2011 11th IEEE International Conference on Data Mining Workshops

Scalable Link Prediction on Multidimensional Networks

Giulio Rossetti
KDDLab, ISTI-CNR
Pisa, Italy
giulio.rossetti@isti.cnr.it

Michele Berlingerio
KDDLab, ISTI-CNR
Pisa, Italy
michele.berlingerio@isti.cnr.it

Fosca Giannotti
KDDLab, ISTI-CNR
Pisa, Italy
fosca.giannotti@isti.cnr.it

***Abstract*—Complex networks have been receiving increasing attention by the scientific community, also due to the availability of massive network data from diverse domains. One problem largely studied so far is Link Prediction, i.e. the problem of predicting new upcoming connections in the network. However, one aspect of complex networks has been disregarded so far: real networks are often multidimensional, i.e. multiple connections may reside between any two nodes. In this context, we define the problem of Multidimensional Link Prediction, and we introduce several predictors based on structural analysis of the networks. We present the results obtained on real networks, showing the performances of both the introduced multidimensional versions of the Common Neighbors and Adamic-Adar, and the derived predictors aimed at capturing the multidimensional and temporal information extracted from the data. Our findings show that the evolution of multidimensional networks can be predicted, and that supervised models may improve the accuracy of underlying unsupervised predictors, if used in conjunction with them.**

***Keywords*-Link Analysis; Graph Mining; Link Prediction; Network Analysis;**

I. Introduction

Network Science, Graph Data Mining and Social Network Analysis are receiving large attention in the last years, also thanks to the increasing availability of real network data, mostly concerning human behaviors. One hot topic of research in these fields is studying dynamic networks. Researchers have been investigating, from the global to the local level, problems such as the analysis of structural changes during time [1],the evolution of communities [2], the extraction of frequent local patterns of evolution [3], and the prediction of new nodes and links joining the network structure in the future [4], [5], [6], [7], i.e., the Link Prediction problem [8]. So far, these problems have been studied on *monodimensional network*, i.e. networks where only one connection between two nodes is possible. Real world networks, however, are often multidimensional: two nodes may be connected by more than one relation, that we call *dimensions*, expressing either different types of relationship (e.g. friends, colleagues, relatives), or different quantitative values of the same kind of relationship (e.g. different ranks, or different publication venues for the same co-authorship relation). The additional degree of freedom that different dimensions add to the classical problems of network analysis, makes it difficult, when not impossible, to treat these kind of networks with the available tools. Questions such as "can we predict the evolution of a multidimensional network?", "what is the probability that a new link between two specific nodes will form in dimension one?", require new tools of analysis that take the interplay among dimensions into account.

In this context, we introduce the Multidimensional Link Prediction problem. Following the approach of a large family of studies based on structural properties of the network such as, for example, Common Neighbors [5], Preferential Attachment [6] or Adamic-Adar [4], we introduce several new classes of predictors able to exploit the knowledge that can be learned from the multidimensional structure of networks to predict new links in specific dimensions. In our vision, the evolution of a multidimensional network depends on three factors: i) the underlying theoretical model of node interactions (e.g. nodes with high degree tend to attract more connections); ii) the interplay among dimensions (e.g. links may form in a specific dimension with a higher likelihood); iii) the complete temporal history of a link (e.g. links always present during the network history may be more likely to appear also in the future). In oder to reflect this, we build predictors that combine the contribution of three basic measurements used in conjunction: i) multidimensional versions of Common Neighbors and Adamic-Adar, ii) global measures capturing the interplay of multiple dimensions at different levels, and iii) measures based on the complete history of the presence of a link within a network.

Our contribution can be then summarized as follows: we introduce structural measures for capturing the interplay among dimensions and a few measures on the temporal history of a link between two nodes (Section II); we define the Multidimensional Link Prediction problem, and we propose several scalable predictors based on combinations of the introduced concepts (Section III); we give experimental evaluation of the proposed approaches on real world multidimensional networks (Section IV). Our results show that we are able to predict the evolution of a multidimensional network, and that multidimensional and temporal information, used either individually or in combination, can improve the accuracy of the classical theoretical predictors based on structural properties of networks.

978-0-7695-4409-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDMW.2011.150

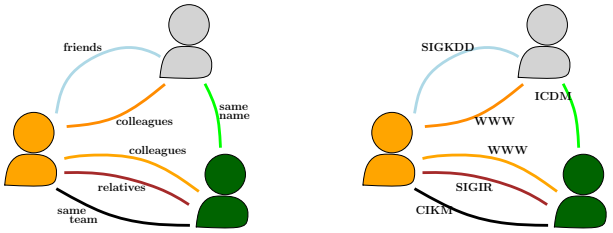


Figure 1. Example of multidimensional networks

II. Multidimensional Networks

Figure 1 depicts two possible multidimensional networks, where on the left we have different types of links connecting the three nodes (they can be friends, colleagues, and so on), while on the right we have different values (publication venues) for one relationship (for example, co-authorship). In this setting, multidimensional analysis is needed to distinguish among different kinds of interactions, or equivalently to look at interactions from different perspectives. Dimensions in network data can be either explicit or implicit. In the first case, where usually the nodes explicitly set their connections (two users of a social network become friends, two computers exchange a message, and so on), the dimensions directly reflect the various interactions in reality; in the second case, the dimensions are defined (mostly in a passive way from the perspective of the nodes) by the analyst to reflect different interesting qualities of the interactions, that can be inferred from the available data.

