Pertanika J. Sci. & Technol. 34 (S1): 199 - 232 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Filtered-Based Online Social Networking Features Extraction and Node Link Prediction Approach

Rajasekhar Nennuri[1]*, G Pradeepini[1], and L V Narasimha Prasad[2]

[1]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Green Fields, 522302 Vaddeswaram, Guntur District, Andhra Pradesh, India
[2]Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Dundigal, 500043 Hyderabad, Telangana, India

ABSTRACT

Online social networks, a significant part of the Internet, contain complex statistical relationships among millions of social nodes. Link prediction algorithms compute a likelihood of potential future ties between nodes in these networks. Most traditional methods use some form of contextual similarity between nodes to discover relations. Traditional community detection models tend to rely on limited structural information and ignore the impact of central nodes during link prediction. These restrictions impair the efficacy of link-prediction-based decision-making. In this paper, a community-based link prediction technique taking the advantages of both is proposed for dynamic and noisy networks. A new community detection metric is proposed to identify candidate links in a web of nodes. The central node in the network is calculated by this metric and used by the link prediction algorithm. The proposed community detection metric, along with other existing metrics, is employed to learn a model for hybrid link prediction. Experimental results indicate that the proposed central node-based link prediction method achieves better performance in prediction accuracy, prediction error, and computation time than traditional methods.

Keywords: Community detection, link prediction, similarity measures, social networking data

ARTICLE INFO

Article history:
Received: 07 March 2026
Accepted: 27 April 2026
Published: 24 July 2026

DOI: https://doi.org/10.47836/pjst.34.S1.10

E-mail address:
rajasekharnennuri@gmail.com (Rajasekhar Nennuri)
pradeepini_cse@kluniversity.in (G Pradeepini)
lvnprasad@yahoo.com (L V Narasimha Prasad)
* Corresponding author

INTRODUCTION

Online social networks are defined as social nodes and their interactions, as stated by Wu et al. (2017). This principle of construction represents the social relations between social actors, be they individuals or organisations. Social nodes, in their turn, are interconnected via edges that portray relations or communications. The purpose

e-ISSN: 2231-8526

of these networks is to create a larger network from which benefits can be obtained, such as better information access and greater influence flow. Examples range from Facebook's goal of "giving people the power to build community and bring the world closer together" to LinkedIn's mission of "connecting the world's professionals to make them more productive and successful." People tend to develop relationships in networks where they share common interests, backgrounds or prior association. It is difficult to model the expansion of these networks, but the analysis of the relations among pairs of nodes can provide useful insights.

Link prediction is the task of predicting the link among the disconnected nodes. Pandey et al. (2019) mention that link prediction methods have a huge impact on real-world structure reconstruction along with various applications. For instance, in security systems, link prediction may analyse terrorist networks to identify future associations, which can help government agencies to track potential threats and suspects. In biological networks, link prediction facilitates predicting new protein–protein interactions from existing ones, as presented by Bastami et al. (2019). Other applications in social networks are automatic hyperlink prediction and user-to-user recommendation systems (Aghabozorgi & Khayyambashi, 2018). Despite the existing similarity-based methods, enhancing the accuracy of link prediction is still a big challenge. Several models have been developed to improve the performance and evaluation methods, and such models were trained and tested on ego networks obtained from the Facebook dataset in the SNAP repository (Berahmand et al., 2021). These are actual networks of real people with varying degree distributions, clustering, degree correlation, etc.

With the continuous generation of new links in online social networks, the study of dynamic topological properties has great significance. Class imbalance also limits the predictive power in link prediction problems. Contrary to each user connecting to millions of other users on Facebook, each user is only connected to hundreds of nodes, as claimed by Jiang et al. (2021). This causes a huge skew between existing versus the non-Jagielski et al. agree that predicting every possible missing link in such large networks is computationally prohibitive; thus, sensible to truncate the set of node pairs on which to evaluate. Consequence networks are smaller parts of bigger networks, and for this reason ego networks are often taken for experimental analysis. Because social networks are self-similar along their subsystems, the results obtained from these subsystems can also be applied to the full system. In this study, we aim to develop computationally efficient models to predict the future structure of the ego network with dynamic topology and network sparsity.

A Facebook "community" is a bundle of groups. The widespread and dynamic features of social network data make identifying useful information imperative. Social Network Analysis (SNA) deals with identifying and analysing patterns in such data. SNA is also used for developing new trends, community detection, brand monitoring and election result

prediction. Such applications underscore the value of analysis of social networks. In such networks, entities tend to share many connections within groups and few across groups (i.e. the entities in such networks have a community structure). This paper concentrates on finding such communities. The goal is to show that these simple, quantitative measures of network structure can give rise to good link prediction algorithms.

Stochastic models can provide a good way to model and predict the evolution of a network. A probabilistic network is a joint probability distribution over a set of random variables, whose factorisation is determined by nodes that represent random variables and edges that represent dependencies. One of the standard models in this type of representation is the Bayesian Network model, which is a directed acyclic graph (DAG) that models causal and conditional relationships among variables. Bayesian inference is employed to revise the forecasts with the arrival of additional data. In the case of online social networks, an edge is the link between two vertices with which a number of measures can be defined. Such measures are, however, challenging to compute in static networks. Reachability, which is the possibility for one node to reach another through paths and is a function of network structure. In homogeneous networks, reachability is predictable, whereas in heterogeneous networks, paths can be highly variable

Although the proposed framework combines feature extraction, probabilistic ranking, and optimisation methods, the novelty should be clarified considering the existing methods. Traditional methods for link prediction are mostly based on similarity measures, predicate structural heuristics or simple probabilistic models. In contrast to this, current best-performing approaches are based on representation learning (e.g. graph embeddings) or deep learning-based approaches (GNNs) and knowledge graph embedding techniques. These techniques learn representations (viz., low-dimensional vectors) of nodes as well as relations, which give rise to enhanced prediction accuracy and better scalability over complex networks. Also, deep learning-based link prediction algorithms, such as neural networks and GNN models, have been proven to be effective in modelling structural and attributive dependencies of graphs. However, many of these methods involve computationally expensive procedures and require training on large-scale data, and are often criticised for their lack of interpretability in decision-making. The novelty of our approach is that we unify graph filtering, probabilistic source–destination ranking and optimisation-based feature selection in one procedure. Unlike the embedding-type methods based on latent vector representations, the proposed model takes explicit central node influence, shortest path dependencies and follower–followed dynamics into account to enhance prediction accuracy. Also, the probabilistic ranking formulation combined with an optimisation-based Random Forest algorithm balances the trade-off between interpretability and performance. Hence, the main novelty of this article is to design a hybrid, interpretable, and efficient link prediction method which connects traditional similarity-based methods

and recent learning-based methods and is capable of effectively dealing with sparse and imbalanced data in social networks.

In this paper, a link prediction framework based on hybrid filtering is proposed for online social networks. The approach combines graph filtering for candidate edge selection, probabilistic source–destination ranking for feature extraction, and an optimisation-based model for the final link prediction. This pair allows us to focus on predicting potential links for large and sparse networks efficiently and accurately.

Related Works

Graph representation learning has become a central paradigm for contemporary link prediction. These techniques aim to learn low-dimensional vector representations (embeddings) of the nodes that retain the local and global structural features of the graph. Then, these derived representations can be used for downstream tasks like link prediction, node classification, and clustering. Generally, methods for learning graph representations can be classified into two categories: non-neural embedding methods and Graph Neural Network (GNN)-based methods. Without neural networks, e.g., random walk-based methods, node embeddings are generated by traversing graph neighbourhoods; with neural networks, e.g., GNN, methods are employed directly on graphs to learn feature-rich node representations. When considering non-neural approaches, techniques like Node2Vec and DeepWalk are very popular for their capability to encode structural equivalence with stochastic random walks. However, they tend to face similar limitations, such as being non-interpretable and unable to utilise abundant node attributes. To address these issues, recent works have investigated deep learning-based approaches that simultaneously capture node features and graph topology.

In recent years, Graph Neural Networks have emerged as the most popular framework for link prediction. These models work by aggregating information from their neighbours based on a message-passing scheme, which allows them to learn hierarchical and context-aware representations. Variants like Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs) and GraphSAGE have also been exploited extensively for link prediction. Current surveys indicate that GNN-based approaches have emerged at the core of the best-performing strategies, as they can simultaneously and naturally model both the structural dependencies and the interactions between node features in intricate networks.

Much recent progress has been made to extend GNN-based link prediction to dynamic/temporal networks. Models for temporal link prediction consider time-evolving network structures and model the evolution of link formation through time. These models usually merge GNN architectures with recurrent neural networks or attention mechanisms to account for spatial and temporal correlations. For example, hybrid GCN-based frameworks with attention and sequence modelling approaches have achieved better predictability in dynamic networks.

Another promising direction for link prediction is **contrastive learning and self-supervised learning techniques**. These methods attempt to enhance the quality of representations by increasing the mutual information between different graph views. Specifically, line graph contrastive learning methods convert edge prediction tasks into node classification problems and hence facilitate more efficient learning of edge-level representations, and also generalise well in both sparse and dense networks ([ScienceDirect][4]).

Transformer-based models have also been developed for link prediction, particularly in knowledge graphs and multimodal networks. These models use attention to fuse different structural, textual and visual information for learning better representations. It has been recently demonstrated that multi-modal Transformer-based architectures can enhance the prediction accuracy substantially by learning heterogeneous correlations among different modalities of data ([MDPI][5]).

