2015 IEEE 15th International Conference on Data Mining Workshops

Link Prediction in Large Networks by Comparing The Global View of Nodes in The Network

Mustafa Coşkun
Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, OH 44106, USA.
Email: mustafa.coskun@case.edu

Mehmet Koyutürk
(1) Department of Electrical Engineering and Computer Science
(2) Center for Proteomics and Bioinformatics
Case Western Reserve University
Cleveland, OH 44106, USA.
Email: mehmet.koyuturk@case.edu

***Abstract*—Link prediction is an important and well-studied problem in network analysis, with a broad range of applications including recommender systems, anomaly detection, and de-noising. The general principle in link prediction is to use the topological characteristics of the nodes in the network to predict edges that might be added to or removed from the network. While early research utilized local network neighborhood to characterize the topological relationship between pairs of nodes, recent studies increasingly show that use of global network information improves prediction performance. Meanwhile, in the context of disease gene prioritization and functional annotation in computational biology, "global topological similarity" based methods are shown to be effective and robust to noise and ascertainment bias. These methods compute topological profiles that represent the global view of the network from the perspective of each node and compare these topological profiles to assess the topological similarity between nodes. Here, we show that, in the context of link prediction in large networks, the performance of these global-view based methods can be adversely affected by high dimensionality. Motivated by this observation, we propose two dimensionality reduction techniques that exploit the sparsity and modularity of networks that are encountered in practical applications. Our experimental results on predicting future collaborations based on a comprehensive co-authorship network shows that dimensionality reduction renders global-view based link prediction highly effective, and the resulting algorithms significantly outperform state-of-the-art link prediction methods.**

I. Introduction

Large scale real-world networks and their characteristics have been soaring popularity and gaining considerable attention of researchers in a broad range of applications. Designing methods to understand the evolution of these networks and predict their future behavior is a major challenges for researchers. One recurring theme in this line of challenges is known as the link prediction problem. Link prediction is usually defined as the task of predicting the links that are likely to emerge/dissappear in an evolving network or identifying the missing/spurious links in a static network [1].

Earlier link prediction methods are designed to characterize the relationship between pairs of nodes in a network based on the local network neighborhood of the nodes, and use this information to assess the likelihood of the emergence of an edge between nodes [2]–[4]. To assess the likelihood of the emergence/existence of an edge between a pair of nodes, these methods usually utilize the individual properties of the two nodes or the sets of their neighbors (incident nodes). Clearly, such methods treat the network as a "bag of interactions" rather than a true network, since they do not take into account the potential flow of information across the network through indirect paths. However, the topology of the entire network may have an influence on the existence or emergence of an edge between a pair of nodes. For example, in the case of co-authorship networks, researchers may gain collaborations through a chain of collaborators. Similarly, in biological networks, paths often represent the flow of information in the cell, including transduction of signals through a series of interactions or synthesis of metabolites through a chain of reactions [5], thus pairs of proteins involved in similar biological processes may be likely to gain interactions [6].

"Global" link prediction methods aim to account for the influence of the overall network topology on the emergence of edges. The main idea behind these approaches is that pairs of nodes that are close to each other in the network are likely to gain edges. Existing methods differ in terms of how they assess the proximity of (or distance between) nodes in the network. In the context of link prediction, algorithms that are utilized to assess proximity (or distance)) include shortest paths [7], random walks, rooted page rank, and hitting time [8].

Network-based disease gene prioritization can be considered an application of link prediction. In this application, the network is composed of genes and diseases. The task in disease gene prioritization is to use this network to rank a given set of candidate genes in terms of their likelihood of association with a given disease [9]. In the context of this problem, global methods are shown to be significantly more effective than local methods [10]. However, these methods are vulnerable to ascertainment bias in that they favor high-degree nodes over nodes with lower degree [11]. To alleviate this problem, "global view based" methods assess the closeness of nodes in the network by comparing the views of the network from the perspective of nodes [12]. To be more precise, these methods compute the topological profile of each node in terms of proximity to other nodes, and use those profiles to assess the positional similarity of nodes in the network.

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.195

In computational biology, "global view based" methods are shown to be effective in disease gene prioritization [12], network de-noising [6], and prediction of protein function [13].

In this paper, we show that the performance of "global view based" methods in the context of link prediction in large social networks can be adversely affected by high-dimensionality. To address this problem, we propose two dimensionality reduction techniques that take advantage of the sparsity and modularity of the networks. Namely, to exploit the sparsity of the network, we compute reduced topological profiles for each node in the network that can parsimoniously represent the position of the node in the network. Second, we exploit the modularity of the network by clustering the nodes and using node clusters as the dimensions in the topological profile. This approach is motivated by the observation that node clusters may have semantic meanings that may contribute to the emergence of links between nodes. For example, in a co-authorship network, two authors can be more likely to publish papers if they have a common research interest, attend the same conference, or reside in the same institution [15]. Such 'discrete" entities can be better captured by considering clusters of nodes as opposed to considering proximity to each individual nodes.

