Available online at sjuoz.uoz.edu.krd

Science Journal of University of Zakho

Vol. 10, No. 4, pp. 235 –257, October-December 2022

p-ISSN: 2663-628X
e-ISSN: 2663-6298

LINK PREDICTION IN CO-AUTHORSHIP NETWORKS: A REVIEW

Hajar Ali Hasin [a,*], and Diman Hassan [a]

[a] Dept. of Computer Science, CCNP Research Lab, Faculty of Science, University of Zakho, Zakho, Kurdistan Region, Iraq – (email: hajar.hasan@staff.uod.edu.krd)

Received: *12 Oct., 2022 /* ***Accepted:*** *17 Oct., 2022 /* ***Published:*** *07 Nov., 2022* ***https://doi.org/10.25271/sjuoz.2022.10.4.1040***

ABSTRACT:

Besides social network analysis, the Link-Prediction (LP) problem has useful applications in information retrieval, bioinformatics, telecommunications, microbiology, and e-commerce as a forecast of future links in a given context to find what possible connections are based on a local and global statistical analysis of the given graph data. However, in Academic Social Networks (ASNs), the LP issue has recently attracted a lot of attention in academia and called for a variety of link prediction techniques to predict co-authorship among researchers and to examine the rich structural and associated data. As a result, this study investigates the problem of LP in ASNs to forecast the upcoming co-authorships among researchers. In a systematic approach, this review presents, analyses, and compares the primary taxonomies of topological-based, content-based, and hybrid-based approaches, which are used for computing similar scores for each pair of unconnected nodes. Then, this study ends with findings on challenges and open problems for the community to work on for further development of the LP problem of scholarly social networks.

KEYWORDS: Link Prediction, Co-authorship Networks, Topological-based measures, Content-based measures. INTRODUCTION

1. INTRODUCTION

Social networks (SN) are used to share thoughts, preferences, and interests. They produce a complex graph that reveals the relationships inside the network as a result. It is defined as a finite set of nodes and the edges that connect between them focuse on the properties of the nodes and the connections between them (E Fonseca et al., 2016)(Tabassum et al., 2018). The Scientific Collaboration Network (SCN) is a popular social network that is regarded as a feature of current academic research in which scientists are regarded as part of such networks (Sonnenwald, 2007)(Newman, 2004). These members are seeking solutions to many challenging problems that demand multidisciplinary approaches, such as social, political, economic, and technological issues (Sonnenwald, 2007). By pooling resources, ideas, and information, cooperative researchers in the SCN can improve research efficiency and produce innovations (Smith & Sotala, 2011). One of the common SCNs is the co-authorship network which is frequently used to evaluate and analyze scientific collaboration patterns where nodes represent authors or research groups. These nodes are connected when they share the authorship of a given article (Newman, 2004).

In network theory, the problem of link prediction (LP) was coined first in 2007 (Liben-Nowell & Kleinberg, 2007) and defined as the problem of predicting the presence of a connection between two entities in a co-authorship network, known before as the problem of Preferential Attachment (PA) (Jaya Lakshmi & Durga Bhavani, 2017)(Barabási & Albert, 1999)(T. Zhou et al., 2009), the triangle concept (Newman, 2001) and Adamic/Adar algorithm (AA) (Adamic & Adar, 2003). As a result, statistical graph analysis methods were presented such as the common neighbor (CN) method was created and the proposal of additional algorithms, including the Hub Promoted Index (HPI), the Hub Depressed Index (HDI) (Ravasz et al., 2002), the Leicht-Holme-Nerman-1 (LHN1) index (Leicht et al., 2006), Resource Allocation (RA) (T. Zhou et al., 2009), Jaccard Similarity Coefficient (JC) and Salton Cosine Similarity (SA) (T. Zhou et al., 2009)(Jaccard, 1982), which have affected the LP research.

The LP problem can be formulated as follows: considering a real network G = (V, E) at time t, where G represents a graph or network, V represents a set of nodes or vertices, and E denotes a set of edges or links at time t. V(G) and E(G), represent the group of nodes and edges in a graph, respectively. Predicting the edges that are most likely to form between t and t + 1 (t < t + 1) is the goal of the LP (Liben-Nowell & Kleinberg, 2007)(Yuliansyah et al., 2020)–(X. Liu et al., 2013). Thus, a node pair will be generated to predict the future or missing link during interval time t to t+1, as shown in Figure 1.

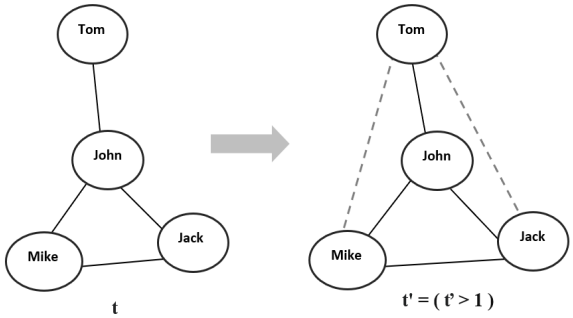

Figure 1. Graphical Representation of Missing LP; Dashed Lines Depict Possible

Predicting friendship links among users in a social network (Bhattacharyya et al., 2011), and co-authorship links in a citation network (Börner et al., 2004; Martin et al., 2013; Wallace et al., 2012), are all examples of LP. It is also employed in e-commerce (Bahabadi et al., 2014), where it is used to recommend items to users. LP can be used for recording deduplication in citation database curation. It also has been used to predict protein-protein interactions in bioinformatics (PPI) (Crichton et al., 2018). In security applications, LP is used to identify hidden groups of terrorists and criminals (Hasan et al., n.d.)(Assouli et al., 2021; Berlusconi et al., 2016). The prediction methods employed in the link prediction studies have been detailed in the earlier reviews, for example, in the literature (Mohammad Al Hasan, Zaki, 2011)–(Martínez et al., 2016), and others. However, in this

* Corresponding author

review, we explore the problem of predicting co-authorships among researchers in research communities and academic social networks by presenting a new primary taxonomy of topological-based, content-based, and hybrid-based approaches systematically, besides their time-complexities. To compare our evaluation with theirs and learn from them, all the reviews and surveys that have been completed on LP are presented. Finally, the main objectives of this paper are as follows:

1. To present all the related reviews and surveys that have been accomplished on the LP problem.
2. To propose a taxonomy of link prediction-based approaches, namely the topological-based, content-based, and hybrid methods between the two.
3. To investigate and evaluate the literature based on the time-complexity and accuracy of the utilized approaches and proposed systems.

This paper is structured as follows: In Section 2, the related surveys and reviews regarding LP are listed. Our proposed survey approach is presented in Section 3. After that, the LP approaches based on the taxonomy demonstrated in Figure 2 are illustrated in Section 4. The evaluation measures used in the LP problem are described in section 5. Finally, the discussion section followed by the conclusion is introduced in sections 6 and 7, respectively.

1. RELATED WORKS

Several efforts on the topic have been performed to review the LP problem and the approaches proposed in the literature to solve this social graph analysis problem. The structure and functions of networked systems have been reviewed by Newman for the fields of the Internet, social networks, and biological networks (Newman, 2003). For example, the developments of the small-world effect, degree distributions, clustering, network correlations, random graph models, models of network growth and preferential attachment, and dynamical processes taking place on networks. However, the author has not considered the problem of LP in his revision which could be addressed in the fields that the author refers to.

Al Hassan et. al., 2011) examined various typical LP approaches for social networks by dividing them into three categories of models introduced in recent years: binary classification model, probabilistic model, and linear algebraic model. From the survey, it is concluded that the accuracy of the LP can be significantly improved if the time is considered in corporation with the social network data as well as graph topology.

Lu and Zhou (2011) surveyed classified LP approaches into three major categories, such as Similarity-based approaches, Maximum Likelihood Methods, and Probabilistic Models. Their work emphasized the random walk and maximum likelihood methods, and they quantified bonding strength using a connection-weight score for each node pair. The survey concluded that the studies of LP and complex networks will benefit each other since an in-depth understanding of network structure can be used to design advanced LP algorithms. Moreover, the performance of an LP algorithm could provide evidence about structural features as well as the algorithms themselves can be used to improve the estimates of real networks' properties.

Haghani and Keyvanpour conducted a review of the earliest scoring-based methodologies for the LP problem and extended them into the most recent methodologies, which are based on deep learning methods (Haghani & Keyvanpour, 2019). The review considered the dynamic behavior of a social network, which is primarily determined by two important features: node information and linkage information about the relationship between two nodes, as well as categorized the LP methods based on their technical approach, and discussed the strengths and weaknesses of various methods. This work ignored the text content of the authors when performing node-level analysis.

Wang et al. published extensive literature on LP techniques and discussed the current problems such as temporal LP, heterogeneous networks, and LP scalability (P. Wang et al., 2015). The authors claimed that predicting the missing or unobserved links in current social networks and newly added or deleted links in future social networks is very important for understanding the evolution of social networks. Furthermore, the literature concluded that the LP problem is necessary to mine and analyze social networks since it tackles the problems of incompletion and the dynamic behavior of the networks, but again ignoring the research interest of the collaborators.

In the short review of (Kushwah & Manjhvar, 2016), the authors discussed the current developments in LP algorithms and conducted a study of all existing LP systems. The study discussed the LP problem complexity, accessible solutions effective group communication management, and social link consciousness. Moreover, it summarizes recent growth in LP algorithms and surveys of all the prevailing LP techniques, but ignoring the hybrid approaches.

Martínez et al. (2016) focused on LP strategies in undirected networks using derived topological properties of the network. Because topology-based techniques are not domain-specific, it is argued that they were more versatile than attribute-based methods. The authors created a taxonomy to classify LP algorithms based on the methodology and the quantity of information used. Furthermore, the authors carried out an empirical analysis of the strategies, applying the most relevant methods to a variety of networks with varying attributes and assessing outcomes. The study observed that new links can be better predicted using only local or quasi-local information in most networks without including what papers actually include by the terms of topics.

A review was undertaken by Kong et al., (2019) to study the background, present state, and tendencies of academic social networks. The authors examined analytical approaches, including pertinent metrics, network features, and accessible academic analysis tools, and investigated models based on node types and timeliness. Likewise, they identified certain major mining methods for academic social networks and conducted an analysis of sample research tasks in this domain at three levels: actor, relationship, and network. The review concluded that the problems of mining useful and effective information, the complexity of the academic social networks, and sharing and lack of data need to be addressed in any research about academic social networks.

Hemkiran and Sudha (2000) presented an analysis of similarity metrics, approaches used in forecasting future linkages, and LP applications with an emphasis on dynamic networks in 2020. In addition, an examination of existing models for LP and their appropriateness for heterogeneous, massive, static, or dynamic networks was provided. The authors revealed that numerous studies have addressed static and homogeneous networks whereas very few have investigated dynamic and heterogeneous networks. It is revealed that the importance of time-dependent dynamic nature over static social networks in terms of increasing accuracy and efficiency needs to be addressed as well.

Daud et al., (2020) have conducted an LP survey, which included the most recent methodologies and applications for solving contemporary challenges such as large-scale networks, multi-dimensional networks, scalability, and network dynamicity. This research made recommendations and gave new insights for improving LP analysis in social networks as well. Furthermore, the authors claimed that the review paper was the first attempt to describe LP applications in various situations and analyses, with a particular focus on social networks and not investigating the SCN networks.

Mutlu et al., (2020), in their study provided comprehensive information on the issue of LP for big networks, which aided in

the discovery of the most related LP algorithms, and purposefully classified them into the suggested taxonomy. This paper examined similarity-based methods, such as local, global, and quasi-local approaches, probabilistic and relational methods as unsupervised solutions to the LP problem, and learning-based methods, such as matrix factorization, path, and walk-based LP models, and using neural networks for LP. However, the study has not examined the content-based measures such as the text in the research papers and the corporation between the topological and content-based features.

Wang and Le, (2020) presented an analysis of LP from static to dynamic networks, homogeneous to heterogeneous networks, and unsigned to signed networks. From the standpoints of informatics and physics, the hierarchical design concept was brought into the LP categorization system. The hierarchical model consisted of seven layers, namely the network, metadata, feature classification, selection input, processing, selection, and output layers, and seven aspects referred to as similarity-based, probabilistic, likelihood, unsupervised learning, semi-supervised learning, supervised learning, and reinforcement learning methods. These methods included many classic and up-to-date LP techniques. However, the review has not addressed the aspect of the content-based approaches and the hybrid methods between similarity-based and content-based.

Lastly, a survey on all the methods of topic modelling was conducted (Alghamdi & Alfalqi, 2015). The survey provided two categories of topic modelling. The first category consists of methods of topic modelling, such as Latent semantic analysis (LSA)(Abdul & Mastan, 2013), Probabilistic latent semantic analysis (PLSA)(Abdul & Mastan, 2013), Latent Dirichlet Allocation (LDA)(Abdul & Mastan, 2013)(PARIMI, 2010), and Correlated Topic Model (CTM)(Abdul & Mastan, 2013)(Blei & Lafferty, 2007). The second category is called topic evolution models, which model topics by considering an important factor time. It included the methods of the Topic Over Time (TOT), dynamic topic models (DTM), multiscale topic tomography, dynamic topic correlation detection, and detecting topic evolution in scientific literature. In our review, we concentrated on the LDA method which has been used by most researchers to predict links in social networks. We analyzed the literature that used the content-based features in combination with other similarity-based approaches to provide the most suitable LP formula in the future.

