Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi

Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **22** (2022) 045103 (778-788) AKU J. Sci. Eng. **22** (2022) 045103 (778-788)

DOI: 10.35414/akufemubid.1127509

Araştırma Makalesi / Research Article

Discovering Link Prediction Methods' Performances by Network Topology Relation

Günce Keziban ORMAN[1,*]

[1]*Galatasaray University, Faculty of Engineering and Technology, Department of Computer Engineering, İstanbul.*

*e-posta: : *Corresponding Author: korman@gsu.edu.tr ORCID ID: http://orcid.org/0000-0003-0402-8417*

Geliş Tarihi: 07.06.2022 Kabul Tarihi: 23.08.2022

Keywords
Link prediction; Performance evaluation; Network topology; Graph embedding

Abstract

One of the prominent topics in complex network analysis is link prediction, which is a key component of network-based recommendation systems or finding missing connections. There are several different link prediction methods in the literature based on measuring the likelihood of the existence of a link between two nodes. These methods use different topological properties of the network. Although there are methods using different strategies, previous studies have focused only on method success but have not adequately examined the relationship between the performance of these methods and the topology of the network. The main motivation for this study is to reveal the role of different network topologies in link prediction. Thus, the choice of link prediction method can be customized according to the topological characteristics of the network. The two main contributions of the study are, firstly, comparing different link prediction methods with well-known performance measures in social, biological, and information networks with different topological properties in a large experimental setup; and second, examining the possible relationship between the performance of link prediction methods and the network topology. Based on the experimental results, the global methods are more successful than others, regardless of the network topology. In addition, it was concluded that the high eigenvector centralization in the network may affect the missing link prediction performance.

Ağ Topolojisi İlişkisi ile Bağlantı Tahmin Yöntemlerinin Performanslarının Keşfi

Anahtar kelimeler
Bağlantı tahmini; Performans değerlendirmesi; Ağ topolojisi; Graf gömme

Öz

Karmaşık ağ analizinde öne çıkan konulardan biri, ağ tabanlı öneri sistemlerinin veya eksik bağlantıların bulunmasının önemli bir bileşeni olan bağlantı tahminidir. Literatürde iki düğüm arasında bağlantı bulunma şansını ölçümlemeye dayanan birçok farklı bağlantı tahmini yöntemi vardır. Bu yöntemler ağın farklı topolojik özelliklerini kullanır. Çok farklı stratejiler kullanan yöntemler bulunmasına rağmen, önceki çalışmalar yalnızca yöntem başarısına odaklanmış ama bu yöntemlerin performansının ağın topolojisi ile ilişkisini yeteri kadar incelememiştir. Bu çalışmanın ana motivasyonu farklı ağ topolojilerininin bağlantı tahminindeki rolünü bir ortaya koymaktır. Böylece ağın topolojik özelliklerine göre bağlantı tahmin yöntemi seçimi özelleştirilebilir. Çalışmanın iki temel katkısı, ilk olarak, büyük bir deney düzeneğinde farklı topolojik özelliklere sahip sosyal, biyolojik ve bilgi ağlarında iyi bilinen performans ölçümleriyle farklı bağlantı tahmin yöntemlerini karşılaştırmak ve ikincisi, bağlantı tahmin yöntemlerinin performansı ile ağ topolojisi arasındaki olası ilişkinin incelenmesi olarak sıralanabilir. Sonuçlara göre, ağ topolojisine bakılmaksızın küresel yöntemlerin diğerlerinden daha başarılı olduğunu gördük. Ayrıca, ağda özvektör merkezileşmesinin yüksek olmasının eksik bağlantı tahmin performansını etkileyebileceği sonucuna ulaşıldı.

1. Introduction

Link prediction is one of the most studied topics in complex network analysis. Finding missing links in a network allows one to solve many problems in various applications. For example, predicted links in a network of item connections can be used for making product recommendations (Kaya 2020, Li *et al*. 2014). Or, on a social network, link prediction allows for the formation of new friendships (Zareie and Sakellariou 2020, Liben-Nowell and Kleinberg 2007). Due to this popularity, numerous studies have been conducted on it.

We can categorize the methods into three parts. First, traditional methods calculate a score for all possible links, i.e. pairs of nodes, that are not seen in the network, based on a strategy (Martínez *et al*. 2016, Lü and Zhou 2011). Afterwards, these scores are ranked from largest to smallest, and the desired number of links is selected in order to be predicted. Methods that find scores, adamic adar, resource allocation, etc., are calculated with mathematical formulas using network topology. The score depends on the likelihood of a studied link. Those methods differ from each other according to the strategies they use. Second, as in the traditional methods, the scores of all missing links are calculated by using different methods. However, this time, instead of making a simple ranking, all of these scores are solved together with a machine-learning model (De Sá and Prudêncio 2011, Malhotra and Goyal 2021). In other words, the link prediction problem is tackled by constructing a machine-learning model of different features of possible links. These methods differ in terms of both the supervised learning technique and the artificial learning algorithms they use. Finally, in recent works, graph embedding techniques are held (Wang *et al*. 2021, Rossi *et al*. 2021). In many studies, researchers develop new metrics or methods according to the needs of their application. Those methods use different topological elements in the network. Nevertheless, the comparison of the performance of link prediction methods has not been extensively studied in the literature. However, real-world networks have different topological structures. The link prediction method, which gives successful results in some types of networks, may not be successful in other types. Thus, there is a need for an exhaustive study to reveal the possible relation between link prediction methods with network structure.

