# Node-Pair Feature Extraction for Link Prediction

Tesh Feyessa, Marwan Bikdash (Prof) and Gary Lebby(Prof.)
Department of Electrical and Computer Engineering
North Carolina A&T State University
Oct 9, 2011

# Motivation

- Current link prediction approaches focus on the structure of the network and don’t consider the objects behind the network to perform prediction. In addition some of graph theoretic inter-nodal relationships used to perform prediction require global visibility.

# Outline

- Introduction
- Network indices
  - degree
  - Clustering coefficient
- Node-pair features
  - Visibility
  - Reciprocity
  - Degree correlation
  - Common neighbors
- Learner
- Trust network
  - Set up
  - Results
- Conclusion

# Introduction

- Link prediction is the problem of predicting the
  - existence of undiscovered links or
  - probability of formation of links in the future
  - between two given nodes in a network.
- Analysis of node proximity  to each-other
- Structural property of the underlying graph is used to implement statistical relational learning
  - Degree, centrality, clustering coefficient
  - Shared neighborhood, paths
- Modeling by studying the underlying process such as object centered sociality
- Modeling via physical properties of the network, such as temporality and spatiality

# Network indices

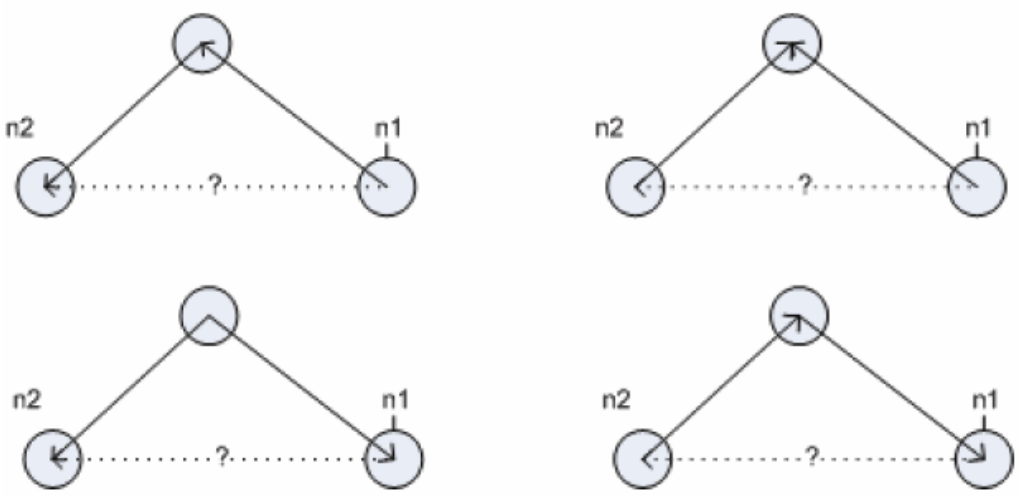
- Most indices in a network are defined either in the network level or node level, e.g., node centrality, degree, clustering, assortativity, …
- Path based measures, such as path, shortest path, and connectivity, describe relationship between two nodes
- This work assumes limited visibility and path computation usually require global view
- Node-level and network level measures are modified to node-pair features
- The clustering coefficient of a node i is the ratio between number of triangles it belongs to and the number of triangles that could have been formed with i as a vertex
- Assortativity (degree mixing) of a network is a Pearson correlation coefficient of the degrees at either ends of a link

$$C_i = \frac{\sum_{j \neq i} \sum_{h \neq i,j} a_{ij} a_{ih} a_{hj}}{d_i(d_i - 1)}$$

$$r = \frac{\sum_i j_i k_i / M - \left[\sum_i \frac{1}{2}(j_i + k_i)/M\right]^2}{\sum_i (j_i^2 + k_i^2)/2M - \left[\sum_i \frac{1}{2}(j_i + k_i)/M\right]^2}$$

# Node-Pair Features

- Reciprocity: if there is a direct link from node n1 to node n2 there is a reciprocal link
- In directed graphs four types of triangles can be formed if a link between n1 and n2 exists
- Some of these triangles represent transitivity property of the network while others represent cycles.
- the clustering between nodes i and j, is the ratio of triangles both nodes i and j are vertices of and the minimum of degrees of i and j