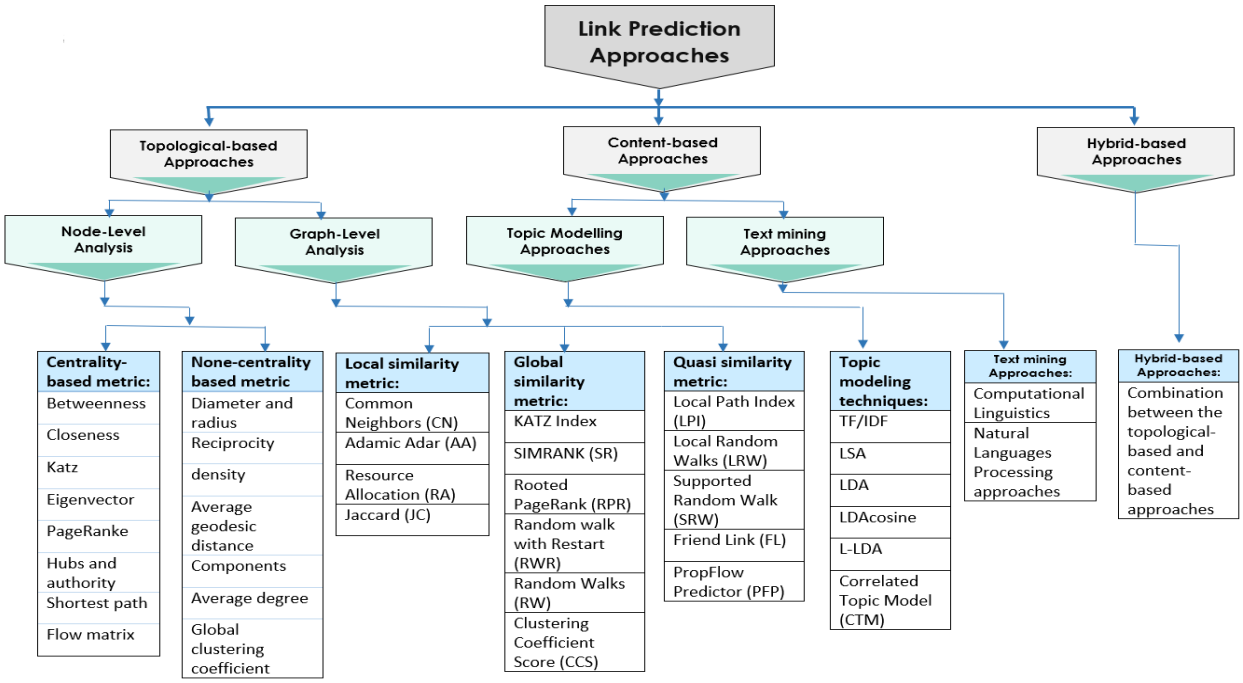

Figure 2. The Diagram of the Taxonomy LP Approaches

2. THE PROPOSED SURVEY APPROACHES

In light of the above literature, several reviews and surveys regarding LP have been presented, however very few have attempted to explore both the content-based LP, topological-based as well as hybrid measures (Hasan et al., n.d.),(Raut et al., 2020). In addition to that, the topological metrics have helped in solving the link prediction problem to a good degree due to the ability to explicitly model the LP problem in the term of social components and interactions. On the other hand, beside the topological features the academic collaboration networks include other characteristics that must be taken into account, such a content of a paper that two authors share (title, abstract, keywords etc.), which represents the knowledge fusion among the members of a given scientific collaboration network. Motivated by this fact and since the social networks also have to be concerned with the analysis of the embedded information, in this review paper, academic social context is categorized into three classes: structured-based information, which represents the network's topological structure, content-based information representing the features associated with entities and their relationships and the hybrid information of both, as it is clear in Figure 2. The analysis of networks using the three types of information can aid in discovering and quantifying the interesting facts on both the individual and group levels of academic collaboration networks. Additionally, this type of information will support the hypothesis that the text and graph mining techniques can be used to increase overall prediction accuracy for collaboration networks in order to formulate the best LP approach for forecasting future cooperation patterns and trends in academic networks.

3. REVIEWED LINK PREDICTION APPROACHES

In any collaboration network, LP techniques and approaches can be examined based on the type of information used to forecast a connection. There are three types of information approaches topological approaches, content-based approaches, and hybrid approaches that combine the two. Topology-based metrics, in particular, are widely used to decipher LP problems due to their ease of use and applicability to simpler networks with fewer node and edge attributes. Content-based features (non-topological features) have the advantage of improving the LP problem's performance. They are, however, not always accessible and perhaps difficult to attain. Finally, the third type of information used in LP is hybrid measures of the previous two measures

(topological and non-topological), which require aggregate features derived from existing score functions.

3.1 Topological-Based Similar Measures

Because of their simplicity, similarity approaches appear to be more promising than other approaches in terms of LP. A node pair with a higher similarity score is more likely to form a link in the future. Earlier studies on similarity approaches focused on using widely used graph-based measures due to their simplicity of use and applicability to simpler networks without many node and edge attributes. The topological-based approaches can be classified into two levels of analysis: node level analysis and graph (group) level analysis. The node-level analysis consists of centrality-based metrics and non-centrality-based metrics, while the group-level analysis contains the local indices, global indices, and quasi-local indices (Hassan, n.d.).

3.1.1 Node-Level Statistical Analysis: These metrics are valuable because they reveal the functions of the network's nodes. The goal of researching how individuals' function and interact in networks is typically to comprehend the behavior of the social systems that created those networks. The measures introduced here can be divided into two types (Tabassum et al., 2018): Centrality-Based Measures: these investigate generic measures of centrality as a means of understanding how a vertex's position fits into the overall structure of the graph and, as a result, aids in identifying the network's major participants. The most often used were degree, closeness, local clustering coefficients, and eigenvector centrality. Non-Centrality Measures: The non-similarity-based measurements offer more concise data and enable assessment of the network's general structure, providing insights into crucial features of the underlying social phenomena. Diameter and radius, Reciprocity, Density, Average geodesic distance, Component, Average degree, and Global clustering coefficient are the widely used examples of this type of measure.

3.1.2 Graph-Level Analysis: There are three types of graph-based approaches: local, global, and quasi similarity algorithms (Hassan, n.d.)(Yuliansyah et al., 2020). Local indices are one of the simplest methods used for calculating the similarity score in link prediction, which take the number of neighbor nodes and the degree of neighbor into account. In cases where the path distance is under two, a node is frequently regarded as a neighbor node. Examples of local indices are illustrated in Table 1 with their formulas and the main purposes that were used. Common Neighbors, Salton index, Jaccard index, Sorensen index, hub promoted index, Leicht-Holme-Newman index, Preferential Attachment index, Adamic Adar index, and Resource Allocation index are the most commonly used measures by researchers. Furthermore, local similarity indexes are frequently employed in practical applications, since they reduce resource consumption and computational complexity while preserving the greatest prediction performance.

In contrast, global indices use the graph's global link structure to calculate similarity scores when nodes have a path distance of more than two. In other words, global indices score each link by considering the entire network of topological data. Global indices, as opposed to local index methods, identify all interesting direct and indirect paths that should be factored into the similarity score. Table 2 depicted the existing and most widely used global measures such as the Katz index, the Leicht-Holme-Newman2 index, and the Matrix Forest index with their formulas and purposes. Due to the high dimensionality of networks, global similarity indices in the context of link prediction in large networks are time-consuming and computationally intensive.

Like global indices, quasi-local methods make use of additional topological data, and similar to local index techniques, they compute the score using nodes with a maximum path distance of two. A trade-off between a model's complexity and prediction accuracy is offered by quasi-local indices, which include the local path index, local random walk, and superposed random walk. Table 3 illustrated the examples of these measures where they are more accurate in their predictions than local techniques because they take into account more topological data with less computational complexity. Furthermore, the dataset and the application have a significant impact on the performance of quasi-local techniques. Therefore, by creating an algorithm that can compute the similarity index for the entire network with more precision and stability, the quasi-local technique can be further enhanced.

3.2 Content-Based Similar Measures

The performance of the LP issue can be improved by using content-based (non-topological) characteristics (Y. Zhang et al., 2012). They are, however, not permanently available and perhaps hard to obtain. More importantly, the majority of content-based features are domain-specific, where recognizing and finding them necessitates well domain knowledge. As a result, while a general LP learning model typically considers generic features such as node, network, and topological features, non-topological features should similarly be considered, specifically for a practical LP application (P. Wang et al., 2015). The content-based approaches can be divided into topic modelling approaches and text mining approaches.

3.2.1 Topic Modelling Approaches: There are different topic modelling approaches used in literature, Term Frequency/Inverse Document Frequency (TF/IDF) (Bartal et al., 2009) (Quercia et al., 2012), Latent Semantic Analysis (LSA) (Abdul & Mastan, 2013), LDA(Abdul & Mastan, 2013)(PARIMI, 2010), LDAcosine (Chuan et al., 2018), Labelled-LDA (L-LDA) (Quercia et al., 2012), and Correlated topic model (CTM) (Abdul & Mastan, 2013). However, this review will concentrate on the LDA approach which is widely used by researchers as it is proved to be the simplest topic model that improved the performance of the LP problem (Chuan et al., 2018), Table 4. The majority of content-based prediction methods follow a machine learning approach; that is, they use classification-based methods to make predictions, such as Decision Tree, Support Vector Machine (SVM), and Naïve Bayes.

Therefore, and according to the content similarity of the papers, the researchers in (Chuan et al., 2018) proposed a new weighted metric called LDAcosin with mathematical notation for LP in the co-authorship network, and the new metric is being experimentally validated on in the public bibliographic collection. Furthermore, the time when the relationship was observed is saved for an edge, as are the publication year, title, and abstract of the article, all of which are used as features for the new metric. Authors with multiple high similarity papers but no common papers have shown a gain in link degree using the new metric, when using the binary classification method (weighted SVM) for the LP.

Hassan (n.d.) proposed a new method for supervised LP in co-authorship networks using predictors. The predictors are extracted by computing the similarity between the research interests (the keywords that generally express their research interests) of each two author node in the network, the similarity between their affiliations, the sum of their research performance indices, and the similarity between the two author nodes themselves. The author proposed a method that utilized the author's research performance indices to predict potential future links between them by introducing new predictor variables for solving the LP problem using supervised learning in academic SN. To represent the set of predictor variables used for training the supervised learning algorithm for LP, the new predictors were combined with those computed from computing the similarity between two author nodes.

Muniz et al.(2018) in 2018 proposed a combination of global similarity indices and content-based measures in their work to

improve the performance of LP. The authors showed that the performance of the LP method could be improved by combining contextual and temporal information with topological data in weight computation using weighted similarity indices in unsupervised LP with temporal, contextual, and topological attributes. A supervised topic model for assigning "topics" to a collection of documents (e.g., Twitter profiles) has been proposed (Quercia et al., 2012). The model is known as Labelled-LDA, and it has proven to be effective at the task of Twitter profile classification, outperforming the competitive SVM. L-LDA could accurately classify a profile with topics for which it has only observed small amounts of training data and outperforms SVMs in determining how similar a pair of profiles is. It is also demonstrated that L-inference LDA's techniques are preferable to each SVM's linear classification when dealing with rich, mixed-topic documents such as Twitter profiles. In (Parimi & Caragea, 2011), the authors utilized machine learning algorithms to predict friendship links based on the content (data from user profiles) and graph structure of social network sites such as Live Journal. The research employed a topic modelling approach, specifically the Latent Dirichlet Allocation (LDA), which provides a simple and efficient way of capturing the semantics of user interests by grouping them into categories, also known as topics, and thus reducing the problem's dimensionality. In (Zhao et al., 2017), a co-authorship network dataset known as NIPS234 was utilized to test a fully Bayesian approach that models large, sparse, and unweighted relational networks with arbitrary node attribute encoded in binary form, with 234 authors and 598 links extracted from NIPS 1-17 conferences. The proposed model outperforms other models in terms of LP, especially when training data is scarce. All of the dataset's papers written by the same author were merged into a single document, and an LDA model with 100 topics was trained. The attributes were chosen from the top five most frequently occurring topics, yielding a 234*100 attribute matrix with 1170 non-zero entries.
Sachan and Ichise (2010) proposed a semantic measure based on the work of researchers in the Digital Bibliography Library Project (DBLP) network. A semantic approach called Abstract-Keywords Match Count (AKMC) to improve the accuracy of the proposed link predictor by utilizing abstract information, research titles, and event information is utilized.

3.2.1 Text Mining Approaches: Information Retrieval (IR), Natural Language Processing (NLP), Vector Space Model (VSM), Information Extraction from the text (IE), Text Summarization, Unsupervised Learning Methods, Supervised Learning Methods, Probabilistic Methods for Text Mining, Text Streams and Social Media Mining, Opinion Mining and Sentiment Analysis, and Biomedical Text Mining are examples of the text mining approaches utilized in different fields including LP in social networks (Allahyari et al., 2017). However and to the best of literature knowledge, there is no work conducted to predict links in co-authorship networks using one of the text mining approaches except the work of (Bartal et al., 2009), which combined the text mining methods (i.e., NLP and VSM) with many topological measures of SNA.

3.3 Hybrid Methods

Table 1 describes the hybrid approaches in the literature for solving the link prediction problem. Prediction of the link between nodes is influenced by several score functions and they differ from one another in their importance. It is possible to aggregate features derived from existing score functions in a hybrid way. This is why features were viewed from a hybrid perspective. As a result, multiple score functions were integrated to utilize different measurements to categorize two classes for a pair of nodes (Mishra & Nandi, 2015).
Muniz et al. (2018) in 2018 proposed a combination of global similarity indices and content-based measures in their work to improve the performance of LP. The authors could improve the performance of the LP method by combining contextual and temporal information with topological data in weight computation using weighted similarity indices in unsupervised LP with temporal, contextual, and topological attributes.
Chuan et al (2018) presented a supervised LP approach for predicting links between authors in a co-authorship network based on measuring the similarity of two paper contents written by a pair of authors in that network. The researchers proposed LDAcosin, a new content similarity measure that employs the Latent Dirichlet Allocation (LDA) method to determine the content similarity between two papers. The greater the similarity, the greater the possibility of a future link between the two authors of the two papers. This new content similarity metric is utilized as a new metric feature in the co-authorship network to perform LP.
Ibrahim and Chen (I2015) proposed a method for dynamic network LP that incorporates temporal information, community structure, typological information, and node centrality. The authors use eigenvector centrality to predict a node's future importance and then to predict links. In other words, their approach was based on a reduced static graph using a modified reduced adjacency matrix to reflect the frequency of each link, while the other approach integrates similarity indices of the nodes to exploit both temporal and topological information such as community structure and centrality of the nodes. The approach employs a damping factor to integrate the similarity indices in the time steps to emphasize the importance of more recent topological information. Furthermore, they take into account existing links by using an