In this study, we test sixteen well–known link prediction methods on eleven real-world networks having different topologies and evaluate their performances. We separate the link prediction methods into three categories: local, global, and embedding. We did not focus on developing fine-tuned algorithms but rather applied the methods roughly as their traditional ranking strategy. Our main goal is to reveal which type of method can be successful in networks with which topological properties. As a matter of fact, we do not only compare the success of the methods but also focus on finding the possible relationship between these successes and the topological properties of the networks.

Our main contributions are first, comparing different link prediction methods with well-known performance metrics on both social, biological, and information networks having different characteristics in a large experimental setup; and second, examining the possible relationship between the performance of link prediction methods with the network topology. In the following, we first explain the details of link prediction methods, and then we explain the experimental setup and the results in detail. In this part, we also discuss the relationship between network topology and link prediction performance. Finally, we summarize the work by giving some future perspectives.

2. Material and Method

The link prediction experiments in this work are done with traditional semi-supervised learning techniques (Lü and Zhou 2011). First, a training network is assigned by removing an amount of randomly selected links from the original network. Then the link prediction metrics are calculated for all

missing links in the training network, including the removed links and already unseen links from the original network. For each metric, the first links that receive the highest scores are assigned as predicted links. The predicted links are evaluated as true or false predictions by determining whether they are included in the original network. Finally, the performance of the link prediction methods is measured by their precision, or AUC scores (Lü and Zhou 2011). In the next part, we explain the similarity/distance metrics that assess the likelihood of having a link between any pair of nodes u,v. We categorized those metrics according to their essential techniques, which are used for link prediction tasks.

2.1 Link Prediction with Local Information

The common strategy behind the link prediction methods we describe here is triadic closure principle (TCP) (Kovács *et al*. 2019). This principle favors the tendency of having links between two nodes if they have more common connections. TCP looks for completely local information related to the first-level neighborhood of the compared nodes. Let us denote that neighborhood $N(u)$, or (N_u), of a node u is the set of nodes directly connected to u, $N(u) = \{v \in V | (u,v) \in L\}$ with L is the set of links.

Definition 1. (Newman 2001) Common Neighbors (CN) is the size of the set of common neighbors between any two nodes.

$$s(u,v) = |N_u \cap N_v| \quad (1)$$

More generally, the higher the number of degrees, the more possible to have higher CN for the nodes. Thus, CN has a tendency of being high for any two hub nodes.

Definition 2. (Adamic and Adar 2003) Adamic Adar (AA) counts the total number of neighbors of all common neighbors. However, it depresses the score by logarithmic function for demoting the scores of higher degree nodes. Shortly, it penalizes the scores for hub neighbors.

$$s(u,v) = \sum_{i \in N_u \cap N_v} \frac{1}{\log_2(|N_i|)} \quad (2)$$

Definition 3. (Zhou and Zhang 2009) Resource Allocation (RA) is almost the same with AA. It also counts the total number of neighbors of all common neighbors. However, differently from AA, it considers the degrees not their logarithms.

$$s(u,v) = \sum_{i \in N_u \cap N_v} \frac{1}{|N_i|} \quad (3)$$

Definition 4. (Jaccard 1912) Jaccard Coefficient (JC) originally developed for comparing two sets. It is the ratio of the number of common neighbors to the number of all neighbors of two nodes.

$$s(u,v) = \frac{|N_u \cap N_v|}{|N_u \cup N_v|} \quad (4)$$

Definition 5. (Dice 1945, Sørensen 1948) Sørrenson/Dice Index (Dice) measures the common parts of the neighborhoods and normalizes it with the size of the neighborhoods of two studied nodes. If the neighborhoods have many nodes in common but also the common neighbors have many other links to the outside of the common neighborhood, Dice becomes lower than JC. It penalizes being a hub as well.

$$s(u,v) = \frac{2 \cdot |N_u \cap N_v|}{|N_u| + |N_v|} \quad (5)$$

Definition 6. (Cannistraci *et al*. 2015) Cannistraci-Alanis-Ravasi index (CAR) is the sum of the number of common neighbors of two nodes each having neighbors in common with those nodes.

$$s(u,v) = \sum_{i \in N_u \cap N_v} 1 + \frac{|N_u \cap N_v \cap N_i|}{2} \quad (6)$$