$$C_{ij} = \frac{\sum_{h \neq i,j} a_{ih} a_{hj}}{d_{ij}}$$

$$d_{ij} = \min(d_i, d_j)$$

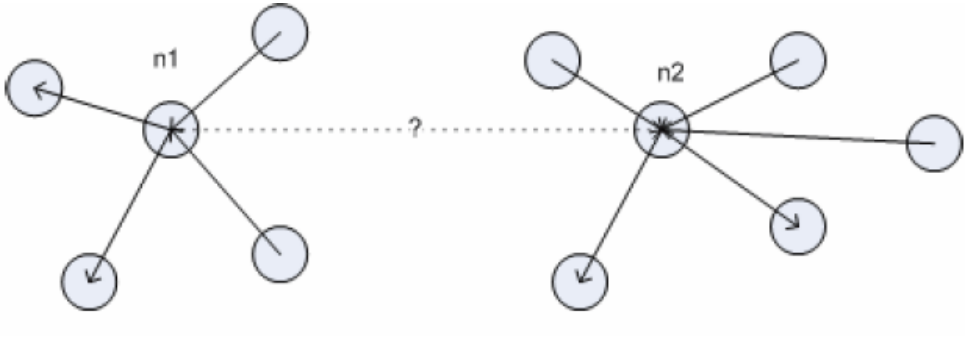
di, degree of node i

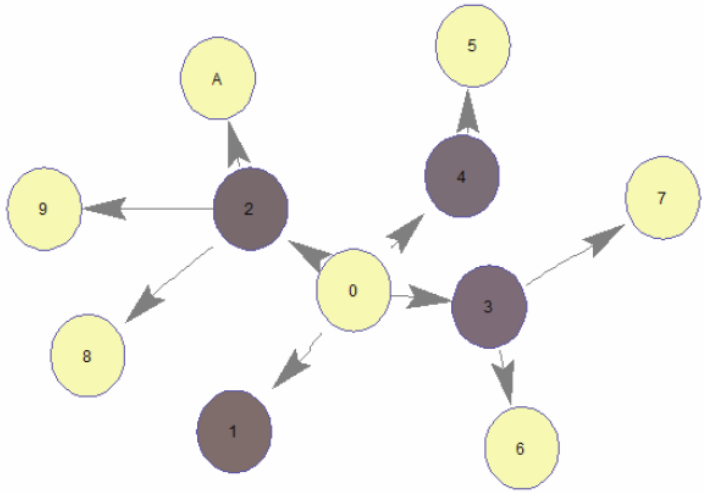
# Continued…

- The local assortativity between two nodes at either end of link i can be approximated
- The local assortativity between the node pairs on either end of link i can be approximated by dropping all summations and plugging ji+ki for M,

