IJCSS
International Journal of Complex Systems in Science

Int. J. Complex Systems in Science
vol. **3(1)** (2013), pp. 77–82

Link prediction in multiplex bibliographical networks

Manisha Pujari[1,†] and Rushed Kanawati[1]

[1] *Laboratoire d'Informatique de Paris Nord (LIPN), Université Sorbonne Paris Cité, CNRS UMR 7030 Villetaneuse, France*

Abstract. In this work we present a new approach for co-authorship link prediction based on leveraging information contained in general bibliographical multiplex networks. A multiplex network, also called multi-slice or multi-relational network, that we consider here is a graph defined over a set of nodes linked by different types of relations. For instance, the multiplex network we are studying here is defined as follows : nodes represent authors and links can be one of the following types : co-authorship links, co-venue attending links and co-citing links. Other types of links can also be considered involving bibliographical coupling, research theme sharing and so on.

We show here a new approach for exploring the multiplex relations to predict future collaboration (co-authorship links) among authors. The applied approach is a supervised-machine learning approach where we attempt to learn a model for link formation based on a set of topological attributes describing both positive and negative examples. While such an approach has been successfully applied in the context on simple networks, different options can be applied to extend it to the multiplex network context. One option is to compute topological attributes in each layer of the multiplex. Another one is to compute directly new multiplex-based attributes quantifying the multiplex nature of dyads (potential links).

We show our first results on experiments conducted on real datasets extracted from the famous bibliographical database; DBLP that has been enriched with citation informations.

Keywords: Link prediction, multiplex networks, supervised machine learning

† **Corresponding author:** manisha.pujari@lipn.univ-paris13.fr
Received: December 5th, 2013
Published: December 31th, 2013

1. Introduction

Link prediction plays an important role in the analysis of complex networks with its wide range of application like identification of missing links in biological networks, identification of hidden and new criminal links, recommendation systems, prediction of future collaborations or purchases in e-commerce. It can be defined as the process of identifying missing or new links in a network by studying the history of the network.

A popular category of link prediction approaches is *dyadic topological approaches* which consider only the graph structure and compute a score for unconnected nodes pairs. In a seminal work of Liben-Nowell et al.[8], authors have shown that simple topological features characterising pairs of unlinked nodes, can be used for predicting formation of new links.

They propose to sort a list of unconnected node pairs according to the values of a topological measure. The top k node pairs are then returned as the output of the prediction task. Here an assumption is made that the topological measure should be able to rank the most probable new links on the top.

Many other works have been published focusing on combining different topological metrics to enhance prediction performances which convert the link prediction task to a binary classification problem and hence use machine learning algorithms[9, 7].

But all these work address the link prediction in only simple networks which have homogeneous links. To our knowledge, not much have been explored to add multiplex information for the task of link prediction. Although there are a few recent work proposing methods for prediction of links in heterogeneous networks, networks which have different types of nodes as well as edges[1]. There have also been few work on extending simple structural features like degree, path etc. to the context of multiplex networks [4, 6] but none have attempted to use them for link prediction. We propose a new approach for exploring the multiplex relations to predict future collaboration (co-authorship links) among authors. The applied approach is a supervised-machine learning approach where we attempt to learn a model for link formation based on a set of topological attributes describing both positive and negative examples. While such an approach has been successfully applied in the context on simple networks, different options can be applied to extend it to the multiplex network context. One option is to compute topological attributes in each layer of the multiplex. Another one is to compute directly new multiplex-based attributes quantifying the multiplex nature of dyads (potential links). Both approaches will be discussed in the next section.

2. Link prediction approach

Our approach includes computing simple topological scores for unconnected node pairs in a graph. Then we extend these attributes to include information from other dimension graphs. This can be done in three ways: First we compute the simple topological measures in all dimensions; second is to take the average of the scores; and third we propose an entropy based version of each topological measures which gives importance to the presence of a non-zero score of the node pair in each dimension. In the end all these attributes can be combined in various ways to form different sets of vectors of attribute values characterizing each example or unconnected node pair. Formally, if we have a multiplex graph $G =< V, E_1, \ldots, E_m >$ which in fact is a set of graphs $< G_1, G_2, \ldots, G_m >$ and a topological attribute X. For any two unconnected nodes u and v in graph G_i (where we want to make a prediction), $X(u, v)$ computed on G_i will be *direct* attribute and the same computed on all other dimension graphs will be *indirect* attributes. The second category computes an average of the attribute over all the dimension i.e. $X_{average} = \frac{\sum_{\alpha=1}^{m} X(u,v)^{[\alpha]}}{m}$ for $u, v \in V$ and $(u, v) \notin E_i$. where m is the number of types of relations in the graph (dimension or layer). In the third category we propose a new attribute called *product of node degree entropy (PNE)* which is based on *degree entropy*, a multiplex property proposed by F. Battistion et al. [4]. If degree of node u is $k(u)$, the degree entropy is given by: $E(u) = -\sum_{\alpha=1}^{m} \frac{k(u)^{[\alpha]}}{k_{total}} \log(\frac{k(u)^{[\alpha]}}{k_{total}})$ where $k_{total} = \sum_{\alpha=1}^{m} k(u)^{[\alpha]}$ and we define *product of node degree entropy* as

$$PNE(u, v) = E(u) * E(v)$$

We also extend the same concept to define entropy of a simple topological attribute, say X_{ent}

$$X_{ent}(u, v) = -\sum_{\alpha=1}^{m} \frac{X(u, v)^{[\alpha]}}{X_{total}} \log(\frac{X(u, v)^{[\alpha]}}{X_{total}})$$

where $X_{total} = \sum_{\alpha=1}^{m} X(u, v)^{[\alpha]}$. The entropy based attributes are more suitable to capture the distribution of the attribute value over all dimensions. A higher value indicates uniform distribution attribute value across the multiplex layers. We address average and entropy based attributes as *multiplex attributes*.