We now present a formal model for multidimensional networks, and a set of measures for them.

A. A model for multidimensional networks

We use a *multigraph* to model a multidimensional network and its properties. For the sake of simplicity, in our model we only consider undirected multigraphs and since we do not consider node labels, hereafter we use *edge-labeled undirected multigraphs*, denoted by a tuple $\mathcal{G} = (V, E, L, T, \tau)$ where: V is a set of nodes; L is a set of labels; E is a set of labeled edges, i.e. the set of triples (u, v, d) where $u, v \in V$ are nodes and $d \in L$ is a label; T is a set of timestamps; $\tau : E \rightarrow \mathcal{P}(T)$ is a function returning the set of timestamps of presence of a given edge (and $\mathcal{P}(T)$ denotes the power set of T). Where we are not interested in the temporal history of an edge, we refer to it by simply using a triple (u, v, d). Whenever, in turns, we need to specify the temporal information, we use the pair $((u, v, d), \tau(u, v, d))$. Moreover, if we write $(u, v, d) \in E$ we assume: $\tau(u, v, d) \neq \emptyset$, i.e. there exist at least one timestamp t in which the edge (u, v, d) is present, or, in other words, in which the dimension d connects u and v. Hereafter, we omit this note in the definitions of our structural measures when this is not needed.

Also, we use the term *dimension* to indicate *label*, and we say that a node *belongs to* or *appears in* a given dimension d if there is at least one edge labeled with d adjacent to it. We also say that an edge *belongs to* or *appears in* a dimension d if its label is d. We assume that given a pair of nodes $u, v \in V$ and a label $d \in L$ only one edge (u, v, d) may exist. Thus, each pair of nodes in $\mathcal{G}$ can be connected by at most $|L|$ possible edges. Hereafter $\mathcal{P}(L)$ denotes the power set of L.

B. Connectivity measures for multidimensional networks

Here we define new measures on the multidimensional structure of networks, and we assume: $\tau(u, v, d) \neq \emptyset$, thus we omit the temporal information in our measures. This section defines only a small set of the possible measures that can be defined over multidimensional networks, and this list is not meant to be exhaustive, as this is not the purpose of this paper. However, in Section IV we see how even the following few concepts can be used effectively to build a multidimensional predictor.

*1) **Neighbors**:* In classical graph theory the *degree* of a node refers to its connections in a network: it is defined as the number of edges adjacent to a node. In a simple graph, each edge is the sole connection to an adjacent node. In multidimensional networks the degree and the number of nodes adjacent to the node are no longer related, since there may be more than one edge between any two nodes. For instance, in Figure 2, the node 4 has five neighbors and degree equal to 7. In order to capture this we define the following:

Definition 1 (Neighbors): Let $v \in V$ and $D \subseteq L$ be a node and a set of dimensions of a network $\mathcal{G} = (V, E, L, T, \tau)$, respectively. The function $Neighbors : V \times \mathcal{P}(L) \rightarrow \mathcal{P}(V)$ is defined as $Neighbors(v, D) = \{u \in V \mid \exists (u, v, d) \in E \ \wedge \ d \in D\}$, where $\mathcal{P}(V)$ denotes the power set of V. This function returns the set of all the nodes directly reachable from node v by edges labeled with dimensions belonging to D.

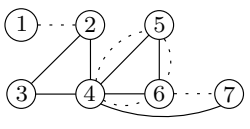
While this measure might be used directly into the formulas of Common Neighbors and Adamic-Adar to adapt their formulation for the multidimensional setting, we define a variant of it, aimed at capturing the interplay among the dimensions w.r.t the exclusivity of the connections.

Definition 2 (Neighbors$_{XOR}$): Let $v \in V$ and $D \subseteq L$ be a node and a set of dimensions of a network $\mathcal{G} = (V, E, L, T, \tau)$, respectively. The function $Neighbors_{XOR} : V \times \mathcal{P}(L) \rightarrow \mathcal{P}(V)$ is defined as

$$Neighbors_{XOR}(v, D) =$$
$$\{u \in V | \ \exists d \in D : (u, v, d) \in E \wedge \nexists d' \notin D : (u, v, d') \in E\}$$

It returns the set of neighboring nodes connected by edges belonging only to dimensions in D.

*2) **Dimension Connectivity**:* While the two measures above are local to nodes, here we define four global measures on the sets of nodes and edges. To this end we introduce the *Dimension Connectivity* and *Average Correlation* measures on both the sets of nodes and edges.



$\tau(1,2,2) = \{1,3\}$ $\tau(2,3,1) = \{2,3\}$
$\tau(2,4,1) = \{1\}$ $\tau(3,4,1) = \{3\}$
$\tau(4,5,1) = \{2,3\}$ $\tau(4,5,2) = \{2,3\}$
$\tau(4,6,1) = \{3\}$ $\tau(4,6,2) = \{2,3\}$
$\tau(4,7,1) = \{1,2\}$ $\tau(5,6,1) = \{1,2,3\}$
$\tau(5,6,2) = \{1,3\}$ $\tau(6,7,2) = \{1,2,3\}$

Figure 2. Toy example. Solid line is dimension 1, dashed is dimension 2.

