# CONNECTING USER PROFILES OF SOCIAL NETWORKS USING PROXIMITY-BASED CLUSTERING

***Rashmi C[1*], Mallikarjun M Kodabagi[2]***

[1,2]Faculty of Computing and Information Technology, REVA University, 560064 Bengaluru, India

Email: rashmi.c@reva.edu.in[1*](corresponding author)

DOI: https://doi.org/10.22452/mjcs.sp2022no2.1

***ABSTRACT***

*The establishment of connections among social network users using their profile information is an important task in social network analysis, which facilitates the development of various technological solutions such as stock market analysis, crime detection, tracking system of fraudulent events, etc. In this work, a proximity-based clustering method for networking LinkedIn profiles is presented. The proposed system computes proximity values between users using various attributes of user profiles. The proximity measures are computed by analyzing unstructured data of user profiles to connect users. The method addresses various issues such as comparison of familiar sentences, finding patterns, and sub-patterns among user profiles, assigning weights on attributes similarity, and computing total similarity which is associated with unstructured data. After computing proximity measures on various attributes of user profiles, the connecting edges between nodes are determined by employing artificial intelligence and a network graph is formed. The method is evaluated on a LinkedIn data set to form a connected graph. The strength of the proposed methodology lies in the formation of multi-layered network graphs, as it uses various attributes of the user profiles to connect them. The proposed methodology helps various applications like recommendation systems to form network graphs of selected attributes and perform the social network analysis. The method achieves an accuracy of 96%. However, the profiles containing abbreviations of important information are not matched and the system accuracy drops down in such cases.*

***Keywords: User Profile, Social Network Analysis, Network Graph, Clustering, Unstructured Data, Local Efficiency, Global Efficiency***

## 1.0 INTRODUCTION

In recent years, social networks have become popular and are widely used by people for the communication of information [1]. Some of the most widely used social networks are Facebook, LinkedIn, Quora, Google+, and Twitter. People are connected to these online social networks based on their profile information and interest. Using these online social networks, People share millions of unstructured data [2], like images, videos, text, and so on. Such unstructured data is quite useful for the development of various applications. like recommendation systems in terms of job postings or job recommendations, product recommendations, stock predictions [3], sales predictions [4], crime analysis [5],[6], and many more. All these applications are social network-dependent and require user profile information. Hence, the formation of a network of social network users using their profiles becomes essential in social network analysis.

In recent literature, many techniques have been reported related to establishing social networks of people using their profile information. The reported techniques are user profile information, which consists of various attributes like age, gender, skills, location, education, and many more for forming a network of social network users. But, on most social media networks like Twitter, Facebook, and LinkedIn, users may not fill in the details of all attributes and are sometimes unavailable [7]. There are various issues associated with the connectivity of a user profile. First, is the Unavailability of complete data where users provide some basic information like name, and gender, but rarely give other detailed information. Second, dealing with unstructured data, where the user profile data can contain information in terms both qualitative and quantitative like dates, numbers, and facts. Third, the most important reason is privacy issues, most online social network sites limit access to some personal information [8]. Fourth, to explore and extort meaningful information from these infinite online social networking data, special tools are used to build the structure of the online social networks which are graph-based. Multiple tools are used, but selecting the

best tool for a specific task is difficult to decide [9]. Other challenges in dealing with profile information include converting unstructured data into meaningful structural data for further analysis.

These issues need to be addressed to create a network of social network users. Such connected networks are used in various platforms such as network creation [10], community detection [11],[12] structural clustering [13],[14] influential node detection [15], [16] link prediction[17], useful for the development of various technological solutions such as recommendation system [18],[19], recruitment system [20],[21] crime detection [22], and demography [23], etc. Therefore, establishing a connection between LinkedIn profiles of any online social network is a paramount task. A method to find the relation between LinkedIn social network profiles and connect them using a graph is introduced in this paper. To create a linked graph of LinkedIn users, the approach computes proximity measures on different LinkedIn profile attributes. In this method, Levenshtein distance [24] is applied to calculate the differences between sequences in which we get the similarity index using a threshold value for every feature that is compared to every user profile.

The organization of the paper is as follows. Section 2 reports the critical literature review on the creation of network structure. The methodology for network structure creation is described in section 3. The results and analysis are given in section 4. The conclusion of the work is given in Section 5.

## 2.0 LITERATURE REVIEW

All social user profiles can be linked and networked to multiple applications and fields. A description of some of the most notable works is given below.

An efficacious team formation using social connections of the network is presented in [25]. The proposed method developed an effective team to follow through with a particular job. The method considered a skill set with each level of proficiency for a specific skill. Based on this concept, an effective team was formed with a group of nodes that commonly performed a particular job with the required skillfulness and with low interaction costs. The main aim of the methodology is to propose an efficient team with a set of skills with different levels of expertise.

In the multi-layer network formation, a unique relationship represented between the nodes of different layers of the network is expressed in [26]. This method describes characteristics of certain attributes of the network to form a multiple-layer network. Each layer represents a participant by selecting its interaction connectivity edge in reaction to the interaction connectivity edge of the different participants using different functions.

The topics-based network formation using location-based social networks is implemented in [27]. To build a network structure model of social network link creation that has pairwise user similarities and individual characteristics. Finally, the method uses proximity measures such as geography, biography, short messages, and mobility. To construct proximity measures from text information where most information is unstructured, the model builds a topic model approach such as latent Dirichlet allocation [28].

A LinkedIn personalized expertise search, where the problem of personalized expertise using a LinkedIn search is addressed, especially for skills-containing search queries, is discussed in [29]. The method proposes heuristics to derive training data sets from the search logs by considering the members' skills when handling position and sample selection biases.

In [30], the interest graph for social networks is created by considering user-generated tags. The technique is to create an initial network based on user interest along with user tag representation and then cluster the network based on interest to produce a model of hierarchical interest using the enhanced Girvan Newman algorithm [31].

The effective formation of graph clusters that are attribute-based is discussed in [32]. The method explains node features with the topographical features, using the graph attributes such as nodes of a network and edges that connect the nodes in a network. The model illustrates cluster formation dynamically with familiar inner properties that make use of a self-learning algorithm. The initial configuration of the cluster can be formed at any arbitrary point by discovering the balanced result of graph clustering that is attribute-based.

