The 14th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty

Evidential *k*-NN for Link Prediction

Sabrine Mallek[†‡], Imen Boukhris[†], Zied Elouedi[†], Eric Lefèvre[‡]

[†]Université de Tunis, LARODEC
Institut Supérieur de Gestion de Tunis, Tunisia
[‡]Université d'Artois, LGI2A, Béthune, France

Lugano

July 2017

LARODEC
Laboratoire de Recherche Opérationnelle de Décision et de Contrôle de Processus

LGI2A

Outline

Outline

Motivation

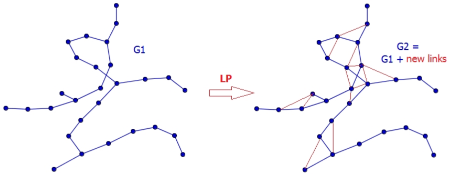
- **Link prediction** (LP) is an important scientific issue in SNA that studies network dynamics and evolving.
- Social network (SN) **data** are prone to **observation errors**, they are usually **noisy** and **missing**.
- Supervised machine learning techniques have been intensively applied to LP. However, most covered algorithms lack functionality to properly manipulate and deal with noisy and **imperfect SN data**.

Objective and hypothesis

1. Propose a new approach for supervised link prediction that handles social network data imperfection.
2. Handle uncertainty via the belief function theory framework.
3. Improve classification accuracy by integrating topological information of the network.

Link prediction

Link prediction addresses the problem of predicting the existence of new/missing relations.

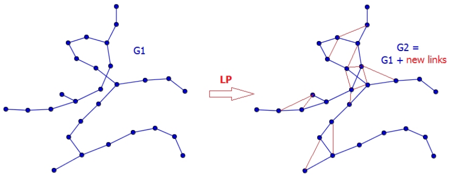


Applications:

- Infer new relations to be formed in the future
- Expose links that already exist but are not apparent
- Assist users to make new connections

Link prediction

Link prediction addresses the problem of predicting the existence of new/missing relations.



Most methods compute similarity scores of node-neighborhoods based on network topology.
Popular measures: Common neighbors, Adamic-Adar, Jaccard Coefficient, Ressource allocation, Preferential attachment.

Belief function theory (BFT)

- A general framework for reasoning with uncertainty

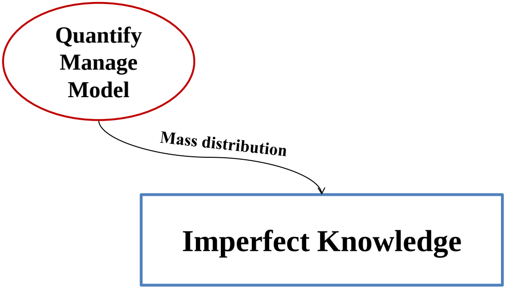
Belief function theory (BFT)

- A general framework for reasoning with uncertainty

Imperfect Knowledge

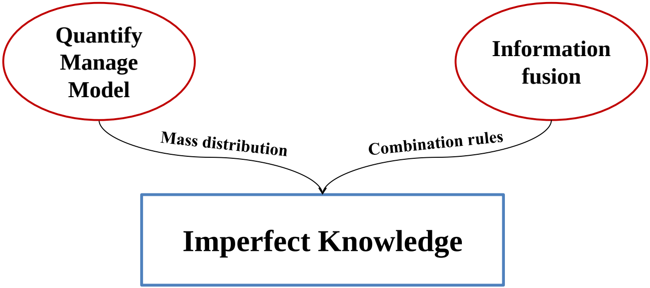
Belief function theory (BFT)

- A general framework for reasoning with uncertainty



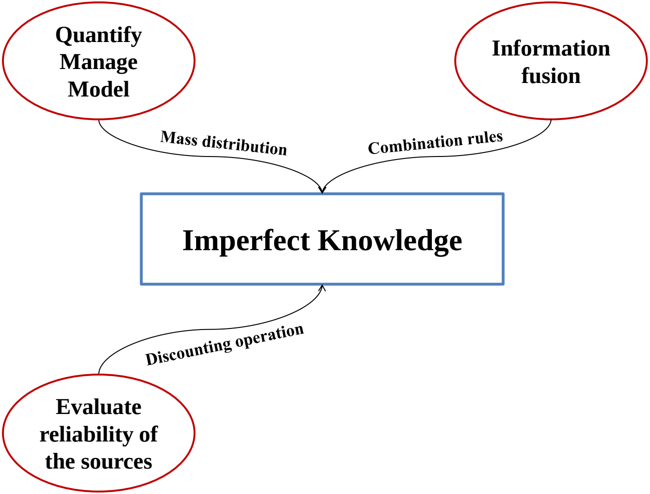
Belief function theory (BFT)

