



An effective keyword search co-occurrence multi-layer graph mining approach

Janet Oluwasola Bolorunduro¹ · Zhaonian Zou¹ · Mohamed Jaward Bah²

Received: 14 March 2023 / Revised: 27 December 2023 / Accepted: 14 February 2024 /
Published online: 2 April 2024

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2024

Abstract

A combination of tools and methods known as "graph mining" is used to evaluate real-world graphs, forecast the potential effects of a given graph's structure and properties for various applications, and build models that can yield actual graphs that closely resemble the structure seen in real-world graphs of interest. However, some graph mining approaches possess scalability and dynamic graph challenges, limiting practical applications. In machine learning and data mining, among the unique methods is graph embedding, known as network representation learning where representative methods suggest encoding the complicated graph structures into embedding by utilizing specific pre-defined metrics. Co-occurrence graphs and keyword searches are the foundation of search engine optimizations for diverse real-world applications. Current work on keyword searches on graphs is based on pre-established information retrieval search criteria and does not provide semantic linkages. Recent works on co-occurrence and keyword search methods function effectively on graphs with only one layer instead of many layers. However, the graph neural network has been utilized in recent years as a branch of graph model due to its excellent performance. This paper proposes an Effective Keyword Search Co-occurrence Multi-Layer Graph mining method by employing two core approaches: Multi-layer Graph Embedding and Graph Neural Networks. We conducted extensive tests using benchmarks on real-world data sets. Considering the experimental findings, the proposed method enhanced with the regularization approach is substantially excellent, with a 10% increment in precision, recall, and f1-score.

Keywords Keyword search · Co-occurrence graph · Graph mining · Graph neural network · Multi-layer graph embedding · Deep learning

1 Introduction

Computers can comprehend language because it is the medium humans use for communication; hence, Search Engine Optimization (SEO) is optimizing websites to increase their visibility in Google's natural ranking and other search engines. It can model how

Editors: Dino Ienco, Robert Interdonato and Pascal Poncelet.

Extended author information available on the last page of the article

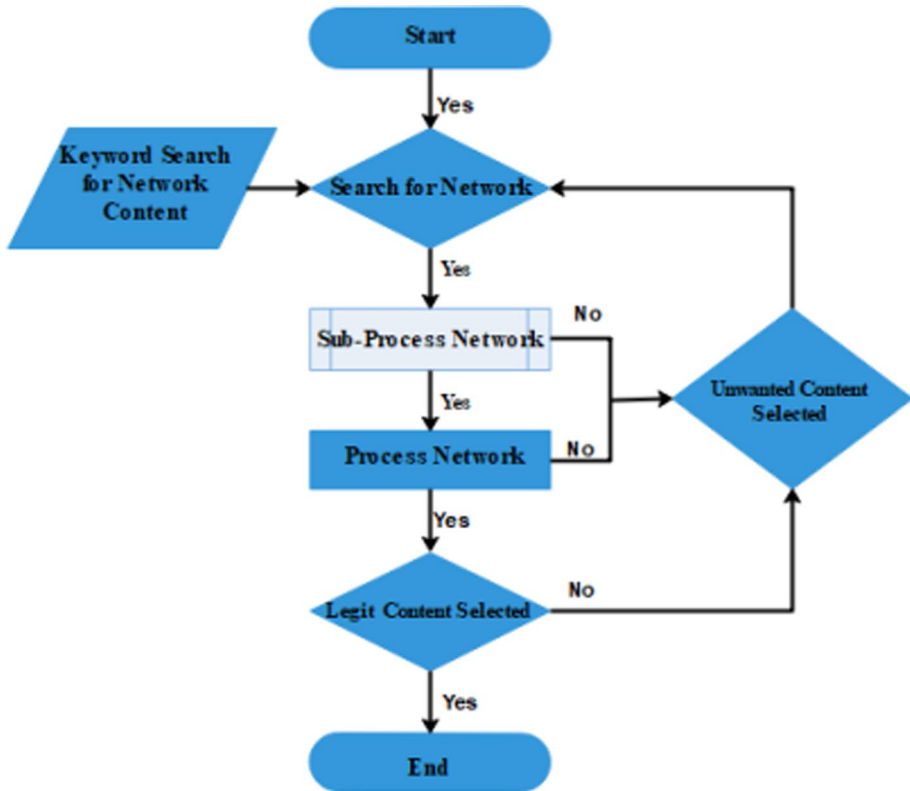


Fig. 1 Keyword search representation flow chart

individuals acquire and discover information on practically any topic. Keyword search is finding the relevance of words, queries, and phrases to a website and its folios so that the user can find the best folio to answer their query on real-world applications, known as search intent see Fig. 1 for more details.

One of the most valuable uses of pattern recognition (PR), machine learning (ML), artificial intelligence (AI), social computing (SC), and recommender systems (RS) is to help make informed decisions and provide a more realistic representation of multiple relations that characterize an entity in the system. However, optimizing content or creating possible searches from search engines is possible if what people are searching for and what they want to see can be accessed easily (Han et al., 2022; Aggarwal, 2016). Yet another approach to finding Co-occurrence (CO) patterns is revealed through corpus linguistics and statistical analyses in which extensible Markup Language (XML) and graph structures in hypertext corpora extract specific data attributes.

Co-occurrence networks, sometimes called semantic networks, Segev (2021) are graphical methods for solving ambiguity problems and analyzing text, including potential relationships among entities, concepts, and organisms like bacteria (Freilich et al., 2010) using a graphic visualization. Co-occurrence networks are collections of terms that are connected together because they occur together in a certain text, concept, or structure. By linking words together according to a set of co-occurrence strategies and searching the format of

scientific communication, co-citation analysis, multinomial model, and graph neural networks (Han et al., 2022; Aggarwal, 2016; Yang et al., 2021; Garg, 2021) networks are created, which have significantly improved the techniques nevertheless still have flaws. There is great interest in relational database keyword searches (Yang et al., 2021; Garg, 2021; Bast et al., 2016), and the most critical aspect of relational data access is a Structured Query Language (SQL). Accessing a significant volume of relational data has become more challenging for prospective users due to the requirement that relational data schema be well-known to use SQL. Graphs, also known as social graphs, are being used in social media for information organization, structure, storage, and retrieval, for node categorization, connections prediction, clustering, and visualization (Cai et al., 2018; Goyal & Ferrara, 2018). Graph clustering groups the nodes of a graph into clusters using the graph structure or node attributes. Numerous research works (Ma et al., 2021) in the node distribution approach are proposed, and the denoted nodes can be transformed into traditional clustering algorithms. Search Engine Optimization (SEO), such as Google, still represents an influential and trustworthy resource for discovering practical website information.

The context relevant of the user query and the search engines indexed folios were the primary factors used by early search engines to return pertinent folios for the user. The information retrieval (IR) techniques were directly implemented in the retrieval and ranking algorithms. Conventional information retrieval (IR) presumes that the fundamental unit of information is a document and that a vast array of documents can be accessed to create the text database. Researchers have used IR to extract knowledge from structured data for community identification and search. A list of keywords sometimes referred to as terms, is the most widely used query format. Information in the text is unstructured, whereas data in databases is highly structured and kept in relational tables; thus, information retrieval from text varies from retrieved data from databases using SQL queries. The primary goal of interest is retrieval and related activities that can increase the accuracy or efficiency of retrieval since text retrieval lacks a structured query language like SQL, and the IR community has not focused much on real-world data applications like false news.

Keyword research is the first and most crucial step in any search engine optimization strategic plan (Yang et al., 2021; Garg, 2021). The most popular approach to solving the keyword search problem is Graph-Based Keyword Search (GBKS), which identifies a set of closely linked nodes in the graph that may match a specific keyword based on the query (Bhalotia et al., 2002; Kacholia et al., 2005; He et al., 2007), BANKS-I (Bhalotia et al., 2002) considers the shortest route from a tree's root to a node that contains keywords, BANKS-II (Kacholia et al., 2005) suggests using a forward search to approximate a solution, and BLINKS (He et al., 2007) tries to identify the set of all different sub trees with the best scores to improve the BANKS-II approach. These retrieval techniques are centered on nodes while using keyword search engines and semantic relationships (Wang et al., 2008) can link keyword inquiries and formal questions. Therefore, classical manual reading for information extraction and knowledge acquisition cannot keep up with the needs of the complex data age.

Researchers on machine learning (ML) and graph mining have used various branches of artificial intelligence, from recommendation systems, computer vision, natural language processes, and graph-based, for solving standard processes through graph-based machine learning. In conventional ML, researchers have been working on alternative clustering problems on graphs, and comparing the similarity of objects of the same kind is crucial in many applications (Han et al., 2022; Aggarwal, 2016). A sustainable cluster is designated as a collection of nodes in a multiplex network that is concurrently coupled to one another across all of the distinct layers (Baxter et al., 2016). Moreover, sustainability corresponds

to several paths that connect the same pair of nodes in the feasible cluster, but each exists on a different multiplex layer. Therefore, understanding fundamental search co-occurrence correlation through multi-layer graph representations is an essential methodology from literature to intelligence analysis (Fig. 2).

Multiple layers are a feature of realistic systems. Multi-layer graphs (MLGs) are widely accepted as such (Boccaletti et al., 2014; Kivelä et al., 2014; Kumar et al., 2020) differ from single-layer graphs SLGs by their multi-relational structure that offers a range of resources for making good decisions, with an inter-relational corporation structure that provides various resources for decision-making, as well as entities that can have different types of relationships between them. When modeling several real-world applications among the same group of people, for example, MLGs provide an expressive method where layers represent various online and offline relations (e.g., following, co-authorship, co-working relations, and so on), keyword research is the first and most crucial step in any search engine optimization strategic plan where various academia and the business community have utilized it in helping users maximize network resources where Label Propagation (LP) (Nickel et al., 2015; Alimadadi et al., 2019) Random Walks (RW) (Bojchevski et al., 2018; Valdeolivas et al., 2019), E-Commerce Recommendation (E-CR) (Aggarwal, 2016) Multi-layer graph embedding (MLGE) (Rossi et al., 2021; Makarov et al., 2021), Deep Neural Network have been well studied to forecast the relational link between entities and keyword search on multi-layer graphs to represent complex relationships accurately (Wu et al., 2020; Perozzi et al., 2014). However, the common usage of MLG representations of various vertices, edges, and critical world search methods find relevant components in a network system. Current methods focus on specific multi-layer graphs, such as multiplex and heterogeneous structures of interconnected complex systems. At the same time, most affirmation approaches have their merit and demerits despite challenges like

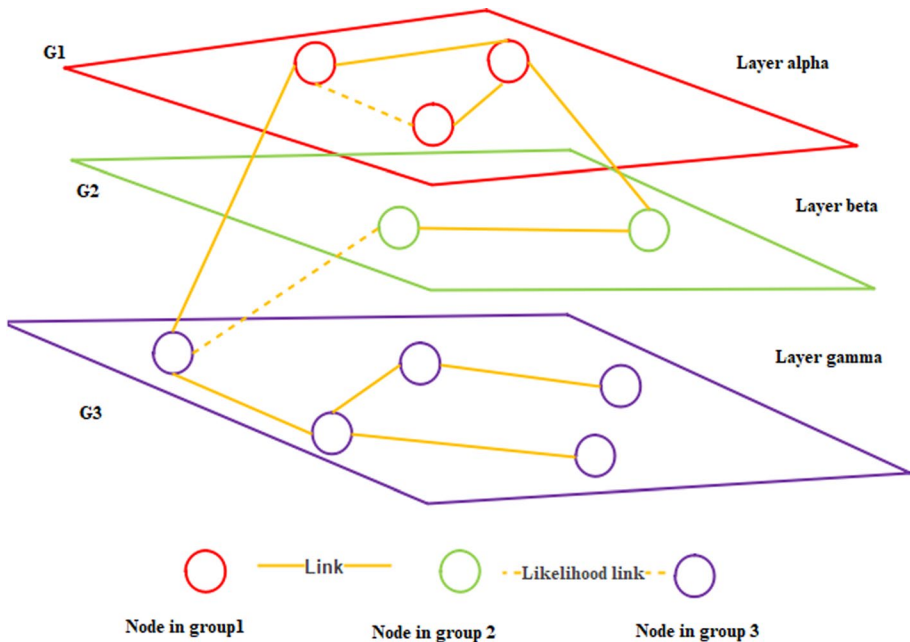


Fig. 2 Multi-layer graphs representative

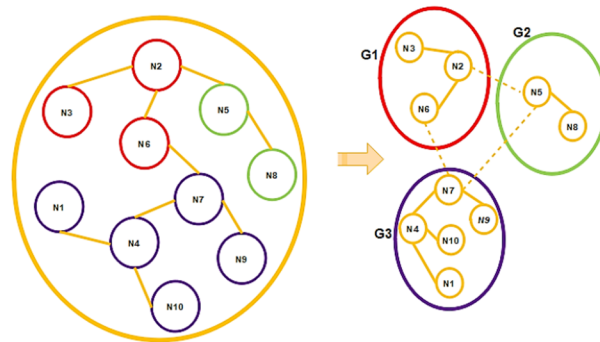
memory cost and time complexity, graph embedding known as representative of network learning offers (Grover & Leskovec, 2016; Hamilton et al., 2017) an effective solution by changing the representation form and mapping nodes into a low-dimensional space, maintaining consistent and enhancing understanding of network entities. The increasing accessibility of complex networks with billions of vertices and edges has significantly advanced network analysis, where Multi-layer Graph Embedding (MLGE) attempts to describe the vertices and edges in vector space while maintaining the structure of the graph and information within and across layers in overcoming the complex network representation and analysis challenges of the graph embedding network.