Experimental results on predicting future collaboration based on a large co-authorship network derived from DBLP show that our techniques render global view based methods highly effective in link prediction, and resulting algorithms significantly outperform Random Walk with Restart algorithm [14] and global topological similarity based link prediction without dimensionality reduction (GLOBAL) [12].

The remainder of the paper is organized as follows. We first discuss existing global view based link prediction approaches in Section 2. In Section 3, we present the algorithmic details of the proposed dimensionality reduction methods. We give systematic experimental studies using DBLP, a large-scale real-world dataset,[1] in Section 4. We conclude our discussion in Section 5.

II. Related Work

Link prediction can be supervised or unsupervised, depending on whether training samples are available for the "future" instances of the network. Here, our focus is on unsupervised link prediction, where no such training data is available. Unsupervised link prediction methods need to make assumptions on which topological characteristics of the earlier instances of the network indicate the likelihood of the emergence of an edge. The most commonly used assumption can be roughly defined as "guilt-by-associations" (or generalizations of this principle thereof), where the expectation is that nodes that have many indirect connections or in general are close to each other in the network are likely to become adjacent to each other in the network.

Unsupervised link prediction methods mainly fall into two categories in terms of how they quantify the "closeness" of two nodes $v_i, v_j \in \mathcal{V}$: local neighborhood methods [8] and global perspective methods [16]. Local neighborhood based methods mostly work on the sets of nodes that are adjacent to the nodes being considered for link prediction. A general principle that is employed by these methods is that pairs of nodes that share a large set of neighbors are likely to become adjacent in the future. The performance of these methods have been broadly examined by Nowell *et al.* [8] and it was found that the Adamic and Adar's method that takes into account node degrees [2] outperforms other local neighborhood-based methods.

Global perspective based methods aim to account for the entire topology of the network. Earlier global perspective methods, such as shortest path (the most reliable path) or $Katz$ measure [17], were shown to be adversely affected by the "small wold phenomenon" [8]. Recently, information flow based methods have been commonly used to measure global proximity, including random walk with restarts [18]. These methods are also used in various applications, including community detection [19], disease gene prioritization [12], and modeling the evolution of social networks [20]. Other global view based prediction methods include meta-path based approaches [21], where the path count and random walk around the given meta paths is used to assess the proximity between pairs of nodes.

III. Method

In this section, we first describe the link prediction problem within a formal framework. Subsequently, we formulate the concept of global topological similarity of pairs of nodes in terms of their proximity to other nodes in the network. Then, we introduce our dimensionality reduction techniques that aim to alleviate the effect of high dimensionality in very large networks and we discuss how we employ these dimensionality reduction techniques to score node pairs in the context of the link prediction problem.

A. Link Prediction Problem

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{T})$ be a large network that evolves over time through emergence of new edges. Here, $\mathcal{V}$ denotes the set of nodes, $\mathcal{E}$ denotes the set of edges, and $\mathcal{T} : \mathcal{E} \rightarrow \mathbb{N}$ is a function that specifies the creation times of edges. For some real-world networks (e.g., friendships in social networks, protein-protein interaction networks at the evolutionary time scale) edges can emerge or disappear. For many other real-world networks (e.g., co-authorship networks, acquaintances in social networks), edges can emerge over time, but they do not disappear. While the methods proposed in this paper apply to either formulation; we focus here on the latter model (i.e., edges cannot disappear) for simplicity.

Let $\mathcal{G}^{(t)} = (\mathcal{V}, \mathcal{E}^{(t)})$ denote the instance of the network at a given time $t \in \mathbb{N}$, i.e., $\mathcal{E}^{(t)} = \{uv \in \mathcal{E} : \mathcal{T}(uv) \leq t\}$. In the classical problem of link prediction, the input is such an instance of the network at time t [8]. The output is a scoring or ranking of the pairs of nodes in $(\mathcal{V} \times \mathcal{V} \backslash \mathcal{E}^{(t)})$ in terms of the likelihood of the emergence of an edge between these nodes

[1] http://www.informatik.uni-trier.de/ley/db/

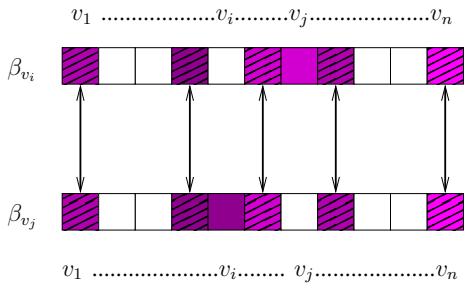

Figure 1. **Illustration of link prediction using global topological similarity of nodes.** The two vectors represent the global topological profiles of two nodes, v_i and v_j. Shades of magenta indicate the proximity of v_i and v_j to the node represented by each dimension in the vector. Dimensions on which both nodes have patterned shades, highlighted by bidirectional arrows, indicate the nodes that are close to both v_i and v_j. Global topological similarity aims to utilize such nodes to capture the positional similarity of v_i and v_j in the network.

at a future time $t' > t$. Here, the prediction is unsupervised, i.e., no training data is available from time t'.