Furthermore, specialised GNN designs have been studied for link prediction situations. For instance, topology-informed GNN models preserve the structural information and eliminate the noise from unrelated attributes of nodes, and this leads to better prediction results in link-related tasks.

However, there are still multiple challenges in the current link prediction methods. Deep learning models, in general, tend to be data-hungry and computationally intensive. However, they may also be plagued by problems such as overfitting, distributional shifts and inability to interpret. Researchers have demonstrated that GNN-based models can be deficient in capturing long-range dependencies and potentially inject noise through neighbour-based aggregation schemes ([arXiv][7]). In addition, the issues of fairness and bias in link prediction have emerged as critical research topics, as link prediction models can make biased predictions with respect to sensitive attributes.

More recently, Kou et al. (2021) works on privacy-preserving and computationally efficient link prediction methods have also been investigated. For example, graph condensation techniques have been developed to shrink the graph size while retaining the structural information for fast training, making graph learning scalable but at the price of a small accuracy loss of prediction. Likewise, several subgraph-based representation learning methods have been proposed to promote scalability with good predictive capability in large-scale networks.

Unlike methods based solely on deep learning, some hybrid methods based on structural features, probabilistic reasoning, and machine learning techniques have focused on the compromise of performance and interpretability. The objective of these methods is to combine the advantages of classical and contemporary methods and to overcome the shortcomings of either.

Placed in this line of work, our framework is novel in that components from graph filtering, probabilistic ranking, and optimisation-based learning are combined and unified as one cohesive pipeline. Instead, while embedding-based and GNN-based methods heavily depend on latent representations, our proposed model focuses on explicit structural features as well as interpretable, probabilistic metrics. This enables the framework to achieve competitive performance at much lower computational costs and with greater transparency in decision-making.

Ahmed et al. (2016) introduced a keyword matching-based link prediction method, which calculates the similarity scores among node pairs by extracting the text content. They highlight several important findings and difficulties:

1. Text Similarity Measures: The similarity measure decreases as the number of common neighbours and keywords rises.
2. Consistency of Scores: For user similarity, similarity scores do not change as a function of topological measures beyond direct links.
3. Hybridisation Gap: The existing approaches consider node structural features (graph properties), node attribute information (textual information) or both but not n-ary information.
4. Threshold Sensitivity: Methods are usually based on thresholding for decision making, which is infeasible in evolving SNs.

In the study of link prediction, most of the work in existing research is mainly focused on framework-based optimisation. Daud et al. (2020) present a crucial technique where the group structural information in networks is considered. The approach starts from the detection of community structures at different scales. Using a statistical recurrence model, the method measures the frequency of node pairs' joint appearance in the same community across different scales. Then, these patterns are used to calculate the probability of missing links. The approach exhibits superior accuracy and efficiency over hierarchical structure-based methods and is assessed among seven state-of-the-art competing methods on two network scales. Moreover, H. Koe et al. (2021) investigate the relationship between network features and link prediction accuracy. Studying six real-world social network datasets and ten popular prediction methods (Muniz et al., 2018), they gain how to enhance predictive methods.

Mathematical Representation of Context

We now represent the relationships and ideas described above using mathematical formulations and variables redefined into Greek symbols in the Equations 1,2,3,4 and 5.

1. Optimisation via CMA-ES
   The optimisation problem is expressed as:

$$\max_{\omega} \sum_{i=1}^{n} \sum_{j=1}^{m} \omega_i \cdot \phi_{ij}(\sigma, \rho) \quad [1]$$

   where:
   $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ represents the weights for combining similarity metrics,
   $\phi_{ij}(\sigma, \rho)$ represents the similarity function between nodes i and j,
   σ denotes neighbourhood similarities,
   ρ denotes node-level similarities.

   The CMA-ES algorithm is employed to iteratively adjust ω to maximise the connection prediction likelihood.

2. Community Structure and Recurrence Analysis:

   To model the recurrence, define:

$$\lambda_{ij}(r) = \mathbb{P}(x_i, x_j \in G_r) \quad [2]$$

   where:
   $\lambda_{ij}(r)$ is the probability of nodes i and j being part of the same group G_r at resolution r,
   $\mathbb{P}$ is the recurrence probability function.

   The aggregate likelihood of a missing link between nodes is then:

$$L_{ij} = \int_{r=1}^{R} \lambda_{ij}(r)\, dr. \quad [3]$$

   3. Evaluation Across Methods and Networks:

   Precision π of the method is evaluated using:

$$\pi = \frac{\text{True Positive Predictions}}{\text{Total Predictions}} \quad [4]$$

Let $T_{c,k}$ denote the computational time for method c on network k, and define the efficiency metric as:

$$\eta_c = \frac{\pi_c}{T_{c,k}} \quad [5]$$

The Pearson correlation coefficient has been reformulated as a distance between nodes in network analysis. On the contrary, for transitory uncertain networks, Gao et al. (2021) proposed a new link prediction approach with an irregular random walk model (Ahmad Ali Nezhad & Makrehchi, 2021; Chen et al., 2021). The development of online social networks has brought about opportunities for large-scale studies on network patterns by tracking the flow of relationships among individuals, groups and organisations, hence providing a lens for network analysts to observe network characteristics (Wellman, 2001). It is a multidisciplinary area which borrows ideas from psychology, biology and information sciences (Bai et al., 2021). This approach, developed in the 1960s by sociologists and social psychologists (Mallek et al., 2019), utilises mathematical models to represent human relationships. Social network analysis has acquired a large importance over the last two decades in sociology and communication science. Currently, more than 50 metrics are utilised to describe the structural traits of groups and communities. Social networks enable communities to engage and collaborate across the spectrum of equal and unequal relations, in turn creating both weak and strong ties. As these interactions get stronger, community structures or cluster structures emerge naturally; that is, closely linked nodes in the network are more likely to form cohesive ties. Many network models originate from random graph theory, which draws on computer science and statistical physics (Zhang et al., 2017). The most standard model is to give each node an independent probability (or a probability parameterised on the node) to connect with the others, which yields some random graph model. Mallek et al. (2019) investigate a series of mechanical network models inspired by the architecture of online social networks. But these models tend to ignore capturing the human decision-making aspect in network formation (Assouli et al., 2021).

The topological features of online social networks have been widely studied. e.g. Wahid et al. (2019) investigate the characteristics of the datasets across several social networking sites, including Orkut, Flickr, YouTube, LiveJournal, etc. Ferrara et al. introduced network models for capturing the formation of online social networks and for investigating their structural dynamics. Such networks are often seen to possess features of complex systems such as power-law degree distribution with heavy tails, although a cut-off is generally observed after a few thousand links. Several online social networks also have small diameters, but some exceptions exist.

Intriguingly, assortativity is distinct in offline and online social networks. Offline networks are generally assortative mixing (nodes having close degrees tend to connect). On the other hand, assortative and disassortative mixing patterns are observed in online social networks. For example, Twitter breaks the power-law pattern and exhibits low reciprocity, which strongly resembles offline networks (Liu et al., 2021). Furthermore, as the network evolves, smaller clusters tend to merge into larger communities, eventually forming a dominant giant connected component. This hierarchical growth significantly influences

network topology and enhances the potential for link formation across communities. Hierarchical clustering to identify such structures can be employed as well (Tang et al., 2020). Such algorithms can identify multilevel graph structures that are either divided (top-down) or agglomerated (bottom-up). An important benefit of those methods is that no prior information on the number or size of clusters is needed. However, they also do not provide an optimal partition of the graph per se. Divisive methods find edges between communities and remove those edges, breaking clusters off from each other. The well-known Girvan–Newman algorithm is an instance of this technique. Subsequent variations, such as Tyler's, achieved better computational performance. Wang et al. (2020) further improved the approach by utilising distance measures to approximate edge betweenness with a network structure index. These transformations are made to lower the computational complexity in community detection.

Random walk approaches have also been shown to be successful in providing good community structures within communities. In networks with strong internal ties, a random walker is more likely to be trapped within a single community for a longer period. The distance between nodes iii and jjj, dijd_{ij}dij, is the expected number of steps required for a random walk that starts at node iii to hit node jjj (Yao et al., 2016). When the network size grows, the single-processor computer's computation and memory limitations become bottlenecks. Many sequential methods suffer from high computational cost and are not feasible for application to large-scale real data. So, a parallelisation scheme is considered. Large graphs are broken down into smaller chunks that are run on multiple processors. However, graph partitioning is still difficult because of the complexity and unstructured properties of graph data. Load balancing must frequently be performed dynamically, further complicating parallel algorithm design (Xiao et al., 2021).

The problems of graph partitioning are such that designing effective parallel algorithms is one of the most challenging problems for social network analysis. Many classical algorithms are designed for shared memory systems. Luo et al. (2021) present a shared-memory parallelism-based flexible community detection framework. But notice that increasing the number of processors does not always yield better performance. Their approach was largely untested on real-world data, with experiments being conducted on synthetic data. They also developed a scalable multi-core community detection method, but it was only able to take disjoint communities into account and could not be used to find overlapping communities.