Definition 7. (Cannistraci *et al*. 2015) CAR-based Adamic and Adar (CAA), is a hybrid metric of the AA with CAR strategy. It merges two strategies of favoring clique-like neighborhoods with the penalization of being hub.

$$s(u,v) = \sum_{i \in N_u \cap N_v} \frac{|N_u \cap N_v \cap N_i|}{\log_2(|N_i|)} \quad (7)$$

Definition 8. (Cannistraci *et al*. 2015) Another hybrid metric is CAR-based Resource Allocation (CRA). It merges the two strategies of CAR with RA, which are explained previously.

$$s(u,v) = \sum_{i \in N_u \cap N_v} \frac{|N_u \cap N_v \cap N_i|}{|N_i|} \quad (8)$$

2.2 Link Prediction with Global Information

This category is dedicated to the link prediction methods, which do not use TCP, but they still use network-related topological properties. In local methods, the metrics completely focus on the common neighborhood, which was based on the TCP idea for link prediction. Here, we explain the metrics using other strategies related to network topology.

Definition 9. (Newman 2001) Preferential Attachment (PA), is the multiplication of degrees of two nodes. PA promotes the nodes having higher degree. It assumes that the famous nodes should have more probability of connecting with each other.

$$s(u,v) = |N_u| \cdot |N_v| \quad (9)$$

Definition 10. (Cannistraci *et al*. 2015) CAR-based Preferential Attachment (CPA) merges the strategies of CAR and preferential attachment.

$$s(u,v) = e_u \cdot e_v + e_u \cdot CAR(u,v) + e_v \cdot CAR(u,v) + CAR(u,v)^2 \quad (10)$$

With $e_u = |N_u \setminus (N_u \cap N_v)|$ and $e_v = |N_v \setminus (N_u \cap N_v)|$ is the number of the neighbors of u that are not common neighbors of u and v, and $CAR(u,v)$ is the CAR score between nodes u and v.

Definition 11. (Kovács *et al*. 2019) L3 link predictor (L3), considers network paths of length three.

$$s(u,v) = \sum_{i,j} \frac{a_{ui} \cdot a_{ij} \cdot a_{jv}}{\sqrt{k_i \cdot k_j}} \quad (11)$$

Here, a_{ui} is 1 if there is a link between the nodes u and i. And k_i is the degree of node i. Since the third level neighbors numbers are exponentially larger than the second level ones, the metric applies a degree normalization strategy. It also avoids the biased high scores coming from the hub nodes, which are naturally building shortcuts, and increases the number of third level neighbors for entire nodes.

Definition 12. (Clauset *et al*. 2008) Hierarchical Random Graph model (HRG), is originally a method of producing general hierarchical structure of a given network. Differently from producing an overfitted one single dendrogram, which only explains the hierarchical structure of the studied state of the network, HRG uses MCMC sampling of hierarchical models around the optimum one and produces the likelihoods for each member from the sample. In fact, those members are the dendrogram with associated probabilities. The model decomposition is then used for link prediction. For any node pairs, their prediction score is the average probability of connection within these dendrograms.

Definition 13. (Lü *et al*. 2008) Structural perturbation method (SPM) focuses on perturbing the adjacency matrix and observing the change of eigenvalues provided the fixed eigenvectors. This technique is similar to the first-order perturbation in quantum mechanics. It produces the scores, which are similar to previously explained similarities, for all links based on the perturbation of removal links from the adjacency matrix of the original network.

2.3. Link Prediction with Embedding

Beyond the usage of TCP principle or network structural information, there are other techniques of link prediction, which transform the network into the lower dimensional Euclidean space. Such a transformation is called graph embedding. There are several different techniques of graph embedding. Here we focus on the ones, which are used for link predictions.

Definition 14. (Tenenbaum *et al*. 2000, Kuchaiev *et al*. 2009) Isometric mapping (ISOMAP) uses one of the traditional graph embedding techniques. The studied network, $G = (V, L)$, is first transformed to a distance matrix D of its nodes in which each member d_{uv} of D is the shortest distance between the nodes u and v from V. Then D is transformed to a lower dimensional matrix $L \in \mathbb{R}^l$ with Multidimensional scaling based on non-lineaire embedding method, MDS. Here l is the new dimension that G is transformed to. MDS tries to

keep original distance d_{uv} between the node pairs and generates new vectors $x_1, x_2, \cdots, x_n$ for each node whose lengths are l. $x_1, x_2, \cdots, x_n$ is found as a minimizer of some cost function $min_{x_1,x_2,\cdots,x_n} \sum_{u,v}(d_{uv} - \|x_u - x_v\|)^2$.
Once MDS generates new lower dimensional vectors for each node, then ISOMAP calculates basic euclidean distance between the nodes as their dissimilarities.

Definition 15. (Belkin and Niyogi 2001) Laplacian Eigenmaps (LEIG) uses a minimization function that can be solved by the generalized eigenvalue problem. Hence, it first generates the laplacian matrix of the original network, and then spectral decomposition of the corresponding laplacian matrix is computed. LEIG finds l eigenvalues and eigenvector with l is the number of new dimensions. After embedding, the link prediction is again done by regarding euclidean distance of the node pairs.