Diverse techniques have been put out to learn graph representations. Graph Neural Networks (GNN) (Battaglia et al., 2018), the most known network that Google recently introduced, extends popular networks like RNN and CNN to graph-structured data (Scarselli et al., 2008; Duvenaud et al., 2015; Niepert et al., 2016; Defferrard et al., 2016). One study area is building neural networks as an RNN variant that functions on graphs. (Li et al., 2015) extended the GNN model by proposing a brand algorithm of RNN in the original GNN model. A significant pull of works that have attracted fast-ripening goal is the GCNs (Kipf & Welling, 2016), centered on spectral graph theory, which was initiated (Bruna et al., 2013) and then extended by Defferrard et al. (2016) with fast localized convolution. Most neural networks transverse deep to get a unique performance. Recent GNNs that deal with node categorization on graphs are unable to achieve high performance on a variety of data sets because they are shallow networks and tend to concentrate on node-wise scores.

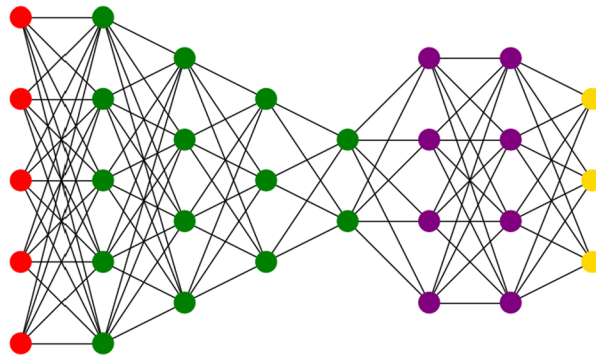
GNNs are becoming famous in multi-layer learning. Wu et al. (2020); Hamilton (2020) However, prior methods have yet to thoroughly investigate these graphical interactions since they have not combined information from several links concurrently. Researchers have proposed to utilize a multi-omics data analysis by embedding multiple knowledge into graph neural networks to solve this problem (Xiao et al., 2023) To buttress the benefit of structural diversity and deep GNN Architectures, GNN model a pipeline with two-stage novel space is proposed by Feng et al. (2023) which aim to generate high performance. In contrast, transferable deep GNN models in a block-wise manner are utilized, Liang et al. (2021) and He et al. (2021) make use of the multilevel embedding framework MILE and a distributed multilevel framework (Dist MILE) for scalable graph embedding. Our proposed keyword search co-occurrence multi-layer graph mining (EKSCOMLGs) considers implementing association based on multi-layer graph embedding and graph neural networks based on multiple knowledge for mining of features network. Thus, Its fundamental is to learn co-occurrence relations between real-world data sets.

Figure 3a considers a scenario where, in a certain community, there are researchers, and recommendations of individuals who have never cooperated seem more valuable. Suppose ten researchers are skilled in different fields and assume there is a talent hunt for a project requiring Mathematicians, Architecture, and Computer Analysis. Since a graph can be used as a pictorial drawing for easy illustration, a social graph mapping based on Co-membership can be used to indicate the model of bringing together information from two or more people who belong to the same community of researcher but different areas of expertise groups (G). Using Fig. 3a to illustrate, where (Red α) represents researchers who are well skilled in Mathematician (G1), (Green β) represents researchers who are well skilled in Architecture (G2), and (Purple γ) represents researcher who is well skill in Computer Analysis (G3).

A graph neural network representation example is shown in Fig. 3b, where the circles indicate nodes and their functions on the data are represented by the edges, which represent weights or information passing along where certain layers may be hidden. The



(a) Keyword search co-occurrence representation of researchers in a community



(b) Graph Neural Network Representative

Fig. 3 Graphs using keyword co-occurrence and graph neural network representation

structural role of the circles of a node can be represented by Red, Green, Purple, and Gold color, When the layer is few, it is called a shallow neural network and when the hidden layer is many, they are called a Deep Neural Network. For the proper execution, there must be a mutual linkage or interest between nodes and edges in Fig. 1.

In this way, we are particularly interested in two research questions: (1) What is the relatedness between nodes and edges within the same community type or different community types using real-world data? (2) Whether the proposed model will perform better using our proposed model? To solve these questions, search engine optimization (SEO) based on content information properties using elements of Multi-layer Graph Embedding (Rossi et al., 2021; Makarov et al., 2021) and Graph Neural Networks have gained helpful information (Wu et al., 2020; Hamilton, 2020). However, a practical keyword search co-occurrence multi-layer graph mining approach (EKSCOMLG) is an NP-complete problem. Thus, the proposed EKSCOMLGs are driven by enhanced multi-layer graph embedding and graph neural networks, which could revolutionize practical keyword search co-occurrence tasks in real-world applications, fully utilizing the network's capabilities to enhance user experience. The following is a novelty of this paper's contributions:

- An effective keyword search co-occurrence multi-layer graph mining approach is proposed. The proposed method is built on multi-layer graph embedding and graph neural networks with highly adaptive real-world processes to build intelligent solutions.
- We performed extensive experiments using four evaluation metrics on distinct data sets against other benchmark methods. Our proposed model shows improved performance and offers the advantage of providing links that guide the classification process, which enhances existing techniques by examining and learning co-occurrence relations, social association, deformity prediction, and recommendation.

The remaining section of the manuscripts is sorted as follows: Sect. 2 describes the preliminary and problem definition, Sect. 3 denote the materials and methods, Sect. 4 denotes the experiment 5 denote the results and discussions, Sect. 6 represents the related works, Sect. 7 is the conclusion.

2 Preliminaries and problem definition

The preliminaries are introduced in this section, including the definitions and notations used (Table 1), and then the problem definition where directed or undirected edges can represent a graph's real-world network. To introduce the terminology, for a graph G , the node-set is represented by N and the edge-set with E ; thus $G = (N, E)$ where N is the vertex or node set of size $n = |N|$, E is the edge list of size $m = |E|$. Note N is defined as a subset $N_u = \{u_1, u_2, \dots, u_n\}$ and $N_v = \{v_1, v_2, \dots, v_n\}$ and a set of edges between this vertex $E = \{e_{11}, e_{12}, \dots, e_{mm}\}$ where $e_{uv} = u_i, v_j \in E, 1 \leq i, j, \leq n$.

Another way to describe graph G is as an adjacency matrix A with $A(u, v) = 1$ if $(u, v) \in E$ and 0 otherwise. if $A(u, v) \neq A(v, u)$, G is a directed network, otherwise If the graph is undirected, the matrix $A(u, v) = A(v, u)$ for all nodes $u, v \in N$ is symmetric. If $A(u, v)$ is weighted by $w(u, v) \in W$, $G = (N, E, W)$ is a weighted network; otherwise, it is an unweighted network. An improved graph with vital information from simple graphs can be created using attributed graphs, multi-relational graphs (Hamilton, 2020), and Multi-layer graphs (Kivelä et al., 2014).

Definition 2.1 Simple graphs are expanded into attributed graphs. The node attributes X , and the edge attributes X^e are added to obtain them. For example, $X \in R^{n \times d}$ represents a node feature matrix, and $X^e \in R^{m \times c}$ represents an edge matrix, with $x_{u_i, v_j}^e \in R^c$ representing the vector of an edge $e_{u, v}$.

Definition 2.2 An extension version of basic graphs with edges having many kinds of relations τ are called multi-relational graphs. $e_{uv} = (u_i, v_j) \in E \rightarrow e_{uv} = (u_i, \tau, v_j) \in E$ is the situation in question. One related adjacency matrix A^τ exists for each edge. It is possible to construct the complete graph as an adjacency tensor $A \in R^n \times r \times n$. Heterogeneous and multiplex graphs are two sub-types of multi-relational graphs.

Definition 2.3 Multi-Layer Graphs (MLGs) have multiple edges between nodes. Denoting a MLGs where $G_1, G_2, \dots, G_m = (N, E_1, E_2, E_m)$ considering that the graph has m layers. Accordingly, Ma et al. (2021); Bhalotia et al. (2002) can likewise be modeled as an EKSCOMLGs $M = (G, C)$. The MLGs $= (G^\alpha, \alpha \in \{1, 2, \dots, n\})$ are the pair of graphs in this case $G^\alpha = (N^\alpha, E^\alpha)$, G^α is set of layer α of G . The CO among nodes of various layers

Table 1 List of notations

| Notation | Meaning | Notation | Meaning |
|--------------------------------|--|----------------|--|
| MLGs | Multi-layer graph | G | Graph |
| N_u | Set of vertices for graph u_1, u_2, \dots, u_k | N_v | Set of vertices for graph v_1, v_2, \dots, v_k |
| G_M | Graph MLG | N_M | Vertex MLG |
| \iff | Long if and only if arrow | \implies | Long Double right arrow |
| E | Set of edges for graph e_1, e_2, \dots, e_m | δ | Delta |
| N | Number of Node or N | E | Number of edges or M |
| Ω, C | MLGs for co-occurrence graph | u_i | Vertex u on i th |
| $e_{u,v}$ | Edge between vertex u and vertex v | v_j | Vertex v on j th |
| $W_{u,v}$ | Weight on edge $e_{u,v}$ | k_k | Vertex u on k th |
| $W_{u,u,k}$ | Weight on three vertices | m | Graph Layer m |
| A | 0-1 Adjacency matrix | G_1 | Graph in G_1 |
| X, Y, Z | Set of vertices or node | S_{uv} | Adjacency matrix on sequence |
| $y_{u,v}$ | Entry in matrix Y | $k(u)$ | Member of unique (u_i, v_j) of S |
| N | Number of Vertex or Nodes | E | Number of edges, connections or links |
| N^α | Set of nodes on layer α | E^α | Edges set present in layer α |
| N | Set of vertex or nodes | E | Set of Edges, connections or Links |
| N^α | the set of nodes on Layer α | E^α | the set of edges present in Layer α |
| G^α/G^β | CO between node of different layer | C | Cross-Layer |
| ρ | Parameter for binomial distribution | P | Probability distribution |
| $E^{\alpha,\beta}$ | Interlayer linkage | $h(qv)$ | Relatedness score on layer |
| G_Q | Query Graph | $\hat{K}(u,v)$ | Number of times unique entity pair (u, v) appears in S |
| $S(u_1, v_1) \dots (u_n, v_n)$ | Sequence ordered of entities in N | $k(u, v)$ | $\hat{K}(u, v) + \hat{K}(v, u)$ if $u \neq v$ |
| S_1 | Sequence of $S(u_1, v_1) \dots (u_n, v_n)$ | $k(u, v)$ | $\hat{K}(u, v)$ if $u = v$ |
| S_2 | unique entity S_1 | | |

$C = \{E^{\alpha,\beta} \in N^\alpha * N^\beta; \alpha, \beta \in \{1, 2, \dots, n\} \alpha \neq \beta\}$ where G^α is set of layer α of G and G^β is set of layer β of G with $\alpha \neq \beta$. The MLG M 's element E^α is the set of connections that make up the α layer, and the elements $E^{\alpha,\beta}$ is the set of edges linking α and β layers. The nodes and edges that comprise the layer α are collectively called N^α and E^α , respectively.

Definition 2.4 Graph Embedding (GE). A functional definition for the graph embedding-based with a mapping function F is defined by $f \in N \times R \times R$. Thus, an object mapping function for nodes $f : N \rightarrow X$ and an object link mapping function: $p : E \rightarrow Y$ are both included in MLGs. In object type X , each object node $n \in M$ corresponds to a particular object type or $f(n) \in X$. Each link object in the collection of object types, $e \in M$ or $f(e) \in Y$, corresponds to a certain object type. When two links are members of the same relationship type, their start and end object types are the same for both links.

