Link Prediction in Bipartite Network Using Composite Similarities

Bijay Gaudel[1], Deepanjal Shrestha[1,2], Niosh Basnet[1], Neesha Rajkarnikar[2], Seung Ryul Jeong[3*] and Donghai Guan[1]

[1]Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2]School of Business, Pokhara University, Pokhara, Nepal
[3]Graduate School of Business IT, Kookmin University, Seoul, South Korea
[e-mail: gaudelbijay10@gmail.com, deepanjal@hotmail.com, chandralal.basnet@gmail.com, neesha_rk@hotmail.com, srjeong@kookmin.ac.kr, dhguan@nuaa.edu.cn]
*Corresponding author: Seung Ryul Jeong

Received May 23, 2023; revised July 11, 2023; accepted July 18, 2023; published August 31, 2023

Abstract

Analysis of a bipartite (two-mode) network is a significant research area to understand the formation of social communities, economic systems, drug side effect topology, etc. in complex information systems. Most of the previous works talk about a projection-based model or latent feature model, which predicts the link based on singular similarity. The projection-based models suffer from the loss of structural information in the projected network and the latent feature is hardly present. This work proposes a novel method for link prediction in the bipartite network based on an ensemble of composite similarities, overcoming the issues of model-based and latent feature models. The proposed method analyzes the structure, neighborhood nodes as well as latent attributes between the nodes to predict the link in the network. To illustrate the proposed method, experiments are performed with five real-world data sets and compared with various state-of-art link prediction methods and it is inferred that this method outperforms with ~3% to ~9% higher using area under the precision-recall curve (AUC-PR) measure. This work holds great significance in the study of biological networks, e-commerce networks, complex web-based systems, networks of drug binding, enzyme protein, and other related networks in understanding the formation of such complex networks. Further, this study helps in link prediction and its usability for different purposes ranging from building intelligent systems to providing services in big data and web-based systems.

Keywords: Bipartite network; ensemble; link prediction; potential link

This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education and National Research Foundation of Korea.

http://doi.org/10.3837/tiis.2023.08.004 **ISSN : 1976-7277**

1. Introduction

Network analysis is a significant research area in modern network-based systems, enabling the prediction of relationships between interconnected nodes based on various parameters. Many of these networks exhibit a bipartite structure [1], characterized by two distinct types of nodes organized into separate clusters connected through links. Bipartite networks are prevalent in diverse domains, including scientific collaborations [2], human sexual interactions [3], and metabolic processes [4]. Analyzing complex bipartite networks for link prediction is an exciting and challenging field of study, as it allows for the prediction of future links and the understanding of unknown interactions, which holds vital importance across multiple application domains. Examples of such domains include e-commerce networks [5, 6], biological networks [7, 8], social networks [9–11], and drug side effect networks [12].

Link prediction in bipartite networks employs various approaches, such as the projection of bipartite networks into unipartite networks and the application of local, quasi-local, and global similarity methods [13]. Local similarity methods primarily rely on the local contact structure and utilize the formation of triangle closing as a key principle. Global similarity indices utilize global topological information, considering the shortest path measures between nodes and assigning less weight to longer paths. However, the computation complexity of these methods may render them unfeasible for large networks. Quasi-local methods have emerged as a balanced solution, leveraging topological information as a global approach, thereby efficiently calculating associations and predicting links. In 2007, Liben Nowell conducted a comprehensive analysis of link prediction [14], providing a foundation for subsequent research in this field. Extensive research has been dedicated to link prediction in bipartite networks [15–24], employing mechanism-based models [15–24] and latent feature models [25–29]. However, latent feature models often overlook the presence of latent groups, potentially omitting crucial links within the network. Mechanism-based models can be further categorized as projection-based models [22, 23], which sacrifice structural information, and local community paradigm (LCP)-based models [18, 19]. Projection-based models are not efficient for link prediction in bipartite networks as they lose structural information and local community paradigm.

To address the limitations of mechanism-based and latent feature-based models, this paper proposes a novel approach based on a composite similarity measure. The proposed method initially projects the bipartite network into a unipartite network based on a predefined threshold, extracting strong links (potential links) by eliminating redundant connections. Moreover, the proposed model incorporates two structural features: the number of butterfly enclosures formed by potential links and the average number of patterns between potential links. The number of butterfly enclosures counts the different ways information can travel between nodes within potential links, while the average number of patterns indicates the average number of nodes needed to connect one node to another in the projected network. The proposed method also extracts two additional features: the reciprocal of the sum of neighbors of the butterfly's enclosure by potential link and internal link. The former suggests that increased restriction in the information travel path raises the likelihood of link formation between nodes, while the latter classifies links into two types based on the presence of latent features. Subsequently, these extracted features are utilized as input for a classifier algorithm to predict links in the network. One potential approach is to employ a composite similarity measure that combines multiple similarity metrics, such as cosine similarity, Euclidean distance, and Jacquard similarity. This approach incorporates both mechanism-based and

latent feature-based similarities, aiming to enhance model accuracy. Additionally, techniques like feature selection, dimensionality reduction, and ensemble learning can be integrated to further improve model performance. By introducing a novel approach based on a composite similarity measure, this research aims to enhance the performance of mechanism-based and latent feature-based models by considering semantic similarity, structural similarity, and behavioral similarity to determine entity similarity. The composite similarity measure offers a comprehensive and accurate evaluation of similarity, leading to improved accuracy and robustness in addressing problems in mechanism-based and latent feature-based models.

1.1 Problem Definition

The core problem addressed in the above scenario is link prediction in complex bipartite networks. Bipartite networks consist of two distinct types of nodes organized into separate clusters connected through links. The objective is to predict future links or interactions between the nodes based on various parameters. The problem is challenging yet significant, as it allows for the understanding of unknown interactions and has applications in various domains such as e-commerce networks, biological networks, social networks, and drug side effect networks. The proposed approach based on a composite similarity measure aims to overcome the limitations of mechanism-based and latent feature-based models in link prediction. It combines multiple similarity metrics, considers both mechanism-based and latent feature-based similarities, and incorporates structural and additional features to enhance accuracy. By utilizing a classifier algorithm trained on known links and leveraging the composite similarity measure, the approach strives to predict the likelihood of link formation in the bipartite network accurately. The objective is to provide a more comprehensive and accurate evaluation of similarity, leading to improved accuracy and robustness in addressing link prediction problems in bipartite networks.

1.2 Conceptual Theoratical Framework

In this work, we first acknowledge the significance of link prediction in bipartite networks, which consist of distinct nodes organized into separate clusters connected by links and the wide-ranging applications of link prediction in these networks across various domains. We then explore different approaches used in link prediction, including the projection of bipartite networks into unipartite networks and the application of local, quasi-local, and global similarity methods. However, its is observed that there are limitations of existing mechanism-based and latent feature-based models, such as their oversight of latent groups and the loss of structural information. To overcome these limitations, we propose a novel approach based on a composite similarity measure. The work involves a systematic approach to link prediction in bipartite networks. This approach begins by projecting the bipartite network into a unipartite network, thereby simplifying the analysis. The next step involves extracting structural features, including butterfly enclosures and patterns, which capture important information about the network's connectivity. These features, along with additional measures, are then used as inputs to a classifier algorithm for link prediction. In order to determine entity similarity, the approach emphasizes the consideration of semantic similarity, structural similarity, and behavioral similarity. This is achieved by utilizing a composite similarity measure that combines multiple similarity metrics. The ultimate goal of the proposed approach is to enhance the performance of link prediction models by improving accuracy and providing a comprehensive evaluation of similarity.

2. Related Works

Link prediction in bipartite networks is a prominent research area within network dynamics analysis. Various approaches have been proposed to address this challenge, encompassing mechanism-based models [5][15][18,19][21–24], latent feature models [25–29], embedding methods [30–33], and community detection mechanisms [34]. Additionally, research has explored different network formations, including the integration of crowd-users rating information in social media platforms [35], trust-based selection network models of web services [36], and trajectory distance algorithms based on segment transformation distance [37]. Mechanism-based models can be further categorized as projection-based mechanisms [5, 22, 23], LCP mechanisms [18, 19], PA mechanisms [15], and homophily mechanisms [21, 24]. Projection-based mechanisms involve projecting the bipartite network into a unipartite network and utilizing various similarity measures to predict links. LCP mechanisms emphasize the common neighbor index and local community structure in link formation. PA mechanisms are based on the preferential attachment principle, where higher degree nodes easily connect to other nodes. Homophily mechanisms operate based on the concept of triangular enclosures in unipartite networks transformed into quadrangular enclosures. Latent feature-based models focus on the latent groups present in the network. Community detection-based models aim to identify the number of communities but are often limited by the number or types of networks, commonly referred to as layers. Embedding methods involve converting the network into a lower-dimensional vector space, facilitating similarity search and supporting machine learning through low-dimensional representations. However, a drawback of embedding methods is the lack of interpretability.