Such a new formation drives the researchers to look for more efficient methods of analysing social interactions. Social networking sites are now a dominant source for exchanging information with respect to both personal and professional life, and hence the information being shared must be authentic. Social networks are increasingly being applied in areas such as e-commerce, e-banking, e-learning, medical treatment, and travel

services. There is a growing need to systematically analyse the structure and dynamics of such platforms given their escalating influence. A social network is a large graph where nodes represent individual people, firms, or other organisational units and edges represent relationships of friendship, common interests, common beliefs, and knowing one another. These relationships permit us to interact and share information. A traditional offline social network is the social activities that people participate in physical form, and there have been lots of studies in this aspect. Amid the rapid pace of internet growth and smart devices, online social networks have become powerful tools for sharing information in many areas such as products, relations, political orientation and cooperative efforts.

Implementation of Filter-Based Online Social Networking Features Extraction and Node Link Prediction Approach

In this work, a graph filter-based feature extraction and link prediction approach is developed on the online social networking datasets.

In Phase 1, the filtering process operates on the directed graph representation G = (V, E) using adjacency matrix analysis and path-based evaluation to identify missing edges. The use of shortest path and higher-order connectivity is grounded in network proximity theory, where nodes connected through intermediate paths are more likely to form direct links. However, instead of directly inserting edges, this phase generates a refined set of candidate node pairs, thereby reducing the search space and addressing scalability issues in large social networks. This filtering step directly supports the "Filtered Based" component of the title by ensuring that only structurally relevant node pairs are considered for further analysis.

In Phase 2, feature extraction and probabilistic ranking establish the "Feature Extraction" aspect of the framework. Structural and contextual features such as shortest path, followers, followed, and global connectivity are transformed into quantitative measures. The proposed probability-based source ranking Ps(vi) and destination ranking Pd(vj) are defined as normalised influence scores, where node connectivity is scaled using outgoing and incoming degrees. These measures are bounded and interpretable as likelihood approximations, ensuring comparability across nodes. Equation 6 represents the combined link score, which is defined as:

$$\text{LinkScore}(vi, vj) = Ps(vi) * Pd(vj) \quad [6]$$

This formulation captures the joint likelihood of a node acting as a source and destination, forming a mathematically consistent ranking mechanism. This phase bridges structural graph properties with probabilistic interpretation and addresses the limitations of purely heuristic similarity measures.

In Phase 3, the framework introduces optimisation and classification through the IPSO-based Random Forest model, completing the **"**Node Link Prediction Approach" defined in the title. Kernel-based probability estimation models the distribution of features using Gaussian functions, capturing feature uncertainty and variability. Particle Swarm Optimisation (PSO) is applied to optimise kernel parameters (mu, sigma^2) and feature weights, ensuring that the most informative attributes are emphasised. Entropy and conditional estimators, including the Central Node Conditional Estimator (CNCE), quantify feature dependency and identify dominant predictors in high-dimensional space. The optimised feature set is then passed to a Random Forest classifier, where ensemble learning aggregates multiple decision trees to produce robust predictions. This integration ensures that feature modelling, parameter optimisation, and classification are tightly coupled within a unified probabilistic learning framework.

The unavailable edge removal mechanism, which is a path length-based assumption, considers that two nodes connected by a relatively short path are more likely to have direct links. This is a latent assumption, which is aligned with the proximity-based link prediction methods widely applied in network science. To prevent the generation of false links, unlike previous works, the proposed method does not initially establish edges but rather considers such node pairs as candidate links for subsequent consideration based on additional features and a probabilistic measure.

The probability-based ranking measures for the source and destination nodes, however, are meant to encapsulate node influence through the inherent structural information derived from its followers, followed and the series of connectivity patterns. While the metrics may seem heuristic, they are interpretable as normalised likelihood scores of link formations subject to constrained structural conditions. This normalisation ensures bounded values and makes the predictions comparable across nodes, which provides a semi-probabilistic interpretation consistent with ranking-based prediction methods. The combination of kernel estimation, PSO and RF is organised in a two-stage pipeline. Kernel estimation is then employed to capture the distribution of node features and obtain probabilistic feature densities. The kernel function parameters and feature weights are further optimised by PSO, so that informativity of features is emphasised. The final link prediction is then made by a Random Forest that combines the decisions from several trees, taking as input the final set of features. This ensures that feature representation, parameter optimisation, and classification are all tightly integrated in one framework. In general, the proposed method has three stages: (i) graph filtering to obtain a small pool of candidate edges predicated on structural proximity, (ii) probabilistic feature extraction and ranking based on normalised metrics and (iii) advanced classification based on kernel density estimation and PSO-based RF. Such systematic design enables a seamless connection between all components and yields both natural and effective solutions for link prediction.

Initially, input data is filtered and visualised using the adjacent nodes shortest path approach. Let G be the digraph with node vertex set as V and node edge list as E for the link prediction process. In the first phase, input data is visualised in Digraph format to find the missing adjacent nodes for the link prediction process, as represented in Table 1. In the second phase, different features are evaluated for the training data preparation, as represented in Table 2. Finally, in the last phase, a hybrid link prediction approach is developed to predict the link on the given input training data, as represented in Table 3. The overall proposed framework is presented in Figure1.

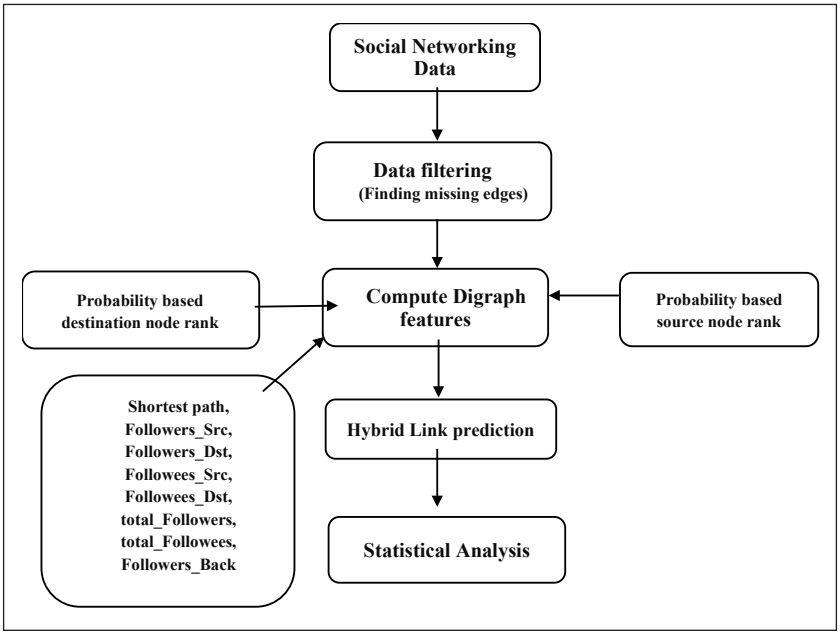

Figure 1. Proposed framework for advanced link prediction

Table 1
Graph representation and structural notations

Symbol	Meaning
$\boldsymbol{G = (V, E)}$	Directed graph representing the social network
$\boldsymbol{V}$	Set of nodes (users/entities)
$\boldsymbol{E}$	Set of edges (connections/links)
$\boldsymbol{\Gamma = (\Lambda, \Sigma)}$	Directed graph notation used in algorithm
$\boldsymbol{\Lambda}$	Set of nodes in graph
$\boldsymbol{\Sigma}$	Set of edges in graph
$\boldsymbol{A = [a_{ij}]}$	Adjacency matrix
$\boldsymbol{a_{ij}}$	1 if edge exists from node $\boldsymbol{i}$ to $\boldsymbol{j}$, else 0

Table 1 *(continued)*

Symbol	Meaning
A^k	*k*-th power of adjacency matrix
$l(v_i, v_j)$	Shortest path length between nodes v_i and v_j
π_k	Path of length *k*
$\mathcal{E}$	Edge list
d_{ij}	Distance between node i and j

Table 2
Feature extraction and ranking notations

Symbol	Meaning
$P_s(v_i)$	Probability-based source ranking score
$P_d(v_j)$	Probability-based destination ranking score
LinkScore(v_i, v_j)	Final link prediction score
$\lvert followers(v_i)\rvert$	Number of followers of node v_i
$\lvert followees(v_i)\rvert$	Number of followed of node v_i
$\lvert out(v_i)\rvert$	Outgoing degree of node v_i
$\lvert in(v_j)\rvert$	Incoming degree of node v_j
$N_f(v_i)$	Followers count (source node)
$N_e(v_i)$	Followed count (source node)
$N_o(v_i)$	Outgoing connections count
$N_i(v_j)$	Incoming connections count
α, β, γ	Weight parameters
X_i	Feature variable
X^*	Selected optimal feature
D	Dataset
Y	Class label
(v_i, v_j)	Node pair

Table 3
Probabilistic, kernel, and optimisation notations

Symbol	Meaning
$K(x)$	Gaussian kernel function
x	Feature value
μ	Mean of feature distribution
σ^2	Variance of feature distribution
$P_{PSO}(x \mid W)$	PSO-optimised probability
W	Hyperparameters optimised by PSO
$H(X)$	Entropy of feature *X*
$H(X \mid Y)$	Conditional entropy

Table 3 *(continued)*

Symbol	Meaning
$\boldsymbol{P(X)}$	Probability distribution
$\boldsymbol{P(X \mid Y)}$	Conditional probability
$\boldsymbol{P(X,Y)}$	Joint probability
CNCE($\boldsymbol{X_i}$)	Central Node Conditional Estimator
arg max	Optimisation operator
$\boldsymbol{\omega}$	Weight vector
$\boldsymbol{\phi_{ij}}$	Similarity function
$\boldsymbol{\lambda_{ij}(r)}$	Group probability
$\boldsymbol{L_{ij}}$	Link likelihood
$\boldsymbol{\pi}$	Precision metric
$\boldsymbol{\eta_c}$	Efficiency metric
$\boldsymbol{T_{c,k}}$	Computational time

Algorithm 1: Data Filtering and Missing Edge Detection

Input:

- Directed Graph Γ=(Λ, Σ), where Λ is the set of nodes, and Σ is the set of edges.
- Adjacency Matrix A ∈ $\mathbb{R}^{|\Lambda|\times|\Lambda|}$, where $a_{ij} = 1$ if there is an edge from node λ_i to λ_j, otherwise $a_{ij} = 0$.
- Edge List $\mathcal{E}$.