A method to extract professional data from LinkedIn using LinkedIn API and normalizing it by eliminating redundancies is implemented in [33]. Additionally, data is normalized again based on the locations of LinkedIn

connections with the help of Geo-Coordinates. Then, the normalized data set is clustered using K-means, Hierarchical, and Greedy clustering algorithms based on company names, job titles, and geographic locations.

Link prediction of multilayer in online social networks using certain features of social networks is narrated in [34]. The proposed method had better accuracy by considering different node pair features that require community formation which makes use of the clustering algorithm such as InfoMap [35]on the auxiliary graph and also on the attributes of the selected network on which it is used to predict the link formation using multiple machine learning algorithms.

A community network was formed based on the user attributes such as location, high-density interactions between users, user lifespan, and user weight. The proposed method Recently largest Information [36] had better performance in identifying the network structure for time complexity and accuracy based on identified user attributes over the existing methods in a dynamic social network.

After a critical analysis of the literature, it is observed that some of the problems are addressed, such as the development of connected networks that are distance-based properties, present network characteristics, the users' opinion, the factor of location, and prediction of ties within social networks. Nevertheless, the issues such as network structure formation in real-world scenarios, various active models for creating a different layout of the network, many nodes under cogitation with many factors for building a network of many social networks, cost edges for creating network graph, predicting jobs based on details on user profiles information remain unanswered. The research study is done to overcome these questions in the proposed model.

## 3.0 SOCIAL NETWORK USERS PROFILING

To create a network of connected people in social network subscribers, the technique uses proximity measures determined over different profile attributes such as description, qualifications, position-title, position-summary, education-school-name, position-company name, education-degree, location, area, etc. The created network graph is represented by equations (1) and (2).

$$\boldsymbol{G = \{U, C \quad where U = u_1, u_2 \ldots . . u_n\}} \tag{1}$$

Here, $\boldsymbol{U}$ denotes, the set of nodes and

$$\boldsymbol{C = \{C_{ij},\ where\ C_{ij} = (u_i, u_j) \forall u_i, u_j \in U\}} \tag{2}$$

$\boldsymbol{C_{ij}}$ represent connections between the nodes.

The method works in three stages: the computation of the similarity/proximity value between nodes, the discovery of edges between nodes, and the creation of a graph of social networks. The outline schematic for the approach proposed is represented below.

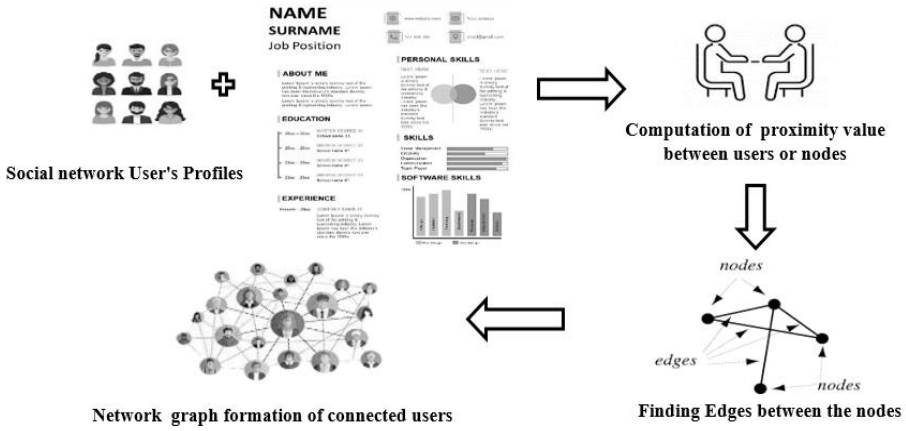


Fig. 1: Outline of the schematic of the approach proposed

The description of each of the processing steps of the methodology is detailed in the following subsections.

### 3.1 Proximity Value Computation

The proximity value between the nodes or user profiles computations carried out on various attributes of the user profiles as detailed in equation (3). The proximity value indicates the affinity of the edge between the nodes. More affinity will carry a higher proximity value on the edge and less affinity will have a lower proximity value. The proximity value lies in the range ***0 to M***. Where, ***M*** indicates the number of attributes of the user profile. It is computed as a sum of all similarity values of attributes. The computation for similarity values of attributes is given in the following subsections.

$$\boldsymbol{P_{roximity(i,j)} = \sum S_{ij}(A_k) W_k} \quad \textbf{(3)}$$

Where k = 1,2……..M

$P_{roximity(i,j)}$ Indicates the total similarity between the nodes on all M attributes.

$S_{ij}$ Indicates the similarity between the nodes on the Kth attribute.

The M attributes considered for similarity computation are as below

$A_1$ = Summary (it is a biography of the users)

$A_2$= skills

$A_3$= position-summary

$A_4$= position-title

$A_5$ = position-company-name

$A_6$ = location

$A_7$= education-school-name

$A_8$ = education-degree

$A_9$= education-field

$A_{10}$= industry

$W_k$= Indicates the weight of the attribute which is initialized to 1.

#### 3.1.1 Summary Similarity

The summary of the person/user contains the biography of the person in the unstructured text form. Comparing the unstructured text biography of one person with another is a challenging problem. Due to the unstructured text nature, sentences cannot be directly compared, as they may not appear in the same order. However, the idea of a set of word matches can be used to find the similarity between them. The match percentage is thresholded to find the similarity. The same has experimented with the research work. The similarity computation is as below.

Let $\boldsymbol{W}$ be the set of words representing the biography of the user. In this summary similarity, it is computed to determine whether the biography of one user is a subset of a biography of another user denoted by ($\boldsymbol{W_1 \in W_2}$).

The similarity value is represented by equation (4)

$$\boldsymbol{S_{ij}(A_1) = \begin{cases} 1, if(W_i \in W_j) \wedge (wordmatch\%) > T_1 \\ 0, otherwise \end{cases}} \quad \textbf{(4)}$$

Where $T_1$ a threshold is empirically chosen.