Definition 3 (Node Dimension Connectivity): Let $d \in L$ be a dimension of a network $\mathcal{G} = (V, E, L, T, \tau)$. The function $NDC : L \to [0, 1]$ defined as

$$NDC(d) = \frac{|\{u \in V | \exists v \in V : (u, v, d) \in E\}|}{|V|}$$

computes the ratio of nodes of the network that belong to the dimension d.

Definition 4 (Edge Dimension Connectivity): Let $d \in L$ be a dimension of a network $\mathcal{G} = (V, E, L, T, \tau)$. The function $EDC : L \to [0, 1]$ defined as

$$EDC(d) = \frac{|\{(u, v, d) \in E | u, v \in V\}|}{|E|}$$

computes the ratio of edges of the network labeled with the dimension d.

While the two measures above regard the importance that a single dimension has w.r.t. the connectivity of the network, we now define other two measures aimed at capturing the global interplay among dimensions, by looking at their average correlation.

Definition 5 (Average Node Correlation): Let $d \in L$ be a dimension of a network $\mathcal{G} = (V, E, L, T, \tau)$. The function $ANC : L \to [1/|L|, 1]$ is defined as

$$ANC(d) = \frac{\sum_{d' \in L} NJaccard(d, d')}{|L|}$$

where $NJaccard(d, d')$ is the Jaccard correlation index on the node sets $\frac{|N(d) \cap N(d')|}{|N(d) \cup N(d')|}$, where $N(\bar{d}) = \{u \mid \exists\ (u, v, \bar{d}) \in E\}$. It computes the average node correlation of a dimension with all the others.

Definition 6 (Average Edge Correlation): Let $d \in L$ be a dimension of a network $\mathcal{G} = (V, E, L, T, \tau)$. The function $AEC : L \to [1/|L|, 1]$ is defined as

$$AEC(d) = \frac{\sum_{d' \in L} EJaccard(d, d')}{|L|}$$

where $EJaccard(d, d')$ is the Jaccard correlation index on the edge sets $\frac{|E(d) \cap E(d')|}{|E(d) \cup E(d')|}$, where $E(\bar{d}) = \{(u, v) \mid \exists (u, v, \bar{d}) \in E\}$. It computes the average edge correlation of a dimension with all the others.

Example 1: In Figure 2 the EDC of dimension d_1 (solid line) is $7/12$ since it has 7 edges out of the 12 total edges of the network, while the EDC of d_2 (dashed line) is $5/12$. The NDC for d_1 is $5/7$ and NDC for d_2 is $6/7$. The AEC of d_1 is $(1 + 3/12)/2 = 0.625$. For the same dimension, ANC is $(1 + 5/7)/2 = 0.857$.

*3) **Temporal Link Information**:* Besides the analysis of the multidimensional structure at both the local and global levels, we also want to take into account the complete temporal history of an edge of the network. In order to do this, we define four different measures. Thus, here we make use of the τ function.

The first measure simply counts the number of temporal snapshots in which an edge is present in a dimension:

Definition 7 (Frequency): Let $(u, v, d) \in E$ be an edge of a network $\mathcal{G} = (V, E, L, T, \tau)$. The function $Freq : E \to [1, |T|]$ defined as

$$Freq(u, v, d) = \mid \tau(u, v, d) \mid$$

computes the frequency of an edge in terms of the number of temporal snapshots in which it appears.

We can aggregate the above by dimensions, counting the number of snapshots in which a pair of nodes is connected:

Definition 8 (Over All Frequency): Let (u, v) be two nodes in V in a network $\mathcal{G} = (V, E, L, T, \tau)$. We define $OAFreq : V \times V \to [1, |L| \times |T|]$ as:

$$OAFreq(u, v) = \left| \bigcup_{\{d \in L | (u,v,d) \in E\}} \tau(u, v, d) \right|$$

As time has a natural ordering, we may want to be able to give more (or less) importance to more recent interactions when predicting new ones. To this end, we define two weighted measures on the temporal history of an edge:

Definition 9 (Weighted Presence): Let $(u, v, d) \in E$ be an edge of a network $\mathcal{G} = (V, E, L, T, \tau)$. The function $WPres : E \to [1, +\infty]$ is defined as

$$WPres(u, v, d) = \sum_{\{t \in \tau(u,v,d)\}} w_t$$

where w_t is the weight of the temporal snapshot t. For simplicity, given the temporal ordering, we assume $w_{t_i} = i$.

As done above, we can also aggregate $WPres$ by dimensions:

Definition 10 (Over All Weighted Presence): Let (u, v) be two nodes in V in a network $\mathcal{G} = (V, E, L, T, \tau)$. The function $OAWPres : V \times V \to [1, +\infty]$ defined as

$$OAWPres(u, v) = \sum_{\{d | (u,v,d) \in E\}} WPres(u, v, d)$$

Example 2: In the toy example in Figure 2, where we reported also the complete history of each edge in the table, we have: Freq(4,5,1)=2; OAFreq(4,5)=4; WPres(4,5,1)=5; OAWPres(4,5)=10.