Definition 2.5 Graph Neural Networks (GNNs) are developed by applying deep learning models to graph structure data. It implies that although deep learning models work with data in Euclidean space, some GNNs operate in non-Euclidean domains. Suppose a graph $G = N, E$ with adjacency matrix A and vertex matrix (or edge matrix) X (or X^c).

Given A and X as inputs, the goal of a GNN is to discover the output, i.e., node embedding and node classification, after the m -th layer is: $H^m = F(A, H^{(m-1)}, \theta^{(m)})$, where F is a mapping(propagation) function, θ is a parameter function F , and m denote the index of the layer so when $m = 1$, then $H^{(0)} = X$. Assume $\sigma(\cdot)$ is a non-linear function e.g ReLu, w^m is the weight matrix of layer m . A simple form of the mapping function is often used: $F(A, H^m = \sigma(AH^{(m-1)}W^m)$. The mapping function can be enhanced for suitable GNN tasks such as the node classification task and node embedding task (Kipf & Welling, 2016; He et al., 2021). More information on general graph representatives using embedding and GNNs can be found in Hamilton (2020).

3 Materials and methods

3.1 Overview of keyword search

Keyword search creates a friendly interface for information retrieval from complex data structures. Likewise, information retrieval suggests content to users of web services during interactions. Over the years, Tags have become increasingly popular to categorize web and online social network content known as folksonomy (Bai et al., 2009) and are a well-studied topic in information retrieval, computer science, and the recommender system field.

Using a folksonomy, it is possible to use a 3-dimensional array $F = [f_{uvk}]$ of items with a tag. Folksonomy is defined over the group of vertices called users $U=(u_1, u_2, \dots, u_n)$ the group of items $I=(I_1, I_2, \dots, I_m)$ with a tags $T=(t_1, t_2, \dots, t_k)$ where the element $f_{u,v,k}$ is a unary value indicating whether the user u has tagged the item v with the k th tag. Two tags may be strongly related if their co-occurrence frequency is high; however, their co-occurrence frequency should be shallow if the two are unrelated.

Consequently, ML algorithms extract meaningful themes from a corpus of documents such as probabilistic topic models(PTM). PTM is a common semantic representation method used for the social network node. The straightforward approach utilizes Latent Dirichlet Allocation (LDA) to extrapolate the topic from the generative model. This strategy can also be divided into ranked search and conventional search. Most search algorithms used in conventional search are conjunctive keyword searches, which return all documents containing the search terms without considering the semantic linkages between them or centered on node interactions. A Link prediction based on Keyword Search on structural similarity or dynamic correction has been presented to estimate the propensity of a connection between two nodes as standard search is inadequate for ranked search; however, it has its flaws (Han et al., 2022; Aggarwal, 2016; Yang et al., 2021; Garg, 2021; Kumar et al., 2020).

Limitation of Keyword Search using Social Tag and Probabilistic Model

The user language's homonyms, polysemies, synonyms, and other user tagging practices might sometimes make the recommendation process challenging. As a result, social networking services like Flickr might have hundreds of millions of users, objects, and tags. Most topic modeling research does not explicitly employ multi-layer graphs, while several studies disregarded categorical delivery and cross-validation outcomes from the balanced population data set presented.

3.2 Keyword search using graphs embeddings and multi-layer graphs (MLGs)

MLGs allow users to enter several search terms for the best relevant results. Though it can be complex, keyword research, content creation, and link development are the three main components of SEO. Of those three, keyword research is the most crucial. For instance, we may produce the best content and generate amazing links that propel us to the top of Google results. Still, If a wrong keyword is targeted in terms of real-world applications, there won't be benefits in terms of e-commerce growth and technological aspects. Effective keywords can make or break an SEO application in the real world. Key actions to initiate keyword research are as follows:

Step 1: Using important terms and related searches, develop keyword ideas.

Step 2: Determine the actual keyword difficulty and searches.

Step 3: As shown in Fig. 1, ascertain the user's goal.

Cao et al. (2013) is a method that protects privacy and ranks documents using coordinate matching. Searching documents in the dictionary-scale vectors describes the keywords where the links of different keywords in the document are not considered thus the retrieval result obtained by the schema lacks accuracy. Aggarwal (2016) developed the influence limiter algorithm to study trustworthy recommender systems. A global measure of each user's reputation is utilized in the suggestion process, but it cannot expressly endorse trustworthiness without user feedback thus this method needs help obtaining more requests for trustworthy dimensions. GE and Co graphs as a feature can support updates on the data set, to use CO graphs as features, the interrelationship is needed and it is often addressed as a boolean feature.

GE is a family of ML and DL approaches that take advantage of the inherent graph structure of data types to transform high-dimensional vectors into continuous vector representations of low-dimensional discrete variables. To capture structural information, GE models (Rossi et al., 2021; Makarov et al., 2021) offer a global picture of latent relationships. For instance, the node-embedding method utilized a node-wise method such that $e_{uv} = h(y_u, y_v)$, where y_u and y_v are the node-wise embeddings and h is the decoder function ranging in complexity from a parameter-free inner product of a multi-layer MLP. In contrast, the constituted node embedding themselves is typically computed with some form of trainable GNN encoder model g of the form $y_u = g(x_u, G_u)$ and $y_v = g(x_v, G_v)$ where G_u and G_v are the subgraphs containing nodes u_i and v_j respectively. Turning to edge-wise methods, the edge representation e_{uv} relies on the subgraph G_{uv} defined by both u_i and v_j . In this case $e_{uv} = h_e(u_i, v_j, G_{uv})$, where h_e is an edge encoder GNN whose predictions can generally not be decomposed into a function of individual node embedding method. With ML systems, we note that while the embeddings from node-wise subgraph for all nodes in the graph can be produced by a single GNN forward pass, node classification, node clustering, link prediction, and community discovery and keyword search which are often focused on finding a group of nodes in the graph that match the keywords, which is more of a search task, edge-wise subgraph and corresponding forward pass and multi-layer linkage graphs are needed to make predictions for each candidate edge.

The node embeddings are implemented by DeepWalk (Perozzi et al., 2014) and node2vec (Grover & Leskovec, 2016); both rely on random node co-occurrence to train the models. Since their objective function is non-convex, initializations of this kind may become trapped in local optima. Thus, using node embedding directly in keyword searches is not natural. Most node embedding methods rely on network distance; nevertheless, the resulting edge-wise embedding specifies a relationship between nodes.

In graph-based machine learning, shallow embedding methods have proven effective in capturing the relationships between nodes. We delve deeper into the exciting world of multi-relational graphs. Multi-relational graphs are complex networks consisting of nodes and edges, where each edge represents a specific relation between two nodes.

Formally, a knowledge graph is denoted as $G = (N, E, R)$, where R is a relation type, entities $u_i \in N$, and edges $(u_s, \tau, v_o) \in E$ are the entities. In ascertaining the likelihood that such edges correspond to E , the task assigns scores for legit ideas (i.e., triple-like subject, relation, and object). Since they hold factual information as tuples of the form (u, τ, v) , which represent a relation τ between nodes u and v and can be selected from a range of GNNs, these graphs are frequently referred to as knowledge graphs. Numerous decoder functions, such as ComplEx, RotatE, RESCAL, TransE, and TransX, have been proposed. Every decoder has its method for encoding and decoding relations between nodes, although they all have advantages and disadvantages (Hamilton, 2020).

The extension of graph mining has created multi-layer graphs. Liu et al. (2017) suggested three techniques to build a multi-layer network into a continuous vector space: "layer co-analysis," "results aggregation," and "network aggregation." To find a vector space for a multi-layer network, "network aggregation" and "results aggregation" apply the conventional network embedding method on the merged graph or each layer; our proposed method differs from this approach.

3.3 Proposed method

3.3.1 Keyword search co-occurrence model

Our model uses both directed and undirected multi-layer graphs. We use an edge-wise approach for a multi-layer graph denoted by MLGs = (G, N, E, M) Where $G_M = N_M, E_M$, N is represented as nodes and E is represented as edge or links with M denoting the Layer. In a graph $G = (N, E)$ with the node $N|N| = m$ with a link set, E is considered directed if $(u, v) \in E \implies (v, u) \in E$ whereas an undirected edge implies that $(u, v) \in E \iff (v, u) \in E$. Suppose information is observed on a selected subset of nodes in N , denoted by $N_o|N_o| = m_o$ and G_o represent the subset of G induced by N_o , let X and Y be a set of nodes and edges such that if $x \in X$ and $y \in Y$, then (x, y) represents the pair of x and y thus, a variable $Y_{u,v}$, where $u, v = 1, 2, \dots, m, u_i \neq v_j$ to show whether a link exists among two nodes u and v in G or not where Y is defined as the graph G 's adjacency matrix. For any edge (x, u_i) , its equivalent edge could be represented by $e = (x, y)$. Note that for undirected edges, $Y_{u,v} = Y_{v,u}$. In the case of the edge-wise approach, the edge representation is denoted as $e_{uv} = h_e(u_i, v_j, G_{u,v})$, where h_e is an edge encoder GNNs whose predictions cannot be generated as the node embedding mention previously. This basic idea can be generated through a query using a synthetic example.

Synthetic Example Suppose the community consists of the set of Researcher $R = \{M, A, C\}$ and make up of group of expertise denoted as $G_E = \{G_m, G_a, G_c\}$ as shown in Fig. 3a. The Researcher can be grouped according to the area of expertise and modeled as a bipartite graph $G = \{N_r, N_e, E\}$. A bipartite graph in this regard is used in MLG to organize entities based on their relationships where N_r represents the entire researcher and N_e is the group of experts where the edge E is denoted as $(r, e) \in E$.

A complete bipartite graph on Researcher (nodes) R and Expert Group G_E contains all possible edges between the researcher and the expert group thus an edge $(r, e) \in E$ is established from r to e if r performs an action in e . An edge between r and e is

linked by a relation R_{re} based on performed action or the weight between them. Let us assume that each researcher in an expert group is associated with a programming language, the idea is to compute relationship scores with respect to certain focus areas that demand area of specialty and location and can be passed based on keyword query $Q = \{rq_1, rq_2, \dots, rq_n\}$ to the relationship algorithm. The detailed mechanism for extracting the query Q from the research profile is not detailed in this work.

In knowledge graph representation, a multi-relation graph can be represented as $G = (N, E, R)$ where the edges are modeled as tuples $e = (u, \tau, v)$ signifies the type of a relation $\tau \in R$ occur among two entities. Let X be the $m \times n$ matrix of real-world network expression value from n samples and R denoted the $m \times q$ matrix of the relationship links, then $Z_{u,v,k} = S_k(Y, X_{u_i}, X_{v_j}, R_{u_i}, R_{v_j})$ where X_{u_i} and R_{u_i} denote i th row of X and R . This notion emphasizes that the k th co-occurrence graph is a function of the feature of the network (Y), the expression level of the corresponding network (X_{u_i} and X_{v_j}) as well as other network linkages (R_{u_i} and R_{v_j}) and the function S_k denotes any sequential measure based on different data sources. Our goal is to study the KSCOMLGs in a simple framework to relate the values of Y_{uv} to $Z_{u,v,k}$ in the settings where Y_{uv} can be of different types. A link $R(u, v, k)$ in each layer denotes the associations among nodes u , v , and k in a given community, and the sequence of the distribution all serves as the training data for the real-world data contain some vital information useful for the analysis.

3.4 Computational complexity of the proposed EKSCOMLGs

EKSCOMLGs could deduce navigation graphs denoted in the search engine query logs to comprehend the relationships between search engine inquiries. We investigate structures with layers in addition to nodes and edges to describe networks with many types of edges (or with other comparable features) in systems. A graph can be represented as a color problem; a similar procedure is called "graph coloring" on an undirected graph G , where the nodes serve as the colored regions, and the edges serve as the neighboring pairs.

Considering a scenario in the given Keyword search co-occurrence representative in Fig. 3a where a community of experts is to set up an activity that requires at least one additional expert from group 1 the *Red* = N_3, N_2, N_6 , Group 2 the *Green* = N_5, N_8 and Group 3 the *Purple* = $N_1, N_4, N_7, N_9, N_{10}$. Since N_2 personally knows N_3 and N_6 from previous collaboration (reflected by social relation), N_2 is well connected to group 1 the mathematical expertise group but N_2 does not know any member from group 3 the computer analysis group but there is a link between group 3 member the computer analysis N_7 and a member a mathematician member N_3 , and likewise a link between group 1 member N_2 and group 2 member N_5 the expertise in architecture group.