### 3.1.2 Skill Similarity

The skill set of every individual is an important attribute in the profile. The skillset will help to find the similarity between the profiles. To form a community of people having the same skill set. This attribute may carry more weight when compared to some of the other attributes of the user profiles. Various communities can be formed using the skill set. Therefore skill set is also used in our method to determine the similarity. The computation of similarity is shown in equation (5). The idea of a set of word matches is also explored to find the skill set similarity.

Let $\boldsymbol{X}$ be the set of words representing the skills of the user. In this skill-based similarity, it is computed to determine whether the skills of one user are a subset of another user's skills denoted by ($\boldsymbol{X_1 \in X_2}$).

$$\boldsymbol{S_{ij}(A_2)} = \begin{cases} \mathbf{1}, & \boldsymbol{if(X_i \in X_j) \wedge (wordmatch\%) > T_2} \\ \mathbf{0}, & \boldsymbol{otherwise} \end{cases} \quad \textbf{(5)}$$

Where the threshold $T_2$ is empirically chosen.

### 3.1.3 Position-Summary

The position summary is considered an important asset of the user profile. The position summary describes the brief information of a user profile. It specifies the information about the roles and responsibilities of current and previous jobs of a user. This attribute helps us to find the similarity of users based on user profile roles and responsibilities where we connect people working on similar roles. Therefore position summary adds strength to find out how close the user profile is which is computed using the equation (6).

Let $\boldsymbol{Y}$ be the set of words representing the overall roles and responsibilities of a user described as Position_summary. In this position's summary-based similarity, it is computed to determine whether one user is a subset of another user denoted by ($\boldsymbol{Y_1 \in Y_2}$).

$$\boldsymbol{S_{ij}(A_3)} = \begin{cases} \mathbf{1}, & \boldsymbol{if(Y_i \in Y_j) \wedge (wordmatch\%) > T_3} \\ \mathbf{0}, & \boldsymbol{otherwise} \end{cases} \quad \textbf{(6)}$$

Where the threshold $T_3$ is empirically chosen.

### 3.1.4 Position-Title

The Position title describes the specific information of the job title of the user's job profile. The Position title is also an important asset to user profile matching where we can list or identify users with similar job titles. Using this information we can connect people with similar job titles as a user. This attribute provides weightage when compared with other attributes of a user profile. Therefore the similarity is computed in equation (7).

Let $\boldsymbol{Z}$ be the set of words representing the position title of a user profile. In this position-title-based similarity, it is computed to determine whether one user is a subset of another user denoted by ($\boldsymbol{Z_1 \in Z_2}$).

$$\boldsymbol{S_{ij}(A_4)} = \begin{cases} \mathbf{1}, & \boldsymbol{if(Z_i \in Z_j) \wedge (wordmatch\%) > T_4} \\ \mathbf{0}, & \boldsymbol{otherwise} \end{cases} \quad \textbf{(7)}$$

Where the threshold $T_4$ is empirically chosen.

### 3.1.5 Position-Company-Name

The company name plays a major role to identify people with similar interests. Using this information, a network connection can be formed based on the similar company name of a user profile. This attribute provides weightage when computed with other attributes and also a separate community can be formed using company-name as a single attribute. Therefore the similarity is computed in equation (8).

Let $\boldsymbol{P}$ be the set of words representing the company name of a user. In this, we compute similarity to determine whether one user is a subset of another user using a company name as one of the features and is denoted by ($\boldsymbol{P_1 \in P_2}$).

$$S_{ij}(A_5) = \begin{cases} 1, & if(P_i \in P_j) \wedge (wordmatch\%) > T_5 \\ 0, & otherwise \end{cases} \quad (8)$$

Where the threshold $T_5$ is empirically chosen.

### 3.1.6 Location

The location of a user profile is used as one of the main features of finding similarities. A location determines the place of a user working which helps to segregate users mainly based on their location. A location of a user is determined from the user profile which will be in unstructured text format. By extracting this information, one can identify the similarity of one user profile with another user. The location attribute contributes more weightage when compared with other attributes of a user profile. A networking community can be formed based only on location, whereas other features can be considered to form a group within a community. Therefore, the similarity is computed in equation (9).

Let $Q$ be the set of words representing the location of a user. Using location-based similarity, it is computed to determine whether one user is a subset of another user denoted by $(Q_1 \in Q_2)$ in terms of location.

$$S_{ij}(A_6) = \begin{cases} 1, & if(Q_i \in Q_j) \wedge (wordmatch\%) > T_6 \\ 0, & otherwise \end{cases} \quad (9)$$

Where the threshold $T_6$ is empirically chosen.

### 3.1.7 Education-School-Name

The user who pursued his/her education can also play a major role to distinguish his academic performance from the other user. The university name signifies the validation of a user for overall academic performance. It can be used as a validation of a user where the school name plays a major role to identify the user. It has weightage when computed with the remaining attributes of a user profile. Therefore, similarity based on school name is denoted by equation (10).

Let $R$ be the set of words representing the school name of a user. In this school name-based similarity, it is computed to determine whether one user is a subset of another user denoted by $(R_1 \in R_2)$.

$$S_{ij}(A_7) = \begin{cases} 1, & if(R_i \in R_j) \wedge (wordmatch\%) > T_7 \\ 0, & otherwise \end{cases} \quad (10)$$

Where the threshold $T_7$ is empirically chosen.

### 3.1.8 Education-Degree

Education-degree feature signifies the degree obtained by the user for the corresponding academic discipline. It helps us to categorize the users based on their degrees even if they belong to the same academic discipline. This attribute is important when computed with the school name and field-of-study attribute to validate the user profile. A network connection can be formed using these attributes and strengthens the similarity computation, which is represented in equation (11).

Let $D$ be the set of words representing the degree obtained by the user. In this degree-based similarity, it is computed to determine whether one user is a subset of another denoted by $(D_1 \in D_2)$, in terms of Educations_degree.

$$S_{ij}(A_8) = \begin{cases} 1, & if(D_i \in D_j) \wedge (wordmatch\%) > T_8 \\ 0, & otherwise \end{cases} \quad (11)$$

Where the threshold $T_8$ is empirically chosen.

