LINK PREDICTION IN SOCIAL NETWORKS

Independent Project Final Report

CIS 830 Advanced Topics in Data Mining

Instructor: Prof. William H. Hsu

MS Student: Svitlana Volkova

Table of Contents

> *…It seems clear that the first epoch of the 3rd Millennium, the Network Age, was already asserting itself in increasingly powerful ways by the early 1990s, if not sooner, through the literal reorganization of the social, political, and economic infrastructure around electronic networks… [Smi01]*

Introduction

The person who built the modern social network theory was the Stanley Milgram [Was94].

> *[Social network] is a map of the individuals, and the ways how they are related to each other.*

A single person is the node of the network while edges, that link nodes and are called also "connections", "links", correspond to relationships between people as represented on Fig.1. There are a lot of examples social network services, such as:

- MySpace;
- Facebook;
- LiveJournal, etc.

Why we are more interested in Facebook? Because, it is doubling in size once every six months by 100,000 users per day. More than 60 percent of users are outside of college age. Zuckerberg attributed the power of Facebook to the *"social graph"* the network of connections and relationships between people on the service. He said, "It's the reason Facebook works." [Far07]

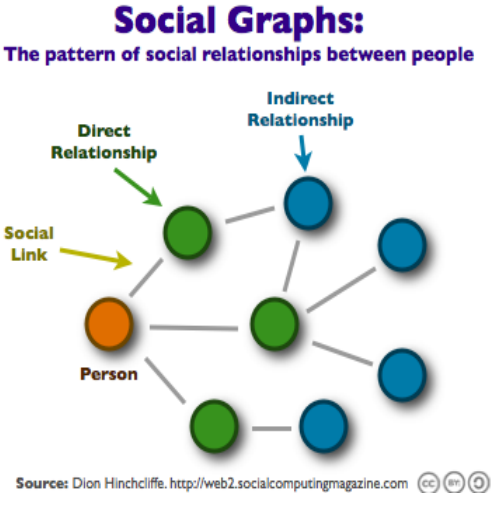


Figure 1: Social graph representation

There are other popular social networks that created and maintained by commercial companies such as: Friendster or LinkedIn. Another social network example is the structure of the web, especially the network of hyperlinks between home pages – Figure 2 [Ada03].

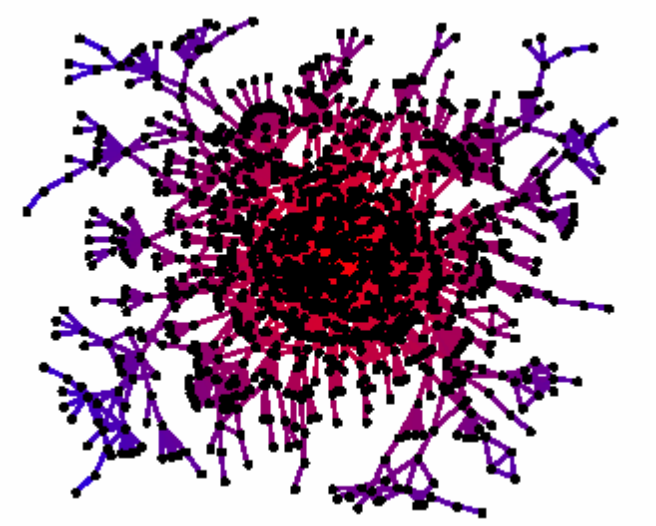


Figure 2: The social network based on hyperlinks at Stanford University

Why is it difficult to predict links in social networks?

Social networks are highly dynamic, sparse and have collective structure, that's why its outcome is difficult to predict. Moreover, because of early stage popularity, it is possible to estimate the popularity at later stage.

The *task of accurate prediction the presence of links or connections in a domain* is on the one hand important task and on the other hand is very challenging. So, can we predict more on social network?

Since the links from the network, their maintenance and quality, reflect social behaviors of individuals, the research on them can be helpful at the quantitative and qualitative assessment of human relationships in the age of information society. Moreover, link prediction is applicable to a wide variety of areas like bibliographic domain, molecular biology, criminal investigations and recommender systems.

Related Studies

Link analysis and social network analysis (SNA) tasks are essential in such areas of study as: sociology, criminology, and intelligence because of using graph-theoretic representations [Jen98, Was94].

On the one hand, most current data-mining methods assume the *data is from a single relational table* (e.g. as in market-basket analysis). So, such data representation is "propositional" (aka "feature vector" or "attribute value") representation of examples [Wit99].

Relational data mining (RDM) [Dˇze01], on the other hand, concerns mining *data from multiple relational tables* that are connected.

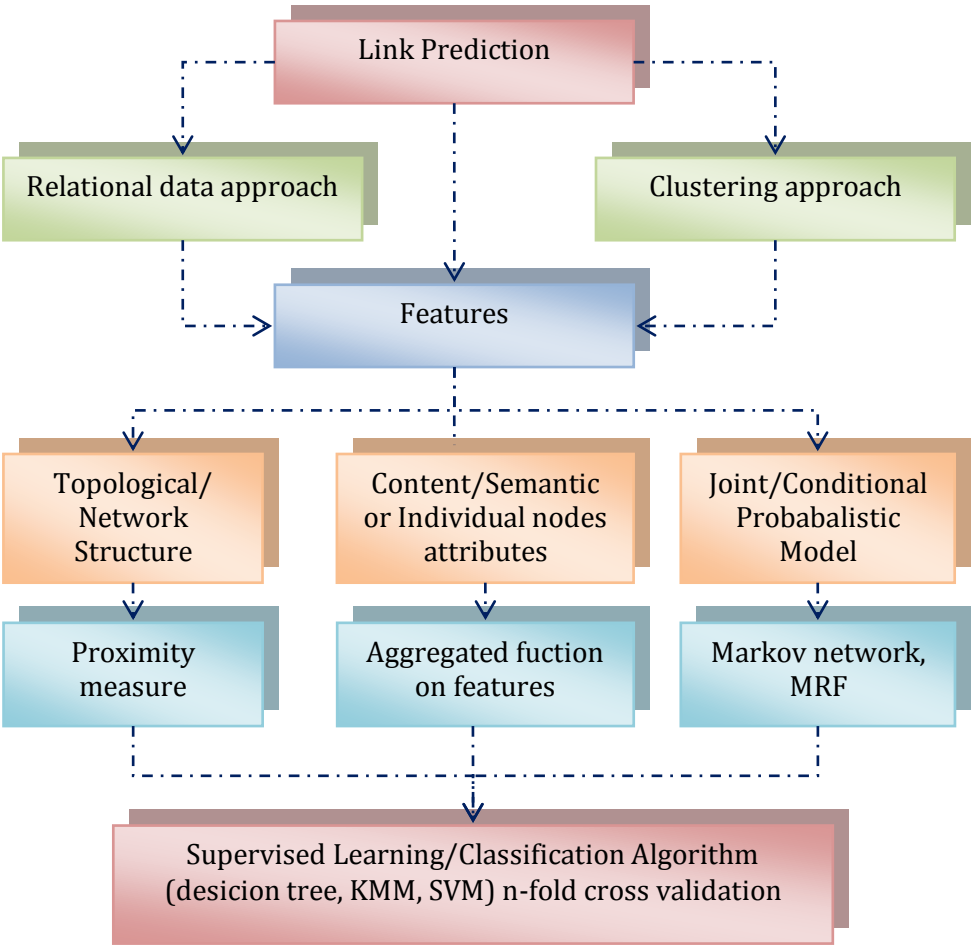


Figure 3: Different approaches to link prediction task [Var08]