In this scenario, N_2 of the mathematical group may collaborate with N_5 which is linked with N_2 , hence N_2 may act as an invitation to join the architecture group. Likewise, since N_7 member of the computer analysis has a link with N_3 the mathematician group, it is most likely that N_3 will serve as the invitation to join the computer group since N_7 is the focal node in the Computer analysis group-which is linked with N_3 . Thus the graph mining technique supports the discovery of emerging social relations which is the logic that is utilized in the discovery of keyword search co-occurrence multi-layer graphs.

3.4.1 Scenario: a keyword by typing a URL and searching the co-occurrence graphs

Let Researcher R suggest a URL, U , that has been previously visited, the system identifies the relationship between researcher R , and the experts who have searched the URL, U . The co-occurrence is represented as a Researcher Co-occurrence Matrix (RCM) and is evaluated based on relatedness between researchers as shown in Table 2. This stems from the fact that co-occurrence is considered a more general representation of the URLs since they are descriptor of the project being addressed as compared to the URL address themselves. Each Researcher is represented by a vector of co-occurrence he/she has utilized linked by the frequency vectors of each pair of expertise on a given researcher topic.

For instance, if N_1 and N_2 are the number of experts in groups 1 and groups 3 and N is the sum total of researchers, the expected number of co-occurrences as proposed by Forbes is $E(X) = n_1, n_2/N$.

3.4.2 Multi-layer activity

Assume the nodes is swap between colors α , β , and γ in a given community as shown in Fig. 2. In this case, N_u in Layer α can communicate to node N_v in Layer β , and N_u in Layer α can communicate to N_v in Layer β and N_k in Layer γ , respectively. Let $G_{\alpha,\beta,\gamma}$ represent the induced sub-graph of G 's nodes, colored α, β, γ . The operation of a (α, β, γ) swap concerning G is as follows:

Lemma: Let $G \in G'$ be appropriately colored and assume x be any node of G , Suppose nodes $y, z \in Adj(x)$ be colored (α, β, γ) respectively with $\alpha \neq \beta$ or $\alpha, \beta \neq \gamma$. if (α, β, γ) connection connects y and z in G , then (α, β) or (α, γ) in $G_{Adj(x)}$ is then connected.

Proof Let $C = [y = x_0, x_1, x_2, \dots, x_m = z]$ be a $[\alpha, \beta, \gamma]$ Link in G sequence of communication between y and z . Thus, if every edge has at least one end, m is vertex cover.

We assert that the equation $\{x_0, x_1, x_2, \dots, x_m\} \subseteq Adj(m)$ The statement is correct if either $m=1$ in Layer β or Layer γ . Suppose $m \geq 3$ and the link is correct for all minimum α, β, γ links less than m . G is a K -edge connected subgraph if subgraph $G' = (N, E)$ is connected for all $S \subseteq E$ where $|S| < K$. The highest value of k , such that G is k -edge-links, is the edge association of G .

4 Experiments

4.1 Experimental settings

The EKSCOMLGs model's performance is assessed in this section; followed by the description of the experimental setup and presentation of findings. The algorithms were implemented using Python 3.0 with Anaconda and UCINET 6.733. The tests are performed

Table 2 Researcher co-occurrence matrix analysis

| Researcher | N_1 | N_2 | N_3 | N_4 | N_5 | N_6 | N_7 | N_8 | N_9 | N_{10} |
|--------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|
| Group ₁ | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Group ₂ | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| Group ₃ | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |

on a Legion System (GPU/RTX) running Windows 10 and equipped with an Intel(R) Core(TM) i7-11800 H processor clocked at 2.30GHz, 2304MHz, six cores, 12 logical processors, 8 GB of RAM, and a 512GB SSD.

4.1.1 Data acquisition and descriptions

Six real-world data sets with distinct qualities were utilized. The primary reasons we considered the data sets are the wide range of characteristics, accessibility, and potential to make the study results repeatable, and they consist of techniques to do supervised and unsupervised learning on graph structure data where predictive, recommendation, and analytic approach are flexible with the real-world data set.

4.1.2 Data description

1. The Cora data set (Kipf & Welling, 2016), a citation network, comprises 2708 scientific publications. The nodes are categorized into one of seven subject classes. There are 5429 connections in the citation network. Nodes represent science articles, and the left node mentions the right node when an edge connects the two nodes. Each publication in the data set is described by a 0/1-valued word vector indicating the absence/presence of the corresponding word from the dictionary. One thousand four hundred thirty-three distinct words make up the dictionary.
2. Dolphin Data set, authored by Lusseau (2006) identifies bottle nose dolphin point locations in Doubtful Sound. It consists of 62 nodes and 159 undirected edges with three community numbers, where a link represents frequent associations between dolphins.
3. Jruvika has assembled the Fake News Detection (Kumar et al., 2021) data set on the Kaggle platform. Its four properties are site URL, Headline, Body, and Label (Real/Fake News). There were 4009 new occurrences in the data set at first. Following the first data cleaning steps, which included deleting entries with incorrect labels, missing headlines, and body content, 3988 rows containing 1867 Real and 2121 Fake news samples were obtained. The majority of articles focus on political and World news topics.
4. Kyphosi is a spin-related unusually large convex curvative. The 81 records with four attributes for each patient that underwent corrective spinal surgery in the kyphosis data set (John & Trevor, 1992), which was retrieved from Kaggle <https://www.kaggle.com/abbasit/kyphosisdata>. A factor denoted present indicates a type of deformation was present after the surgery, suggesting that the patient may be recommended to undergo another surgery. Some key attributes of the data employed for analysis include the ages of patients, the number of patients involved, and the start date, which indicates the day a patient is operated upon.
5. The supermarket data set, containing historical sales data from three branches for 3 months includes 1000 rows and 17 columns from Kaggle. It includes information on invoice ID, branch, city, customer type, gender, product line, unit price, tax, total, date, time, payment, COGS(cost of goods sold) gross margin percentage, gross income, and rating.
6. Zachary's Karate Club data, a university karate social network, was developed by Zachary (1977) is the final data set used in this investigation. Michelle Girvan's 2002, makes use of a variant of Zachary's data, popularized multi-layer graphs for illustrating community structures in networks. The data has 34 pairs of nodes and 78 edges. Each node represents a karate member, and a pair indicates the two members had interacted.

The characteristics for each data are listed in Table 3, describing the six data used to assess the efficiency of the Keyword Search Co-occurrence technique. The sets N and E in Table 3 correspond to the MLGs' Nodes and edges, with class denoting the respective class.

The specific features of the data set and extension of Table 3 are listed in Table 4, where N and E are nodes or rows and edges numbers or columns, respectively, with class representing their respective classes, k being the average degree of the graphs, and DD and DU for directed and undirected, respectively, denoting density. The diameter (DIA), radius (RA), and average path length (APL) make up Graph DD and Graph DU , respectively, which indicate the Graph distance for directed and undirected operations. For directed and undirected networks, respectively, CCD and CCU make up the clustering coefficient (CC) of the network. Every network has both directed and undirected linkages, which is important to note.

4.1.3 Baseline methods

1. The principle of multi-layer embedding (Kumar et al., 2021) proposed three methods of multi-layer network into a continuous vector space.
2. Kumar et al. (2020) employ vertex attributes using the degree of overlapping between keywords research embedded with other features in the co-authorship work.
3. Deep learning is a member of the machine learning family of techniques, which is a subset of artificial intelligence and artificial neural networks, which are modeled after biological neural networks. Chauhan et al. (2023) presented a supervised machine learning and deep learning model for diagnosing kyphosis disease.
4. A research direction is to explore (Ma et al., 2021) whether to design a multi-layer graph embedding method that can naturally learn distance/similarity.

4.1.4 Model parameter settings and training

The basic size of the data sets can vary from hundreds of thousands of nodes, and edges can interact simultaneously; the elements u , v , and k (nodes) are represented in binary form with the values 1 and 0, respectively. Eighty percent of each data set was used to train the model, while the final twenty percent was used as the test set.

Table 3 Description of Multi-Layer Graphs Data Sets

| Topological feature of real-world data sets | | | |
|---|------|-------|-------|
| Networks | N | E | Class |
| Cora data set | 2708 | 10557 | 7 |
| Dolphin data set | 62 | 159 | 3 |
| Fake New data set | 3988 | | 2 |
| Kyhposi data set | 81 | 4 | 2 |
| Supermarket data set | 1000 | 17 | 3 |
| Zachary Karate data set | 34 | 78 | 3 |

Table 4 Characteristic of multi-layer graphs data sets

| N | N | E | Class | K | Density | | Graph DD | | | Graph DU | | | CC | |
|---|------|-------|-------|-------|---------|-------|----------|----|-------|----------|----|-------|-------|-------|
| | | | | | DD | DU | DIA | RA | AVL | DIA | RA | AVL | CCD | CCU |
| C | 2708 | 10557 | 7 | 2.693 | 0.001 | 0.002 | 1 | 0 | 1.0 | 6 | 1 | 3.82 | 0.0 | 0.0 |
| D | 62 | 159 | 3 | 5.129 | 0.000 | 0.084 | 0 | 0 | 0 | 8 | 5 | 3.36 | 0 | 0.303 |
| F | 3988 | – | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| K | 81 | 4 | 2 | 1.152 | 0.015 | 0.030 | 1 | 0 | 1.0 | 4 | 1 | 2.408 | 0.0 | 0.0 |
| S | 1000 | 17 | 3 | 2.728 | 0.000 | 0.001 | 1 | 0 | 1.0 | 4 | 1 | 3.622 | 0.0 | 0.0 |
| Z | 34 | 78 | 3 | 4.588 | 0.070 | 0.139 | 3 | 0 | 1.273 | 5 | 3 | 2.408 | 0.285 | 0.588 |

4.1.5 Hypothesis

The interdependence structure between nodes and edges might contain helpful information that leads to conclusive and supporting decision-making. Let P be the probability relation that meets certain requirements: a multi-layer graph may be a directed relation or a symmetric and transitivity relation. True with hypothesis if support $H_\theta(x) \geq 0.5$ and False if support $H_\theta(x) < 0.5$ the proposed approach is $0 \leq H_\theta(x) \leq 1$.

4.2 Evaluation metrics

Understanding the Effective Keyword Search Co-occurrence on Multi-Layer Graphs is the main goal of our evaluation. The performance of EKSCOMLGs and baseline methods are validated using quantitative measurements. Wilcoxon Rank Sample, Accuracy, Precision, Recall, F-measures, and support are used to assess the quality of trained classifiers.

The area under the curve (AUC) can be used to enumerate the vertices of a graph but cannot capture certain aspects of user satisfaction. Precision, recall, and f1 measures receive more attention than accuracy in our study. The confusion matrix of a binary classifier is shown in the Eqs. 1 to 4 (Yang et al., 2021; Kumar et al., 2020).

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

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

$$\text{F - measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

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

Intuitively, recall measures how well the search engine finds all the co-occurrence graph items for a query, and precision measures how well it rejects non-occurrence graph items.

True Negative (TN). It is recommended to set the Class 0 non-occurrence data item to 0 rather than co-occurrence (the pattern does not link.)

True Positive (TP). The appropriate data item (Class 1) is advised as 1 and co-occurrence (pattern corresponding to links exist.)

False Negative (FN). A connection that is a part of the Graph's co-occurrence data item (Class 1) is advised to be 0 and not co-occurrence (the pattern does not link.)

False Positive (FP). It is advised to treat the Class 0, not co-occurrence data item, as 1 and co-occurrence (pattern corresponding to links exist.)

5 Results and discussion

5.1 Performance evaluation of EKSCOMLGs using Wilcoxon rank sum test

The Wilcoxon signed-rank test is a non-parametric numerical hypothesis test that can be employed to evaluate two networks' regions using two corresponding samples, assess a network using a sample of data, or carry out a paired difference test of recurring quantities on a single sample to ascertain whether the network mean ranks differ.

5.1.1 Test procedure

Two versions of the signed-rank test exist. The one-sample test is essential since it allows for the linked sample test to be obtained by modifying the data to correlate with the one-sample test's criterion. Linked data, however, is where most of the signed-rank test's practical claims originate. The data includes samples $\{(X_1, Y_1), \dots, (X_n, Y_m)\}$ for a paired sample test.