### 3.1.9 Education-Field-of-Study

In this feature, the similarity is identified based on the terms of academic discipline concerning higher education. This feature is one of the principal components of identifying similar users based on their academic discipline which

is referred to as Education-field-of-study. As described earlier, this attribute has importance when considered with the school name, degree, and industry of a person. Therefore, the similarity is computed in equation (12).

Let $\boldsymbol{F}$be the set of words representing the academic discipline for the field of study. Here, the similarity of an academic discipline is computed based on one user being a subset of another user denoted by $(\boldsymbol{F_1} \in \boldsymbol{F_2})$, in terms of field-of-study.

$$\boldsymbol{S_{ij}(A_9)} = \begin{cases} \mathbf{1}, & \boldsymbol{if}(\boldsymbol{F_i} \in \boldsymbol{F_j}) \wedge (\boldsymbol{wordmatch}\%) > \boldsymbol{T_9} \\ \mathbf{0}, & \boldsymbol{otherwise} \end{cases} \quad \textbf{(12)}$$

Where the threshold $T_9$ is empirically chosen.

### 3.1.10 Industry

The industry is an important asset to classify people based on the domain they work in. It is used to compare one user profile with the other based on their domain which is classified using industry as its feature. A community of users can be formed based only on a domain. This attribute highlights the similarity when computed with other attributes of a user profile. Therefore, the similarity is denoted in equation (13).

Let $\boldsymbol{I}$ be the set of words representing the domain of the user working. In this domain-based similarity, it is computed to determine whether the domain of one user is a subset of another user domain denoted by $(\boldsymbol{I_1} \in \boldsymbol{I_2})$ the domain knowledge.

$$\boldsymbol{S_{ij}(A_{10})} = \begin{cases} \mathbf{1}, & \boldsymbol{if}(\boldsymbol{I_i} \in \boldsymbol{I_j}) \wedge (\boldsymbol{wordmatch}\%) > \boldsymbol{T_{10}} \\ \mathbf{0}, & \boldsymbol{otherwise} \end{cases} \quad \textbf{(13)}$$

Where the threshold $T_{10}$ is empirically chosen.

It is assumed that all the attributes carry the same weight. So each attribute similarity is computed based on the chosen threshold value and all the attributes are considered to have the same weight.

---

***Algorithm***: ***Network graph of social media users***

---

**Input**: Social media user's profile data set.

**Output**: Network Graph of Social Media users.

**Initialization**: Initialize Matrix *M*(similarity matrix) to zero.
*Begin*
*Step 1:*For every $i^{t\square}$ profile of the data set.
*Step 2*: For every $j^{t\square}$ profile of the data set.
*Step 3*: Compute the Similarity value $S_{ij}$ between $i^{t\square}$ and $j^{t\square}$ profile on all attributes $\boldsymbol{A_k}$

$$\boldsymbol{S_{ij}(A_k)} = ((\boldsymbol{i^{th}} \in \boldsymbol{j^{th}attribute}) \wedge (\boldsymbol{wordmatch}\%)) > \boldsymbol{Thresholdvalue}$$

*Step 4*: $\boldsymbol{M_{ij}} = \sum \boldsymbol{S_{ij}(A_k)}$
*Step 5*: End of Step 2
*Step 6*: End of Step 1
*Step 7*: Construct Network Graph using $M_{ij}$ matrix.
*End*

The proposed methodology is evaluated on the LinkedIn data set where we have considered multiple attributes to identify the similarity of users and establish a social network structure, the results of the methodology are discussed in the following section.

## 4.0 RESULTS AND DISCUSSIONS

The proposed system is evaluated using the LinkedIn, dataset. LinkedIn is a largely appealing environment describing the skillful domain of its members. By exploring LinkedIn network data and information, one can analyze the etiquette and other assets of groups as well as individuals [37]. Data from LinkedIn and its API is unique and represented differently from data accessible on other social networks. The dataset was used in JSON style and retrieved using the API from LinkedIn to get the Secret Key, API key, OAuth User Secret Identifications, and Token

of the application [38]. Table 1 displays the different features obtained from the LinkedIn API in the dataset and the corresponding definition.

Table 1: Features Description

| Features | Information |
|---|---|
| Summary | Biography of a user |
| Skills | The skillset of an individual user |
| Position-summary | Overall position held by an individual user |
| Position-title | Specifying a role in a specific business |
| Position-company-name | Specifying the name of the company |
| Location | Location details of the work area |
| Education-school-name | Specifying the school name of an individual user |
| Education-degree | Specifying the Degree of an individual user |
| Education-field-of-study | Overall Academic discipline of an individual user |
| Industry | Industry detail of the Individual user |

The dataset contains several files with 200 user profiles for each file, which are called nodes. 10 user profile attributes are considered in this approach, as shown in Table 1. The N-gram similarity [39], Levenshtein Distance [40], Jaccard distance [41], etc., some general proximity measures are much useful for comparing profile characteristics such as company name, location, and role title names. A loop iterates on the entire dataset in the proposed technique and clusters them together according to the stated threshold value specified using the metric of similarity, and data is saved as a matrix. When the data for every user profile is contrasted with all other user profiles., the similarity of the user profile characteristics using the Levenshtein distance is determined. For all the features, the distance measure is used to quantify the discrepancies between sequences in which the proximity value is obtained based on the threshold value, compared to one for all user profiles. Connectivity is defined using proximity values and the network model is constructed using the NetworkX [42]open-source framework, a graph-based platform for mining, that easily models the construction of social networks, as shown in figure 2.