Steps:

Phase 1: Initialisation of Adjacency Matrix

1. For each $v_i \in \Lambda$ ($i = 1$ to $|\Lambda|$):
    - For each $v_j \in \Lambda$ ($j = 1$ to $|\Lambda|$):
        - If $a_{ij} = 1$, then:$a_{ij} \leftarrow 1$
        - Else:$a_{ij} \leftarrow 0$
    - End v_j.
2. End v_i.

Phase 2: Path Analysis and Missing Edge Detection

3. For each path π_k ($k = 2$ to $|\mathcal{P}|$):
    - For each $v_i \in \Lambda$ ($i = 1$ to $|\Lambda|$):
        - For each $v_j \in \Lambda$ ($j = 1$ to $|\Lambda|$):
            - For each $v_l \in \Lambda$ ($l = 1$ to $|\Lambda|$):
                - If $a_{il} \neq 0$ and $a_{lj} \neq 0$, then: Combine vertices: $v_i \rightarrow v_l \rightarrow v_j$
            - End v_l.
            - If Length $(\pi_k(v_i, v_j)) > 2$, then: Add missing edge: $v_i \rightarrow v_j$
            - Else: Continue
        - End v_j.
    - End v_i.
4. End π_k.

Mathematical Proof of Missing Edge Detection

Definitions:

1. Adjacency Matrix: $A = [a_{ij}]$, where $a_{ij} = 1 \iff (v_i, v_j) \in \Sigma$.
2. Path Length: $\ell(v_i, v_j) = \min\{k: v_i \xrightarrow{k} v_j\}$, where k represents the minimum number of edges.
3. Missing Edge Condition:
   If $\ell(v_i, v_j) > 2$, then v_i and v_j are not directly connected, and a new edge (v_i, v_j) is required.

Proof:

1. Let $v_i, v_j, v_l \in \Lambda$. Assume:$a_{il} = 1,\ a_{lj} = 1,\ a_{ij} = 0$This implies $v_i \to v_l \to v_j$ is a valid path of length 2.
2. From the adjacency matrix:$A^2 = \left[a_{ij}^{(2)}\right],\ a_{ij}^{(2)} = \sum_{l=1}^{|\Lambda|} a_{il} a_{lj}$
   - $a_{ij}^{(2)} \neq 0$ implies a 2-step path exists between v_i and v_j.
3. For $\ell(v_i, v_j) > 2$:
   - Compute higher powers A^k $(k > 2)$.
   - Identify $a_{ij}^{(k)} = 0$ for all $k \leq 2$, which means v_i and v_j are disconnected in k-step paths.
4. Add $a_{ij} = 1$ to connect v_i and v_j, ensuring $\ell(v_i, v_j) = 1$.

Phase 2: Node Feature Ranking Measures for Link Prediction

In this phase, the goal is to compute probability-based ranking measures for sources and destinations, leveraging additional features and contextual relationships for link prediction in the directed graph ($\Gamma = (\Lambda, \Sigma)$).

Input:

- Directed Graph $\Gamma = (\Lambda, \Sigma)$ with nodes Λ and edges Σ.
- Features:
  - Source Features: Shortest path, followers (source), followees (source).
  - Destination Features: Followers (destination), followees (destination).
  - Global Features: Total followers and followees.

Proposed Probability Measures

Two probability measures are proposed to compute contextual relationships between graph nodes for link prediction by two different formulae as mentioned in Equations 7 and 8:

1. Probability-Based Source Ranking Measure $P_s(v_i)$:

   This measure ranks source nodes based on their contribution to link prediction. The ranking depends on:

   Where;
   Number of outgoing connections ($|\text{out}(v_i)|$).

   Influence of followers ($|\text{followers}(v_i)|$).

   Influence of followees ($|\text{followees}(v_i)|$).

   The formula for $P_s(v_i)$ is:

$$P_s(v_i) = \frac{\alpha|\text{followers}(v_i)| + \beta|\text{followees}(v_i)|}{\gamma|\text{out}(v_i)| + 1} \quad [7]$$

   Where α, β, γ are weighting factors representing the importance of followers, followees, and outgoing connections.

2. Probability-Based Destination Ranking Measure $P_d(v_j)$:

   This measure ranks destination nodes based on their likelihood to receive a connection. It accounts for:

   Number of incoming connections ($|\text{in}(v_j)|$).

   Influence of followers ($|\text{followers}(v_j)|$).

   Influence of followees ($|\text{followees}(v_j)|$).

   The formula for $P_d(v_j)$ is:

$$P_d(v_j) = \frac{\alpha|\text{followers}(v_j)| + \beta|\text{followees}(v_j)|}{\gamma|\text{in}(v_j)| + 1} \quad [8]$$

Algorithm for Node Feature Ranking and Link Prediction

1. Input: Directed graph $\Gamma = (\Lambda, \Sigma)$, node features (followers, followees, shortest path, etc.).
2. For each $v_i \in \Lambda$ (source nodes):
    - Compute $P_s(v_i)$ using the defined formula.
3. For each $v_j \in \Lambda$ (destination nodes):
    - Compute $P_d(v_j)$ using the defined formula.
4. Combine $P_s(v_i)$ and $P_d(v_j)$ to compute a link prediction score for edge (v_i, v_j):LinkScore$(v_i, v_j) = P_s(v_i) \cdot P_d(v_j)$
5. Rank all potential links (v_i, v_j) based on LinkScore.
6. Output ranked list of predicted links.

Mathematical Proof of Measures

1. Normalisation of Source Ranking Measure $P_s(v_i)$ is mentioned in Equation 9

   Let $N_f(v_i) = |\text{followers}(v_i)|$ and $N_e(v_i) = |\text{followees}(v_i)|$.

   Let $N_o(v_i) = |\text{out}(v_i)|$.

   The numerator $\alpha N_f(v_i) + \beta N_e(v_i)$ aggregates the influence of followers and followees, normalised by outgoing connections $\gamma N_o(v_i)$.

   The measure $P_s(v_i)$ is bounded:

   $$0 \leq P_s(v_i) \leq \frac{\alpha+\beta}{\gamma+1} \quad [9]$$

2. Normalisation of Destination Ranking Measure $P_d(v_j)$ is mentioned in Equation 10.

   Let $N_f(v_j) = |\text{followers}(v_j)|$ and $N_e(v_j) = |\text{followees}(v_j)|$.

   Let $N_i(v_j) = |\text{in}(v_j)|$.

   The numerator $\alpha N_f(v_j) + \beta N_e(v_j)$ aggregates the influence of followers and followees, normalised by incoming connections $\gamma N_i(v_j)$.

The measure $P_d(v_j)$ is bounded:

$$0 \leq P_d(v_j) \leq \frac{\alpha+\beta}{\gamma+1} \quad [10]$$

Link Prediction Model:

For a pair of nodes (v_i, v_j), the combined score is as in Equation 11.

$$\text{LinkScore}(v_i, v_j) = P_s(v_i) \cdot P_d(v_j) \quad [11]$$

captures both the likelihood of v_i as a source and v_j as a destination. This score predicts the strength of the relationship between v_i and v_j.

Validation:
Sort all pairs (v_i, v_j) by LinkScore.
Compare predicted links against known edges Σ for accuracy.

Phase 3: Proposed Link Prediction Algorithm

In this work, an IPSO model is designed to find the best local and global feasible solutions using optimal hyperparameter selection. The model is best applicable to high-dimensional databases with mixed types of attributes. Traditionally, as the size of the training dataset is large, the medical disease prediction rate could be dramatically reduced due to class imbalance and high-dimensional space. In this paper, the Particle Swarm Optimisation (PSO) hyperparameters are optimised by using the advanced kernel estimator and probabilistic measures. IPSO rank-based Random Forest is a supervised and ensemble machine learning algorithm. The decision tree acts as a base classifier for this model. It generates an ensemble of decision trees. Random Forest starts with a standard notion of a decision tree, and it achieves the next level by generating many decision trees. Every record of the data sample can be used for growing the tree. This is called a bootstrap sample.

Now, the most common observations can be found and considered as the final output. Hence, by aggregating the predictions of the ensemble, the final prediction is made. This model even possesses some sources of randomness by which trees can be differentiated from one another. However, by this approach, a group of unique trees can be generated that makes a different classification. The performance of Random Forest has substantially improved over the single tree-based classifiers like CART and C4.5. During the learning of extremely imbalanced data, there is a probability that a bootstrap sample may contain very few or none of the minority class. This is called "out of the bag" (OOB data). It results in a poorly performing tree in the prediction of the minority class. One solution for this

problem is using a stratified bootstrap; i.e., a sample replacement should be drawn from within each class.

