# Link Prediction
# in
# Affiliation Networks

A Literature Review by

Christopher Carrino

Phys 597A

April 18, 2006

# Definition of the Problem

*In social networks:*
“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'*. [Liben-Nowell & Kleinberg 2003]

*In product networks:*
“Are the future interactions between consumers and products *inherently predictable* given the past consumer-product interactions?” [Huang & Zeng 2005]

# Applications of Link Prediction

1) Retail Recommender Systems

Used by e-commerce and marketing companies to recommend potential products to customers.

*Amazon.com*: recommends books, music and retail products to potential customers.

*SkyMall.com*: recommends retail products to potential customers.

*Netflix.com* and *MovieLens.org:* recommend movies to potential viewers.

[Riedl & Konstan 2002]

# Applications of Link Prediction

2) Information Filtering

Used by news forums, academic publication repositories and web search engines.

*GroupLens*: recommends current event news articles to subscribers.

*ReferallWeb*: recommends authors and information to people based on a trusted chain of referrals from other users.

*CiteSeer:* recommends scientific publications to researchers.

[Resnick & Varian 1997]

# Applications of Link Prediction

## 3) Biological Systems – Protein Interactions

Due to the fact that protein-protein interaction networks contain hub proteins and contain triads and tetrads at a higher frequency than random networks, it is possible to predict future protein interactions.

***"We show that we can leverage the information encoded in consensus interaction patterns to generate high relevance predictions for new interaction partners of any given protein." [Albert & Albert 2004]***

# Applications of Link Prediction

## 4) Social Network Analysis - Criminology

***"Research in security has recently begun to emphasize the role of social network analysis, largely motivated by the problem of monitoring terrorist networks; link prediction in this context allows one to conjecture that particular individuals are interacting even though their interaction has not been directly observed" [Liben-Nowell & Kleinberg 2003].***

*CrimeNet Explorer*: detects subgroups in criminal networks. [Xu & Chen, 2005]

*NETEST*: determines network structure in sparse criminal and terrorist networks. [Dombroski & Carley 2002]

*i2inc Analyst Suite:* identifies patterns in social, phone call, and financial networks involved in criminal investigations. [i2inc.com]

# Algorithms for Link Prediction

## 1) Centrality

a) Degrees
b) Betweenness
c) Closeness

## 2) Node Neighborhoods

a) Common neighbors
b) Preferential attachment
c) Jaccard's coefficient
d) Adamic/Adar

