DATA ANALYSIS II

Link Prediction Techniques

2021/22

Outline

- Link prediction problem
- Main approaches
- Local similarity-based techniques

Problem setting

- Link prediction is *prediction of social connections between users…*
- Generally, the prediction problem is studied from two perspectives:
    - network structure,
    - attributes of nodes and connections.
- Structure refers to the way in which nodes that compose the network are interconnected; it reflects the information about network topology.

Link prediction based on attribute information

- The attribute information refers to description of the features of nodes and/or edges.

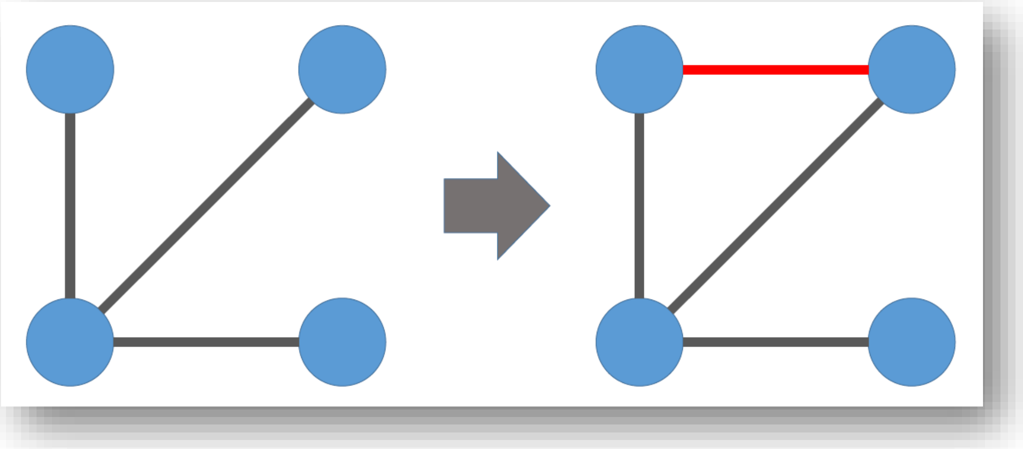
- Attribute-based prediction methods follow a machine learning approaches.

- Methods include e.g. decision trees, support vector machine(SVM), Naïve Bayes, etc.

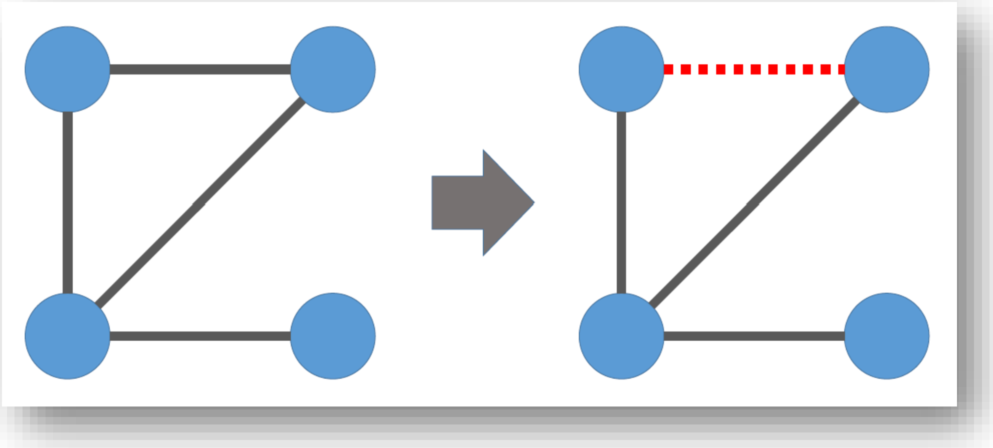
- Results show that the performance of link prediction improves when machine learning approaches are used; however, this additional network information is not always available.

Formally

Adding links



Removing links



Let *G(V, L)* be a network within the time period of *G[t, t1]* where *V* represents the set of nodes and L represents the set of links. For the next time period *G(t1, t2]*, the network might change. The link prediction focuses on how to predict the evolution of links, i.e. how *L[t, t1]* will differ from *L(t1, t2]*.

- Adding Links
  - …means that in the next time window a new link will be created between existing nodes. There can be one or more newly created links.
  - This problem is preferred among link prediction techniques (also in this presentation).

- Removing Links
  - …means that the link will disappear in the next time window. Similar to the situation when new links are added, one or more link can be removed in one time step.

- Adding and Removing Links
  - …is about the combination of two previously described problems. It means that from one time window to another both appearance and disappearance of links can be predicted.

