Non-backtracking walks for link prediction

Candidate Number 1004226 *,

*University of Oxford, C5.4 Networks Mini-Project

With the explosion of network data, link prediction has become a fruitful area of research in numerous domains. In particular, link prediction may be applied to large biological, social and economic networks in order to glean new insight into the complex interactions between agents. Further, our ability to predict new interactions is deeply connected to our understanding of how networks evolve. Multiple strategies, both local and global, have been suggested for evaluating the likelihood of a new link. Here we propose a novel metric, the Non-Backtracking Index, based on the theory of non-backtracking walks. We first give proof of concept by applying the index to some simple toy networks and then proceed to evaluate performance on three real-world networks, highlighting the comparative advantage of the index over the current methods.

Non-Backtracking | Katz | Prediction

Significance Statement

The discovery that most networks are scale-free is highly problematic for the Katz method of edge prediction. Shown to be heavily biased towards hubs, the Katz Index needs to be updated. In line with the application of non-backtracking walks to centrality, community detection and various other aspects of Network Science, we apply them here to the task of edge prediction. The proposed index can be calculated similarly to the Katz Index and at no extra cost, and indeed we show this index performs moderately better on some small networks ($N < 10000$). We conjecture that the difference in performance would be amplified on larger networks due to advantageous behaviour in the sparse limit.

Introduction

The question of how to predict new edges, using only the topology of a network, is a fundamental one. In one way or another, all network models incorporate the probability of an edge existing given the rest of the network structure. For example, the Erdos-Renyi model assumes independence whereas the Barabasi-Albert model assumes proportionality to $k_i k_j$. Most would agree the Barabasi-Albert model is usually a much better model for real-world networks. So developing more accurate methods of link prediction can directly inform our ability to model networks.

Perhaps more concretely, link prediction has a wide range of applications in the physical, biological and social sciences. For example, One rapidly growing area of interest is the human protein interactome (1). Studying this network is likely to lead to a richer understanding of protein functionality, however, it is estimated that the protein interactome is only 11% complete (2). Filling in the missing links will be a huge challenge as there are around 20,000-15,000 different proteins. Testing each of the $\binom{20000}{2}$ possible pairs of proteins to see if they interact is unfeasible. Accurately identifying proteins which are likely to react could save an incredible amount of time and resources. In addition to completing partially mapped networks, other applications include predicting the evolution of networks and identifying the incorrect, or unlikely interactions within them.

Several algorithms for link prediction based on the topology of a network have been proposed (3). One class of such algorithms takes two nodes, x and y, as input and gives a similarity score $S_{x,y}$ as output where a higher similarity indicates a more likely edge. (4) Categorises these into local, global and quasi-global indices. A local index requires only local information whereas a global index takes into account the structure of the whole network. Whilst local indices perform reasonably well on small highly connected networks, onglobal indices generally tend to do much better, the trade-off being the expense of computation.

One of the best performing global indices is the Katz Index, conceptually similar to Katz centrality. The Katz Index is rooted in counting walks on the graph: essentially the similarity of x and y is modelled as proportional to the ability to travel from x to y.

$$S^{Katz}_{x,y} = \sum_{l=1}^{\infty} \alpha^l (A^l)_{x,y}$$

where A is the adjacency matrix and α is a parameter often called the attenuation. It is well known that $(A^l)_{x,y}$ is the number of walks between x and y of length l. Therefore, the Katz Index counts the number of walks from x to y, giving a larger weight to smaller walks. The attenuation parameter permits the user strong control on how much weight one wishes to give long walks. Note that the sum can alternatively be expressed

$$S^{Katz} = (I - \alpha A)^{-1}$$

It may be shown the sum is convergent in the range of $0 \leq \alpha \leq 1/\rho(A)$ Where $\rho(A)$ denotes the largest eigenvalue of A.

The Katz index and the related Katz centrality have been highly regarded as providing both deep and general information about the topology of a network. Recently though in (4,5), it has been observed that the inclusion of backtracking walks can induce undesirable localisation effects. Essentially, by allowing backtracking paths, the Katz index tends to exaggerate the importance of hubs (nodes of high degree). One can intuitively view the hubs as exerting a trapping force preventing the walkers from escaping. Clearly this is sub-optimal as most real-world networks have been observed to show power law behaviour, further, suggesting a node is highly probable to a link with a hub is not so insightful. We propose an index analogous to the Katz index in which only the number of non-backtracking walks are counted.