B. Assessing Global Topological Similarity

For a given network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and node $v_i \in \mathcal{V}$, the *global topological profile* β_{v_i} of v_i is defined as a $|\mathcal{V}|$-dimensional vector that is composed of the proximity of v_i to the nodes in $\mathcal{V}$, i.e., $\beta_{v_i}(j)$ contains the proximity of v_i to v_j for all $v_j \in \mathcal{V}$. Erten et al. [12] quantify the proximity between two nodes using random walk with restarts, and we follow the same procedure in this paper. However, the methods proposed here can be applied to any other measure of proximity as well.

Random walk based proximity is defined as follows:

$$\beta_{v_i} = (1-\alpha)W\beta_{v_i} + \alpha r_{v_i}. \tag{1}$$

Here, W denotes the stochastic matrix derived from the adjacency matrix of $\mathcal{G}$, r_{v_i} denotes the restart vector that contains a 1 at its ith entry and a 0 in all other entries, and α denotes the damping factor, which is a parameter that sets the probability of restarting at v_i in a random walk of the network.

Given the global topological profiles of all nodes in the network, the global topological similarity of v_i and $v_j \in \mathcal{V}$ is defined as [12]:

$$\rho(\beta_{v_i}, \beta_{v_j}) = \frac{\sum_{v_t \in \mathcal{V}} (\beta_{v_i}(v_t) - \frac{1}{|\mathcal{V}|})(\beta_{v_j}(v_t) - \frac{1}{|\mathcal{V}|})}{\sqrt{\sum_{v_t \in \mathcal{V}} (\beta_{v_j}(v_t) - \frac{1}{|\mathcal{V}|})^2} \sqrt{\sum_{v_t \in \mathcal{V}} (\beta_{v_j}(v_t) - \frac{1}{|\mathcal{V}|})^2}} \tag{2}$$

In other words, $\rho(\beta_{v_i}, \beta_{v_j})$ is defined as the Pearson correlation coefficient of β_{v_i} and β_{v_j}. While $\rho(\beta_{v_i}, \beta_{v_j})$ is useful in quantifying the positional similarity of v_i and v_j in the network, high dimensionality of these vectors adversely affects its usefulness in very large networks. To remedy this problem, we propose two dimensionality reduction techniques.

C. Dimensionality Reduction for Topological Profiles

The core idea behind the proposed dimensionality reduction techniques is to identify a small set of nodes or other entities in the network that can be used to represent the position of each node in the network. The first approach we propose exploits the modular nature of real-world networks, while the second exploits their sparsity.

1) Modularity-Based Dimensionality Reduction: Many real-world networks are organized in a modular manner; i.e., they are composed of identifiable clusters where the nodes in a cluster are densely connected to each other, but are somewhat sparsely connected to the rest of the network [15]. Since the nodes in a cluster densely interact with each other, one can naturally expect that the proximity of the nodes in a cluster to any other node in the network will be similar to each other. Motivated by this insight, we identify clusters in $\mathcal{G}^{(t)}$ and use these clusters to compute a modularity-based topological profile for each node in the network. This process is illustrated in Figure 2.

We first identify the clusters in $\mathcal{G}^{(t)}$ using an established algorithm for clustering networks based on the connectivity of the nodes [22]. This algorithm identifies clusters in a given network by hierarchically partitioning the network into K clusters based on node proximities computed using page rank. An important parameter here is the number of clusters that are utilized, since a smaller number of clusters leads to smaller number of dimensions, whereas a larger number of clusters provides a more accurate representation of the topology of the network. We denote this parameter with K and investigate the effect of this parameter on the performance in link prediction in the next section.

Once clusters in the network are identified, we utilize the random walk with restarts procedure to compute modularity-based topological profiles for all nodes in the network. Let $C_1, C_2, ...C_K$ denote the K clusters identified where each cluster is a set of nodes. We first compute $|\mathcal{V}|$-dimensional topological profiles for clusters as

$$\beta_{C_i} = (1-\alpha)W\beta_{C_i} + \alpha r_{C_i}, \tag{3}$$

where

$$r_{C_i}(j) = \begin{cases} 1/|C_i| & \text{if } v_j \in C_i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Subsequently, for each node $v_i \in \mathcal{V}$, we construct a K dimensional modularity-based topological profile γ_{v_i}, by setting

$$\gamma_{v_i}(j) = \beta_{C_j}(i). \tag{5}$$

We then compute the modularity-based topological similarity between v_i and v_j as the Pearson correlation between the vectors γ_{v_i} and γ_{v_j}, and use this score as an indicator of the likelihood of the emergence of a link between v_i and v_j.