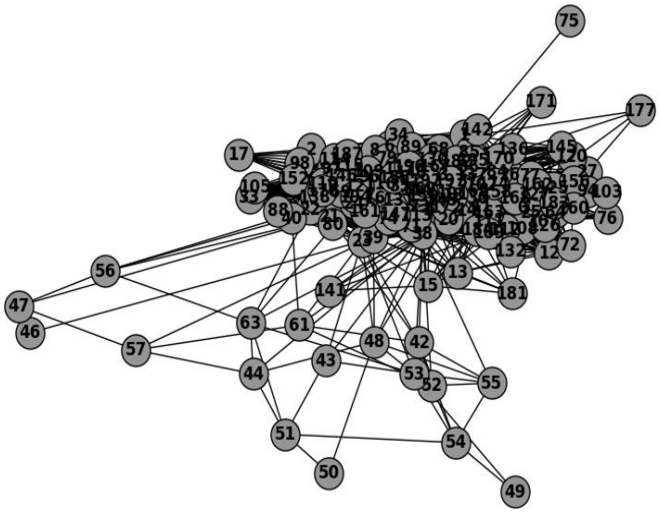


Fig. 2: Network graph using a similarity measure

The first node is compared with the remaining nodes in the second, the third, the fourth, etc. A comparison of a node to every other node is made. In total, 198 nodes are considered in the dataset. As shown in Table 1, a comparison of each node with characteristics such as similarity, abilities, location, and many more. Later, using the Levenshtein interval, the measure of similarity is defined with a value threshold of 40. The graphs below represent the affinity between vertices/nodes that uses an index of similarity, as shown in the following figures.

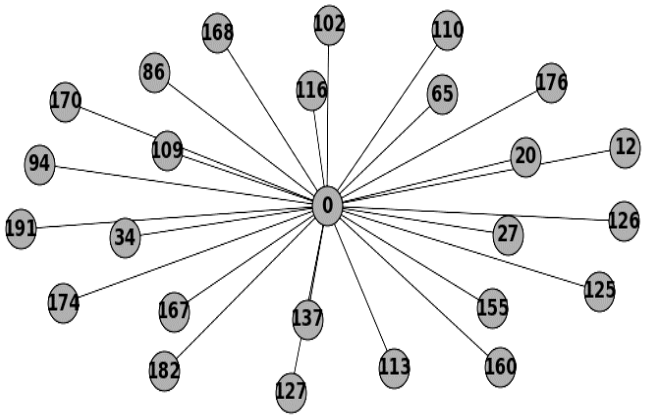


Fig.3: The first node's connectivity using the similarity measure

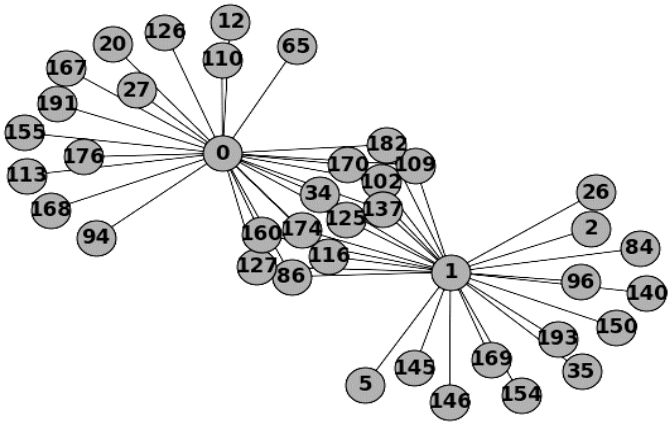


Fig. 4: Connectivity of nodes 0 and 1 with all other nodes using the similarity measure

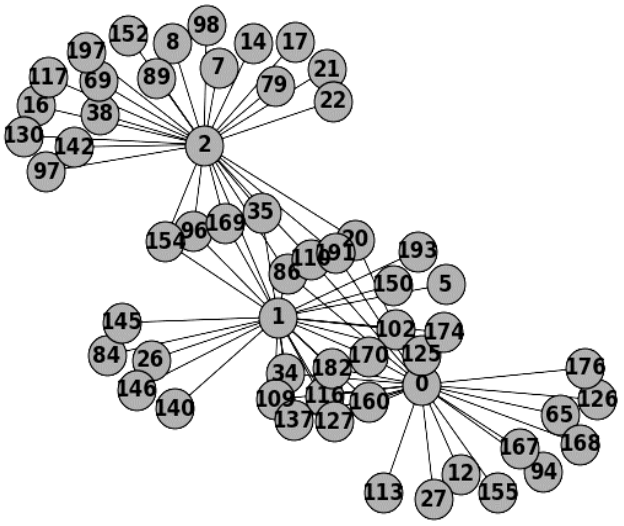


Fig. 5: Connectivity of nodes 0, 1, and 2 with all other nodes using the similarity measure

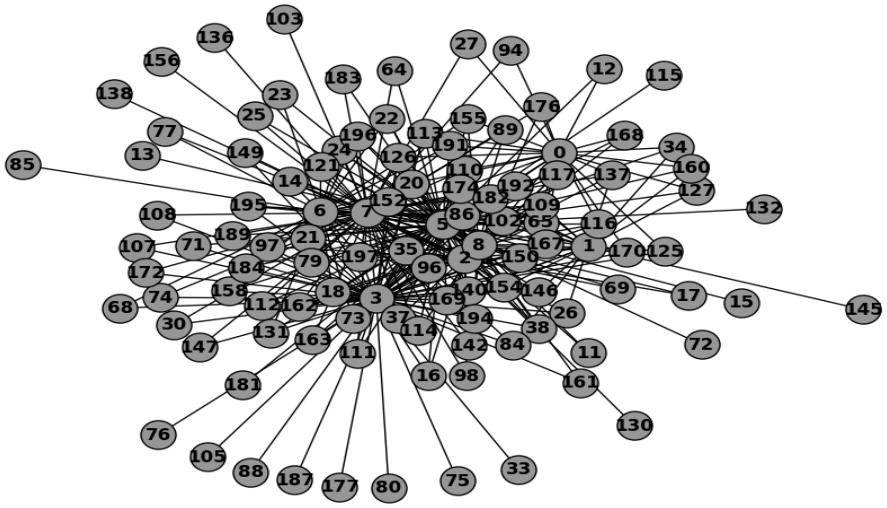


Fig. 6: Connectivity of first ten nodes with all other nodes using the similarity measure

In the above figure such as Fig 3, the first node 0 is compared with all other remaining nodes by considering all the characteristics of the node as mentioned in Table 1, if there is an affinity between the nodes then the connectivity is established and a network graph is constructed using Levenshtein distance based on the threshold value. The same procedure is repeated for node 0 and node 1, for node 0, node 1 and node 2, and finally for the first ten nodes as shown in Fig 4, Fig 5, and Fig 6 respectively. The proximity between the nodes varies based on the threshold value used in the Levenshtein interval.