- A general framework for reasoning with uncertainty



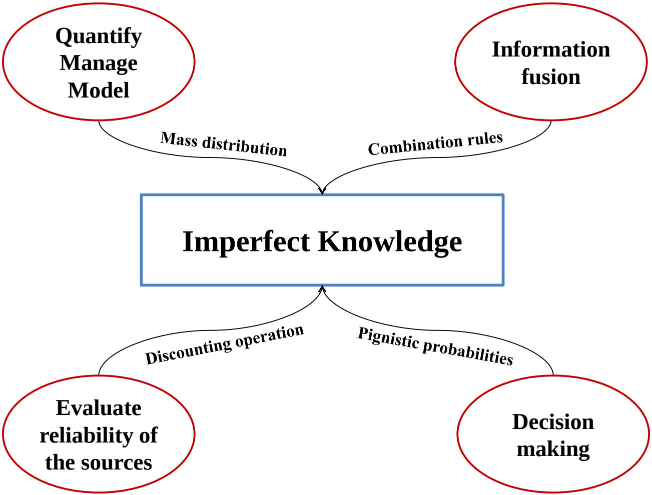
Belief function theory (BFT)

- A general framework for reasoning with uncertainty



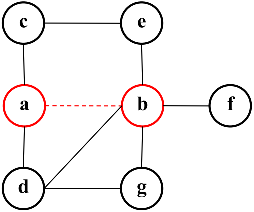
Belief function theory (BFT)

- A general framework for reasoning with uncertainty



Outline

Detecting *k*-nearest neighbors

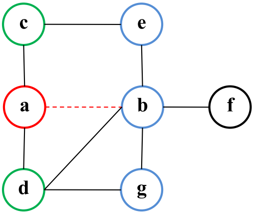


Subgraph of a social network

Assume that **ab** is the query link

The first step is to detect the neighborhood of ab

Detecting *k*-nearest neighbors



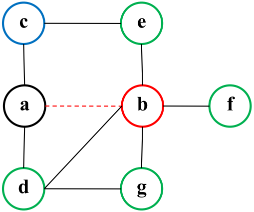
Subgraph of a social network

-Query link: **ab**

-First level neighbors of a={c,d}

-Second level neighbors of a={e,g,b}

Detecting *k*-nearest neighbors



Subgraph of a social network

-Query link: **ab**

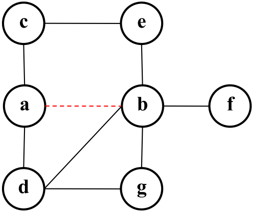
-First level neighbors of a={c,d}

-Second level neighbors of a={e,g,b}

-First level neighbors of b={e,d,g,f}

-Second level neighbors of b={c}

Detecting *k*-nearest neighbors



Subgraph of a social network

-Query link: **ab**

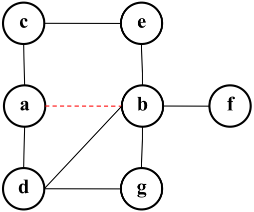
-First level neighbors of a={c,d}
-Second level neighbors of a={e,g,b}
-First level neighbors of b={e,d,g,f}
-Second level neighbors of b={c}

Neighboring links:

-Links shared with first level neighbors of **a** and first level neighbors of **b** are **N^1={ac, ad, be, bg, bd, bf}**

-Links unshared with second level neighbors of **a** and **b** are **N^2={ae, ag, bc}\{ab}**

Detecting *k*-nearest neighbors



Subgraph of a social network

-Query link: **ab**

-N^1={ac, ad, be, bg, bd, bf}

-N^2={ae, ag, bc}

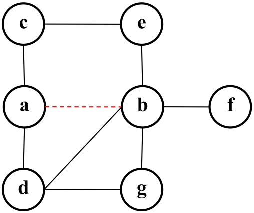
-Let Ω={exist, not exist} be the set of classes

-The classes of the links in N^1 are exist

-The classes of the in N^2 are not exist

-Similarity between ab and their neighboring links in N^1 and N^2 is evaluated according to the Euclidean distance where strutural metrics are used as features.