2) Sparsity Based Dimensionality Reduction: Many real-world networks are sparse, i.e., nodes have connections to only a very small fraction of all nodes in the large networks. This translates into a skewed distribution in the proximity of a node to other nodes in the network: A node is very close to its few neighbors, somewhat close to its two or three-hop neighbors, and the proximities dissipate quickly as the number of hops grow further. For this reason, the proximity of a node to many of the nodes in the network may not be informative on the position of that node in the

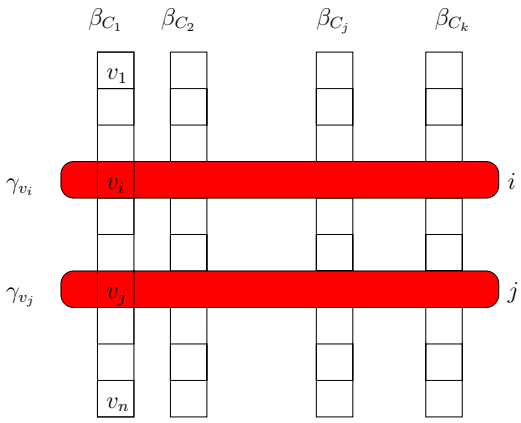

Figure 2. **Modularity-based dimensionality reduction for computing global topological similarity.**The column vectors labeled $\beta_{C_1}, ..., \beta_{C_k}$ $|\mathcal{V}|$-dimensional topological profiles for K clusters identified using Gmine, an established network clustering algorithm. The red row vectors γ_{v_i} and γ_{v_j} show the modularity-based topological profiles for the nodes v_i and v_j derived from these cluster topological profiles.

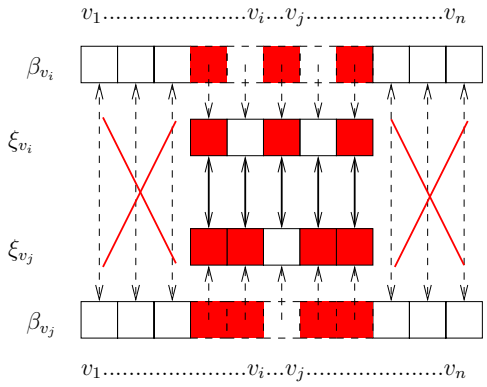

Figure 3. **Sparsity-based dimensionality reduction for computing global topological similarity.** β_{v_i} and β_{v_j} represent the global topological profiles of v_i and v_j. Red boxes indicate larger values in each vector. A smaller number of dimensions are chosen by selecting the dimensions in which at least one of β_{v_i} or β_{v_j} have a sufficiently large value (as determined using a threshold parameter denoted ϵ). ξ_{v_i} and ξ_{v_j} show *sparsity based topological profiles* of v_i and v_j which are computed by projecting the global topological profiles on these selected dimensions.

network. Motivated by this insight, we propose to reduce the dimensionality of the global topological profiles based on the magnitude of entries. This sparsity-based dimensionality-reduction technique is illustrated in Figure 3.

In sparsity-based dimensionality reduction, the dimensions utilized depend on the pair of nodes that are being compared, i.e., we select a different set of dimensions for each pair of nodes v_i and v_j. The motivation behind this approach is that the nodes that are most informative about the relative positions of v_i and v_j with respect to each other are the nodes that are in close proximity to at least one of these nodes. Namely, for a given pair of nodes v_i and v_j, we select a subset $D^{(i,j)} \in \{1, 2, ..., |\mathcal{V}|\}$ of the networks as follows:

$$D^{(i,j)} = \{v_k \in \mathcal{V} : \beta_{v_i}(k) \geq \epsilon \text{ OR } \beta_{v_j}(k) \geq \epsilon\}. \quad (6)$$

Here, ϵ is a parameter that is used to balance the trade-off between the number of dimensions and the accuracy of the approximation provided by the reduced profiles. We also investigate the effect of ϵ on link prediction performance in the next section. Then, for each pair of nodes v_i and v_j, we compute reduced topological $\xi_{v_i}^{(j)}$ and $\xi_{v_j}^{(i)}$ by projecting β_{v_i} and β_{v_j} on the dimensions specified by $D^{(i,j)}$. We compute the sparsity based topological similarity of nodes based on the Pearson correlation of ξ_{v_i} and ξ_{v_j}, and use this score as an indicator of the likelihood of the emergence of a link between v_i and v_j.

Table I
DESCRIPTIVE STATISTIC OF DBLP DATASETS

DBLP Data Set	DBLP-1	DBLP-2
Number of Nodes	10704	6250
Training Links	49750	30125
Maximum Degree	115	72
Average Degree	4.65	4.82
Test Links	12741	5592

IV. RESULTS

In this section, we systematically evaluate the performance of the proposed dimensionality reduction techniques. We start by describing the datasets and our experimental setting. Next, we analyze the performance of the two dimensionality reduction techniques and the effect of parameters. Subsequently, we compare the performance of our proposed methods with random walk with restart based link prediction and global topological similarity without dimensionality reduction.

A. Datasets

We test and compare the proposed methods on two comprehensive sets of real-world collaboration networks extracted from DBLP Computer Science Bibliography [23] [2]. For training data, we consider authors who have published papers between 2006 and 2008. In these networks, the authors are represented by nodes and there is a undirected link if two authors published at least one paper from 2006 to 2008. As test data, we use new co-author links that emerge between 2009 and 2010. These datasets are described in Table I.

