Published in final edited form as:

Soc Networks. 2015 July ; 42: 72–79. doi:10.1016/j.socnet.2015.03.002.

A Note on Using the Adjusted Rand Index for Link Prediction in Networks

Michaela Hoffman^A,

University of Missouri – Columbia

Douglas Steinley^B, and

University of Missouri – Columbia

Michael J. Brusco^C

Florida State University

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.

Keywords

adjusted Rand index; link prediction; missing links; network analysis

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

^ACorresponding author: 107 Psychology Building, Columbia, MO 65211, USA, mmhb7f@mail.missouri.edu +1 (816) 977 – 3453.

^B19 McAlester Hall, Columbia, MO 65211, USA, steinleyd@missouri.edu

^C600 W. College Avenue, Tallahassee, FL 32306, USA, mbrusco@cob.fsu.edu

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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 & 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. M. E. J. 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 & 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 type of analysis. Section 2 of this paper describes a selection of existing methods for prediction links in networks using proximity measures (2.1). This is followed by a description of the proposed method, using the ARI (2.2), as well as a number of similar proximity measures used for comparison (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 (3.1), a comparison on real network data (3.2) and finally a simulation considering an additional application of the ARI in detecting incorrect links (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 & 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 i^{th} object, and $i = 1, \dots, N$. 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 i^{th} and j^{th} 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

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

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.

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 $N \times (N - 1)$ distances¹, 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

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

where CN_{ij} is the common neighbor score between nodes i and j , $|\cdot|$ 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, equation (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 & 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 & Brennan, 1980; Steinley, 2004) is a conditional probability of forming a triangle across the local neighborhoods of node i and j

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

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 coefficient has long been used in cluster validation studies (Milligan et al., 1983; Dubes, 1987; Steinley, 2004).

¹The diagonal elements of $\mathbf{D}^{(GD)}$ are automatically set equal to zero

2.1.4 Preferential Attachment—Preferential attachment has been used as a solution to model growth (Mitzenmacher, 2004) and relies on the rationale that the probability of a new link being connected to node i is proportional to the size of node i 's neighborhood. If node neighborhoods are thought to be independent, then the probability of two nodes having a missing link (or forming a link in the future) is proportional to the product of their individual neighborhood sizes,

$$PA_{ij} = |\Gamma(i)| |\Gamma(j)| = (a + c)(a + b) \quad (3)$$

In 2004, Liben-Nowell and Kleinberg compared these proximity measures, along with a few others. They found that all were able to outperform a random predictor. More specifically, they found that *CN* was among the class of best performing link prediction methods (e.g., there was no significant difference between *CN* and the other best performing methods) regardless of whether they were similarity measures, as presented here, or more complicated, algorithmic approaches. The authors also found that *GD* was among the class of worst performing methods, with *JC* not being significantly different than *CN* or *GD*. Consequently, the set of similarity measures chosen here represent the entire performance range of link prediction methods; surprisingly though, the best of the procedures they investigated was only correct 16% of the time, demonstrating that more precise methods are needed.

2.2 Proposed Method: Adjusted Rand Index

As mentioned above, Jaccard's coefficient has long been used to compare two partitions of data. In some sense, the set of links can be thought of as the definition of a partition for node n_i , where every node divides the remaining set of $n - 1$ nodes into two clusters: (a) those that node n_i are connected to and (b) those that node n_i are not connected to. Subsequently, the similarity between clusterings could be calculated for each pair of nodes, n_i and n_j . Steinley (2004) compared Jaccard's method to several other methods of cluster agreement/similarity, finding that the adjusted Rand index (*ARI*, Hubert & Arabie, 1985) is the best performing method for comparing two partitions. The *ARI* obtains a value of unity when the two partitions perfectly agree (in the current situation, when two nodes are connected to the exact same set of other nodes) and a value of zero when there is only chance agreement between the partitions. At face value, while the *ARI* might be known to outperform other methods of partition agreement, it still suffers from the problem of having to establish an arbitrary threshold to decide if a pair of nodes should be linked or remain unlinked.

