Using network evolution theory and singular value decomposition method to improve accuracy of link prediction in social networks

Qinxue Meng  Paul J. Kennedy

Centre for Quantum Computation and Intelligent Systems  
Faculty of Engineering and Information Technology  
University of Technology, Sydney  
PO Box 123 Broadway NSW 2007 Australia  
Email: Qinxue.Meng@student.uts.edu.au  
Paul.Kennedy@uts.edu.au

Abstract

Link prediction in large networks, especially social networks, has received significant recent attention. Although there are many papers contributing methods for link prediction, the accuracy of most predictors is generally low as they treat all nodes equally. We propose an effective approach to identifying the level of activities of nodes in networks by observing their behaviour during network evolution. It is clear that nodes that have been active previously contribute more to the changes in a network than stable nodes, which have low activity. We apply truncated singular value decomposition (SVD) to exclude the interference of stable nodes by treating them as noise in our dataset. Finally, in order to test the effectiveness of our proposed method, we use co-authorship networks from an Australian university from between 2006 and 2011 as an experimental dataset. The results show that our proposed method achieves higher accuracy in link prediction than previous methods, especially in predicting new links.

Keywords: Link prediction, network evolution, Katz measure, singular value decomposition, social network analysis.

1 Introduction

Social networks are a type of graph structure whose nodes represent people or other entities embedded in a social context and relationships are indicated by links between nodes. There are many natural examples of social networks (Scott & Carrington 2011). This paper considers academic collaboration networks. In this type of network, researchers are linked via media, for example co-authored publications, projects or grants. Meanwhile, abstract concepts can also work as media to establish connections and relationships in social networks. Examples of useful abstract concepts to illustrate this point are common interests and friendship.

A common feature of social networks is that they are generally dynamic rather than static compared with other networks, such as gene networks. They grow and change quickly over time by adding, strengthening or removing links and nodes. Under these circumstances and since examples of social networks abound in the real world, identifying patterns of evolution and predicting changes in links becomes an urgent and fundamental question that is still not well understood.

Traditionally, the analysis of social networks (Carrington et al. 2005) focuses on a single snapshot of a network and generally makes predictions based on local neighbourhood algorithms: the more common neighbours two nodes have, the higher the possibility they have to develop a link in the future. An obvious weakness of this is in treating every node equally. As a result, accuracy of prediction is not very satisfactory (Liben-Nowell & Kleinberg 2007).

The reason for this is straightforward: link prediction in social networks is a non-stationary problem. The underlying relationships between people are ever changing. Changes in relationships in social networks are the result of many contributions as the nodes are people. Consequently, during the evolution of networks, not all nodes tend to generate new links over time. For example, nodes labeled as “stable” may maintain their old relationships for a long time without any changes while “active” nodes tend to establish new links with others even if there is a long path between the nodes in the existing network. Finally, “regular” nodes have a tendency to expand their social circle, but more slowly.

Based on these considerations, we propose a new method to improve the accuracy of link prediction. We group and label nodes in a social network based on their level of behaviour during network evolution. Then singular value decomposition (SVD) is applied to assess the contribution of different groups to the results of link prediction.

In our experiment, we choose co-authorship networks as our experimental dataset. We form prediction problems in two categories: (i) predicting all links and (ii) predicting new links. Co-authorship networks are applicable to both of these kinds of problem. Unlike friendship networks, co-authorship networks can be seen as a kind of temporary network, where coauthored publications are transitory. Whether two authors are connected depends on predicting their co-publications in the future. Meanwhile, co-authorship networks represent academic collaboration and therefore it is meaningful to predict new links.

The experimental dataset is gathered from the research management database of an Australian university, namely the University of Technology, Sydney, over the five year period from 2006 to 2010. The experimental results show that the accuracies of predictors using a succession of years of data for network evolution exceed those on a single snapshot and that smoothing of the network using truncated singular value decomposition improves prediction.
The rest of this paper is organized as follows. Section 2 gives a brief literature review of some previous research on link prediction. The experimental dataset and methodology are given in sections 3 and 4, followed by the experimental results in section 5 and conclusion and future work in section 6.