DBLP-1 The first DBLP dataset consists of 15 representative conferences in 6 computer science research areas (Databases, Data Mining, Artificial Intelligence, Information Retrieval, Computer Vision and Machine Learning) [23].

DBLP-2 The second DBLP dataset is constructed from 16 representative conferences in the 6 different computer science fields namely (Algorithms and Theory, Natural Language Processing, Bioinformatics, Networking, Operating Systems and Distributed and Parallel Computing) [23].

[2]http://www.informatik.uni-trier.de/ley/db/

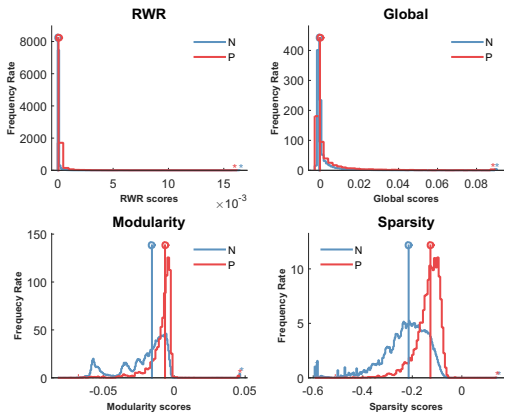

Figure 4. **Comparison of the distribution of prediction scores for positive and negative node pairs:** The distribution of prediction scores for the positive node pairs is shown in red, the mean of 10 randomly chosen negative test of equal size to the positive set is shown in blue. This figure is generated using the DBLP-1 dataset.

B. Experimental Setting

The objective of link prediction is to predict links that will emerge in the network in the future. For this reason, a "positive" label in this setup refers to a new link that emerges in the future version of a network, whereas a "negative" label refers to two nodes that remain unconnected in the future version. Since the real-world networks are highly sparse, the number of negative pairs is much larger than the number of positive pairs. To have a balanced set of positive and negative labels in the test data, we subsample negatives to construct our test set. Namely, let $\mathcal{W}$ denote the set of authors who published at least one paper in the testing interval [2009, 2010], but have not published together in the training interval ([2006, 2008]). We construct our positive and negative node pairs from this set as follows:

- The positive test set $\mathcal{P}$ is composed of $u, v \in \mathcal{W}$ such that u and v published a paper between 2009 and 2010.
- The negative test set $\mathcal{N}$ is composed of $|\mathcal{P}|$ randomly chosen pairs $u, v \in \mathcal{W}$ such that u and v did not publish a paper in 2009 and 2010.
- We generate 100 negative test sets $\mathcal{N}^{(1)}, \mathcal{N}^{(2)}, ..., \mathcal{N}^{(100)}$ and perform all experiments in each of the positive and negative test pair sets ($\mathcal{P}$ vs. $\mathcal{N}^{(i)}$ for $1 \leq i \leq 100$).
- For all experiments, we report the mean and the standard deviation of the performance figures among these 100 runs.

To evaluate the performance of different methods in scoring the testing pairs for link prediction, we use the area under ROC curve (AUC) as the performance criterion.

Comparison to other algorithms. We compare the performance of the proposed methods against random-walk with restart based link prediction (RWR) [24] and global topological similarity based link prediction without dimensionality reduction (GLOBAL) [12]. RWR is selected for comparison since it is shown to outperform local network topology based link prediction algorithms [24]. We also compare against

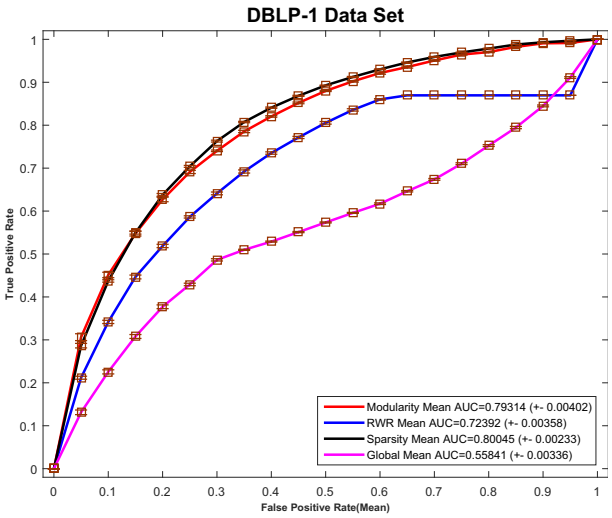

Figure 5. The performance of global topological similarity based link prediction using sparsity and modularity based dimensionality reduction, as compared to random walk with restarts and global topological similarity without dimensionality reduction, on the **DBLP-1** dataset.