Definition 16. (Cannistraci *et al*. 2013) Centered and non-centered Minimum Curvilinear Embedding (MCE) and (ncMCE) respectively, are two network embedding techniques using the distances in the minimum spanning tree of studied networks. Both methods first generate the minimum spanning tree, MST of corresponding G, and then computes the distances of every pair of nodes in the MST. These distances unders the form of distance matrix are called the kernel. In the algorithm if centering is not chosen, the ncMCE performs an economy size singular value decomposition of the distance matrix. Otherwise, an algebraic operation is performed for kernel centering at first and then the decomposition is done. Finally the new lower dimensional space of G is produced by the transpose of the product of computed singular values with right singular vectors with the algebraic corrections.

3. Experiments and Results

3.1. Datasets and their Topological Properties

We used eleven famous networks in our experiments, which are taken from (Rossi and Ahmed 2015;, Kunegis 2013). Some of them are directly social or anthropological networks representing the relation between a group of humans while some others are biological or transport. Table 1 shows their names and topological properties. Details of these topological properties can be obtained from (Watts and Strogatz 1998, Albert and Barabási 2002, Newman 2003). We evaluate the node number, a.k.a. network size (n), the link number (m), the average path length (l), the transitivity (T), the average (⟨k⟩), minimum (min (k)) and maximum (max (k)) degrees, the diameter (diam), the radius (rad), the density (δ) , the degree of centralization (DC), the betweenness centralization (BC), the closeness centralization (CC) and the eigenvector centralization (EC) metrics. Some have high link density, while others have high transitivity. Some of them have nodes with high EC, that is, they are popular in the network. . None of the topological properties listed in this table individually describe the network, but a few do give us an idea of its structure. For example, the Tribes network has a structure of local clusters with low EC and high T. It probably has no central or hierarchical formation. However, gene-fusion or C-Elegans networks, on the contrary, have more recursive and centralized dynamics, with high ECs and relatively low Ts. We will examine the effects of these possible topological differences on connection estimation in the next sections. Briefy, Table 1 shows us that our experiments are performed on a large set with a wide variety of properties since the networks show different characteristics.

3.2. Link Prediction Results

First, we examine the results of link prediction metrics in our preliminary experiments. In these experiments, 20% of the links in each network were randomly selected and removed by using the cross-validation method. The score of each link prediction metric is then calculated for every possible node pair, as a missing link. The missing links that receive the highest scores are output as estimated links. Afterwards, estimated links were compared with the previously extracted links. The number of links to be predicted is too small when compared to the total number of possible links in the network. It is to find out about rarely occurring events. One of the

most suitable performance metrics for these cases is the precision score and AUC (Lü and Zhou 2011). The precision and AUC scores are shown in Table 2 and Table 3.

In general, precision scores are low. This is an expected result in rank-based link prediction. Many nodes in the network may have a similar topological position. Their link prediction scores may be the same, or too close. As a result, many correct links can be ignored because the ranking is done and only a certain number of links can be selected. Therefore, precision results are good to compare the metrics without focusing on the values. Accordingly, not a single metric or a family of metrics stands out. Metrics with different strategies such as RA, CAA, PA, HRG, and SPM yield different results in different networks.

The AUC, on the other hand, does not take true positives or false Negatives into account. Instead, it measures the performance based on the scores the candidate links get. AUC results can have high values, as can be seen in Table 3. Accordingly, RA and HRG stand out compared to other metrics. In many networks, these two metrics allow for more successful link prediction. In contrast, embedding methods using the spectral properties of the networks, especially LEIG and MCE, give the lowest scores in most of the studied networks.

Evaluating both precision and AUC scores together, we find that first, different methods perform differently in the same network, and second, the same method performs differently in different networks. The first finding may be due to the fact that the methods have different link prediction strategies. However, the fact that the same method gives different results in different networks may be due to the possible topological difference between the networks. If these networks were composed of regular lattice graphs, there would be no structural difference between them. There would only be a difference in the number of nodes and links. In lattice, since all nodes will be equivalent in terms of their positions in the network, it was expected that the link prediction performances made in different networks would be equal. However, the networks we work with have different centrality measures, different mean degrees, different transitivity, etc., which means the nodes have different structural positions. It seems the performance of a method is affected by such a difference. We examine this fact in detail in the next sections.

