NATIONAL RESEARCH UNIVERSITY

Machine Learning on graphs. Link prediction

Network Science
Lecture 13

Leonid Zhukov

lzhukov@hse.ru
www.leonidzhukov.net/hse/2022/networks

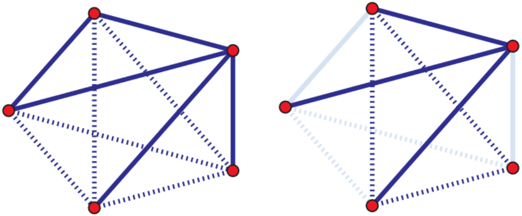
National Research University Higher School of Economics
School of Data Analysis and Artificial Intelligence, Department of Computer Science

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 $O(\frac{1}{|V|^2})$

Similarity based algorithms - unsupervised

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
4. Quality measurements - precision $TP/(TP + FP)$, precision at top N

Local similarity indices

Local neighborhood of v_i and v_j

- Number of common neighbors:

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

- Jaccard's coefficient:

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

- Resource allocation:

$$s_{ij} = \sum_{w \in \mathcal{N}(v_i) \cap \mathcal{N}(v_j)} \frac{1}{|\mathcal{N}(v)|}$$

- Adamic/Adar:

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

Liben-Nowell and Kleinberg, 2003

Local similarity indices

- Preferential attachment:

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

or

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

- Clustering coefficient:

$$s_{ij} = CC(v_i) \cdot CC(v_j)$$

or

$$s_{ij} = CC(v_i) + CC(v_j)$$

Path based methods

Paths and ensembles of paths between v_i and v_j

- Shortest path:

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

- Katz score:

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

- Personalized (rooted) PageRank:

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

Liben-Nowell and Kleinberg, 2003

Path based indeces

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

- 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

Community based methods

- Within-inter community/cluster of $v_i, v_j \in C$

$$\sum_{w \in \mathcal{N}(v_i) \cap \mathcal{N}(v_j)} \frac{|w \in C|}{|w \notin C|}$$

- Common neighbors with community information, $v_i, v_j \in C$, $f(w) = 1$ if $w \in C$

$$|\mathcal{N}(v_i) \cap \mathcal{N}(v_j)| + \sum_{w \in \mathcal{N}(v_i) \cap \mathcal{N}(v_j)} f(w)$$

- Resource allocation index with community information (soundarajan-hopcroft), $v_i, v_j \in C$, $f(w) = 1$ if $w \in C$

$$\sum_{w \in \mathcal{N}(v_i) \cap \mathcal{N}(v_j)} \frac{f(w)}{|\mathcal{N}(w)|}$$

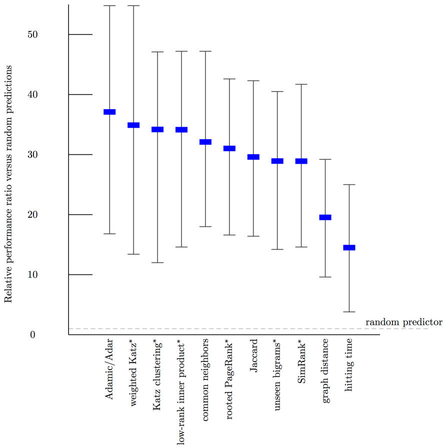
Low-rank approximations

- Low-rank approximation (truncated SVD)

$$A = \sum_k^n U_k S_k V_k^T \to \sum_k^r U_k S_k V_k^T = A', r < n$$

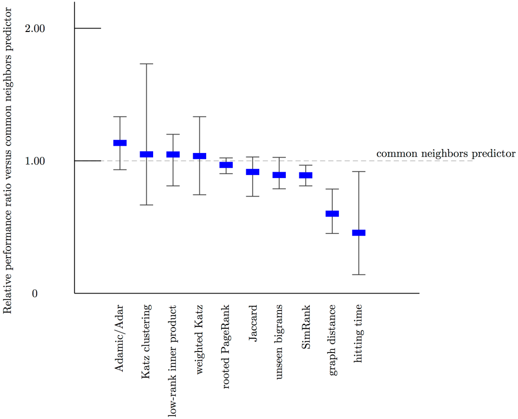
$$\underset{m \times n}{\overset{\hat{X}}{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix}}} \approx \underset{m \times r}{\overset{U}{\begin{pmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{pmatrix}}} \underset{r \times r}{\overset{S}{\begin{pmatrix} s_{11} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & s_{rr} \end{pmatrix}}} \underset{r \times n}{\overset{V^\top}{\begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn} \end{pmatrix}}}$$