The connectivity of a social network can be gauged using multiple measures. The connectivity of a network can be measured through network efficiency. This method was first introduced to measure the efficiency of vertices in a graph [43]. The established social network as shown in Fig. 2can be measured using network efficiency methods such as Global efficiency and Local efficiency. If the value is closer to 1, the established connected network is deemed more efficient, and if it is closer to 0, it is considered less efficient.

Let Gbe a graph with a vertex $V(G)$and the connection between the two vertices is represented by $C(G)$. Let $d(p, q)$ represent the distance that is the number of edges in the shortest path from 'p' to 'q'. If there is no connection between 'p' and 'q', then $d(p, q) = \infty$. The efficiency between two vertices 'p' and 'q' can be defined as

$$E_{(p,q)}(G) = \frac{1}{d(p,q)} \; \forall \; (p \neq q) \qquad [14]$$

The efficiency of a pair of nodes in a graph is the multiplicative inverse of the shortest path distance between the nodes [44]. The Global Efficiency $G_E$ of a graph is the average efficiency overall $p \neq q$. Therefore $G_E$ is represented as

$$G_E(G) = \frac{1}{n(n-1)} \sum_{p \neq q} \frac{1}{d(p,q)} \qquad [15]$$

Where 'n' represents the number of nodes in a network. In theoretical networks, because of the nature of networks with diameters greater than 2, average global efficiency can range from 0 to 0.5. Because of the larger interactions compared to the number of nodes, the global efficiency value drops, and the highly-dense network can also decrease the global efficiency value [45], [45]. In our established social network graph we obtain the average $G_E$ = 0.575865, since the graph is a densed network and also it has higher interactions compared to the number of nodes.

The average global efficiency of the subgraph caused by the node's neighbors is the local efficiency $L_E$ of a node in a graph. Therefore, the average local efficiency is the average of local efficiencies of each node is represented as

$$L_E(G) = \frac{1}{n} \sum_{s \in G} G_E(G_s) \qquad [16]$$

Where,$G_s$ is the subgraph induced by the neighbors of 's'. In our established social network we obtain the average $L_E$= 0.966608. Our established social network is considered to be the most efficient network since the average efficiency value is near 1.

The methodology is compared with existing methods of profiling network users/network structure, the comparative analysis is given in table 2.

Table 2: Comparative Analysis Study

| Authors | Methodology | Method Description | Advantages | Limitations |
|---|---|---|---|---|
| [25] | Effective Team Using Social Network | Team/Network formation using skills as a feature, based on the level of expertise. The size of the network increases as skills increase. | A set of nodes with skillful expertise guarantees low communication costs. Smaller the team/network size better communication between the nodes/team members. | It focused only on unweighted edges of the graph and it was not applied to real-time datasets. |
| [26] | Strategic network formation of Multi-Layer Networks | Distance-based utility network formation between specific pair of nodes, where the first layer is formed between the nodes that are neighbors in a different layer by optimizing it. | Establishes a path only to a specific set of nodes using distance-based network formation, whereas traditional methods only focused on the minimum distance between all the setoff nodes in a network. | It focused only on distance-based utilities rather than additional classes of effective functions in the multiple-layer network formation. |
| [27], [28] | Location-based strategic network formation. | Structural network formation using location-based. The network is formed by considering the characteristics of nodes and comparing the similarity pair-wise between the nodes using the latent Dirichlet allocation model. | the method uses four user proximity measures such as geography, biography, short messages, and mobility. | The link formation depends only on pairwise user similarity measures and each-individual user characteristic. |
| [29] | Personalized expertise search on LinkedIn | The collaborative filtering approach for personalized expertise search is based on matrix decomposition. | Estimates scores for both the mentioned and not mentioned skills of the user-profiles list. | The weights of the attributes are manually adjusted to obtain the personalized search of a LinkedIn user. |
| [30], [31] | Social Networks graph formation is based on interest in user-generated Tags. | Interest network formation using user tags, later applying the hierarchical clustering method to build a matching graph based on interest. | clustering the network structure-based interest in creating a model of hierarchical interest through the implementation of an enhanced Girvan Newman algorithm. | It can be improvised in modeling efficiency and also applied in real-time applications. |
| [32] | Dynamic cluster game formation for clustered attributes of a graph. | Dynamic cluster creation with properties that develop with great individual variability in real-world networked systems, as well as using a self-learning algorithm (SLA). | Both the topological structure and attribute information are merged to obtain a dynamic cluster formation. SLA is comparatively a bit faster to dynamically identify clusters as it considers two different sources of input. | Concerning scalability, SLA is a bit lesser than traditional community detection algorithms. |
| [33] | Extraction of user-profile data from | The extraction of data is centered on the positions of | To study and analyze the LinkedIn network properties, | It can be upgraded to a predictive system that |

| Authors | Methodology | Method Description | Advantages | Limitations |
|---|---|---|---|---|
| | LinkedIn. | connections to LinkedIn and clustering of the data based on user profile by normalizing it and applying hierarchical, Greedy, and finally K-Means clustering algorithms. | and group with the necessary security and confidentiality of user data on these networks. | recommends user job profiles. |
| [34], [35] | Prediction of multilayer connections in online social networks using topological features of the network. | to anticipate link formation using machine learning algorithms such as KNN, Naive Bayes, and many more. | It provides an accurate estimation of links in a social network by considering the network-based similarity between friends considering both geographical-based and content-based similarities. | to explore the estimation accuracy when using several node pair features |
| [36] | Community detection based on user attributes | Formation of communities based on the attributes of the user's high-density interaction, and location using the Recently Largest Algorithm method. | It provides better accuracy and time complexity in identifying communities over existing methods and can be used effectively in recommender systems. | It focused only on very few attributes of a user profile where more weight was given to high-density interaction, compared to other user attributes. |

The methods described above focused only on network formation with very few sets of nodes for unweighted edges of the graph, and analysis of network structure using only a few sets of attributes. Finally, it focused only on synthetic data sets rather than real-time data sets. In the proposed methodology, the proximity measures on common user-profile attributes are computed by considering various sets of attributes and applied to real-time data sets from the LinkedIn application. Later, connectivity is established between the attributes employing artificial intelligence techniques and a network graph is formed.