Table 1. Local Similarity Measures

Measure	References	Purpose	Formula	Time complexity	Normalized Similarity Score	References (Accuracy, AUC, AUROC)
Common Neighbors (CN)	(Raut et al., 2020), (Ghorbanzadeh et al., 2021a), (Zhu et al., 2012), (Aghabozorgi & Khayyambashi, 2018), (Muniz et al., 2018), (Assouli et al., 2021), (Q. Zhang et al., 2020), (J. H. Liu et al., 2016), (Chuan et al., 2018), (Kumar et al., 2020), (Berlusconi et al., 2016), (Ibrahim & Chen, 2015), (H. Liu et al., 2019), (Jaya Lakshmi & Durga Bhavani, 2017), (Gao et al., 2015), (R. Lichtenwalter & Chawla, 2012), (R. N. Lichtenwalter et al., 2010), (Bartal et al., 2009), (L. Dong et al., 2013)	Obtaining the number of common neighbors	$S(a,b) = \lvert\Gamma_a \cap \Gamma_b\rvert$	O(VK^3), V is the number of nodes, K is the largest node degree on the graph	No	(Raut et al., 2020) accuracy = 0.8890 (Ghorbanzadeh et al., 2021a) N/A (Zhu et al., 2012) N/A (Aghabozorgi & Khayyambashi, 2018) accuracy = 0.9033 (Muniz et al., 2018) N/A (Assouli et al., 2021) AUC= 0.7417 (Berlusconi et al., 2016) N/A (Martínez et al., 2017) AUC=0.9523 (Aghabozorgi & Khayyambashi, 2018) AUC= 0.9661 (Q. Zhang et al., 2020) Precision =0.5559 (J. H. Liu et al., 2016) AUC= 0.9354 (Chuan et al., 2018) AUC=0.6626 (Kumar et al., 2020) AUROC= 0.9918 (Ibrahim & Chen, 2015) N/A (H. Liu et al., 2019) AUC= 0.9993 (Jaya Lakshmi & Durga Bhavani, 2017) AUROC = 0.6811 (Gao et al., 2015) AUC= 0.8900 (R. Lichtenwalter & Chawla, 2012) N/A (R. N. Lichtenwalter et al., 2010) N/A (Bartal et al., 2009) accuracy = 0.9773 (L. Dong et al., 2013) N/A
Adamic Adar (AA)	(Raut et al., 2020), (Aghabozorgi & Khayyambashi, 2018), (Muniz et al., 2018), (Q. Zhang et al., 2020), (J. H. Liu et al., 2016), (Chuan et al., 2018), (Kumar et al., 2020), (Ibrahim & Chen, 2015), (Jaya Lakshmi & Durga Bhavani, 2017), (Gao et al., 2015), (R. Lichtenwalter & Chawla, 2012), (R. N. Lichtenwalter et al., 2010), (Bartal et al., 2009), (L. Dong et al., 2013)	Common neighbors with fewer neighbors	$s(a,b) = \sum_{\lvert f \in \Gamma_a \cap \Gamma_b\rvert} \frac{1}{log\lvert\Gamma_f\rvert}$	O(VK^3)	No	(Raut et al., 2020) accuracy = 0.8890 (Aghabozorgi & Khayyambashi, 2018) accuracy =0.8964 (Muniz et al., 2018)N/A (Ibrahim & Chen, 2015) N/A (Jaya Lakshmi & Durga Bhavani, 2017) AUROC = 0.7011 (Gao et al., 2015) AUC= 0.8922 (R. Lichtenwalter & Chawla, 2012) N/A (R. N. Lichtenwalter et al., 2010) N/A (Bartal et al., 2009) accuracy = 0.9773 (L. Dong et al., 2013) N/A (Martínez et al., 2017) AUC=0.9553 (Aghabozorgi & Khayyambashi, 2018) AUC=0.9631 (J. H. Liu et al., 2016) AUC= 0.9432 (Chuan et al., 2018) AUC=0.6626 (Kumar et al., 2020) AUROC= 0.9338 (Q. Zhang et al., 2020) Precision = 0.5557

Resource Allocation (RA)	(Raut et al., 2020), (K. Zhou et al., 2019), (Q. Zhang et al., 2020), (J. H. Liu et al., 2016), (Berlusconi et al., 2016), (Samad et al., 2019) , (L. Dong et al., 2013)	Similar to AA	$RA(a,b) = \sum_{\lvert f \in \Gamma_a \cap \Gamma_b \rvert} \frac{1}{\lvert \Gamma_f \rvert}$	O(VK^3)	No	(Raut et al., 2020) accuracy = 0.8890 (K. Zhou et al., 2019) N/A (Berlusconi et al., 2016) N/A (Samad et al., 2019) accuracy = 0.9200 (L. Dong et al., 2013) N/A (Martínez et al., 2017) AUC=0.9561 (J. H. Liu et al., 2016) AUC= 0.8976 (Q. Zhang et al., 2020) Precision = 0.5576
Resource Allocation Based on Common Neighbor Interactions (RA-CNI)	(Zhang, Jianpei and Zhang, Yuan and Yang, Hailu and Yang, 2014)	A hybrid of RA and CN	$S(a,b) = \sum_{f \in \Gamma_{ba} \cap \Gamma_b} \frac{1}{\Gamma_f} + \sum_{e_{i,j} \in E, \lvert \Gamma_i \rvert < \lvert \Gamma_j \rvert, i \epsilon \Gamma_a, j \epsilon \Gamma_b} \left(\frac{1}{\lvert \Gamma_a \rvert} - \frac{1}{\lvert \Gamma_b \rvert} \right)$	O(VK^4)	No	(Zhang, Jianpei and Zhang, Yuan and Yang, Hailu and Yang, 2014) N/A (Martínez et al., 2017) AUC= 0.9564
Jaccard (JC)	(Raut et al., 2020), (Aghabozorgi & Khayyambashi, 2018), (Assouli et al., 2021), (Q. Zhang et al., 2020), (J. H. Liu et al., 2016), (Chuan et al., 2018), (Kumar et al., 2020), (Ibrahim & Chen, 2015), (H. Liu et al., 2019), (Jaya Lakshmi & Durga Bhavani, 2017), (Gao et al., 2015), (R. N. Lichtenwalter et al., 2010), (Samad et al., 2019)	Ratio of the intersection to the union of the common neighbors	$JC(a,b) = \frac{\lvert \Gamma_a \cap \Gamma_b \rvert}{\lvert \Gamma_a \cup \Gamma_b \rvert}$	O(VK^3)	Yes	(Raut et al., 2020) accuracy = 0.8890 (Aghabozorgi & Khayyambashi, 2018) accuracy = 0.8509 (Assouli et al., 2021) AUC= 0.8767 (Ibrahim & Chen, 2015) AUC= N/A (H. Liu et al., 2019) 0.9993 (Jaya Lakshmi & Durga Bhavani, 2017) AUROC = 0.6012 (Gao et al., 2015)AUC= 0.8702 (R. N. Lichtenwalter et al., 2010) N/A (Samad et al., 2019) accuracy = 0.8200 (Martínez et al., 2017) AUC= 0.9492 (Raut et al., 2020) AUC= 0.8861 (J. H. Liu et al., 2016) AUC= 0.8853 (Chuan et al., 2018) AUC=0.6626 (Kumar et al., 2020) AUROC= 0.9895 (Q. Zhang et al., 2020) Precision = 0.5107
Sørensen (SO)	(K. Zhou et al., 2019), (Gao et al., 2015), (Samad et al., 2019)	Nodes with lower degrees are more likely to establish links	$S(a,b) = \frac{2 . \lvert \Gamma_a \cap \Gamma_b) \rvert}{\lvert \Gamma_a \rvert + \lvert \Gamma_b \rvert}$	O(VK^3)	Yes	(K. Zhou et al., 2019) N/A (Gao et al., 2015) AUC= 0.8650 (Samad et al., 2019) accuracy = 0.8100 (Martínez et al., 2017) AUC=0.9492
Preferential Attachment (PA)	(Raut et al., 2020), (Zhu et al., 2012), (Aghabozorgi & Khayyambashi, 2018), (J. H. Liu et al., 2016), (Kumar et al., 2020), (Ibrahim & Chen, 2015), (Gao et al., 2015), (R. Lichtenwalter & Chawla, 2012), (R. N. Lichtenwalter et al., 2010), (Bartal et al.,	establish links between nodes with high degrees.	$PA(a,b) = \lvert \Gamma_a \rvert . \lvert \Gamma_b \rvert$	O(VK^2)	No	(Raut et al., 2020) accuracy = 0.8890 (Zhu et al., 2012) N/A (Aghabozorgi & Khayyambashi, 2018) accuracy = 0.8706 (Ibrahim & Chen, 2015)N/A (Gao et al., 2015) AUC= 0.9109 (R. Lichtenwalter & Chawla, 2012) N/A (R. N. Lichtenwalter et al., 2010) N/A (Bartal et al., 2009) accuracy = 0.9773 (L. Dong et al., 2013) N/A

	2009), (L. Dong et al., 2013)					(Martínez et al., 2017) AUC= 0.9386 (Aghabozorgi & Khayyambashi, 2018) AUC= 0.9480 (J. H. Liu et al., 2016)AUC= 0.8945 (Kumar et al., 2020) AUROC= 0.9342
Hub Promoted (HPI)	(Ghorbanzadeh et al., 2021a), (Q. Zhang et al., 2020), (Gao et al., 2015)	Link probability is specified by lower node degrees.	$HP(a,b) = \frac{\lvert\Gamma_a \cap \Gamma_b\rvert}{min\{\lvert\Gamma_a\rvert, \lvert\Gamma_b\rvert\}}$	O(VK^3)	Yes	(Ghorbanzadeh et al., 2021a) N/A (Gao et al., 2015) AUC= 0.8440 (Martínez et al., 2017) AUC= 0.9484 (Q. Zhang et al., 2020) Precision = 0.2334
Hub Depressed (HDI)	(Ghorbanzadeh et al., 2021a), (Q. Zhang et al., 2020), (Gao et al., 2015)	Link probability is specified by higher node degrees	$HD(a,b) = \frac{\lvert\Gamma_a \cap \Gamma_b\rvert}{\max\{\lvert\Gamma_a\rvert, \lvert\Gamma_b\rvert\}}$	O(VK^3)	Yes	(Ghorbanzadeh et al., 2021a) N/A (Gao et al., 2015) AUC= 0.8511 (Martínez et al., 2017) AUC= 0.9483 (Q. Zhang et al., 2020) Precision = 0.4627
Salton Index (SA)	(J. H. Liu et al., 2016)	For calculation of the cosine similarity between two nodes	$Salton\ (a,\ b) = \frac{\lvert\Gamma_a \cap \Gamma_b\rvert}{\lvert\Gamma_a\rvert.\lvert\Gamma_b\rvert}$	O(VK^3)	Yes	(Martínez et al., 2017) AUC=0.9501 (J. H. Liu et al., 2016) AUC=0.9049
Salton Cosine Similarity And Sorensen Cosine index	(Samad et al., 2019)	For calculation of the cosine similarity between two nodes	$Salton\ (a,\ b) = \frac{\lvert\Gamma_a \cap \Gamma_b\rvert}{\sqrt{\lvert\Gamma_a\rvert.\lvert\Gamma_b\rvert}}$	O(VK^3)	Yes	(Samad et al., 2019) =0.9200
Cosine similarity	(Gao et al., 2015)	used to compare documents in text mining	$cosine\ sim(ab) = \frac{\lvert\Gamma_a\rvert.\lvert\Gamma_b\rvert}{\lVert\Gamma_a\rVert * \lVert\Gamma_b\rVert}$	O(VK^3)	Yes	(Gao et al., 2015) AUC= 0.8812 (SA, Cosine similarity): (Martínez et al., 2017) AUC= 0.9501
Leicht-Holme-Newman (LLHN)	(Gao et al., 2015)	Establish links between couples of nodes with many common neighbors.	$S(ab) = \frac{\lvert\Gamma_a \cap \Gamma_b\rvert}{\lvert\Gamma_a\rvert.\lvert\Gamma_b\rvert}$	O(VK^3)	Yes	(Gao et al., 2015) AUC= 0.7881 (Martínez et al., 2017) AUC=0.9411
Mutual Information (MI)	(Li et al., 2021), (X. Liu et al., 2013), (Ghorbanzadeh et al., 2021a)	Use of probability rules and common neighbors for obtaining the similarity between couples of nodes	$S(a,b) = -I(e_{x,y}\lvert\Gamma_a \cap \Gamma_b\rvert) = -I(e_{x,y}) + \sum_{z\in\lvert\Gamma_a\cap\Gamma_b\rvert} I(e_{x,y};z)$	*O(VK^6)*	Yes	(Li et al., 2021) AUC = 09891 (X. Liu et al., 2013) N/A (Ghorbanzadeh et al., 2021a) N/A (Martínez et al., 2017) AUC=0.9379
CAR-Based Indices (CAR) Common Neighbour version	(Kumar et al., 2020)	Use of a notion known as a local communit y (LC) and redefinitio	$S(a,b) = \sum_{f\in\Gamma_a\cap\Gamma_b} 1 + \frac{\lvert\Gamma_a \cap \Gamma_b \cap \Gamma_f\rvert}{2}$	*O(VK^4)*	Yes	CAR-CN: (Martínez et al., 2017) AUC= 0.9541 CAR-AA: (Martínez et al., 2017) AUC= 0.9557 CAR-RA: (Martínez et al., 2017) AUC= 0.9561

		n of the AA, RA, and CN measures on that basis				(Kumar et al., 2020) AUROC= 0.9447
Functional Similarity Weight (FSW)	(Samad et al., 2020), (Ghorbanzadeh et al., 2021a)	Derived from the Sørensen measure	$S(a,b) = \left(\frac{2\lvert\Gamma_a \cap \Gamma_b\rvert}{\lvert\Gamma_a\rvert - \lvert\Gamma_b\rvert + 2\lvert\Gamma_a \cap \Gamma_b\rvert +$ $\gamma = \max(0, <\Gamma$ $> -(\lvert\Gamma_a - \Gamma_b\rvert$ $+ \lvert\Gamma_a \cap \Gamma_b\rvert))$	O(VK^3)	Yes	(Samad et al., 2020) N/A (Ghorbanzadeh et al., 2021a) N/A (Martínez et al., 2017) AUC= 0.9473
Individual Attraction Index (IA)	(Samad et al., 2020), (Ghorbanzadeh et al., 2021a), (Y. Dong et al., 2011), (Mutlu & Oghaz, 2020)	Establish a link between two nodes with common neighbors that firm connection between their common neighbors.	$S(a,b) = \sum_{f \in \Gamma_a \cap \Gamma_b} \frac{\lvert e_{\Gamma_a \cap \Gamma_b}\rvert + 2}{\lvert\Gamma_f\rvert\lvert\Gamma_a \cap \Gamma_b\rvert}$	O(VK^3)	Yes	(Samad et al., 2020) N/A (Ghorbanzadeh et al., 2021a) N/A (Y. Dong et al., 2011) accuracy = 0.9675 (Mutlu & Oghaz, 2020) N/A IA1: (Martínez et al., 2017) AUC= 0.9562 IA2: (Martínez et al., 2017) AUC= 0.9562
Local Na¨ıve Bayes (LNB)	(Li et al., 2021), (Ghorbanzadeh et al., 2021a)	Establish a link, so the probability of the link can be estimated by the theories of likelihood.	$S(a,b) = \sum_{m \in \Gamma_a \cap \Gamma_b} f(m)\log(oW_m)$ $where\ o = \frac{punconnected}{pconnected}$ $= \frac{\frac{1}{2}\lvert v\rvert(\lvert v\rvert) - 1}{\lvert E\rvert} - 1$ $W_m = \frac{2\lvert\{e_{a,b}: a,b \in \Gamma_m, e_{a,b} \in$ $2\lvert\{e_{a,b}: a,b \in \Gamma_m, e_{a,b} \notin$	*O*(v *O*(*f*(*z*))+ V*K*^{3})	Yes	(Li et al., 2021) AUC = 0.9930 (Ghorbanzadeh et al., 2021a) N/A LNB-CN: (Martínez et al., 2017) AUC= 0.9541 LNB-AA: (Martínez et al., 2017) AUC= 0.9557 LNB-RA: (Martínez et al., 2017) AUC= 0.9561
Negated Shortest Path (NSP)	(Kumar et al., 2020), (Liben-Nowell & Kleinberg, 2007)	compute the shortest path between a pair of nodes	$S(a,b) = -\lvert shortest\ path_{a,b}\rvert$	O (ev log v)	----	(Liben-Nowell & Kleinberg, 2007) N/A (Martínez et al., 2017) AUC=0.9423 (Kumar et al., 2020) AUC= N/A
Extended Resource Allocation	(Linyuan & Zhou, 2011)	adds longer paths to the RA index	$ERA(a,b) = \sum_{x=2}^{3} \sum_{s \in s^x(a,b)} \prod_{n(p)}$	between O (*N (k)*^{2}) (RA) and *O*(*N*(*k*)^3) (LP)	No	RA: (Martínez et al., 2017) AUC= 0.9561 RA: (Linyuan & Zhou, 2011) AUC= 0.9550