III. Multidimensional Link Prediction

A. Problem statement

Given a pair of nodes in an evolving network, the literature on monodimensional network analysis defines Link Prediction (LP, hereafter) as the problem of estimating the likelihood that an edge will form between two nodes [8], [7]. There can be several ways to reformulate it in the multidimensional setting. For example, the classical definition may be preserved as it is, disregarding the dimensions, only focusing on new connections between any two nodes. Another possible way is to specify a set of dimensions for which we want to estimate the likelihood. A more specific formulation, that we use in the rest of the paper, is estimating the likelihood that an edge will form between two nodes in a specific dimension. That is, we add an additional parameter to the classical definition. More formally we define:

Definition 11 (Multidimensional Link Prediction): Given a multidimensional network modeled as a multigraph $\mathcal{G} = (V, E, L, T, \tau)$, the Multidimensional Link Prediction problem (from now on, MLP) requires to return a function

score : $V \times V \times L \rightarrow [0, +\infty[$ of scores measuring the likelihood that any two pairs of nodes will connect in a specific dimension, in the future.

We now present several possible solutions for MLP, introducing a list of functions to use as scores. It is clear how, in analogy with the LP problem in the monodimensional case, there can be a taxonomy of solutions, divided in supervised or unsupervised approaches, based on structural analysis or on the extraction of frequent patterns of evolution, based on statistical analysis of temporal series, and so on. In the rest of this section we present solutions based on the structural analysis of the network. We start from the multidimensional reformulation of two classical approaches based on neighborhood (Common Neighbors and Adamic-Adar), then we introduce other measures to be taken into account in the final list of scoring functions. Our resulting solutions are then combinations of supervised and unsupervised approaches, aimed at capturing all the possible strong and weak signals of the non-trivial interplay of multidimensionality and temporal evolution.

B. Predictive models based on structural analysis

We now combine all the available theoretical basic bricks to build our set of predictors for MLP. For convenience, in this section we use the notation $N(\circ, \bullet)$ for $Neighbors(\circ, \bullet)$, and, in analogy, $N_{XOR}(\circ, \bullet)$ for $Neighbors_{XOR}(\circ, \bullet)$.

1) ***Base predictors****:* We wanted to have basic predictors for our experiments, and we choose Common Neighbors [5] and Adamic-Adar [4], as they are among the best w.r.t predictive performances [8]. We can introduce a multidimensional version of them by using our function Neighbors:

Definition 12 (Multidimensional Common Neighbors): Let $\mathcal{G} = (V, E, L, T, \tau)$ be a network and $(u, v, d) \notin E$ be a candidate future edge. We define:

$$Multidimensional\ Common\ Neighbors(u, v, d) = \mid N(u, d) \cap N(v, d) \mid$$

Hereafter, we often use M-CN to refer to this predictor.

Definition 13 (Multidimensional Adamic Adar): Let $\mathcal{G} = (V, E, L, T, \tau)$ be a network and $(u, v, d) \notin E$ be a candidate future edge. We define:

$$Multidimensional\ Adamic\ Adar(u, v, d) = \sum_{z \in \{N(u,d) \cap N(v,d)\}} \frac{1}{log(|N(z,d)|)}$$

Hereafter, we often use M-AA to refer to this predictor.

In the following, instead, we replace Neighbors with Neighbors$_{XOR}$, by following the intuition that more sophisticated multidimensional information may lead to better predictive performance. As we see in Section IV, this intuition was proved to be incorrect in the networks used.

Definition 14 (Multidimensional Common Neighbors$_{XOR}$): Let $\mathcal{G} = (V, E, L, T, \tau)$ be a network and $(u, v, d) \notin E$ be a candidate future edge. We define:

$$Multidimensional\ Common\ Neighbors_{XOR}(u, v, d) = \mid N_{XOR}(u, d) \cap N_{XOR}(v, d) \mid$$

Definition 15 (Multidimensional Adamic Adar$_{XOR}$): Let $\mathcal{G} = (V, E, L, T, \tau)$ be a network and $(u, v, d) \notin E$ be a candidate future edge. We define:

$$Multidimensional\ Adamic\ Adar_{XOR}(u, v, d) = \sum_{z \in \{N_{XOR}(u,d) \cap N_{XOR}(v,d)\}} \frac{1}{log(|N_{XOR}(z,d)|)}$$

As for above, we use M-CN$_{XOR}$ and M-AA$_{XOR}$ hereafter to refer to these two predictors, respectively.

2) ***Multidimensional scores****:* In principle, it is possible to define several scores on the basis of the multidimensional measures presented above. For example, it is possible to multiply the $Neighbors_{XOR}$ of two nodes in one dimension to obtain a score, ending up with a Preferential-Attachment like model [6]. We tried several combinations, but, due to extremely poor predictive performances as tested during our experimental stage, we do not report their definition. According to our experiments, in fact, the multidimensional information gathered by our measures in the networks used is not enough to predict new edges. This negative result is analog to the one obtained by the authors of [7], who reported that their supervised model was not performing well when used alone for prediction. In analogy with their strategy, we tried then to combine the information learned from the data, with unsupervised model, as we see in 4).

