Link prediction

Leonid E. Zhukov

School of Data Analysis and Artificial Intelligence
Department of Computer Science
National Research University Higher School of Economics

Structural Analysis and Visualization of Networks

NATIONAL RESEARCH
UNIVERSITY

Lecture outline

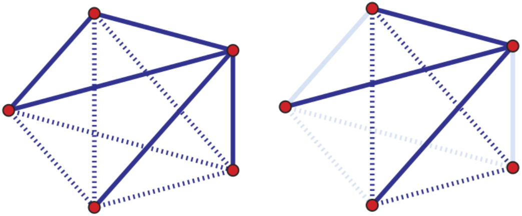
1. Link prediction problem
2. Proximity measures
3. Prediction by supervised learning
4. Other methods

Link prediction

- **Link prediction**. A network is changing over time. Given a snapshot of a network at time t, predict edges added in the interval (t, t')
- **Link completion** (missing links identification). Given a network, infer links that are consistent with the structure, but missing (find unobserved edges)
- **Link reliability**. Estimate the reliability of given links in the graph.

- Predictions: link existence, link weight, link type

Link prediction



- Graph G(V,E)
- Number of "missing edges": $|V|(|V|-1)/2 - |E|$
- In sparse graphs $|E| \ll |V|^2$, Prob. of correct random guess $O(\frac{1}{|V|^2})$

Scoring algorithm

Link prediction by proximity scoring

1. For each pair of nodes compute proximity (similarity) score $c(v_1, v_2)$
2. Sort all pairs by the decreasing score
3. Select top n pairs (or above some threshold) as new links

Scoring functions

Local neighborhood of v_i and v_j

- Number of common neighbors:

$$|\mathcal{N}(v_i) \cap \mathcal{N}(v_j)|$$

- Jaccard's coefficient:

$$\frac{|\mathcal{N}(v_i) \cap \mathcal{N}(v_j)|}{|\mathcal{N}(v_i) \cup \mathcal{N}(v_j)|}$$

- Adamic/Adar:

$$\sum_{v \in \mathcal{N}(v_i) \cap \mathcal{N}(v_j)} \frac{1}{\log |\mathcal{N}(v)|}$$

Liben-Nowell and Kleinberg, 2003

Scoring functions

Paths and ensembles of paths between v_i and v_j

- Shortest path:

$$-\min_s \{path_{ij}^s > 0\}$$

- Katz score:

$$\sum_{l=1}^{\infty} \beta^l |paths^{(l)}(v_i, v_j)| = \sum_{l=1}^{\infty} (\beta A)_{ij}^l = (I - \beta A)^{-1} - I$$

- Personalized (rooted) PageRank:

$$PR = \alpha (D^{-1} A)^T PR + (1 - \alpha) I$$

Liben-Nowell and Kleinberg, 2003

Scoring functions

- Expected number of random walk steps:
  - hitting time: $-H_{ij}$
  - commute time $-(H_{ij} + H_{ji})$
  - normalized hitting/commute time $(H_{ij}\pi_j + H_{ji}\pi_i)$

- SimRank:

$$SimRank(v_i, v_j) = \frac{C}{|\mathcal{N}(v_i)| \cdot |\mathcal{N}(v_j)|} \sum_{m \in \mathcal{N}(v_i)} \sum_{n \in \mathcal{N}(v_j)} SimRank(m, n)$$

Liben-Nowell and Kleinberg, 2003

Vertex feature aggregations

- Preferential attachment:

$$k_i \cdot k_j = |\mathcal{N}(v_i)| \cdot |\mathcal{N}(v_j)|$$

  or

$$k_i + k_j = |\mathcal{N}(v_i)| + |\mathcal{N}(v_j)|$$

- Clustering coefficient:

$$CC(v_i) \cdot CC(v_j)$$

  or

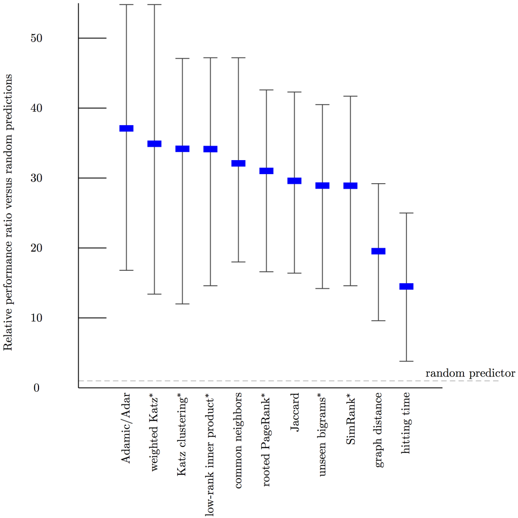
$$CC(v_i) + CC(v_j)$$

Low-rank approximations

- Low-rank approximation (truncated SVD)

