Enhanced Clustering Technique for Efficient Identification of Independent Groups in Social Networks

R. Srinivasan[1], V. Ramachandran[2] and Nagaraju Baydeti[1,*]

[1]*Department of Computer Science and Engineering, National Institute of Technology Nagaland, Chumukedima, Dimapur – 797 103, Nagaland, India*
[2]*Department of Computer Science and Engineering, DMI College of Engineering, Affiliated by Anna University, Chennai – 600 123, Tamilnadu, India*
E-mail: rajusrini@hotmail.com; rama5864@gmail.com; baydetinagaraju@nitnagaland.ac.in
[*]*Corresponding Author*

Received 21 September 2021; Accepted 18 September 2022;
Publication 24 November 2022

Abstract

The main aim of this paper is to develop a new approach for identifying independent groups among users communicating in social networks using social media applications at any instant. Grouping of users as independent clusters is of dynamic nature as communication between known and unknown users can happen randomly at any point of time. It is becoming inherent to identify the groups, where the members of the group have strong relationship who communicate frequently and consistently via social media applications. Louvain's algorithm will identify the clusters in the community detection process but keeps the lightweight nodes in the original groups without making them into one group by considering the dependence relations. The concept of Bernstein conditions is enhanced and applied to identify the dependency among the users of social networks by formulating equivalence relations, which adhere to the properties of Reflexivity, Symmetricity and Transitivity.

Journal of Web Engineering, Vol. 21_7, 2011–2032.
doi: 10.13052/jwe1540-9589.2171

Then, the equivalence classes are identified which denote the individual groups of clusters where the users of one cluster are loosely coupled with the users of any other cluster but tightly coupled among the users of the same group. The strength of relationship among the users within the same and different clusters is identified with respect to the quantum of messages being propagated among the users using Louvain's algorithm and the results of equivalence class approach are compared using the same set of communication sequences to show the relation dependency among the members in various clusters.

Keywords: Equivalence relation, equivalence class, social networks, social media and clustering.

1 Introduction

Social network penetration is increasing worldwide exponentially with 71 percent of the Internet population are the users of the social media applications [1]. At present, communication via social media applications plays a vital role with respect to the higher user engagement rates and expanding mobile possibilities. Social networks enable the users not only to communicate beyond local or social boundaries but also to provide various options for sharing user generated contents like pictures and videos. The increased worldwide usage of smartphones and other mobile devices has opened up possibilities of mobile social networks with features such as location-based services. The rise of smart devices and mobile Internet has also lead to the increase in the mobile content consumption [2]. The global mobile data traffic has already crossed 11.2 exabytes per month as per the estimation at the end of 2017 due to the increasing spread in the usage of mobile applications which are exclusively providing social networking services [3]. The statistics also shows that the global average social network penetration rate is 42 percent [4] where the social networking usage has also become increasingly mobile. The statistics further provides information on the most popular networks worldwide as of October 2018, ranked by the number of active accounts. Facebook is leading the list with 2.23 billion monthly active users followed by YouTube having 1.9 billion, WhatsApp with 1.5 billion, Facebook Messenger with 1.3 billion, WeChat and Instagram reaching 1 billion monthly active users [5].

The popularity and prominence gained by the digitized mode of managing the social relations over the last two decades have changed the dimensions of communication scenario which are being entirely dominated by online social networks and media applications [6, 7]. The revolutionary increase in the users of social media [8] brought drastic changes in terms of the amount of data exchange across online social network applications. The data exchange via social networking services are not limited only to spreading news, sharing of viral videos, entertainment, political satirical videos, uploading pictures and personal updates. At present, people are getting connected more through online social networking services when compared to traditional Web portals or newsgroup services.

For any social network service provider, it is a challenging task to identify the relationship among their users. The users of social media applications are highly interconnected with strong and light relationship with each other. A pair of users are said to have a strong relationship when the interactions among them are significantly more and the relationship can be termed as light or weak when the interactions are minimal among the users [17]. Aggregation of users of social media form a structure called as cluster. The users may communicate in social networks via social media applications through intra and inter mode of sharing messages between clusters without any restriction. It is essential for the social media service providers to identify the independent groups of communication where the sender and receiver relationships from different groups are absolutely nil or light-weight communication between each group. This initiative of identification of independent groups will improve the security measures and eliminate the transfer of irrelevant messages between the groups to some extent.

2 Related Work

Many research works are being carried out in the area of community detection in the social networks to identify the users having similar interests and to find the strength of relationships among the users of social media based on the amount of messages shared between them and the period of consistent communication among them. Grouping of such users into independent clusters will be helpful to portray the role played by the members in each group so as to minimize the amount of inappropriate and irrelevant information being transferred. The identification of independent groups will aid in the prediction of actual originator of any information viz., false memes, spurious news etc., thus lead to improve the security in social media applications.

A social network can be considered as a *'graph'* where the vertices denotes the users of any social media application and the *'edges'* represent the communication between any two users. In online social networks, a community can be treated as a group of users who are frequently interacting with each other. It is obvious that the interaction among one group of users might be very close when compared to other set of users. So in this case, it is also possible to represent the online social network users as a '*directed weighted graph*' where the interactions among the users can be labelled as weight of the edges. A weighted edge can be one of the measures which indicate the strength of the relationship among the users. An edge with more weight represents the strong relationship among the users to which the edge is pointing to and an edge with relatively less weight denotes the least connectedness among the users. It is obvious that the strength of the relationship can be one-sided when weight w_{ij} between ($\mathrm{User_i}$, $\mathrm{User_j}$) is less than that of the weight w_{ji} between ($\mathrm{User_j}$, $\mathrm{User_i}$). As part of the social network analysis, users can be grouped into dense clusters known as communities. Louvain's method [11, 12] is the popular community detection algorithm that is based on modularity. Community detection is a process of segregating the users of a network in to various partitions where the intra-cluster communication strength of the users is stronger than the inter-cluster communication strength. Communities are also be grouped based on the common interests and similarities shown within the same group of users.