## 5.0 CONCLUSION AND FUTURE WORK

In this work, a proximity value-based clustering method for connecting LinkedIn subscribers/users are presented. The similarity is used in the methodology where values are computed to create a graph of social network users on different attributes of user profiles. The proximity measures are computed by analyzing unstructured data of user-profiles addressing various problems associated with unstructured information. After computing proximity measures using attribute values of user profiles, the connecting edges between nodes are determined by employing artificial intelligence, and a network graph is formed. The method is evaluated on LinkedIn data and has given promising results. As the method uses various user profile attributes for building network graphs, several network graphs can be obtained on selected attributes. The proposed methodology helps various applications to form network graphs of selected attributes and perform the social network analysis. As a future research work, we can further build personalized recommendations based on user profile attributes, we can also apply natural language processing (NLP) techniques to identify the similarity between users.

## REFERENCES

[1] A. Mislove, M. Marcon, K. P. Gummadi, P. Drushel and B. Bhattacharjee et al., "Measurement and analysis of online social networks", in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement - IMC '07*, 2007, pp. 29-42, doi: 10.1145/1298306.1298311.

[2] H. Xue and D. Zhang, "A Recommendation Model Based on Content and Social Network", in *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, May 2019, pp. 477–481, doi: 10.1109/ITAIC.2019.8785729.

[3] S. Tiwari, A. Bharadwaj and S. Gupta, "Stock price prediction using data analytics", in *2017 International Conference on Advances in Computing, Communication and Control (ICAC3)*, Dec. 2017, pp. 1–5, doi: 10.1109/ICAC3.2017.8318783.

[4] A. Sheshasaayee and H. Jayamangala, "A study on the new approaches for social network based recommendations in digital marketing", in *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Feb. 2017, pp. 627–632, doi: 10.1109/ICIMIA.2017.7975537.

[5] K. Biron, W. Mansoor, S. Miniaoui, S. Atalla, H. Mukhtar and K. F. B. Hashim, "Data Science Tools for Crime Investigation, Archival, and Analysis", in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2019, pp. 1263–1266, doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00235.

[6] A. S. Srinath, H. Johnson, G. G. Dagher and M. Long, "BullyNet: Unmasking Cyberbullies on Social Networks", *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 2, Apr. 2021, pp. 332–344, doi: 10.1109/TCSS.2021.3049232.

[7] D. K. Srivastava, B. Roychoudhury and H. V. Samalia, "Importance of User's Profile Attributes in Identity Matching Across Multiple Online Social Networking Sites", in *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Jan. 2018, pp. 14–15, doi:10.1109/CONFLUENCE.2018.8442455.

[8] O. Hasan, B. Habegger, L. Brunie, N. Bennani and E. Damiani, "A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case", in *2013 IEEE International Congress on Big Data*, Jun. 2013, pp. 25–30, doi: 10.1109/BigData.Congress.2013.13.

[9] N. Akhtar, "Social Network Analysis Tools", in *2014 Fourth International Conference on Communication Systems and Network Technologies*, Apr. 2014, pp. 388–392, doi: 10.1109/CSNT.2014.83.

[10] A. Fabrikant, A. Luthra, E. Maneva, C. H. Papadimitriou and S. Shenker, "On a network creation game," in *Proceedings of the twenty-second annual symposium on Principles of distributed computing - PODC '03*, 2003, pp. 347–351, doi: 10.1145/872035.872088.

[11] S. Fortunato, "Community detection in graphs", *Phys. Rep.*, vol. 486, no. 3–5, Feb. 2010, pp. 75–174, doi: 10.1016/j.physrep.2009.11.002.

[12] C. Rashmi and M. M. Kodabagi, "A review on overlapping community detection methodologies", in *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, 2017, pp. 1296-1300. doi: 10.1109/SmartTechCon.2017.8358576.

[13] X. Xu, N. Yuruk, Z. Fheng and T. A. J. Schweiger, "SCAN: a structural clustering algorithm for networks", in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*, 2007, pp. 824-833, doi: 10.1145/1281192.1281280.

[14] Y. X. Qiu, R. H. Li, J. Li, S. Qiao, G. Wang, J. X. Yu and R. Mao, "Efficient Structural Clustering on Probabilistic Graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 10, Oct. 2019, pp. 1954–1968, doi: 10.1109/TKDE.2018.2872553.

[15] Y. Wang, G. Cong, G. Song and K. Xie, "Community-based greedy algorithm for mining top-K influential nodes in mobile social networks", in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*, 2010, pp. 1039-1048, doi: 10.1145/1835804.1835935.

[16] A. Ali, V. R. Hulipalled and S. S. Patel, "Centrality Measure Analysis on Protein Interaction Networks", in *2020 IEEE International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET)*, 2020, pp. 1-5, doi: 10.1109/TEMSMET51618.2020.9557447.

[17] R. Michalski, P. Kazienko and D. Krol, "Predicting Social Network Measures Using Machine Learning Approach", in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Aug. 2012, pp. 1056–1059, doi: 10.1109/ASONAM.2012.183.

[18] G. Suryateja and S. Palani, "Survey on efficient community detection in social networks", in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, Dec. 2017, pp. 93–97, doi:10.1109/ISS1.2017.8389304.

[19] S. Dhelim, H. Ning, N. Aung, R. Huang and J. Ma, "Personality-Aware Product Recommendation System Based on User Interests Mining and Metapath Discovery", *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 1, Feb. 2021, pp. 86–98, doi: 10.1109/TCSS.2020.3037040.

[20] Z. Wang, J. Zhao, J. Hu, T. Zhu, Q. Wang, J. Ren and C. Li, "Towards Personalized Task-Oriented Worker Recruitment in Mobile Crowdsensing", *IEEE Trans. Mob. Comput.*, vol. 20, no. 5, May 2021, pp. 2080–2093, doi: 10.1109/TMC.2020.2973990.

[21] S. Jere, L. Jayannavar, A. Ali and C. Kulkarni, "Recruitment graph model for hiring unique competencies using social media mining", in *Proceedings of the 9th International Conference on Machine Learning and Computing*, 2017, pp. 461–466, doi: 10.1145/3055635.3056575.