Table 2. Global Similarity Measures

Measure	References	Purpose	Formula	Time complexity	normalized similarity score	References (accuracy, AUC, AUROC)
KATZ Index	(Zhu et al., 2012), (K. Zhou et al., 2019), (Kumar et al., 2020), (Ibrahim & Chen, 2015), (Gao et al., 2015), (R. N. Lichtenwalter et al., 2010), (L. Dong et al., 2013)	sums all the paths between two nodes and decreases the contribution of paths with high lengths using a damping factor	$Score(a,b) = \sum_{m=1}^{\infty} \beta^m * \lvert Paths_{a,b}^m \rvert$	$O(V^3)$	-	(Zhu et al., 2012) N/A (Ibrahim & Chen, 2015) N/A (K. Zhou et al., 2019) N/A (Linyuan & Zhou, 2011)AUC = 0.9880 (Kumar et al., 2020) AUROC = 0.9466 (Martínez et al., 2017) AUC = 0.9630 (Gao et al., 2015) AUC = 0.9312 (R. N. Lichtenwalter et al., 2010)N/A (L. Dong et al., 2013) N/A
SIMRANK(SR)	(Zhu et al., 2012), (Samad et al., 2020), (Linyuan & Zhou, 2011)	general similarity measure that considers two nodes are similar if their neighbors are similar	$s(x,y) = \beta \frac{\sum_{i \in \Gamma_x} \sum_{j \in \Gamma_y} s(i,j)}{\lvert\Gamma_x\rvert \lvert\Gamma_y\rvert}$	$O(V^2K^{2l+2})$	-	(Zhu et al., 2012) N/A (Samad et al., 2020) N/A (Martínez et al., 2017)AUC= 0.9414
LEICHT-HOLME NEWMAN2 Index (LHN2)	(Linyuan & Zhou, 2011)	Similar to Katz Score	$S(a,b) = D^{-1}\left(I - \frac{\emptyset A}{\gamma_1}\right)^{-1} D^{-1}$	$O(cv^2d)$	-	(Linyuan & Zhou, 2011) AUC = 0.986 (Martínez et al., 2017) LLHN: AUC = 0.9411
Rooted PageRank (RPR)	(Mutlu & Oghaz, 2020)	type of PageRank centrality, which is used to rank the search results	$RPR = (1-\alpha)(I - \alpha N\hat{}) - 1$ $N\hat{} = D - 1A$ $D[i,i] = \Sigma j\, A[i,j]$	$O(vlk^l)$	-	(Mutlu & Oghaz, 2020) N/A (Linyuan & Zhou, 2011) AUC = 0.9930
Random walk with Restart (RWR)	(J. H. Liu et al., 2016), (Coskun & Koyuturk, 2016), (L. Dong et al., 2013)	Random walk where picks a node and move a random walk with probability α or we return to the starting node with probability (1−α).	$\min_{P^i} \alpha \sum_{a,b \in V} M_{a,b}^T \left(\overrightarrow{p_a^i} - \overrightarrow{p_b^i}\right)^2 + (1 - \alpha) \sum_{a \in V} \left(\overrightarrow{p_a^i} - \overrightarrow{s_a^i}\right)^2$	$O(cv^2d)$	-	(J. H. Liu et al., 2016) N/A (Coskun & Koyuturk, 2016) N/A (L. Dong et al., 2013) N/A (Linyuan & Zhou, 2011) AUC = 0.9800 (Martínez et al., 2017) AUC = 0.9681
PropFlow Predictor (PFP)	(R. Lichtenwalter & Chawla, 2012), (R. N.	same as Rooted PageRank, however, it	$S_{a,j} = S_{a,j} + \frac{S_{a,i}}{\lvert\Gamma_i\rvert}$	$O(vlk^l)$	-	(R. Lichtenwalter & Chawla, 2012) N/A

	Lichtenwalter et al., 2010)	is more localized				(R. N. Lichtenwalter et al., 2010) N/A (Martínez et al., 2017) AUC = 0.9636
Pseudoinverse of the Laplacian Matrix (PLM)	(Mutlu & Oghaz, 2020)	widely used in spectral graph theory	$s(a,b) = \frac{L^+_{a,b}}{\sqrt{L^+_{a,a}\, L^+_{a,b}}}$	$O(V^3)$	-	(Mutlu & Oghaz, 2020) N/A (Martínez et al., 2017) AUC = 0.9439
Blondel Index (BI)	(Mutlu & Oghaz, 2020)	Initially proposed to measure similarity for a pair of vertices in different graphs	$S(k) = \frac{AS(k-1)A^T + A^TS(k-1)A}{\|AS(k-1)A^T + A^TS(k-1)A\|_F}$	$O(cv^2k)$	-	(Mutlu & Oghaz, 2020) N/A (Martínez et al., 2017) AUC = 0.9081
Random Forest Kernel Index (RFK)	(Yuliansyah et al., 2020)	In graph theory, a spanning tree of a graph G is defined as a connected undirected subgraph with no cycles that includes all the vertices and some or all the edges of G	$S = (I + L)^{-1}$	$O(V^3)$	-	(Yuliansyah et al., 2020) N/A (Martínez et al., 2017) AUC = 0.9593
Maximal Entropy Random Walk (MERW)	(Mutlu & Oghaz, 2020)	Nodes tend to be linked to central nodes in structured networks	$\mu = \lim_{\iota \leftarrow \infty} \frac{-\sum_{path^{\iota}_{a,b}} p(path^{\iota}_{a,b}) \ln p(path^{\iota}_{a,b})}{\iota}$	$O(cv^2k)$	-	(Mutlu & Oghaz, 2020) N/A (Martínez et al., 2017) AUC= 0.9575
Matrix Forest Index (MFI)	(S. Liu et al., 2017)	similarity score defined as ratio of the number of spanning rooted forests	$S(v_i, v_j) = (I + L)^{-1}$		-	(S. Liu et al., 2017) AUC = 0.9790
Random Walks (RW)	(K. Zhou et al., 2019), (J. H. Liu et al., 2016), (Coskun & Koyuturk, 2016)	Randomly select a neighbor of the node and move to it; then, repeat this process for each reached node.	$\overrightarrow{P^a}(t) = M^t\overrightarrow{P^a}(t-1)$	$O(cv^2d)$	-	(K. Zhou et al., 2019) N/A (J. H. Liu et al., 2016) AUC = 0.9409 (Coskun & Koyuturk, 2016) N/A
Clustering Coefficient Score (CCS)	(W. Zhang & Wu, 2014), (Shao et al., 2013), (De Tre et al., 2014)	topological measures that summarize the global structure of a graph	$C(m) = \frac{(number\ of\ cycles\ of\ length\ m)}{(number\ of\ paths\ of\ length\ m)}$	$O(v^3)$	-	(W. Zhang & Wu, 2014) N/A (Shao et al., 2013) accuracy = 0.92611 (De Tre et al., 2014) N/A

Vertex Collocation Profile (VCP)	(P. Wang et al., 2015), (R. N. Lichtenwalter & Chawla, 2012)	is a vector describing the relationship between two vertices, in terms of their common membership	$VCP^{n,r} = 2^{\frac{n(n-1)r}{2}-1}$	O(|E|/|V|)	-	(P. Wang et al., 2015) N/A (R. N. Lichtenwalter & Chawla, 2012) AUROC = 0.951
Flow Propagation	(Martínez et al., 2017)	Iterative definition corresponds to a propagation process	$M = D^l A D^r$ $D^l_{i,i} = 1 \Big/ \sqrt{\sum_j A_{i,j}}$ $D^r_{i,i} = 1 \Big/ \sqrt{\sum_j A_{j,i}}$	O(cv^{2}d)	Yes	(Martínez et al., 2017) AUC = 0.9674

Table 3. Quasi Similarity Measures

Measure	References	Purpose	Formula	Time complexity	normalized similarity score	References (accuracy, AUC, AUROC)
The Local Path Index (LPI)	(Kumar et al., 2020), (Mutlu et al., 2020), (Srilatha & Manjula, n.d.)	Strongly based on the Katz index but it only considers a finite number of path lengths	$S = \sum_{i=2}^{l} \beta^{i-2} A^i$	O(lv^{2}k)	-	(Mutlu et al., 2020) N/A (Kumar et al., 2020) AUC = 0.9912 (Martínez et al., 2017) AUROC = 0.9591
Local Random Walks (LRW)	(J. H. Liu et al., 2016), (Ghorbanzadeh et al., 2021b), (Fu et al., 2018), (Linyuan & Zhou, 2011)	Exploit the concept of random walks but limit the number of iterations to a fixed a priori small number l	$S^{a,b}(t) = \frac{\lvert\Gamma_a\rvert}{2\lvert E\rvert}\overrightarrow{p_b^a}(t) + \frac{\lvert\Gamma_b\rvert}{2\lvert E\rvert}\overrightarrow{p_a^b}(t)$	O(lv^{2}k)		(Fu et al., 2018) accuracy = 0.6650 (Linyuan & Zhou, 2011) N/A (Ghorbanzadeh et al., 2021b) N/A (J. H. Liu et al., 2016) LRW accuracy = 0.8412 (Martínez et al., 2017) AUC = 0.9647
Superposed Random Walks (SRW)	(Ahmed et al., 2016)	Based on the local random walk method, has been proposed to counteract this issue by continuously releasing the walker at the starting node	$S^{a,b}(t) = \sum_{i=1}^{t}\left(\frac{\lvert\Gamma_a\rvert}{2\lvert E\rvert}\overrightarrow{p_b^a}(t) + \frac{\lvert\Gamma_b\rvert}{2\lvert E\rvert}\overrightarrow{p_a^b}(t)\right)$	O(lv^{2}k)	-	(Ahmed et al., 2016) N/A (Martínez et al., 2017) AUC = 0.9726
Third-Order Resource Allocation Based on Common Neighbor Interactions (ORA-CNI)	(Mutlu & Oghaz, 2020)	Extension of resource allocation based on common neighbor interactions	$S(a,b) = \sum_{f\in\Gamma_a\cap\Gamma_b}\frac{1}{\lvert\Gamma_f\rvert} + \sum_{e_{i,j}\in E,\lvert\Gamma_i\rvert<\lvert\Gamma_i\rvert,i\in\Gamma_a,j\in\Gamma_b}\left(\frac{1}{\lvert\Gamma_i\rvert} - \frac{1}{\lvert\Gamma_j\rvert}\right) + \beta\sum_{[a,p,q,b]\in paths^3_{a,b}}\frac{1}{\lvert\Gamma_p\rvert\lvert\Gamma_q\rvert}$	$O(vk^6)$	-	(Mutlu & Oghaz, 2020) N/A (Martínez et al., 2017) AUC = 0.9682
FriendLink (FL)	(Papadimitriou et al., 2012)	Similar to the local path index	$S(a,b) = \sum_{i=2}^{l}\frac{1}{i-1}\frac{(A^i)_{a,b}}{\prod_{j=2}^{i}(\lvert V\rvert - j)}$	O(lv^{2}k)	yes	(Papadimitriou et al., 2012) N/A (Martínez et al., 2017) AUC = 0.9619

PropFlow Predictor (PFP)	(Mutlu et al., 2020), (R. Lichtenwalter & Chawla, 2012)	This method is similar to the random walk with restart or rooted PageRank algorithms	$S_{a,j} = S_{a,j} + \frac{S_{a,i}}{\lvert \Gamma_i \rvert}$	O(vlk^l)	-	(Mutlu et al., 2020) N/A (R. Lichtenwalter & Chawla, 2012) N/A (Martínez et al., 2017) AUC = 0.9636

Table 4. The LP Approaches that Used Content-Based Features as Text Similarity Matrices

Reference	Text Analysis Techniques	Points of Strength	Limitations	Recall	Precision	F1-score	AUC
(Chuan et al., 2018)	• LDAcosin with mathematical notation for LP	• Use the paper's contents in the proposed metric LDA Cosine for LP • A mathematical notion of the LP in the co-authorship network and a LP algorithm based on topic modeling	• high computational time in comparison with the relevant algorithms since they have to calculate the content similarity through the LDA method. • The values of AUC were not high • More information from the content is needed such as the user's information (affiliation, study subjects).	0.8235	0.2096	0.3250	0.6626
(Parimi & Carage a, 2011)	• topic modeling approach- the Latent Dirichlet Allocation (LDA)	• Better performance results. • Using Interest features alone are better than other features	• high memory size and computational time • only the static image of the LiveJournal social network is considered.	-	-	-	0.9046 ± 0.01
(Hassan , n.d.)	• used predictors extracted by computing the similarity between the research interests, their affiliations, the sum of research performance indices, and the similarity between the two author nodes themselves	• introducing new predictor variables for solving LP problems using supervised learning in academic SN. • The new predictors are the research interests and affiliations of each author pair and the research performance indices of each author pair.	• Huge dataset used makes the memory size limited	0.2020	0.3190	0.2330	0.6800
(Zhao et al., 2017)	• fully Bayesian approach model for LP. Used a trains LDA model for all papers of the dataset	• The proposed method is effectively modeling node attributes • The model is scalable for large but sparse relational networks with large sets of node attributes • the proposed models work on directed and undirected relational networks with flat and hierarchical node attributes.	No limitation is recorded	-	-	-	-

