2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing

Node-pair Feature Extraction for Link Prediction

Teshome Feyessa, Marwan Bikdash and Gary Lebby
Department of Electrical and Computer Engineering
North Carolina A&T State University Greensboro, North Carolina 27411
tmfeyess@ncat.edu, bikdash@ncat.edu, lebby@ncat.edu

Abstract— **In social networks, one of the most essential problems is predicting existence or formation of a link between nodes. Traditional structure based link predicting algorithms leverage node properties such as degree and centrality and relation between nodes such as common neighbors and paths. Most of these algorithms rely on visibility of the entire or significant portion of the network structure; node centrality and shortest distance between nodes often require global knowledge. This work uses a back propagation neural network to predict existence or emergence of a link between pairs of nodes using node pair properties such as reciprocity, transitivity and shared neighbors. A limited network visibility by individual nodes is assumed, hence the size of the node pair feature vector varies with the given visibility range. This approach is tested on a large social object centered trust network where visibility is limited to two hops, 828 accurate predictions out of 1000 pair of nodes is achieved.**[1]

I. Introduction

Study of social networks has been a center of research with an ever growing demand for data mining of social network data from the business sector. Among various important concepts in social networks, centrality and link prediction stand out as the most significant ones. The two concepts are not entirely exclusive; in fact various structural link predicting algorithms make use of centrality [1, 2, 3]. 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.

Link prediction algorithms generally leverage the structural property of the graph representing the network to implement a statistical relational learning. Other approaches include modeling by studying the underlying process such as object centered sociality or other physical properties of the network, such as temporality and spatiality.

The problem of link prediction is often addressed by developing measures to analyze proximity of nodes in the network. Most social network researchers are concerned with modeling the evolution of a social network using features that are inherent to the network itself using structural properties of the graph one way or another. Algorithms based on the concept of shared neighborhood include common neighbors, Jacard's coefficient, Adamic/Adar and preferential attachment [1, 2]. Katz, page-rank and simRank use ensemble of all paths to determine proximity between two nodes [1, 4, 5]. Low-rank approximation [6] and graph clustering [7] are also used to refine some of the other structural methods. Techniques such as supervised learning [8] and probabilistic approaches [2, 4] are used to complement the proximity formulation from graphs.

The nature of the relationship modeled by the network, such as friendship network, author collaboration network, employee network and the likes can give further insight for prediction. Such attributes are often applicable when there is more than one type of relationship represented by links in the network [9] or when network is unavailable but some other information regarding the nodes is available; for instance, cold start method [10]. The cold start method builds network by first generating a network as a probabilistic graph and later refining using graph theoretic measures. Another probabilistic method [4] uses maximum entropy Markov random model to estimate the joint probability of the nodes comprising the central neighborhood set between nodes whose link is to be determined.

Combining time series models based on temporal evolution of links with static link prediction algorithms is shown to improve the prediction model [5]. There are also spectral-based link prediction approaches that use low-rank approximation, spectral transformation, spectral regression and a combination of these [11, 12].

The performance of link prediction approaches is evaluated by constructing a receiver operating characteristics (ROC) curve as most classification algorithms. The x-axis is the ratio of false positive to the total negatives in the test set while the y-axis is ratio of true positive to total positives in the test set. A false positive is when the classifier classifies a given event negative event incorrectly as positive. In this case, positive indicates existence of a link. The area under the ROC curve (AUC) is the standard measure to assess the quality of the curve. A random algorithm has an AUC measure of 0.5 while perfect algorithm has an AUC measure of 1 [5, 13]. For instance by combining 'co-occurrence probability' and topological measures the authors of [4] have achieved an AUC of 0.87 and precision of 57%.

For large complex networks, having a global knowledge of the network is impossible and at times processing the entire network is impractical. In these instances, being able to determine existence of links between nodes using only local information becomes very important. When the view of the network is limited, global measures such as path and centrality will not be available for prediction, unless defined locally. This work pursues an alternative path for this problem. The concepts of clustering [14], neighborhood and distribution

[1] This work is supported by Pennsylvania State University and The Defesne Threat Reduction Agency under contract DTRA01-03-D-0010/0020 and Subcontract S03-34.

978-0-7695-4578-3/11 $26.00 © 2011 IEEE
DOI

sequence [15] and assortativity [16] are extended to determine proximity of two nodes among other node-pair properties. The node pair properties are compiled as a feature vector for randomly selected node pairs whose connectivity is known. Then a supervised learner is used to learn the deduction of existence of connection from a given feature vector.