While the previous description highlights various approaches in the field of network dynamics, it is important to address the underlying needs and understanding of new network dynamics proposed in this study. To bridge this gap, our paper proposes a novel approach that goes beyond the existing techniques mentioned above. By considering the limitations and challenges associated with cluster analysis, we aim to overcome these drawbacks and introduce a method that effectively identifies strong potential links in bipartite networks. Cluster analysis is a widely used technique for discovering patterns and structures in large datasets [1]. However, it is sensitive to initial conditions and the choice of clustering algorithm [2]. This sensitivity can lead to different results and hinder replicability and validation. Moreover, the computational complexity of cluster analysis increases with larger datasets, limiting its scalability [3]. Overfitting is also a concern, where clusters may not generalize well to new data if the number of clusters is not appropriately chosen [4].

To address these limitations, our proposed method introduces new network dynamics that overcome the challenges of cluster analysis. By focusing on the underlying needs and understanding of bipartite networks, we provide a comprehensive approach to identify strong potential links. Through a detailed analysis and consideration of network dynamics, our method aims to enhance link prediction in bipartite networks, offering improved accuracy and scalability compared to existing techniques. By addressing the root needs of network dynamics and overcoming the limitations of traditional cluster analysis, our proposed approach contributes to advancing the field and providing meaningful insights into bipartite network dynamics.

2.1 Bipartite Network Projection

The bipartite graph can be represented as G = (U, V, E), where, U and V are two different types of sets of nodes and E is the edge or link between the element of U and V: $E \subseteq U \times V$.

The bipartite graph can be projected into three types of unipartite networks, which consist of unweighted, weighted, and strengthening projections. The unweighted U projection of G is $G_u = (U, E_u)$, in which $(a,b) \in E_u$ if a and b have at least one common neighbor in G i.e $\gamma(a) \cap \gamma(b) \neq \varphi$, then E_u can be written as: $E_u = \{(a,b)|a,b \in U, \exists p \in V, p \in \gamma(a) \cap \gamma(b)\}$. The unweighted V-projection is defined dually. The unweighted U and V projection of the bipartite network shown in **Fig. 1** is extended in **Fig. 2a**, and **Fig. 2b** respectively.

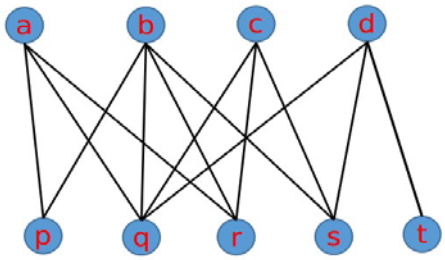
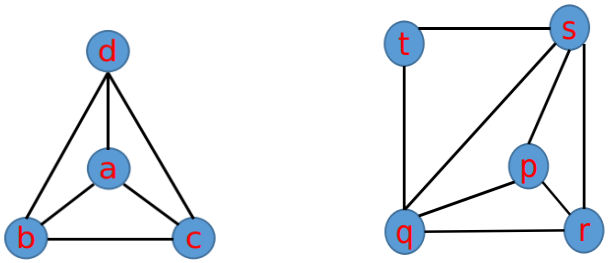
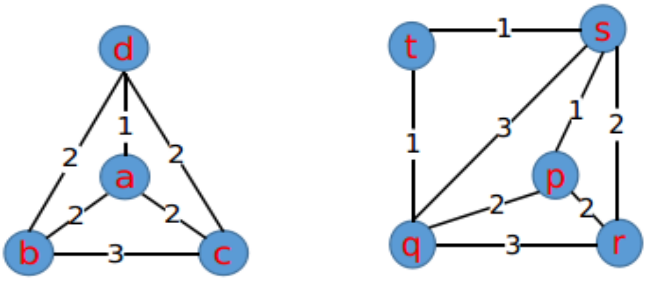


Fig. 1. Dummy bipartite network G



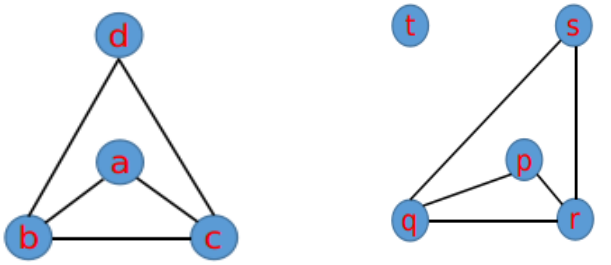
(a (b

Fig. 2. (a) Unweighted U-projection and (b) Unweighted V-projection of bipartite network **Fig. 1**.



(a) (b)

Fig. 3. (a) Weighted U-projection and (b) Weighted V-projection of bipartite network **Fig. 1**.



(a) (b)

Fig. 4. (a) Strengthening U-projection of the bipartite network **Fig. 1** when, T=1, (b) strengthening V-projection of the bipartite network **Fig. 1** when, T=1

In general, the unweighted network projections work, if this network has at least one common neighbor in G, irrespective of the number of common neighbors. This has a problem that other vital connected neighbors maybe be discarded. To overcome this problem, weighted projection networks were proposed [38]. Projections of unipartite networks are projected with the weight function W. For G_u projection network $W(u_i, v_j)$ is $|\gamma(u) \cap \gamma(v)|$ in bipartite network G. The V -weighted projection of network G is defined dually. The weighted U and V-projections of the bipartite network G in **Fig. 1** are described in **Fig. 3a**, and **Fig. 3b** respectively. Again, there are so many weakly connected ties that contain redundant information as much as strong links that include essential information. To extract these strong links, strengthening the projection of a bipartite networks, is proposed [39]. The projection network here is strengthened with a predetermined threshold value T. The edge whose weight $|W(u_i, v_j)| > T$ are kept, and others are removed in the U and V part of the network in **Fig. 1**, which is projected in **Fig. 4a**, and **Fig. 4b** respectively.

3. Proposed Model

Consider a bipartite network G = (U, E, V), where U and V are two different types of sets of nodes and E is the edge or link between the element of U and V such that; $E \subseteq U \times V$. The information in the bipartite network can be represented as $M \in \{0,1\}m \times n(m=|U|, n=|V|)$. Where,

$$Mij = \{^{1}_{0} if i \in U \text{ and } j \in V, \text{ are connected otherwise} \quad (1)$$

The formation of a link in the bipartite network is not solely dependent on the neighborhood of the nodes. It needs to analyze the neighborhood as well as other factors like the formation of the pattern and the way to reach one node from another, etc. The possibility of connecting all the nodes is also very low and calculating the link prediction score for all the possible links is computationally expensive. Therefore, this work proposes an efficient link prediction method to deal with bipartite networks, which removes some less important links based on the projection of the network and keeps only potential links that tend to exist in the future. Based on the potential link, information is extracted and fed to the link predictor to predict the formation of the link.

Table 1. Mathematical symbols used in this paper

Symbols	Meaning
G	Bipartite Network
G_u, G_v	U and V are projections of network G
T	Weight threshold
G_u^T, G_v^T	U and V strengthening projections of the network G by weight threshold T
PL	Potential link
$a, b, c,\ldots$	Nodes in part U part of Network G
$p, q, r,\ldots$	Nodes in the V part of the network
$\gamma(x)$	Neighbor of node x in the network G
$\gamma_u^T(x)$ $\gamma_v^T(x)$	Neighbor of node x in the network G_u^T and G_v^T respectively
∞	Butterfly enclosure by PL

The architecture of the proposed model is given in **Fig. 5**. The proposed model has three functional modules: potential link extractor, feature extractor, and link predictor. The potential link extractor module extracts some strong links by analyzing the structural property of the graph with the help of the projected network. Feature extractor module extracts four features: The number of butterflies (∞) enclosures between two nodes (∞ (i,j)i$\in$U and j$\in$V), the average number of patterns between the potential link (avg{N_u(PT), N_v(PT)}), reciprocal of the sum of the neighbor of butterfly nodes (RSNN), and internal link (IL). The link predictor module uses an ensemble classifier to classify the link to 1 or 0 based on the formed and unformed link respectively. The details of these three parts are as follows:

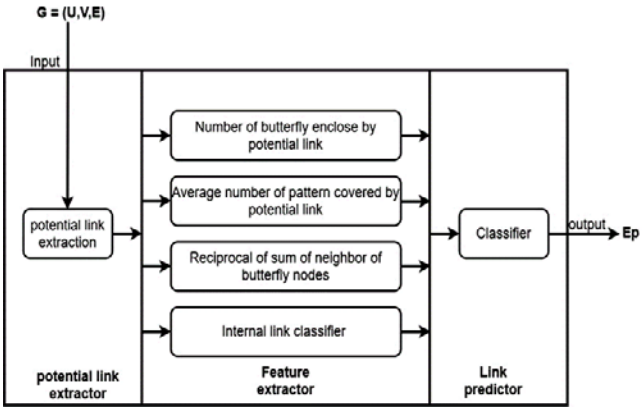


Fig. 5. Architecture of the proposed bipartite link prediction model

3.1 Potential Link Extractor

A network has many weak links that contain redundant information and strong links that contain important information. The weak links are the ones that may not have a probability to exist in the future and only the strong links are vital in these formations. The potential link extractor extracts these strong links by filtering weak links. For bipartite network G = (U, V, E), where U-part is strengthening unipartite projection network) G_uT= (U, E_uT) with threshold value T and V part strengthening projection network) G_vT = (U, E_vT) with a threshold T.Value of T is determined by the grid search through cross-validation in training sets AUC-PR value. For example, let X_1 belong to U in G_u = (U, E_uT) strengthening projection network shown in **Fig. 4**. $\gamma_u^T(X_1) = \{k_1, k_2, k_3\ldots, k_n\}$ is the neighbor of X_1 in G_uT and $\gamma(X_1)$ is the neighbor of

X_1 in the bipartite network. Potential link for node X_1 based on U-projection of G can be written as: $PL(X_1) = \gamma(X_1) \times \{\gamma(k_1) \cup \gamma(k_2) \cdots \cup \gamma(k_n)\}$. Assume that there are $X_1, X_2, \cdots, X_n$ nodes in G_uT, therefore the total potential links based on the U-projection of G can be written as: $PL = \{PL(X_1) \cup PL(X_2) \cdots \cup PL(X_n)\}$. Based on the strengthening U-projection of the network as in **Fig. 4a**, a set of potential links for node c is extracted as $u^T(c) = \{a,b,d\}$, $\gamma(a) \cup \gamma(b) \cup \gamma(d) = \{p,q,r,s,t\}$ therefore, potential links are $(PL)=\{(c,p),(c,q),(c,r),(c,s),(c,t)\}$. Similarly, the potential link (PL) is extracted based on the V-projection of the graph. Further, the work does feature extraction and link prediction based on these potential links.

Algorithm 1: Potential link extraction algorithm based on U-projection of the graph

Input: G = (U, V, E), U and V are the two types of a vertex set and E is the edge set of the graph.

Parameters: T is a threshold value for network projection, m, and n are the size of node set U and V.

Output: Potential Link; PL ⊆ E; is the set of node pair: $u_i \times v_j$, where, $u_i \in U$ and $v_i \in V$.

```
1   U = {u1,u2,⋯,um}
2   V = {v1,v2 ⋯,vn}
3   EuT ← φ
4   for i = 1 to n do
5       for j = i + 1 to n do
6           if |γ(ui) ∩ γ(uj)| > T then
7               EuT ← EuT ∪ {ui × uj}
8           end
9       end
10  end
11  Gu = (U, Eu )
12  PL ← φ
13  for i = 1 to n do
14      for j = 1 to m do
15          if |γuT (ui) ∩ γ(vi)| ≠φ then
16              PL ← PL ∪ {γuT (ui) × γ(vj)}
17          end
18      end
19  end
```

Algorithm 1 gives the potential link extraction algorithm used in the proposed method. It is based on the U-projection of the given bipartite network, which evaluates the structural property of the node pair and extracts the potential link.

3.2 Feature Extractor

Feature extractor is a mechanism, which extracts the structural features, neighborhood-based features, and latent features based on the potential link.

3.2.1 Number of Butterfly Enclosures by the Potential Link

For any potential link, if the potential link is replaced with the real link nodes, then the number of ∞ patterns formed between these nodes is the number of butterfly enclosures. For a potential link (c,p); as described above in the potential link extractor part; there are five numbers of butterflies enclosures: {∞(c,q, a,p),∞(c,q,b,p),∞(c,r, a,p),∞(c,r,b,p),∞(c,s,b,p)}, (**Fig. 6**).

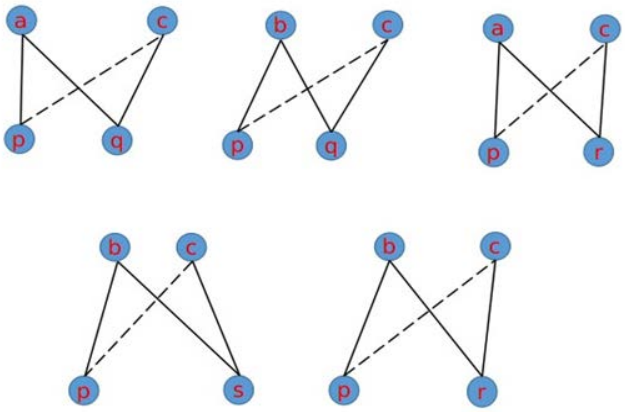


Fig. 6. Number of butterfly encloser for the potential link (c,p)

3.2.2 Pattern Covered by the Potential Link

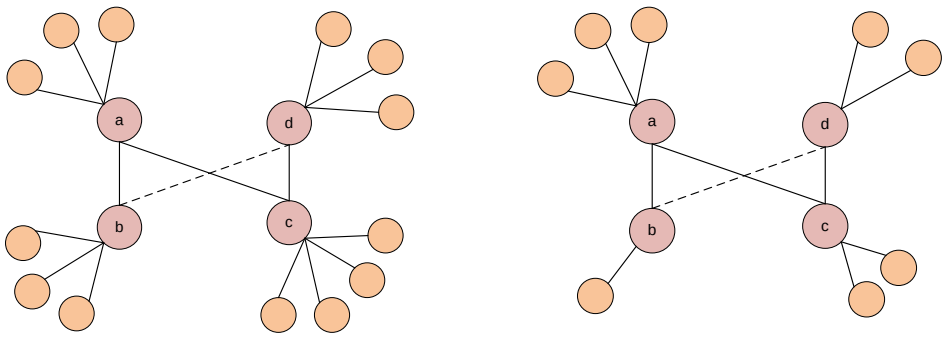
Let b and c belongs to U of the bipartite graph G = (U,V,E), if there exist a node q ∈ V such that (b,q) ∈ E and (c,q) ∈ E then the node pair {b,c} is a pattern in U-part of the bipartite graph G. In other words, if ɣ(b)∩ɣ(c) ϕ then the node pair {b,c} is a pattern in the U-part of the bipartie graph G = (U,V,E), where b,c ∈ U. Therefore, there must exist an edge (b,c) in the U-projection of the bipartite graph G. Similarly, let p and q belong to the V part of the bipartite graph G = (U,V,E), if there exists a node a ∈ U such that (p,a) ∈ E and (q,a) ∈ E then the node pair {p,q} is a pattern in the V-part of the bipartite graph G. In other words, if ɣ(p) ∩ ɣ(q) ϕ then the node pair {p,q} is a pattern in the V-part of the bipartite graph G = (U,V,E), where p,q ∈ V . Therefore, there must exist an edge (p,q) in the V-projection of the bipartite graph G.

Let (c,p) be a potential link in bipartite graph G: G_u and G_v which are the U-part and V-part of a projected graph with threshold T. For each node $g_i \in ɣu(c) \cap ɣ(p)$, we call {c, g_i} is the pattern covered by potential link (p,c) in the U-part of the network, for each node $s_i \in ɣv(p) \cap ɣ(c)$. and call {p, s_i} is the pattern covered by potential link (p,c) in the V-part of the network. **Fig. 2a** and **Fig. 2b** represent the U-projection and V-projection of the network in **Fig. 1**. The neighbor of node c in G_u, $ɣ_u(c)$ = {a,b,d}, and neighbor of p in bipartite graph G, γ(p) = {a, b}, therefore patterns covered by potential link (c,p) in U-part of the network are, {{c, a},{c,b}}. The neighbor of node p in G_v, $γ_u(p)$ = {q,r,s}, and neighbor of c in bipartite graph G, γ(c) = {q,r,s}, therefore patterns covered by potential link (c,p) in V-part of the network are, {{p,q},{p,r},{p,s}}. Potential link (c,p) covered two patterns in the U-part and three patterns in the V-part, therefore, the patterns covered by the potential link in the U and V part of the network can be similar or different.

3.2.3 Reciprocal of the Sum of Neighbor of Butterflies

Let x_1,x_2,x_3, and x_4 is the nodes in butterfly, the sum of the neighbor of butterfly can be written as: $|γ(^{\infty}(x_1,x_2,x_3,x_4))| = |γ(x_1)| + |γ(x_2)| + |γ(x_3)| + |γ(x_4)| - 8$. Here 8 is subtracted because every node in the butterfly has two neighbors in the same butterfly. It is assumed that the higher the neighbors of nodes in the butterfly, the lower the chance of the formation of a link between the potential nodes. That is why the reciprocal of the sum of neighbors of butterflies be fed to the link predictor. Here it considers these neighbors of nodes in butterflies as noise. In **Fig. 7a** the number of neighbors of butterfly nodes is high compared to **Fig. 7b** so

the information travel path in the potential link (b,d) may get distracted with a higher probability in **Fig. 7a** due to a large number of neighbors open for a possible path.



