Learning Links in MeSH Co-occurrence Network

Preliminary Results

Andrej Kastrin[1] and Dimitar Hristovski[2]*

[1]Faculty of Information Studies, Novo mesto, Slovenia

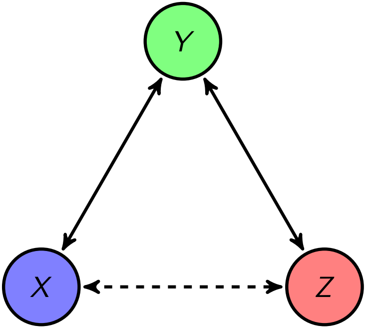
[2]Institute of Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

*Presenting author

The First International Workshop on Large-Scale Graph Storage and Management, GraphSM 2014
April 20-24, 2014
Chamonix, France

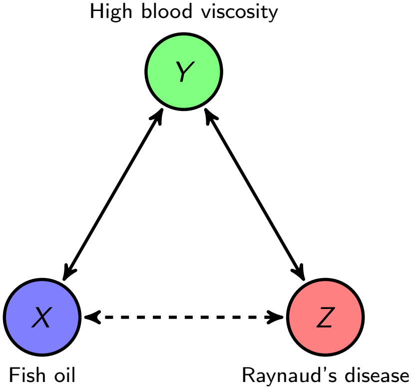
Literature-Based Discovery

- Find implicit relations between entities.
- Propose implicit relations as potential scientific hypoteses.
- Swanson's XYZ model:
  - Relations XY and YZ are known
  - Implicit relation XZ is (putative) new discovery



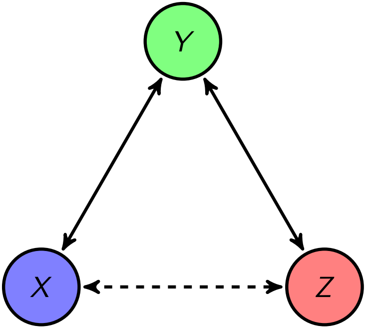
Swanson's Example

- Blood viscosity was found to co-occur with Raynaud's disease.
- Fish oil reduces blood viscosity.
- Fish oil was proposed as a new treatment for Raynaud's disease.

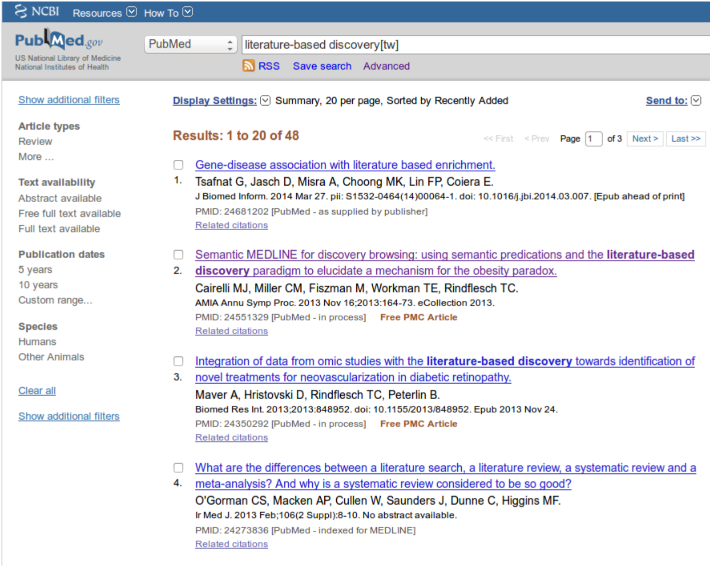


Literature-Based Discovery as Link Prediction Problem

- We can model biomedical literature as a network of biomedical concepts.
- Link prediction refers to the prediction of future links between concepts that are not directly connected in the current snapshot of a network.



MEDLINE/PubMed



www.ncbi.nlm.nih.gov/pubmed

Medical Subject Headings

- Comprehensive controlled vocabulary for indexing in the life sciences.
- The 2013 version of MeSH contains 26 853 descriptors.
- Every article in MEDLINE/PubMed is indexed with about 10-15 descriptors.
- Some descriptors are designated (*), indicating the article's major topic.

