KDD Laboratory, ISTI-CNR & University of Pisa

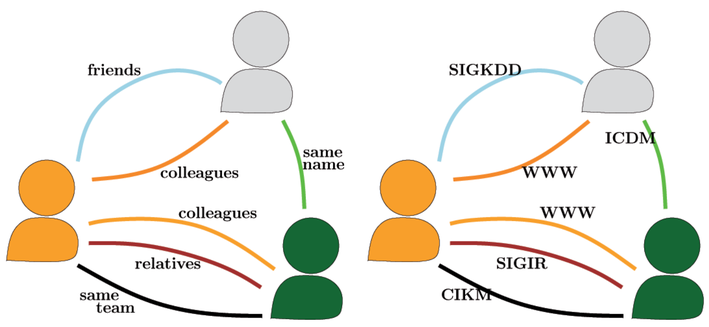
Scalable Link Prediction on Multidimensional Networks

Giulio Rossetti, Michele Berlingerio and Fosca Giannotti

Multidimensional Networks

Real world networks are often multidimensional: two nodes may be connected by more than one relation, that we call *dimensions*, expressing:

- **different types** of relationship
  - friends, colleagues, relatives
- or **different quantitative values** of the same kind of relationship
  - different ranks
  - different publication venues



Existing problems need to be reformulated.

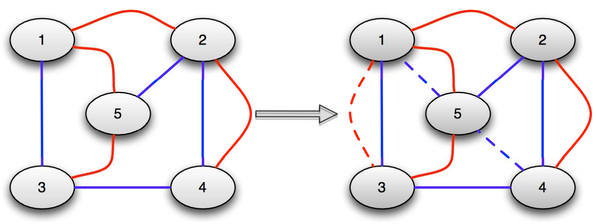
For instance, we can decide to take into account

- the relevance of a specific dimension.

Link Prediction

Problem Definition

Given a multidimensional network modeled as a multigraph $\mathcal{G} = (V, E, L, T, \tau)$[1], the Multidimensional Link Prediction problem 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 choose to extend well-known *unsupervised* predictors:

- *Common Neighbors*:

$$CN = |\Gamma(u) \cup \Gamma(u)|$$

- *Adamic Adar*:

$$AA = \sum_{z \in \Gamma(u) \cup \Gamma(u)} \frac{1}{\log|\Gamma(z)|}$$

We build our baseline modifying those predictors in order to take care of multiple dimensions. As a second step we introduce new measures that estimate the relevance of each dimension as well as information relating to the frequency of interaction between each pair of nodes.

Multidimensional Baselines

We execute the chosen link prediction algorithms separately on each dimension of the graph: in this way each predicted edge keep information of the dimension in which is likely to appear.
The resulting predictors are our baselines:

Multidimensional Adamic Adar (**M-AA**) and Multidimensional Common Neighbors (**M-CN**).

Dimension relevance

We define four measure in order to evaluate the importance of each dimension:

- **Node (Edge) Dimension Connectivity (NDC-EDC):**
  Let $d \in L$ be a dimension of a network $\mathcal{G}$. The functions $NDC : L \rightarrow [0, 1]$ and $EDC : L \rightarrow [0, 1]$ defined as

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

  computes the ratio of nodes (or edges) of the network that belong to the dimension d.

- **Average Node (Edge) Correlation (ANC-AEC):** Let $d \in L$ be a dimension of a network $\mathcal{G}$. The functions $ANC : L \rightarrow [1/|L|, 1]$ and $AEC : L \rightarrow [1/|L|, 1]$ is defined as

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

  where $(N|E)Jaccard(d, d')$ is the Jaccard correlation index on the node (edge) sets. It computes the average node (edge) correlation of a dimension with all the others.

Temporal Information

Besides the analysis of the multidimensional structure we also want to take into account the complete temporal history of an edge of the network.