(a) (b)

Fig. 7. butterfly nodes with (a) a large number of neighbors and (b) a small number of neighbors

3.2.4 Internal Link Classifier

In [22], the author has described the internal link and proved the sufficient condition to be an internal link. The work further studied the latent features and came up with the idea of the sufficient condition and proved it. The internal link classifier classifies the potential link into two classes: 0 and 1. If the link is internal, then this link belongs to class 1 else class 0. Consider G = (U, V, E) is a bipartite network G, x ∈ U, and y ∈ V are the two nodes in G and the edge (x,y) E. The works create a new bipartite network G' = (U, V, E') by adding (x,y) belongs to U × V or (x,y) ∈ E', where E' = E ∪ {(x × y)}. Let G_u = (U, E_u) be the U-projected graph of G and G'_u = (U, E'_u) be the U-projected graph of G'. If G_u = G'_u then (x,y) an internal link by U-projection. In other words, (x,y) is the pair of nodes in the bipartite graph G, such that adding a link (x,y) to G does not change the U-projection of G, then the node pair (x,y) is an internal link by U-projection. Similarly, if adding the link in graph G does not change the V-projection of graph G then this link is the internal link by V-projection. In **Fig. 1**, node pair (c,p) is an internal link by both U and V-projection. All the neighbors of node p, γ(p) = {a,b} are already connected to the node c in G_u, After adding a link (c,p), the neighbor between (c, a) and (c,b) will increase to two, but there is no addition of new link to G_u thus G_u does not change. Similarly, all the neighbors of node c, γ(c) = {q,r,s} are already connected to the node p in G_v. After adding a link (c,p), the neighbor between (p,q), (p,r), and (p,s) will increase but the V-projection of graph G does not change.

Theorem 1. (Necessary condition): A node pair (x,y) in a bipartite graph G = (U, V, E) is an internal link by U-projection if and only if it satisfies γu(x) ∩ γ(y) φ. Here, φ is the empty set.

Proof: Let (x,y) E, x ∈ U y ∈ V in G = (U, V, E) is an internal link. Let graph G' = (U, V, E')= E_u ∪ {(x,y)} is the new bipartite graph by adding the link (x,y) in G. Then according to internal link definition, E'_u =E_u U {(x,z), z ϵ y(y)} Suppose that (x,y) is an internal link, i.e. E_u =E'_u then all links (x,z) are already in Eu. Therefore, each z ∈ {γ(y) and γ_u(x)} so γ(y) ∩ γ_u(x) ≠φ.

(Sufficient condition): A node pair (x,y) in a bipartite graph G = (U, V, E) is an internal link by U-projection if it satisfies γ(y) ⊆ γ_u(x), and x γ(y).

Proof: Let (x,y) E, and x ɣ(y), x ∈ U y ∈ V in G = (U, V, E) be an internal link. Let graph G' = (U, V, E)' = E_u ∪{(x,y)} is the new bipartite graph by adding the link (x,y) in G. Then

according to internal link definition, $E'_u = E_u \cup \{ (x,z), z \in \gamma(y)\}$. Suppose that (x,y) is an internal link, i.e. $E_u = E'_u$ then all links (x,z) are already in Eu. Therefore, for each $z \in \gamma(y)$, there exists β in U. By symmetry, $z \in \gamma(\beta)$ and $z \in \gamma_u(\beta)$. Therefore, $z \in \gamma(x)$ and so $\gamma(y) \subseteq \gamma_u(x)$.

3.3 Link Predictor

In this research, different machine learning classification algorithms are used as a basic model to extract classes which are represented as 1 when a link is formed and 0 when a link is not formed.

4. Experiments

In the experiment, the connected link and possible link are used in equal proportion to training our model, as depicted in the proceeding sections.

4.1 Data Sets

Experiments are performed on five real-world data sets. (i) Enzyme [7]: a biological bipartite network of drugs binding enzyme protein. (ii) Ion Channel [7]: a biological bipartite network of drugs binding ion protein referred to here as (IC). (iii) G-protein Coupled Receptor (GPCR) [7]: a biological bipartite network of drugs binding G-protein Coupled Receptor. (v) Books [34]: a bipartite network of user who rates books. (vi) Drug side-effect association network (http://snap.stanford.edu/decagon): This is a drug side-effect association network that contains information on side effects caused by drugs that are on the U.S. market here referred to as (DSE). The topological statistics of these five data sets are presented in **Table 2**.

Table 2. The detailed information of the Five data sets used to verify the proposed model

Network	|U|	|V|	|E|
Enzyme	445	664	2926
IC	204	210	1476
GPCR	95	223	635
Books	445801	105278	1149739
DSE	640	10185	174978

* |V|, |U| denotes the number of two types of nodes. |E| indicates the number of edges between the nodes.

4.2 Baseline Algorithms

The work compares the proposed link prediction methods with other popular link prediction methods in the bipartite network as a baseline. Four of them are node neighborhood similarity-based methods (CN, JC, AA, PA) [18], three of them are LCP mechanism-based similarity methods (CAR, CJC, CAA) [18], and three embedding-based methods (line [30], deep walk [31], node2vec [32]). **Table 3** shows the algorithmic expression of all the baseline methods. Where $\gamma(x)$ and $\gamma(y)$ represent the first-order neighbor, $\gamma(\gamma(x))$ and $\gamma(\gamma(y))$ represent the second-order neighborhood of the node x and y respectively. S^{LCL} is the total link between the common neighbor between the nodes and $\alpha(z)$ is the local community degree of z. Relatively similar work to this paper is presented in [1]. The author focuses on the concepts of potential energy and mutual information. The approach taken is a three-step process: converting the bipartite graph into a unipartite graph using a weighted projection, computing the potential energy and mutual information between each node pair in the projected graph.

However, our model benefitted from the information in the unipartite network projected from the bipartite network and the structural information in an original bipartite network.

4.3 Evaluation Measures

The experiment is conducted to compare the proposed method with the various state of the art; link prediction methods in the bipartite network. Here experiment results with three evaluation indexes: area under the precision-recall curve (AUC-PR), F1-score (f), and Precision (p) are shown. The goal of the AUC-PR measure is to have a model at the right top corner, which means getting only true positives without false positives and false negatives. F1-score is the weighted average or harmonic mean of precision and recall.

Table 3. Overview of the baseline method

Class	Method	Formulae
Node-neighbor similarity	CN[18]	$S_{xy}^{CN} = \lvert(\gamma(x) \cap \gamma(\gamma(y))) \cup \gamma(\gamma(y) \cap \gamma(\gamma(x)))\rvert$
	JC[18]	$S_{xy}^{JC} = \frac{S_{xy}^{JC}}{\lvert\gamma(x) \cup \gamma(y)\rvert}$
	AA[18]	$S_{xy}^{AA} = \sum_{z \in ((\gamma(x) \cap \gamma(\gamma(y))) \cup \gamma(\gamma(y) \cap \gamma(\gamma(x))))} \frac{1}{log_2(\gamma(z))}$
	PA[18]	$S_{xy}^{PA} = \lvert\gamma(x)\rvert . \lvert\gamma(y)\rvert$
LCP mechanism similarity Embedding method	CAR[18]	$S_{xy}^{CAR} = S_{xy}^{CN} . S_{xy}^{LCL}$
	CJC[18]	$S_{xy}^{CJC} = \frac{S_{xy}^{CAR}}{\lvert\gamma(x) \cup \gamma(y)\rvert}$
	CAA[18]	$S_{xy}^{CAA} = \sum_{z \in ((\gamma(x) \cap \gamma(\gamma(y))) \cup \gamma(\gamma(y) \cap \gamma(\gamma(x))))} \frac{\alpha(z)}{log_2(\gamma(z))}$
Embedding Methods	line[30]	
	deepwalk[31]	
	node2vec[32]	

* The table represents the baseline algorithm used in this work to compare the results.

It measures the prediction precision and coverage, and the relative contribution of precision and recall to the F1-score is equal. F1-score can be calculated by the formula:

$$F1 = 2 \times \left(\frac{precision \times recall}{precision + recall} \right) \tag{2}$$

Where,

$$\text{Precision} = \frac{tp}{(tp + fp)}$$

$$\text{Recall} = \frac{tp}{(tp + fn)} \tag{3}$$

Where, tp, fp, and fn are the number of true-positive samples, number of false-positive samples, and number of false-negative samples respectively. Precision is the ability of a classifier not to label a negative sample as positive.

4.4 Experiment Result

All the experiments are performed on Intel core i5-7500 CPU at 3.40 GHZ×4 with 8-GB DDR4 RAM. The programming language used is python 3.6 under Ubuntu 16.04.6 LTS 64-bit. **Table 4**, **Table 5**, and **Table 6**, show the comparative values of the measure, AUC-PR, F1-score, and precision respectively for the proposed method, which is named a CSE, and all the other baseline methods. In this many machine-learning classifiers such as Random Forests

Classifier (RFC), Support Vector Machine Classifier (SVC), Decision Tree Classifier (DTC), and Gaussian Naive Bayes Classifier (GNB) are tested. All the results in the table are based on 10-fold cross-validation after removing 10% of the edges randomly from the extracted potential links. 10-fold cross-validation is used to verify the performance of all the baselines as well as the proposed model. In 10-fold cross-validation, the data set is randomly divided into ten parts, one part is fixed as a test data set, and the other nine parts are used for the training data sets.