Fortunately, we can take advantage of a clever proof given by Warrens (2008) that demonstrates that the *ARI* is equivalent to Cohen's κ (1960) when there are two partitions and the cluster labels are known *a priori*. Both of those conditions are met as the two partitions are defined by the edge sets of node n_i and node n_j , while the labels are static across all nodes as either "connected" or "not connected" (avoiding the problem of unknown cluster labels discussed in Steinley, 2004). Under these two constraints, and given the values in Table 1, the *ARI* (and κ) can be computed as

$$ARI_{ij} = \frac{2(ad - bc)}{(a + b)(b + d) + (a + c)(c + d)}, \quad (4)$$

In this application, *ARI* gives each potential link a score between -1 and 1 , where a score greater than zero indicates that the probability of a link being there is greater than random chance. Using zero as a defined cutoff value to detect links is an inherent benefit of *ARI* over other methods where the cutoff must be determined ahead of time.

2.3 Related Indices

While the number of similarity measures that have been used in the literature has been relatively few (primarily, the four mentioned above), making connections to the literature for comparing clusterings opens up a wide range of potential indices for comparison (see Albatineh & Niewiadomska-Bugaj, 2011). As such, we included several other indices in our comparison to aid in the probing of why successful indices are indeed successful.

2.3.1 Maxwell & Pilliner—An index very similar to the Cohen's Kappa formulation of the adjusted Rand index, is the index proposed by Maxwell and Pilliner 1968:

$$MP_{ij} = \frac{2(ad - bc)}{(a + b)(c + d) + (a + c)(b + d)}. \quad (5)$$

Both (4) and (5) are corrected for chance in the same manner in the numerator; however, they are normalized slightly differently in the denominator, with the difference between the two denominators if (5) was subtracted from (4) being $(b - c)^2$. Thus, (4) and (5) will be equivalent when the number of misattributions are symmetric between n_i and n_j , as captured in the quantities b and c . Otherwise, when they are not symmetric (4) will be a smaller value than (5) because its denominator is larger. Depending on the degree of asymmetry, there could be a large difference between the two. Including *MP* in the study allows for an assessment of how critical the patterning of mismatches are in predicting links.

2.3.2 Phi Coefficient—The Phi-coefficient is the equivalent of the product-moment correlation as applied to binary data.

$$\Phi_{ij} = \frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}} \quad (6)$$

Effectively, this is using the angle between $\mathbf{a}'_i$ and $\mathbf{a}'_j$ to determine the similarity between a pair of nodes.

2.3.3 Rand Index—Including the Rand (1971) index allows for the probing of how crucial the correction for chance is in determining the predictive relationship between a pair of nodes in a network. The formula for the Rand index is given as

$$R_{ij} = \frac{a + d}{a + b + c + d}, \quad (7)$$

which for two binary vectors is equivalent as the simple matching coefficient when a , b , c , and d are defined as above.

2.3.4 Kulczynski—Kulczynski (1928) proposed an interesting measure based on conditional probabilities:

$$K_{ij} = \frac{1}{2} \left(\frac{a}{a + b} + \frac{a}{a + c} \right). \quad (8)$$

This measure can be interpreted as the conditional probability that a link is present in one observation given that the link is present in the other observation. This measure is averaged over both observations acting as predictors, and K ranges between zero and unity.

3. Method

3.1 Detecting Missing Links - Simulation

For our simulation we generated data sets in the form of a cohesive subgroup blockmodel based on structural equivalence to best represent the clusters found in real social networks. The density *within* blocks determines the number of links within each cluster of nodes. Three levels of within density were used: 60%, 75% and 90%. The density *between* blocks is the amount of links between different clusters, also set at three levels: 10%, 25% and 40%. The size of the data sets were varied to contain different numbers of clusters: 2, 4, or 6 (each containing 20 nodes, resulting in networks of size 40, 80, and 120, respectively).