Every sample comprises two capacities; these capacities can be converted to absolute numbers or an interval scale in the most basic scenario. The linked sample test can be modeled to a one-sample test by changing every edge of values (X_{u_i}, Y_{v_j}) with their difference, $X_{u_i} - Y_{v_j}$. Generally speaking, the alterations between the pairs must be ranked reasonably plausible. An ordered metric scale, which may have less evidence than an interval scale but carries more than an ordinary scale, is required for the data. Consequently, four real-world data sets—The Dolphin, Kyphosis, Supermarket, and Zachary Karate—are used to sample the Wilcoxon signed-rank test. Tables 5, 6, 7, 8, and 9 represent the general data set for all the real-world data.

5.1.2 Wilcoxon rank sum test using dolphin data set

For the Dolphin data sets, we sample the top 20 elements from the data sets as shown in Table 6.

Claim: The probability of Social Dolphin Co-occurrence using 20 sample data sets where the degree of the sample dolphin is used to rank the entire data sets from the lowest to the highest degree.

Claim $n_1 = n_2$, $H_o = n_1 = n_2$, $H_A = n_1 \neq n_2$, $\alpha = 0.05$. The value of $T_1 = n_1 = 20$, $T_2 = n_2 = 20$.

Claim: n_1 Co-occurrence among social dolphins = n_2 Absent of Co-occurrence among social dolphins

$H_o = n_1$ Co-occurrence present = n_2 Co-occurrence absent

$H_A = n_1$ Co-occurrence present \neq n_2 Co-occurrence absent.

Ranking the sample data sets using $n_1=20$ and $n_2=20$, from the lowest to the highest degree, we observe that the value for the 20 samples from the top selected data sets is grouped into 2 communities. Ranking the entire sample we observe that some samples dolphin, such as Degree 3, appear twice, 4 degrees appear five times, 5 appear three times, 6 degrees appear nine times, 7 degrees appear seven times, 8 degrees appear six times, 9 degrees appear five times and 12 degrees appear 3 times with a total of the

Table 5 Wilcoxon rank sum test using dolphin data set

| Dolphin data set comprises of 3 communities | | | | | |
|---|--------|-------------|--------|-----------|--------|
| Group 1 | Degree | Group 2 | Degree | Group 3 | Degree |
| Beak | 6 | Beescratch | 8 | CCL | 3 |
| Bumper | 4 | DN16 | 4 | Cross | 1 |
| Fish | 5 | DN21 | 6 | Double | 6 |
| 'Fork | 1 | DN63 | 5 | Five | 1 |
| Grin | 12 | Feather | 7 | Haecksel | 7 |
| Hook | 6 | Gallatin | 8 | Jonah | 7 |
| Kringel | 9 | Jet | 9 | MN105 | 6 |
| SN4 | 11 | Knit | 4 | MN60 | 3 |
| SN63 | 8 | MN23 | 1 | MN83 | 6 |
| SN9 | 8 | Mus | 3 | Patchback | 9 |
| SN96 | 6 | Notch | 3 | SMN5 | 1 |
| Scabs | 10 | Number1 | 5 | Topless | 11 |
| Shmuddel | 5 | Oscar | 5 | Trigger | 10 |
| Stripes | 7 | PL | 5 | Vau | 2 |
| TR120 | 2 | Quasi | 1 | Zap | 5 |
| TR77 | 6 | Ripplefluke | 3 | | |
| TR88 | 1 | SN90 | 5 | | |
| TR99 | 7 | TR82 | 1 | | |
| TSN103 | 4 | Upbang | 7 | | |
| TSN83 | 2 | Wave | 2 | | |
| Thumper | 4 | Web | 9 | | |
| Whitetip | 1 | Zig | 1 | | |
| Zipfel | 3 | | | | |

entire data sets = 820. Where our $N_1 = 20$ and $N_2 = 20$ $T_1 = 491.5$ and $T_2 = 328.5$. $T_2 = 328.5$ is chosen for testing the two groups.

One set of vital values for one-tail $\alpha = 0.025$ and two-tail $\alpha = 0.05$ and another set for one-tail $\alpha = 0.05$ and two-tail $\alpha = 0.10$ exist for every pair of sample scopes (m, n), according to the Wilcoxon Rank-Sum Test Table of Critical Values. The sample size for the smallest sample is shown in column m , while the sample for the largest is in column n . Either sample can be named m if the sample sizes are equal. Assume $m = 20$ and $n = 20$ for a two-tailed test at $\alpha = 0.05$. Both $n = 20$ and $m = 20$ are given. It is asserted that the social dolphins' probability distribution is comparable. The following numbers can be found in the relevant row and column: 483, 337. The minimum and maximum critical values for WX , the testing statistic H_0 : $MX = MY$, are 337 and 483. H_0 would be rejected if $WX \leq 337$ or $WX \geq 483$ while Fig. 4 shows the association between dolphin social networks and friends.

5.1.3 Wilcoxon rank sum test using kyphosis data set

Claim: The probability of kyphosis present or absent using 20 samples from kyphosis data sets using the age of the patient as the factor to rank the entire data sets starting from the minimum age to the maximum in the selected Table 7.

Table 6 Wilcoxon rank sum test using dolphin data set

| Top 20 Samples from Dolphin Dataset | | | |
|-------------------------------------|--------|------------|--------|
| Sample A | Degree | Sample B | Degree |
| Double | 6 | CCL | 3 |
| Feather | 7 | DN16 | 4 |
| Feather | 7 | DN21 | 6 |
| Fish | 5 | Beak | 6 |
| Fish | 5 | Bumper | 4 |
| Gallatin | 8 | DN16 | 4 |
| Gallatin | 8 | DN21 | 6 |
| Gallatin | 8 | Feather | 7 |
| Grin | 12 | Beak | 6 |
| Grin | 12 | CCL | 3 |
| Haeckel | 7 | Beak | 6 |
| Hook | 6 | Grin | 12 |
| Jet | 9 | Beescratch | 8 |
| Jet | 9 | DN21 | 6 |
| Jet | 9 | Feather | 7 |
| Jet | 9 | Gallatin | 8 |
| Jonah | 7 | Haeckel | 7 |
| Knit | 4 | Beescratch | 8 |
| Knit | 4 | DN63 | 5 |
| Kringel | 9 | Double | 6 |

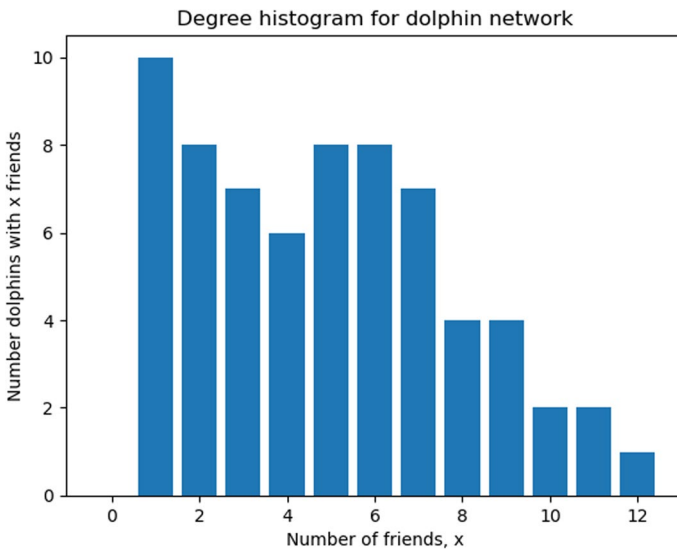


Fig. 4 Dolphin social network representative-based associations

Claim $n_1 = n_2$, $H_o = n_1 = n_2, H_A = n_1 \neq n_2, \alpha = 0.05$. The value of $T_1 = n_1 = 10$, $T_2 = n_2 = 10$.

Table 7 Wilcoxon Rank Sum Test using Kyphosi data set

| Kyphosi data set comprises deformation disease | | | | |
|--|----------|-----|--------|-------|
| Number | Kyphosis | Age | Number | Start |
| 24 | Present | 15 | 7 | 2 |
| 39 | Present | 91 | 5 | 12 |
| 38 | Absent | 20 | 6 | 9 |
| 50 | Absent | 68 | 5 | 10 |
| 31 | Absent | 125 | 2 | 11 |
| 49 | Absent | 177 | 2 | 14 |
| 42 | Absent | 143 | 9 | 3 |
| 19 | Absent | 27 | 4 | 9 |
| 4 | Absent | 1 | 4 | 15 |
| 71 | Absent | 127 | 4 | 12 |
| 20 | Absent | 22 | 2 | 16 |
| 29 | Absent | 151 | 2 | 16 |
| 14 | Absent | 168 | 3 | 18 |
| 64 | Absent | 118 | 4 | 16 |
| 26 | Absent | 8 | 3 | 6 |
| 37 | Present | 52 | 5 | 6 |
| 41 | Absent | 35 | 3 | 13 |
| 66 | Absent | 195 | 2 | 17 |
| 11 | Absent | 148 | 3 | 16 |
| 52 | Present | 139 | 10 | 6 |

Claim: n_1 Kyphosis present = n_2 Kyphosis absent

H_o = n_1 kyphosis present = n_2 kyphosis absent

H_A = n_1 kyphosis present \neq n_2 kyphosis absent

Ranking the sample data sets using $n_1=10$ and $n_2=10$, from the minimum to maximum, we observe that the value for the 20 samples is 210 consisting of 4 samples showing patients with kyphosis disease present and 16 samples showing that kyphosis disease is absent summing the total ranking number we have where $T_1 = 105$ and $T_2 = 105$ for the 20 sample data. Using the Wilcoxon Rank-Sum Test Critical Values Table, assume a two-tailed test at $\alpha = 0.05$, we have $m = 10$ and $n = 10$. The claim is that the probability distribution associated with the kyphosis disease is equivalent. In the appropriate row and column, we find 78, 132, 78, and 132, the minimum and maximum critical values for WX ; the testing statistic $H_0: MX = MY$. If $WX \leq 78$ or $WX \geq 132$, H_0 would be rejected while Fig. 5 shows the Kyphosi disease representation based on the patients' age.

5.1.4 Wilcoxon rank sum test using supermarket data set

Claim: The association denote the Co-purchase of product in a supermarket using Supermarket data sets grouped into three branches, A, B, and C, using the branch and product purchase as the factor to rank the entire data sets starting from the minimum to the maximum where only 10 samples whereas selected whereas selected from Table 8, we realized that using the 3 branches, branch A has the highest number of top 10 sample data followed by Branch C and B. Ranking the entire sample from the minimum to the maximum, we

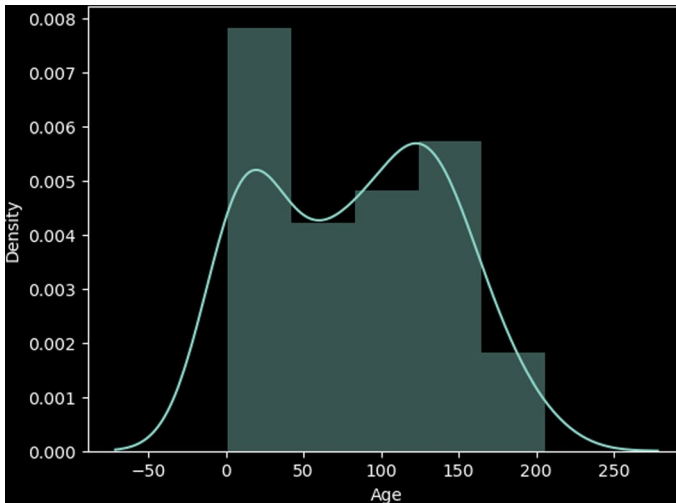


Fig. 5 kyphosis disease representative-based age

observe that some product purchases in branch A appear 6 times, C appears 3 times and B once with a product such as Health and Beauty and electronic and accessories appear 3 times, each, Home and lifestyle appear 2. In contrast, sports and traveling and food and beverages appear once each. Claim: $n_1 = n_2, n_3$ $H_o = n_1 = n_2$ or $n_1 = n_3$, $H_A = n_1 \neq n_2$ $H_A = n_1 \neq n_3$, $\alpha = 0.05$ The value of $T_1 =$ is the 3 branch of supermarket, $T_2 =$ The product purchased from the 3 branches. Claim: n_1 Co-purchase of produce exists in the three branches $= n_2$ Co-purchase didn't exist among members. n_3 There is a Likelihood of mutual existence among buyers. $H_o = n_1$ Co-purchase present $= n_2$ Co-purchased absent, n_3 Likelihood of mutual purchase of products. $H_A = n_1$ Mutual purchase occur $\neq n_2$ Mutual purchase is absent $T_1 = 55$ and $T_2 = 55$. Ranking the sample data sets using $n_1 = 10$ and $n_2 = 10$, and $n_3 = 10$ from the minimum to maximum, we observe that the value for the 10 samples is 55 where $T_1 = 55$ and $T_2 = 55$. The following numbers, 78 and 132, can be found by using the Table of Critical Values for the Wilcoxon Rank-Sum Test and assuming that, for a two-tailed test at $\alpha = 0.05$, we have $m = 10$ and $n = 10$. The argument is that the probability of sales distribution in the three branches is identical. The statistic testing $H_0: MX = MY$ has lower and higher critical values of 78 and 132 for WX . H_0 would be denied if $WX \leq 78$ or $WX \geq 132$ Fig. 6 shows the co-purchase between three branches of a supermarket based on gender.