Table 4. AUC-PR of the algorithm on different data sets

AUC-PR	Enzyme	IC	GPCR	DSE	Book
CN	0.757	0.780	0.698	0.748	0.781
JC	0.787	0.781	0.728	0.832	0.823
AA	0.780	0.771	0.689	0.814	0.795
PA	0.702	0.696	0.489	0.546	0.691
CAR	0.805	0.794	0.738	0.764	0.826
CJC	0.816	0.805	0.741	0.846	0.842
CAA	0.818	0.808	0.692	0.794	0.806
line	0.879	0.880	0.763	0.862	0.888
deepak	0.827	0.853	0.799	0.843	0.887
node2vec	0.899	0.899	0.782	0.873	0.898
CSE-KNC	0.905	0.922	0.860	0.958	0.957
CSE-DTC	0.921	0.911	0.831	0.954	0.945
CSE-SVC	0.862	0.873	0.781	0.899	0.848
CSE-GNB	0.843	0.853	0.791	0.941	0.883
CSE-RFC	0.925	0.927	0.849	0.959	0.958

Table 5. F1-score on different data sets with different

F	Enzyme	IC	GPCR	DSE	Book
CN	0.726	0.757	0.613	0.746	0.779
JC	0.814	0.760	0.641	0.814	0.823
AA	0.812	0.749	0.502	0.812	0.796
PA	0.524	0.650	0.227	0.524	0.683
CAR	0.779	0.762	0.675	0.762	0.826
CJC	0.816	0.776	0.678	0.848	0.844
CAA	0.818	0.782	0.582	0.793	0.807
Line	0.879	0.878	0.784	0.921	0.894
deepak	0.874	0.853	0.798	0.924	0.898
node2vec	0.898	0.895	0.793	0.932	0.926
CSE-KNC	0.883	0.910	0.800	0.957	0.957
CSE-DTC	0.907	0.892	0.772	0.954	0.945
CSE-SVC	0.839	0.854	0.712	0.899	0.837
CSE-GNB	0.798	0.808	0.673	0.941	0.877
CSE-RFC	0.916	0.921	0.805	0.959	0.957

Table 6. Precision on different data sets with different model

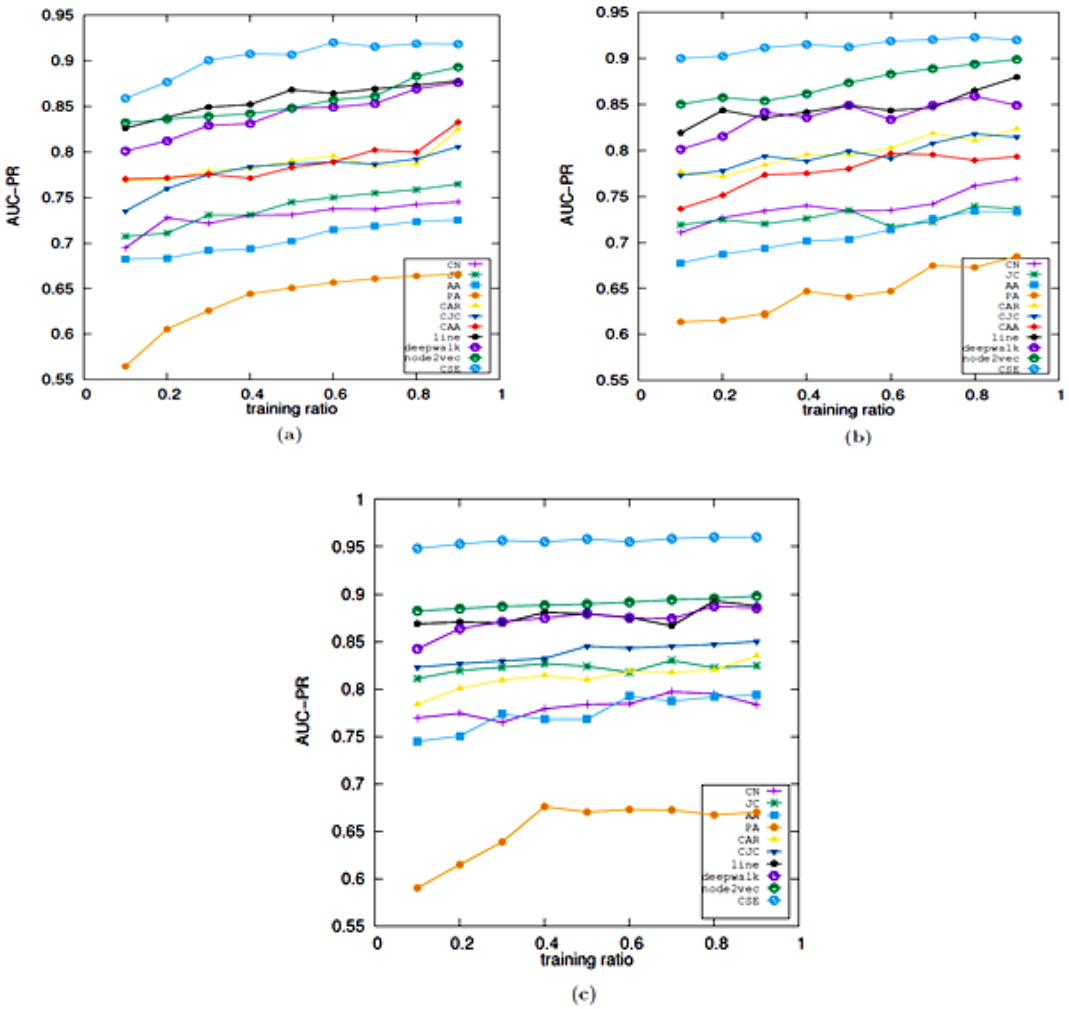
Precision	Enzyme	IC	GPCR	DSE	Book
CN	0.713	0.785	0.739	0.753	0.802
JC	0.757	0.757	0.689	0.842	0.823
AA	0.767	0.760	0.639	0.810	0.805
PA	0.693	0.659	0.472	0.562	0.705
CAR	0.802	0.790	0.768	0.766	0.821
CJC	0.803	0.788	0.733	0.842	0.842
CAA	0.805	0.788	0.643	0.800	0.811
line	0.874	0.883	0.821	0.925	0.903
deepak	0.875	0.881	0.854	0.902	0.904
node2vec	0.884	0.905	0.824	0.934	0.826
CSE-KNC	0.888	0.950	0.808	0.958	0.957
CSE-DTC	0.906	0.893	0.828	0.949	0.957
CSE-SVC	0.871	0.907	0.840	0.900	0.926
CSE-GNB	0.800	0.801	0.686	0.923	0.951
CSE-RFC	0.923	0.922	0.857	0.958	0.958

The process is repeated ten times. For embedding-based methods, the embedding dimensions of 12 for dataset Enzyme, IC, GPCR, and 128 for Book and DSE dataset are used to achieve the best result. After getting embedding for every node, a random sample of some node is used that subtracts their embedding matrix and feeds it into a classifier (in this experiment RFC works best) to classify the link. In node neighbor-based similarity and LCP-based similarity methods, features are extracted according to the formula and fed these features into a classifier (in this experiment RFC works best in most of the cases) to classify the link. In Tables 4, 5, and 6, the first column represents all the baseline and this method, and the first row represents the data sets with which the experiment is carried. Each cell in the table represents the results of the corresponding method in the corresponding data sets. The highest value in the column is shown in the bold text.

The measure AUC-PR is improved by this proposed method with the best -performing classifier for all the datasets, as shown in **Table 4**. For example, the improvement from the nearest algorithm is 2.89% (from node2vec) on Enzyme, 3.11% (from node2vec) on IC, 6.1% (from the deep walk) on GPCR, 8.6% (from node2vec) on DSE, and 7% (from the line) in Book dataset. This indicates that this proposed method surpasses the results of other baseline methods in selecting only true positives by ignoring the false positives and false negatives.

The measure F1-score of the proposed method is also higher than the other baseline methods, as shown in **Table 5**. For example, an improvement from the nearest algorithm is 2% (from node2vec) on Enzyme, 3.06% (from node2vec) on IC, 1% (from a deep walk) on GPCR, 2.9% (from node2vec) on DSE, and 3.34% from (node2vec) on Book datasets. Similar to measuring AUC-PR and F1-Score, the measurement precision by the proposed method is also higher than

the baseline methods as shown in **Table 6**. Improvement in precision is ∼1% to ∼4% on different datasets from the best-performing baseline algorithm. In most cases, Random Forest Classifier (RFC) outperforms another classifier in the method. The work has also analyzed the experiment results by varying training datasets for three bipartite networks: Enzyme, IC, and Books, from 10% to 90% shown in **Fig. 8**, **Fig. 9**, and **Fig. 10**.