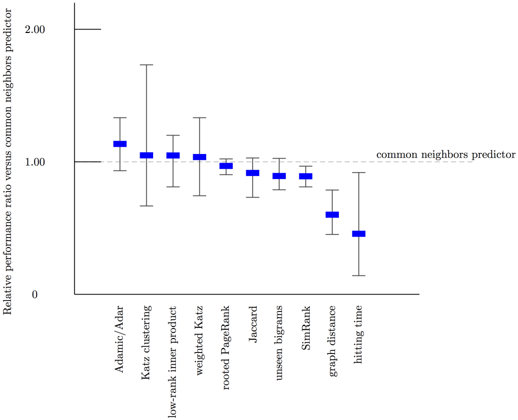
$$A \approx \sum_k U_k S_k V_k^T$$

Link prediction



Liben-Nowell and Kleinberg, 2007

Link prediction



Liben-Nowell and Kleinberg, 2007

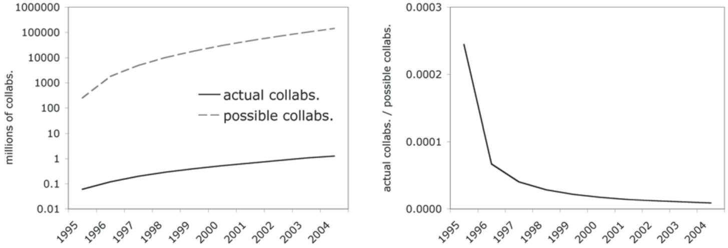
Binary classification

Challenging classification problem:

- Computational cost of evaluating of very large number of possible edges (quadratic in number of nodes)
- Highly imbalanced class distribution: number of positive examples (existing edges) grows linearly and negative quadratically with number on nodes

Prediction difficulty

Actual and possible collaborations between DBLP authors



Extreme class imbalance

from Rattigan and Jensen, 2005

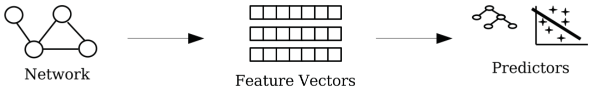
Link prediction with supervised learning

Supervised learning:

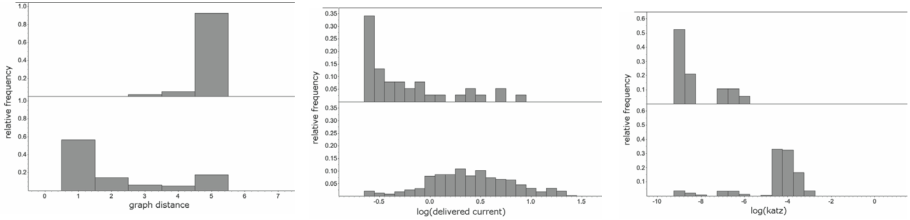
1. Features generation
2. Model training
3. Testing (model application)

Features:

- Topological proximity features
- Aggregated features
- Content based node proximity features



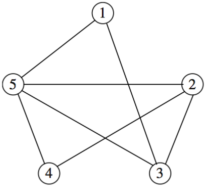
Discriminative abilities of features



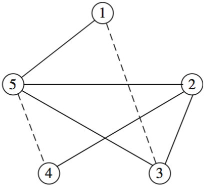
from Rattigan and Jensen, 2005

Simple evaluation

Simple "hold out set" evaluation



Whole graph



Training graph

Training and testing

Evaluation for evolving networks

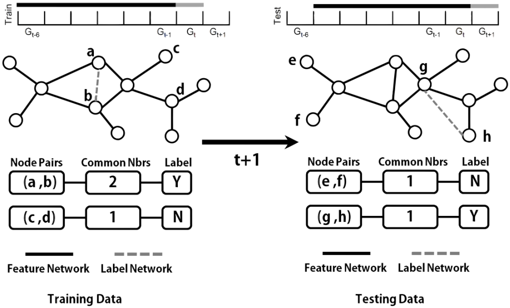


image from Y. Yang et.al, 2014

Evaluation metrics

- Precision and Recall, F-measure

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

- True positive rate (TPR), False positive rate (FPR), ROC curve, AUC

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$

Performance of classification algorithms

BIOBASE database (research publications)