Table 1. The topological properties of the networks

Network name	n	m	l	T	$\langle k \rangle$	$min(k)$	$max(k)$	$diam$	rad	δ	DC	BC	CC	EC
Adjnoun	112	425	2.54	0.16	7.59	1	49	5	3	0.068	0.37	0.23	0.43	0.84
C Elegans	453	2025	2.66	0.12	8.94	1	237	7	4	0.020	0.50	0.48	0.54	0.92
Contact	274	2124	2.42	0.57	15.50	1	101	4	2	0.057	0.31	0.14	0.38	0.78
Dolphins	62	159	3.36	0.31	5.13	1	12	8	5	0.084	0.11	0.21	0.23	0.74
Gene fusion	292	279	3.90	0.00	1.91	1	34	9	3	0.007	0.11	0.08	0.46	0.97
Jazz	198	2742	2.24	0.52	27.70	1	100	6	4	0.141	0.37	0.15	0.38	0.74
Karate	34	78	2.41	0.26	4.59	1	17	5	3	0.139	0.38	0.41	0.30	0.65
Les Miserable	77	254	2.64	0.50	6.60	1	36	5	3	0.087	0.39	0.56	0.52	0.77
Moreno	242	923	2.47	0.25	7.63	2	28	5	3	0.032	0.08	0.02	0.22	0.95
Tribes	16	58	1.54	0.53	7.25	3	10	3	2	0.483	0.18	0.04	0.21	0.31
US-Air	332	2126	2.74	0.40	12.81	1	139	6	3	0.039	0.38	0.20	0.46	0.86

Table 2. The link prediction methods' precision results

	Local Information									Global Information				Graph Embedding		
Network name	**CN**	**AA**	**RA**	**JC**	**DICE**	**CAR**	**CAA**	**CRA**	**PA**	**CPA**	**L3**	**HRG**	**SPM**	**ISOMAP**	**LEIG**	**MCE**
Adjnoun	0.05	0.03	0.05	0.02	0.02	0.03	0.03	0.03	**0.06**	0.04	0.05	0.02	**0.06**	0.01	0.02	0.01
C Elegans	0.05	0.04	**0.10**	0.03	0.06	0.03	0.04	0.06	0.02	0.02	0.00	0.01	0.04	0.00	0.00	0.00
Contact	0.22	0.23	0.71	0.06	0.07	0.63	**0.67**	0.64	0.30	0.64	0.35	0.31	0.17	0.01	0.01	0.01
Dolphins	0.11	0.11	0.03	0.06	0.07	0.08	0.04	0.02	0.01	0.05	0.03	**0.24**	0.20	0.02	0.04	0.01
Gene fusion	0.00	0.00	0.00	0.00	0.00	0.00	**0.02**	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Jazz	0.41	0.43	0.45	0.39	0.37	0.36	0.50	0.48	0.06	0.35	0.01	0.17	**0.59**	0.02	0.01	0.03

Karate	0.10	0.02	0.12	0.01	0.02	0.08	0.01	**0.37**	0.06	0.25	0.06	0.06	0.13	0.02	0.08	0.01
Les Miserable	0.52	0.48	**0.64**	0.16	0.15	0.58	0.52	0.66	0.07	0.43	0.55	0.34	0.62	0.04	0.15	0.03
Moreno	0.08	**0.09**	**0.09**	0.07	0.07	**0.09**	0.07	0.07	0.01	0.02	0.01	0.07	0.08	0.02	0.01	0.04
Tribes	0.06	0.06	0.08	0.20	0.60	0.08	0.04	0.09	0.04	**0.11**	0.09	0.06	0.08	0.37	0.06	0.04
US-Air	0.08	0.08	0.15	0.03	0.04	0.13	0.16	0.09	0.05	0.32	0.14	**0.38**	0.12	0.03	0.03	0.02

Table 3. The link prediction methods' AUC results

	Local Information									Global Information				Graph Embedding		
Network name	**CN**	**AA**	**RA**	**JC**	**DICE**	**CAR**	**CAA**	**CRA**	**PA**	**CPA**	**L3**	**HRG**	**SPM**	**ISOMAP**	**LEIG**	**MCE**
Adjnoun	0.71	0.65	0.69	0.66	0.67	0.70	0.68	0.60	0.73	0.73	**0.75**	0.72	0.63	0.56	0.67	0.51
C Elegans	0.91	0.93	**0.96**	0.80	0.78	0.88	0.88	0.88	0.77	0.75	0.51	0.85	0.84	0.66	0.50	0.65
Contact	0.95	0.95	**0.97**	0.92	0.92	0.96	0.96	0.96	**0.97**	0.96	0.95	0.95	0.83	0.66	0.53	0.63
Dolphins	0.87	0.89	0.73	0.75	0.82	0.84	0.75	0.73	0.75	0.75	0.86	**0.90**	0.78	0.83	0.88	0.70
Gene fusion	0.53	0.69	0.65	0.62	0.80	0.76	0.83	0.81	0.96	0.92	0.72	**0.97**	0.75	0.91	0.53	0.88
Jazz	0.96	0.97	**0.98**	0.96	0.97	0.96	0.96	0.97	0.80	0.94	0.50	0.88	**0.98**	0.78	0.51	0.76
Karate	0.86	0.60	**0.92**	0.57	0.73	0.49	0.48	0.77	0.51	0.29	0.91	0.74	0.85	0.72	0.43	0.58
Les Miserable	0.95	0.97	**0.98**	0.95	0.94	0.96	0.95	0.96	0.85	0.77	0.94	0.95	0.96	0.82	0.91	0.71
Moreno	0.89	0.89	0.90	0.91	0.90	0.81	0.77	0.77	0.61	0.54	0.63	**0.92**	0.84	0.87	0.54	0.91
Tribes	0.51	0.56	0.62	0.75	**0.87**	0.63	0.21	0.60	0.34	0.54	0.68	0.41	0.63	0.84	0.49	0.32
US-Air	0.65	0.73	0.71	0.69	0.69	0.45	0.45	0.37	0.75	0.85	0.84	**0.93**	0.88	0.71	0.52	0.50