$$r_{jk} = \frac{4j_i k_i - j_i - k_i}{2j_i^2 + 2k_i^2 - j_i - k_i}$$

- ji and ki are the degrees of the nodes on either side of link i



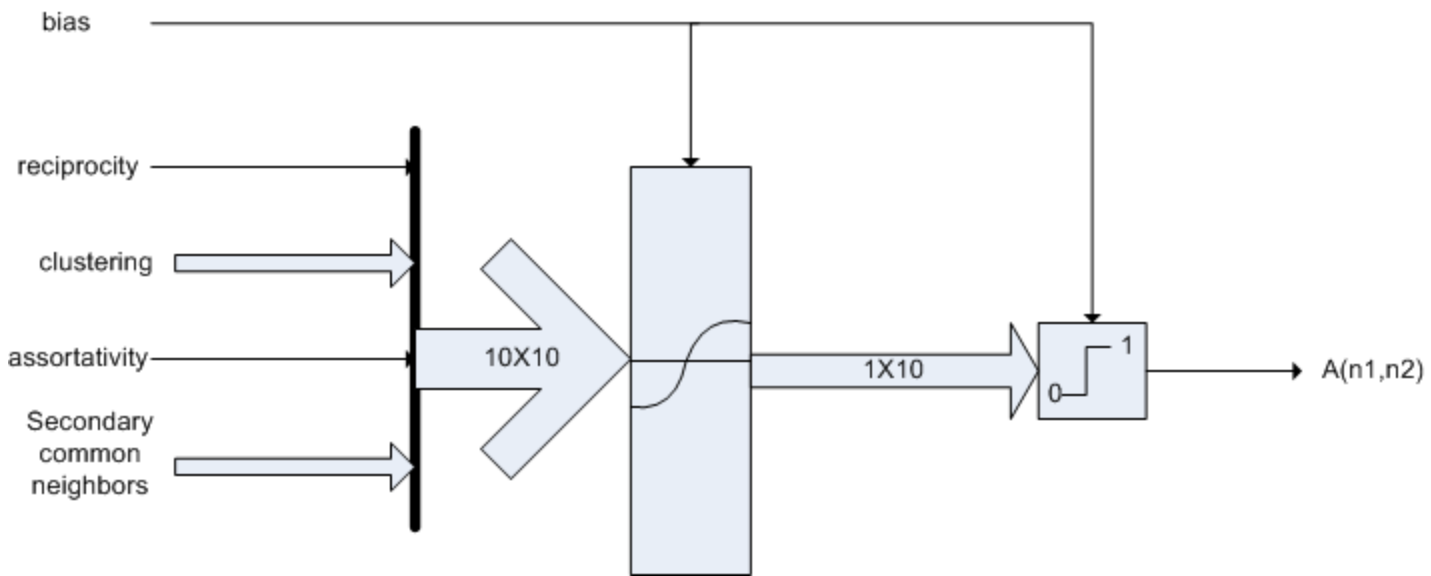



# Continued…

| Property | Description |
|---|---|
| Reciprocity | $a_{ji}$ |
| Clustering | $\frac{\sum_{h\neq i,j} a_{ih}a_{hj}}{\min(d_i^0,d_j^i)}$<br>$\frac{\sum_{h\neq i,j} a_{ih}a_{jh}}{\min(d_i^0,d_j^o)}$<br>$\frac{\sum_{h\neq i,j} a_{hi}a_{hj}}{\min(d_i^i,d_j^i)}$<br>$\frac{\sum_{h\neq i,j} a_{hi}a_{jh}}{\min(d_i^i,d_j^o)}$ |
| Assortativity | Degree correlation coefficient |
| $2^{nd}$ Common neighbors | $\frac{\sum_{h,k,u\neq i,j,a_{ih}=1,a_{kj}=1} a_{hu}a_{uk}}{\min(d_h^o,d_k^i)}$<br>$\frac{\sum_{h,k,u\neq i,j,a_{ih}=1,a_{kj}=1} a_{hu}a_{ku}}{\min(d_h^o,d_k^o)}$<br>$\frac{\sum_{h,k,u\neq i,j,a_{ih}=1,a_{kj}=1} a_{uh}a_{uk}}{\min(d_h^i,d_k^i)}$<br>$\frac{\sum_{h,k,u\neq i,j,a_{ih}=1,a_{kj}=1} a_{uh}a_{ku}}{\min(d_h^i,d_k^o)}$ |

# Classifier

- Link prediction defined as a classification problem
- Two layer feed-forward back propagation neural network (BPN)
  - with sigmoid transfer function.
  - The output of the network is fed to a hard decision function using predefined threshold values.
  - The threshold is varied to obtain the receiver operating characteristics (ROC)curve



# Trust Network

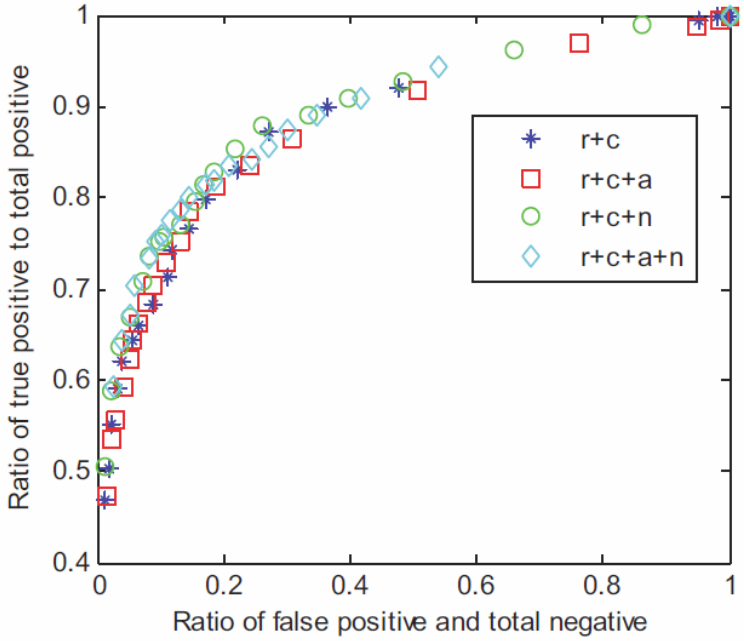
- A social network is always tied together by one or more objects of sociality.
- An online who-trust-whom network of general consumer review site Epinions.com
  - the consumer reviews are the objects of sociality
  - Members of the site, nodes of the network, interact with each other using the ”trust” relationship
  - the ”trust” relationship, links of the network

# Experiment and Results

- To avoid outliers in the extracted data, samples are taken only from the largest connected component
- Four combination of features
  - reciprocity and clustering,
  - reciprocity, clustering and assortativity
  - reciprocity, clustering and 2nd common neighbors
  - all extracted features
- All the features, except $2^{nd}$ common neighbors, only require 1-hop view

# Continued…

- Assortativity is the least important feature in determining existence of a link
- The network has an almost neutral degree mix, -0.01.




| Features | AUC | Accuracy |
|---|---|---|
| r+c | .881 | .813 |
| r+c+a | .873 | .811 |
| r+c+n | .893 | .820 |
| r+c+a+n | .882 | .828 |

# Conclusion

- A link prediction method that relies only on local information (as far as the neighbors of neighbors of a node) to characterize quantitatively the relationship between pairs of nodes is developed
- The features extracted from the network were used to train and test a supervised neural networks.
- A promising prediction accuracy in an object centered social network was obtained from this approach