The most widely studied methods for inducing relational patterns are those in *inductive logic programming (ILP)* [Lav94]. ILP concerns the induction of Horn-clause rules in first-order logic (i.e., logic programs) from data in first-order logic.

There are several basic methods for dealing with link prediction task in social networks: supervised vs. unsupervised.

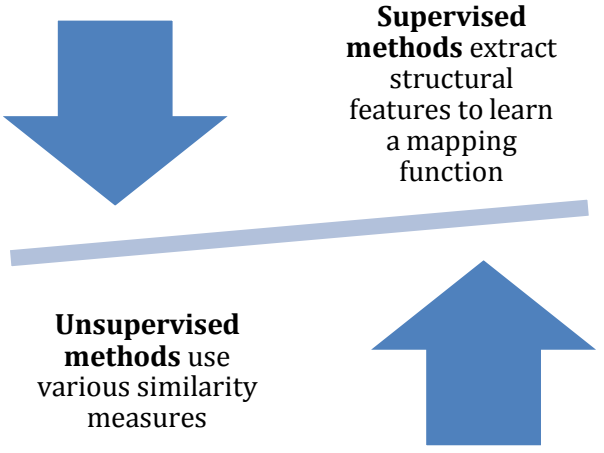


Figure 4: Application of supervised vs. unsupervised methods

Supervised Learning Approaches (SLA) are based on learning a binary classifier that will predict whether a link exists between a given pair of nodes or not [Hassan et al., 2006]

Moreover, classification whether a link exists or not can be performed using various supervised learning/classification algorithms like decision tree, K-nearest neighbors or support vector machines (SVM).

Let's consider in details link prediction task as classification, where we are given two entities in a network, and we should define if there be a link between them?

Classification based on features of entities: entity attributes and relational (graph-based) features (indirect relations).

- ✓ *Attributes:* number of neighbors, interests, topic model, affiliations, demographic data (geographical location)
- ✓ *Graph-Based Features:* length of shortest path, neighborhood overlap, relative importance, mean first passage time

Directed Graphical Models vs. Undirected Graphical Models (e.g. Markov Networks), such as *Bayesian networks and PRMs* [Get02] allow easily captures the dependence of link existence on attributes and constraint probabilistic dependency graph be a directed acyclic.

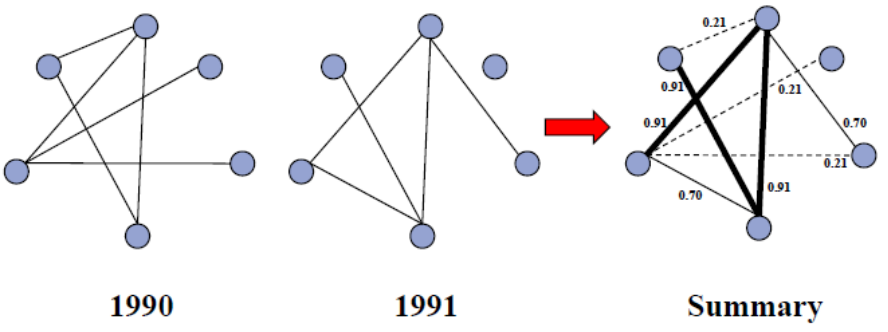


Figure 5: Application of probabilistic models to link prediction task

In order to resolve main issue let's consider existed approaches that can be helpful:

- exploring relational structure, clustering [Jen03, Get02];
- using links to predict classes/attributes of entities [Get04, Tas03, Jen98];
- predicting link types based on known entity classes [Tas03]
- predicting links based on location in high-dimensional space [Hof03];
- ranking potential links using a single graph-based feature [Kle04].

In addition, link discovery problem was deeply investigated at Kansas State University:

- [Hsu07] considered the problems of predicting, classifying annotating friends' relations in social networks by application feature constructing approach.
- [Wen08] proposed genetic programming-based symbolic regression approach to the construction of the relational features for link analysis task in social network domain.

Theoretical Background

Metrics in social network analysis [Wikipedia]

Betweenness

Measure reflects the number of people who a person is connecting indirectly through their direct links.

Bridge

An edge is said to be a bridge if deleting it would cause its endpoints to lie in different components of a graph.

Centrality

This measure gives a rough indication of the social power of a node based on how well they "connect" the network. "Betweenness", "Closeness", and "Degree" are all measures of centrality.

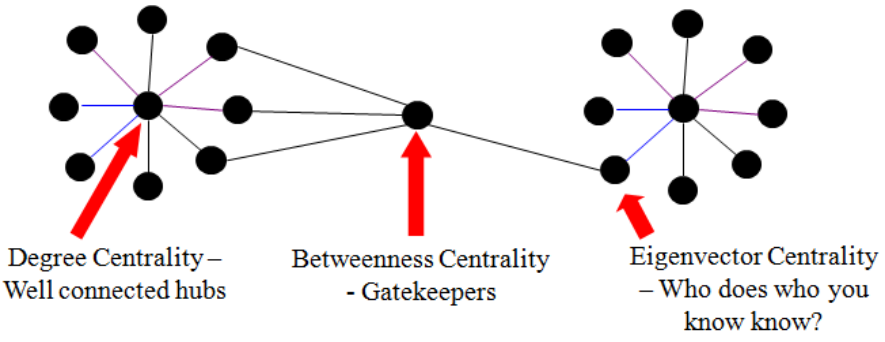


Figure 6:

Centralization

The difference between numbers of links for each node divided by maximum possible sum of differences.

Closeness

The degree an individual is near all other individuals in a network (directly/indirectly). Closeness is the inverse of the sum of the shortest distances between each individual and every other person in the network.