3.3. Link Prediction Methods' Sensitivity

In our second experiment, for the train set, we removed from 10% to 90% of the links in the network, respectively, with 10% increments. We run all the methods in each link extraction and observe the performances. Related plots can be seen in Figure 1 and Figure 2. In these figures, the x-axis shows the link ratio extracted from the trained network, and the y-axis shows the value of the performance metric. According to precision results, all metrics, without exception, yield more successful results as the link ratio in the test set increases. This is due to the precision metric itself. The true positives increase as the number of links that need to be estimated increases. When more links need to be predicted, multiple links showing the same score can now be included in the true positive set. Therefore, an increase in precision results is an expected result. Here, in order to evaluate the true success of the methods, it is necessary to focus on the curve of the increase. A simple linear increment is due to the precision properties. A good method is expected to give a result that is as stable or as close to no slope as possible. When the plots are examined, no method produces such a result. A less-linear increase is obtained by CN, HRG, and L3 for Dolphin, Karate, and US-Air networks, respectively.

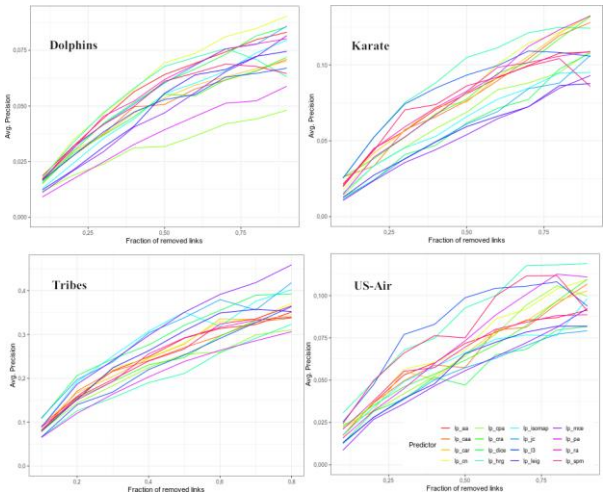

Figure 1. The average precision results over four networks based on different sized test data. The X-axis is the rate of links reserved for testing; the y-axis is the average precision value.

AUC scores are more stable (see Figure 2). Here, all the methods for Dolphin and Tribe networks show a more stable performance, while the success of the methods differs as the predicted link number increases for the other two networks. In these plots, the more stable and higher the performance curve of a method, the better it is. Accordingly, we observe that HRG is the most successful method in all networks except Tribe. It seems that the L3 method performs more unsuccessfully as the link ratio increases in almost all networks. It seems to be strongly affected by missing links in the network

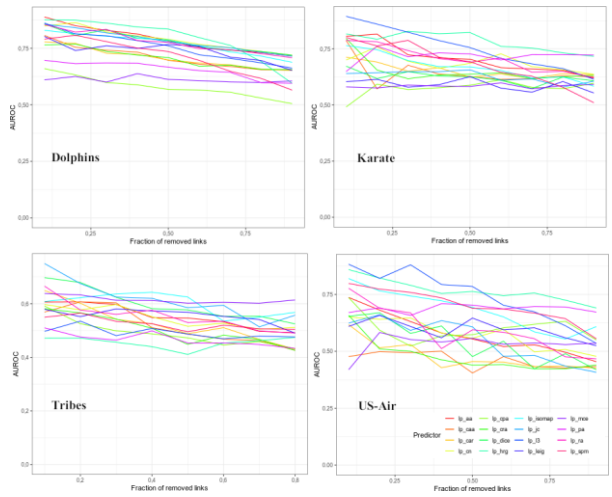

Figure 2. The AUC results over four networks based on different sized test data. The X-axis is the rate of links reserved for testing; the y-axis is the AUC value.

In the experiments conducted in this section, we observed that the precision of all methods increases as the removed link ratio is increased, while the AUC declines smoothly. As we mentioned in the previous section, since the studied networks are not regular lattice graphs and each has its own topological structure, removing links from these networks causes changes in their topology. In our experiments, we both used two different performance metrics and different networks. In this way, changing input and performance metrics can cause a cross-effect when evaluating the success of methods. However, in this setup, we see that the success of all sixteen link prediction methods without exception changes when the removed link amount changes, that is, when the topology of the network is gradually degraded/affected. Accordingly, we can claim that all methods are clearly affected by the topographical change, even if there might be a crossover effect coming from input or metric change. We do not measure this possible crossover effect for now because it is out of the scope of this work.