Reserved for Publication Footnotes

The Non-backtracking Operator

A non-backtracking walk is defined as a walk in which one does not return to the node which one has just left i.e. a sequence of vertices $v_1, v_2, ..., v_n$ in which the sequence u, v, u does not occur. Connections to the theory zeta functions, specifically the Ihara zeta function, have helped us to explore the properties of non-backtracking walks and non-backtracking random walks alike (6). Let us define

$$S^{NB} = \sum_{l=1}^{\infty} t^l p_l(A)$$

Where $p_l(A)_{x,y}$ denotes the number of walks from x to y of length l. In (4) it is shown that S^{NB} can be calculated easily by using the deformed laplacian. $M(t) = I - At + (\Delta - I)t^2$. They show

$$S^{NB} = (1 - t^2)M(t)^{-1}$$

A radius of convergence is also given. If G is a tree then we get an infinite radius of convergence. If G is not a tree, under certain weak conditions, the radius of convergence is the smallest eigenvalue of $M(t)$. This is shown to be equivalent to the inverse of the largest eigenvalue of

$$C := \begin{bmatrix} A & I - \Delta \\ I & 0 \end{bmatrix}$$

With these results in hand, we proceed to propose a novel index, the Non-Backtracking Index.

A New Non-backtracking Index

Given two nodes x and y, let their similarity be

$$S^{NB}_{x,y} := \sum_{l=1}^{\infty} t^l p_l(A)_{x,y} = [(1 - t^2)M(t)^{-1}]_{x,y}$$

counting backtracking walks between x and y. Where t is a parameter analogous to the attenuation parameter α in the Katz Index. This index requires calculating the inverse of an $N \times N$ matrix with a similar sparsity to the Katz matrix whence having a similar computational complexity to the Katz Index. Moreover, it has been shown that non-backtracking systems exhibit better behaviour in the sparse limit (8). Given the increase in availability of very large sparse networks, this could prove a considerable advantage over the Katz Index.

We now test the index on some toy examples. First, a small Erdos-Renyi graph, $G(10, 1/2)$.

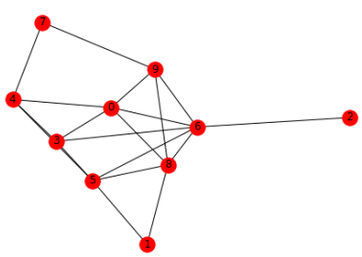

Fig. 1. Erdos-Renyi $G(10, 1/2)$

Clearly link prediction in an Erdos-Renyi Graph (due to the assumption of random, independent edges) is a fruitless task, however, it serves as a first proof of concept.

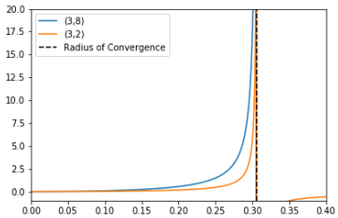

Fig. 2. A plot of two similarity scores $S_{3,8}$ and $S_{3,2}$ varied over the attenuation parameter. The dotted line represents the largest eigenvalue of the deformed laplacian, the radius of convergence.

Our first observation is the asymptotics at the radius of convergence, which is what we expect to see. For t within the radius of convergence $S_{3,2}$ is well separated from $S_{3,8}$ i.e. the index suggests a new edge (3,8) is more probable than an edge (3,2) which conforms with our intuition given Fig.1.

Next, we test the index on a Barabasi-Albert 'Scale-free' model, choosing to increase the size of the network ($N = 100$).

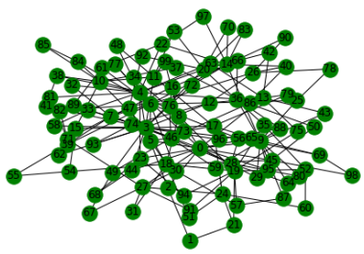

Fig. 3. Barabasi-Albert with 100 nodes.

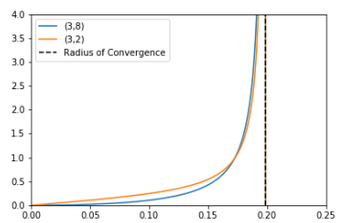

Fig. 4. A plot of two similarity scores $S_{3,8}$ and $S_{3,2}$ varied over the attenuation parameter. The dotted line representing the largest eigenvalue of the deformed laplacian, the radius of convergence.

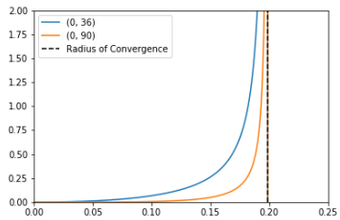

Fig. 5. A plot of two similarity scores $S_{0,36}$ and $S_{0,90}$ varied over the attenuation parameter. The dotted line representing the largest eigenvalue of the deformed laplacian, the radius of convergence.

In Figure 4, we examine two missing edges with close similarity profiles. Note the intersection of the two profiles, this intersection highlights the importance of the attenuation parameter. For small values of t we decide the edge (3,8) is more important, however, when we consider longer walks (2,3) we change our mind. This suggests optimisation of the attenuation parameter needs to be performed on a case-by-case basis.