Path Length

Measure represents distances between pairs of nodes in the network.

Clustering coefficient

A measure of the likelihood that two associates of a node are associates them.

Cohesion

The degree to which actors are connected directly to each other by cohesive bonds.

Degree

The count of the number of ties to other actors in the network ("geodesic distance").

(Individual-level) Density

The degree a respondent's ties know one another/proportion of ties among an individual's nominees (sparse versus dense networks).

Flow betweenness centrality

The degree that a node contributes to sum of maximum flow between all pairs of nodes (not that node).

Eigenvector centrality

Measure of the node's importance a network.

Local Bridge

An edge is a local bridge if its endpoints share no common neighbors.

Prestige

In a directed graph prestige is the term used to describe a node's centrality. "Degree/Proximity/Status Prestige", are all its measures

Radiality

Degree an individual's network reaches out into the network and provides novel information and influence.

Reach

The degree any member of a network can reach other members.

Structural cohesion

The minimum number of members who if removed from a group, would disconnect the group.

Structural equivalence

Refers to the nodes have a common set of linkages to other nodes.

Mining tasks in network-structured data

There several fundamental tasks of link mining defined by [Get02]

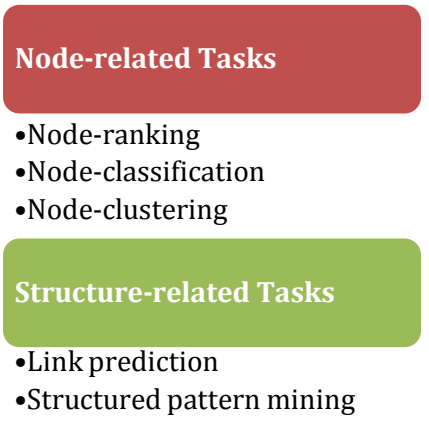


Figure 7: Representation of different link mining tasks

Link prediction addresses four different problems as shown in the figure below. Most of the research papers on link prediction, concentrate on *problem of link existence* (whether a "new" link between two nodes in a social network will exist in the future or not). This is because the link existence problem can be easily extended to the other two problems of *link weight* (links have different weights associated with them) and *link cardinality* (more than one link between same pair of nodes in a social network). The fourth problem of *link type* prediction is a bit different which gives different roles to relationship between two objects [Var08].

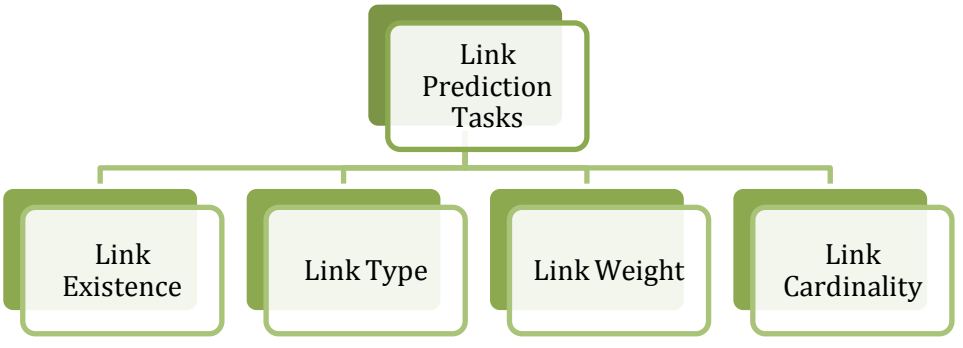


Figure 8: Differentiation of link prediction tasks

Link prediction problem

Link prediction problem is equal to network structure prediction problem, especially if we are considering social networks. There are several types of well-know link prediction methods that based on:

- node information;
- structural information.

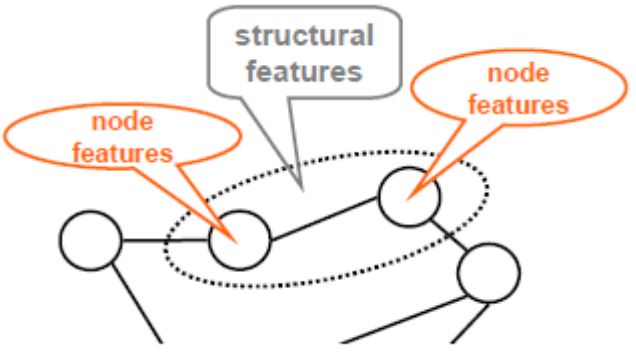


Figure 9: Node attributed vs. structural features [Get04]

So, in social network relations among data are represented as a graph structure, where each node represents a data and a link represents a relation between two data, in other words nodes represent constituent elements and links represent relations among them.

In addition, each node can also have an associated vector-structured data in the network model. In comparison to machine learning standard tasks settings: data is represented as tables, where rows represent observations and columns represent features/attributes.

Moreover, let's distinguish two link prediction tasks in networked data:

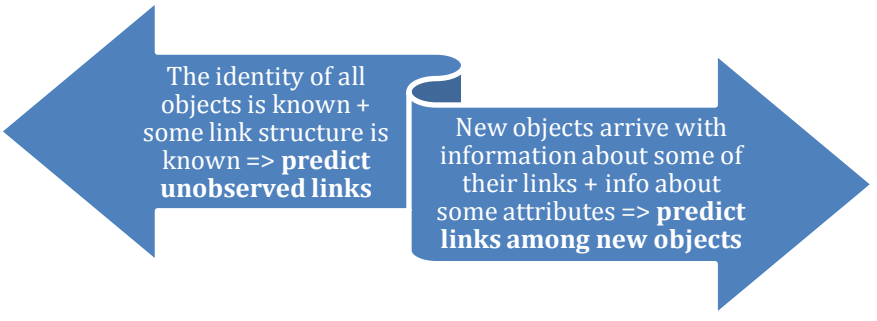


Figure 10: Unobserved link prediction task vs. new links prediction task

Mathematical description for problem

The link prediction problem is usually described as:

> *Given a set of data instances* $V = \{v_i\}_{i=1}^{n}$, *which is organized in the form of a social network* $G = (V, E)$, *where* E *is the set of observed links, then the task to predict how likely an unobserved link* $e_{ij} \notin E$ *exists between an arbitrary pair of nodes* $\langle v_i, v_j \rangle$ *in the data network.*

Given a snapshot of a social network at time t (or network evolution between t_1 and t_2), 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'.