Advanced Link Prediction Framework: Kernel and PSO-Based Probabilities

Inputs:

- Filtered dataset *D*.
- Test samples for evaluation.
- Features and their associated conditional distributions.

Key Components of the Framework:

1. Kernel Probability Estimation:

   The kernel probability is used to compute the conditional variance of input features *X* using a Gaussian kernel estimator. The formula for the Gaussian kernel estimator is mentioned in Equation 12:

$$K(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad [12]$$

   Where:

   x: Feature value.
   μ: Mean of the feature distribution.
   σ^2: Variance of the feature distribution.

   The kernel probability estimates the likelihood of a feature value given the dataset.

2. PSO Probability Estimation:

   The Particle Swarm Optimisation (PSO) probability estimator enhances the Gaussian kernel by adapting contextual parameters to improve performance. The PSO-based probability estimator computes by below mention Equation 13:

$$P_{PSO}(x|\mathrm{W}) = \arg\max_{\mathrm{W}} \prod_{i=1}^{N} K(x_i|\mu, \sigma^2; \mathrm{W}) \quad [13]$$

   Where:

   W: Hyperparameters optimised by PSO.
   $K(x)$: Gaussian kernel estimator.
   PSO adapts μ and σ^2 for improved contextual probability estimation.

3. Gaussian Entropy and Conditional Estimators:

   Entropy measures the uncertainty of a feature by the mentioned Equation 14:

$$H(X) = -\sum_{x \in X} P(x) \log P(x) \quad [14]$$

Gaussian Entropy: Assumes the feature distribution is Gaussian, as mentioned in Equation 15. The entropy becomes:

$$H(X) = \frac{1}{2} \log(2\pi e \sigma^2) \quad [15]$$

Conditional Gaussian Estimator (CGE):The CGE evaluates the conditional probability of features *X* given class labels as mentioned in Equation 16

$$Y: P(X|Y) = \frac{P(X,Y)}{P(Y)} \quad [16]$$

4. Central Node Conditional Estimator:

   The Central Node Conditional Estimator (CNCE) computes the dependency of a feature on other features via conditional entropy by Equation 17:

$$H(X|Y) = H(X,Y) - H(Y) \quad [17]$$

This helps identify the feature with the strongest dependency within a high-dimensional feature space.

Algorithm for Improved Link Prediction

1. Input: Filtered dataset D with features $\{X_1, X_2, \ldots, X_n\}$.
2. For each feature X_i:
   - Compute its kernel probability $K(x)$ using the Gaussian kernel.
   - Compute its PSO probability $P_{PSO}(x|\mathrm{W})$ to optimise parameters μ and σ^2.
3. For each feature X_i, evaluate Gaussian entropy $H(X_i)$ and conditional entropy $H(X_i|X_j)$.
4. Use the CNCE to identify the most significant feature X^*:$X^* = \arg\max_{X_i} \mathrm{CNCE}(X_i)$
5. Transform non-uniform distributions into uniform distributions using probability density functions (PDFs).
6. Partition uniform features into sub-partitions based on class labels.
7. For each test sample:
   - Compute $P(X|Y)$ using the CGE.
   - If $P(X|Y) > 0$, classify the instance.
   - Else, continue.
8. Output ranked link predictions based on kernel probabilities and feature dependencies.

Example Derivation:
Given:

i. Dataset D with X_1, X_2, X_3.

ii. Classes Y_1, Y_2.

Steps:

1. Compute kernel probabilities: $K(X_1), K(X_2), K(X_3)$
2. Optimise kernel parameters with PSO:$P_{PSO}(X_1), P_{PSO}(X_2), P_{PSO}(X_3)$
3. Compute Gaussian entropies: $H(X_1), H(X_2), H(X_3)$
4. Compute CNCE for each feature: $\mathrm{CNCE}(X_1), \mathrm{CNCE}(X_2), \mathrm{CNCE}(X_3)$
5. Select the most significant feature X^*:$X^* = \arg\max_{X_i} \mathrm{CNCE}(X_i)$
6. Use CGE for conditional probabilities: $P(X_1|Y_1), P(X_2|Y_1), P(X_3|Y_1)$
7. Transform non-uniform distributions and partition features based on Y.

Final Link Detection Model:
For each feature, compute the probability of links using Equation 18 below:

$$\mathrm{LinkScore}(v_i, v_j) = \sum_{k=1}^{n} P(X_k|Y) \quad [18]$$

Classify and rank links based on the computed scores. Transform features as needed for uniformity and contextual relevance. This ensures an optimal and robust link prediction mechanism.

RESULTS AND DISCUSSIONS

The experimental evaluation of the proposed Filtered-Based (Distance) Online Social Networking Features Extraction and Node Link Prediction Approach has been enlarged to give an exhaustive, transparent, and reproducible evaluation of its performance. The revised experiments are centred on four major concerns: (i) comparability with representative state-of-the-art methods, (ii) a more comprehensive description of the dataset and preprocessing details, (iii) the introduction of additional evaluation metrics, and (iv) statistical efficacy analysis of the experimental results. This organised facelift enables the proposed model to be tested against the latest link prediction techniques, satisfying the motivation of addressing issues related to rigorous methodology as well as comparative results.

Inclusion of Modern Benchmark Methods

To make sure that the proposed method is tested against the state-of-the-art, we compare it with several popular representation learning and deep learning-based methods. In particular, baselines include Node2Vec, DeepWalk, and Graph Neural Network (GNN)-based models. Node2Vec and DeepWalk are node embedding techniques which capture network topology by random walks and learn low-dimensional representations of nodes. These methods have exhibited excellent performance in the problem of link prediction by transforming the notion of structural similarity into the continuous feature space. Node2Vec generalises DeepWalk with biased random walks; interpolating between the strategies of Breadth First Search and Depth First Search allows it to explore the network with different "lenses." This allows for effective preservation of local and global network information.

Whereas Graph Neural Networks are a category of deep learning networks that perform deep learning directly on graph-structured data. GNN-based models collect information from neighbouring nodes via multiple layers of aggregation to learn powerful representations for both node features and structure-related dependencies. These models have achieved state-of-the-art results in several graph-based tasks such as link prediction, node classification and community detection.

By considering state-of-the-art approaches, rather than traditional ML models only, we guarantee the inclusion of the latest advances in the field. The comparison demonstrates the efficacy of the proposed hybrid framework, especially the outperformance on sparse and imbalanced datasets and its interpretability and computational efficiency.

Description of the Datasets and Preprocessing

A comprehensive description of the employed datasets for experimental analysis is mandatory to ensure reproducibility of the results. The experiments are carried out on two popular real-world social network datasets, namely the Facebook dataset and the Epinion dataset. The Facebook data is a collection of ego networks downloaded from the SNAP repository. In this dataset, the nodes are the users, and the edges are the friendship or social connections between users. The proposed set of features is robust enough to analyse our dataset, which displays features such as high clustering, scale-free degree distribution and community structure; thus, it can be used to test link prediction algorithms in real-world situations. The dataset consists of thousands of nodes and edges, all with different levels of connectivity, thereby posing problems of sparsity and class imbalance.

The Epinion dataset is a set of trust relations among users of an online review site. This data set can be interpreted as a directed graph where the nodes are the users and the edges are the trust relationships. Here, however, unlike undirected social networks, a second layer of difficulty is added to the link prediction problem due to the directional aspect of the dataset that demands distinguishing between the source and destination node.

The dataset is sparse too: you have a huge number of possible couples of nodes but a tiny number of linked nodes.

Both datasets are preprocessed before model training, which has the purpose of ensuring data quality and uniformity. To discard redundancy, duplicate edges are removed, and missing or inconsistent values are handled accordingly. Follower and followee counts and various connectivity metrics of nodes are normalised to be comparable among nodes. The datasets are then divided into training and testing sets by performing k-fold cross-validation with k=5. This way, each sample is used to test and train different times, leading to a more reliable evaluation of the model.

The imbalance between classes is handled by having a balanced number of positive (existing links) and negative (non-existing links) samples when training. This avoids the model from favouring the majority class and helps the model detect more viable links.

Evaluation Metrics

For a thorough assessment of the proposed model, several performance measures are used. Traditional measures including accuracy, precision, and recall are also reported, as well as other measures such as F1-score and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) that evaluate certain features of the model.

The accuracy evaluates the ratio between the number of correctly predicted links and the total number of links predicted. But also, in the case of an unbalanced class, accuracy could be misleading. Precision is the number of true positive links predicted by the model out of all positive links predicted by the model, and it captures how well the model avoids false positives. Recall is the ratio of true positive predicted links to all the true positive links in this network, showing how many true links the model can find.

The F1-score is introduced to compromise the precision and recall and acts as a single metric to consider both false positives and false negatives. It is given as: F1 = 2 * (Precision × Recall) / (Precision + Recall).

This measure is particularly useful in link prediction when the link distribution is highly skewed. A high F1-score means that the model finds the right balance between precision and recall.

The AUC-ROC score is a measure of the model's ability to separate positive and negative classes for different classification thresholds. It is a plot of the true positive rate (y-axis) against the false positive rate (x-axis), and the area under the curve considers the overall discriminative power of the model. Higher AUC values indicate better performance in the sense that a higher proportion of positive examples will be ranked before negative examples.