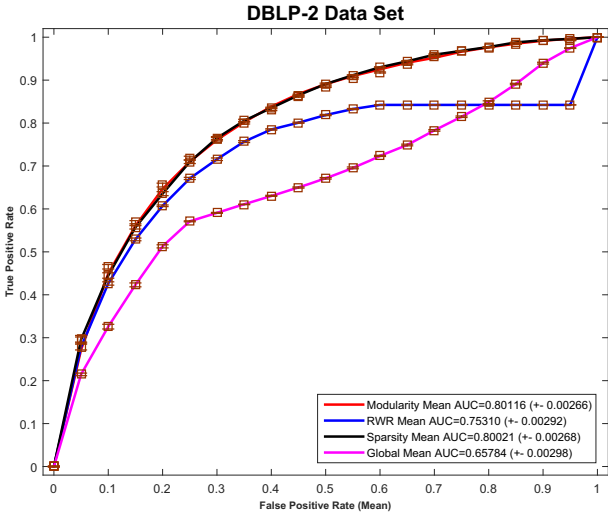

Figure 6. The performance of global topological similarity based link prediction using sparsity and modularity based dimensionality reduction, as compared to random walk with restarts and global topological similarity without dimensionality reduction, on the **DBLP-2** dataset.

GLOBAL to assess the contribution of dimensionality reduction to the performance of this algorithm.

C. Performance Evaluation

Performance of dimensionality reduction techniques. We first compare the distribution of prediction scores for the positive and negative test pairs for each algorithm. The results of this analysis are shown in Figure 4. As can be seen in the figure, the scores assigned to negative and positive pairs by both RWR and GLOBAL are highly overlapped around 0, making it very difficult to distinguish the scores of positive and negative pairs. The scores computed using both the sparsity and modularity based dimensionality reduction

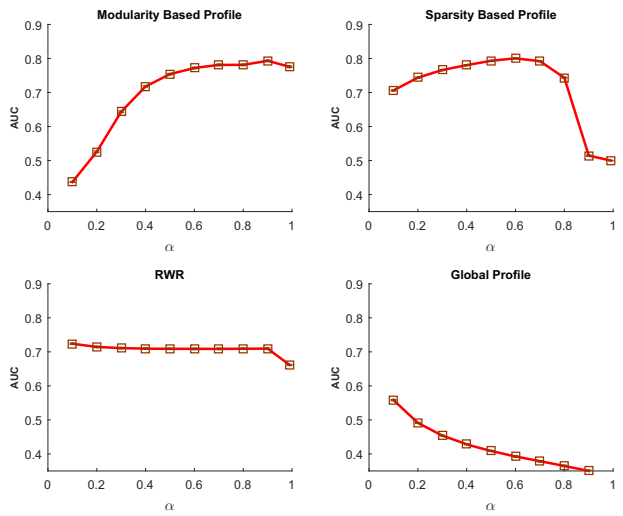

Figure 7. Effect of restart probability α on the predictive performance of link prediction methods for data set **DBLP-1**. Optimal values of α are respectively $0.9, 0.6, 0.1, 0.1$ for modularity-based, sparsity based, RWR and GLOBAL methods.

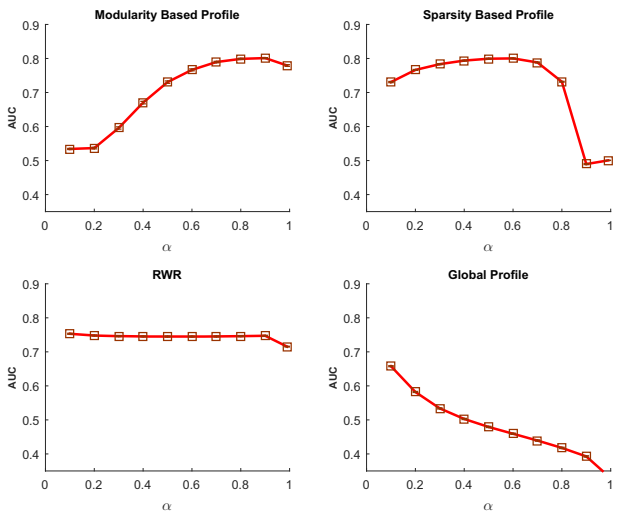

Figure 8. Effect of restart probability α on the predictive performance of link prediction methods for data set **DBLP-2**. Optimal values of α are respectively $0.9, 0.6, 0.1, 0.1$ for modularity-based, sparsity based, RWR and GLOBAL methods.

provide a wider distribution, making the separation between the positive and negative sets more visible.

Next, we compare the prediction performance of the four methods using ROC curves. The results of this analysis for both datasets is shown in in Figures 5 and 6. The results reported in these figures are the best results provided by each algorithm based on optimization of the relevant parameters (restart probability, threshold on retaining dimensions, number of clusters). As seen in the figure, the prediction performance of the two dimensionality reduction techniques on both datasets is similar, but sparsity based dimensionality reduction achieves best performance on larger dataset. Both methods drastically outperform RWR and GLOBAL, showing the value of using the global view of the nodes along with dimensionality reduction. Strikingly, GLOBAL performs poorly as compared to RWR on these datasets, showing that the global view based approach does not provide any value in predictive performance unless dimensionality reduction is performed.