For additional analysis, we consider completely random graphs, where there is no cluster structure. To create these we generated graphs where the density between and density within were both set at 50%, so between any pair of nodes on the entire graph there is a 50/50 chance of a link. We varied the percentage of links removed and number of clusters at the same levels as the first analysis.

After the graphs were created, each was split by removing a percentage of the links randomly to form a test data, leaving the training data. This was done on three different levels, removing 10%, 30% and 50% of the links. With this setup, the test data are the “missing links” that we are trying to predict by analyzing the training data sets they were removed from. Ten data sets were created at every combination of levels of cluster number, density within and between and number of links removed (a completely crossed design), resulting in

$$3(\text{Number of Clusters}) \times 3(\text{Density Within}) \times 3(\text{Density Between}) \times 3(\text{Links Removed}) = 81$$

different scenarios.

The *ARI* gives a definite cutoff for the number of links it predicts, since any score greater than zero indicates a predicted link and less than zero is indicative of no link. This is not the case with the other indices; meaningful cutoff values are not provided. Typically, to compare the links “predicted” by these algorithms, the top scores are chosen as the links, with the number corresponding to the number of links originally removed. Consequently, this would bias the study against *ARI*, in that *ARI* has to ascertain automatically how many links it needs to impute and evaluate; whereas, the other methods are given the target amount prior to assessment of accuracy².

Due to this problem, we assessed the effectiveness of each algorithm using the area under the ROC curve statistic (Fawcett, 2006; Hanley & McNeil, 1982). The area under the ROC curve (AUC) is a statistic that is often used to evaluate the performance of link prediction methods. It is calculated by first creating the ROC curve which compares the rate of true positives to false positives at increasing cutoffs levels for the scores. The area under this curve is taken to find the AUC statistic. The AUC can be interpreted as the probability that the algorithm will detect a true link over a false link. This can evaluate how much better than chance this method is predicting, because an AUC of .5 is expected from random guessing. Because it evaluates each method this way, it allows all methods to be compared without the issue of arbitrarily selecting a cutoff value to help mitigate the bias mentioned above.

3.2 Detecting Missing Links - Real Networks

Next we considered the application of these measures on three real world data sets. This analysis was set up similarly to the simulations above. Each real network was randomly split into a test and training set with the percentage of links removed again at three levels: 10%, 30% and 50%. The random split at each level of links removed was repeated 10 times. The same set of 10 proximity measures were applied to the training sets, using the AUC to evaluate the performance.

To further examine the utility of the *ARI*, we consider the classification of all the potential links (all dyads unlinked in the test data) using the defined cutoff value given by *ARI*. Here we recorded the rates of false positives, cases where *ARI* gives a score greater than 0 to a link not in the test data set, false negatives, where κ gives a score less than 0 to a link present in the test data, and correct classifications.

We consider three network data sets commonly found in social network analysis literature:

Zachary's Karate Club is a network of associations between members of a collegiate club collected in the 1970's (Zachary, 1977). After a disagreement the club split into two groups, creating a cluster structure in the network. This network is small with $n = 34$ nodes and $m = 78$ edges.

²If this were a simulation evaluating the performance of clustering algorithms, it would be equivalent to forcing one method to both determine the number of clusters and perform a clustering while the other methods were provided the number of clusters prior to the analysis.

The *American College Football* network is comprised of the Division IA college football teams from the 2000 season, where an edge indicates the two teams played against each other in a game (Girvan & Newman, 2002). It contains $n = 115$ teams, belonging to 12 conferences (defining the cluster structure).

Coauthorships in Network Science was compiled by Newman in 2006 using references from two network science review articles with a few more authors added by hand (M. E. Newman, 2006). We use an unweighted version of this network, and remove authors with zero neighbors before proceeding with the algorithm. The final network used is comprised of 1461 authors.