- **Frequency (Freq-OAFreq):**
  Let $(u, v, d) \in E$ be an edge of a network $\mathcal{G}$. The two functions $Freq : E \rightarrow [1, |T|]$ and $OAFreq : V \times V \rightarrow [1, |L| \times |T|]$ defined as

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

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

- **Weighted Presence (WPres-OAWpres):**
  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.
  Let $(u, v, d) \in E$ be an edge of a network $\mathcal{G}$. The functions $WPres : E \rightarrow [1, +\infty]$ and $OAWPres : V \times V \rightarrow [1, +\infty]$ defined as

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

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

Combinations

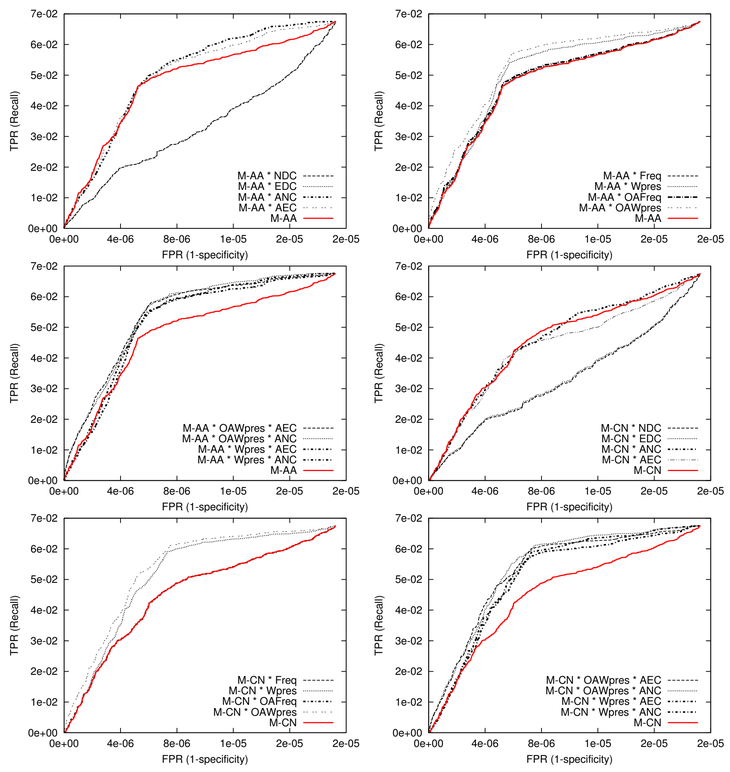
Baselines, multidimensional and temporal measures represent different kinds of "*informative waves*" so they can be used in combination in order to obtain better performances.
We tested the multidimensional baselines in conjunction with the following combination of our measures:

Baseline	Multidim.	Temporal
M-AA\M-CN	NDC	
M-AA\M-CN	EDC	
M-AA\M-CN	AEC	
M-AA\M-CN	ANC	
M-AA\M-CN		Freq
M-AA\M-CN		OAFreq
M-AA\M-CN		WPres
M-AA\M-CN		OAWpres

Baseline	Multidim.	Temporal
M-AA\M-CN	AEC	WPres
M-AA\M-CN	AEC	OAWPres
M-AA\M-CN	ANC	WPres
M-AA\M-CN	ANC	OAWPres
M-AA\M-CN		
M-AA\M-CN		
M-AA\M-CN		
M-AA\M-CN		

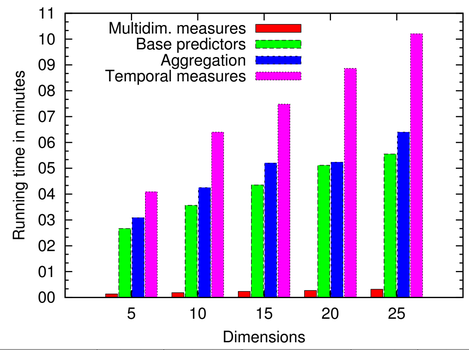
Results

Here we show the plot of ROC curves for all the combination of our measures. In red are reported the baselines.



Scalability

In order to verify the scalability of our predictors we build a few network with different node and edge sizes.



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

The figure 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.
As we see the running time grows **linearly** with the number of edges, with a maximum time of 30 minutes (in a single run were computed all the introduced aggregations).

Conclusion

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.

Rossetti, Berlingerio and Giannotti, "Scalable Link Prediction on Multidimensional Networks", ICDM Workshops 2012.

[1] V set of nodes, E set of edges, L set of dimension labels, T set of timestamps, τ function that assign to each edge $(u, v, d) \in E$ a subset of T.