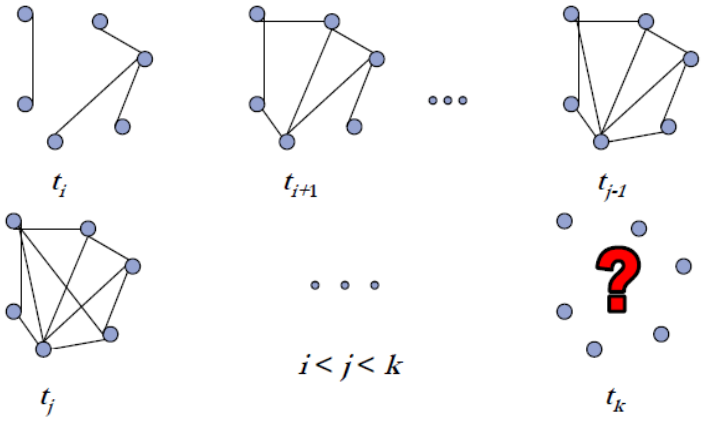


Figure 11: Visual representation of link prediction task

All the methods assign a connection weight $score(x, y)$ to pairs of nodes x, y, based on the input graph, and then produce a ranked list in decreasing order of $score(x, y)$. They can be viewed as computing a measure of proximity or "similarity" between nodes x and y. Negated length of shortest path between x and y. All nodes that share one neighbor will be linked.

Classification of measures for LP approaches

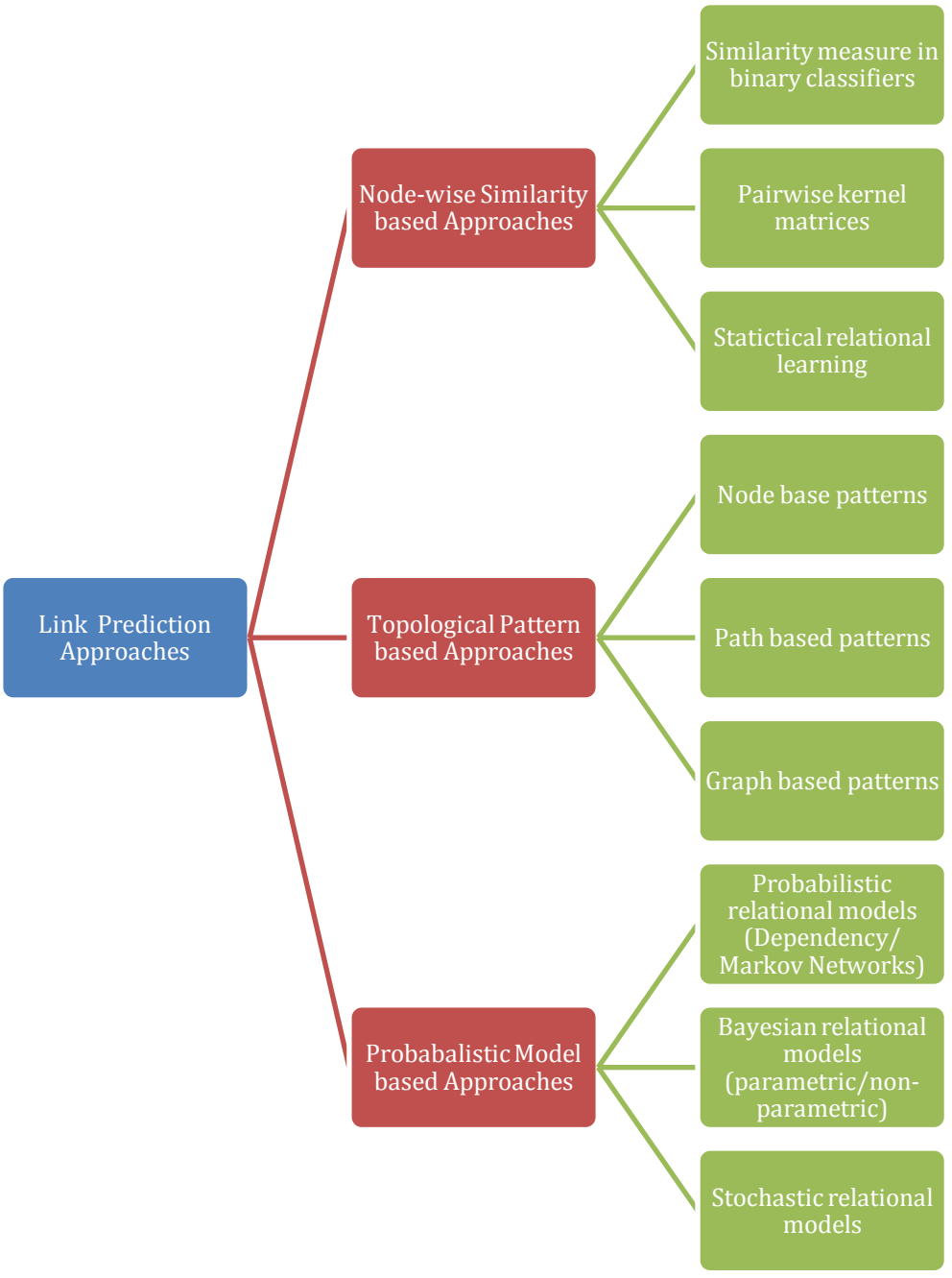


Figure 12: Classification of approaches for link prediction task [Xia08]

Node Based Topological Patterns [Xia08]

Common Neighbors [Newman 2001]

The measure states that the probability of scientists collaborating increases with the number of other collaborators they have in common:

$$score(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

Jaccard Similarity

The measure defines the issue that scientists may have in common neighbors because each has a lot of neighbors, but not because they are strongly related to each other:

$$score(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

Adamic/Adar [Adamic, 2003]

The measure allows counting common neighbors but gives more weight to neighbors that are not shared with many others:

$$score(x, y) = \sum_{z \in P(x) \cap P(y)} \frac{1}{log|\Gamma(y)|}$$

Ex. neighbors who are linked with 2 nodes are give weight - $1/\log(2) = 1.4$, while the neighbors who are linked with 5 nodes - $1/\log(5) = 0.62$

$$score^*_{unweig\ hted}(x, y) = \left|\left\{z: z \in \Gamma(y) \cap S_x^{\langle\delta\rangle}\right\}\right|$$

$$score^*_{weig\ hted}(x, y) = \sum_{z \in \Gamma(y) \cap S_x^{\langle\delta\rangle}} score(x, z)$$

Preferential attachment [Barabasi, 2003]

The probability of co-authorship of x and y is correlated with the product of the number of collaborators of x and y. This proposal corresponds to the measure:

$$score(x, y) = |\Gamma(x)| \cdot |\Gamma(y)|$$

Path Based Topological Patterns [Xia08]

Katz [Katz, 1953]