Applications of Link Prediction

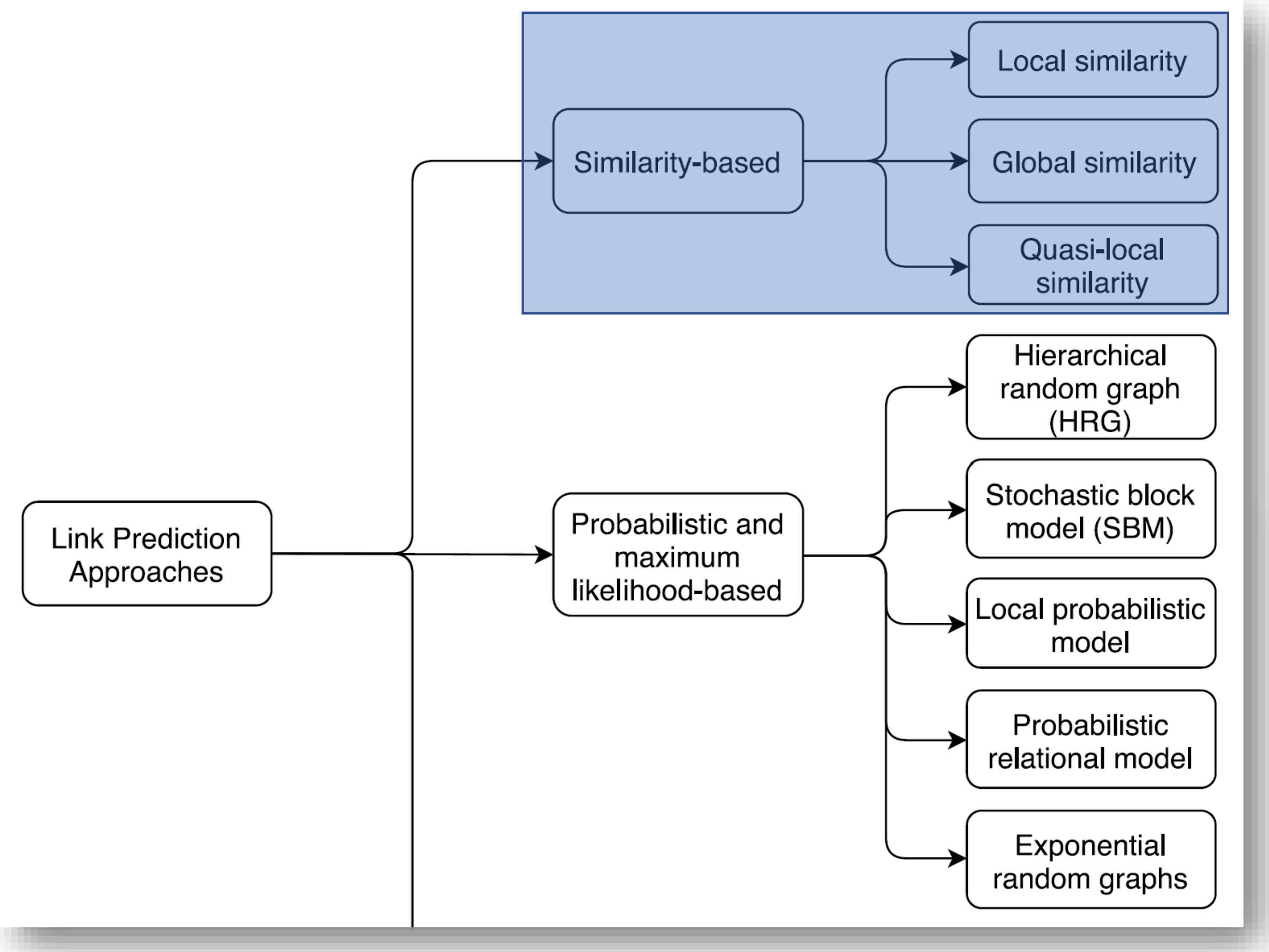
- Improving similar users’ selection in recommender systems that follow a collaborative approach. E.g., most social networks use link prediction techniques to automatically suggest friends.

- Methods to find possible interactions between pairs of proteins in a protein-protein interaction network (PPI network).