We note that due to the sparsity of links in the Karate and Coauthorship networks, the division of the networks' links into training and test sets often resulted in nodes in the training set with no recorded neighbors. In these cases, to remain as close as possible to a realistic link prediction situation, these nodes were removed from both the test and training sets before calculating any measures.

3.3 Detecting Incorrect Links

We conducted a second set of simulations to look into another potential use of the *ARI* index: namely, finding incorrect links in the data. In the link prediction context, the *ARI* values at or below 0 indicate that a link should not be added. If links present in the test data are assessed with *ARI*, it may point out incorrect links. To test this, data sets were simulated in the same manner as before with the same levels of between and within density and number of clusters. Then random edges were generated, adding links to 10%, 30% and 50% of the unlinked dyads. *ARI* values were then calculated for the random links added and the links already present in the test set.

4. Results

4.1 Detecting Missing Links - Simulation

The AUC scores for the simulation can be found in Table 2. First we consider *ARI*'s performance over the various generated networks. To determine which conditions had the most influence on the variation of κ , an analysis of variance was conducted, the results of which are presented in Table 3. As Steinley (2006) indicates, the significance of particular factors are often irrelevant in simulation studies because the sample size can be arbitrarily increased to make all of the factors significant; rather, the more informative measure is the percentage of variance accounted for by each factor (indicated by the traditional measure of effect size, η^2). The most influential factor was the density between clusters, where an increased density between clusters resulted in worse link prediction. This can be expected because, as the between cluster density approaches 50%, it is just as likely that there is a connection between clusters as not and the network is approaching a more random state. The second most influential factor was the within cluster density, following the expected pattern of improved performance as cluster density increased. The third most influential factor was the number of clusters; while it was seven times less influential than between cluster density, it does follow the same patterns found in general cluster analysis studies (Steinley, 2003, 2006) where it is more difficult to correctly resolve the partition structure as the number of

clusters increase. The least influential factor was the percentage of links removed; however, it followed the predicted pattern of worse performance the more links were removed. All together, these four main effects account for 92% of the variance, and when evaluated as one, indicate that the most important predictor of successful link prediction using similarity indices is an underlying, discernible cluster structure.

Analysis of the potential for interactions among the network characteristic values can be seen in Table 3. Looking at the η^2 values it can be seen they had little influence. The two way interaction between the number of clusters and the density within has a small effect ($\eta^2 = .02$). Considering interaction plots it seems the two cluster case is slightly more influenced by the density between than the four or six cluster cases.

As can be seen in Table 2, *ARI* outperforms the other indices in eight of the conditions and in the overall average. An ANOVA comparing all nine link prediction approaches revealed there is a significant difference between their AUC scores ($\eta^2 = .34$). A follow-up Tukey's HSD found graph distance and preferential attachment significantly different from all other methods. Common neighbors was significantly different than all but Rand. The remaining seven measures were not found to be significantly different from each other.

To further investigate the relationship between these algorithms, a rank correlation was performed on the scores they produced. Table 4 shows the correlations, averaged over all iterations. Considering the four existing node similarity measures, Cohen's *ARI* is most closely related to Jaccard's Coefficient and common neighbors. This is an expected result given that Jaccard's Coefficient and Cohen's *ARI* use the common neighbors information in their scores. *ARI* has near perfect agreement with Maxwell and Pilliner and ϕ . These three include all the same information but in slightly different relationships. Jaccard's Coefficient and the Kulczynski measure are in near perfect agreement with each other. These are also highly correlated with the other three, the differences likely due to the fact that they include d in their formulas.

Finally, the situation where the density of links within blocks and between blocks were set to be equal was investigated, corresponding to a random graph model where the probability of a link was held constant for all pairs of nodes, regardless of group membership. The average AUC results for all link prediction approaches across all conditions were between .49 and .51, which is as anticipated since the expected value of the AUC is 50% for guessing.