5.1.5 Wilcoxon rank sum test using Zachary's data set

Claim: The association of Zachary's relationship using Zachary's data grouped into three communities using the degree as the factor to rank the entire data sets starting from the lowest to the highest age. Ranking the entire sample from the minimum to the maximum, as shown in Table 9, we observe that some sample data such as 1, 9, 10, 12, 16, and 17 appear once in the sample data sets, 2 appear eleven times, 3 appear 6 times, 4 appear 6 times, 5 appear 3 times, and 6 appear 2 times.

Claim: $n_1 = n_2, n_3$ $H_o = n_1 = n_2$ or $n_1 = n_3$, $H_A = n_1 \neq n_2$ $H_A = n_1 \neq n_3$, $\alpha = 0.05$ The value of $T_1 = n_1 = 10$, $T_2 = n_2 = 10$ and $T_3 = n_3 = 14$

Table 8 Wilcoxon rank sum test using supermarket data set

Supermarket data set comprises of 3 branches

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|-------------------------|-------------------|------------------------|--------------------|-------------------|-------------------|------------------------|------------------------|--------------------|-------------------|--------------------|
| Invoice ID | 750-67-8428 | 226-31-3081 | 631-41-3108 | 123-19-1176 | 373-73-7910 | 699-14-3026 | 355-53-5943 | 315-22-5665 | 665-32-9167 | 692-92-5582 |
| Branch | A | C | A | A | A | C | A | C | A | B |
| City | Yangon | Naypyitaw | Yangon | Yangon | Yangon | Naypyitaw | Yangon | Naypyitaw | Yangon | Mandalay |
| Customer type | Member | Normal | Member | Member | Normal | Normal | Member | Normal | Member | Member |
| Gender | Female | Female | Male | Male | Male | Male | Female | Female | Female | Female |
| Product line | Health and beauty | Electronic accessories | Home and lifestyle | Health and beauty | Sports and travel | Electronic accessories | Electronic accessories | Home and lifestyle | Health and beauty | Food and beverages |
| Unit price | 74.69 | 15.28 | 46.33 | 58.22 | 86.31 | 85.39 | 68.84 | 73.56 | 36.26 | 54.84 |
| Quantity | 7 | 5 | 7 | 8 | 7 | 7 | 6 | 10 | 2 | 3 |
| Tax 5% | 26.1415 | 3.82 | 16.2155 | 23.288 | 30.2085 | 29.8865 | 20.652 | 36.78 | 3.626 | 8.226 |
| Total | 548.9715 | 80.22 | 340.5255 | 489.048 | 634.3785 | 627.6165 | 433.692 | 772.38 | 76.146 | 172.746 |
| Date | 1/5/2019 | 3/8/2019 | 3/3/2019 | 1/27/2019 | 2/8/2019 | 3/25/2019 | 2/25/2019 | 2/24/2019 | 1/10/2019 | 2/20/2019 |
| Time | 13:08 | 10:29 | 13:23 | 20:33 | 10:37 | 18:30 | 14:36 | 11:38 | 17:15 | 13:27 |
| Payment | Ewallet | Cash | Credit card | Ewallet | Ewallet | Ewallet | Ewallet | Ewallet | Credit card | Credit card |
| cogs | 522.83 | 76.4 | 324.31 | 465.76 | 604.17 | 597.73 | 413.04 | 735.6 | 72.52 | 164.52 |
| gross margin percentage | 4.761905 | 4.761905 | 4.761905 | 4.761905 | 4.761905 | 4.761905 | 4.761905 | 4.761905 | 4.761905 | 4.761905 |
| gross income | 26.1415 | 3.82 | 16.2155 | 23.288 | 30.2085 | 29.8865 | 20.652 | 36.78 | 3.626 | 8.226 |
| Rating | 9.1 | 9.6 | 7.4 | 8.4 | 5.3 | 4.1 | 5.8 | 8 | 7.2 | 5.9 |

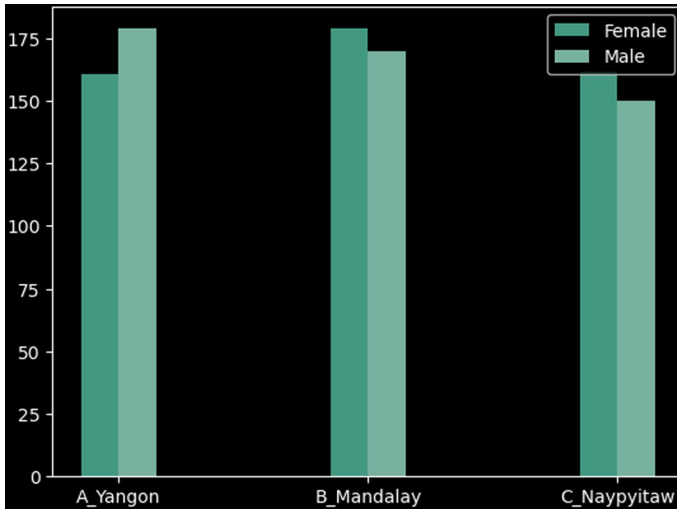


Fig. 6 Supermarket representative based on three branches and genders

Table 9 Wilcoxon rank sum test using zachary karate data set

| Zachary Karate data set comprises of 3 communities | | | | | |
|--|--------|-------|--------|-------|--------|
| Node1 | Degree | Node2 | Degree | Node3 | Degree |
| 9 | 5 | 2 | 9 | 1 | 16 |
| 15 | 2 | 3 | 10 | 5 | 3 |
| 16 | 2 | 4 | 6 | 6 | 4 |
| 19 | 2 | 8 | 4 | 7 | 4 |
| 21 | 2 | 10 | 2 | 11 | 3 |
| 23 | 2 | 13 | 2 | 12 | 1 |
| 24 | 5 | 14 | 5 | 17 | 2 |
| 25 | 3 | 18 | 2 | 20 | 3 |
| 26 | 3 | 22 | 2 | | |
| 27 | 2 | | | | |
| 28 | 4 | | | | |
| 29 | 3 | | | | |
| 30 | 4 | | | | |
| 31 | 4 | | | | |
| 32 | 6 | | | | |
| 33 | 12 | | | | |
| 34 | 17 | | | | |

Claim: n_1 Co-occurrence exist among member = n_2 Co-occurrence didn't exist among member. n_3 There is a likelihood of mutual existence among members.

$H_o = n_1$ Co-occurrence present = n_2 Co-occurrence absent, n_3 Likelihood of mutual occurrence. $H_A = n_1 \neq n_2$ Mutual relation is absent.

Ranking the sample data sets using $n_1=17$ and $n_2=9$, and $n_3=8$ from the minimum to maximum, we observe that the value for the 34 samples is 605 where $T_1 = 309.5$ and $T_2=$

165 and $T_3 = 130.5$. Assume for a two-tailed test at $\alpha = 0.05$ that we have $m = 17$ and $n = 9$ and $n = 8$. The Table of Critical Values for the Wilcoxon Rank-Sum Test is utilized. It is asserted that the Zachary Karate Club's probability distribution is comparable.

For the two-tailed test, we make use of $n_1=17$ and $n_2=9$, $T_1 = 309.5$ and $T_2 = 165$ followed by $n_1=17$ and $n_3=8$ utilizing $T_1 = 309.5$ and $T_3 = 130.5$ where the minimum value serves as our test value for the sample data sets. We find 84 and 159 numbers for n_1 and n_2 in the appropriate row and column. The 84 and 159 are the minimum and maximum critical values for WX ; the testing statistic $H_0: MX = MY$. If $WX \leq 84$ or $WX \geq 159$, H_0 would be denied.

We find the following numbers, 70 and 138, for n_1 and n_3 . The 70 and 138 are the lower and upper critical values for WX ; the statistic testing $H_0: MX = MY$. If $WX \leq 70$ or $WX \geq 138$, H_0 would be denied. Figure 6 shows the Zachary karate network representative using club members (Fig. 7).

5.2 Performance using graph machine learning and deep learning algorithm

Using machine learning models, such as (a) Logistic Regression (LR), (b) Gradient Boosting Classifier GBC (C) Random Forest (RF) classifier, and (D) K Nearest Neighbor (Alimadadi et al., 2019), we examine the performance of EKSCOMLGs using Logistic regression and Gradient boosting classifier. Machine learning relies heavily on LR, especially when dealing with categorization issues. This algorithm performs exceptionally well in situations where one of two possible outcomes is a diagnosis of a medical problem or the behavior of an application in the real world. In real-world applications, minimizing the loss function in logistic regression is often achieved through gradient-boosting classifiers. Eighty percent of each data set was used for training, and the remaining twenty percent was utilized as the test set to compare the performance of the proposed technique with that of the existing methods using metrics like Precision, Recall, and F1-Score. Precision-Recall

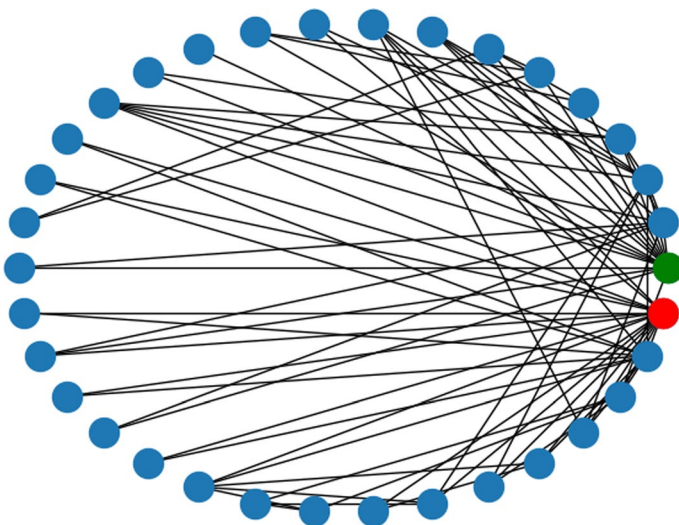


Fig. 7 Zachary social network representative

Table 10 Analysis description using cora data set

| Cora data set | | | | | |
|---------------|---------|-----------|--------|---------|----------|
| Networks | Results | Precision | Recall | F-Score | Accuracy |
| LR | 0 | 0.92 | 0.90 | 0.91 | 0.91 |
| | 1 | 0.90 | 0.92 | 0.91 | 0.91 |
| | W | 0.91 | 0.91 | 0.91 | |
| GBC | 0 | 0.87 | 0.87 | 0.87 | 0.88 |
| | 1 | 0.89 | 0.89 | 0.89 | 0.88 |
| | W | 0.88 | 0.88 | 0.88 | |
| RFC | 0 | 0.82 | 0.87 | 0.85 | 0.85 |
| | 1 | 0.88 | 0.84 | 0.86 | 0.85 |
| | W | 0.86 | 0.85 | 0.86 | |
| KNN | 0 | 0.84 | 0.88 | 0.86 | 0.86 |
| | 1 | 0.89 | 0.85 | 0.87 | 0.86 |
| | W | 0.87 | 0.86 | 0.87 | |

Table 11 Analysis description using dolphin data set

| Dolphine data set | | | | | |
|-------------------|---------|-----------|--------|---------|----------|
| Networks | Results | Precision | Recall | F-Score | Accuracy |
| LR | 0 | 0.86 | 0.97 | 0.91 | 0.91 |
| | 1 | 0.97 | 0.86 | 0.91 | 0.91 |
| | W | 0.92 | 0.91 | 0.91 | |
| GBC | 0 | 1.00 | 1.00 | 1.00 | 1.00 |
| | 1 | 1.00 | 1.00 | 1.00 | 1.00 |
| | W | 1.00 | 1.00 | 1.00 | |
| RFC | 0 | 1.00 | 0.73 | 0.84 | 0.85 |
| | 1 | 0.75 | 1.00 | 0.86 | 0.85 |
| | W | 0.89 | 0.85 | 0.85 | |
| KNN | 0 | 1.00 | 0.73 | 0.84 | 0.85 |
| | 1 | 0.75 | 1.00 | 0.86 | 0.85 |
| | W | 0.89 | 0.85 | 0.85 | |

and F1 Measure are used to summarize the various machine-learning models for real-world data sets in Tables 10 through Table 16. The cross-validation represents a stratified study that is ten-fold and five-fold.