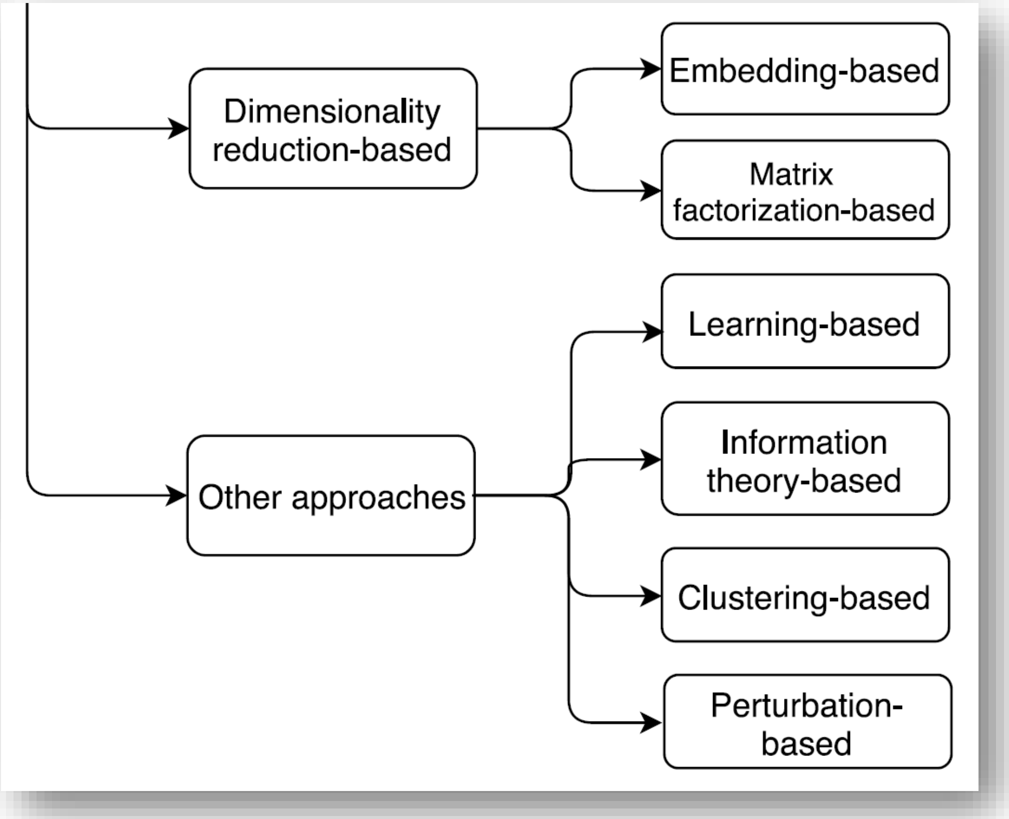
- Collaboration prediction in scientific co-authorship networks. It is about to better understand how some research fields will evolve by predicting which authors or groups could potentially collaborate in the future.

- Analysis of the structure of criminal and terrorist networks in order to fight against organized crime; link prediction can reveal non-observed links in criminal networks, allowing us to anticipate certain criminal actions.

- Link prediction can be used to analyze how tendencies spread across society. It was shown how link prediction techniques can be used in viral marketing in order to achieve better marketing plans


Link Prediction Approaches
Similarity-based
Local similarity
Global similarity
Quasi-local similarity
Probabilistic and maximum likelihood-based
Hierarchical random graph (HRG)
Stochastic block model (SBM)
Local probabilistic model
Probabilistic relational model
Exponential random graphs

Four groups of recent approaches include many techniques with different performances.



Similarity-based approaches

- Similarity-based metrics are the simplest ones in link prediction, in which for each pair x and y, a similarity score *S(x, y)* is calculated.

- The score *S(x, y)* is based on the structural properties of the considered pair. The pair of nodes having a higher score (above some threshold) represents the predicted link between them.

- The similarity measures between every pair can be calculated using several properties of the network; *we assume only structural properties.*

Local similarity indices

- Local indices are generally calculated using information about common neighbors and node degree.

- These indices consider immediate neighbors of a node. Examples of such indices contains common neighbor, preferential attachment, Adamic/Adar, resource allocation, etc.

Common Neighbors (CN)

- In a given network or graph, the size of common neighbors for a given pair of nodes *x* and *y* is calculated as the size of the intersection of the two nodes neighborhoods.

- $S(x, y) = |\Gamma(x) \cap \Gamma(y)|$, where $\Gamma(x)$ and $\Gamma(y)$ are neighbors of the node *x* and *y* respectively.

- The likelihood of the existence of a link between *x* and *y* increases with the number of common neighbors between them.

- It has been observed that the common neighbor approach performs well on most real-world networks and beats other, often very complex, methods.

Jaccard Coefficient

- This metric is similar to the common neighbor. Additionally, it normalizes the above score, as given below.

$$S(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}.$$

- The Jaccard coefficient is defined as the probability of selection of common neighbors of pairwise vertices from all the neighbors of either vertex. The pairwise Jaccard score increases with the number of common neighbors between the two vertices considered.