4.2 Detecting Missing Links - Real Networks

Comparing the performance of the distance measures on the three different real networks, we find varying results. Table 5 shows the average AUC results for each of the algorithms with different percentages of links removed. None of the algorithms are able to do very well finding the links removed from the Karate data set. The Rand index's AUC scores indicate it performed worse than random chance. Preferential attachment, graph distance, and common neighbors had the best AUCs. The top performing algorithms on the Football network were ϕ , Rand, MP and *ARI*. Tukey's HSD shows no significant difference between these four, and that they are all significantly different from the other six, with preferential attachment showing a particularly poor performance. All of the algorithms were able to provide good

results on the Coauthorship network. Here with the highest AUC scores were given by graph distance, but these were only significantly different from preferential attachment, ϕ and Rand.

Since *ARI* provides a cutoff value built into the algorithm that the other methods lack, we examine these results alone in Table 6. For all three networks, the predictions are correct more than half the time. All of the data sets show the same patterns with increasing percentages of links removed: false negative rates decrease and the other three measures increase.

4.3 Detecting Incorrect Links

Table 7 shows the average percentage of correctly and incorrectly classified links. The main focus for this simulation is on the correctly classified random links, which are those that were given a negative *ARI*. As with link prediction, *ARI* seems to do best identifying random links when there are fewer clusters. This effect was smaller for the correct classification of true links (with a positive *ARI*). As more random links were added the performance did decrease as expected. However, this effect is small in the classification of both random and true links; for random links and true links.

Looking at the cluster structure of the data, the density within clusters also seems to be an important factor for the random links. It is easiest to detect incorrect links when the original clusters are better defined within. Density between, however, does not seem to have this same amount of influence. The ANOVA indicated the effect was significant but less strong. Looking at the original links that were correctly classified, the relationships with density within and between is almost the opposite, where the density within clusters has a smaller effect, but the density between clusters is very influential.

5. Discussion

The current manuscript has several notable advances. Methodologically, this is the first attempt at a broad-based, systematic simulation study of link prediction methods in social network analysis. Prior studies have focused solely on the manipulation of “real-world”, observed data sets. While these types of investigations can be informative, they can also be misleading because the true structure of the network (e.g., data set) is unknown; consequently, it is unclear whether the obtained results are due to consistent effects present across a class of data sets or they are merely by-products of data specific idiosyncrasies. Further, the reformulation of link prediction indices in terms of the four quantities *a*, *b*, *c*, and *d* show that they fall into a class of traditional similarity indices (see Albitaneh et al., 2011). Consequently, these indices will perform well when there is some underlying cluster (e.g., community) structure within the network.

After evaluating the performance of several link prediction methods on generated blockmodel data, it becomes apparent that *ARI* is among the best performing methods. Cohen’s *ARI* then provides two unique advantages over the other methods. First, there is a built in threshold of zero to determine whether a link should be added or not – a feature not present in other methods of link prediction; however, we do show that other methods that are

corrected for chance could perform just as well as *ARI*. Second, the *ARI* provides the beginning foundations of the developing the theoretical conditions under which link prediction might perform well; namely, networks with at least some level of cluster structure. Thus, rather than just “throwing” link prediction methods at the problem, methods can now be designed that leverage the types of local, internal structures that promote improved performance.

Often, cluster structures are formalized as being internally cohesive and externally isolated (Cormack, 1971), which can be translated as being tight clusters that are well-separated. The link prediction (predicting missing links) and link detection (detecting incorrect links) problems indicate that *ARI* might address both of these conditions; albeit separately. Link prediction has the biggest performance variation when the between cluster density changes, or in Cormack’s phrasing, as the internal cohesiveness of the clusters changes. On the other hand, link deletion is most influenced by the within cluster density, or as the external isolation of the cluster changes. The current research is being extended in two distinct directions: (1) adapting of the presented formulation of the *ARI* to cluster detection algorithms in networks, and (2) developing link prediction methods that are not reliant on the presence of community structure within the network.

Acknowledgments

The second author was supported by 1R21AA022074-01 from the National Institute on Alcohol Abuse and Alcoholism.