Measure that directly sums over this collection of paths, exponentially damped by length to count short paths more heavily. This notion leads to the measure where $paths_{x,y}^{\langle l \rangle}$ is the set of all length-l paths from x to y, and $\beta > 0$ is a parameter of the predictor. There are two variants of this Katz measure: (a) unweighted, in which $paths_{x,y}^{\langle l \rangle} = 1$ and y have collaborated and 0 otherwise, and (b) weighted, in which is the number of times that x and y have collaborated.

$$score(x, y) = \sum_{l=1}^{\infty} \beta^l \cdot \left| path_{x,y}^{\langle l \rangle} \right|$$

Hitting time

A random walk on G starts at a node x and iteratively moves to a neighbor of x chosen uniformly at random from the set $\Gamma(x)$. The hitting time $H_{x,y}$ from x to y is the expected number of steps required for a random walk starting at x to reach y. Because the hitting time is not in general symmetric, it also is natural to consider the commute time $C_{x,y} = H_{x,y} + H_{y,x}$. Both of these measures serve as natural proximity measures and hence (negated) can be used as $score(x, y)$.

PageRank [Brin, 1998]

It defines $score(x, y)$ under the rooted PageRank measure with parameter $\alpha \in [0,1]$ to be the stationary probability of y in a random walk that returns to x with probability α each step, moving to a random neighbor with probability $1 - \alpha$.

SimRank [Jeh, 2002]

SimRank is a fixed point of the following recursive definition: Two nodes are similar to the extent that they are joined to similar neighbors. Numerically, this quantity is specified by defining $score(x, x) = 1$ for a parameter $\gamma \in [0.1]$.

$$score(x, y) = \gamma \cdot \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} score(a, b)}{|\Gamma(x)| \cdot |\Gamma(y)|}$$

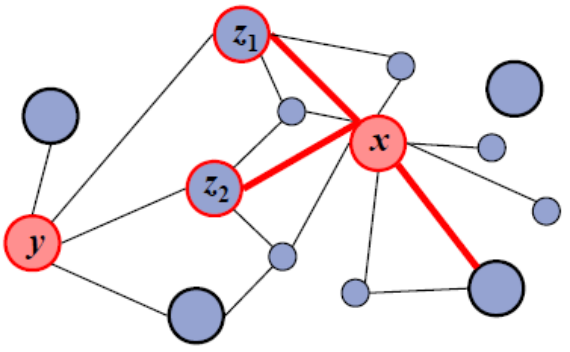
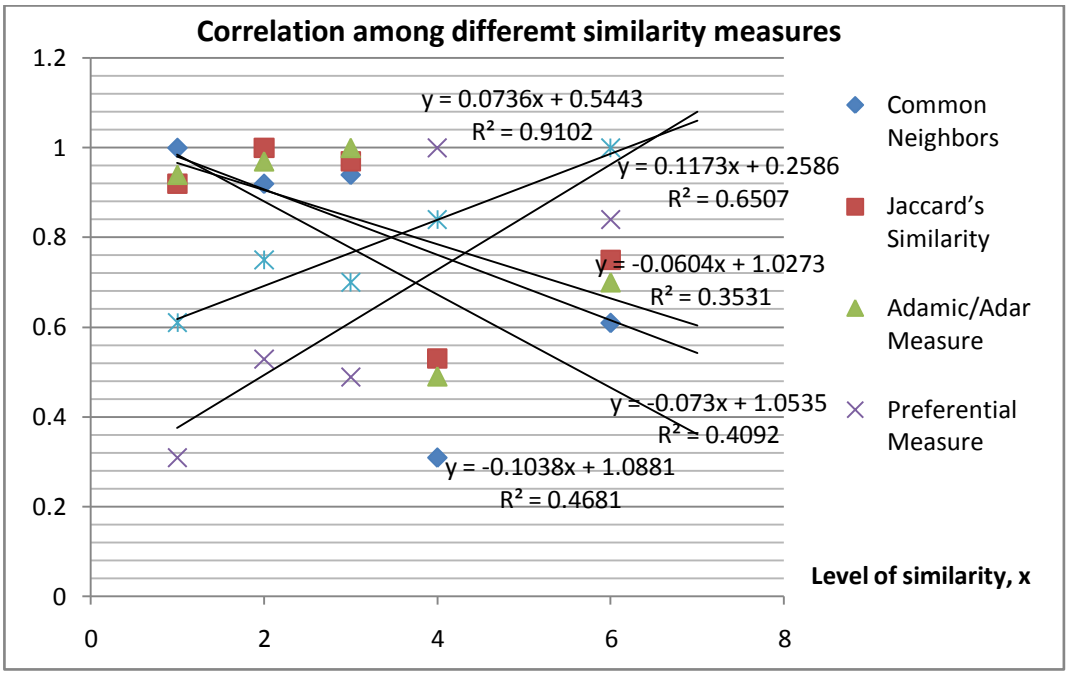


Figure 13: The level of similarity between nodes (red edges - strong)

	Common Neighbors	Jaccard's Similarity	Adamic/Adar Measure	Preferential Measure	Kartz Measure
Common Neighbors	1	0.92	0.94	0.31	0.61
Jaccard's Similarity	0.92	1	0.97	0.53	0.75
Adamic/Adar Measure	0.94	0.97	1	0.49	0.70
Preferential Measure	0.31	0.53	0.49	1	0.84
Katz Measure	0.61	0.75	0.70	0.84	1

The data for table was retrieved from [Nov03].



Experiment Planning

Crawling Social Network

Crawler is automatic programs which explores the World Wide Web, following the links and searching for information or building a database such programs are often used to build automated indexes for the Web, allowing users to do keyword searches for Web documents. Web crawlers are programs that exploit the graph structure of the Web to move from page to page [Cha07].