3) ***Temporal scores****:* It is possible to define temporal scores based on modifications of the above measures. We tried a few of them but, in analogy with the multidimensional scores, their predictive power when used alone was very poor on our networks.

4) ***Combinations****:* Finally, we can define a scoring function by combining all the basic bricks presented in our theory. In particular, we can aggregate the information provided by the baseline models with the information provided by the multidimensional measures or the temporal ones. This is exactly the line followed in [7], where the authors combine the information provided by the model defined by the complete set of frequent evolution rules mined from the network with the information provided by the baseline models. In analogy with their paper, we tried several combinations of our proposed measures. Table I shows the non-XOR versions of all the solutions we tested. Each line represents which basic bricks we used for building one scoring function, for a total of 26 predictors. The basic bricks were combined by multiplying their scores. Clearly, other aggregates or combinations are possible and we tried some of them, but, due to poor predictive power and to lack of space, here we only report the best ones.

5) ***Implementation and complexity****:* All the measures defined, and the predictors presented, may be implemented by trivially scanning the list of edges linearly, thus making the approach scalable (see Section IV for an empirical evaluation of scalability). We omit the implementation details due to lack of space.

Base	Multidim. Measure	Temporal Measure
M-AA		
M-AA	NDC	
M-AA	EDC	
M-AA	AEC	
M-AA	ANC	
M-AA		Freq
M-AA		OAFreq
M-AA		WPres
M-AA		OAWpres
M-AA	AEC	WPres
M-AA	AEC	OAWPres
M-AA	ANC	WPres
M-AA	ANC	OAWPres

Base	Multidim. Measure	Temporal Measure
M-CN		
M-CN	NDC	
M-CN	EDC	
M-CN	EC	
M-CN	NC	
M-CN		Freq
M-CN		OAFreq
M-CN		WPres
M-CN		OAWPres
M-CN	AEC	WPres
M-CN	AEC	OAWPres
M-CN	ANC	WPres
M-CN	ANC	OAWPres

Table I
TAXONOMY OF THE PROPOSED APPROACHES (NON-XOR VERSIONS)

IV. EXPERIMENTS

In this section we report the results obtained by applying our predictors on real networks. The predictive performance is measured via ROC (Receiver Operating Characteristic) curves computed on the results of the predictors. We use ROC curves instead of Precision/Recall plots for their better comparability among different networks and predictors.

All the tests were ran on a server with an AMD Phenom II X4 processor at 3.2GHz, with 8GB of RAM, running Linux 2.6.35. The predictors were implemented in Java.

A. Datasets

We built two networks coming from different real world sources: the bibliographic database DBLP[1], from which we extracted a co-authorship network, and the movie database IMDb[2], from which we extracted a collaboration network. More in details, we built the following two networks:

- **DBLP.** We extracted author-author relationships if two authors collaborated at least in one paper. The dimensions are defined as the venues in which the paper was published. We took only the publications in the most important 28 conferences in computer science, which include VLDB, SIGKDD, WWW, AAAI and more. For the training set we narrowed the temporal span to the 1999-2008 years and chose year 2009 as test set.
- **IMDb.** We extracted a collaboration network of the actors involved in Indian movie productions. Two actor (nodes) are connected by an edge if they took part in at least one movie together in a given year: as training set we considered the years from 1999 to 2008 and the year 2009 as test set. To introduce multidimensionality we took care, for each actor-actor edge, of the genres of the movie, ending up with 25 different dimensions.

Table II reports, for each network and set considered, the number of nodes, edges, and neighbors, reporting for each of them min, max and average computed over the different dimensions, and their global values computed disregarding the multidimensional information (where "gl. avg" is the average degree).

[1]http://dblp.uni-trier.de
[2]http://www.imdb.com

Dataset	$\|V\|$				$\|E\|$				Neighbors			
	min	max	avg	global	min	max	avg	global	min	max	avg	gl. avg
DBLP train.	378	3,891	1,718.3	33,329	560	7,792	3,418.4	95,727	14	77	35.5	5.07
DBLP test	26	1,126	404.8	8,507	13	1,963	672.9	17,496	1	24	12.9	3.87
IMDb train.	9	9,219	1,581.4	12,146	36	310,811	39,568.3	989,208	8	885	228.1	62.84
IMDb test	3	2,181	354.1	2,844	3	36,658	4676,4	116,910	2	161	61.7	31.43