GML analysis Table 10 shows that LR has the highest value in terms of precision Recall F1-Score and Accuracy, followed by GCB, KNN, and RFC. LR and GBC perform well using precision one and Recall zero, while RFC and KNN have similar outcomes in three analyses. LR and GBC performed well in almost all the data sets, with LR performing excellently using the Cora and Zachary karate data set. The general result shows that using LR and GCB yields 2–10% increment compared to other approaches as shown in Tables 10, 11, 12, 13, 14, 15, 16. While the Fake New Data set operated for days to provide an overall result, it was unsuccessful when employing KNN with its two neighbors to achieve accuracy. It explains the reason why the KNN result for the Fake Data set is not

Table 12 Analysis description using kyphosi data set

| kyhposi data set | | | | | |
|------------------|--------------|-----------|--------|---------|----------|
| Networks | Results | Precision | Recall | F-Score | Accuracy |
| LR | Absent | 0.89 | 0.92 | 0.90 | 0.84 |
| | Present | 0.64 | 0.54 | 0.58 | 0.84 |
| | Weighted avg | 0.84 | 0.84 | 0.84 | |
| GBC | Absent | 1.00 | 1.00 | 1.00 | 1.00 |
| | Present | 1.00 | 1.00 | 1.00 | 1.00 |
| | Weighted avg | 1.00 | 1.00 | 1.00 | |
| RFC | Absent | 0.75 | 0.92 | 0.83 | 0.71 |
| | Present | 0.00 | 0.00 | 0.00 | 0.71 |
| | Weighted avg | 0.57 | 0.71 | 0.63 | |
| KNN | Absent | 0.73 | 0.85 | 0.79 | 0.65 |
| | Present | 0.00 | 0.00 | 0.00 | 0.65 |
| | Weighted avg | 0.56 | 0.65 | 0.60 | |

Table 13 Analysis description using fakew data set

| Fake new data set | | | | | |
|-------------------|---------|-----------|--------|---------|----------|
| Networks | Results | Precision | Recall | F-Score | Accuracy |
| LR | 0 | 0.99 | 0.99 | 0.99 | 0.99 |
| | 1 | 0.98 | 0.99 | 0.99 | 0.99 |
| | W | 0.99 | 0.99 | 0.99 | |
| GBC | 0 | 1.00 | 0.99 | 1.00 | 1.00 |
| | 1 | 0.99 | 1.00 | 1.00 | 1.00 |
| | W | 1.00 | 1.00 | 1.00 | |
| RFC | 0 | 0.99 | 0.99 | 0.99 | 0.99 |
| | 1 | 0.99 | 0.99 | 0.99 | 0.99 |
| | W | 0.99 | 0.99 | 0.99 | |

available. The keyword co-occurrence graph found the proposed strategy to be helpful in accurately and consistently guiding the potential MLG linkages across data sets and methodologies, according to experimental results.

6 Related work

6.1 Graph theory and vital application

A graph represents binary, multiple associations among a person's contents and thus is a prevalent data structure. Several essential tools are typically used for real-world applications, like the Greedy Search technique for Graph Mining, the Inductive Database Search technique for Graph Mining, and the Graph Clustering technique for Graph Mining, which describes achieving more enhancements (Han et al., 2022; Scarselli et al., 2008). When well-educated heuristics are available to direct the search, greed search is a

Table 14 Analysis description using supermarket

| Supermarket data set | | | | | |
|----------------------|---------|-----------|--------|---------|----------|
| Networks | Results | Precision | Recall | F-Score | Accuracy |
| LR | 0 | 0.97 | 0.96 | 0.97 | 0.97 |
| | 1 | 0.96 | 0.98 | 0.97 | 0.97 |
| | W | 0.97 | 0.97 | 0.97 | |
| GBC | 0 | 1.00 | 0.99 | 0.99 | 0.99 |
| | 1 | 0.99 | 1.00 | 1.00 | 0.99 |
| | W | 1.00 | 0.99 | 0.99 | |
| RFC | 0 | 0.98 | 0.98 | 0.98 | 0.98 |
| | 1 | 0.98 | 0.98 | 0.98 | 0.98 |
| | W | 0.98 | 0.98 | 0.98 | |
| KNN | 0 | 0.94 | 0.94 | 0.94 | 0.94 |
| | 1 | 0.94 | 0.94 | 0.94 | 0.94 |
| | W | 0.94 | 0.94 | 0.94 | |

Table 15 Analysis description using zachary karate datasets

| Zachary Karate Data set | | | | | |
|-------------------------|---------|-----------|--------|---------|----------|
| Networks | Results | Precision | Recall | F-Score | Accuracy |
| LR | 0 | 0.98 | 1.00 | 0.99 | 0.99 |
| | 1 | 1.00 | 0.97 | 0.99 | 0.99 |
| | W | 0.99 | 0.99 | 0.99 | |
| GBC | 0 | 0.90 | 0.90 | 0.90 | 0.90 |
| | 1 | 0.90 | 0.90 | 0.90 | 0.90 |
| | W | 0.90 | 0.90 | 0.90 | |
| RFC | 0 | 0.90 | 0.90 | 0.90 | 0.90 |
| | 1 | 0.90 | 0.90 | 0.90 | 0.90 |
| | W | 0.90 | 0.90 | 0.90 | |
| KNN | 0 | 0.90 | 0.90 | 0.90 | 0.90 |
| | 1 | 0.90 | 0.90 | 0.90 | 0.90 |
| | W | 0.90 | 0.90 | 0.90 | |

Table 16 Cross-validation for keyword co-occurrence graph

| Network | Training (0) (%) | Testing (0) (%) | Training (1) (%) | Testing (1) (%) |
|---------------------|------------------|-----------------|------------------|-----------------|
| Cora data | 91 | 0 | 86 | 100 |
| Dolphin data | 88 | 0 | 94 | 0 |
| Kyphosis data | 82 | Absent | 95 | Absent |
| Supermarket | 98 | 100 | 96 | 100 |
| Zachary Karate data | 100 | 0 | 100 | 0 |

successful and effective technique for searching an intractably ample space. When optimizing or minimizing an objective function is required, greed searches are employed. Greedy algorithms, in contrast to backtracking, must determine the best option all at once and are unable to reverse their conclusion.

The idea of searching databases of graphs for (subgraph) patterns and the application of particular data structures that reflect the space of solutions define the inductive database technique for graph mining. For the former, it is required to have a query language for defining the patterns of interest. Although most applications of the latter focus on small molecule structure-activity relationships (SARs), they still attempt to provide a concise representation of the solution patterns. The graph mining strategy on multi-layer networks usually focuses on varying granularity depending on the job.

Graph clustering is an active technique for grouping data into different collections or clusters based on the similarity of the attributes and characteristics of the data points (Aggarwal, 2016; Boccaletti et al., 2014; Kivelä et al., 2014). Graph clustering is divided into two categories of tasks: (1). Developing a model to forecast a graph's class is the first task (2). Predicting node labels in big graphs is the second. However, considering the vast diversity of graph types and the information they can convey, the labeling costs associated with graph data are relatively significant. Multi-layered networks represent intricate connections found in contemporary networked information technology systems. Each pair of nodes in such a network may have multiple edges connecting them, each representing a distinct user activity related to cooperation or communication. For instance, the study (Huang et al., 2021) presents multi-layered degree centrality for multi-layered social networks, and (Bolorunduro & Zou, 2023) describes a practical application of centrality and depth-first search for community detection on multi-layer graphs based on intra-layer and inter-layer linkage graphs.

Graph Neural Networks (GNNs) are special neural networks or neural message-passing networks originally proposed for learning molecular graph representation that work with a graph data structure (Wu et al., 2020). They are highly influenced by Convolution Neural Networks (CNNs) and Graph Embedding. Graph Neural Networks (GNNs) have been extensively employed in graph illustration learning, attaining cutting-edge results in Node categorization, Link Prediction, and graph-based assignments. The essential idea of most of these methods is to formulate previous GNNs as a framework of neural message transmission among nodes or designed to learn node representations on fixed single graphs. At the same time, Graph Convolutional Networks (GCNs) that utilize aggregations are a distinct type of GNNs, and other models of GNNs based on different aggregations such as gated graph neural networks (Li et al., 2015) and graph attention networks (Velickovic et al., 2017) exist. The limitations especially become problematic when learning representations on a multi-layer graph consisting of various nodes and edges (Hamilton, 2020). GNNs were introduced when CNNs failed to accomplish optimal outcomes due to the arbitrary size of the graph and complex structure. Both shallow neural networks and deep neural networks face challenges despite their enormous success in learning graph representations; the existing GNN model has shown how susceptible they are to hostile examples that may exist in graph structure data. While (Yang et al., 2020) uses two network information-topology and node attributes-to collect semantic variance from the privileged group of actual and false samples, it must address the over-fitting issue. Although the proximity can represent underlying linkages within communities, there are not enough edges in sparsely connected real-world networks.

6.2 Keyword search co-occurrence graph

Searching over graphs has attracted much attention recently (Yang et al., 2021; Garg, 2021; Bast et al., 2016) because it gives helpful information without being aware of the underlying entities, schema, or access techniques. Search for information over massive, complicated graphs and various sophisticated keyword search algorithms have developed a connection between keyword search co-occurrence and an artificial index classification (Han et al., 2022; Rossi et al., 2021; Makarov et al., 2021). Thus, keyword search is fundamental to retrieving information most relevant to the query keywords. Latent Semantic Indexing (LSI) (Han et al., 2022; Aggarwal, 2016), known as Singular Value Decomposition, utilizes a matrix to the bipartite network of keywords and documents to assess similarity and generalized searches. However, LSI has two fundamental problems with vector space retrieval. (1) LSI cannot be expressed in Negation. (2) Boolean conditions cannot enforce it, and the SVD has a high computational cost.

By combining computer science and statistics, machine learning creates graph mining models that work better when exposed to relevant data than when given specific instructions. The benefit of Co-occurrence is focused on relatedness rather than similarity, which expresses how many traits two items share. Feature extraction is a primary problem in classical machine learning models, where the programmer must precisely specify the features that the computer is to be trained to detect. These attributes will facilitate decision-making. Deep neural networks are an option if simple pattern recognition remains problematic as pattern complexity increases. Capturing the latent information of the Keyword Search Co-occurrence analysis, our proposed method employs a multi-layer graph embedding and graph neural network for Effective keyword Search Co-occurrence Multi-layer Graph Mining.

6.3 Multi-layer graph embedding and graph neural networks

Property graphs are converted into a vector or a collection of vectors through graph embedding. Instead of focusing on a local structure, embedding approaches offer a global picture of latent relationships (Rossi et al., 2021; Makarov et al., 2021). Three basic inference tasks can be easily implemented in space using graph embedding: Finding a query vertex's closest neighbors in the embedding space is the first step in Node classification. The second step is to Link suggestions of nodes that will be connected in the future or missing, and Community Detection finding potential edges from the input graph is the third step. Before learning multi-layer representations, graph-based representation learning to graph embedding (such as Deep Walk (Perozzi et al., 2014), LINE(Tang et al., 2015), and node2vec (Grover & Leskovec, 2016)) characterize vertex neighborhoods using random walks techniques. However, these techniques are based on a single graph. As far as we know, a thorough investigation has yet to be done on the graph embedding technique for multi-layer graphs.

The decoder is an early technique for learning multi-relational embedding, called RESCAL, as described by Hamilton (2020). A critical family of decoders labels relationships as translations in the embedding space; TransE published their model in 2013. A second well-known type of research generalizes the dot-product decoder from graphs to build multi-relational decoders, as opposed to developing a decoder based on translating embedding. The method is commonly referred to as DISTMULT. One major limitation

of the DISTMULT decoder is that it cannot encode directed and diversified relations, which include most multi-relational graph relation types. Each relation type is individually embedded by Deep Graph Infomax for attributed Multiplex network embedding (DMGI) (Park et al., 2020), which then computes network embeddings to maximize globally shared features to detect communities. Through the discriminator, constructive learning takes place on each layer between the original network and a corrupted network.