3.4. Link Prediction Methods' Correlation with Network Topology

We analyze whether different topological properties of networks have an effect on link prediction methods by examining their correlations.

The correlation results between the performance score of each method as AUC or Precision in different networks and the different topological properties, and the related heat maps are shown in Figure 3, left and right plots, respectively. According to AUC results, HRG and EC are highly positively correlated. PA and MCE are also correlated with EC. A high EC shows us that some nodes are quite popular in the network while many others are not, so there may be a tree-like structure. HRG uses a strategy relying on a hierarchical structure in the network. PA increases the score according to the number of neighbors. It is thus reasonable for these methods to be highly correlated with EC. On the other hand, we can say that the success of HRG and PA decreases as the network density increases.

Correlation results with precision show the relationship between the ability of link prediction methods to detect missing links and topological properties. Accordingly, the success of PA and CAA increases as the density and the minimum degree increase, and the success of these two methods decreases as EC increases. The most interesting result here is that AUC scores of PA are positively correlated with EC, while Precision scores are negatively correlated. PA favors putting links between popular nodes. It is therefore reasonable that the score result increases with EC increase. However, the negative correlation on precision means that there is a decrease in the true positive link prediction rate with the increase of EC. Most of the links predicted by PA are tied to higher degree nodes. This can result in popular nodes being linked to unpopular ones if there are fewer popular node pairs than the number of missing links in the network. As a result, this may cause a negative correlation with EC.

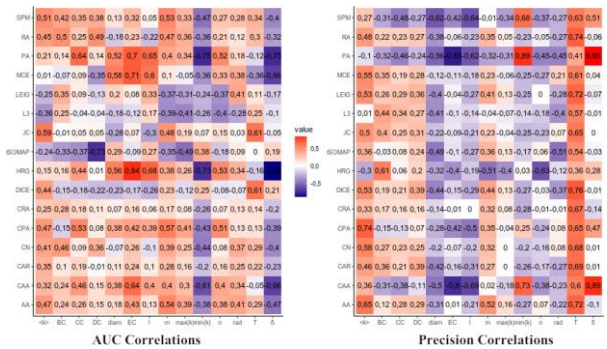

Figure 3. Heat maps of the Pearson correlation coefficient. Precision (right) and AUC (left) correlations between all link prediction methods and all topological properties. Experiments are carried out with the 20% test data of the links.

We showed which topological properties each method might be related to with correlation heat maps. As explained, some methods were found to be related one-to-one with some topological properties. But many methods do not seem to be correlated to any topological properties. In particular, the correlations between local methods and topological properties always seem low. Local methods use TCP, which takes into account only the first-level neighborhood of two nodes. Therefore, it may be possible that there is no relation between the centralizations, which gives information about the global structure of the network. However, the fact that they are not particularly related to transitivity should be investigated because transitivity is a property related to triangular connections in the first-level neighborhood of the node. Here, in order to measure correlation, we consider the global values such as the average of the local properties such as transitivity or degree. This may obscure the possible relationship between transitivity and link prediction. In addition, while the topological properties individually can provide information about the network structure, a few of them together can create cross-effects, allowing us to explain the structure of the network (for example, high EC and low T together can mean a hierarchical structure). The correlations we measured here only reflect the possible relations between singular properties and singular methods. This may also cause us not to see this result in correlation, even if some methods are affected by the combination of some topological properties.

3.5. Discussion

When all the experiments are evaluated together, it has been observed that there may be a relationship between the topological properties of the studied network and the performance of the link prediction method. When the details of this relationship are examined, it is concluded that some methods, such as HRG, may be affected by single topological properties such as EC. In the study, possible relationships were examined between each method and each topological property, on the networks in the experimental setup. In networks, it is possible that more than one topological property together creates new dynamics. However, it is not easy to measure this fact. Probably for this reason, the relationship between the performance of methods using local information and the topology of the network could not be resolved. A detailed analysis of this possible relation can be done with a regression. More specifically, the precision or AUC score of each method can be described by a regression line that takes input from the topological properties studied here. However, the fact that several topological properties together create a non-linear dynamic may still not be measurable even with regression. This difficulty arising from the nature of complex networks can be solved by examining each system in detail and temporally, with new topological properties (and/or hybrid topological properties) that will explain the structural features of that system. This study is the basis for revealing the topology-link prediction relationships. Later on, these relationships can be detailed in many respects. Some experiments conducted in this study can form the basis for studies in many different disciplines, each of which can be a separate subject in physics, mathematics, social or computer sciences.