(Sachan & Ichise, 2010)	• A semantic approach named as Abstract-Keywords Match Count (AKMC)	• An improved method for LP by utilizing node attributes like abstract information and local network density is built.	No limitation is recorded	-	-	-	-
(Muniz et al., 2018)	• A combination of global similarity indices and content-based measure.	• Three weighting criteria have been proposed that combine contextual, temporal, and topological information to improve results in unsupervised LP.	No limitation is recorded	-	-	-	-
(Quercia et al., 2012)	• A model is known as Labeled-LDA	• understanding how well a fairly new version of topic modeling (i.e., L-LDA) works in the specific context of Twitter, an increasingly useful source of informative textual data. • L-LDA is a suitable profile classification method when only a small number of tweets exist for each training profile.	L-LDA effectiveness may be limited when the number of Tweet training profiles is small.	-	-	-	-

Table 5. The Literature of Proposed Hybrid Methods for LP

Paper title	Graph Analysis Formula	Text Analysis Techniques	Points of Strength	Limitations	Recall	Precision	F1-score	AUC	Accuracy
Link prediction in co-authorship networks based on hybrid content similarity metric (Chuan et al., 2018)	$SIM_{LDAcosin(u,v)} = S(P_u, P_v) \times \frac{1}{\lvert\Gamma_u \cap \Gamma_v\rvert} \times \sum_{z \in \Gamma_u \cap \Gamma_v} S(P_{uz}, P_{vz})$	the proposed method is based on the measurement of content similarity to estimate the similar scores of author pairs for the LP.	Mathematical notions of the LP in the co-authorship network and a LP algorithm based on topic, modeling is proposed	high computational time in comparison with the relevant algorithms	0.8235	0.2096	0.3250	0.6626	-
Combining contextual, temporal and topological information for unsupervised link prediction in social networks (Muniz et al., 2018)	$W^{CTT}(u,v) = \lvert E(u,v)\rvert * \beta \frac{CTime - max(t_{(u,v)})}{CTime - \min(t)} * \alpha^{1-\cos(u,v)}$	A combination of global similarity indices and content-based measure	Three weighting criteria have been proposed that combine contextual, temporal, and topological information to improve results in unsupervised LP.	No limitations are recorded	-	-	-	-	-
ConPredict a Method for Link Prediction in Co-authored	No formula mentioned	ConPredict combines the standards-based approach	A hybrid approach is proposed that consider the network topology and the contents of the nodes	No limitations are recorded	0.8421	0.9350	0.8602	-	-

Content-Based Networks (Antunes et al., 2013)		structural network from and the approach based on the similarity between nodes	(title and abstract) of researchers' articles						
LP-UIT: A Multimodal Framework for Link Prediction in Social Networks (Wu et al., 2022)	For word representation $TF-IDF(i,j) = TF(i,j) \times IDF(i)$	use the TF-IDF method to extract W textual words from the correspondin g activities	The proposed model outperforms state-of-the-art methods in terms of accuracy.	Not mentioned	-	-	-	0.951 6	-
Predicting links in social networks using text mining and SNA (Bartal et al., 2009)	No specific formula presented	Text Mining and SNA (NLP-Natural Language Processing and VSM-vector space model)	solve the LP problem in an academic co-authoring network by combining Text Data Mining methods to evaluate and represent authors' interest topics by extracting key concepts from common articles titles and SNA methods.	The prediction without Text data mining gives less accuracy compared to that with it	-	-	-	-	0.97 73

augmented adjacency matrix to calculate the similarity indices at each time step. The results of their experiments demonstrated that their methods produce higher quality results for LP in temporal social networks.

Antunes et al. (2013) proposed the ConPredict method for LP in co-authored content-based networks, which utilized the structure and content represented by each network node as a source of information. The title and abstract of the articles published by researchers were utilized as content similarity. For the proposed method, the authors used a hybrid approach (based on structural patterns in the network and similarity between nodes) by exploring the network topology using two metrics: the shortest path distance between two nodes and their similarity. The authors concluded that there is a possibility to predict a new relationship when using the hybrid method for the co-authored network nodes.

Parimi and Caragea, (2011) presented a new multimodal framework for LP (referred to as LP-UIT) in 2022. It made use of a broad collection of features collected from multi-modal data (such as user information and topological features) (i.e., textual information, graph information, and numerical information). The model utilized a graph convolutional network to process network information to capture topological features, natural language processing techniques (i.e., TF/IDF and word2Vec) to model users' short-term and long-term interests, and numerical features to identify social influence and "weak links." The link between textual and topological variables using an attention mechanism is described as well. Finally, for LP, a Multi-Layer Perceptron (MLP) is built to merge the representations from three modalities.

The LP problem has been examined by the authors in 2009 using an academic co-authorship network using text mining methods (Bartal et al., 2009). The authors addressed the problem by combining the Text Data Mining (TDM) methods to evaluate and represent authors' interest topics by extracting key concepts from common article titles and SNA methods. Two TDM methods are compared (NLP-Natural Language Processing and VSM-vector space model) and the relevance of this knowledge to the prediction accuracy is examined. The goal of this research was to find which measures of the network can lead to the most accurate LPs and to examine whether TDM methods can contribute to the overall prediction accuracy. It has been shown through empirical testing that the two predictions, with TDM and without TDM, have different results in favor of using TDM in the prediction algorithm. Table 4 demonstrated the recent literature that utilized the content-based measures as a text similarity measure.

In their article Hasan et al. (n.d.), considered a supervised machine learning approach for the prediction of non-existing links. They created several machine learning models to capture the topological information associated with network links and nodes. They extracted the proximity feature based on the node content and confirmed that using social network data as well as graph topology can significantly improve the prediction result.

4. EVALUATION MEASURE IN LINK PREDICTION

The evaluation metrics that are commonly used in LP are similar to those utilized in any binary classification task in general. However, other metrics are scale-dependent and used for regression models such as RMSE (Root Mean Square Error) and MAE (Mean Absolute Error), where they estimate the fitness of the model. Both are representing the differences between predicted link-prediction score values and observed link-prediction values to measure the accuracy of the model. Thus, in the case of using any regression technique such as, Linear Regression to predict links in a network, these two measurements can be used to evaluate the similarity score of a pair of nodes.

The evaluation measurements in LP are classified into two types: fixed threshold metrics and threshold curves (Haghani & Keyvanpour, 2019), as depicted in Figure 3. Top-N prediction precision and recall are typically fixed threshold metrics. These measures frequently suffer from the limitation of relying on a reasonable threshold. This is rarely true in research contexts for performance without being tied to any particular domain or deployment (Antunes et al., 2013). Threshold curves are the alternative to the fixed ones such as the receiver operating characteristic (ROC) curve (Davisu & Goadrich, 2016)(R. Lichtenwalter & Chawla, 2012)(Hanley & McNeil, 1982), derived curves like cost curves (Drummond & Holte, 2006) and precision-recall curves (PR) (Davisu & Goadrich, 2016) are widely used in LP evaluation. Furthermore, in the presence of imbalanced data, the F1-Score and ROC are considered the best metrics (Samad et al., 2020)(Hanley & McNeil, 1982). AUC can be defined as the likelihood that a randomly chosen missing link will have a higher score than a randomly chosen non-existent link. Since it is more complicated to specify and explain LP evaluation strategies than standard classification, where it is sufficient to fully specify a dataset, new evaluation methods or performance metrics must be proposed (Antunes et al., 2013).

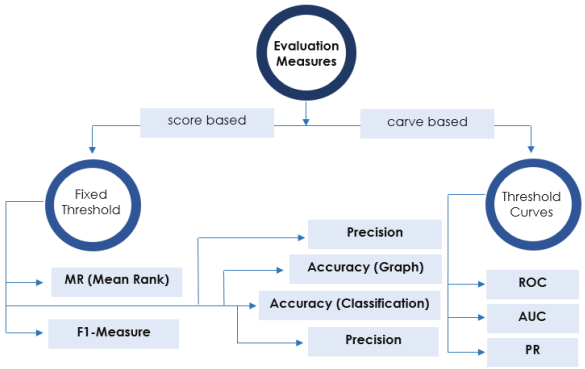

Figure 3. Taxonomy of Performance Evaluation Measures, adapted from (Samad et al., 2020)

4.1 Fixed-Threshold

In general, fixed-threshold metrics can utilize different types of thresholds, such as those based on prediction score, percentage of instances, or a number of instances. There are additional constraints in LP in particular. In the literature, accuracy, precision, recall, and top-k equivalents are frequently used in LP. To present the evaluation metrics, it is necessary to introduce the following common terms:

1. True Positive (TP): The number of node pairs that have links is correctly identified as positive.
2. False Positive (FP): The number of node pairs that do not have links is incorrectly classified as positive.
3. False Negative (FN): The number of node pairs that have links is incorrectly recognized as negative.
4. True Negative (TN): The number of node pairs that do not have links is correctly judged as negative.

Precision indicates the proportion of the actual number of node pairs with links in the node pairs predicted as positive instances. The higher the value of Precision, the better the prediction performance. The calculation formula is:

$$Precisian = \frac{TP}{TP + FP} \quad (1)$$

The recall is the proportion of correctly predicted node pairs among all the actual linked node pairs. Mathematically

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

F1-score is the harmonic average of Precision and Recall, which is

$$F_1 - score = \frac{2Precisian \times Recall}{Precisian + Recall} \quad (3)$$

Accuracy is the proportion of all correctly predicted node pairs, which is defined as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

MAE (Mean Absolute Error) is utilized to reflect the prediction error rate. The smaller the value, the better the prediction performance. The mathematical expression of MAE reads as:

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |y_i - \hat{y}_i|, \quad (5)$$

where y_i and b $\hat{y}_i$represent the actual class label and predicted class label of instance i, respectively. m is the number of instances to be predicted.

RMSE (Root Mean Square Error) is employed to measure the deviation between the predicted value and the actual value. More explicitly:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2} \quad (6)$$

Accuracy (Graph): Graph accuracy is the same as classification accuracy. However, graph accuracy takes into account the original graph and predicted graph.

$$Accuracy = 1 - \frac{E(G_p) + E(G_0) - 2E(G_p \cap G_0)}{MAX\left(E(G_p), E(G_0)\right)} \quad (7)$$

Where E corresponds to the edges of a graph, G_p corresponds to the predicted graph and G_0 denotes the original graph.

4.2 Threshold-Curves

Threshold curves are commonly used in the binary classification community to express results due to the rarity of cases where researchers have reasonable fixed thresholds. They are particularly popular when the class distribution is highly imbalanced, and as a result, they are increasingly utilized in LP evaluation (Lichtenwalter et al., 2010). Threshold curves similarly accept scalar measures, which serve as a single performance summary statistic. The ROC curve depicts the true positive rate with the false positive rate at all classification thresholds, and its area (AUC) is equivalent to the probability of a randomly chosen positive instance appearing above a randomly chosen negative instance. However, ROC curves can present an overly optimistic view of an algorithm's performance if there is a large skew in the class distribution whereas, in such a situation, PR is an alternative. A significant distinction between ROC space and PR space is the visual representation of the curves, with PR curves revealing differences between algorithms that are not visible in ROC space. The researchers in (Drummond & Holte, 2006) went into great detail about the relationship between ROC space and PR.

ROC: ROC is an abbreviation of receive operation characteristics. It narrates fragments of false-positive rates versus true-positive rates on different thresholds. Where the true positive rate is

$$TPR = \frac{TP}{TP + FN} \quad (8)$$

and the false positive rate is as follows:

$$FPR = \frac{FP}{TN + FP} \quad (9)$$

Where TPR estimates the portion of correctly predicted positive links. While FPR estimates the misinterpreted negative links.

PR: The precision-recall curve is abbreviated as PR. It represents precision and recall at various thresholds. It only considers positive links and ignores negative links. Because it is necessary to predict removed links in periodic LP, the PR curve is unsuitable.

AUC: The area under the ROC is abbreviated as AUC. In this case, a high AUC represents superior classification results, whereas a low AUC represents poor classification results.

5. DISCUSSION

This study provides a different and up-to-date survey of the literature on LP analysis in the scientific collaboration network. To offer a broad grasp of LP, the paper commenced by providing a number of review and survey papers in various LP areas that covered a wide variety of elements and approaches, such as binary classification models, probabilistic models, and linear algebraic models in (Mohammad Al Hasan, Zaki, 2011), random walk and maximum likelihood methods in (Linyuan & Zhou, 2011), deep learning methods in (Haghani & Keyvanpour, 2019), and others in (Wang et al., 2015) (Martínez et al., 2016)(Mutlu & Oghaz, 2020) (Wang & Le, 2020) (Liben-Nowell & Kleinberg, 2007).

The three-dimensional taxonomy incorporates three methodologies, named topological-based approaches, content-based approaches, and hybrid-based approaches as well. The topological-based approaches are the simplest among the LP approaches, as it is provided as a score ranking for each unobserved pair of nodes. The techniques can be used effectively in certain networks (Linyuan & Zhou, 2011). These methods are based on the closeness of similarities between disconnected node pairs (Wang et al., 2015). The similarity is derived based on the identification of prospective pairing node candidates, which is defined as E (u, v), where u and v of unconnected node pairings are computed as the index similarity score using the specified graph-based methods. The index scores are ordered from highest to lowest and the pair of nodes with the highest probability of producing new or missing connections receives the highest score. The advantages of topological-based approaches are the simplicity and the low computational time compared to the other types of metrics. On the other hand, the main drawback is that these metrics could be complicated for some networks where it can work for some networks and fail for another. For the investigation of social networks using topological techniques at various levels, many metrics were used: Node-level (measures of centrality and non-centrality) [36] as well as graph-level (Local, Global and Quasi measures, see Tables 1, 2 and 3).