* The graphs are plotted from the data obtained from experiments.

Fig. 8. AUC-PR under different methods with different sizes of training sets on three real networks. (a) Enzyme Channel. (b) Ion channel. (c) Books.

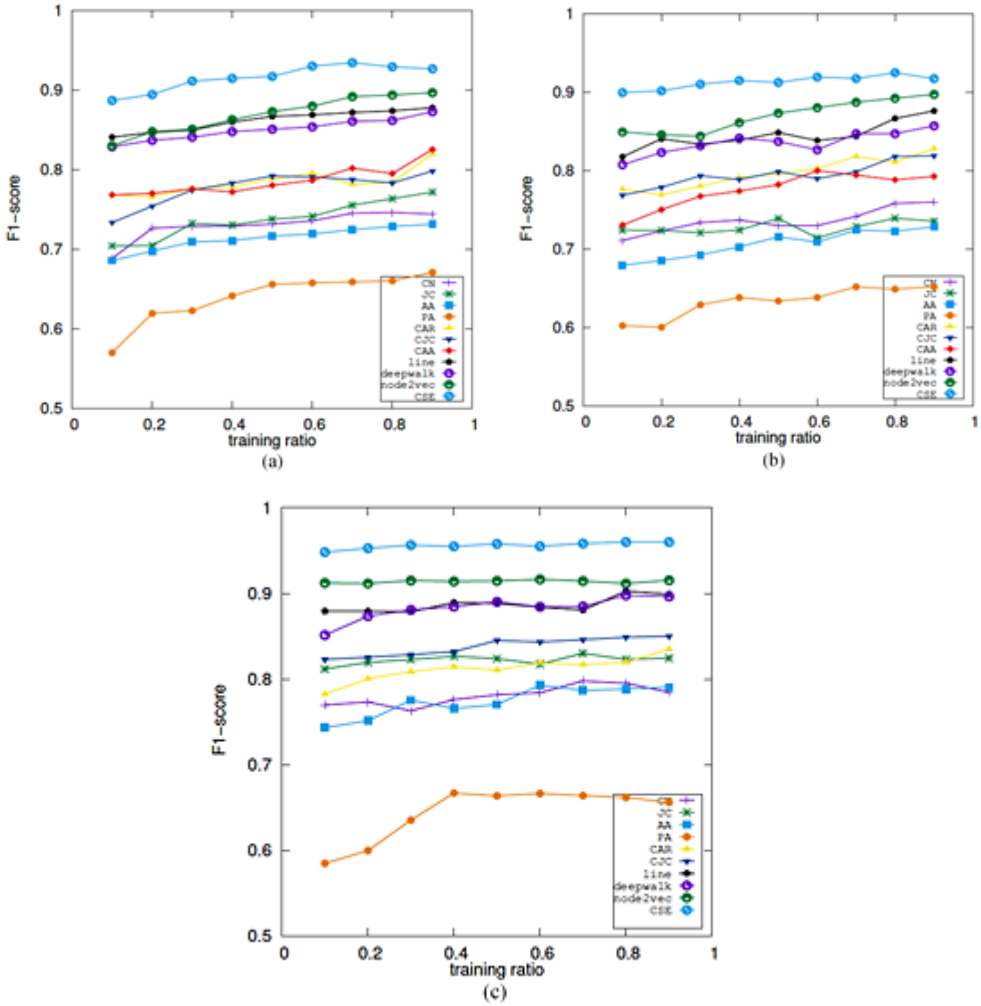


Fig. 9. F1-score under different methods with different sizes of training sets on three real networks. (a) Enzyme Channel. (b) Ion channel. (c) Books.

The work uses a random forest classifier in this method and the method is represented by name (CSE). In most cases, increasing the training ratio improves the performance of the methods this is because the increase in training ratio increases the information needed for the classifier to classify unobserved links.

The work also carried out principal component analysis (PCA) [40] in extracted features for all the datasets. The main purpose of doing PCA is to find new variables that are a linear function of those in the extracted datasets and that successively maximize variance and are uncorrelated with each other. The cumulative explained variance for all the datasets is presented in **Fig. 11**. It shows each component is not related to each the others as the cumulative explained variance reaches 100% in the fourth component in any dataset. This indicates that all four features play a role independently in link prediction on these datasets.

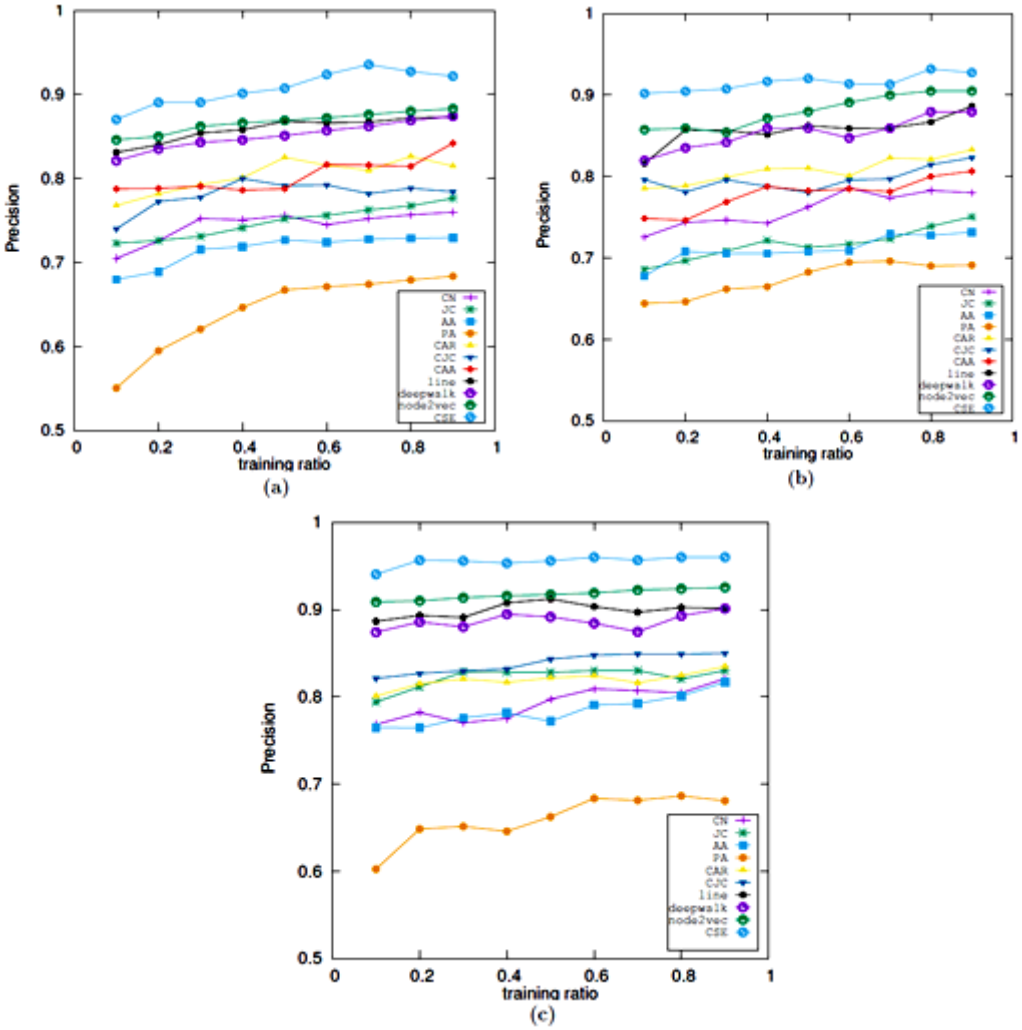


Fig. 10. Precision under different methods with different sizes of training sets on three real networks. (a) Enzyme channel. (b) Ion channel. (c) Books.

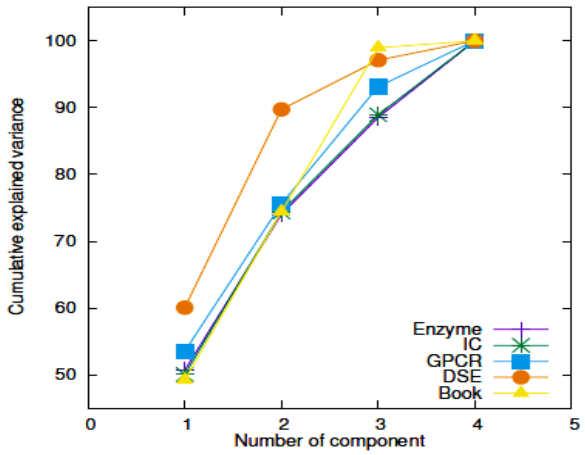


Fig. 11. Principal component analysis on extracted features

This work has also performed experiments without removing weak redundant links using a random classifier as a predictor and the result for three metrics: AUC-PR, F1-score, and Precision is presented in **Table 7** which shows that the proposed method performs better (can be seen in **Table 4**, **5, 6**) after the weak links were removed.