## 3) Path Lengths

a) Graph Distance
b) Katz

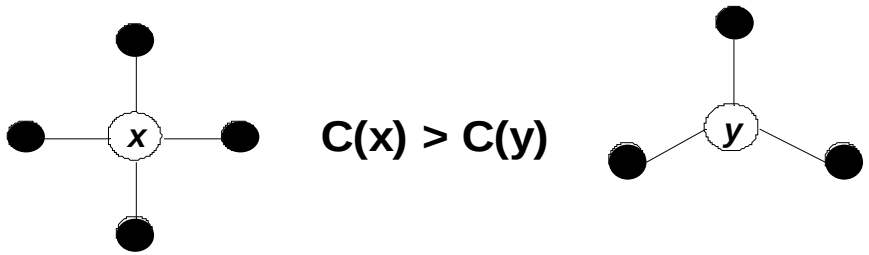
## 4) Collaborative Filtering

a) User Based
b) Item Based

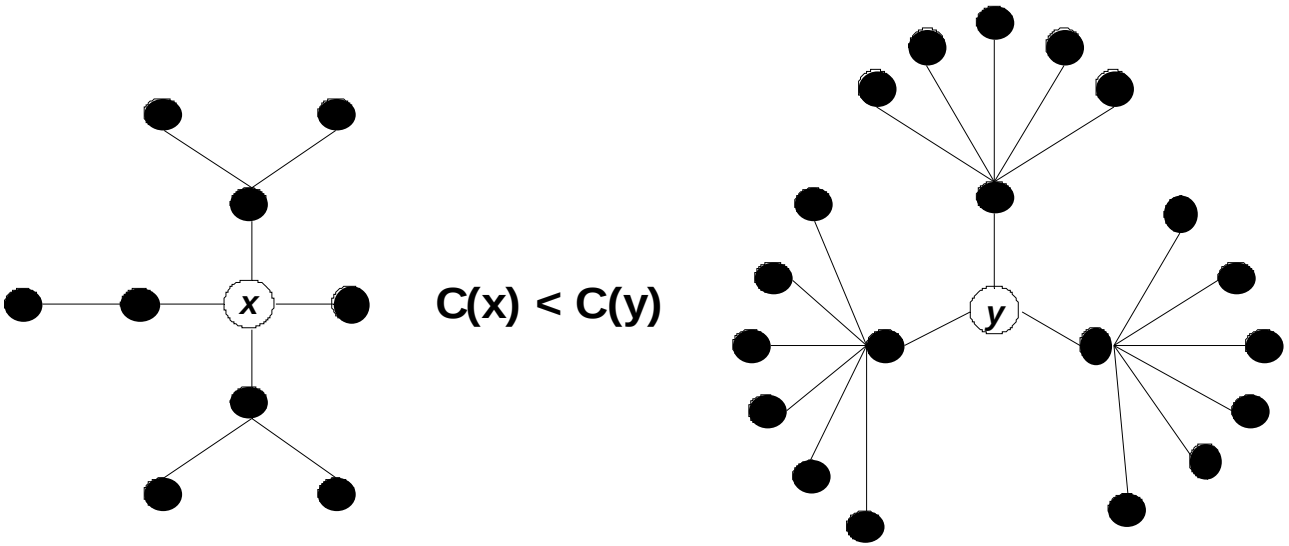
# Algorithms for Link Prediction

## 1a) Centrality: Degrees

***a) Freeman's Approach** [Freeman 1978]*




b) Bonacich's Approach [Bonacich 1972a], [Bonacich 1972b]



# Algorithms for Link Prediction

## 1b) Centrality: Betweenness

a) Freeman's Approach [Freeman 1978].

$$C(x) = \sum_{y}^{N} \sum_{w}^{N} \frac{g_{yw}(x)}{g_{yw}} \quad (y < w)$$

Where $g_{yw}$ is the number of geodesics between nodes y and w.

b) Flow Centrality [Hanneman & Riddle 2005].

$$C(x) = \sum_{y}^{N} \sum_{w}^{N} \frac{p_{yw}(x)}{p_{yw}} \quad (y < w)$$

Where $p_{yw}$ is the total number of distinct paths between nodes y and w.

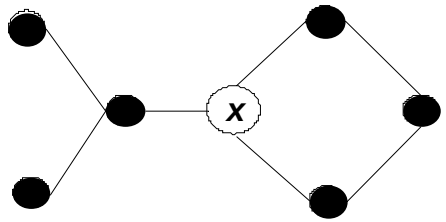
c) Local Betweenness [Thadakamalla et al. 2005].

$$C(x) = \sum_{y}^{N} \sum_{w}^{N} \frac{g_{yw}(x)}{g_{yw}} \quad (y < w)$$

Similar to Freeman's approach, but nodes *y* and *w* are limited to be 1st or 2nd degree neighbours of the root node, *x*.

# Algorithms for Link Prediction

## 1c) Centrality: Closeness




a) Path Distances [Sabidussi 1966].

$$C(x) = \left[ \sum_{i=1}^{N} d(x, y_i) \right]^{-1}$$

Where $d(x,y_i)$ is the number of edges in the geodesic connecting nodes x and y.

b) Reach [Hanneman & Riddle 2005].

The percentage of nodes in the network that can be reached from a given node in a given number of steps. The number of steps can vary from 1,2,...,N. The maximum valud of centrality is attained when the entire network can be reached in 1 step.

# Algorithms for Link Prediction

## 2) Node Neighborhoods

Let Γ(*x*) denote the set of neighbors of *x* in *G*, and | Γ(*x*) | denotes the count of these neighbors.

a) Common Neighbors: $sim_{(x,y)} = | \Gamma(x) \cap \Gamma(y) |$ [Newman 2001]

b) Prefential Attachment: $sim_{(x,y)} = | \Gamma(x) | \bullet | \Gamma(y) |$ [Barabasi et al. 2002]

c) Jaccard's Coefficient: $sim_{(x,y)} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$ [Salton & McGill 1984]

d) Adamic/Adar: $sim_{(x,y)} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$ [Adamic & Adar 2003]

# Algorithms for Link Prediction

## 3) Path Lengths

a) Graph Distance: $sim_{(x,y)}$ = Length of the geodesic b/w x and y. [Newman 2003]

b) Hubbell, Katz, Taylor, and Zelen influence [Katz 1953], [Hanneman & Riddle 2005].