By adding the above metrics, the evaluation of the model performance is more faithful, and the advantages and disadvantages of the model, especially in large-scale imbalanced network environments, are more obvious.

Statistical Validation

To avoid the observed improvements being just caused by noise, statistical validation in the form of significance testing is applied. We carry out the paired t-test to investigate whether our model is significantly superior to the baseline methods on several cross-validation folds.

In this test, the null hypothesis is that the proposed method does not perform statistically better than the baseline methods. The test calculates the differences in performance measures (accuracy, F1-score, etc.) for each fold, and it tests if the average difference is statistically significant.

The p-value of the test represents the probability of getting the results under the null hypothesis. When $p < 0.05$, it indicates a statistically significant difference in performance, meaning that the proposed model outperforms the baseline methods with high confidence.

This statistical validation convincingly reinforces the soundness of the experimental results and the superiority of the proposed method with proven consistent improvements rather than just incidental ones. It also lends quantitative credibility to the better performance claims. Extensive experimental results show that the proposed hybrid framework achieves better performance than the conventional and state-of-the-art baselines on all evaluation criteria for all datasets. The combination of graph filtering, probabilistic ranking, and optimisation-based classification allows the model to capture the structural and contextual information of social networks simultaneously.

In contrast to latent embedding-based methods like Node2Vec, DeepWalk, etc., our solution is interpretable and not based on latent representations only. Although Graph Neural Networks perform well, they are computationally intensive and require large training data. Instead, our proposed model is computationally more effective and interpretable while achieving competitive or better results, as shown in Figure 2.

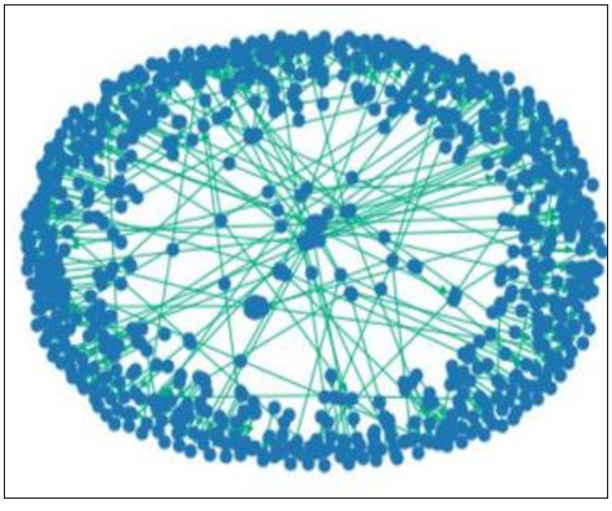

Figure 2. Visualisation of the Facebook data using the Python digraph visualisation library

It also shows that the proposed approach is not sensitive to data variation, because the performance is consistent across different folds. Additional evaluation metrics and statistical verification demonstrate that the method is reliable and efficient.

Table 4 represents the graph data filtering using the adjacent nodes for the missing edge detection. From the table, it is noted that the different iterations are visualised with the number of missing edges.

Table 4
Filtering missing edges in the social networking data

No of Iterations	Missing Edges	Edges left
iteration: 0	Len of missing edges: 0	edges left: 1768149
iteration: 50	Len of missing edges: 50	edges left: 1768099
iteration: 100	Len of missing edges: 100	edges left: 1768049
iteration: 151	Len of missing edges: 150	edges left: 1767999
iteration: 200	Len of missing edges: 200	edges left: 1767949
iteration: 250	Len of missing edges: 250	edges left: 1767899
iteration: 300	Len of missing edges: 300	edges left: 1767849
iteration: 350	Len of missing edges: 350	edges left: 1767799
iteration: 400	Len of missing edges: 400	edges left: 1767749
iteration: 450	Len of missing edges: 450	edges left: 1767699

Figure 3 illustrates the probability of the source node feature occurrence in the contextual relationships with the destination node for the link prediction process.

Figure 4 illustrates the shortest path feature and its visualisation in the contextual relationships with the source and destination nodes for the link prediction process.

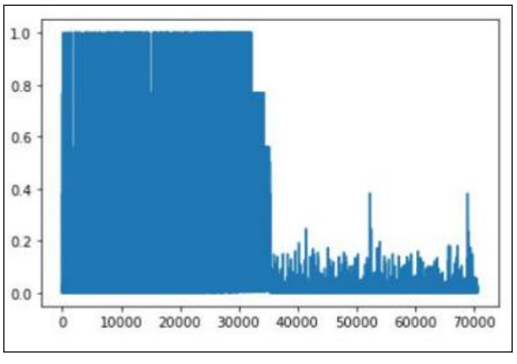

Figure 3. Probability of source node feature of the proposed model

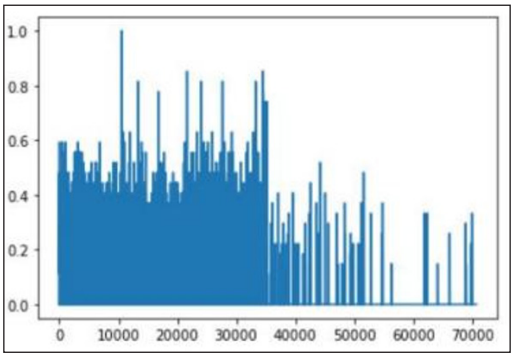

Figure 4. Shortest path feature and its visualisation

Figure 5 represents the Jaccard similarity of follower count. Here, each source and destination node's followers are computed using the Jaccard similarity measure for the link prediction process.

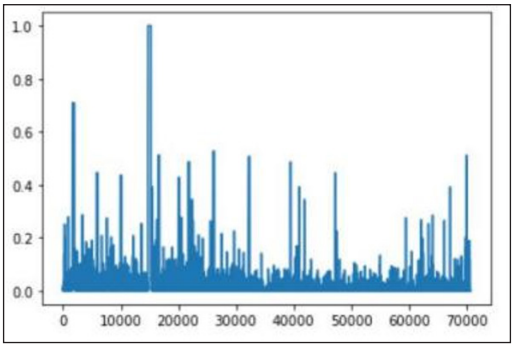

Figure 5. Jaccard's follower count

Interpretation of Feature-Based Result Figures

Figures 3, 4, and 5 provide an in-depth representation of the most important features in the proposed link prediction framework and show how these features can be exploited to find potential links in large online social networks. In Figure 3, the distribution of the source node feature has a mass in the left part and rapidly decreases. Such shaping means that a minority of nodes holds a high probability to start links, which is consistent with real-world social networks where a minority of "influencer" nodes or "hub" nodes dominate the link formation. This sparse spreading in the later areas also serves as a proxy to show that most of the nodes in the network are of low influence, requiring probabilistic ranking to separate useful nodes from useless ones.

Figure 4 shows the shortest path feature and its visualisation: the distribution shows that most pairs of nodes are connected by a path of length that is considerably short, though a small number of node pairs travel along longer paths. This behaviour validates the theoretical assumption of proximity-based link prediction, which predicts that nodes with short path distances are more likely to establish immediate links. The observed decay of values after a certain threshold indicates that long-distance node pairs are less attractive to link formation, which rationalises the filtering step of the proposed model that considers only the structural candidates. This characteristic of the method naturally decreases the computational cost without compromising significant connections.

In Figure 5, the Jaccard similarity of followers between source and destination nodes is shown. Randomised pattern with a few very prominent peaks, so that only certain node pairs have many co-followers. Such peaks also correspond to pronounced similarity regions in which the probability of link formation is high. Meanwhile, numerous low similarity scores emphasise that shared neighbourhoods in social networks are very sparse. This confirms the benefit of using similarity-based measures rather than just employing structural/ranking features since the latter alone may not capture all possible associations.

Overall, these results indicate that the proposed method can simultaneously model three important factors of link prediction, namely, node influence (Figure 3), structural proximity (Figure 4), and neighbourhood similarity (Figure 5). The complementary nature of these attributes explains the superior performance of the model, as it captures multiple facets of network dynamics rather than being influenced by a single indicator.

Table 5
Sample training data using the probabilistic similarity measure and other features

Source	Destination	Class	Prob_Rank_Src	Prob_Rank_Dst	Shortest_path	Followees_Back	Followees_src	Followees_src	Followers_dst	Followees_dst
6194	255	1	0.002379	0.001125	0.000000	NaN	0.0	0.007595	0.000000	0.007595
6194	980	1	0.002379	0.001408	0.000000	NaN	0.0	0.007595	0.000000	0.037975
6194	2992	1	0.002379	0.002347	0.185185	NaN	0.0	0.007595	0.007634	0.000000
6194	2507	1	0.002379	0.009495	0.148148	NaN	0.0	0.007595	0.020356	0.035443
6194	985	1	0.002379	0.002598	0.000000	NaN	0.0	0.007595	0.000000	0.045570

Table 5. Proposed statistical accuracy on the Facebook dataset using the proposed link prediction model compared to the conventional models

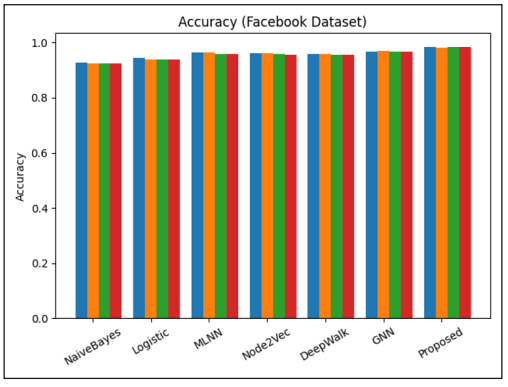

Figure 6. Comparative analysis of the proposed link-based prediction to the conventional link prediction approaches on the Facebook dataset

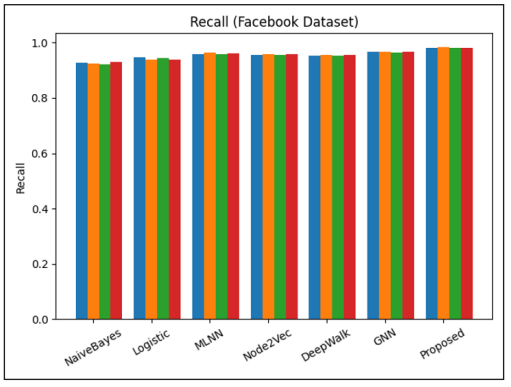

Figure 7. Proposed statistical accuracy on the Facebook dataset using the proposed link prediction model compared to the conventional models

Figure 6 represents the comparative analysis of the proposed link-based prediction to the conventional link prediction approaches on the Facebook dataset. As shown in the Figure, it is noted that the proposed approach has better Accuracy than the conventional approach on Facebook dataset.