Table 7. Evaluation metrics on book data set using a combination of extracted features

	Enzyme	IC	GPCR	DSE	Book
AUC-PR	0.881	0.873	0.842	0.912	0.910
F1-score	0.873	0.882	0.792	0.902	0.921
Precision	0.878	0.891	0.823	0.911	0.923

The research also analyzed the result of our proposed methods using each single extracted feature and the combination of these features for some datasets shown in Tables 8 and 9. From these tables, it is clear that adding every extracted feature improves the link prediction ability of the proposed model. It has represented several butterfly enclosures, pattern covered, reciprocal of the sum of neighbors of butterfly nodes, and internal link features as C1, C2, C3, and C4 respectively. In this paper, the extracted feature C1 is the most important feature as it predicts a link with a high score alone the other 3 features. C2 and C3 are the second and third important features that contribute to link prediction as shown in **Table 8** and **Table 9** for sample datasets. The work demonstrates how the result changes if all these single features and a combination of prominent features are fed. The feeding of C1 and C2 features into the classifier shows that the result for link prediction is better than compared to feeding single features. The same is valid for a combination of C1, C2, and C3.

The work achieves the best result by feeding all these extracted features. This indicates that all these extracted features somehow play role in the improvement of the model link prediction ability.

Table 8. Evaluation metrics on book data set using a combination of extracted features

	C1	C2	C3	C4	C1+C2	C1+C2+C3
AUC-PR	0.891	0.894	0.736	0.871	0.927	0.948
F1-score	0.890	0.892	0.715	0.878	0.925	0.948
Precision	0.922	0.919	0.788	0.809	0.923	0.946

Table 9. Evaluation metrics on enzyme data set using a combination of extracted features

	C1	C2	C3	C4	C1+C2	C1+C2+C3
AUC-PR	0.826	0.817	0.716	0.46	0.890	0.921
F1-score	0.789	0.767	0.617	0.11	0.874	0.909
Precision	0.855	0.788	0.646	0.13	0.889	0.909

5. Discussions and Contribution of the Research

The provided research focuses on addressing the limitations of existing mechanism-based and latent feature-based models for link prediction in bipartite networks. The paper introduces a novel approach based on a composite similarity measure, which aims to enhance the accuracy and robustness of the prediction models by considering semantic similarity, structural similarity, and behavioral similarity to determine entity similarity. One of the key strengths of our proposed approach is the integration of multiple similarity metrics, such as cosine similarity, Euclidean distance, and Jacquard similarity, into a composite measure. By combining these metrics, our model captures the different aspects of similarity and provides a more comprehensive evaluation. This not only improves the accuracy of link prediction but also enhances the model's ability to handle complex network structures and relationships. Another important contribution of the research is the incorporation of structural features in the prediction process. The introduction of features such as the number of butterfly enclosures and the average number of patterns adds valuable information about the network structure and the ways in which information can travel between nodes. These features provide a deeper understanding of the relationships within the bipartite network and contribute to the overall prediction accuracy. Furthermore, this work suggests the utilization of techniques like feature selection, dimensionality reduction, and ensemble learning to optimize the model's performance. These approaches help to identify the most relevant features, reduce the computational complexity, and improve the generalization ability of the model. The proposed research has potential implications and applications across various domains, including e-commerce networks, biological networks, social networks, and drug side effect networks. The ability to accurately predict links in bipartite networks has practical significance in understanding unknown interactions, predicting future connections, and providing valuable insights into the underlying relationships. The research addresses a vital aspect of network analysis in modern network-based systems and contributes to the advancement of knowledge in this field. However, one area that could be further discussed is the evaluation and validation of the proposed approach. While our paper mentions the use of a classifier algorithm and potential techniques for improving model performance, it lacks some more work in the details on the specific evaluation metrics used, the choice of benchmark datasets, and comparative analyses with existing methods. This can also be the extended work for this research. Overall, the research presents an innovative approach to link prediction in bipartite networks by integrating mechanism-based and latent feature-based similarities through a composite similarity measure. The incorporation of structural features, the utilization of multiple similarity metrics, and the consideration of various domains make this research significant in advancing the understanding and prediction of relationships within complex network structures.

The study significantly advances the field by presenting a unique method for link prediction in bipartite networks based on a composite similarity measure. By including both structural and latent feature-based similarities, this strategy solves the drawbacks of previous mechanism-based and latent feature-based models. Accurate predictions are enhanced by the addition of structural components like butterfly cages and patterns. The composite similarity measure synthesizes many similarity measures to offer a thorough assessment of similarity. The suggested method improves the precision and robustness of current models by taking into account semantic, structural, and behavioral similarities. The study also proposes merging feature selection, dimensionality reduction, and ensemble learning approaches to further improve model performance. A thorough framework for evaluating complex networks across a variety of disciplines is provided by this research.

6. Conclusions

This research analyzes the multiple similarities directly extracted from the original bipartite network after extracting potential links from the weighted projected network. The work focuses on preserving structure-based, neighbor-based, and latent features to make the final prediction of the link. By summing up these features, the proposed method achieved superior results compared to other state-of-art methods. Experiments were performed by taking different real-world data sets which consist of a biological bipartite network of drugs binding enzyme protein, a biological bipartite network of drugs binding ion protein, a biological bipartite network of drugs binding G-protein Coupled Receptor, and a bipartite network of a user who rates books and compared the results of same data sets with other state-of-art link prediction methods. The propose proposed method performed better than other methods and we were able to reach more than 95% of AUC-PR, F1-score, and precision hence proving the algorithm to be better. This work holds great significance in the study of complex bipartite networks to predict link and unknown interactions formed by nodes and edges in a given network, which can be utilized and applied in various fields of research and business. The work has an utmost application in the study of biological networks, e-commerce networks, complex web-based systems, the network of drug binding, enzyme protein, and other related networks to understand the formation of such complex networks. Further, the results achieved from this work will help in link prediction usability for different purposes ranging from building intelligent systems to providing service in big data and web-based systems.

Future research in bipartite networks should concentrate on scalable and effective methods for dealing with sizable network datasets, integrating various data sources, and investigating hybrid models that mix mechanism-based and latent feature-based strategies. Deep learning, graph neural networks, and reinforcement learning are examples of advanced machine learning algorithms that may capture complicated patterns and non-linear correlations that standard models would find challenging. The effectiveness of suggested techniques will be established, and their benefits will be emphasized, through evaluation studies and comparison analyses utilizing benchmark datasets. Finally, real-world applications in e-commerce, social networks, or biology may show how the generated models perform in practice and confirm how well they work to solve certain problems.

References

[1] P. Holme, F. Liljeros, C. R. Edling and B. J. Kim, "Network bioactivity," *Physical Review E*, 68(5), 2003. Article (CrossRef Link)

[2] M. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America*, 98, 404-409, 2001. Article (CrossRef Link)

[3] F. Liljeros, C. Edling, L. Amaral, H. Stanley and Y. Aberg, "The web of human sexual contacts," *Nature*, 411, 907-908, 2001. Article (CrossRef Link)

[4] H. Jeong, B. Tombor, R. Albert, Z. Oltvai and A.-L. Barabasi, "The large-scale organization of metabolic networks," *Nature*, 407, 651–654, 2000. Article (CrossRef Link)

[5] T. Zhou, J. Ren, M. Medo, and Y. C. Zhang, "Bipartite network projection and personal recommendation," *Physical Review. E, statistical, nonlinear, and soft matter physics*, 76, 046115, 2007. Article (CrossRef Link)

[6] H. Chen, X. Li and Z. Huang, "Link prediction approach to collaborative filtering," in *Proc. of the 5th ACM/IEEE-CS,Joint Conference on Digital Libraries*, pp. 141–142, 2005. Article (CrossRef Link)

[7] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda and M. Kanehisa, "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, 24 (13), pp. i232–i240, 2008. Article (CrossRef Link).

[8] E. M. Airoldi, D. M. Blei, S. E. Fienberg, E. P. Xing and T. Jaakkola, "Mixed membership stochastic block models for relational data with application to protein-protein interactions," in *Proc. of the International Biometrics Society Annual Meeting*, 2006.

[9] N. Benchettara, R. Kanawati and C. Rouveirol, "Supervised machine learning applied to link prediction in bipartite social networks," in *Proc. of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, pp. 326–330, August 2010. Article (CrossRef Link).

[10] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines et al., "Friendship prediction and homophily in social media," *ACM Transactions on the Web*, 6 (2), 9:1–9:33, 2012. Article (CrossRef Link).