[22] U. M. Butt, S. Letchmunan, F. H, Hassan, M. Ali, A. Baqir and H. H. R. Sherazi, "Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review," *IEEE Access*, vol. 8, 2020, pp. 166553–166574, doi: 10.1109/ACCESS.2020.3022808.

[23] K. Keerthika and T. Saravanan, "Enhanced Product Recommendations based on Seasonality and Demography in Ecommerce", in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Dec. 2020, pp. 721–723, doi: 10.1109/ICACCCN51052.2020.9362760.

[24] S. Zhang, Y. Hu and G. Bian, "Research on string similarity algorithm based on Levenshtein Distance", in *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Mar. 2017, pp. 2247–2251, doi: 10.1109/IAEAC.2017.8054419.

[25] K. Kamel, N. Tubaiz, O. AlKoky and Z. AlAghbari, "Toward forming an effective team using social network", in *2011 International Conference on Innovations in Information Technology*, Apr. 2011, pp. 308–312, doi: 10.1109/INNOVATIONS.2011.5893839

[26] E. M. Shahrivar and S. Sundaram, "The Strategic Formation of Multi-Layer Networks", *IEEE Trans. Netw. Sci. Eng.*, vol. 2, no. 4 , Oct. 2015, pp. 164–178, doi: 10.1109/TNSE.2015.2500162.

[27] G. M. Lee, L. Qiu and A. B. Whinston, "Strategic Network Formation in a Location-Based Social Network: A Topic Modeling Approach", in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, Jan. 2016, pp. 5249–5258, doi: 10.1109/HICSS.2016.649.

[28] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation", *J. Mach. Learn. Res.*, vol. 3, 2003, pp. 993–1022.

[29] V. Ha-Thuc, G. Venkataram, M. Rodriguez and S. Sinha, "Personalized expertise search at LinkedIn", in *2015 IEEE International Conference on Big Data (Big Data)*, Oct. 2015, pp. 1238–1247.

[30] P. Xu and K. Xia, "Modeling interest graph of social networks with user-generated tags", in *2013 IEEE 4th International Conference on Software Engineering and Service Science*, May 2013, pp. 680–683, doi: 10.1109/ICSESS.2013.6615398.

[31] L. Despalatovic. T. Vojkovic and D. Vukicevic, "Community structure in networks: Girvan-Newman algorithm improvement", in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2014, pp. 997–1002, doi: 10.1109/MIPRO.2014.6859714.

[32] Z. Bu, H. Li, J. Cao, Z. Wang and G. Gao, "Dynamic Cluster Formation Game for Attributed Graph Clustering", *IEEE Trans. Cybern.*, vol. 49, no. 1, Jan. 2019, pp. 328–34, doi: 10.1109/TCYB.2017.2772880.

[33] P. Garg, R. Rani and S. Miglani, "Mining Professional's Data from LinkedIn", in *2015 Fifth International Conference on Advances in Computing and Communications (ICACC)*, Sep. 2015, pp. 98–101, doi: 10.1109/ICACC.2015.35.

[34] H. Mandal, M. Mirchev, S. Gramatikov and I. Mishkovski, “Multilayer Link Prediction in Online Social Networks”, in *2018 26th Telecommunications Forum (TELFOR)*, Nov. 2018, pp. 1–4, doi: 10.1109/TELFOR.2018.8612122.

[35] P. Held, B. Krause and R. Kruse, “Dynamic Clustering in Social Networks Using Louvain and Infomap Method”, in *2016 Third European Network Intelligence Conference (ENIC)*, Sep. 2016, pp. 61–68, doi: 10.1109/ENIC.2016.017.

[36] A. Mahmoudi, A. A. Bakar, M. Sookhak and M. R. Yaakub, et al., “A Temporal User Attribute-Based Algorithm to Detect Communities in Online Social Networks”, *IEEE Access*, vol. 8, 2020, pp. 154363–154381, doi: 10.1109/ACCESS.2020.3018941.

[37] M. A. Russell and M. Klassen, *Mining the Social Web*, 3rd edition. O’Reilly Media, Inc., 2019.

[38] LinkedIn Developer Products, *https://developer.linkedin.com/.*

[39] R. Kumar and R. P. Mathur, “Short text clustering using numerical data based on n-gram”, in 2014 5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence), Sep. 2014, pp. 274–276, doi: 10.1109/CONFLUENCE.2014.6949257.

[40] C. Rashmi and M. M. Kodabagi, "Profiling of Social Network Users using Proximity Measures", *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, 2020, pp. 24-28, doi: 10.1109/ICSTCEE49637.2020.9277129.

[41] T. Kongsin and S. Klongboonjit, “Machine Component Clustering with Mixing Technique of DSM, Jaccard Distance Coefficient and k-Means Algorithm”, in *2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA)*, Apr. 2020, pp. 251–255, doi: 10.1109/ICIEA49774.2020.9101912.

[42] L. Broasca, V. -M, Ancusa and H. Ciocarlie, “A Qualitative Analysis on Force Directed Network Visualization Tools in the Context of Large Complex Networks”, in *2019 23rd International Conference on System Theory, Control and Computing (ICSTCC)*, Oct. 2019, pp. 656–661, doi: 10.1109/ICSTCC.2019.8885641.

[43] V. Latora and M. Marchiori, “Efficient Behavior of Small-World Networks”, *Phys. Rev. Lett.*, vol. 87, no. 19, Oct. 2001, p. 198701, doi: 10.1103/PhysRevLett.87.198701.

[44] B. Ek, C. VerSchneider and D. A. Narayan, “Global efficiency of graphs”, *AKCE Int. J. Graphs Comb.*, vol. 12, no. 1, Jul. 2015, pp. 1–13, doi: 10.1016/j.akcej.2015.06.001.

[45] M. Mohammadi, M. Sarmad and N. R. Arghami, "An Extension of the Outlier Map for Visualizing the Classification Results of the Multi-Class Support Vector Machine." *Malaysian Journal of Computer Science*, vol. 34, no. 3, pp. 308-323, 2021, doi: 10.22452/mjcs.vol34no3.5.