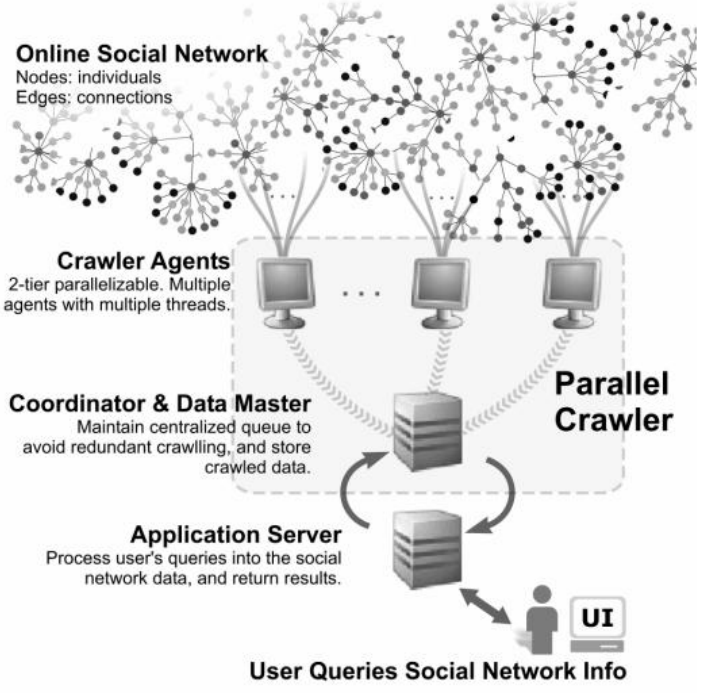
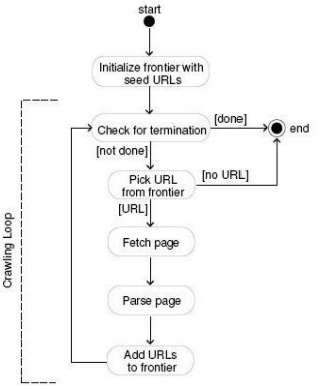


Figure 14: Parallel crawling of social networks example

For retrieving Facebook database will users' profiles it is planning to use free open-source SQL full-text search engine Sphinx that available on http://www.sphinxsearch.com/ .



Facebook Database

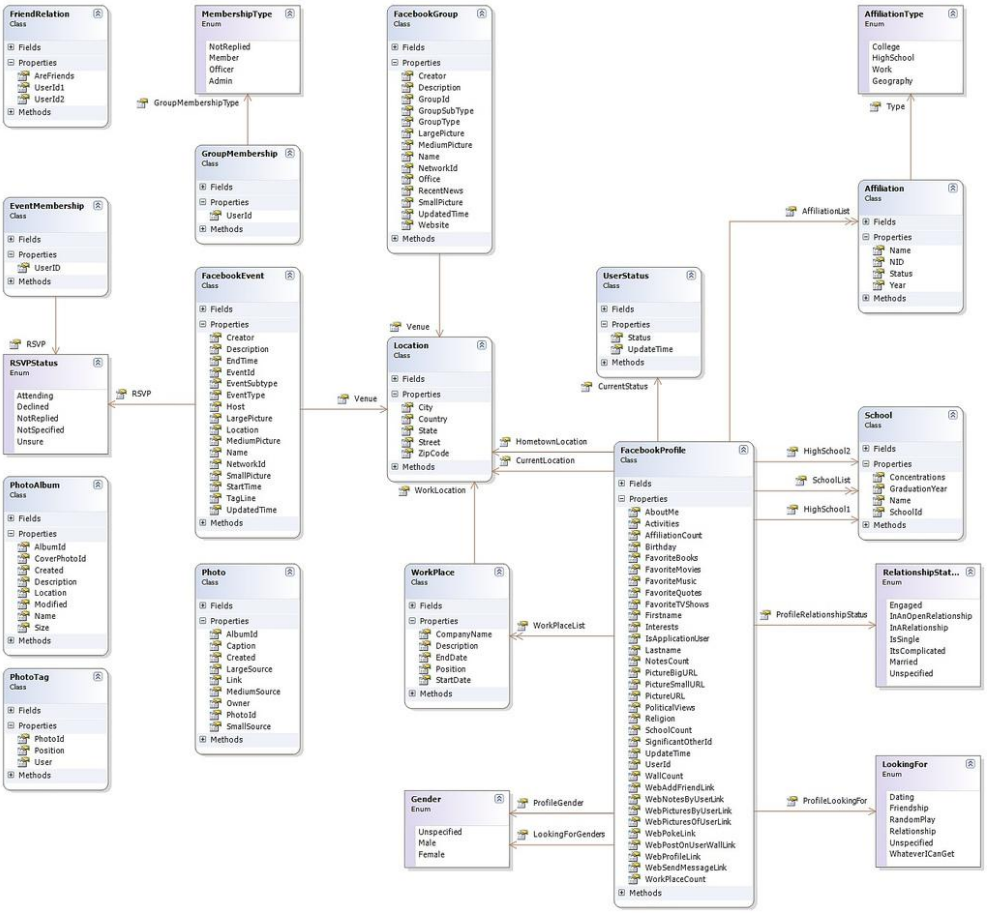


Figure 15: Facebook database representation

Facebook social graph visualization tools

Nexus displays the friend links and commonalities between your friends. It finds how your friends are connected, and which interests they share - even find which of your friends you have the most in common with.

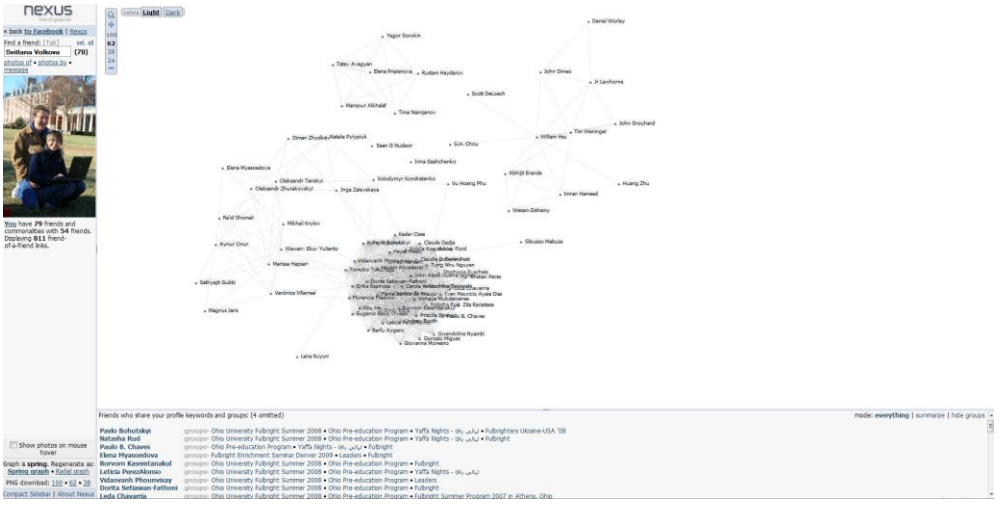


Figure 16: User's profile visualization tool (http://nexus.ludios.net/view/)