While the majority of approaches are based on network topological properties such as degree, clustering coefficient, path index, and so on (Kumar et al., 2020), (Ma et al., 2021), the structural information is less susceptible to noise. Therefore, several studies in SNA have focused on using supervised machine learning algorithms to develop an optimized model for predicting links in real-world networks by analyzing the content associated with nodes and edges (L. Dong et al., 2013), (Haghani

& Keyvanpour, 2019), (Chuan et al., 2018), (Gao et al., 2015), (Shahrabi Farahani, Alavi, Ghasem, 2020). However, these are likely to be less accurate because the content might contain noisy and irrelevant data, necessitating 90 percent of the time spent only on data cleansing (Kumari et al., 2020). On the other hand, content-based characteristics (non-topological features), have the benefit of enhancing the performance of the LP issue. Nevertheless, they are sometimes available and perhaps hard to attain. Furthermore, the vast majority of content-based capabilities are domain-specific such that identifying and uncovering them demands an understanding of the domain. As a result, while a general LP learning model normally only takes into account generic data such as node, network, and topological information, non-topological features need to be taken into account for a practical LP application as well (Wang et al., 2015). The content-based approaches can be divided into topic modelling approaches and text mining approaches. In terms of topic modelling techniques, the LDA method is utilized in this study to list all related articles, as presented in Table 4, and it is the most common and simplest method that improved the performance of the LP problem based on the literature. Text mining methods, on the other hand, play a minor part in the LP problem. This is based on scant research on the application of these strategies for LP in social networks.

As a result, it is preferable to combine the information generated from current score functions in a hybrid manner to obtain more comprehensive score functions that integrate the advantages of the aforementioned methodologies. Therefore, the analysis of networks using the three types of information was necessary to discover and quantify the interesting facts on both the individual and group levels of collaboration networks, such as conferences. The goal of the analysis is to come across the fact that the best LP approach for predicting the invisible upcoming collaboration patterns and trends over time can be attained using text and graph mining methods to improve overall prediction accuracy for collaboration networks.

In the terms of accuracy of performance, the majority of the topological metrics produced good accuracy, while others did not, as shown in Tables 1, 2, and 3. According to the literature, the local similarity measures in Table 1 is providing high accuracy and AUC for Jaccard Coefficient (JC) metric with = 0.9993 accuracy (Liu et al., 2019) and 0.9895 AUCROC (Kumar et al., 2020), but with a high time-complexity of O(VK3) , this is due to not just considering a deeper analysis of neighborhood analysis than just counting the number of shared common neighbors. The majority of the literature work in the Table 2 used the AUC measure to assess the efficacy of their model, though. Thus, the global metric Rooted PageRank had the best AUC, with value of 0.9930 (Linyuan & Zhou, 2011) using the famous Rank searching algorithm, where the individual researcher nodes in the graph contribute in determining the best candidate as members of the future co-authorship collaborations. Similar to Table 2, most of the metrics in Table 3 used the AUC measure for evaluation. In comparison to other quasi-similarity metrics, the Local Path Index (LPI) had the greatest AUC with value of 0.9912 (Kumar et al., 2020), due to use of the Katz index by considering a finite number of path lengths among the researchers in the global collaboration network. Except for references (Hassan, n.d.), (Chuan et al., 2018), and (Parimi & Caragea, 2011), the majority of the literature, where the LDA metric was used to measure the similarity of a text, did not provide any assessment metrics for their works in Table 4, as the evaluation is done using the AUC metric, and the work of reference (Parimi & Caragea, 2011) produced the best results with a value of 0.9046 by using the researcher's interest features alone for predicting the future collaboration links.

6. CONCLUSION

The LP problem has applications in a wide range of domains. One of the main applications is to forecast future scientific and collaboration trends in the scientific collaboration network. This review presented an approach-based methodology to categorize the work done in the literature. To conclude, the main efforts were spent on utilizing the topological analysis of the collaboration networks, such as considering the number of common neighbors/collaborators among the nodes/researchers, thus ignoring the actual academic content of the research materials written by the research collaborators.

In terms of time complexity; most of the topological metrics produce the run-time of O(VK3), but the local similarity metric of Negated Shortest Path (NSP) gives the best runtime complexity of O (ev log v), which excludes the most of the traversing path in the social graph being analyzed. Whereas, the global similarity metrics the Vertex Collocation Profile (VCP) produced the best runtime of O(|E|/|V|), which is a vector-based metric for describing the relationship between two vertices, in terms of their common membership. However, more promising runtimes are yet to be proposed in LP literature. In light of the above information and since the formation of the link between nodes is influenced by several score functions and due to the lack of such scores, it is concluded that the ability to aggregate features (topological and content-based) derived from existing score functions in a hybrid methodology should yield the ability to predict more robust relationships in the network under analysis. This is the reason why the aggregated features were viewed through a hybrid lens which leads that multiple score functions being introduced into the literature to categorize two classes for a pair of nodes using different comprehensive measurements.

ACKNOWLEDGEMENTS

The authors would like to express their sincere thanks to the department of computer science and the CCNP research lab at the faculty of science in the University of Zakho for offering the laboratory facilities and providing financial support.

REFERENCES

Abdul, M., & Mastan, N. (2013). *REVIEW A Survey on LDA Approach in Predicting Link Behavior in Social Networks*. *2*(3), 176–180.

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, *25*(3), 211–230. https://doi.org/10.1016/S0378-8733(03)00009-1

Aghabozorgi, F., & Khayyambashi, M. R. (2018). A new similarity measure for link prediction based on local structures in social networks. *Physica A: Statistical Mechanics and Its Applications*, *501*, 12–23. https://doi.org/10.1016/j.physa.2018.02.010

Ahmed, N. M., Chen, L., Wang, Y., Li, B., Li, Y., & Liu, W. (2016). Sampling-based algorithm for link prediction in temporal networks. *Information Sciences*, *374*, 1–14. https://doi.org/10.1016/j.ins.2016.09.029

Alghamdi, R., & Alfalqi, K. (2015). A Surv e y o f Topic Mode ling i n Text Mining. *Nternational Journal of Advanced Computer Science and Applications*, *6*(1), 7.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. http://arxiv.org/abs/1707.02919

Antunes, J. B., Antunes, J. B., Filho, H. F. B. P., Maia, R. D., De Queiroz, R. B., Da Silva, C. M. R., Rodrigues, R. B., & De Almeida Barros, F. (2013). ConPredict: A method for link prediction in co-authored content-based networks. *Proceedings of the IADIS International Conference WWW/Internet 2013, ICWI 2013*, *January*, 11–18.

Assouli, N., Benahmed, K., & Gasbaoui, B. (2021). How to predict crime

— informatics-inspired approach from link prediction. *Physica A: Statistical Mechanics and Its Applications*, *570*. https://doi.org/10.1016/j.physa.2021.125795

Bahabadi, M. D., Golpayegani, A. H., & Esmaeili, L. (2014). *A Novel C2C E-Commerce Recommender System Based on Link Prediction: Applying Social Network Analysis*.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Bartal, A., Sasson, E., & Ravid, G. (2009). Predicting links in social networks using text mining and SNA. *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2009*, 131–136. https://doi.org/10.1109/ASONAM.2009.12

Berlusconi, G., Calderoni, F., Parolini, N., Verani, M., & Piccardi, C. (2016). Link prediction in criminal networks: A tool for criminal intelligence analysis. *PLoS ONE*, *11*(4). https://doi.org/10.1371/journal.pone.0154244

Bhattacharyya, P., Garg, A., & Wu, S. F. (2011). Analysis of user keyword similarity in online social networks. *Social Network Analysis and Mining*, *1*(3), 143–158. https://doi.org/10.1007/s13278-010-0006-4

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17–35. https://doi.org/10.1214/07-aoas114

Börner, K., Maru, J. T., & Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(SUPPL. 1), 5266–5273. https://doi.org/10.1073/pnas.0307625100

Chuan, P. M., Son, L. H., Ali, M., Khang, T. D., Huong, L. T., & Dey, N. (2018). Link prediction in co-authorship networks based on hybrid content similarity metric. *Applied Intelligence*, *48*(8), 2470–2486. https://doi.org/10.1007/s10489-017-1086-x

Coskun, M., & Koyuturk, M. (2016). Link Prediction in Large Networks by Comparing the Global View of Nodes in the Network. *Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015*, 485–492. https://doi.org/10.1109/ICDMW.2015.195

Crichton, G., Guo, Y., Pyysalo, S., & Korhonen, A. (2018). Neural networks for link prediction in realistic biomedical graphs: A multi-dimensional evaluation of graph embedding-based approaches. *BMC Bioinformatics*, *19*(1), 1–11. https://doi.org/10.1186/s12859-018-2163-9

Daud, N. N., Ab Hamid, S. H., Saadoon, M., Sahran, F., & Anuar, N. B. (2020). Applications of link prediction in social networks: A review. In *Journal of Network and Computer Applications* (Vol. 166). Academic Press. https://doi.org/10.1016/j.jnca.2020.102716

Davisu, J., & Goadrich, M. (2016). *The relationship between precision-recall and ROC curves*. 233–240.

De Tre, G., Hallez, A., & Bronselaer, A. (2014). Performance optimization of object comparison. *International Journal of Intelligent Systems*, *29*(2), 495–524. https://doi.org/10.1002/int

Dong, L., Li, Y., Yin, H., Le, H., & Rui, M. (2013). The algorithm of link prediction on social network. *Mathematical Problems in Engineering*, *2013*. https://doi.org/10.1155/2013/125123

Dong, Y., Ke, Q., Wang, B., & Wu, B. (2011). Link prediction based on local information. *Proceedings - 2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011*, 382–386. https://doi.org/10.1109/ASONAM.2011.43

Drummond, C., & Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, *65*(1), 95–130. https://doi.org/10.1007/s10994-006-8199-5

E Fonseca, B. de P. F., Sampaio, R. B., Fonseca, M. V. de A., & Zicker, F. (2016). Co-authorship network analysis in health research: Method and potential use. In *Health Research Policy and Systems* (Vol. 14, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12961-016-0104-5

F Shahrabi Farahani, M Alavi, M Ghasem, Bt. (2020). Scientific Map of Papers Related to Data Mining in Civilica Database Based on Co-Word Analysis. *International Journal of Web Research*, *3*(1), 11–18.

Fu, C., Zhao, M., Fan, L., Chen, X., Chen, J., Wu, Z., Xia, Y., & Xuan, Q. (2018). Link Weight Prediction Using Supervised Learning Methods and Its Application to Yelp Layered Network. *IEEE Transactions on Knowledge and Data Engineering*, *30*(8), 1507–1518. https://doi.org/10.1109/TKDE.2018.2801854

Gao, F., Musial, K., Cooper, C., & Tsoka, S. (2015). Link prediction methods and their accuracy for different social networks and network metrics. *Scientific Programming*, *2015*(i). https://doi.org/10.1155/2015/172879

Ghorbanzadeh, H., Sheikhahmadi, A., Jalili, M., & Sulaimany, S. (2021a). A hybrid method of link prediction in directed graphs. *Expert Systems with Applications*, *165*(February 2020), 113896. https://doi.org/10.1016/j.eswa.2020.113896

Ghorbanzadeh, H., Sheikhahmadi, A., Jalili, M., & Sulaimany, S. (2021b). A hybrid method of link prediction in directed graphs. *Expert Systems with Applications*, *165*. https://doi.org/10.1016/j.eswa.2020.113896

Haghani, S., & Keyvanpour, M. R. (2019). A systemic analysis of link prediction in social network. *Artificial Intelligence Review*, *52*(3), 1961–1995. https://doi.org/10.1007/s10462-017-9590-2

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hasan, M. Al, Chaoji, V., Salem, S., Zaki, M., & York, N. (n.d.). *Link Prediction using Supervised Learning*.

Hassan, D. (n.d.). *SUPERVISED LINK PREDICTION IN CO-AUTHORSHIP NETWORKS BASED ON RESEARCH PERFORMANCE AND SIMILARITY OF RESEARCH INTERESTS AND AFFILIATIONS*.

Hemkiran, S., & Sudha Sadasivam, G. (2020). A review of similarity measures and link prediction models in social networks. *International Journal of Computing and Digital Systems*, *9*(2), 239–248. https://doi.org/10.12785/IJCDS/090209

Ibrahim, N. M. A., & Chen, L. (2015). Link prediction in dynamic social networks by integrating different types of information. *Applied Intelligence*, *42*(4), 738–750. https://doi.org/10.1007/s10489-014-0631-0

Jaccard, P. (1982). Etude de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de La Murithienne*, *XXXVII*, 81-92. https://doi.org/10.5169/seals-266450

Jaya Lakshmi, T., & Durga Bhavani, S. (2017). Link prediction in temporal heterogeneous networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10241 LNCS*, 83–98. https://doi.org/10.1007/978-3-319-57463-9_6

Kong, X., Shi, Y., Yu, S., Liu, J., & Xia, F. (2019). Journal of Network and Computer Applications Academic social networks : Modeling , analysis , mining and applications. *Journal of Network and Computer Applications*, *132*(December 2018), 86–103. https://doi.org/10.1016/j.jnca.2019.01.029

Kumar, A., Mishra, S., Singh, S. S., Singh, K., & Biswas, B. (2020). Link prediction in complex networks based on Significance of Higher-Order Path Index (SHOPI). *Physica A: Statistical Mechanics and Its Applications*, *545*, 123790. https://doi.org/10.1016/j.physa.2019.123790

Kumari, A., Behera, R. K., Sahoo, K. S., Nayyar, A., Kumar Luhach, A., & Prakash Sahoo, S. (2020). Supervised link prediction using structured-based feature extraction in social network. *Concurrency Computation , February*, 1–16. https://doi.org/10.1002/cpe.5839

Kushwah, A. K. S., & Manjhvar, A. K. (2016). A review on link prediction in social network. *International Journal of Grid and Distributed Computing*, *9*(2), 43–50. https://doi.org/10.14257/ijgdc.2016.9.2.05

Leicht, E. A., Holme, P., & Newman, M. E. J. (2006). Vertex similarity in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *73*(2), 1–10. https://doi.org/10.1103/PhysRevE.73.026120

Li, L., Wang, L., Luo, H., & Chen, X. (2021). Towards effective link prediction: A hybrid similarity model. *Journal of Intelligent and Fuzzy Systems*, *40*(3), 4013–4026. https://doi.org/10.3233/JIFS-200344

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019–1031. https://doi.org/10.1002/asi.20591

Lichtenwalter, R., & Chawla, N. V. (2012). Link prediction: Fair and

effective evaluation. *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, 376–383. https://doi.org/10.1109/ASONAM.2012.68