The rest of this paper is organized into two parts. The first section presents node level features that are used to describe relationship between pairs of nodes whose link existence is in question. The later section discusses the design of the learner and presents the performance of the prediction approach. The test network used in this work is obtained from an online who-trust-whom network of general consumer review site Epinions.com [17].

II. Node-pair Feature Extraction

Locality here is described by the number of hops that are visible to a node. The number of hops is the least number of arcs between two nodes disregarding direction. The least locality considered here is a degree, i.e., a node can only see its neighbors, nodes that are one hop away. The largest locality, an entire view of the graph, is the diameter of the graph. Node pair features are used to describe the relationship between two nodes. The exact nature of the relationship between the node-pair features and existence of a direct link between the nodes is not clear, hence, here, a black box modeling using a supervised learner is used to map the features to prediction.

A. Reciprocity and Clustering

Link reciprocity states that if there is a direct link from node n_1 to node n_2 there is a reciprocal link. Reciprocity is a very common phenomena in social networks such as friendship and coauthorship networks. These types of networks also have a tendency to form tightly connected, clustered, neighborhood. The clustering coefficient of a node i, C_i, is the ratio between number of triangles i it belongs to and the number of triangles that could have been formed with i as a vertex [14]. Let A be the adjacency matrix of a network where a_{ij} is 1 if there exists a direct link from node i to j and d_i be the degree of node i, then the clustering coefficient of i is

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

In undirected graphs there is only one type of triangle, and the order of nodes is irrelevant. In directed graphs four types of triangles can be formed if a link between n_1 and n_2 exists as shown in figure 1. Some of these triangles represent transitivity property of the network while others represent cycles. The total number of these triangles is also the number of 2-hop paths between nodes $n1$ and $n2$ in undirected networks. The clustering coefficient between node pairs can be defined in a similar fashion. Let d_{ij} be the the minimum of d_i and d_j, then C_{ij}, the clustering between nodes i and j, is the ratio of triangles both nodes i and j are vertices of and the number of triangles they could have shared.

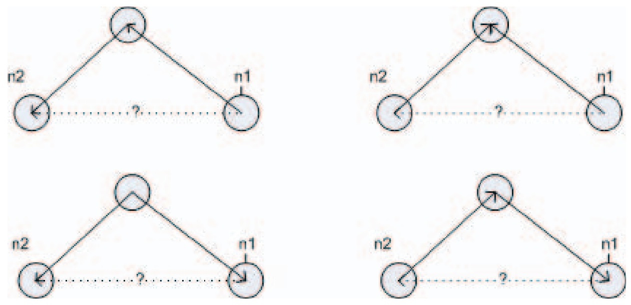

Fig. 1. Four types of triangles that can be formed by existence of direct link

$$C_{ij} = \frac{\sum_{h\neq i,j} a_{ih}a_{hj}}{d_{ij}} \tag{2}$$

Equation 2 shows the node-pair clustering is ratio of the common neighbors of nodes i and j to the minimum degree of both. Both clustering coefficient and reciprocity can be determined with only a single hop visibility in the network. Since most social networks are directed, the clustering coefficients can be separated into four coefficients based on link directions as shown in figure 1

B. Degree Correlation

Another node pair property considered is the node degree correlation between two nodes, assortativity. Most networks are known to exhibit either assortative or disassortative mixing. Assortativity, r, of a network is a Pearson correlation coefficient of the degrees at either ends of a link. Generally, for an observed undirected network with M edges, assortativity is evaluated as [16]

$$r = \frac{\sum_i j_i k_i/M - \left[\sum_i \frac{1}{2}(j_i+k_i)/M\right]^2}{\sum_i (j_i^2+k_i^2)/2M - \left[\sum_i \frac{1}{2}(j_i+k_i)/M\right]^2} \tag{3}$$

where, $i = 1, 2, ..., M,$ is a link in the network, j_i and k_i are the degrees of nodes on either side of link i. Equation (3) returns a coefficient that is an average over all links. The local assortativity between the node pairs on either end of link i can be approximated by dropping all summations and plugging $j_i + k_i$ for M, the assortativity based node-pair feature is thus

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

r_{jk} ranges $0 \leq r_{jk} \leq 1$ while $-1 \leq r \leq 1$. Evaluating r_{jk} presupposes the existence of link i and hence the learner learns which assortativity values actually exist in the network and which doesn't. Computing the local assortativity only requires a single hope visibility in the network.

