data Mining (SDM)*, pages 618–629, 2011.

[4] T. Murata and S. Moriyasu. Link prediction of social networks based on weighted proximity measures. In *Web Intelligence*, pages 85–88, 2007.

[5] L. Tang and H. Liu. Scalable learning of collective behavior based on sparse social dimensions. In *Proceedings of International Conference on Information and Knowledge Management (CIKM)*, 2009.

[6] X. Wang and G. Sukthankar. Extracting social dimensions using Fiedler embedding. In *Proceedings of IEEE International Conference on Social Computing*, pages 824–829, 2011.

[7] E. W. Xiang. A survey on link prediction models for social network data. *Science and Technology*, 2008.

[8] A. Zuccala. Modeling the invisible college. *Journal of the American Society for Information Science and Technology*, 57(2):152–168, 2005.