References

- Brusco MJ, Steinley D. An evaluation of a variable-neighborhood search method for blockmodeling of two-mode binary matrices based on structural equivalence. *Journal of Mathematical Psychology*. 2007; 51:325–338.
- Cohen J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*. 1960; 20(1):37–46.
- Cormack RM. A review of classification. *Journal of the Royal Statistical Society. Series A (General)*. 1971:321–367.
- Dijkstra EW. A note on two problems in connexion with graphs. *Numerische mathematik*. 1959; 1(1): 269–271.
- Dowton M, Brennan T. Comparing classifications: An evaluation of several coefficients of partition agreement. Paper presented at the meeting of the Classification Society; Boulder, CO. 1980.
- Duan D, Li Y, Li R, Lu Z. Incremental k-clique clustering in dynamic social networks. *Artificial Intelligence Review*. 2012; 38(2):129–147.
- Dubes RC. How many clusters are best?-an experiment. *Pattern Recognition*. 1987; 20(6):645–663.
- Fawcett T. An introduction to roc analysis. *Pattern recognition letters*. 2006; 27(8):861–874.
- Girvan M, Newman ME. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*. 2002; 99(12):7821–7826.
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*. 1982; 143:29–36. [PubMed: 7063747]
- Harary F. *Graph theory*. Addison-Wesley; 1969.
- Hubert L, Arabie P. Comparing partitions. *Journal of classification*. 1985; 2:193–218.
- Krebs V. Uncloaking terrorist networks. *First Monday*. 2002; 7(4)
- Kulczyński S. *Die pflanzenassoziationen der pieinen*. Imprimerie de l’Université; 1928.

- Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. *Journal of the American society for information science and technology*. 2007; 58(7):1019–1031.
- Maxwell A, Pilliner A. Deriving coefficients of reliability and agreement for ratings. *British Journal of Mathematical and Statistical Psychology*. 1968; 21(1):105–116. [PubMed: 5726239]
- Milligan GW, Soon SC, Sokol LM. The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 1983; 5(1)
- Mishra N, Schreiber R, Stanton I, Tarjan RE. Finding strongly knit clusters in social networks. *Internet Mathematics*. 2008; 5(1):155–174.
- Mitzenmacher M. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*. 2004; 1(2):226–251.
- Newman ME. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*. 2006; 74(3):036104.
- Newman MEJ. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*. 2001; 98(2):404–409.
- Rand WM. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*. 1971; 66(336):846–850.
- Ressler S. Social network analysis as an approach to combat terrorism: Past, present, and future research. *Homeland Security Affairs*. 2006; 2(2):1–10.
- Shibata N, Kajikawa Y, Sakata I. Link prediction in citation networks. *Journal of the American Society for Information Science and Technology*. 2012; 63(1):78–85.
- Steinley D. Local optima in k-means clustering: what you don't know may hurt you. *Psychological methods*. 2003; 8(3):294–304. [PubMed: 14596492]
- Steinley D. Properties of the hubert-arable adjusted rand index. *Psychological methods*. 2004; 9(3): 386–396. [PubMed: 15355155]
- Steinley D. Profiling local optima in k-means clustering: Developing a diagnostic technique. *Psychological methods*. 2006; 11(2):178–192. [PubMed: 16784337]
- Steinley D, Brusco MJ, Wasserman S. Clusterwise p* models for social network analysis. *Statistical Analysis and Data Mining*. 2011; 4(5):487–496.
- Warrens MJ. On the equivalence of cohens kappa and the hubert-arabie adjusted rand index. *Journal of Classification*. 2008; 25(2):177–183.
- Wasserman S, Faust K. *Social network analysis: Methods and applications*. Vol. 8. Cambridge University Press; 1994.
- Wasserman S, Steinley D. Sensitivity analysis of social network data and methods: Some preliminary results. *Dynamic social network modeling and analysis: Workshop summary and papers*. 2003:197–208.
- Zachary WW. An information flow model for conflict and fission in small groups. *Journal of anthropological research*. 1977:452–473.
- Zhu X, Gerstein M, Snyder M. Getting connected: analysis and principles of biological networks. *Genes & development*. 2007; 21(9):1010–1024. [PubMed: 17473168]