C. Common Neighbors

Clustering and reciprocity take shared neighbors between the node-pairs into consideration. When considering non-immediate neighbors in networks that has visibility range of greater than one hop, further information can be extracted

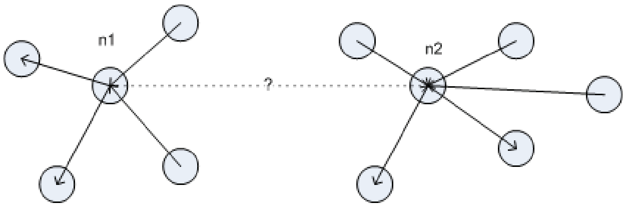

Fig. 2. Mixing of different degree nodes

by observing shared nodes in the neighborhood of each of the nodes. The neighborhood of node $n, \Gamma(n)$, is the set of nodes adjacent to node n and the i^{th} neighborhood of a node n, $\Gamma^i(n)$ is the set of nodes that are adjacent to all nodes that are in the $(i-1)^{th}$ neighborhood and nodes that do not belong to any of the previous neighborhoods, i.e., $\Gamma^1(n)$ to $\Gamma^{i-1}(n)$ [15]. For instance in figure 3 the dark nodes are 1^{st} neighbors to node 0 while the lighter nodes are 2^{nd} neighbors. For nodes n_1 and n_2, the intersection of $\Gamma^i(n_1)$ and $\Gamma^j(n_2)$ for i, j less than the visibility range of the network, can be extracted as a node-pair feature. Clustering features capture the intersection of $\Gamma^1(n_1)$ and $\Gamma^1(n_2)$, hence in common neighbors features only other combination of neighborhoods should be considered. The possible features due to common neighbors are, intersection of $\Gamma^1(n_1)$ and $\Gamma^2(n_2)$, $\Gamma^2(n_1)$ and $\Gamma^1(n_2)$, ..., $\Gamma^m(n_1)$ and $\Gamma^m(n_2)$, where m is the network visibility range. Second common neighbors are equivalent to clustering between neighbors of original node-pair, i.e., second neighbor clustering between nodes i and j, S_{ij}, is clustering between nodes h and k such that $a_{ih} = 1$ and $a_{kj} = 1$

$$S_{ij} = \frac{\sum_{h,k,u \neq i,j,a_{ih=1},a_{kj}=1} a_{hu} a_{uk}}{d_{hk}} \tag{5}$$

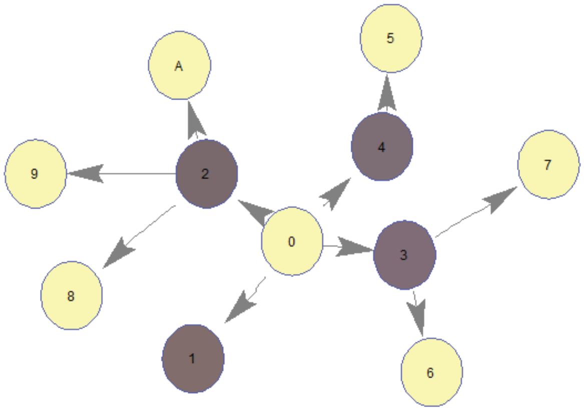

Fig. 3. First and second neighborhood of a node

III. Link Prediction

A. Learning

The node to node features discussed here can be generated for a generic network and do not depend on the social object. But a social network is always tied together by one or more objects of sociality. For instance, in the online who-trust-whom network of general consumer review network used in this work the consumer reviews are the objects of sociality. Members of the site, nodes of the network, interact with each other using the ”trust” relationship, links of the network.

This work treats the link prediction problem as a classification of node-pair features into one of two classes, existence or non-existence of a link between the node-pair. For this experiment, node visibility is limited to only two hops, therefore the neighborhood of node n is limited to $\Gamma^1(n)$ and $\Gamma^2(n)$. Table 1 summarizes the available node pair features of a network represented by adjacency matrix A for classification describing the relationship between nodes i and j

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

Table 1. Summary of node-pair features

The assortativity can also be split into two by defining in-degree and out-degree mixing, but here the common general assortativity is implemented. Therefore, there are a total of 10 input features for the classifier whose value ranges is $[0, 1]$. More number of features can be extracted, especially from common neighbor features, but most of the information is already captured in the ten features discussed and hence doing so will only be redundant and computationally cumbersome.

Once the features are determined training and testing data set are extracted from who-trust-whom network, mentioned in the introduction. To avoid outliers in the extracted data, samples are taken only from the largest connected component of the network for this experiment. These features are fed to a neural networks. The network used in this work is, a 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 ROC curve.