2 Related work

2.1 Models of network evolution

There have been a number of methods proposed and analyzed in the area of network evolution. The majority focus on changes in topological structure of networks. As graphs can be abstracted as a probability distribution, one typical approach (Guo et al. 2007) models dynamic graphs with a sequential linear model, namely Markov chains, to form a series of probability distributions (Snijders 2002). Another similar method (Song et al. 2009) introduces a Bayesian network to model the process of network evolution. Probabilistic methods of building network models such as those outlined above are generally based on a single snapshot of a network. Cortes et al. (2001) proposed the idea of summarizing a series of networks from different time points to represent the network model. This model is extended by Sharan and Neville (2008) to develop a representation that captures the changes in social networks by adding different kernel functions to summarize the weights of links. In our experiment, we follow their idea of a summary representation but propose a practical algorithm to summarize the weights of links from different time periods.

2.2 Link prediction

The link-prediction problem for social networks can be described from a data-mining point of view in the following way: given a snapshot of a social network at time t, the goal is to predict new links that will be added to the network during the interval from time t to a given future time t’. This problem can be viewed as a simple binary classification problem. That is, for any two potentially linked objects o_i and o_j, predict whether l_{ij} is 1 or 0. However, it is still difficult to completely solve this binary classification problem. Current research aims to measure the degree of similarity and closeness between two target nodes. In terms of the networks this means that not only should they be similar to each other, but they must also be reachable through the network. In other words, the closer and more similar are they, the higher the possibility they have to be connected in the future. Generally, approaches to addressing this problem come from two sides: the attribute information of nodes and the structural properties of social networks.

In the first of these approaches, attribute information is used for link prediction. Popescul and Ungar (2003) introduces a structured logistic regression model that can make use of relational features to predict the existence of links on citation datasets from CiteSeer. In that experiment, link prediction on citation networks is cast into a citation recommendation problem by gauging the similarities between target publications and existing publications. However, those methods have the limitation that attributes of nodes can only reveal similarity with other nodes, but fail to take into account the concept of “distance”. For example, two people may be quite similar to each other in terms of habits, interests and backgrounds. However, they cannot be friends if they are located far from each other in geography as they have no chance to meet.

Links are predicted on the basis of graph proximity measures. Among different proximity measures, nearest neighbourhood algorithms have been widely applied. In the experiment of Murata and Moriyasu (2008), they introduced a weighted common neighbour approach and compared its prediction results with common neighbour and the Jaccard coefficient method. Unfortunately, their experimental results have low accuracies, all under 50%. The reason for this is because neighbourhood algorithms can make correct predictions when two nodes are quite close to each other. Some relationships, such as friendship, co-authorship and co-citation are transitive: nodes may be connected in the future if there is a path among them, but these methods may ignore this.

The second approach takes into account the structure of social networks, namely the “distance” between nodes, when devising measures of closeness. One famous approach is the Katz measure (1997) which defines a measure that directly sums over the collection of paths, exponentially damped by their length so as to count short paths more heavily. Another approach uses a random walk (Rudnick & Gaspari 2004), to calculate the number of steps from a start point s to the end point e. Because the time to arrive is not in general symmetric, a common way to detect closeness from this probabilistic approach is to consider the commute time \( C_{s,e} = T_{s,e} + T_{e,s} \), where \( T_{s,e} \) and \( T_{e,s} \) are times to move from start to end and end to start respectively. Liben-Nowell and Kleinberg (2007) present an experiment comparing predictors on large co-authorship networks. Their work suggests that information about future interactions can be extracted from the network topology alone and that subtle measures for detecting node proximity can outperform more direct measures. However, the results show that among all predictors, the best is the Katz measure, combining neighbourhood and distance concepts. However, the derived accuracy of 16% is still quite low.

The above research confirms that current methods of link prediction have much room for improvement. In our experiment, we gather information concerning the evolution of a co-authorship network at an Australian university so as to treat nodes differently. Through observing their past actions, we label nodes by their activity and those nodes with high values in activity have a high possibility to connect with others. In order to improve the accuracy further, matrix theory is applied to detect major patterns for predicting results.

3 Dataset

The dataset used for analysis, visualization and explanation of our approach is from the research master enterprise (RME) database of the University of Technology, Sydney. It is a collection of over 60000 records covering all faculties. It provides information about all publications from UTS researchers during the recent six years (2006-2011) including journals, conference papers and proceedings, chapters and books.

The dataset consists of researcher, publication and researcher-publication relationship files for each year, containing 56,535 records in total (Table 1). In our experiments, datasets from 2006 to 2010 are used to summarize the evolution of the co-authorship network on which link prediction is based and the data from
2011 is used to estimate the accuracy of predictors. The advantage of this data source over other scientific bibliographic databases such as CiteSeer and the computer sciences bibliographic data source, DBLP, is its integrity. Although the number of records in our dataset is relatively small compared to those scientific bibliographic databases, it contains all types of authors and publications (see Table 2). This means it is easier to capture the process of network evolution. With this information, it is straightforward to detect the activities of researchers.