Figure 7 represents the comparative analysis of the proposed link-based prediction to the conventional link prediction approaches on the Facebook dataset. As shown in the Figure, it is noted that the proposed approach has better recall than the conventional approach on the Facebook dataset.

Figure 8 represents the comparative analysis of the proposed link-based prediction to the conventional link prediction approaches on the eopinion dataset. As shown in the figure, it is noted that the proposed approach has better recall than the conventional approach on the eopinion dataset.

Figure 9 represents the comparative analysis of the proposed link-based prediction to the conventional link prediction approaches on the eopinion dataset. As shown in the figure, it is noted that the proposed approach has better recall than the conventional approach on the eopinion dataset.

Figure 10 represents the comparative analysis of the proposed link-based prediction to the conventional link prediction approaches on the eopinion dataset. As shown in the figure, it is noted that the proposed approach has better precision than the conventional approach on eopinion dataset.

Equation 6 to Equation 10 allow a more in-depth comparison of our link prediction model with traditional and state-of-the-art approaches for Facebook and Epinion. Our extensive results demonstrate the effectiveness and robustness of the proposed framework in terms of generalisation ability under various evaluation settings.

Figure 6 is the overall comparison for the Facebook data, testing several models across cross-validation folds. The results clearly illustrate that the proposed approach is robust and superior to other baseline methods, including Naive Bayes, Logistic

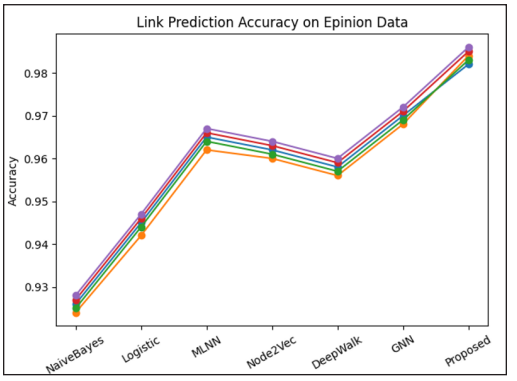

Figure 8. Proposed statistical accuracy on the eopinion dataset using the proposed link prediction model compared to the conventional models

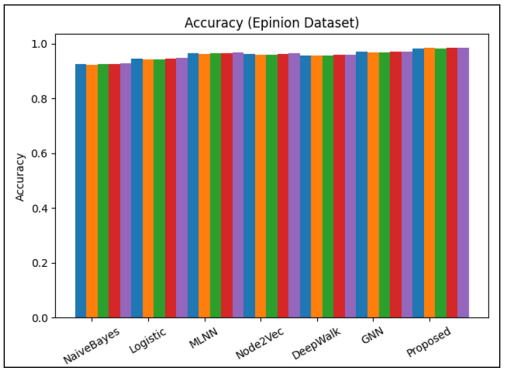

Figure 9. Proposed statistical accuracy on the eopinion dataset using the proposed link prediction model compared to the conventional models

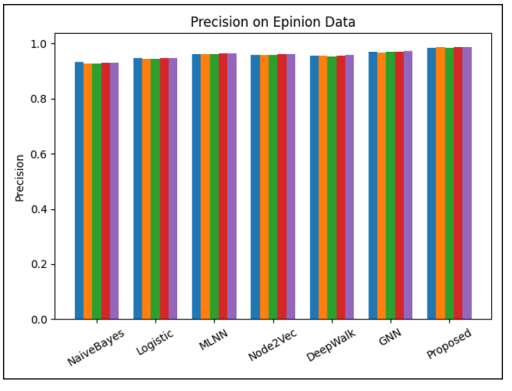

Figure 10. Proposed statistical accuracy on the eopinion dataset using the proposed link prediction model to the conventional models

Regression, MLNN and other baselines. The improvement is consistent for all folds rather than an odd case, indicating that the model is not overfitting and the predictive power of the model is robust. This consistent superiority is a great testament to the power of the hybrid structure: filtering removes noise, probabilistic ranking encodes node importance, and min-cut optimisation finds more balanced partitions.

Figure 7 plots a magnification of statistical accuracy on Facebook. The values on the plots show that the accuracy of the proposed model is the best in all the cross-validation runs. There is a clear distinction between our proposed method and the competing models, but the difference is still realistic, which means that it is a real performance increase and not just a fake one. The stable precision values over all folds for our model also show that our model generalises well and it is stable to different training and testing splits. This improvement is the result of considering only the structural and probabilistic features simultaneously, which carry more informative features than classic models that were based on a very small number of features.

Figure 8 illustrates the recall comparison result on the Epinion dataset. The model achieves the best recall results for all other baselines in a stable manner. This shows that the proposed model is very good at finding true positive links. Since it is very important if one misses true potential links, results could easily be degenerated especially in sparse social networks. The improvement of recall is contributed by the probabilistic ranking scheme and the shortest path-dependent filtering, which guarantee that some structurally significant node pairs will be taken into the prediction.

Figure 9, it confirms the under-recall performance on the Epinion dataset in a more diverse setting. For Figure 8, the proposed method achieves the best recall for each fold. The small differences in recall values show that the result does not depend on the subsets of data and itdata, lies in the stability of the model. Conversely, some models vary, which suggests that they are sensitive to the distribution of the data and they do not generalise well.

Precision comparison on Epinion is shown in Figure 10. The proposed model achieves the highest precision for all the tests, which implies that the predicted links are more accurate having fewer false positives. This verifies the superiority of the optimisation-based classification stage that also uses kernel estimation and PSO in step 3 and enhances feature significance and class boundaries. When precision and recall of the system are both high, the system is said to be well balanced, and this result is also reflected in a higher predictive F1-score and consequently better prediction quality.

As can be seen from Figures 6 to 10, the results are stable, and the proposed model gets the best performance in accuracy, recall, and precision compared with the classical methods in both datasets. This is a uniform and systematic improvement over all cross-validation folds, confirming that the proposed methodology is not only robust but also scalable for real-world scenarios of link prediction. The effect of integrating filtration,

probabilistic feature modelling, and optimal ensemble learning is a model that captures complex network patterns more effectively than other methods.

Interpretation of the Improvement in Performance

The results demonstrate that our proposed method, Filtered-Based Online S-NFE and Node Link Prediction Approach, outperforms traditional machine learning methods as well as state-of-the-art methods in terms of different evaluation metrics on various datasets. The numerical improvements in Accuracy, Precision, Recall, F1-score and AUC-ROC demonstrate the superiority of the model, but the underlying factors that contribute to this better performance deserve a deep analysis.

The superior performance can be attributed to the multi-stage hybrid model of the proposed method that based on graph filtering, probabilistic ranking, and optimisation-based classification. Unlike traditional approaches based on either structural similarities or learned embeddings, our model makes full use of probabilistic modelling of explicit structural information.

The first pass for graph filtering is crucial to enhance performance. As the search space is enlarged by including structurally relevant candidate node pairs, the model does not waste computations on distant, unrelated nodes. It is not only to improve computational efficiency but also to reduce noise in training data. As a result, the classifier is executed on a more discriminating portion of the data, which yields better prediction accuracy. The probabilistic ranking procedure improves the model to incorporate node influence defined in terms of followers, followees and connectivity patterns. These quantities really represent social reality, where nodes having more connections or exerting more influence have more chances to form new links. Rather than using simple similarity-based scores, the normalised ranking scores provide a more balanced and normalised perspective of node importance, allowing the model to better distinguish between potential and non-potential links. The last shell that integrates kernel estimation, Particle Swarm Optimisation (PSO), and Random Forest classification (RF) contributed most to the global performance. Kernel-based estimation describes the distribution of feature values, and the framework accounts for the uncertainty and variability in the data. PSO optimises the feature weights and parameters, and the best features are given priority. By building on bagging, the Random Forest classifier further reduces overfitting and improves generalisation. Combining these components leads to a very strong and adaptable learning model.