Highlights

- We propose the use of the adjusted Rand index to predict links in network data
- Nine proximity measures were compared on simulated and real networks
- The adjusted Rand index was the top performer out of the nine proximity measures considered
- The adjusted Rand index can also be used to detect unusual or incorrect links in a network

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

Table 2

AUC statistics for similarity indices, averaged 10 iterations. An asterisk indicates the highest AUC for each row.

Clusters	GD	CN	PA	JC	ARI	MP	Φ	R	K
2	0.46	0.66	0.47	0.69	0.70*	0.70	0.70	0.69	0.69
4	0.47	0.62	0.48	0.64	0.65*	0.65	0.65	0.64	0.64
6	0.47	0.60	0.48	0.61	0.62	0.62*	0.62	0.60	0.61
Density Within									
60	0.48	0.56	0.49	0.58	0.59	0.59*	0.59	0.58	0.58
75	0.47	0.63	0.48	0.65	0.65*	0.65	0.65	0.64	0.65
90	0.47	0.63	0.48	0.65	0.66*	0.66	0.66	0.64	0.65
Density Between									
10	0.49	0.75	0.46	0.77	0.77*	0.77	0.77	0.75	0.77
25	0.45	0.60	0.48	0.62	0.63	0.63*	0.63	0.62	0.63
40	0.46	0.54	0.49	0.55	0.56*	0.56	0.56	0.55	0.55
Links Removed									
10	0.44	0.66	0.47	0.67	0.68*	0.68	0.68	0.68	0.67
30	0.46	0.63	0.48	0.65	0.66	0.66*	0.66	0.65	0.65
50	0.49	0.60	0.48	0.62	0.63*	0.63	0.63	0.60	0.62
Average	0.47	0.63	0.48	0.65	0.66*	0.66	0.66	0.64	0.65

ANOVA results comparing the influence of different network conditions on the AUC values for ARI algorithm. The factors are number of clusters (NC), percentage of links removed (LR), density within clusters, (DW) and density between clusters (DB).

Table 3

Source	SS	df	MS	F	p-value	η^2
NC	0.87	2	0.44	756.36	0.00	0.079
LR	0.39	2	0.20	338.95	0.00	0.035
DW	2.50	2	1.25	2169.46	0.00	0.227
DB	6.35	2	3.17	5505.22	0.00	0.576
NC*LR	0.02	4	0.01	8.97	0.00	0.002
NC*DW	0.19	4	0.05	81.27	0.00	0.017
NC*DB	0.04	4	0.01	19.31	0.00	0.004
LR*DW	0.01	4	0.00	6.37	0.00	0.001
LR*DB	0.02	4	0.01	10.15	0.00	0.002
DW*DB	0.05	4	0.01	21.92	0.00	0.005
NC*LR*DW	0.00	8	0.00	0.92	0.50	0.000
NC*LR*DB	0.01	8	0.00	2.59	0.01	0.001
NC*DW*DB	0.02	8	0.00	3.80	0.00	0.002
LR*DW*DB	0.09	8	0.01	20.00	0.00	0.008
NC*LR*DW*DB	0.02	16	0.00	2.44	0.00	0.002
Error	0.42	729	0.00			
Total	11.02	809				

Table 4

Rank correlations between the scores given by each of the ten algorithms.

	GD	CN	PA	JC	ARI	MP	Φ	R	K
GD	1	0.53	0.05	0.54	0.53	0.53	0.53	0.34	0.54
CN		1	0.31	0.98	0.90	0.90	0.90	0.53	0.97
PA			1	0.14	-0.08	-0.08	-0.08	-0.55	0.13
JC				1	0.96	0.96	0.96	0.66	1.00
ARI					1	1.00	1.00	0.80	0.96
MP						1	1.00	0.80	0.96
ϕ							1	0.80	0.96
R								1	0.66
K									1

Table 5

AUC scores averaged over 10 random splits for each percentage of links removed to form the test data set. An asterisk indicates the highest AUC for each row.