Effect of restart probability α. Since the prediction scores computed by all link prediction methods considered depends on the restart probability (the parameter α) in the random walk with restarts, we also evaluate the effect of this parameter on the performance evaluation. The results of this analysis for both datasets are shown in Figures 7 and 8. As can be seen in the figures, the performance of RWR is most robust to the value of α, but the performance of this method slightly goes down at larger values of α (i.e., when the topology of the network weighted less) This is consistent with other previous studies [12]. Interestingly, the performance of modularity-based dimensionality reduction improves with increasing α, with a slight decline at very large values. The performance of sparsity-based dimensionality reduction peaks around $\alpha = 0.6$ and declines on either direction, suggesting that the value of this parameter needs to be carefully optimized for sparsity based dimensionality reduction. This is expected because large values of α restrict RWR-based proximity to the immediate neighborhood of nodes, whereas smaller values of α bias RWR-based proximity with overall centrality, diminishing the specificity of individual nodes.

Effects of sparsity threshold and the number of clusters. We also investigate the threshold values for harvesting useful global topological profiles for sparsity based dimensionality reduction, as well as the number of clusters for modularity-based dimensionality reduction. The results of these analyses are respectively shown in Figures 9 and 10.

As seen in Figures 9, we obtain best performance for sparsity based reduction with threshold value $\epsilon = 0.01$. Namely, when dimensionality reduction is very conservative ($\epsilon = 0.1$), the prediction performance of this method is worse than random prediction, i.e., valuable information on the position of the nodes in the network is left out at this value of ϵ. However, when ϵ is sufficiently flexible, the performance of the method is highly robust to the value of ϵ, with a slight decline in performance as ϵ becomes more flexible (i.e., as more dimensions are considered). Also note that, for all the threshold values except $\epsilon = 0.1$, which prunes out valuable global views most, this method significantly improves over global topological profile method.

As seen in Figure 9, the performance of modularity-based dimensionality reduction depends on the number of clusters in a way similar to the dependence of sparsity-based dimensionality reduction on ϵ. The performance is relatively poor if too few clusters are used, the performance peaks when a moderate number of clusters is used, and the performance declines steadily with growing number of clusters (i.e. growing number of dimensions) after this point. Interestingly, the performance

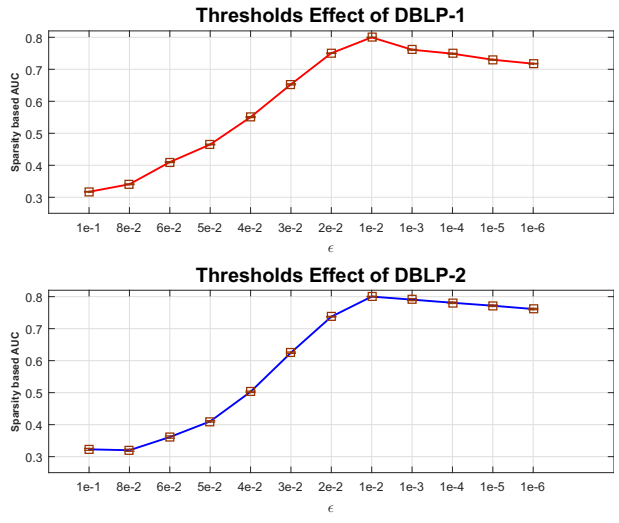

Figure 9. **The effect of the parameter ϵ on the performance of sparsity-based dimensionality reduction in link prediction.** Best performance is obtained for $\epsilon = 0.01$ for both **DBLP-1** and **DBLP-2**datasets

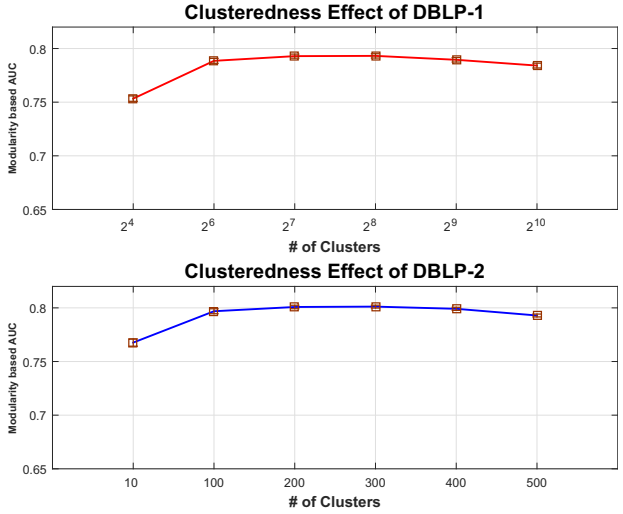

Figure 10. The effect of number of clusters on the link prediction performance of modularity-based dimensionality reduction for the **DBLP-1** and **DBLP-2** datasets

is optimized for both datasets when the average number of nodes in each cluster is around 50 (namely, $\approx 10704/256$ for DBLP-1 and $\approx 6250/100$ for DBLP-2).