Table II
BASIC STATISTICS FOR OUR NETWORKS

B. Evaluation of the results

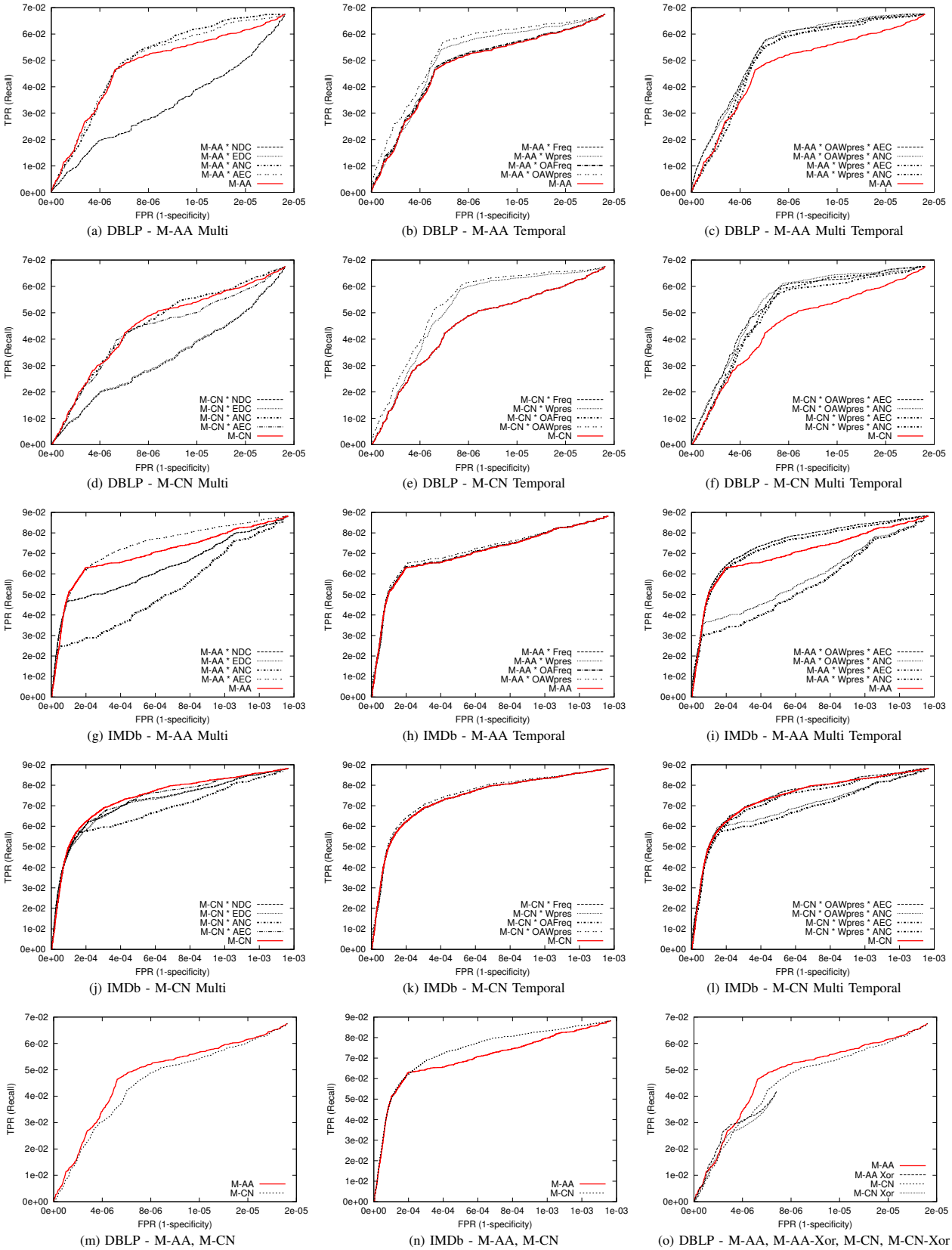
We now want to give a quantitative evaluation of the results. We measure how well M-CN, M-AA and their XOR versions perform, how the two versions of them compare, how much their predictive power can be improved by multidimensional or temporal information, and we want to see if there are global predictors that globally outperform all the others.

We applied all the scoring functions as reported in Table I to our networks. In figure 3 we report the ROC curves computed on a selection of results (due to lack of space, we are not able to report all of them). Figure 3 reports in the first two rows the ROC curves computed in DBLP by using M-CN and M-AA, multiplied by multidimensional information (first column), temporal information (second column) or both of them (last column). The second row report the same, for IMDb. The last row of the figure report different sets of plots. In the first two column of Figure 3 we report the comparison between the M-AA and M-CN on DBLP and IMDb, respectively, while in the third we group all the four multidimensional base predictors on the DBLP network. In Figure 4, we report the comparison between all the four multidimensional base predictors on IMDb, and two examples of the performances given by the predictors based on $Neighbors_{XOR}$.

First, by comparing figures 3 and 4, we report a negative result: the XOR variant of the Neighbors function is destroying part of the information about the neirghbors. This can be seen by the scale on the y axis of all the plots in Figure 4 (in Figure 4, due to lack of space, we report only the best results obtained by the XOR versions). Due to the definition of our measures, the XOR is reducing the number of total predictions issued. This is very clear from the plots in figures 3(o) and 4(a), that compare the normal and XOR versions of the basic predictors for both the networks.

Second, the temporal scores used as multiplier for the base predictors are able to restore (note: not in terms of number of predictions, but in terms of precision of them) part of the predictive power lost with the XOR, which can be additionally recovered by multiplying also by the multidimensional measures. However, the XOR based prediction is globally very poor compared to the normal one.

