Social Networks 42 (2015) 72–79

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

A note on using the adjusted Rand index for link prediction in networks

Michaela Hoffman [a,*], Douglas Steinley [a,1], Michael J. Brusco [b,2]

[a] University of Missouri, Columbia, USA
[b] Florida State University, USA

ARTICLE INFO

Keywords:
Adjusted Rand index
Link prediction
Missing links
Network analysis

ABSTRACT

As network data gains popularity for research in various fields, the need for methods to predict future links or find missing links in the data increases. One subset of the methodology used to solve this problem involves creating a similarity measure between each pair of nodes in the network; unfortunately, these methods can be shown to have arbitrary cutoffs and poor performance. To address these shortcomings, we use the adjusted Rand index to create a similarity measure between nodes that has a natural threshold of zero. The effectiveness of this method is then compared to a number of other similarity measures and assessed on a variety of simulated data sets with block model structure and three real network data sets. Under this particular formulation of the adjusted Rand index, information is also provided on dissimilarity. As such, we then go on to test its use for detecting incorrect links in network data, highlighting the dual use of the approach.

1. Introduction

Network (e.g., graph) data consist of nodes and edges. The nodes can represent any object of interest and the edges are the links between these nodes signifying some form of a connection. This type of data is used in many fields to represent different structures from the neural networks and food chains of the biological sciences (Zhu et al., 2007) to analysis of terrorist networks (e.g. Krebs, 2002; Ressler, 2006). Networks can be either directed, where a link from one node to another is not necessarily reciprocated, or undirected, where two nodes are either connected or not. Links can also be weighted, providing a measure of strength for each connection. For this study, we focus only on undirected/unweighted graphs.

One increasingly popular subset of networks used in fields such as sociology and psychology is social networks (see Wasserman and Faust, 1994, for an overview). These networks are comprised specifically of people or groups and the connections between them, and, more commonly, when measured over time, dynamic networks that add and drop nodes/links as the network evolves. Given the complexity of social networks and their tendency to change over time, detecting links not present in the *observed* data (e.g., the links are truly present, but were not observed) is an important question. This is divided into two problems that would have similar solutions: (a) the missing link problem and (b) the link prediction problem. Searching for missing links is attempting to identify any links that should be in the data set that were not observed, whether from data measurement errors or unknown information. The link prediction problem focuses on links that might occur in the future based on the observed network. Examples often seen in link prediction literature use a network of authors collaborating on papers, where the goal would be to predict collaborations for future papers (e.g. Newman, 2001; Shibata et al., 2012).

In looking at the graph structure of social networks, it is common to see clustering (e.g., tightly connected subgroups that are well separated from each other). For instance, in Newman's 2001 analysis of citation networks, the clustering coefficient indicated that the networks were highly clustered. Similarly, other research has focused on identifying clusters in social networks (e.g. Brusco and Steinley, 2007; Mishra et al., 2008; Steinley et al., 2011; Duan et al., 2012). While the present investigation is not concerned with identifying an underlying cluster structure, the presence of clustering will be taken into account in order to evaluate potential impact among the different link prediction measures.

The purpose of this study is to improve the existing link prediction methodology by proposing the application of a proximity measure, the Adjusted Rand Index (ARI), previously unused for this

* Corresponding author at: 107 Psychology Building, Columbia, MO 65211, USA. Tel.: +1 573 882 6860.
E-mail addresses: mmhb7f@mail.missouri.edu (M. Hoffman), steinleyd@missouri.edu (D. Steinley), mbrusco@cob.fsu.edu (M.J. Brusco).
[1] Address: 19 McAlester Hall, Columbia, MO 65211, USA.
[2] Address: 600 W. College Avenue, Tallahassee, FL 32306, USA.

http://dx.doi.org/10.1016/j.socnet.2015.03.002
0378-8733/

type of analysis. Section 2 of this paper describes a selection of existing methods for prediction links in networks using proximity measures (Section 2.1). This is followed by a description of the proposed method, using the ARI (Section 2.2), as well as a number of similar proximity measures used for comparison (Section 2.3). In Section 3 the procedures used to test and compare the methods of Section 2 are described: a simulation study for link prediction (Section 3.1), a comparison on real network data (Section 3.2) and finally a simulation considering an additional application of the ARI in detecting incorrect links (Section 3.3). The results of these analysis are described in Section 4 and Section 5 contains concluding remarks.

2. Methods for link prediction

There are many methods that have been created to predict links (see Liben-Nowell and Kleinberg, 2007), involving a wide range of techniques from those rooted in the most basic graph theory to more complicated machine learning algorithms and those that require additional substantive information about the network. In this study, we focus on the easiest to implement (and perhaps the most widely used, for that reason): proximity measures. Proximity measures indicate a "distance" between each pair of nodes and solely rely on the structure of the network. This distance (or function thereof) represents how likely it is that a link should be present between two nodes. Common practice is to rank the ensuing distances (usually from smallest to largest – or most similar to least similar) and a certain number or percentage of the best scores of the unlinked pairs are taken to be the "predicted" edges. Naturally, one potential shortcoming of such an approach is determining the percentage of links that should be imputed; unfortunately, thus far, few guidelines have been provided.