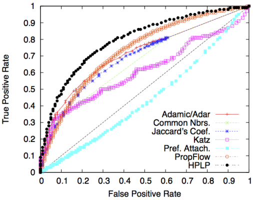
Classification model	Accuracy	Precision	Recall	F-value	Squared Error
Decision Tree	90.01	91.60	89.10	90.40	0.1306
SVM(Linear Kernel)	87.78	92.80	83.18	86.82	0.1221
SVM(RBF Kernel)	90.56	92.43	88.66	90.51	0.0945
K_Nearest Neighbors	88.17	92.26	83.63	87.73	0.1826
Multilayer Perceptron	89.78	93.00	87.10	90.00	0.1387
RBF Network	83.31	94.90	72.10	81.90	0.2542
Naive Bayes	83.32	95.10	71.90	81.90	0.1665
Bagging	90.87	92.5	90.00	91.23	0.1288

DBLP dataset (research publications)

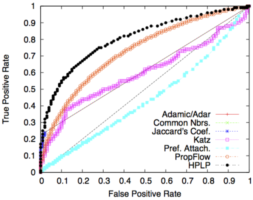
Classification model	Accuracy	Precision	Recall	F-value	Squared Error
Decision Tree	82.56	87.70	79.5	83.40	0.3569
SVM(Linear Kernel)	83.04	85.88	82.92	84.37	0.1818
SVM(RBF Kernel)	83.18	87.66	80.93	84.16	0.1760
K_Nearest Neighbors	82.42	85.10	82.52	83.79	0.2354
Multilayer Perceptron	82.73	87.70	80.20	83.70	0.3481
RBF Network	78.49	78.90	83.40	81.10	0.4041
Naive Bayes	81.24	87.60	76.90	81.90	0.4073
Bagging	82.13	86.70	80.00	83.22	0.3509

from M. Al Hasan, 2010

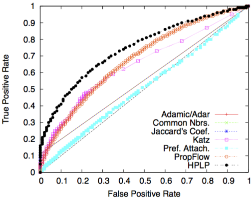
ROC curves



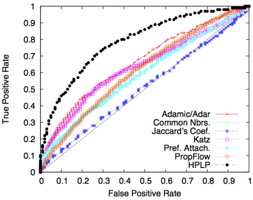
(a) phone $n = 2$



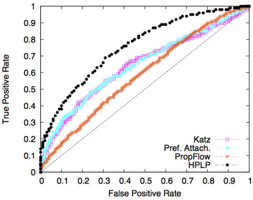
(b) phone $n = 3$



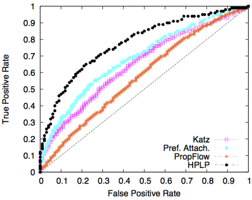
(c) phone $n = 4$



(d) condmat $n = 2$



(e) condmat $n = 3$



(f) condmat $n = 4$

from Lichtenwalter, 2010

Probabilistic models

- Local model, Markov random fields [Wang, 2007]
- Hierarchical probabilistic model [Clauset, 2008]
- Probabilistic relations models:
  - Bayesian networks [Getoor, 2002]
  - relational Markov networks [Tasker, 2003, 2007]

References

- D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. Journal of the American Society for Information Science and Technology, 58(7):1019?1031, 2007
- R. Lichtenwalter, J.Lussier, and N. Chawla. New perspectives and methods in link prediction. KDD 10: Proceedings of the 16th ACM SIGKDD, 2010
- M. Al Hasan, V. Chaoji, S. Salem, M. Zaki, Link prediction using supervised learning. Proceedings of SDM workshop on link analysis, 2006
- M. Rattigan, D. Jensen. The case for anomalous link discovery. ACM SIGKDD Explorations Newsletter. v 7, n 2, pp 41-47, 2005
- M. Al. Hasan, M. Zaki. A survey of link prediction in social networks. In Social Networks Data Analytics, Eds C. Aggarwal, 2011.