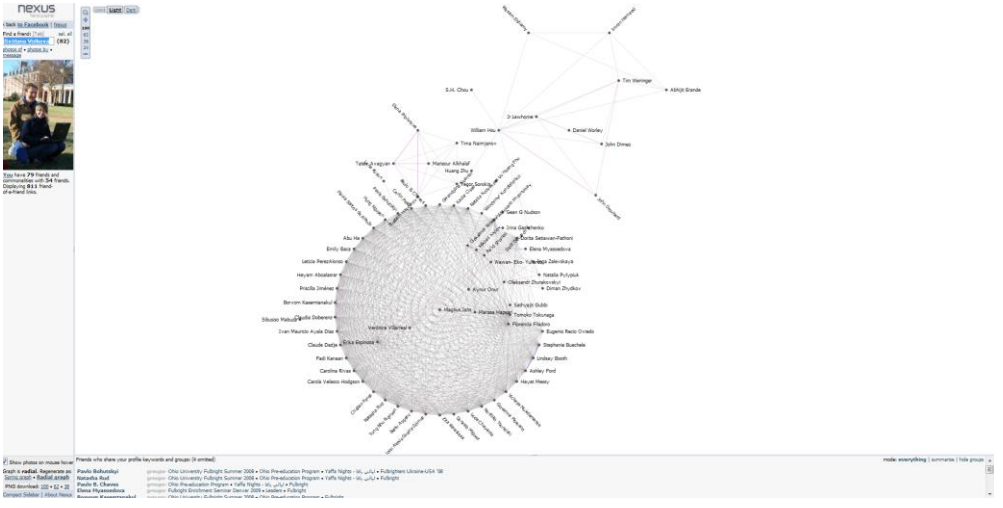


Figure 17: User's profile visualization tool (http://nexus.ludios.net/view/) on radial graph

In the tradition of creating useful and engaging network visualization tools for various systems such as Amazon and Google, **TouchGraph** is now leveraging the power of the booming Facebook platform with its new TouchGraph Facebook Browser. The tool allows users to see how their friends are connected, and who has the most photos together. Users can also explore their own personal networks by graphing photos from anyone's album, or view the connections between members of a group. The interactive tags feature seems also quite interesting.

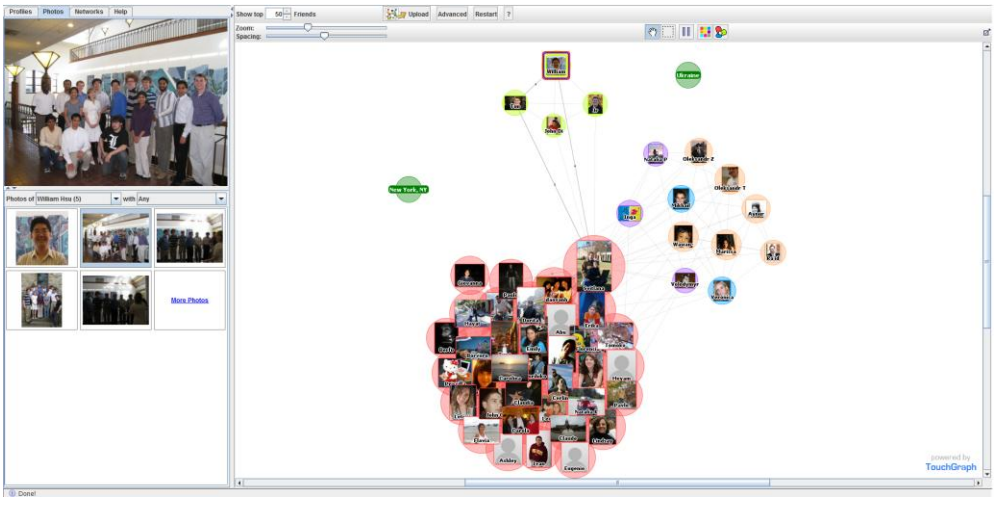


Figure 18: User's profile visualization tool (http://www.touchgraph.com/)

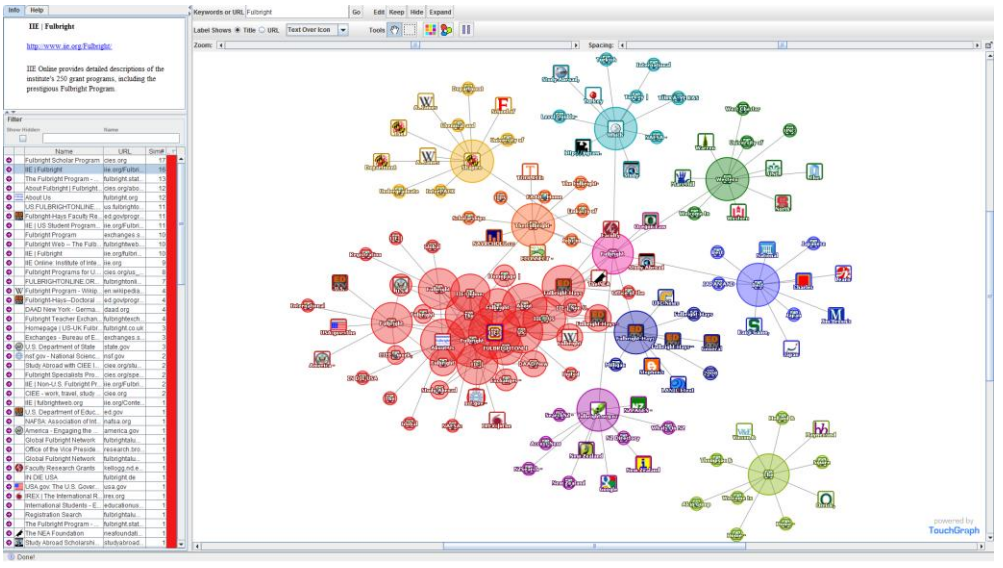


Figure 19: Visualization tool (http://www.touchgraph.com/) for Fulbright Organization based on Google search

Alternative Tools for SNA

It is essential to use number of different social network analysis computer programs. All of these are available in the computer labs. All but UCINET are freely available on the web.

- UCINET, available in computer labs and for purchase from Analytic Technologies at: http://www.analytictech.com/
- PAJEK, available to download at http://vlado.fmf.uni-lj.si/pub/networks/pajek/default.htm
- NETDRAW, available to download at: http://www.analytictech.com/
- STOCNET, available to download at: http://stat.gamma.rug.nl/stocnet/ (see also http://stat.gamma.rug.nl/snijders/siena.html)
- For information about InFlow analysis tools, articles and diagrams http://www.orgnet.com/inflow3.html

The International Network for Social Network Analysis (INSNA) is the international and interdisciplinary professional association for people interested in social network research. Its website (http://www.sfu.ca/~insna/) is a wonderful source of information and resources on social networks, including links to many informative sites and to social network computer programs and data.

