Link Prediction

Lloyd Sanders, Olivia Woolley, Dirk Helbing
Computational Social Science
Overview

- What is link prediction?
- What are some examples of link prediction?
- A How To
  - The setup
  - Local Methods
  - Global Methods
- Challenges of Link Prediction
- References
Abstractly: Given a snapshot of a network, can one predict the next most likely links to form in the network?
Link Prediction in Industry

Proposing Friendships

Proposing dates

Proposing items to purchase
Link Prediction in Science

Artist’s rendition of Human Metabolic network
Link Prediction in Science

Investigating link connections in bio. networks is costly and time consuming.
Link Prediction

Link Prediction in Criminal Networks: A Tool for Criminal Intelligence Analysis

Giulia Berlusconi1, Francesco Calderoni1 *, Nicola Parolini2, Marco Verani2, Carlo Piccardi3 *

1 Universitá Cattolica del Sacro Cuore and Transcrime, Milano, Italy, 2 MOX, Department of Mathematics, Politecnico di Milano, Milano, Italy, 3 Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy

* francesco.calderoni@unicatt.it (FC); carlo.piccardi@polimi.it (CP)

Possible missing links predicted between suspects in organized crime
Link prediction is used to predict future possible links in the network (E.g., Facebook). Or, it can be used to predict missing links due to incomplete data (E.g., Food-webs)
Abstractly: Given a snapshot of a network, can one predict the next most likely links to form in the network?

Graph

\[ G = (V, E) \]

Edge

\[ e = (x, y) \]

Similarity Score

\[ S_{xy} \]
How To: The Set up

- For a given graph
- Split the data into a training set, and a validation set
- Choose a link prediction algorithm
- Run the algorithm on the training set, and test it on the validation set.
- Check the accuracy
- Compare other link prediction algorithms
How To: The Setup

Split Edges into a training and test (probe) set

\[ E = E^T \cup E^P \]

Set of all possible edges on graph

\[ |U| = \frac{|V|(|V| - 1)}{2} \]
Our algorithms consider only the edges that connect the same set of nodes in the training and probe set – usually taken as the nodes of the giant component of the graph.

This assumes the graph is static – no new nodes enter the system.

By definition: The algorithm will not predict edges for nodes not within the giant component.

More about this later.
How To: Output of algorithm

The Link Prediction algorithm will spit out a list, ranked with edges which are most likely to appear at the top, descending.

\[ L : e_L \in U - E^T \]

Taking the first \( n \) links from the list, and calculating the intersection with the probe set of length \( n \), gives a simple measure of accuracy.
Local Methods – a basic example

Links are predicted based solely on your local contact structure. The main idea is that of triangle closing.
How to: Similarity Measures

Common Neighbours

$$S^{CN}_{xy} = |\Gamma(x) \cap \Gamma(y)|$$

Counting number of paths of a certain length

Collaborative filtering is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, et al. It may work well for some networks while fails for some others. In additional, the similarities can be used in a more skilled fashion.
How to: Similarity Measures

Jaccard Index

\[ S_{xy}^{\text{Jaccard}} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \]

Resource Allocation

\[ S_{xy}^{\text{RA}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \]

Many more similarity measures available
Given the adjacency matrix, one can take the global structure of the graph into account when making predictions.

These are often based on number of shortest path measures.

Common neighbors is simply the Adj. matrix squared.

\[ S_{xy}^{CN} = (A^2)_{xy} \]

Try it for yourself to verify it!
Global Methods

- What about paths of greater length? How can we include those? We can weight longer paths less.

\[
\exp(\alpha A) = 1 + \alpha A + \frac{(\alpha^2 A^2)}{2!} + \frac{(\alpha^3 A^3)}{3!} + \cdots
\]

\[
s_{xy} = \exp(\alpha A)\big|_{xy} = \sum_{i=0}^{\infty} \frac{\alpha^i}{i!} A^i\big|_{xy}
\]
Link Prediction on Bipartite Graphs
Link Prediction on Bipartite Graphs

Consider an odd function for the graph kernel

\[ A^3 \]

Or something more sophisticated

\[
\sinh(\alpha A) = \alpha A + \frac{(\alpha A)^3}{3!} + \frac{(\alpha A)^5}{5!} + \cdots
\]

\[
\sinh(\alpha A) = \sum_{i=0}^{\infty} \frac{\alpha^{1+2i}}{(1+2i)!} A^{1+2i}
\]
Your choice of link prediction algorithm makes an implicit assumption of the graph kernel – the mechanism for how the graph grows. Different graph types (social, academic, user-item) grow under different mechanisms. Therefore different link prediction algorithms will work better on other graphs.
Challenges for Link Prediction

- Cold Start Problem
- Temporal Networks
  - Links can be created and destroyed over time
  - New nodes come in and out of the system
- Link prediction with links of a different nature – e.g., ‘negative’ links.
- Eliminating statistical bias in partitioning data – k-fold cross validation.
The Cold Start Problem
The Cold Start Problem

Propose links due to node characteristics
The Cold Start Problem – new nodes
References

- The Link Prediction Problem for Social Networks; Liben-Nowell & Kleinberg
- Link prediction in complex networks: A survey; Lü and Zhou, Physica A; 2011
- Spectral evolution in dynamic networks; Kunegis et al.; Knowl. Inf. Syst.; 2013
- Online Dating Recommender Systems: The split-complex number approach; Kunegis et al.; ACM 2012
- Mean Average Precision
  - https://en.wikipedia.org/wiki/Precision_and_recall
  - Victor Lavrenko: youtube.com/watch?v=pM6DJ0ZZee0