V. Conclusion

In this paper, we investigate the link prediction problem for collaboration networks in a restricted perspective of global views. The global view based link prediction methods ignore the fact that the large-scale network is in sparse form, and treat each links' view equally on a node without considering how negligible they may be. In this paper, we proposed two dimensionality reduction techniques, namely sparsity and modularity based dimensionality reduction, which capture the nodes' intrinsic global views patterns from their global topological profile and elaborate only those restricted views.

Experiments on the DBLP collaboration network demonstrate that the judicious choice of threshold and cluster numbers in sparsity and modularity based reduction methods renders these methods to drastically outperform existing link prediction methods based on global network topology. Our proposed methods are better able to capture the true proximity between node pairs based on the modular structure of the network and improve the performance of unsupervised link prediction methods.

Acknowledgment

The authors would like to thank anonymous reviewers for their valuable comments, which helped improve the paper significantly. This work was supported in part by US National Science Foundation (NSF) award CCF-0953195. Mustafa Coşkun was supported by a scholarship from The Turkish Ministry of National Education.

References

[1] Y. Yang, N. Chawla, Y. Sun, and J. Hani, "Predicting links in multi-relational and heterogeneous networks," in Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, 2012, pp. 755–764.
[2] L. A. Adamic and E. Adar, "Friends and neighbors on the web," Social networks, vol. 25, no. 3, pp. 211–230, 2003.
[3] M. E. Newman, "Clustering and preferential attachment in growing networks," Physical Review E, vol. 64, no. 2, p. 025102, 2001.
[4] A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," Physica A: Statistical mechanics and its applications, vol. 311, no. 3, pp. 590–614, 2002.
[5] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," Nature reviews genetics, vol. 5, no. 2, pp. 101–113, 2004.
[6] C. Lei and J. Ruan, "A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity," Bioinformatics, vol. 29, no. 3, pp. 355–364, 2013.
[7] R. W. Floyd, "Algorithm 97: shortest path," Communications of the ACM, vol. 5, no. 6, p. 345, 1962.
[8] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," Journal of the American society for information science and technology, vol. 58, no. 7, pp. 1019–1031, 2007.
[9] K. Lage, E. O. Karlberg, Z. M. Størling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, N. Tommerup et al., "A human phenome-interactome network of protein complexes implicated in genetic disorders," Nature biotechnology, vol. 25, no. 3, pp. 309–316, 2007.
[10] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," Bioinformatics, vol. 26, no. 8, pp. 1057–1063, 2010.
[11] S. Erten, G. Bebek, R. M. Ewing, and M. Koyutürk, "Dada: Degree-aware algorithms for network-based disease gene prioritization." BioData mining, vol. 4, no. 1, pp. 1–20, 2011.
[12] S. Erten, G. Bebek, and M. Koyutürk, "Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks," Journal of Computational Biology, vol. 18, no. 11, pp. 1561–1574, 2011.
[13] M. Cao, H. Zhang, J. Park, N. M. Daniels, M. E. Crovella, L. J. Cowen, and B. Hescott, "Going the distance for protein function prediction: a new distance metric for protein interaction networks," 2013.
[14] A. Agarwal and S. Chakrabarti, "Learning random walks to rank nodes in graphs," in Proceedings of the 24th international conference on Machine learning. ACM, 2007, pp. 9–16.
[15] L. Backstrom and J. Leskovec, "Supervised random walks: predicting and recommending links in social networks," in Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011, pp. 635–644.
[16] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," Physica A: Statistical Mechanics and its Applications, vol. 390, no. 6, pp. 1150–1170, 2011.

[17] L. Katz, "A new status index derived from sociometric analysis," Psychometrika, vol. 18, no. 1, pp. 39–43, 1953.
[18] L. Lovász, "Random walks on graphs: A survey," Combinatorics, Paul erdos is eighty, vol. 2, no. 1, pp. 1–46, 1993.
[19] K. Macropol, T. Can, and A. K. Singh, "Rrw: repeated random walks on genome-scale protein networks for local cluster discovery," BMC bioinformatics, vol. 10, no. 1, p. 283, 2009.
[20] H. Tong and C. Faloutsos, "Center-piece subgraphs: problem definition and fast solutions," in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006, pp. 404–413.
[21] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han, "Co-author relationship prediction in heterogeneous bibliographic networks," in Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on. IEEE, 2011, pp. 121–128.
[22] J. F. Rodrigues Jr, H. Tong, A. J. Traina, C. Faloutsos, and J. Leskovec, "Gmine: a system for scalable, interactive graph visualization and mining," in Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment, 2006, pp. 1195–1198.
[23] X. Wang and G. Sukthankar, "Link prediction in multi-relational collaboration networks," in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara Falls, Canada, Aug 2013, pp. 1445–1447.
[24] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," Knowledge and data engineering, ieee transactions on, vol. 19, no. 3, pp. 355–369, 2007.