Detecting *k*-nearest neighbors



Subgraph of a social network

-Query link: **ab**

-N^1={ac, ad, be, bg, bd, bf}

-N^2={ae, ag, bc}

-Let Ω={exist, not exist} be the set of classes

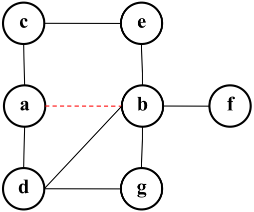
-The classes of the links in N^1 are exist

-The classes of the in N^2 are not exist

-Similarity between ab and their neighboring links in N^1 and N^2 is evaluated according to the Euclidean distance where strutural metrics are used as features.

-The *k* links with smallest distances to ab are considered as the *k*-nearest neighbors.

Information fusion and prediction



Subgraph of a social network

-Query link: **ab**

-N^1={ac, ad, be, bg, bd, bf}

-N^2={ae, ag, bc}

-Let Ω={exist, not exist} be the set of classes

-The classes of the links in N^1 are exist

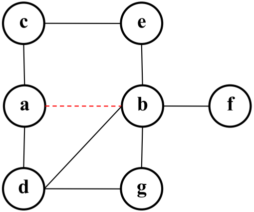
-The classes of the in N^2 are not exist

- Each nearest neighbor represents a source of information regarding the existence of the link ab.

-A mass distribution that quantifies the uncertainty regarding the existence of ab is generated from the distances values.

- The masses are constructed based on the intuition that the closer ab is to its nearest neighbor according to the distance, the more likely for ab to have the same class.

Information fusion and prediction



Subgraph of a social network

-Query link: **ab**

-N^1={ac, ad, be, bg, bd, bf}

-N^2={ae, ag, bc}

-Let Ω={exist, not exist} be the set of classes

-The classes of the links in N^1 are exist

-The classes of the in N^2 are not exist

- Masses given by all the nearest neighbors are fused using the belief function theory conjunctive rule of combination.

- Finally, decision about the membership of ab to one of the classes in Ω is made by comparing the masses on the events exist and not exist
- If the mass on the event exist is higher than the mass on the event not exist, than ab is predicted, it is not predicted otherwise.

Outline

Experimental setup

- Experiments are performed on a component of a real social network of Facebook friendships with 1K actors and 10K links
- A comparative study is made with the classical k-NN classifier
- The accuracy of 10-fold cross validation is used as evaluation by adding randomly generated false links of the same size as the subsamples at each time.
- A preprocessing phase is first conducted to compute local similarity scores of all the links to reduce computational time.
- Different values of the number of nearest neighbors k are tested ranging from 1 to 15.
- The behavior of our algorithms to class imbalance is evaluated by increasing the number of negative instances (non existing links) at each time.

Results: Parameter *k* evaluation

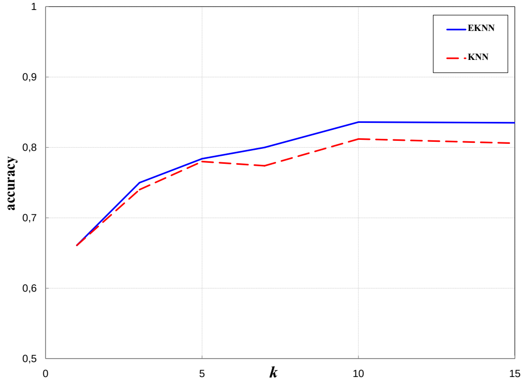


Figure: Accuracy results according to the values of k for evidential k-NN: uncertainty + network topology, and the classical k-NN: network topology, applied to Facebook dataset.

Results: Class imbalance test

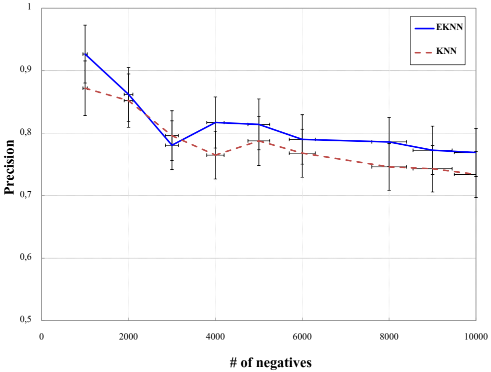


Figure: Precision results for $k = 15$ according to the increase of negative links, for the evidential k-NN: uncertainty + network topology, and the classical k-NN: network topology, applied to Facebook dataset.

Outline

- Link prediction (LP) has been used in many fields of science, as online social networks where links can be recommended as promising friendships.
- Here LP is reformulated into a binary classification problem by extending the evidential k-NN classifier to take network topological properties into account.
- Uncertainty is addressed thanks to the belief function theory tools.
- Experiments confirm the efficiency of the novel framework and show that it handles skewness in social network data.
- In future work, other information could be integrated such as node attributes to add semantics to the LP task.

Thank you