B. Results

To study the impact of the node-pair features on the performance of the prediction, combinations of the 11 features are used to train the BPN. The four combinations generated are:

- reciprocity and clustering,
- reciprocity, clustering and assortativity

- reciprocity, clustering and 2^{nd} clustering
- all extracted features

Figure 4 shows the ROC curves for all combinations of features. The AUC and accuracy of the learner is summarized in table 1. Let reciprocity be r, clustering be c, assortativity be a and 2nd clustering be n.

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

Table 2: Performance summary

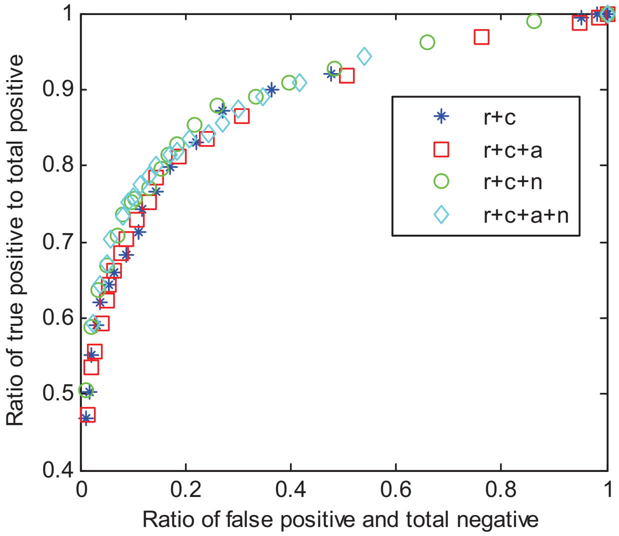

Fig. 4. ROC curves of BPN classifiers. The curves are generated for different combinations of node to node features fed to the BPN learner.

The AUC and accuracy achieved by this method is in par with most other approaches discussed in the introduction section which don't limit node to node visibility. Seemingly, assortativity is the least important feature in determining existence of a link between two nodes. Evaluating the assortativity of the network used in this experiment shows that indeed the network has an almost neutral degree mix. The actual value is -0.01.

IV. Conclusion

We have developed a method that uses only local information, as far as the neighbors of neighbors of a node, to characterize quantitatively the relationship between pairs of nodes whose connection is to be determined. Using local graph features in a network circumvents the barrier that arises due to lack of knowledge of global topology. 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.

Bibliography

[1] D. Liben-Nowell and J. Kleinberg, The Link Prediction Problem for Social Networks, CIKM 12 (2003)

[2] H. H. Song, T. W. Cho, V. Dave, Y. Zhang and L. Qiu, Scalable Proximity Estimation and Link Prediction in Online Social Networks, IMC (2009)

[3] M. E. J. Newman, Scientific Collaboration Networks II. Shortest paths, weighted networks, and centrality Physical Review E 64 (2001)

[4] C. Wang, V. Satuluri, S. Parthasaraty, Local probabilistic Models for Link Prediction, IEEE-ICDM 7 (2007)

[5] Z. Huang, D. Lin, The time series link prediction problem with applications in communication surveillance, Informs J. on Computing, 21 (2009)

[6] C. Fang, J. Lu, A. Ralesu, Graph Spectra Regression with Low-Rank Approximation for Dynamic Graph Link Prediction, NIPS (2010)

[7] Z. Huang, Link Prediction Based on Graph Topology: The Predictive value of the generalized clustering coefficient, SIGKDD, 12 (2006)

[8] M. Hasan, V. Chaoji, S. Salem and M. Zaki, Link Prediction using Supervised Learning, SDIM (2006)

[9] E. Zheleva, L. Getoor, J. Golbeck and U. Kuter, Using Friendship Ties and Family Circles for Link Prediction, SNA-KDD, 2 (2008)

[10] V. Leroy , B. Cambazoglu, F. Bonchi, Cold Start Link Prediction, KDD (2010)

[11] J. Kunegis, J. Lommatzsch, Learning Spectral Graph Transformations for Link Prediction, ICML, 26 (2009)

[12] J. Kunegis, D. Fay, C. Bauckhage, Network Growth and the Spectral Evolution Model, CIKM (2010)

[13] Bradley A., The Use of the Area Under the ROC curve in the Evaluation of Machine Learning Algorithms, Pattern Rcognition, 30 (1997)

[14] G. Fagiolo, Clustering in complex directed networks, Physical Review E 76 (2007)

[15] D.J. Watts, ”Small worlds: the dynamics of networks between order and randomness”, Princeton University Press, 1999

[16] M. E. J. Newman, Assortativity Mixing in Networks, Phys. Rev. Lett. Vol. 89, no. 20, 2002

[17] M. Richardson, R. Agrawal, P. Domingos, Trust Management for the Semantic Web, ISWC (2003)