4 Methodology

This section describes our experimental methodology. The whole process consists of four parts: building co-authorship networks, involvement of network evolution, link prediction and determining the activities of nodes.

4.1 Building co-authorship networks

Firstly, we apply a simple model to build the co-authorship network so as to reveal research collaboration throughout the university. Two researchers are connected if they work on the same publication (see Figure 1). In our graph, co-authorship is a type of mutual relationship and therefore the network is undirected and weighted. The weight for a link represents the number of publications coauthored by the two researchers. Publications are not explicitly represented in the co-authorship network.

4.2 Involvement of network evolution

According to the above model, let us consider a given network \( G = (N, L, W) \) where \( N \) is the set of nodes while links are in the set \( L \subseteq N \times N \) and \( W \) contains the weights of links. Conceptually, network evolution can be represented by a series of networks \( \Omega = (G_1, \ldots, G_t) \), so that \( G_t = (N_t, L_t, W_t) \) represents the network at time \( t \). Then the link prediction problem is to predict \( G_{t+1} \) based on networks \( G_1, \ldots, G_t \). Each graph can be represented by an adjacency matrix \( X \) of size \( n \times n \) where \( n \) is the number of nodes and \( X_{ij} = X_{ji} \) is equal to the weight of the link between node \( i \) and \( j \).

In order to involve the whole history of network evolution into the link prediction problem, the series of matrices associated with the network evolution should be collapsed into a single matrix. The simplest way to do this is to add the matrices together, taking care to ensure that the nodes are the same over the entire time series. However, this does not generally work well in most cases because recent links should have a stronger influence on prediction of future links compared to older links. Here, we propose an approach to collapsing adjacency matrices of networks, motivated by Sharan and Neville’s research (2008) on network evolution, where the link structure is damped backward in time according to the following model

\[
N_{t+1} = N_t \cup N_2 \cup \ldots \cup N_t \\
L_{t+1} = L_t \cup L_2 \cup \ldots \cup L_t \\
W_{t+1} = \sum_{i=1}^{t} \left( \frac{t}{i} \right)^{t-i} W_i
\]

(1)

where \( W_i \) is the adjacency matrix for network \( G_i \). Clearly, this method gives greater weight to more recent links.

4.3 Link prediction

Predicting future links is generally based on closeness between nodes. In order to evaluate the effectiveness of different lengths of sequences on network evolution, we adopt several approaches for measuring closeness between nodes.

1. Common neighbours (CN) (Liben-Nowell & Kleinberg 2007) ranks pairs using the number of common neighbours.

\[
closeness_{CN}(x, y) = |\Gamma(x) \cap \Gamma(y)|
\]

(2)

where \( \Gamma(x) \) denotes the set of neighbours of node \( x \) in a given network and \( |\Gamma(x)| \) is the number of elements in the set.

2. Adamic/Adar (AA) (Adamic & Adar 2003) is an extension of common neighbours that adds weights to the neighbours.

\[
closeness_{AA}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}
\]

(3)
where \( z \) indexes the common neighbours of nodes \( x \) and \( y \).

3. Katz measure (1953) uses a combination of both neighbourhood and distance.

\[
closeness_K(x, y) = \sum_{l=1}^{\infty} \beta^l \cdot |\text{paths}^{l}_{x,y}| \tag{4}
\]

where \( \text{paths}^{l}_{x,y} \) is the set of all \( l \)-length paths from \( x \) to \( y \) and \( \beta > 0 \) is a scale parameter for the function. \( \beta \) can be regarded as a radius around the target node and predictors can only fetch neighbours from inside the circle formed by this radius. A very small \( \beta \) yields predictions much like common neighbours as the long paths contribute very little to the sum. Due to the fact that a network without node attributes can be represented by its adjacency matrix, the corresponding matrix for Katz’s closeness is defined, using the approach in Liben-Nowell & Kleinberg (2007), as

\[
P = (I - \beta A_{t+1})^{-1} - I \tag{5}
\]

where \( A_{t+1} \) is the matrix containing information for the evolution of the network and \( I \) is the identity matrix.

However, values of some elements in \( P \) are quite small, which means that they have low probability of happening in the future. Sometimes, links associated with large values in \( P \) may not happen either as some activity nodes have many close relationships with others. Here we propose a method to improve the prediction accuracy.