In Figure 5, we test two potential edges which, by visual inspection of the graph, are expected to have well separated similarity profiles, and indeed this is the case. Having observed the Non-Backtracking Index is well behaved on some simple models, we proceed to employ rigorous statistical testing on real-world data sets.

Cross-Validation and AUC Scores

In order to measure our success in predicting missing or new links, we must have prior knowledge of which links are missing, or, knowledge of those that will develop in the future. The network data available tends to be static, precluding the latter, and includes all known links, ruling out the former. Hence we must induce the situation artificially.

Assuming our network is $G = (V, E)$ we split the edge set E into the disjoint union of two sets

$$E = E_{train} \sqcup E_{hidden}$$

in a random fashion. We form a new network $G_{train} = (V, E_{train})$. We can now measure our ability to predict E_{hidden} using G_{train}. To assess the quality of our prediction, we opt to use the AUC score (the area under the ROC curve). As noted in (3), the AUC measure may be interpreted as the probability a randomly chosen edge in E_{hidden} is given a higher similarity score than a randomly chosen edge in the network complement G^c_{train}. We also choose to utilise cross-validation (5-fold) to minimise the bias inherent in the random selection of E_{hidden}. Cross-validation involves rotating E_{hidden} through E, ensuring each edge is part of E_{hidden} exactly once. AUC scores of several indices, Katz, Resource Allocation (RA), Jacard (JA), Preferential Attachment (PA) and Adamic-Adar (AA), are calculated for some real-world networks.

Results

As an initial test we use the Rodriguez Madrid train bombing network (9), chosen for its modest size. The task of predicting missing or evolving links has obvious use for this network. Knowing which pairs of terrorists were likely communicating could have potentially allowed for better strategies in preventing the atrocity that occurred.

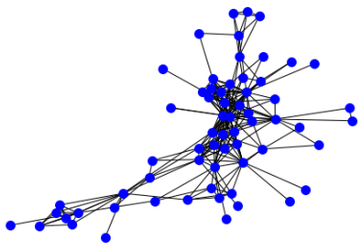

Fig. 6. Madrid 2004 train bombing network. Nodes represent terrorists and edges represent known contact between them.

Table 1. Spanish Train Bombing

Index	AUC
NB	**0.8871**
Katz	0.8848
RA	0.9262
JA	0.9094
PA	0.7792
AA	0.9237

Note the AUC scores for the Katz and Non-Backtracking indices were reasonably robust to variation of the attenuation parameter. On this small network we see most of the local indices perform slightly better than the global indices (Katz and Non-Backtracking). A possible reason for the high performance of local indices is simply, on a small network the global topology is identical to the local topology. Importantly though, the Non-Backtracking Index provides comparable results to the local indices and beats its most important competitor, the Katz Index.

Next, we do the same tests on a network consisting of the face to face interactions between humans at an exhibition. The data was collected in order to model the spread of infectious diseases on networks. This network is significantly larger ($N = 410$) (10). Again, predicting whom contagious individuals are most likely to infect is an important task for epidemiology and controlling the spread of highly infectious diseases.

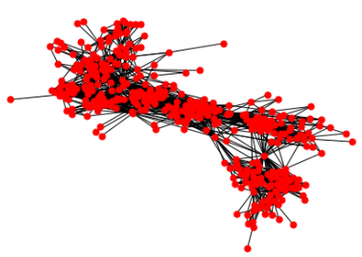

Fig. 7. Infectious Network. Nodes represent people at a conference, edges represent face to face interactions between them.

Table 2. Infectious

Index	AUC
Non Backtracking	**0.9441**
Katz	0.9416
Resource Allocation	0.9273
Jacard	0.9253
Preferential Attachment	0.7053
Adamic Adar	0.9269

The high performance of local measures may again be explained by the size of the network or, perhaps, the high modularity of the network. High modularity indicates most important information is still contained in small well connected components, benefiting the local indices. This time, however, the Non-Backtracking Index outperforms all the other indices, but there is only marginal improvement over the Katz index. The motivation behind the Non-Backtracking Index was to avoid the localisation effects of Katz, so why are we not witnessing larger performance increases? The degree distribution of the network provides some clues.

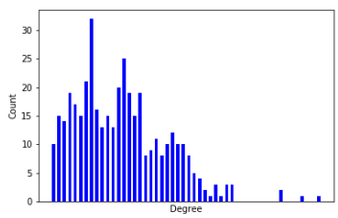

Fig. 8. Degree distribution for the Infectious Network.

The degree distribution indicates a relative non-existence of (localisation inducing) hubs. So for our next network we choose an even fatter tailed distribution. A co-authorship network of papers within the field of Network Science.