Lichtenwalter, R. N., & Chawla, N. V. (2012). *Vertex collocation profiles*. *1019*, 1019–1028. https://doi.org/10.1145/2187836.2187973

Lichtenwalter, R. N., Lussier, J. T., & Chawla, N. V. (2010). New perspectives and methods in link prediction. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 243–252. https://doi.org/10.1145/1835804.1835837

Linyuan, L. L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications*, *390*(6), 1150–1170. https://doi.org/10.1016/j.physa.2010.11.027

Liu, H., Kou, H., Yan, C., & Qi, L. (2019). Link prediction in paper citation network to construct paper correlation graph. *Eurasip Journal on Wireless Communications and Networking*, *2019*(1). https://doi.org/10.1186/s13638-019-1561-7

Liu, J. H., Zhu, Y. X., & Zhou, T. (2016). Improving personalized link prediction by hybrid diffusion. *Physica A: Statistical Mechanics and Its Applications*, *447*, 199–207. https://doi.org/10.1016/j.physa.2015.12.036

Liu, S., Ji, X., Liu, C., & Bai, Y. (2017). Extended resource allocation index for link prediction of complex network. *Physica A: Statistical Mechanics and Its Applications*, *479*, 174–183. https://doi.org/10.1016/j.physa.2017.02.078

Liu, X., Zhang, J., & Guo, C. (2013). Full-text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology*, *64*(9), 1852–1863. https://doi.org/10.1002/asi.22883

Ma, G., Yan, H., Qian, Y., Wang, L., Dang, C., & Zhao, Z. (2021). Path-based estimation for link prediction. *International Journal of Machine Learning and Cybernetics*, *12*(9), 2443–2458. https://doi.org/10.1007/s13042-021-01312-w

Martin, T., Ball, B., Karrer, B., & Newman, M. E. J. (2013). Coauthorship and citation patterns in the Physical Review. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *88*(1), 1–9. https://doi.org/10.1103/PhysRevE.88.012814

Martínez, V., Berzal, F., & Cubero, J.-C. (2017). A Survey of Link Prediction in Complex Networks. *ACM Computing Surveys*, *49*(4), 1–33. https://doi.org/10.1145/3012704

Martínez, V., Berzal, F., & Cubero, J. C. (2016). A survey of link prediction in complex networks. *ACM Computing Surveys*, *49*(4). https://doi.org/10.1145/3012704

Mishra, S., & Nandi, G. C. (2015). A novel hybrid approach for link prediction problem in social network. *International Journal of Social Network Mining*, *2*(2), 115. https://doi.org/10.1504/ijsnm.2015.072281

Mohammad Al Hasan, Zaki, M. J. (2011). A SURVEY OF LINK PREDICTION IN SOCIAL NETWORKS. In *Social Network Data Analytics*. https://doi.org/10.1007/978-1-4419-8462-3

Muniz, C. P., Goldschmidt, R., & Choren, R. (2018). Combining contextual, temporal and topological information for unsupervised link prediction in social networks. *Knowledge-Based Systems*, *156*, 129–137. https://doi.org/10.1016/j.knosys.2018.05.027

Mutlu, E. C., & Oghaz, T. (2020). Review on Graph Feature Learning and Feature Extraction Techniques for Link Prediction. *Machine Learning and Knowledge Extraction*, *2*(4), 672–704. https://doi.org/10.3390/make2040036

Mutlu, E. C., Oghaz, T., Rajabi, A., & Garibay, I. (2020). Review on Learning and Extracting Graph Features for Link Prediction. *Machine Learning and Knowledge Extraction*, *2*(4), 672–704. https://doi.org/10.3390/make2040036

Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *64*(2), 4. https://doi.org/10.1103/PhysRevE.64.025102

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*(2), 167–256. https://doi.org/10.1137/S003614450342480

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(SUPPL. 1), 5200–5205. https://doi.org/10.1073/pnas.0307545100

Papadimitriou, A., Symeonidis, P., & Manolopoulos, Y. (2012). Scalable link prediction in social networks based on local graph characteristics. *Proceedings of the 9th International Conference on Information Technology, ITNG 2012*, 738–743. https://doi.org/10.1109/ITNG.2012.145

PARIMI, R. (2010). *LDA BASED APPROACH FOR PREDICTING FRIENDSHIP LINKS IN LIVE Copyright Rohit Parimi*.

Parimi, R., & Caragea, D. (2011). Predicting friendship links in social networks using a topic modeling approach. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6635 LNAI*(PART 2), 75–86. https://doi.org/10.1007/978-3-642-20847-8_7

Quercia, D., Askham, H., & Crowcroft, J. (2012). TweetLDA: Supervised topic classification and link prediction in Twitter. *Proceedings of the 4th Annual ACM Web Science Conference, WebSci'12*, *volume*, 247–250. https://doi.org/10.1145/2380718.2380750

Raut, P., Khandelwal, H., & Vyas, G. (2020). A Comparative Study of Classification Algorithms for Link Prediction. *2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020 - Conference Proceedings*, *Icimia*, 479–483. https://doi.org/10.1109/ICIMIA48430.2020.9074840

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A.-L. (2002). Hierarchical Organization of Modularity in Metabolic Networks. *Science*, *297*(5586), 1551–1555. https://doi.org/10.1126/science.1073374

Sachan, M., & Ichise, R. (2010). Using Semantic Information to Improve Link Prediction Results in Network Datasets. *International Journal of Engineering and Technology*, *2*(4), 334–339. https://doi.org/10.7763/ijet.2010.v2.143

Samad, A., Qadir, M., & Nawaz, I. (2019). SAM: A Similarity Measure for Link Prediction in Social Network. *MACS 2019 - 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics, Proceedings*. https://doi.org/10.1109/MACS48846.2019.9024762

Samad, A., Qadir, M., Nawaz, I., Islam, M., & Aleem, M. (2020). A Comprehensive Survey of Link Prediction Techniques for Social Network. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, *7*(23), 163988. https://doi.org/10.4108/eai.13-7-2018.163988

Shao, C. X., Dou, H. L., Yang, R. X., & Wang, B. H. (2013). Zero nodes effect: Valid link prediction in sparse networks. *International Journal of Modern Physics B*, *27*(12). https://doi.org/10.1142/S0217979213500525

Smith, C., & Sotala, K. (2011). Knowledge , networks and nations Global scientific collaboration in the 21st century. In *Networks: Vol. 03/11* (Issue RS Policy document 03/11). http://royalsociety.org/uploadedFiles/Royal_Society_Content/Influencing_Policy/Reports/2011-03-28-Knowledge-networks-nations.pdf

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*, 643–681. https://doi.org/10.1002/aris.2007.1440410121

Srilatha, P., & Manjula, R. (n.d.). *Similarity Index based Link Prediction Algorithms in Social Networks: A Survey*.

Tabassum, S., Pereira, F. S. F., Fernandes, S., & Gama, J. (2018). Social network analysis: An overview. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 8, Issue 5). Wiley-Blackwell. https://doi.org/10.1002/widm.1256

Wallace, M. L., Larivière, V., & Gingras, Y. (2012). A small world of citations? the influence of collaboration networks on citation practices. *PLoS ONE*, *7*(3), 1–10. https://doi.org/10.1371/journal.pone.0033339

Wang, H., & Le, Z. (2020). Seven-layer model in complex networks link prediction: A survey. *Sensors (Switzerland)*, *20*(22), 1–33. https://doi.org/10.3390/s20226560

Wang, P., Xu, B. W., Wu, Y. R., & Zhou, X. Y. (2015). Link prediction in social networks: the state-of-the-art. In *Science China Information Sciences* (Vol. 58, Issue 1, pp. 1–38). Science in China Press. https://doi.org/10.1007/s11432-014-5237-y

Wu, H., Wang, S., & Fang, H. (2022). *LP-UIT: A Multimodal Framework for Link Prediction in Social Networks*. http://arxiv.org/abs/2201.10108

Yuliansyah, H., Othman, Z. A., & Bakar, A. A. (2020). Taxonomy of link

prediction for social network analysis: A review. *IEEE Access*, *8*(1), 183470–183487. https://doi.org/10.1109/ACCESS.2020.3029122

Zhang, Jianpei and Zhang, Yuan and Yang, Hailu and Yang, J. (2014). A link prediction algorithm based on socialized semi-local information. *Journal of Computational Information Systems*, *10*(10), 4459--4466. https://doi.org/10.12733/jcis10454

Zhang, Q., Tong, T., & Wu, S. (2020). Hybrid link prediction via model averaging. *Physica A: Statistical Mechanics and Its Applications*, *556*. https://doi.org/10.1016/j.physa.2020.124772

Zhang, W., & Wu, B. (2014). Accurate and fast link prediction in complex networks. *2014 10th International Conference on Natural Computation, ICNC 2014*, 653–657. https://doi.org/10.1109/ICNC.2014.6975913

Zhang, Y., Li, F., Xu, B., Gao, K., & Yu, G. (2012). Using non-topological node attributes to improve results of link prediction in social networks. *Proceedings - 9th Web Information Systems and Applications Conference, WISA 2012*, 141–146. https://doi.org/10.1109/WISA.2012.21

Zhao, H., Du, L., & Buntine, W. (2017). *Leveraging Node Attributes for Incomplete Relational Data*. https://github.com/

Zhou, K., Michalak, T. P., Waniek, M., Rahwan, T., & Vorobeychik, Y. (2019). Attacking similarity-based link prediction in social networks. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *1*, 305–313.

Zhou, T., Lü, L., & Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B*, *71*(4), 623–630. https://doi.org/10.1140/epjb/e2009-00335-8

Zhu, J., Xie, Q., & Chin, E. J. (2012). A hybrid time-series link prediction framework for large social network. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7447 LNCS*(PART 2), 345–359. https://doi.org/10.1007/978-3-642-32597-7_30

Abdul, M., & Mastan, N. (2013). *REVIEW A Survey on LDA Approach in Predicting Link Behavior in Social Networks*. *2*(3), 176–180.

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, *25*(3), 211–230. https://doi.org/10.1016/S0378-8733(03)00009-1

Aghabozorgi, F., & Khayyambashi, M. R. (2018). A new similarity measure for link prediction based on local structures in social networks. *Physica A: Statistical Mechanics and Its Applications*, *501*, 12–23. https://doi.org/10.1016/j.physa.2018.02.010

Ahmed, N. M., Chen, L., Wang, Y., Li, B., Li, Y., & Liu, W. (2016). Sampling-based algorithm for link prediction in temporal networks. *Information Sciences*, *374*, 1–14. https://doi.org/10.1016/j.ins.2016.09.029

Alghamdi, R., & Alfalqi, K. (2015). A Surv e y o f Topic Mode ling i n Text Mining. *Nternational Journal of Advanced Computer Science and Applications*, *6*(1), 7.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. http://arxiv.org/abs/1707.02919

Antunes, J. B., Antunes, J. B., Filho, H. F. B. P., Maia, R. D., De Queiroz, R. B., Da Silva, C. M. R., Rodrigues, R. B., & De Almeida Barros, F. (2013). ConPredict: A method for link prediction in co-authored content-based networks. *Proceedings of the IADIS International Conference WWW/Internet 2013, ICWI 2013*, *January*, 11–18.

Assouli, N., Benahmed, K., & Gasbaoui, B. (2021). How to predict crime — informatics-inspired approach from link prediction. *Physica A: Statistical Mechanics and Its Applications*, *570*. https://doi.org/10.1016/j.physa.2021.125795

Bahabadi, M. D., Golpayegani, A. H., & Esmaeili, L. (2014). *A Novel C2C E-Commerce Recommender System Based on Link Prediction: Applying Social Network Analysis*.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

Bartal, A., Sasson, E., & Ravid, G. (2009). Predicting links in social networks using text mining and SNA. *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2009*, 131–136. https://doi.org/10.1109/ASONAM.2009.12

Berlusconi, G., Calderoni, F., Parolini, N., Verani, M., & Piccardi, C. (2016). Link prediction in criminal networks: A tool for criminal intelligence analysis. *PLoS ONE*, *11*(4). https://doi.org/10.1371/journal.pone.0154244

Bhattacharyya, P., Garg, A., & Wu, S. F. (2011). Analysis of user keyword similarity in online social networks. *Social Network Analysis and Mining*, *1*(3), 143–158. https://doi.org/10.1007/s13278-010-0006-4

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17–35. https://doi.org/10.1214/07-aoas114

Börner, K., Maru, J. T., & Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(SUPPL. 1), 5266–5273. https://doi.org/10.1073/pnas.0307625100

Chuan, P. M., Son, L. H., Ali, M., Khang, T. D., Huong, L. T., & Dey, N. (2018). Link prediction in co-authorship networks based on hybrid content similarity metric. *Applied Intelligence*, *48*(8), 2470–2486. https://doi.org/10.1007/s10489-017-1086-x

Coskun, M., & Koyuturk, M. (2016). Link Prediction in Large Networks by Comparing the Global View of Nodes in the Network. *Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015*, 485–492. https://doi.org/10.1109/ICDMW.2015.195

Crichton, G., Guo, Y., Pyysalo, S., & Korhonen, A. (2018). Neural networks for link prediction in realistic biomedical graphs: A multi-dimensional evaluation of graph embedding-based approaches. *BMC Bioinformatics*, *19*(1), 1–11. https://doi.org/10.1186/s12859-018-2163-9

Daud, N. N., Ab Hamid, S. H., Saadoon, M., Sahran, F., & Anuar, N. B. (2020). Applications of link prediction in social networks: A review. In *Journal of Network and Computer Applications* (Vol. 166). Academic Press. https://doi.org/10.1016/j.jnca.2020.102716

Davisu, J., & Goadrich, M. (2016). *The relationship between precision-recall and ROC curves*. 233–240.

De Tre, G., Hallez, A., & Bronselaer, A. (2014). Performance optimization of object comparison. *International Journal of Intelligent Systems*, *29*(2), 495–524. https://doi.org/10.1002/int

Dong, L., Li, Y., Yin, H., Le, H., & Rui, M. (2013). The algorithm of link prediction on social network. *Mathematical Problems in Engineering*, *2013*. https://doi.org/10.1155/2013/125123

Dong, Y., Ke, Q., Wang, B., & Wu, B. (2011). Link prediction based on local information. *Proceedings - 2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011*, 382–386. https://doi.org/10.1109/ASONAM.2011.43

Drummond, C., & Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, *65*(1), 95–130. https://doi.org/10.1007/s10994-006-8199-5

E Fonseca, B. de P. F., Sampaio, R. B., Fonseca, M. V. de A., & Zicker, F. (2016). Co-authorship network analysis in health research: Method and potential use. In *Health Research Policy and Systems* (Vol. 14, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12961-016-0104-5

F Shahrabi Farahani, M Alavi, M Ghasem, Bt. (2020). Scientific Map of Papers Related to Data Mining in Civilica Database Based on Co-Word Analysis. *International Journal of Web Research*, *3*(1), 11–18.