- It was shown that this similarity metric performs worse as compared to Common Neighbors.

Adamic/Adar Index (AA)

- This metric was proposed to calculate a similarity score between two web pages based on shared features.

$$S(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z},$$

- where k_z is the degree of the node z (a common neighbor).

- More weights are assigned to the common neighbors having smaller degrees. This is also intuitive in the real-world scenario, for example, a person with more number of friends spend less time/resource with an individual friend as compared to the less number of friends.

Preferential Attachment

- Preferential attachment score between two nodes x and y can be computed as follows.

$$S(x, y) = k_x . k_y .$$

- This index shows the worst performance on most networks. The simplicity (as it requires the least information for the score calculation) and the computational time of this metric are the main advantages.

- It requires only degree as information and not the common neighbors. In assortative networks, the performance of the PA improves, while very bad for disassortative networks.

Resource Allocation Index (RA)

- Suppose node x sends some resources to y through the common nodes of both x and y then the similarity between the two vertices is computed in terms of resources sent from x to y. This is expressed mathematically as:

$$S(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}.$$

- This similarity measure and the Adamic/Adar are very similar to each other. The difference is that the RA index heavily penalizes to higher degree nodes compared to the AA index.

Cosine similarity or Salton Index

- The Cosine similarity can be computed as:

$$S(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{(k_x . k_y)}}.$$

Sorensen Index

- As can be observed, it is very similar to the Jaccard index.

$$S(x, y) = \frac{2|\Gamma(x) \cap \Gamma(y)|}{k_x + k_y}.$$

- It was shown that it is more robust than Jaccard against the outliers.

CAR-based Common Neighbor Index (CAR)

- CAR-based indices are presented based on the assumption that the link existence between two nodes is more likely if their common neighbors are members of a local community.

- The likelihood existence increases with the number of links among the common neighbors (local community links LCL).

$$S(x,y) = CN(x,y) \times LCL(x,y) \quad = CN(x,y) \times \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{|\gamma(z)|}{2},$$

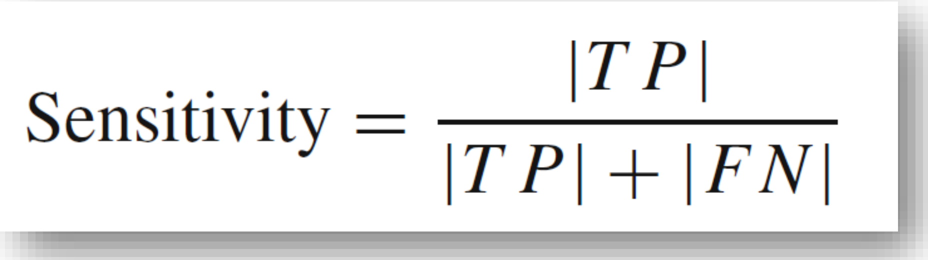
- where $CN(x, y) = |\Gamma(x) \cap \Gamma(y)|$ is number of common neighbors $LCL(x, y)$ refers to the number of local community links which are defined as the links among the common neighbors of nodes x and y. The $\gamma(z)$ is the subset of neighbors of node z that are also common neighbors of x and y.

Link Prediction Performance

- We can transform the link prediction task into binary (link x no-link) classification.
  - **True Positive (TP):** The positive data item (Link Available) predicted as positive (Predicted).
  - **True Negative (TN):** The negative data item (Link Not Available) predicted as negative (Not Predicted).
  - **False Positive (FP):** The negative data item (Link Not Available) predicted as positive (Predicted).
  - **False Negative (FN):** The positive data item (Link Available) predicted as negative (Not Predicted).

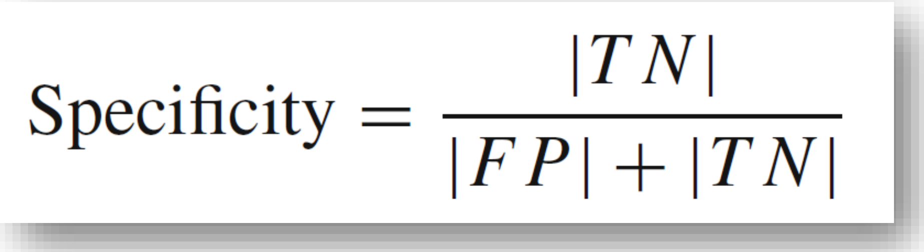
Performance Measures 1/3

- Sensitivity/true positive rate:



$$\text{Sensitivity} = \frac{|TP|}{|TP| + |FN|}$$

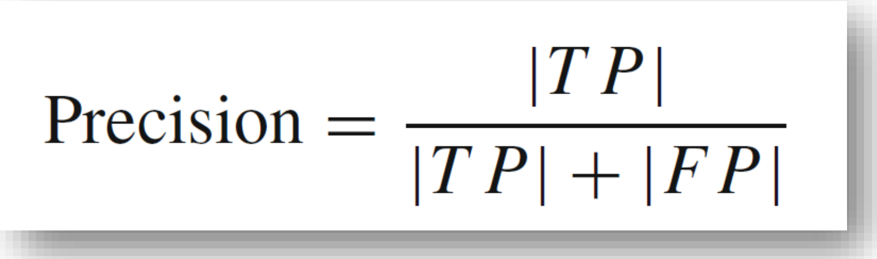
- Specificity/true negative rate:



$$\text{Specificity} = \frac{|TN|}{|FP| + |TN|}$$

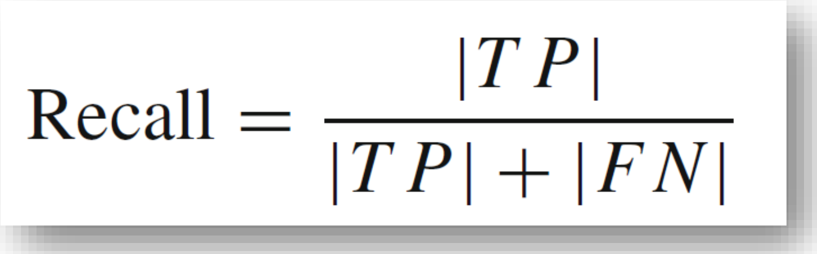
Performance Measures 2/3

- Precision:



$$\text{Precision} = \frac{|TP|}{|TP| + |FP|}$$

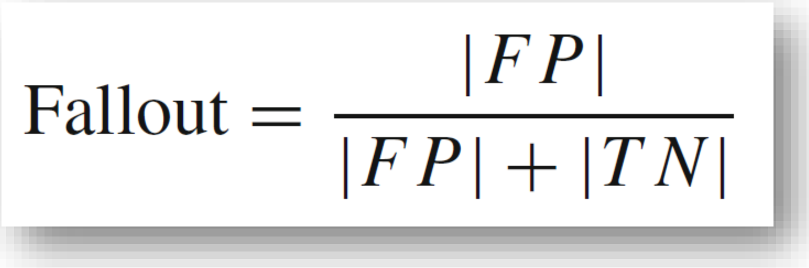
- Recall:



$$\text{Recall} = \frac{|TP|}{|TP| + |FN|}$$

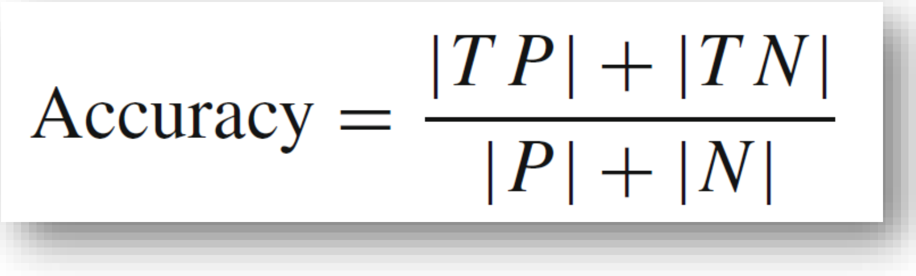
Performance Measures 3/3

- Fallout/false-positive rate:



$$\text{Fallout} = \frac{|FP|}{|FP| + |TN|}$$

- Accuracy:



$$\text{Accuracy} = \frac{|TP| + |TN|}{|P| + |N|}$$

Summary

- Although many link prediction methods have been explored in the literature, it is still an open research problem.

- Several problems are yet to be explored, for example, which structural properties perform better on each technique, also how to deal with the large size.

- Methods based on similarity and analysis of common neighbors have good computational properties (low complexity) and, despite their simplicity, are very powerful.

Seminar Assignments

- Implement at least two link prediction methods and apply them to at least two networks (e.g., Karate Club, Les Misérables, Dolphins). Compute all listed performance measures for all methods and networks.

References

- Gao, F., Musial, K., Cooper, C., Tsoka, S. (2015). Link prediction methods and their accuracy for different social networks and network metrics. *Scientific programming*, 2015.

- Martínez, V., Berzal, F., Cubero, J. C. (2016). A survey of link prediction in complex networks. *ACM Computing Surveys* (CSUR), 49(4), 1-33.

- Kumar, A., Singh, S. S., Singh, K., Biswas, B. (2020). Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, 124289.