Table 3. Co-Authorship Network

Index	AUC
Non Backtracking	**0.9155**
Katz	0.9154
RA	0.9040
JA	0.8818
PA	0.8186
AA	0.9035

Discussion

Link prediction is a fundamental task in Network Science and the improvements in prediction reflects our improved understanding of networks as a whole. With the rapidly growing evidence that non-backtracking walk tools outperform their traditional walk counterparts, we decided to investigate how non-backtracking walks could be used for link prediction. In this article, we presented a novel index which we have shown, outperforms several well-known metrics. In particular, the Non-Backtracking Index presents a direct improvement over the conceptually similar Katz Index. Whilst performance gains were not drastic, we would conjecture there would be marked improvement on larger networks, especially those exhibiting scale-free degree distributions.

We did not try such tests here due to time constraints, and this raises the issue of computational complexity. The Non-Backtracking index requires an $N \times N$ matrix inversion and further we are required to perform this inversion for multiple values of the attenuation parameter t. The method then, does not scale well with network size, the best known matrix inversion techniques are still worse than $O(n^2)$ (12). There are some obvious improvements to be made here, such as using gradient based optimisation techniques to optimise t and exploiting the sparsity of M to find an inverse faster. Redemptively, these problems of complexity are all problems for the Katz Index as well. As previously remarked, the Non-Backtracking Index's better behaviour in the sparse limit could potentially prove a significant advantage over Katz.

Clearly, there are numerous other ways to assess the success of similarity indices, here we focused on one very crude measure. It would be interesting to explore the specific quality of the links proposed. One simplistic way to do this would be to measure the performance of link predictions against known information about the objects represented by the nodes. We might hope suggested links correlate with how 'similar' the two objects are. Another method would be to utilise user feedback, for example if we applied the algorithm on an online social network (suggesting new user connections), we could use the success rate (success if the users become connected) to provide a metric on the quality of our prediction.

A major issue with the method we used to assess prediction quality is its failure to account for how edges actually evolve. If we view the network $G_t = (V, E)$ as evolving, our method can be seen as observing the network at an historic time $G_{t-1} = (V, E_{hidden})$. Intuitively E_{hidden} would not have a random distribution across the network, but nevertheless, we chose to pick E_{hidden} at random. This problem is endemic to the task, for if we knew how E_{hidden} was distributed we would effectively have the perfect link prediction algorithm. Ideally, we would have the evolution data of a Network. This would allow for better testing and understanding of the link prediction task, and of networks in general.

Here we have focused on undirected and unweighted networks without loops. Directed networks and loops are easily dealt with due to (13), so we may use exactly the same techniques with some minor alterations. Extending to weighted networks is slightly more complex. One possibility is to use the Hashimoto Matrix, a $2M \times 2M$ matrix H where $H_{i\to j,k\to l} := \delta_{il}(1 - \delta_{jk})$. This is in some sense equivalent to our use of $p_l(A)$ in that we may use the l^{th} power of H to count the number of non-backtracking walks of length l. Hence similar properties follow, and in (6) the authors show some of these properties may be extended to weighted networks. Whilst some work is required, it seems likely the Non-Backtracking index could be extended to weighted networks.

1. Snider J. et al. (2015). Fundamentals of protein interaction network mapping. Molecular systems biology, 11(12), 848. doi:10.15252/msb.20156351
2. Hart G.T, Ramani A.K., Marcotte E.M. (2006). How complete are current yeast and human protein-interaction networks?. Genome biology, 7(11), 120.
3. L L. and Zhou T. (2011). Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications, 390(6), pp.1150-1170.
4. Grindrod P, Higham D, Noferini V (2018). The Deformed Graph Laplacian and Its Applications to Network Centrality Analysis. SIAM Journal on Matrix Analysis and Applications 39:310-341.
5. Martin T, Zhang X, Newman M (2014). Localization and centrality in networks. Physical Review E 90.
6. Kempton M (2016). Non-Backtracking Random Walks and a Weighted Iharas Theorem. Open Journal of Discrete Mathematics 06:207-226.
7. Krzakala F et al. (2013) Spectral redemption in clustering sparse networks. Proceedings of the National Academy of Sciences 110:20935-20940.
8. Hayes, B. (2006). Connecting the dots. American Scientist 94 (5):400-404.
9. Isella L et al. (2011). What's in a crowd? Analysis of face-to-face behavioral networks. Journal of Theoretical Biology 271:166-180.
10. Williams V (2012). Multiplying matrices faster than coppersmith-winograd. Proceedings of the 44th symposium on Theory of Computing - STOC '12.
11. Arrigo F, Grindrod P, Higham D, Noferini V (2017). Non-backtracking walk centrality for directed networks. Journal of Complex Networks 6:54-78.

Supplemental Information

The python code used to produce the data is left attached. All code produced is original (excluding code from standard libraries scipy, numpy, sklearn and matplotlib).