Third, consider the plots in figures 3(m) and 3(n). Here we report the comparison between M-AA and M-CN for the two networks. As we see, while in DBLP the global better performance of the Adamic-Adar predictor is validated, this is not true for IMDb, where M-CN is performing better. A



(a) DBLP - M-AA Multi (b) DBLP - M-AA Temporal (c) DBLP - M-AA Multi Temporal

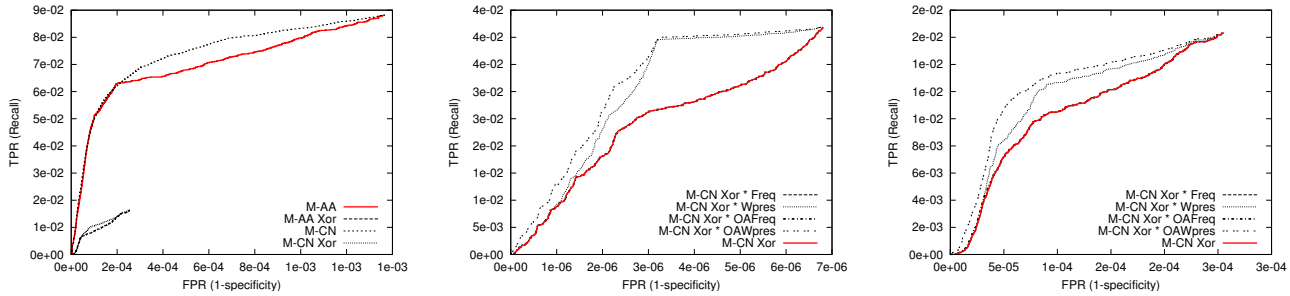
(d) DBLP - M-CN Multi (e) DBLP - M-CN Temporal (f) DBLP - M-CN Multi Temporal

(g) IMDb - M-AA Multi (h) IMDb - M-AA Temporal (i) IMDb - M-AA Multi Temporal

(j) IMDb - M-CN Multi (k) IMDb - M-CN Temporal (l) IMDb - M-CN Multi Temporal

(m) DBLP - M-AA, M-CN (n) IMDb - M-AA, M-CN (o) DBLP - M-AA, M-AA-Xor, M-CN, M-CN-Xor

Figure 3. ROC curves computed on the predictors based on Neighbors



(a) IMDb - M-AA, M-AA-Xor, M-CN, M-CN-Xor (b) DBLP - M-CN-Xor Temporal (c) IMDb - M-CN-Xor Temporal

Figure 4. ROC curves computed on the predictors based on Neighbors$_{XOR}$

possible explanation of this might be found in the structure of the two networks. In addition to the global more dense structure of IMDb, we also note that this network tends to have more cliques, that are also larger, w.r.t. DBLP, given that one movie usually joins together more persons than one scientific paper. In this scenario, the prevalence of the Adamic-Adar intuition (more importance to rarer neighbors) over the Common Neighbors (higher score to nodes with more neighbors in common) seems to lose its strength. In turns, M-CN for IMDb seems to be difficult to boost by means of multidimensional or temporal information, as we can see in the fourth row of Figure 3, that, on the other hand, destroy part of the predictive power of M-CN.

Next, consider the first two rows of Figure 3. In DBLP, the predictive power of both M-CN and M-AA can be boosted by adding multidimensional or temporal information, with an even more powerful conjunction (see (c) and (f)).

Regarding which multidimensional or temporal measures are able to help the prediction, we see that: for the first, ANC and AEC globally tend to add predictive power (especially ANC), while NDC and EDC globally lower the precision; for the second, it is clear from all the plots that the weighted version of all the measures is more accurate in capturing the temporal information, and that the OverAll versions of the measures behave better than the normal ones.

Globally speaking, the best predictors in the networks used result to be the conjunction of OAWpres, ANC, and one of M-CN and M-AA.

C. Comparison with the random predictor

In analogy with previous works [8], we also compare the performance of our predictors with a random one, used as baseline. The performance is given by the precision of the predictive model, and for the random predictor it can be calculated as:

$$Precision_{random} = \frac{|E_{test}|}{|L| \times \frac{|V|(|V|-1)}{2}}.$$

Given that all the introduced predictors are based on common neighborhood, they all output the same set of predicted links, even if the scores might be different. For this reason, we can calculate a single value of performance for all the proposed predictors:

$$Precision = \frac{TruePositive}{TruePositive+FalsePositive}$$

The boost of the performance for the networks analyzed, computed as the ratio $\frac{Precision}{Precision_{random}}$, was 35,150.4 for DBLP, and 7.1 for IMDb. As we see, the gain is much higher for DBLP, which is a much sparser network (see Table II).

D. Scalability

Last open question is how much scalable is our approach. In order to answer it, we built a few network with different node and edge sizes. In particular, we took the training set of IMDb (the largest network), and produced 5 different subnetworks. We started taking the nodes and edges belonging to only 5 dimensions, producing a first small network, then we added 5 dimensions (with the corresponding nodes and edges) at each step, until the entire network was added.Table III reports the basic statistics (only number of nodes and edges) of the subnetworks produced in this way. We decided to divide the 25 dimensions in such a way that the number of edges would have increased (almost) uniformly. On these networks, we computed all the measures, the predictors, and the aggregations as reported above. Figure 5 reports the running times (in minutes) for the experiments. Since we had many aggregations, instead of reporting the total computing time, we split it into four steps: computing the multidimensional measures (first bar in every block of four); computing the multidimensional base predictors M-CN, M-CN$_{XOR}$, M-AA and M-AA$_{XOR}$ (second bar); computing all the aggregations (third bar); and computing the temporal measures. As we see the running time grows linearly with the number of edges, with a maximum time of 30 minutes.