The link prediction matrix \( P \) can be written as \( P = UDV^T \) by the singular value decomposition (SVD). In this formula, \( R \) is the rank of \( P \). Matrices \( U \) and \( V \) are orthogonal matrices of size \( M \times R \) and \( D \) is a diagonal matrix of singular values \( \sigma_1 > \sigma_2 > \ldots > \sigma_R > 0 \). According to truncated matrix theory

\[
P \approx U_k D_k V_k^T \tag{6}
\]

where \( U_k \) and \( V_k \) comprise the first \( k \) columns of \( U \) and \( V \), and \( D \) is the \( k \times k \) principal submatrix of \( D \).

As a result, a matrix of predicted links can be written as \( P' = U_k D_k V_k^T \) which not only contains the main features of matrix \( P \) but also excludes the noise that is entailed in the components greater than \( k \).

4.4 Determining activities of nodes

In our experiment, the parameter \( k \) is calculated in the following way. We classify nodes in the networks into one of three groups: stable, active and regular. This is done based on their contribution to network evolution as follows.

Stable nodes tend to maintain their relationships as time goes. They may be viewed as inactive or nodes that have already left the network. This means they have low probability of generating new connections with others. Identifying them in advance can promote the efficiency of link prediction.

Active nodes represent those who developed new connections more often during past growth of networks. They are sometimes the main contributors to network evolution. Therefore, they are of extreme importance in link prediction. Connection to isolated groups may start from them.

Regular nodes are the remainders, containing the rest.

With this definition, the activities of researchers in the network are defined using

\[
F(x_k) = \sum_{i=1,j=i+1}^{n} \left( \frac{L(x_k, t_j) - L(x_k, t_i)}{n} \right) \tag{7}
\]

where \( x_k \) is a node in the network and the function \( L(x_k, t_j) \) is the number of edges of node \( x_k \) at time point \( t_j \). If the value of activity is not positive, nodes are labeled as “stable”. The value of \( k \) used to form the truncated matrix should be the sum of active nodes and regular nodes or the sum of the active nodes only.

5 Results

We use the University of Technology, Sydney, research dataset to assess the performance of the link predictors discussed in section 4.3. All experiments were performed using R (Ripley 2001).

5.1 Building co-authorship networks

We first build the initial co-authorship networks which contain both authors and publications with connections through co-authorship. From Figure 2, it is clear that the numbers of authors and co-authorship
Figure 3: The compacted co-authorship network at UTS from years 2006 to 2010
Table 3: Co-authorship Networks from 2006 to 2010

<table>
<thead>
<tr>
<th>Year</th>
<th>Nodes</th>
<th>Links</th>
<th>Density of Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>2763</td>
<td>7115</td>
<td>0.0019</td>
</tr>
<tr>
<td>2007</td>
<td>2972</td>
<td>8196</td>
<td>0.0019</td>
</tr>
<tr>
<td>2008</td>
<td>3148</td>
<td>9486</td>
<td>0.0019</td>
</tr>
<tr>
<td>2009</td>
<td>3231</td>
<td>9243</td>
<td>0.0018</td>
</tr>
<tr>
<td>2010</td>
<td>3548</td>
<td>12267</td>
<td>0.0019</td>
</tr>
</tbody>
</table>

relationships grew significantly especially from 2009 to 2010. However the number of publications remained stable, ranging from between 1941 and 2047.

In order to investigate academic collaboration further, we link authors directly by perceiving publications as media according to the network transformation model (see Figure 1). Summary information for the researcher-only co-authorship networks is listed in Table 3. In this table, network densities are calculated as the ratio of the number of edges to the number of possible edges. Although the number of nodes and links increases annually, the network densities stay at almost the same value. This suggests that academic collaboration as measured in co-authorship is relatively stable at UTS.

5.2 Involvement of network evolution

Compacting network evolution into a single matrix is a key step in our experiments. As links from different times contribute differently in link prediction, we assume that the weights, or importance, of links should decrease as time goes on. Using the evolution model presented in Section 4.2, we compacted co-authorship networks from 2006 to 2010 and visualized the resulting network in Figure 3. It shows clear research groups or clusters although there are some isolated authors as well.

5.3 Determining activities of nodes

Determining the activities of nodes in a compacted co-authorship network is a prediction problem that uses the truncated SVD method to exclude noise and improve the accuracies of link prediction. Activities are calculated based on the proposed method described in Section 4.4 and are normalized to lie in [−1, 1]. The distribution of activities is illustrated in Figure 4, which plots the activity of each node ranked from lowest to highest. Stable nodes are those where the tangent of the activity function is less than 0 and takes up more than 50% of the authors (3643 active nodes). These are the nodes who left the network during 2006–2010 or who maintain relationships with the same set of partners. Regular authors are those nodes where the tangent is positive but below the red line which is set at an activity level of 0.74. This threshold was set empirically at the point where the curve grew quickly. The nodes above the threshold are active authors who have the highest activities among the three groups.