2.1. Current methods

To introduce general notation for a network, assume n_i is the ith object, and $i = 1, \ldots, N$. Frequently, network information is collected in the adjacency matrix, $\mathbf{A}_{N\times N} = \{a_{ij}\}$, where $\mathbf{A}$ is a square $N \times N$ matrix that represents the connections between all pairs of objects. Specifically, $a_{ij} = 1$ if n_i is connected to n_j; otherwise, $a_{ij} = 0$. This information can be summarized in row vectors corresponding to the ith and jth row in $\mathbf{A}$, $\mathbf{a}'_i$ and $\mathbf{a}'_j$, (where each are $1 \times n$ vectors, respectively).

If one considers the similarity between n_i and n_j, there are four possible "states" for binary vectors that can be computed when considering the mutual pattern of connections between n_i and n_j when related to all of the other remaining nodes. The counts of these states gives us four quantities: (a) the number of links to other nodes n_m that n_i and n_j have in common, (b) the number of times that n_i has links to other nodes, n_m, and n_j does not, (c) the number of times that n_i does not have links to other nodes n_m, but n_j does, and (d) the number of times that links to other nodes n_m are mutually absent for n_i and n_j. For any pair of row vectors, $\mathbf{a}'_i$ and $\mathbf{a}'_j$, in the adjacency matrix the four quantities can be quickly computed from the following inner products

$$a = \mathbf{a}'_i\mathbf{a}_j$$
$$b = \mathbf{a}'_i(\mathbf{1} - \mathbf{a}_j)$$
$$c = (\mathbf{1} - \mathbf{a}_i)'\mathbf{a}_j$$
$$d = (\mathbf{1} - \mathbf{a}_i)'(1 - \mathbf{a}_j)$$

Often, these four values are collected in a simple 2×2 contingency table, as indicated in Table 1. The values of this contingency table are combined in various ways to form similarity or proximity measures. Four proximity measures currently used for link prediction studies are described in the next section, three of which are calculated from the proximity table.

Table 1
The contingency table for each pair of nodes i and j across all other nodes m.

Node j	Node i	
	Linked to Node m	Not Linked to Node m
Linked to Node m	a	b
Not Linked to Node m	c	d

2.1.1. Graph distance

The first, and likely simplest approach, is what is commonly referred to as the "graph distance" in the link prediction literature. In graph theory, the graph distance would be referred to as the geodesic distance – the shortest path between a pair of nodes. As an example, for nodes n_i and n_j, the geodesic distance represents the shortest path, and if the graph is connected, the geodesic distance will be a metric (Harary, 1969). The geodesic distance between a pair of nodes can be calculated by numerous algorithms (for example, Dijkstra, 1959); however, the simplest approach (although not the fastest) is the method of powers.

The power of a graph's adjacency matrix, $\mathbf{A}^p$, gives the number of walks of length p between all pairs of nodes. Consequently, the geodesic distance matrix, $\mathbf{D}^{(GD)}$ has the entries $d_{ij}^{(GD)} = p$ where p is the smallest p such that $a_{ij}^p > 0$. Calculating the geodesic distance for each pair of nodes creates a set of $N \times (N-1)$ distances,[3] where the predicted (e.g., imputed) links are going to between nodes that have the smallest values of $d_{ij}^{(GD)}$ conditional on not already being directly connected within the observed network.

2.1.2. Common neighbors

Another commonly used method is common neighbors (CN), represented by

$$CN_{ij} = |\Gamma(i) \cap \Gamma(j)| = a \tag{1}$$

where CN_{ij} is the common neighbor score between nodes i and j, $|\bullet|$ is the cardinality of $\bullet$, and $\Gamma(i)$ and $\Gamma(j)$ are the set of nodes that node i and node j share direct links with, respectively. Thus, Eq. (1) represents the cardinality of the intersection of those two sets (e.g., the number of nodes to which each i and j are mutually linked). This measure is equivalent to a in Table 1. An intuitively appealing aspect of this measure is that it would also be monotonically related to the fraction of possible triangles containing nodes i and j that are actually formed, a common measure of density used to determine clustering (Wasserman and Steinley, 2003).

2.1.3. Jaccard's coefficient

If CN is thought of as the probability of forming a triangle containing nodes i and j across the entire graph, then Jaccard's coefficient (JC, Dowton and Brennan, 1980; Steinley, 2004) is a conditional probability of forming a triangle across the local neighborhoods of node i and j

$$JC_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i) \cup \Gamma(j)|} = \frac{a}{a+b+c} \tag{2}$$

where the denominator is now the cardinality of the union between the two neighborhood sets (or the sum of a, b,and c as defined in Table 1). In addition to being used to predict links, Jaccard's

[3] The diagonal elements of $\mathbf{D}^{(GD)}$ are automatically set equal to zero.