[11] W. J. Burk, C. E. Steglich and T. A. Snijders, "Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors," *International Journal of Behavioral Development*, 31 (4), 397–404, 2007. Article (CrossRef Link).

[12] Y. Luo, Q. Liu, W. Wu, F. Li and X. Bo, "Predicting drug side effects based on link prediction in bipartite network," in *Proc. of 2014 7th International Conference on Biomedical Engineering and Informatics*, pp. 729–733, 2014. Article (CrossRef Link).

[13] V. Mart´ınez, F. Berzal and J. C. Cubero, "A survey of link prediction in complex networks," *ACM Computing Surveys*, 49 (4), 69:1–69:33, 2016. Article (CrossRef Link).

[14] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, 58 (7), 1019–1031, 2007. Article (CrossRef Link).

[15] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Physical Review. E.*, 64, 025102, 2001. Article (CrossRef Link).

[16] J. Kunegis, E.W. De Luca and S. Albayrak, "The link prediction problem in bipartite networks," in *Proc. of IPMU 2010: Computational Intelligence for Knowledge-Based Systems Design*, pp. 380-389, 2010. Article (CrossRef Link).

[17] S. Fakhraei, B. Huang, L. Raschid and L. Getoor, "Network-based drug-target interaction prediction with probabilistic soft logic," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11 (5), 775–787, 2014. Article (CrossRef Link).

[18] S. Daminelli, J. M. Thomas, C. Dur´and C. V. Cannistraci, "Common neighbors and the local community-paradigm for topological link prediction in bipartite networks," *New Journal of Physics*, 17, 113037, 2015. Article (CrossRef Link).

[19] C. Dur´an, S. Daminelli, J. M. Thomas, V. J. Haupt, M. Schroeder et al., "Pioneering topological methods for network-based drug-target prediction by exploiting a brain-network self-organization theory," *Briefings in Bioinformatics*, 19 (6), 1183–1202, 2017. Article (CrossRef Link).

[20] S. Aslan and M. Kaya, "Topic recommendation for authors as a link prediction problem," *Future Generation Computer Systems*, 89, 249 – 264, 2018. Article (CrossRef Link)

[21] M. McPherson, L. Smith-Lovin and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, 27 (1), 415–444, 2001. Article (CrossRef Link).

[22] M. Gao, L. Chen, B. Li, Y. Li, W. Liu et al., "Projection-based link prediction in a bipartite network," *Information Science*, 376, 158–171, 2017. Article (CrossRef Link)

[23] M. A. Yildirim and M. Coscia, "Using random walks to generate associations between objects," *PLOS ONE*, 9 (8), 1–9, 2014. Article (CrossRef Link)

[24] K. Lewis, M. Gonzalez and J. Kaufman, "Social selection and peer influence in an online social network," *Proceedings of the National Academy of Sciences*, 109 (1), 68–72, 2011. Article (CrossRef Link).

[25] O. Allali, C. Magnien and M. Latapy, "Internal link prediction: A new approach for predicting links in bipartite graphs," *Intelligent Data Analysis*, 17 (1), 5–25, 2013. Article (CrossRef Link).

[26] D. Kalman, "A singularly valuable decomposition: The side of a matrix," *The College Mathematics Journal*, 27 (1), 2–23, 1996. Article (CrossRef Link).

[27] D. B. Larremore, A. Clauset and A. Z. Jacobs, "Efficiently inferring community structure in bipartite networks," Physical Review. E, *Statistical, Nonlinear, and Soft Matter Physics*, 90(1), 012805, 2014. Article (CrossRef Link)
[28] C. Aicher, A. Z. Jacobs and A. Clauset, "Learning latent block structure in weighted networks," *Journal of Complex Networks*, 3(2), 221–248, 2015. Article (CrossRef Link)
[29] D. Lian, R. Liu, Y. Ge, K. Zheng, X. Xie et al., "Discrete content-aware matrix factorization," in *Proc. of the 23rd ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining*, pp. 325–334, 2017. Article (CrossRef Link)
[30] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan et al., "Line: Large-scale information network embedding," in *Proc. of the 24th International Conference on World Wide Web*, pp. 1067–1077, 2015. Article (CrossRef Link)
[31] B. Perozzi, R. Al-Rfou and S. Skiena, "Deepwalk: Online learning of social representations," in *Proc. of the 20th ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining*, pp. 701–710, 2014. Article (CrossRef Link)
[32] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. of the 22nd ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining*, pp. 855-864, 2016,
[33] W. Hamilton, Z. Ying and J. Leskovec, "Inductive representation learning on large graphs," *Advances in Neural Information Processing Systems*, pp. 1024-1034, 2017.
[34] C. N. Ziegler, S. M. McNee, J. A. Konstan and G. Lausen, "Improving recommendation lists through topic diversification," in *Proc. of International Conference on World Wide Web Conference*, pp. 22-32, 2005.
[35] M. I. Marwat, J. A. Khan, M. D. Alshehri, M. A. Ali, M. Hizbullah et al., "Sentiment analysis of product reviews to identify deceptive rating information in social media: A senti deceptive approach," *KSII Transactions on Internet and Information Systems*, vol. 16, no. 3, pp. 830-860, 2022. Article (CrossRef Link)
[36] M. Hasnain, I. Ghani, M. F. Pasha and S. R. Jeong, "Machine learning methods for trust-based selection of web services," *KSII Transactions on Internet and Information Systems*, vol. 16, no. 1, pp. 38-59, 2022. Article (CrossRef Link)
[37] L. Wang, X. Lv and J. An, "Trajectory distance algorithm based on segment transformation distance," *KSII Transactions on Internet and Information Systems*, vol. 16, no. 4, pp. 1095-1109, 2022. Article (CrossRef Link).
[38] S. Banerjee, M. Jenamani and D. K. Pratihar, "Algorithms for projecting a bipartite network," in *Proc. of Tenth International Conference on Contemporary Computing (IC3)*, Noida, India, pp. 1-3, 2017. Article (CrossRef Link).
[39] S. Aslan and M. I. Y. Kaya, "Predicting links in complex bipartite networks based on strengthening projections," in *Proc. of International Conference on Artificial Intelligence and Data Processing (IDAP)*, 1-6, 2018.
[40] S. Wold, K. Esbensen and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, Vol. 2, no.1–3, pp. 37-52, 1987. Article (CrossRef Link).

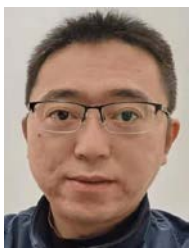
Bijay Gaudel holds a Master of Engineering Degree in Computer Science and Technology from Nanjing University of Aeronautics and Astronautics, and a Bachelor's degree in Electrical Engineering from Institute of Engineering, Tribhuvan University, Nepal. He has done his master's degree research in Graph Representation Learning using deep learning. His research interest includes Computer vision, graph representation learning, and complex network analysis.

Deepanjal Shrestha is an associate professor at the School of Engineering, Pokhara University, Nepal. He holds a Ph.D. from Nanjing University of Aeronautics and Astronautics in Nanjing, China. Assoc. Prof. Shrestha's expertise spans multiple domains, including management information systems, software engineering, digital ecosystems, and service engineering. Currently, He is involved in a groundbreaking research endeavor focused on designing and implementing a digital tourism business ecosystem for Nepal.

Niosh Basnet holds a Bachelors degree in Aeronautical Engineering from Nanjing University of Aeronautics and Astronautics. His research includes logic of modelling; applying a PID (Proportional-Integral-Derivative) controller to systems (magnetic train catching objects); applying MPC (Model Predictive Controller to systems (autonomous car: lane changing maneuvers). His interest lies in space system design; flight vehicle dynamics and control; integration of machine learning algorithms in design optimization and navigation of flight vehicles).

Neesha Rajkarnikar is a Ph.D. Scholar at School of Development and Social Engineering, Pokhara University. She is also an Assistant Professor of Computer Science at School of Business, Pokhara University. She holds Master's degree in Computer Applications and another Masters in Gender & Development. Her area of interest includes Computer Applications, Social Computing and Social Sciences.

Seung Ryul Jeong is a Professor at Graduate School of Business IT, Kookmin University, Seoul, South Korea since 1997. He earned his B.A. in Economics, from Sogang University Seoul, Korea in 1985. He completed his Master of Science degree in MIS from University of Wisconsin - Milwaukee, WI, U.S.A. and PhD in MIS from the University of South Carolina, SC, U.S.A. His research Interests include System Implementation, Process Innovation, Text Mining, Project Management, Information Resource Management and Social Computing, etc



Donghai Guan is an Associate Professor in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He received a Ph.D. from Kyung Hee University (KHU), Suwon, South Korea, in 2009. From 2009 to 2011, he was a Research Professor with the Department of Computer Engineering, KHU. From 2012 to 2014, he was an Assistant Professor with KHU. He has authored more than 70 research papers in related international conferences and journals. His current research interests include machine learning, internet of things, social network analysis, and ubiquitous computing.