	GD	CN	PA	JC	ARI	MP	Φ	R	K
Karate									
10%	0.72*	0.70	0.70	0.64	0.61	0.60	0.60	0.29	0.70
30%	0.70	0.66	0.71*	0.60	0.58	0.58	0.57	0.32	0.63
50%	0.67	0.62	0.72*	0.57	0.53	0.52	0.51	0.31	0.58
Avg	0.70	0.66	0.71*	0.60	0.57	0.57	0.56	0.31	0.64
Football									
10%	0.73	0.82	0.26	0.84	0.88	0.88	0.88	0.91*	0.84
30%	0.77	0.81	0.35	0.82	0.86	0.86	0.86	0.87*	0.82
50%	0.76	0.73	0.40	0.74	0.82	0.82	0.82*	0.78	0.74
Avg	0.75	0.79	0.34	0.80	0.85	0.85	0.85*	0.85	0.80
Coauthorship									
10%	0.99*	0.99	0.68	0.99	0.99	0.99	0.99	0.74	0.99
30%	0.98*	0.95	0.68	0.95	0.95	0.95	0.95	0.62	0.95
50%	0.96*	0.88	0.68	0.88	0.87	0.87	0.87	0.53	0.88
Avg	0.98*	0.94	0.68	0.94	0.94	0.94	0.94	0.63	0.94

Table 6

Counts and percentages of false positives, false negatives, true positives, and true negatives from the ARI scores on the real network data sets.

	10%	30%	50%
Karate			
False Positive	217.5 (44.30%)	156.3 (33.93%)	87.4 (22.84%)
False Negative	1.8 (0.37%)	7.8 (1.69%)	16.1 (4.21%)
True Positive	6.2 (1.26%)	12.8 (2.78%)	13.3 (3.48%)
True Negative	265.5 (54.07%)	283.8 (61.60%)	265.8 (69.47%)
Football			
False Positive	1,845.3 (30.74%)	1,352.9 (22.08%)	753.4 (12.06%)
False Negative	12.4 (0.21%)	46.9 (0.77%)	127 (2.03%)
True Positive	48.6 (0.81%)	137.1 (2.24%)	180 (2.88%)
True Negative	4,096.7 (68.24%)	4,589.1 (74.91%)	5,188.6 (83.03%)
Coauthorship			
False Positive	3,337.4 (0.33%)	2,094.7 (0.24%)	1,176.2 (0.17%)
False Negative	6.2 (0.00%)	67.7 (0.01%)	238.4 (0.03%)
True Positive	242.6 (0.02%)	619 (0.07%)	744.2 (0.11%)
True Negative	1,017,125.7 (99.65%)	884,571.6 (99.69%)	687,432.8 (99.69%)

Table 7

Percentages of original links and random links classified correctly by ARI. Original links with a positive ARI and random links with a negative ARI are correct.

Coauthorship	Original Links <i>ARI</i> > 0	Random Links <i>ARI</i> = 0
Clusters	($\eta^2 = .08$)	($\eta^2 = .24$)
2	64.95%	69.36%
4	59.95%	60.42%
6	57.88%	55.93%
Density Within	($\eta^2 = .15$)	($\eta^2 = .35$)
60%	55.96%	54.17%
75%	61.01%	60.88%
90%	65.81%	70.66%
Density Between	($\eta^2 = .58$)	($\eta^2 = .10$)
10%	71.66%	66.59%
25%	58.75%	61.34%
40%	52.37%	57.78%
Links Added	($\eta^2 = .07$)	($\eta^2 = .07$)
10%	63.90%	65.69%
30%	61.67%	61.52%
50%	57.21%	58.50%
Average	60.93%	61.90%