Evaluation of scoring prediction



Ratio of predictor performance over the baseline, averaged 5 datasets

Evaluation of scoring prediction



Ratio of predictor performance over the baseline, averaged 5 datasets

Liben-Nowell and Kleinberg, 2007

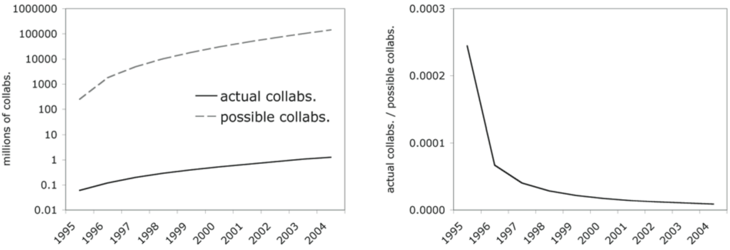
Classification for link prediction

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

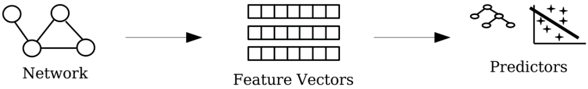
Link prediction with supervised learning

Supervised learning:

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

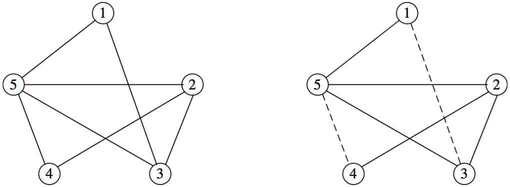
Features:

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



Simple evaluation

Simple "hold out set" evaluation



Whole graph Training graph

Evaluation metrics

- Precision and Recall, F-measure

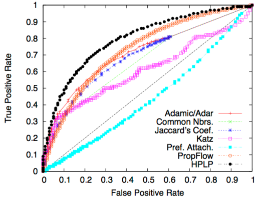
$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}$$

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

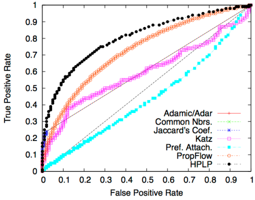
- True positive rate (TPR), False positive rate (FPR), ROC curve, AUC

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

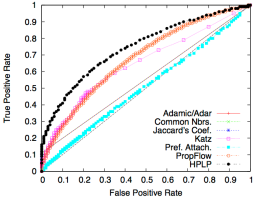
ROC curves



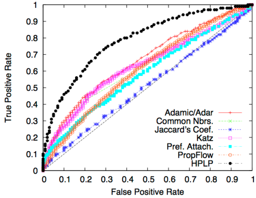
(a) phone $n = 2$



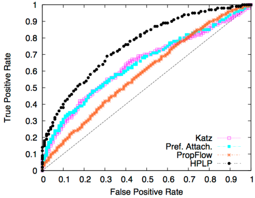
(b) phone $n = 3$



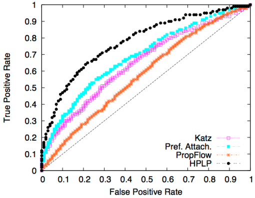
(c) phone $n = 4$



(d) condmat $n = 2$



(e) condmat $n = 3$



(f) condmat $n = 4$

from Lichtenwalter 2010

Training and testing

Evaluation for evolving networks

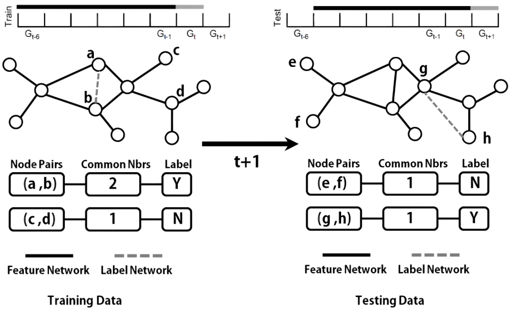


image from Y. Yang et.al, 2014

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.