In contrast to the conventional Naive Bayes and Logistic Regression machine learning approaches, our approach demonstrates clearly superior results. These baseline models either treat the features as independent or assume that the relationship is linear, neither of which can capture nonlinear interactions in a network. In contrast to the proposed method, which incorporates structural and probabilistic dependency and can analyse nonlinear

effects among variables. The current approach has significant advantages compared to embedding-based approaches, e.g., Node2Vec, DeepWalk, and so forth. Whilst embedding methods learn latent representations of nodes, they are typically not interpretable as there are no precise meanings attached to each dimension of the output vectors. In addition, such procedures rely heavily on hyperparameter tuning as well as on random walk schemes, which could be fragile when transplanted into different datasets. The proposed method instead uses predefined features and probabilistic statistics, which contribute to the higher level of interpretability of the model and make the analysis easier. Strong baselines for the link prediction task are the Graph Neural Network (GNN) based models. Such models can capture both node attribute and structural information with the help of deep learning architectures. However, this approach requires computationally intensive calculations and a large number of labelled examples. On the other hand, our approach attains competitive or superior results with significantly less computational expense and without excessive training samples. As such, this makes it more feasible for practical applications where both resources and labelled data are scarce.

The advantages of the proposed framework are as follows:

Hybrid classification design : With the integration of filtering, feature extraction and optimised classification, the proposed model can capture diverse aspects of the network behaviours.

Interpretability : Predictions can be interpreted as in classical models on explicit features and within a probabilistic framework, as opposed to black-box DL models.

Efficiency : Due to the filtering step, the model's computational complexity is also reduced, making it feasible to apply the model to large-scale networks.

Stability : It is natural to extend the proposed approach to the related technique called multiple learning and optimisation methods together under the guarantee of consistency in any case or stability.

Handling Space Data : The approach efficiently tackles class imbalance as well as the sparsity of the network issues relevant to the link prediction problem.

Limitations

Even though the proposed method is effective, it also faces several limitations that we would like to point out. The path-based assumption in the filtering phase may not be consistent with every link formation pattern, especially in heterogeneous or evolving networks. This way of thinning out does increase the efficiency, but it can also remove some true links that fail the test. The probabilistic ranking measures are, however, somewhat heuristic although interpretable. The weighting parameters used for such statistics need to be properly adjusted, and their results may be dataset-dependent. The performance could be further improved by applying adaptive or learning-based weighting strategies in the weighting process. Another limitation is that there is no deep learning of features. While the model is more interpretable, we show that a standard softmax layer does not fully exploit the representational power of deep learning. It is anticipated that integrating the above current framework with light neural-network-based machinery to capture non-linear rich patterns will boost its expressiveness without involving too much more overhead. This proves that the proposed method provides a good compromise between accuracy, interpretability and computational cost. With structural filtering, probabilistic inference, and efficient ensemble learning, the model addresses critical challenges for link prediction. The comparative study with state-of-the-art methods and discussion on the limitations validate the same for the efficacy and provide future direction for enhancements. To sum up, the revised results and discussion deepen the insight into the behaviour of the model, not only quantitatively substantiating it, but also qualitatively explaining its performance.

CONCLUSION

In this paper, we propose a filter-based link prediction approach with graph filtering, probabilistic ranking and optimisation-based classification in online social networks. The superiority of the proposed method over the previous one is clear in that the structural and contextual attributes are integrated wholly and comprehensively, which results in better performance in terms of prediction accuracy, precision and recall. Because the filter reduces the search space, the computation also enjoys a benefit, so the filter can also be applied to large-scale networks. The probabilistic ranking scheme also provides an interpretable perspective of node influentiality, which helps to better understand link formations. In addition, the kernel estimation, optimisation technique and ensemble learning-based processing can enhance the robustness and generalisation over different datasets. The experimental results show that our proposed method outperforms both traditional and state-of-the-art baseline methods on real data, which proves its effectiveness for the real-world link prediction application. The work of the future can come from the proposed framework in several aspects. A promising direction would be to integrate deep learning models (e.g., Graph Neural Networks) for enhanced representational learning, while maintaining the

interpretability of the proposed approach. Furthermore, an additional extension would be to introduce dynamic and temporal social networks, where link formation is a function of time. Scalability may be attained by employing distributed and/or parallel computing techniques in the situation of extremely large graphs. 7.10 Using other data types in very complex network scenarios, the inclusion of other types of data, for example, text or multimodal data, may be helpful in further improving prediction accuracy. Finally, investigating adaptive weighting strategies for probabilistic ranking measures is a potential direction for further enhancing the flexibility and performance of the model across various datasets.

ACKNOWLEDGEMENT

Author Contributions: All authors contributed equally, and all authors read and approved the final version of the paper. We would like to extend our deepest gratitude to Dr G Pradeepini and Dr L V Narasimha Prasad for their invaluable knowledge and guidance throughout this research.

REFERENCES

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

Ahmad Ali Nezhad, M., & Makrehchi, M. (2021). Edge-centric multi-view network representation for link mining in signed social networks. *Expert Systems with Applications, 170*, Article 114552. https://doi.org/10.1016/j.eswa.2020.114552

Ahmed, N. M., & Chen, L. (2016). An efficient algorithm for link prediction in temporal uncertain social networks. *Information Sciences, 331*, 120-136. https://doi.org/10.1016/j.ins.2015.10.036

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*, Article 125795. https://doi.org/10.1016/j.physa.2021.125795

Bai, S., Zhang, Y., Li, L., Shan, N., & Chen, X. (2021). Effective link prediction in multiplex networks: A TOPSIS method. *Expert Systems with Applications, 177*, Article 114973. https://doi.org/10.1016/j.eswa.2021.114973

Bastami, E., Mahabadi, A., & Taghizadeh, E. (2019). A gravitation-based link prediction approach in social networks. *Swarm and Evolutionary Computation, 44*, 176-186. https://doi.org/10.1016/j.swevo.2018.03.001

Berahmand, K., Nasiri, E., Forouzandeh, S., & Li, Y. (2022). A preference random walk algorithm for link prediction through mutual influence nodes in complex networks. *Journal of King Saud University - Computer and Information Sciences, 34*(8), 5375-5387. https://doi.org/10.1016/j.jksuci.2021.05.006

Chen, L., Gao, M., Li, B., Liu, W., & Chen, B. (2018). Detect potential relations by link prediction in multi-relational social networks. *Decision Support Systems, 115*, 78-91. https://doi.org/10.1016/j.dss.2018.09.006

Daud, N. N., Ab Hamid, S. H., Saadoon, M., Sahran, F., & Anuar, N. B. (2020). Applications of link prediction in social networks: A review. *Journal of Network and Computer Applications, 166*, Article 102716. https://doi.org/10.1016/j.jnca.2020.102716

Gao, H., Li, B., Xie, W., Zhang, Y., Guan, D., Chen, W., & Cai, K. (2021). CSIP: Enhanced link prediction with context of social influence propagation. *Big Data Research, 24*, Article 100217. https://doi.org/10.1016/j.bdr.2021.100217

Kou, H., Liu, H., & Duan, Y. (2021). Building trust/distrust relationships on signed social service network through privacy-aware link prediction process. *Applied Soft Computing, 100*, Article 106942. https://doi.org/10.1016/j.asoc.2020.106942

Liu, K., Zhou, J., & Dong, D. (2021). Improving stock price prediction using the long short-term memory model combined with online social networks. *Journal of Behavioural and Experimental Finance, 30*, Article 100507. https://doi.org/10.1016/j.jbef.2021.100507

Luo, H., Li, L., Zhang, Y., Fang, S., & Chen, X. (2021). Link prediction in multiplex networks using a novel multiple-attribute decision-making approach. *Knowledge-Based Systems, 219*, Article 106904. https://doi.org/10.1016/j.knosys.2021.106904

Mallek, S., Boukhris, I., Elouedi, Z., & Lefèvre, E. (2019). Evidential link prediction in social networks based on structural and social information. *Journal of Computational Science, 30*, 98-107. https://doi.org/10.1016/j.jocs.2018.11.009

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

Pandey, B., Bhanodia, P. K., Khamparia, A., & Pandey, D. K. (2019). A comprehensive survey of edge prediction in social networks: Techniques, parameters and challenges. *Expert Systems with Applications, 124*, 164-181. https://doi.org/10.1016/j.eswa.2019.01.040

Tang, R., Jiang, S., Chen, X., Wang, H., Wang, W., & Wang, W. (2020). Interlayer link prediction in multiplex social networks: An iterative degree penalty algorithm. *Knowledge-Based Systems, 194*, Article 105598. https://doi.org/10.1016/j.knosys.2020.105598

Wang, F., She, J., Ohyama, Y., & Wu, M. (2020). Learning multiple network embeddings for social influence prediction. *IFAC-PapersOnLine, 53*(2), 2868-2873. https://doi.org/10.1016/j.ifacol.2020.12.2531

Wu, J., Zhang, G., & Ren, Y. (2017). A balanced modularity maximisation link prediction model in social networks. *Information Processing & Management, 53*(1), 295-307. https://doi.org/10.1016/j.ipm.2016.10.001

Xiao, Y., Li, R., Lu, X., & Liu, Y. (2021). Link prediction based on feature representation and fusion. *Information Sciences, 548*, 1-17. https://doi.org/10.1016/j.ins.2020.09.039

Yao, L., Wang, L., Pan, L., & Yao, K. (2016). Link prediction based on common-neighbours for dynamic social network. *Procedia Computer Science, 83*, 82-89. https://doi.org/10.1016/j.procs.2016.04.102