The listserv, SOCNET, is the main on-line forum for discussion of current topics on social networks. Information on how to join is available through the INSNA site (see above) or at: http://www.heinz.cmu.edu/project/INSNA/socnet.html

Connection is INSNA's newsletter/ informal journal. It is available through the INSNA website or directly at: http://www.sfu.ca/~insna/.

Centrality is a newon-line journal devoted to relationship capital management. http://blog.visiblepath.com. Steve Borgatti's web page is a nice source of introductory material and handouts on various topics on social networks: http://www.analytictech.com/networks/

Conclusions

The prediction task for previously unobserved links in social networks was discussed in the project. The concept of social network was defined as well as social graph representation.

The related works in research area were analyzed. Existed approaches were compared: supervised vs. unsupervised, single table data representation as feature vector vs. relational data mining etc. In case of interpretation link prediction task as classification task the application of range of induces were digested such as: J48, OneR, IB1, Logistic, NaiveBayes as well as other available approaches for resolving given task e.g. SVM, GP, BN, PRMs etc.

Mining tasks in network-structured data were analyzed. Mathematic representation for unobserved link prediction task in social networks was given as well as deep classification and description of measures for link prediction approaches.

The experiment was planned based on crawling technique with application of free open-source SQL full-text search engine Sphinx. The proposed training corpus is Facebook social network web site. The database for selected data was represented in SQL format. The visualization tools for social networks graph representation were described.

Consequently, research in a number of scientific fields have demonstrated that social networks operate on many levels, from families up to the level of nations, and play a critical role in determining the way problems are solved, organizations are run, and the degree to which individuals succeed in achieving their goals.

References

[Ada03a] L.A. Adamic, E. Adar: Friends and Neighbors on the Web. Social Networks, 25 (3) 2003, 211-230.

[Ada03b] L.A. Adamic, O. Buyukkokten, E. Adar: A social network caught in the Web. First Monday, 8 (6) June 2003
http://www.firstmonday.org/issues/issue8_6/adamic/.

[Ada05] L.A. Adamic, E. Adar: How To Search a Social Network. Social Networks 27 (3) 2005, 187-203.

[Bar03] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. Physic A: Statistical Mechanics and its Applications, 311(3-4):590–614, August 2002.

[Bri98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30(1–7): 107–117, 1998.

[Cha07] Duen H. Chau, Shashank Pandit, Samuel Wang, and Christos Faloutsos. Parallel crawling for online social networks. In WWW '07: Proceedings of the 16th international conference on World Wide Web, pages 1283–1284, New York, NY, USA, May 2007. ACM.

[Dˇze01] S. Dˇzeroski and N. Lavraˇc, editors. (2001) Relational Data Mining. Springer, Berlin.

[Far07] Facebook's Zuckerberg uncorks the social graph http://blogs.zdnet.com/BTL/?p=5156.

[Get02] L. Getoor, N. Friedman, D. Koller, & B. Taskar. Learning Probabilistic Models of Link Structure. Journal of Machine Learning Research, 2002.

[Get05] L. Getoor & C. P. Diehl. Link mining: a survey. SIGKDD Explorations, Special Issue on Link Mining, 7(2):3-12.

[Gol05] J. Golbeck: Computing and Applying Trust in Web-Based Social Networks. Ph.D. Thesis, University of Maryland, College Park,

http://www.cafepress.com/trustnet.20473616.

[Han05] Hanneman, R. and Riddle, M. (2005). Introduction to social network methods. Riverside, CA: University of California, Riverside (published in digital form at http://faculty.ucr.edu/~hanneman/).

[Hsu07] Hsu W. H., Lancaster J., Paradesi M. S. R., & Weninger T. "Structural Link Analysis from User Profiles and Friends Networks: A Feature Construction Approach", In Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2007). Boulder, CO, March 26-28, 2007.

[Jeh02] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In KDD, pages 538–543, 2002.

[Jen98] Jensen, D., & Goldberg, H. (1998). AAAI fall symposium on AI and link analysis. AAAI Press.

[Kat53] Leo Katz. A new status index derived from sociometric analysis. Psychometrika, 18(1):39–43, March 1953.

[Kum03] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins: Social networks: From the web to knowledge management. Chapter 17 in Web Intelligence, N. Zhong, J. Liu, Y. Yao (eds), Springer-Verlag, 2003, 367-379.

[Lav94] N. Lavraˇc and S. Dˇzeroski. (1994) Inductive Logic Programming: Techniques and Applications. Ellis Horwood, Chichester, 1994. Freely available at http://www-ai.ijs.si/SasoDzeroski/ILPBook/.

[Lib04] D. Liben-Nowell, J. Kleinberg: The Link Prediction Problem for Social Networks, CIKM 2003, ACM Press, 2004, 556-559.

[New01] M. E. J. Newman. Clustering and preferential attachment in growing networks, April 2001.

[Nov03] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In CIKM ’03: Proceedings of the twelfth international conference on Information and knowledge management, pages 556–559, New York, NY, USA, 2003. ACM Press.

[Sar05] P. Sarkar & A. Moore. Dynamic social network analysis using latent space models. SIGKDD Explorations, Special Issue on Link Mining, 7(2):31-40.

[Smi01] Samuel R. Smith "Postmodernism is dead, now what? Distributed culture and the rise of the network age".

[Spe05] E. Spertus, M. Sahami, O. Buyukkokten: Evaluating similarity measures: a largescale study in the Orkut Social Network. KDD 2005, ACM Press, 2005, 678-684.

[Tas03] Link Prediction in Relational Data, B. Taskar, M. F. Wong, P. Abbeel and D. Koller. Neural Information Processing Systems Conference (NIPS03), Vancouver, Canada, December 2003.

[Var08] Sourabh Vartak, 2008. A Survey on Link Prediction, State University of New York, Binghamton, NY – 13902, U.S.A.

[Was94] S. Wasserman, K. Faust: Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences), Cambridge University Press, New York, 1994.

[Wen08] Weninger T., "Link Discovery in Very Large Graphs By Constructive Induction using Genetic Programming", Masters Thesis. Kansas State University, Manhattan, KS. Dec. 2008.

[Xia08] A survey "A Survey on Link Prediction Models for Networked Data", 2008 Department of Computer Science and Engineering, HKUST.

Journals on Social Networks

Elsevier

- http://www.elsevier.com/wps/find/journaldescription.cws_home/505596/description

WWW International Network for Social Network Analysis

- http://www.insna.org/

Social Network Analysis

- http://semanticstudios.com/publications/semantics/000006.php

Trust in Web Based Social Networks

- http://trust.mindswap.org/cgi-bin/relationshipTable.cgi