According to their definition, and to this empirical evaluation, our proposed predictors are scalable, and the required computing time grows linearly with the number of edges.

Dimensions	$\|V\|$	$\|E\|$
5	9,927	378,675
10	10,987	563,497
15	11,573	711,097
20	11,716	843,506
25	12,146	989,208

Table III
BASIC STATISTICS OF DIFFERENT SUBNETWORKS OF IMDB

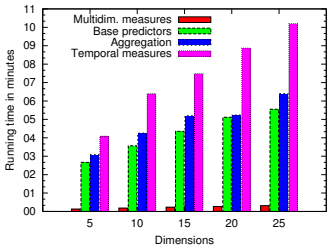


Figure 5. Running times on different IMDB subnetworks.

V. Related work

Many papers studied the problem of Link Prediction, trying both supervised and unsupervised approaches. Among the latter, [6] presented a solution based on the preferential attachment principle, while [4] and [5] introduced models based on the quantitative characteristics of common neighbors. A survey on unsupervised approaches to LP is [8], in which the authors empirically compare many different models. Two supervised approaches are the ones proposed in [7] and [9], where the first one allows also for the prediction of new nodes. In [10] the authors presented a link prediction framework that uses multiple data sources, while [11] proposed an analysis through the use of some graph proximity measure and weight of the existing links. In [12] the authors introduced a semi-supervised learning model for the link prediction problem in multi-relational networks. Like multidimensional networks, multi-relational ones allow different types of interactions between each pair of nodes. However, this model does not allow for multiple simultaneous interactions between two nodes.

Other authors have addressed different problems regarding multidimensional networks. The authors in [13] analysed the degree distributions of the various dimensions, highlighting the need for analytical tools for the multidimensional study of hubs. The authors of [14] introduced a framework for the analysis of multidimensional networks, defining a large set of measures capturing the interplay of the dimensions both at the global and at the local level.

However, to the best of our knowledge, the literature still lacks a definition of the LP problem in multidimensional networks, together with possible solving approaches. In this work, we overcome to this, by defining MLP and several classes of predictors.

VI. Conclusions

We have formulated the Multidimensional Link Prediction problem, and introduced different classes of scalable predictors aiming at capturing the underlying model of node interactions, the multidimensional information and the complete temporal history of a link in the network. We have shown that it is possible to predict new links in multidimensional networks, and our results confirm the literature of monodimensional link prediction: although unsupervised models such as the Adamic-Adar or the Common Neighbors have an high influence in the evolution of a network, their accuracy as predictors may be boosted by the introduction of supervised models (multidimensional and temporal measures) to combine with them, as weaker signals of evolution. We have supported our theory with empirical evaluation on large, real world networks, on which we have also confirmed the scalability of the proposed approach.

References

[1] J. Leskovec, J. M. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *KDD*. ACM, 2005, pp. 177–187.

[2] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu, "Graphscope: parameter-free mining of large time-evolving graphs," in *KDD*, P. Berkhin, R. Caruana, and X. Wu, Eds. ACM, 2007, pp. 687–696.

[3] C. W.-k. Leung, E.-P. Lim, D. Lo, and J. Weng, "Mining interesting link formation rules in social networks," ser. CIKM '10. New York, USA: ACM, 2010, pp. 209–218.

[4] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.

[5] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *PHYS.REV.E*, vol. 64, p. 025102, 2001.

[6] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles, "Winners don't take all: Characterizing the competition for links on the web," *PNAS*, vol. 99, no. 8, pp. 5207–5211, Apr. 2002.

[7] B. Bringmann, M. Berlingerio, F. Bonchi, and A. Gionis, "Learning and predicting the evolution of social networks," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 26–35, 2010.

[8] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," ser. CIKM '03. New York, NY, USA: ACM, 2003, pp. 556–559.

[9] M. Bilgic, G. M. Namata, and L. Getoor, "Combining collective classification and link prediction," ser. ICDMW '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 381–386.

[10] Z. Lu, B. Savas, W. Tang, and I. S. Dhillon, "Supervised link prediction using multiple sources," in *ICDM*. IEEE Computer Society, 2010, pp. 923–928.

[11] T. Murata and S. Moriyasu, "Link prediction of social networks based on weighted proximity measures," in *Web Intelligence*. IEEE Computer Society, 2007, pp. 85–88.

[12] H. Kashima, T. Kato, Y. Yamanishi, M. Sugiyama, and K. Tsuda, "Link propagation: A fast semi-supervised learning algorithm for link prediction," in *SDM*. SIAM, 2009, pp. 1099–1110.

[13] M. Szell, R. Lambiotte, and S. Thurner, "Trade, conflict and sentiments: Multi-relational organization of large-scale social networks," *arXiv.org, 1003.5137*, 2010.

[14] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, and D. Pedreschi, "Foundations of multidimensional network analysis," in *ASONAM*. IEEE Computer Society, 2011, pp. 485–489.