Knowledge graph embedding techniques generate random walks and embedding vectors based on meta-path schemes. Meta Path Aggregated Graph Neural Networks (MAGNN) (Fu et al., 2020) provide a better community discovery method, which uses multi-information semantic meta-pathways to identify multi-layer structures in graph layers. By collecting semantic variants over nodes and meta paths, MAGNN leverages the attention mechanism in its embeddings. Hamilton (2020) embeds network schema and meta path are also gotten through Heterogeneous Information Networks.

The two works mentioned above employ meta pathways to promote community, but creating meaningful ones takes a great deal of topic expertise. By enlarging graph mining into a multi-layer network, a researcher proposes a generic multi-layer graph embedding framework that can be applied to any graph embedding approach model for single-layer graphs. Three approaches have been modeled to project a multi-layer network into a continuous vector space: "network aggregation," "results aggregation," and "layer co-analysis" (Liu et al., 2017). However, to consider the impact of interlayer interactions, "layer co-analysis" extends any single-layer network embedding technique to a multi-layer network. Our work differs from these approaches as we study how to perform Effective Keyword Search Co-occurrence Multi-layer Graph Mining utilizing enhanced Multi-layer Graph Embedding and Graph Neural Networks that provide insights about data and explainable conclusions.

7 Conclusions

While similarity search is helpful in many applications, multi-layer networks make meaningful measures of objects of diverse types more and more crucial. One compelling problem setting that arises from a real-world application, like a keyword search in large publication databases, is when the database can be viewed as an entity relation graph between the paper, authors, and words, or it can be used to characterize various kinds of connections (like clicks, favorites, adds, etc.). A different term for graph representation in low dimensional vectors that can be useful for network research tasks and edge and node prediction is graph embedding. The rapid emergence of graph neural networks, a technical mix of deep learning and graph data mining, illustrates their ability to model and capture complex relationships in graph-based data. Thus, an Effective keyword Search Co-occurrence that considers the significance and relatedness between nodes and edges in real-world applications where Multi-layer Graph Embedding and Graph Neural Network is utilized is presented using graph mining where users can utilize and locate communities that are related to them using our proposed KSCOMLGs. Furthermore, data relations from neighbors, edges, nodes, or multi-layer networks can be concurrently recognized with a particular focus on deep learning to attain efficient outcomes. Graph data are frequently noisy and imprecise in real applications; hence, it is usual to describe them as uncertainty graphs here. Each pair of edges is assigned a worth, indicating the chance it exists. Thus, a likely research direction is to extend this present work using uncertainty in medical data analysis.

Acknowledgements Authors do not take the editor and reviewer for granted and are genuinely grateful that they took the time to read our work. We sincerely appreciate it.

Author Contributions All authors contributed to the study's conception and design. Janet Oluwasola Bolorunduro Conducted the experiments, assembled and evaluated the data, and prepared the materials. Zhao-nian Zou modified Janet Oluwasola Bolorunduro's original text draft. The document was edited and proof-read by Mohamed Jaward Bah. The authors read the concluding manuscript and gave it high praise.

Funding This paperwork was partially supported by the National Natural Science Foundation of China(No 62072138) and the Open Research Projects of Zhejiang Lab(No2021KC0AB02)

Data availability The data sets for the analysis are currently available; for instance Cora data set is obtained from <https://relational.fit.cvut.cz/dataset/COis> obtainedRA, Dolphin data set from <https://sanchom.wordpress.com/tag/average-precision/>, Fake News data set from <https://www.kaggle.com/c/fake-news/data>, kyphosis data set, <https://www.kaggle.com/datasets/abbasit/kyphosis-dataset>, Supermarket dataset from, <https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales>, and Zachary karate data set from, <http://vlado.fmf.uni-lj.si/pub/networks/data/Ucinet/UciData.htm>.

Declarations

Conflict of interest The authors state that none of the research presented in this paper has been influenced by any personal or financial conflicts.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

References

- Aggarwal, C. C. et al. (2016). Recommender systems, Vol. 1, Springer.
- Alimadadi, F., Khadangi, E., & Bagheri, A. (2019). Community detection in facebook activity networks and presenting a new multilayer label propagation algorithm for community detection. *International Journal of Modern Physics B*, 33(10), 1950089.
- Bai, R., Wang, X. & Liao, J. (2009). Folksonomy for the blogosphere: Blog identification and classification. In *2009 WRI World Congress on Computer Science and Information Engineering*. Vol. 3, IEEE, pp. 631–635.
- Bast, H., Buchhold, B., & Haussmann, E. et al. (2016). Semantic search on text and knowledge bases. *Foundations and Trends® in Information Retrieval* 10(2–3), 119–271.
- Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., & Faulkner, R. et al. (2018). Relational inductive biases, deep learning, and graph networks. arXiv preprint [arXiv:1806.01261](https://arxiv.org/abs/1806.01261).
- Baxter, G. J., Cellai, D., Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. (2016). A unified approach to percolation processes on multiplex networks. *Interconnected networks* pp. 101–123.
- Bhalotia, G., Hulgeri, A., Nakhe, C., Chakrabarti, S. & Sudarshan, S. (2002). Keyword searching and browsing in databases using banks. In *Proceedings 18th international conference on data engineering*. IEEE, pp. 431–440.
- Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C. I., Gómez-Gardenes, J., Romance, M., Sendina-Nadal, I., Wang, Z., & Zanin, M. (2014). The structure and dynamics of multilayer networks. *Physics Reports*, 544(1), 1–122.
- Bojchevski, A., Shchur, O., Zügner, D. & Günnemann, S. (2018), Netgan: Generating graphs via random walks. In *International conference on machine learning*. PMLR, pp. 610–619.
- Bolorunduro, J. O. & Zou, Z. (2023). Community detection on multi-layer graph using intra-layer and inter-layer linkage graphs (cdmiilg). *Expert Systems with Applications* p. 121713.
- Bruna, J., Zaremba, W., Szlam, A. & LeCun, Y. (2013). Spectral networks and locally connected networks on graphs. [arXiv:1312.6203](https://arxiv.org/abs/1312.6203).
- Cai, H., Zheng, V. W., & Chang, K.C.-C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9), 1616–1637.
- Cao, N., Wang, C., Li, M., Ren, K., & Lou, W. (2013). Privacy-preserving multi-keyword ranked search over encrypted cloud data. *IEEE Transactions on Parallel and Distributed Systems*, 25(1), 222–233.

- Chauhan, A. S., Lilhore, U. K., Gupta, A. K., Manoharan, P., Garg, R. R., Hajje, F., Keshta, I., & Raahemifar, K. (2023). Comparative analysis of supervised machine and deep learning algorithms for kyphosis disease detection. *Applied Sciences*, *13*(8), 5012.
- Defferrard, M., Bresson, X. & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in Neural Information Processing Systems* **29**.
- Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A. & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*. **28**.
- Feng, G., Wang, H., & Wang, C. (2023). Search for deep graph neural networks. *Information Sciences*, *649*, 119617.
- Freilich, S., Kreimer, A., Meilijson, I., Gophna, U., Sharan, R., & Ruppim, E. (2010). The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Research*, *38*(12), 3857–3868.
- Fu, X., Zhang, J., Meng, Z. & King, I. (2020). Maggn: Metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of The Web Conference 2020*. pp. 2331–2341.
- Garg, M. (2021). A survey on different dimensions for graphical keyword extraction techniques: Issues and challenges. *Artificial Intelligence Review*, *54*, 4731–4770.
- Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, *151*, 78–94.
- Grover, A. & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 855–864.
- Hamilton, W. L. (2020). *Graph representation learning*. Morgan & Claypool Publishers.
- Hamilton, W., Ying, Z. & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*. **30**.
- Han, J., Pei, J. & Tong, H. (2022). Data mining: Concepts and techniques, Morgan kaufmann.
- He, H., Wang, H., Yang, J. & Yu, P. S. (2007). Blinks: Ranked keyword searches on graphs. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. pp. 305–316.
- He, Y., Gurukar, S., Kousha, P., Subramoni, H., Panda, D. K. & Parthasarathy, S. (2021). Distmile: a distributed multi-level framework for scalable graph embedding. In *2021 IEEE 28th International Conference on High Performance Computing, Data, and Analytics (HiPC)*. IEEE, pp. 282–291.
- Huang, X., Chen, D., Ren, T., & Wang, D. (2021). A survey of community detection methods in multilayer networks. *Data Mining and Knowledge Discovery*, *35*, 1–45.
- John, M. C., & Trevor, J. H. (1992). *Statistical models*. Wadsworth and Brooks/Cole.
- Kacholia, V., Pandit, S., Sudarshan, S., Desai, R. & Karambelkar, H. (2005). Bidirectional expansion for keyword search on graph databases.
- Kipf, T. N. & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. [arXiv:1609.02907](https://arxiv.org/abs/1609.02907).
- Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks*, *2*(3), 203–271.
- Kumar, A., Singh, S. S., Singh, K., & Biswas, B. (2020). Link prediction techniques, applications, and performance: A survey. *Physica A: Statistical Mechanics and its Applications*, *553*, 124289.
- Kumar, V., Kumar, A., Singh, A. K. & Pachauri, A. (2021). Fake news detection using machine learning and natural language processing. In *2021 International Conference on Technological Advancements and Innovations (ICTAI)*. IEEE, pp. 547–552.
- Li, Y., Tarlow, D., Brockschmidt, M. & Zemel, R. (2015). Gated graph sequence neural networks. [arXiv:1511.05493](https://arxiv.org/abs/1511.05493).
- Liang, J., Gurukar, S. & Parthasarathy, S. (2021). Mile: A multi-level framework for scalable graph embedding. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15, pp. 361–372.
- Liu, W., Chen, P.-Y., Yeung, S., Suzumura, T. & Chen, L. (2017). Principled multilayer network embedding. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, pp. 134–141.
- Lusseau, D. (2006). The short-term behavioral reactions of bottlenose dolphins to interactions with boats in doubtful sound, New Zealand. *Marine Mammal Science*, *22*(4), 802–818.
- Ma, G., Ahmed, N. K., & Willke, T. LYu. (2021). Deep graph similarity learning: A survey. *Data Mining and Knowledge Discovery*, *35*, 688–725.
- Makarov, I., Kiselev, D., Nikitinsky, N., & Subelj, L. (2021). Survey on graph embeddings and their applications to machine learning problems on graphs. *PeerJ Computer Science*, *7*, e357.
- Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2015). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, *104*(1), 11–33.
- Niepert, M., Ahmed, M. & Kutzkov, K. (2016). Learning convolutional neural networks for graphs. In *International conference on machine learning*. PMLR, pp. 2014–2023.

- Park, C., Han, J., & Yu, H. (2020). Deep multiplex graph infomax: Attentive multiplex network embedding using global information. *Knowledge-Based Systems*, 197, 105861.
- Perozzi, B., Al-Rfou, R. & Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 701–710.
- Rossi, A., Barbosa, D., Firmani, D., Matinata, A., & Merialdo, P. (2021). Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(2), 1–49.
- Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1), 61–80.
- Segev, E. (2021). *Semantic network analysis in social sciences*. Routledge.
- Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J. & Mei, Q. (2015). Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*. pp. 1067–1077.
- Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., Cau, P., Remy, E., & Baudot, A. (2019). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, 35(3), 497–505.
- Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al. (2017). Graph attention networks. *Stat*, 1050(20), 10–48550.
- Wang, H., Zhang, K., Liu, Q., Tran, T. & Yu, Y. (2008). Q2semantic: A lightweight keyword interface to semantic search. In *The Semantic Web: Research and Applications: 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008 Proceedings 5'*, Springer, pp. 584–598.
- Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 4–24.
- Xiao, S., Lin, H., Wang, C., Wang, S. & Rajapakse, J. C. (2023) Graph neural networks with multiple prior knowledge for multi-omics data analysis. *IEEE Journal of Biomedical and Health Informatics*.
- Yang, J., Yao, W., & Zhang, W. (2021). Keyword search on large graphs: A survey. *Data Science and Engineering*, 6(2), 142–162.
- Yang, L., Wang, Y., Gu, J., Wang, C., Cao, X. & Guo, Y. (2020). Jane: Jointly adversarial network embedding. In *IJCAI*. pp. 1381–1387.
- Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4), 452–473.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Janet Oluwasola Bolorunduro¹  · Zhaonian Zou¹  · Mohamed Jaward Bah²

✉ Janet Oluwasola Bolorunduro
janetbolin@hit.edu.cn

Zhaonian Zou
znozou@hit.edu.cn

Mohamed Jaward Bah
easybah@zhejianglab.com

¹ School of Computer Science and Technology, Harbin Institute of Technology, P. Box 75, Harbin, China

² Department of Big Data Intelligence Research Center, Zhejiang Lab, Hangzhou, China