3. Experiments

We evaluated our approach using data obtained from DBLP[1] databases of which we created three datasets, each corresponding to a different period of time. Table.1 summarizes the information about the graphs of each dataset. Each graph has four years for learning or training and next two years are used to label the examples generated from the learning graphs. Examples are unconnected node pairs and they are labelled as *positive* or *negative* based on whether they are connected during the labelling period or not. Table.2 shows the number of examples obtained for each dataset.

Years	**Properties**	**Co-Author**	**Co-Venue**	**Co-Citation**
1970-1973	*Nodes*	91	91	91
	Edges	116	1256	171
1972-1975	*Nodes*	221	221	221
	Edges	319	5098	706
1974-1977	*Nodes*	323	323	323
	Edges	451	9831	993

Table 1: Graphs

Years		**# Positive**	**# Negatives**
Train/Test	**Labeling**		
1970-1973	1974-1975	16	1810
1972-1975	1976-1977	49	12141
1974-1977	1978-1979	93	26223

Table 2: Examples from co-authorship graph

We selected the following topological attributes: Number of common neighbors (CN), Jaccard coefficient (JC), Preferential attachment (PA)[5], Adamic Adar coefficient (AA)[3], Resource allocation (RA)[2] and Shortest path length (SPL). We applied decision tree algorithm on one dataset to generate a model and then tested it on another dataset. We are using data mining tool Orange[2] for that.

We use four types of combinations of the attributes creating five different sets namely: Set_{direct}(attributes computed only in the co-authorship graph); $Set_{direct+indirect}$(attributes computed in co-authorship, co-venue and co-citation graphs); $Set_{direct+multiplex}$(attributes computed from co-authorship

[1] http://www.dblp.org
[2] http://orange.biolab.si

graph with average attributes obtained from three dimension graphs, and also entropy based attributes); Set_{all}(attributes computed in co-authorship, co-venue and co-citation graphs, with average of the attributes, and also entropy based attributes) and $Set_{multiplex}$(average attributes and entropy based attributes). Table.3 shows the result obtained in terms of F1-measure and area under the ROC curve (AUC). We can see that there is improvement in the F1-measure when we use multiplex attributes. AUC is better for all the sets that include multiplex and indirect attributes for both datasets.

Attributes	**Learning:1970-1973 Test:1972-1975**		**Learning:1972-1975 Test:1974-1977**	
	F-measure	**AUC**	**F-measure**	**AUC**
Set_{direct}	0.0357	0.5263	0.0168	0.4955
$Set_{direct+indirect}$	0.0256	0.5372	0.0150	0.5132
$Set_{direct+multiplex}$	0.0592	0.5374	0.0122	0.5108
Set_{all}	0.0153	0.5361	0.0171	0.5555
$Set_{multiplex}$	0.0374	0.5181	0.0185	0.5485

Table 3: Results of decision tree algorithm

4. Conclusion

This paper presents our new approach of link prediction in multiplex networks. We propose some new and extended topological features that can be used for characterizing the unlinked node pairs for link prediction task, including also multiplex relation information. They can be applied to predict links in any of the layers of the network. We tested our method for prediction of co-authorship links on datasets obtained from DBLP databases. The preliminary results show that addition of multiplex information indeed improve the prediction performance and thereby motivates us to continue our research further to better confirm this concept.

References

[1] Yizhou Sun, Rick Barber, Manish Gupta, Charu C. Aggarwa and Jiawei Han, *Advances on social network Analysis and mining (ASONAM)*, (2011).

[2] Tao Zhou, Lu Linyuan and Yi-Cheng Zhang, *Predicting Missing Links via Local Information* , (2009).

[3] Lada Adamic, Orkut Buyukkokten, and Eytan Adar, *A social network caught in web (First Monday)* **8 (6)**, (2003).

[4] Federico Battiston, Vincenzo Nicosia and Vito Latora, *Metrics for the analysis of multiplex networks*, (2013).

[5] Zan Huang, Xin Li and Hsinchun Chen, *Link prediction approach to collaborative filtering (JDLC)* , (2005).

[6] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale and D. Pedreschi, *Foundations of Multidimensional Network Analysis(ASONAM)*, 485–489 (2009).

[7] Nesserine Benchettara, Rushed Kanawati and Céline Rouveirol, *A supervised machine learning link prediction approach for academic collaboration recommendation (RecSys '10)*, 253 (2010).

[8] David Liben-Nowell and Jon M. Kleinberg, *The link prediction problem for social networks (CIKM)*,556–559 (2003).

[9] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem and Mohammed Zaki, *Link prediction using supervised learning (SIAM Workshop on Link Analysis, Counterterrorism and Security with SIAM Data Mining Conference)*, (2006).