Link Prediction in Social Networks

Farzad Zafarani
Fall 2014
Problem Definition:

Given a snapshot of a social network at time $t$, we seek to accurately predict the edges that will be added to the network during the interval from time $t$ to a given future time $t_0$. [Liben-Nowell, Kleinberg]
Applications

• Used in social networks to suggest friends [Leskovec 2010]
• Identifying gangsters network! [Kleinberg 2004]
• Identify user’s navigation history [Zhu 2003]
Proximity measures in networks

• graph distance \( x, y \)  length of shortest path between \( x \) and \( y \)
• common neighbors \( |\tau(x) \cap \tau(y)| \)
• Jaccard’s coefficient \( \frac{|\tau(x) \cap \tau(y)|}{|\tau(x) \cup \tau(y)|} \)
• Adamic/Adar \( \sum_{z \in \tau(x) \cap \tau(y)} \frac{1}{\log |\tau(z)|} \)
• preferential attachment \( |\tau(x)| \cdot |\tau(y)| \)
Unsupervised Learning Methods

• Using 10-fold cross validation
• Let $G'(V, V^*V-E)$ be the complement of the Graph $G(V, E)$
• at each iteration remove 10% of the edges in $G$ and add them to the complement Graph $G'$
• Measure the similarities using the proximity measures and sort them decreasingly
• Choose the first 10% of the edges to be the predicted links
Performance of unsupervised learning

comparing proximity measures on Facebook

- graph distance
- adamic Adar
- Jaccard's Coefficient
- Preferential Attachment
- Common neighbor
- Random Prediction
Supervised Learning

• By using proximity measures as a feature we can convert this problem to a linear classification problem
• edges in the graph are class labeled with 1 and 0 otherwise

If edge e=(x,y) is included in the graph G

\[ w_1 \* \text{Adamic}(x,y) + w_2 \* \text{graphDist}(x,y) + w_3 \* \text{commonNeighbor}(x,y) + w_4 \* \text{preferentialAttachment}(x,y) = 1 \]

otherwise

\[ w_1 \* \text{Adamic}(x,y) + w_2 \* \text{graphDist}(x,y) + w_3 \* \text{commonNeighbor}(x,y) + w_4 \* \text{preferentialAttachment}(x,y) = 0 \]

use linear regression, SVM, or Decision tree to learn the model!
Accuracy of the Supervised Learning Model

Supervised Comparison

- SVM
- Logistic Regression
- Decision Tree
Thank you!