Fu, C., Zhao, M., Fan, L., Chen, X., Chen, J., Wu, Z., Xia, Y., & Xuan, Q. (2018). Link Weight Prediction Using Supervised Learning Methods and Its Application to Yelp Layered Network. *IEEE Transactions on Knowledge and Data Engineering*, *30*(8), 1507–1518. https://doi.org/10.1109/TKDE.2018.2801854

Gao, F., Musial, K., Cooper, C., & Tsoka, S. (2015). Link prediction methods and their accuracy for different social networks and network metrics. *Scientific Programming*, *2015*(i). https://doi.org/10.1155/2015/172879

Ghorbanzadeh, H., Sheikhahmadi, A., Jalili, M., & Sulaimany, S. (2021a). A hybrid method of link prediction in directed graphs. *Expert Systems with Applications*, *165*(February 2020), 113896. https://doi.org/10.1016/j.eswa.2020.113896

Ghorbanzadeh, H., Sheikhahmadi, A., Jalili, M., & Sulaimany, S.

(2021b). A hybrid method of link prediction in directed graphs. *Expert Systems with Applications*, *165*. https://doi.org/10.1016/j.eswa.2020.113896

Haghani, S., & Keyvanpour, M. R. (2019). A systemic analysis of link prediction in social network. *Artificial Intelligence Review*, *52*(3), 1961–1995. https://doi.org/10.1007/s10462-017-9590-2

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hasan, M. Al, Chaoji, V., Salem, S., Zaki, M., & York, N. (n.d.). *Link Prediction using Supervised Learning*.

Hassan, D. (n.d.). *SUPERVISED LINK PREDICTION IN CO-AUTHORSHIP NETWORKS BASED ON RESEARCH PERFORMANCE AND SIMILARITY OF RESEARCH INTERESTS AND AFFILIATIONS*.

Hemkiran, S., & Sudha Sadasivam, G. (2020). A review of similarity measures and link prediction models in social networks. *International Journal of Computing and Digital Systems*, *9*(2), 239–248. https://doi.org/10.12785/IJCDS/090209

Ibrahim, N. M. A., & Chen, L. (2015). Link prediction in dynamic social networks by integrating different types of information. *Applied Intelligence*, *42*(4), 738–750. https://doi.org/10.1007/s10489-014-0631-0

Jaccard, P. (1982). Etude de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de La Murithienne*, *XXXVII*, 81-92. https://doi.org/10.5169/seals-266450

Jaya Lakshmi, T., & Durga Bhavani, S. (2017). Link prediction in temporal heterogeneous networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10241 LNCS*, 83–98. https://doi.org/10.1007/978-3-319-57463-9_6

Kong, X., Shi, Y., Yu, S., Liu, J., & Xia, F. (2019). Journal of Network and Computer Applications Academic social networks : Modeling , analysis , mining and applications. *Journal of Network and Computer Applications*, *132*(December 2018), 86–103. https://doi.org/10.1016/j.jnca.2019.01.029

Kumar, A., Mishra, S., Singh, S. S., Singh, K., & Biswas, B. (2020). Link prediction in complex networks based on Significance of Higher-Order Path Index (SHOPI). *Physica A: Statistical Mechanics and Its Applications*, *545*, 123790. https://doi.org/10.1016/j.physa.2019.123790

Kumari, A., Behera, R. K., Sahoo, K. S., Nayyar, A., Kumar Luhach, A., & Prakash Sahoo, S. (2020). Supervised link prediction using structured-based feature extraction in social network. *Concurrency Computation , February*, 1–16. https://doi.org/10.1002/cpe.5839

Kushwah, A. K. S., & Manjhvar, A. K. (2016). A review on link prediction in social network. *International Journal of Grid and Distributed Computing*, *9*(2), 43–50. https://doi.org/10.14257/ijgdc.2016.9.2.05

Leicht, E. A., Holme, P., & Newman, M. E. J. (2006). Vertex similarity in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *73*(2), 1–10. https://doi.org/10.1103/PhysRevE.73.026120

Li, L., Wang, L., Luo, H., & Chen, X. (2021). Towards effective link prediction: A hybrid similarity model. *Journal of Intelligent and Fuzzy Systems*, *40*(3), 4013–4026. https://doi.org/10.3233/JIFS-200344

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019–1031. https://doi.org/10.1002/asi.20591

Lichtenwalter, R., & Chawla, N. V. (2012). Link prediction: Fair and effective evaluation. *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, 376–383. https://doi.org/10.1109/ASONAM.2012.68

Lichtenwalter, R. N., & Chawla, N. V. (2012). *Vertex collocation profiles*. *1019*, 1019–1028. https://doi.org/10.1145/2187836.2187973

Lichtenwalter, R. N., Lussier, J. T., & Chawla, N. V. (2010). New perspectives and methods in link prediction. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 243–252. https://doi.org/10.1145/1835804.1835837

Linyuan, L. L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications*, *390*(6), 1150–1170. https://doi.org/10.1016/j.physa.2010.11.027

Liu, H., Kou, H., Yan, C., & Qi, L. (2019). Link prediction in paper citation network to construct paper correlation graph. *Eurasip Journal on Wireless Communications and Networking*, *2019*(1). https://doi.org/10.1186/s13638-019-1561-7

Liu, J. H., Zhu, Y. X., & Zhou, T. (2016). Improving personalized link prediction by hybrid diffusion. *Physica A: Statistical Mechanics and Its Applications*, *447*, 199–207. https://doi.org/10.1016/j.physa.2015.12.036

Liu, S., Ji, X., Liu, C., & Bai, Y. (2017). Extended resource allocation index for link prediction of complex network. *Physica A: Statistical Mechanics and Its Applications*, *479*, 174–183. https://doi.org/10.1016/j.physa.2017.02.078

Liu, X., Zhang, J., & Guo, C. (2013). Full-text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology*, *64*(9), 1852–1863. https://doi.org/10.1002/asi.22883

Ma, G., Yan, H., Qian, Y., Wang, L., Dang, C., & Zhao, Z. (2021). Path-based estimation for link prediction. *International Journal of Machine Learning and Cybernetics*, *12*(9), 2443–2458. https://doi.org/10.1007/s13042-021-01312-w

Martin, T., Ball, B., Karrer, B., & Newman, M. E. J. (2013). Coauthorship and citation patterns in the Physical Review. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *88*(1), 1–9. https://doi.org/10.1103/PhysRevE.88.012814

Martínez, V., Berzal, F., & Cubero, J.-C. (2017). A Survey of Link Prediction in Complex Networks. *ACM Computing Surveys*, *49*(4), 1–33. https://doi.org/10.1145/3012704

Martínez, V., Berzal, F., & Cubero, J. C. (2016). A survey of link prediction in complex networks. *ACM Computing Surveys*, *49*(4). https://doi.org/10.1145/3012704

Mishra, S., & Nandi, G. C. (2015). A novel hybrid approach for link prediction problem in social network. *International Journal of Social Network Mining*, *2*(2), 115. https://doi.org/10.1504/ijsnm.2015.072281

Mohammad Al Hasan, Zaki, M. J. (2011). A SURVEY OF LINK PREDICTION IN SOCIAL NETWORKS. In *Social Network Data Analytics*. https://doi.org/10.1007/978-1-4419-8462-3

Muniz, C. P., Goldschmidt, R., & Choren, R. (2018). Combining contextual, temporal and topological information for unsupervised link prediction in social networks. *Knowledge-Based Systems*, *156*, 129–137. https://doi.org/10.1016/j.knosys.2018.05.027

Mutlu, E. C., & Oghaz, T. (2020). Review on Graph Feature Learning and Feature Extraction Techniques for Link Prediction. *Machine Learning and Knowledge Extraction*, *2*(4), 672–704. https://doi.org/10.3390/make2040036

Mutlu, E. C., Oghaz, T., Rajabi, A., & Garibay, I. (2020). Review on Learning and Extracting Graph Features for Link Prediction. *Machine Learning and Knowledge Extraction*, *2*(4), 672–704. https://doi.org/10.3390/make2040036

Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *64*(2), 4. https://doi.org/10.1103/PhysRevE.64.025102

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*(2), 167–256. https://doi.org/10.1137/S003614450342480

Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(SUPPL. 1), 5200–5205. https://doi.org/10.1073/pnas.0307545100

Papadimitriou, A., Symeonidis, P., & Manolopoulos, Y. (2012). Scalable link prediction in social networks based on local graph characteristics. *Proceedings of the 9th International Conference on Information Technology, ITNG 2012*, 738–743. https://doi.org/10.1109/ITNG.2012.145

PARIMI, R. (2010). *LDA BASED APPROACH FOR PREDICTING FRIENDSHIP LINKS IN LIVE Copyright Rohit Parimi*.

Parimi, R., & Caragea, D. (2011). Predicting friendship links in social networks using a topic modeling approach. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6635 LNAI*(PART 2), 75–86. https://doi.org/10.1007/978-3-

642-20847-8_7

Quercia, D., Askham, H., & Crowcroft, J. (2012). TweetLDA: Supervised topic classification and link prediction in Twitter. *Proceedings of the 4th Annual ACM Web Science Conference, WebSci'12*, *volume*, 247–250. https://doi.org/10.1145/2380718.2380750

Raut, P., Khandelwal, H., & Vyas, G. (2020). A Comparative Study of Classification Algorithms for Link Prediction. *2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020 - Conference Proceedings*, *Icimia*, 479–483. https://doi.org/10.1109/ICIMIA48430.2020.9074840

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A.-L. (2002). Hierarchical Organization of Modularity in Metabolic Networks. *Science*, *297*(5586), 1551–1555. https://doi.org/10.1126/science.1073374

Sachan, M., & Ichise, R. (2010). Using Semantic Information to Improve Link Prediction Results in Network Datasets. *International Journal of Engineering and Technology*, *2*(4), 334–339. https://doi.org/10.7763/ijet.2010.v2.143

Samad, A., Qadir, M., & Nawaz, I. (2019). SAM: A Similarity Measure for Link Prediction in Social Network. *MACS 2019 - 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics, Proceedings*. https://doi.org/10.1109/MACS48846.2019.9024762

Samad, A., Qadir, M., Nawaz, I., Islam, M., & Aleem, M. (2020). A Comprehensive Survey of Link Prediction Techniques for Social Network. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, *7*(23), 163988. https://doi.org/10.4108/eai.13-7-2018.163988

Shao, C. X., Dou, H. L., Yang, R. X., & Wang, B. H. (2013). Zero nodes effect: Valid link prediction in sparse networks. *International Journal of Modern Physics B*, *27*(12). https://doi.org/10.1142/S0217979213500525

Smith, C., & Sotala, K. (2011). Knowledge , networks and nations Global scientific collaboration in the 21st century. In *Networks: Vol. 03/11* (Issue RS Policy document 03/11). http://royalsociety.org/uploadedFiles/Royal_Society_Content/Influencing_Policy/Reports/2011-03-28-Knowledge-networks-nations.pdf

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*, 643–681. https://doi.org/10.1002/aris.2007.1440410121

Srilatha, P., & Manjula, R. (n.d.). *Similarity Index based Link Prediction Algorithms in Social Networks: A Survey*.

Tabassum, S., Pereira, F. S. F., Fernandes, S., & Gama, J. (2018). Social network analysis: An overview. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 8, Issue 5). Wiley-Blackwell. https://doi.org/10.1002/widm.1256

Wallace, M. L., Larivière, V., & Gingras, Y. (2012). A small world of citations? the influence of collaboration networks on citation practices. *PLoS ONE*, *7*(3), 1–10. https://doi.org/10.1371/journal.pone.0033339

Wang, H., & Le, Z. (2020). Seven-layer model in complex networks link prediction: A survey. *Sensors (Switzerland)*, *20*(22), 1–33. https://doi.org/10.3390/s20226560

Wang, P., Xu, B. W., Wu, Y. R., & Zhou, X. Y. (2015). Link prediction in social networks: the state-of-the-art. In *Science China Information Sciences* (Vol. 58, Issue 1, pp. 1–38). Science in China Press. https://doi.org/10.1007/s11432-014-5237-y

Wu, H., Wang, S., & Fang, H. (2022). *LP-UIT: A Multimodal Framework for Link Prediction in Social Networks*. http://arxiv.org/abs/2201.10108

Yuliansyah, H., Othman, Z. A., & Bakar, A. A. (2020). Taxonomy of link prediction for social network analysis: A review. *IEEE Access*, *8*(1), 183470–183487. https://doi.org/10.1109/ACCESS.2020.3029122

Zhang, Jianpei and Zhang, Yuan and Yang, Hailu and Yang, J. (2014). A link prediction algorithm based on socialized semi-local information. *Journal of Computational Information Systems*, *10*(10), 4459--4466. https://doi.org/10.12733/jcis10454

Zhang, Q., Tong, T., & Wu, S. (2020). Hybrid link prediction via model averaging. *Physica A: Statistical Mechanics and Its Applications*, *556*. https://doi.org/10.1016/j.physa.2020.124772

Zhang, W., & Wu, B. (2014). Accurate and fast link prediction in complex networks. *2014 10th International Conference on Natural Computation, ICNC 2014*, 653–657. https://doi.org/10.1109/ICNC.2014.6975913

Zhang, Y., Li, F., Xu, B., Gao, K., & Yu, G. (2012). Using non-topological node attributes to improve results of link prediction in social networks. *Proceedings - 9th Web Information Systems and Applications Conference, WISA 2012*, 141–146. https://doi.org/10.1109/WISA.2012.21

Zhao, H., Du, L., & Buntine, W. (2017). *Leveraging Node Attributes for Incomplete Relational Data*. https://github.com/

Zhou, K., Michalak, T. P., Waniek, M., Rahwan, T., & Vorobeychik, Y. (2019). Attacking similarity-based link prediction in social networks. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, *1*, 305–313.

Zhou, T., Lü, L., & Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B*, *71*(4), 623–630. https://doi.org/10.1140/epjb/e2009-00335-8

Zhu, J., Xie, Q., & Chin, E. J. (2012). A hybrid time-series link prediction framework for large social network. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7447 LNCS*(PART 2), 345–359. https://doi.org/10.1007/978-3-642-32597-7_30.