MeSH Terms in an Article

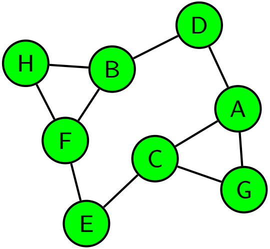
```
PMID- 20091016
TI  - Chi-square-based scoring function for...
AB  - OBJECTIVES: Text categorization has been used...
MH  - Access to Information
MH  - Algorithms
MH  - Artificial Intelligence
MH  - Bayes Theorem
MH  - *Chi-Square Distribution
MH  - Data Collection
MH  - Data Interpretation, Statistical
MH  - *Data Mining
MH  - Humans
MH  - *MEDLINE
MH  - Medical Informatics
MH  - *Natural Language Processing
```

Methods

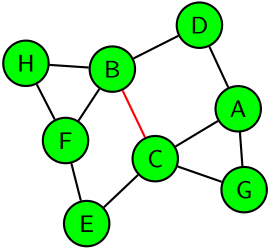
Link Prediction Framework

- We have train network $G[t_1, t_2]$ which contains interactions among nodes that take place in the time interval $[t_1, t_2]$.
- We have test network $G[t_3, t_4]$ which contains interactions among nodes that take place in the time interval $[t_3, t_4]$.
- Learning task: provide a list of edges that are present in test network, but absent in train network.

Train network



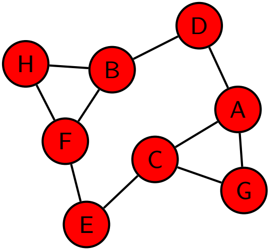
Test network



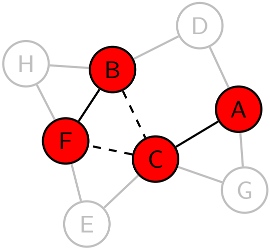
Link Prediction Setup

- Prediction and evaluation was performed on a core subnetwork.
- Core subnetwork consists of nodes with at least 3 neighbors.

Train network



Test network



Data Collection

- We constructed two networks:
  - Train network [2003-2007]
  - Test network [2008-2012]
- Networks were post-processed to remove non-informative edges.
- We applied χ^2 test for independence for each co-occurrence pair to obtain statistic, which indicates whether particular pair occurs together more often than by chance.

Similarity Measures

- For each node pair (u, v) we calculate similarity score $s(u, v)$.
- Score $s(u, v)$ gives the likelihood of link formation between nodes u and v.
- We used two similarity measures:
  - Jaccard coefficient

$$s_{uv} = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

    where $\Gamma(u)$ is set of neighbors of u

  - Adamic-Adar coefficient

$$s_{uv} = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(z)|}$$

Performance Assessment

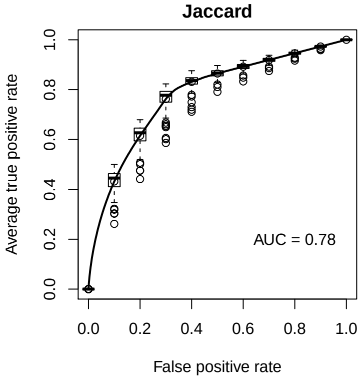
- Major challenge is huge number of possible node pairs.
- We use a bootstrap resampling approach:
  - We draw a random sample of 1000 nodes and create appropriate train and test networks.
  - We compute link prediction score $s(u, v)$ for each node pair that is not associated with any interaction before time t_3.
  - We assign class label "positive" to this node pair if the link occurs in test network and "negative" otherwise.
  - We repeat this procedure 100 times.
- Using class labels and similarity scores we constructed ROC curve.

Results

Topological Characteristics of the MeSH Networks

Parameter	*Train*	*Test*
Nodes	24 225	25 570
Edges	4 897 380	5 615 965
Edges (reduced)	3 328 288	3 810 535
Density	0.01	0.01
Mean degree	274.78	298.05
Average path length	2.23	2.20
Clustering coefficient	0.27	0.26
Small-worldness index	21.57	20.70

Prediction Performance



AUC (*Area under the ROC curve*): 0.90 – 1.00 = excellent, 0.80 – 0.90 = good, 0.70 – 0.80 = fair, 0.60 – 0.70 = poor, 0.50 – 0.60 = fail

Future Work

- Explore the role of node and edge attributes in prediction performance.
- Extend the study to semantic relations instead of co-occurrences.
- Assess prediction performance on large-scale network.
- Develop web application for real-time computing.