4. Conclusion

We carried out a series of experiments to evaluate the performance of sixteen different link prediction methods used in traditional score ranking-based prediction. Experiments were performed with eleven networks having different characteristics. Performances were measured by AUC and Precision. Ultimately, we found that HRG and RA yielded the most successful results. Of these, HRG is less dependent on the number of missing links in the network. In addition, the possible relationship between these methods and the network topology was also examined in the experiments. Again, HRG was found to be correlated with the eigenvector centralization of the network. This method can be

advantageous if prediction is to be done on a network dominated by popular nodes.

The results demonstrate that Precision and AUC scores can exhibit different behaviors. These experiments showed us that an objective evaluation of prediction results is a complex problem with various parameters. In the following stages, new and more different networks can be added to the experimental dataset. The relationship between topological properties and prediction methods can be examined in more depth by using other metrics such as recall or AUPR. Thus, partial hybrid solutions or ensemble solutions resulting in higher success can be developed according to the topology of the network.

5. References

Adamic, L. A., & Adar, E., 2003. Friends and neighbors on the web. *Social networks*, **25**, 211-230.

Albert, R., & Barabási, A. L., 2002. Statistical mechanics of complex networks. *Reviews of modern physics*, **74**, 47-97.

Belkin, M., & Niyogi, P., 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS'01). MIT Press, Cambridge, MA, USA, 585–591.

Cannistraci, C. V., Alanis-Lobato, G., & Ravasi, T., 2013. Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics*, **29**, 199-209.

Cannistraci, C. V., Alanis-Lobato, G., & Ravasi, T., 2015. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Scientific reports*, **3**, 1-14.

Clauset, A., Moore, C., & Newman, M. E., 2008. Hierarchical structure and the prediction of missing links in networks. *Nature*, **453**, 98-101.

De Sá, H. R., & Prudêncio, R. B., 2011. Supervised link prediction in weighted networks. The 2011 International Joint Conference on Neural Networks, IEEE, 2281-2288,

Dice, L. R., 1945. Measures of the amount of ecologic association between species. *Ecology*, **26**, 297-302.

Gleiser, P. M., & Danon, L., 2003. Community structure in jazz. *Advances in complex systems*, **6**, 565-573.

Jaccard, P., 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, **11**, 37-50.

Kaya, B., 2020. Hotel recommendation system by bipartite networks and link prediction. *Journal of Information Science*, **46**, 53-63.

Kovács, I. A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., ... & Barabási, A. L., 2019. Network-based prediction of protein interactions. *Nature communications*, **10**, 1-8.

Kuchaiev, O., Rašajski, M., Higham, D. J., & Pržulj, N., 2009. Geometric de-noising of protein-protein interaction networks. *PLoS computational biology*, **5**, 1-10.

Kunegis, J., 2013. Konect: the koblenz network collection. In Proceedings of the 22nd international conference on world wide web, 1343-1350.

Liben-Nowell, D., & Kleinberg, J., 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, **58**, 1019-1031.

Li, J., Zhang, L., Meng, F., & Li, F., 2014. Recommendation algorithm based on link prediction and domain knowledge in retail transactions. *Procedia Computer Science*, **31**, 875-881.

Lü, L., & Zhou, T., 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, **390**, 1150-1170.

Lü, L., Pan, L., Zhou, T., Zhang, Y. C., & Stanley, H. E., 2015. Toward link predictability of complex networks. *Proceedings of the National Academy of Sciences*, **112**, 2325-2330.

Malhotra, D., & Goyal, R., 2021. Supervised-learning link prediction in single layer and multiplex networks. *Machine Learning with Applications*, **6**, 1-9.

Martínez, V., Berzal, F., & Cubero, J. C., 2016. A survey of link prediction in complex networks. *ACM computing surveys*, **49**, 1-33.

Newman, M. E., 2001. Clustering and preferential attachment in growing networks. *Physical review E*, **64**, 1-13.

Newman, M. E., 2003. The structure and function of complex networks. *SIAM review*, **45**, 167-256.

Rossi, R., & Ahmed, N., 2015. The network data repository with interactive graph analytics and visualization. In Twenty-ninth AAAI conference on artificial intelligence, AAAI Press, 4262-4293

Rossi, A., Barbosa, D., Firmani, D., Matinata, A., & Merialdo, P., 2021. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data*, **15**, 1-49.

Sørensen, T. J., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab Biologiske Skrifter*, **5**, 1-34.

Tenenbaum, J. B., Silva, V. D., & Langford, J. C., 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319-2323.

Wang, M., Qiu, L., & Wang, X., 2021. A survey on knowledge graph embeddings for link prediction. *Symmetry*, **13**, 1-29.

Watts, D. J., & Strogatz, S. H., 1998. Collective dynamics of 'small-world'networks. *Nature*, **393**, 440-442.

Zareie, A., & Sakellariou, R., 2020. Similarity-based link prediction in social networks using latent relationships between the users. *Scientific Reports*, **10**, 1-11.

Zhou, T., Lü, L., & Zhang, Y. C., 2009. Predicting missing links via local information. *The European Physical Journal B*, **71**, 623-630.