The greater the path length between two nodes, the weaker their connection, based upon an attenuation factor. Each step in the path decreases the overall path score.

A path length of 1 has a weight of 1. A path length of 2 has a weight equal to the product of 1 and the attenuation factor. A path length of 3 has weight equal to the product of 1, the attenuation factor and the attenuation factor again. This continues on until all paths have weighted in the same manner.

# Algorithms for Link Prediction

## 4) Collaborative Filtering

a) User-Based algorithms: operate on the assumption that consumers who have bought similar products in the past will prefer to buy similar products in the future [Riedl & Konstan 2002].

b) Item-Based algorithms: operate on the assumption that items that have been co-purchased in the past will continue to be co-purchased in the future [Huang & Zeng 2005].

$\text{sim}(x, y) = \cos(\vec{u}, \vec{v}) = \frac{\vec{u} \bullet \vec{v}}{\|\vec{u}\|_2 \|\vec{v}\|_2}$

where u is the attribute vector for node *x* and v is the attribute vector for node *y*.

$$P(y|x) = \frac{P(x \cap y)}{P(x) \times (P(y))^{\alpha}}$$

Eliminates the bias due to pairing of unpopular items with popular items, by using a weighted condition where α is a parameter from 0 to 1 that represents the popularity of item y [Karypis 2001], [Deshpande & Karypis 2004].

# Implications to My Research

Research Area: Repeated Social Collaborations.

Actors choose to collaborate with each other when they have shared goals.

Since goals are not usually included in network data sets, predictive algorithms must rely on network topology to predict future collaborations.

Despite this limitation, it is possible to infer the key figure(s) in a social group that chose the collaboration.

New prediction algorithms are needed that first seek to identify the key actor(s) in a social group that organized the collaboration, then to derive a prediction set based on the central actor's past collaborations.

By doing this, the algorithms will capture the choice inherent in the collaboration, and the shared interest of the actors.

# References


[Albert & Albert 2004] István Albert and Réka Albert, “Conserved network motifs allow protein-protein interaction prediction”, Bioinformatics 20(18): 3346 - 52 (2004).

[Bonacich 1972a] Phillip Bonacich, “Technique for analyzing overlapping memberships”, in Herbert Costner (ed.) Sociological Methodology, Jossey-Bass, San Francisco, (1972).

[Bonacich 1972b] Phillip Bonacich, “Factoring and weighting approaches to status scores and clique detection”, Journal of Mathematical Sociology, 2: 113-120 (1972).

[Dombroski & Carley 2002] Matthew J. Dombroski and Kathleen M. Carley, “NETEST: Estimating a Terrorist Network’s Structure”, Computational & Mathematical Organization Theory 8(3): 235 - 241 (2002).

[Hanneman & Riddle 2005] Hanneman, Robert A. and Mark Riddle. “Introduction to social network methods”, Riverside, CA: University of California, Riverside ( published in digital form at http://faculty.ucr.edu/~hanneman/ ) (2005).

[Huang & Zeng 2005] Zan Huang and Daniel D. Zeng, “Why Does Collaborative Filtering Work? - Recommendation Model Validation and Selection by Analyzing Bipartite Random Graphs”, Preprint (2005).

[Liben-Nowell & Kleinberg 2003] David Liben-Nowell and Jon Kleinberg, “The Link Prediction Problem for Social Networks”, ACM 1-58113-723-0/03/0011 (2003).

[Resnick & Varian 1997] Paul Resnick and Hal R. Varian, "Recommender Systems", Communications of the ACM, Vol. 40(3):56-58 (1997).

[Riedl & Konstan 2002] John Riedl and Joseph Konstan, “Word of Mouse: The Marketing Power of Collaborative Filtering”, Warner Books, New York, NY (2002).

[Thadakamalla et al. 2005] Hari P. Thadakamalla, Soundar R. T. Kumara and Réka Albert, “Search in weighted complex networks” Physical Review E 72:066128 (2005).

[Xu & Chen, 2005] Jennifer J. Xu and Hsinchun Chen, “CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery”, ACM Transactions on Information Systems 23(2): 201–226 (2005).