5.4 Link prediction

Unlike friendship, co-authorship is a transitory relationship. As a result, link prediction in co-authorship networks can be thought of as two problems: predicting new links and predicting all links (including the existing ones). The first problem aims to find whether two authors who have never cooperated with each other will work together in the future based on their closeness. The second problem not only detects the possible new links but also checks whether co-authored nodes will maintain their current relationships into the future. Here, we first apply several baseline algorithms: common neighbourhood (CN), Adamic/Adar (AA) and the Katz method, and compare their predicted results with and without involvement of network evolution and the truncated SVD method. We have observed that different values of β can affect the performance of the Katz measure greatly. Based on preliminary tests on our datasets with β ∈ [0.005, 0.5], we empirically set β = 0.001 in equations (4) and (5).

We perform link prediction using the three link predictors on different snapshots of the co-authorship networks and compare the predicted results with the real situation in 2011. Accuracies measured by the ratio of correct predictions to all predictions are given in Table 4. These results clearly show, as would be expected, that the more recent networks achieve higher prediction accuracies. Furthermore, involving the evolution of the network by using a compacted network leads to higher accuracies both for predicting new links and all links. Predicting new links is still a challenge as the accuracies are significantly lower than those for predicting all links.

Next, the truncated SVD method is applied to further improve prediction accuracies for the compacted co-authorship network. Results, presented in Table 5, confirm that using the truncated SVD method after identifying the activities of authors can enhance the prediction accuracies in general. For the problem of predicting new links in particular, determining the number of active nodes boosts the prediction accuracies greatly. For prediction of all links, using both active and regular nodes, but not the stable ones, also enhances the prediction accuracy but not by as much. In addition, stable nodes contribute less to the network evolution.

6 Conclusions

This paper describes a novel method for improving the accuracy of link prediction. On the one hand, involving network evolution rather than using a single snapshot of networks provides a more suitable input...
Table 4: Performance of predictors on different co-authorship networks. Figures are the ratio of correct predictions to all predictions. CN = common neighbours, AA = Adamic/Adar, Katz = Katz closeness measure.

<table>
<thead>
<tr>
<th>Co-authorship network</th>
<th>New Links</th>
<th>All Links</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CN</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>0.56</td>
<td>0.54</td>
</tr>
<tr>
<td>Network of 2006</td>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>0.61</td>
<td>0.62</td>
</tr>
<tr>
<td>Network of 2007</td>
<td>0.14</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>0.69</td>
<td>0.66</td>
</tr>
<tr>
<td>Network of 2008</td>
<td>0.17</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>0.72</td>
<td>0.68</td>
</tr>
<tr>
<td>Network of 2009</td>
<td>0.21</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>0.78</td>
<td>0.76</td>
</tr>
<tr>
<td>Network of 2010</td>
<td>0.33</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>0.84</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Table 5: Performance of predictors using truncated SVD method on compacted co-authorship network from 2006 to 2010. CN, AA and Katz are the closeness measures used, node type is the nodes used and $k$ is the parameter for truncated SVD.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Node Type</th>
<th>$k$</th>
<th>New Links</th>
<th>All Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN</td>
<td>Active</td>
<td>278</td>
<td>0.72</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Active and regular</td>
<td>2873</td>
<td>0.49</td>
<td>0.90</td>
</tr>
<tr>
<td>AA</td>
<td>Active nodes</td>
<td>278</td>
<td>0.75</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Active and regular</td>
<td>2873</td>
<td>0.53</td>
<td>0.88</td>
</tr>
<tr>
<td>Katz</td>
<td>Active nodes</td>
<td>278</td>
<td>0.91</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>Active and regular</td>
<td>2873</td>
<td>0.77</td>
<td>0.92</td>
</tr>
</tbody>
</table>

for the prediction methods. On the other hand, the truncated SVD approach can exclude noise and effectively improve prediction accuracies for new links after identifying the amount of activity of authors in networks. The methods were evaluated on six years of research collaboration data from an Australian university. Results demonstrated significant improvements in accuracy using our approach.

In the future, we would like to extend our experimental datasets to larger ones from other sources to test the effectiveness and robustness of our proposed method further. We plan also to develop quantitative approaches to set the threshold between active and regular nodes.

References