Ahmed Alsayat et al. [13] presented a framework for the novel task of detecting communities by clustering messages from large streams of social data. This framework is based on integration of K-means clustering algorithm, Genetic algorithm and Optimized Cluster Distance method for clustering data. The primary goal of this framework is to overcome the limitation of choosing the best initial centroids using Genetic algorithm and maximizing the distance between the clusters by pairwise clustering using Optimized Cluster Distance. The analysis has shown better clustering results and provided a novel use-case of grouping user communities based on their activities. Xu Yang et al. [14] has proposed a classification model to find the proximity amongst the users in terms of registration frequencies to specific cellular towers associated to their working places. The primary aim is to find the correlation between users working in the near proximity. The clustering results successfully reflect the higher proximity at work for the intra-class subjects. Francis T. O'Donovan et al. [15] analyzed the anonymized, scraped data from consenting Facebook users, together with associated demographic and psychological profiles. They have presented five clusters of users like Multimedia-savvy & Engaged, Low Engagement, Private broadcasters, High Engagement, Multimedia specialist users with common observed online behaviors, where the users also show correlated profile characteristics.

Kuldeep Singh et al. [16] have investigated about clustering of users in a social network based on the textual similarity among their tweets extracted from Twitter. Two algorithms are used, the simple K-means which is based on compactness that gives near to accurate results for general numerical datasets and the spectral K-means which is based on connectivity approach that finds textual similarity and the strength matrix which plays an important role in identifying the similarity between people. The results have shown that spectral clustering produces quick results in case of sparse and higher element datasets with high computational cost for larger datasets.

Keerthana N et al. [18] have proposed a novel method for multi-dimensional cluster to indentify the malicious users on online social networks. A new method over the design of the multi-Dimensional Clustering (m-DC) method for Facebook friends, which characterize the weight esteems to utilize the assessments of inactive and dynamic user attacks. Michal Turcanik [19] has analysed the possibility of cluster users of a selected network on the basis of their browsing behaviour. K-means clustering algorithm is used for identification of user groups on the basis of their behaviour on the internet.

Kun He et al. [20] introduced a novel graph-theoretical concept of hidden community for analysing complex networks that contain both stronger and weak communities. The main idea was to detect the hidden communities in complex social networks. Ling Wu et al. [21] have applied Deep Learning techniques for community detection in social networks. Andreas Kanavos et al. [22] addressed the need for an efficient and innovative methodology for community detection that will also leverage users' behavior on emotional level.

The community detection algorithms will not identify the independent clusters, rather they create clusters of users with stronger links among intra-clusters and weaker links among inter-clusters. It is proposed in this paper, to identify the dependencies among communication sequences and to map these relations with the properties of Reflexivity, Symmetricity and Transitivity to formulate the equivalence relations. From these equivalence relations, equivalence classes are extracted that represent the independent clusters of communication sequences.

3 Dependence Analysis of Communication Sequences – Proposed Model

Dependence analysis between communication sequences is the key factor in the identification of independent clusters. The first phase of this paper aims at detecting the dependency among the communication sequences implicitly and in the second phase, a novel approach called equivalence class approach is proposed to formulate independent groups. The dependence analysis of communication sequences is carried out using Bernstein conditions [9, 10]. The Bernstein conditions are enhanced to identify the same set of receivers among any two communication sequences. A relation is formed between communication sequences if there is any dependency among them. The related sequences have to be grouped in the same cluster. The formulated relations satisfy the conditions of the equivalence relations, i.e., reflexivity, symmetricity and transitivity. Each equivalence relation produces a partition in the set of communication sequences. It is proposed to determine the equivalence classes that are disjoint partitions in the set of communication sequences. The number of equivalence classes will represent the number of clusters for the given set of communication sequences. A systematic algorithm has been developed to implement this model and explained.

The proposed model is based on communication scenario where users of a social network are involved in propagation of information among various other users. The information which is propagated in the network can be native or forwarded one i.e., user generated or contents that are widely under public circulation. The information can be in the form of text, images and videos. The proposed model is implemented based on Client-Server model in which the communication being occurred is controlled or rather monitored through a server application. Server application is responsible for the successful communication of the contents that are being transferred across the users of the social network. Each communication sequence is registered in the server application as and when the communication occurs i.e., sharing of messages among the users. These communication sequences are the basis for the evaluation of the proposed model, which serve as an input to obtain the equivalence classes towards the identification of independent clusters. The algorithm for finding the equivalence classes to identify the independent clusters of communication sequences is described below:

Step 1:
The communication sequences are represented as a set $CS = \{cs_1, cs_2, \ldots, cs_n\}$, in which cs_1 shall be "$a \to b \to c$" that represents "user a

is originating a message to user *b*, *b* is forwarding it to user *c*". The set of all the receivers and originating senders are represented as $I_1 = \{\{R_1\}, \{R_2\}, \ldots, \{R_N\}\}$ and $I_2 = \{\{S_1\}, \{S_2\}, \ldots, \{S_N\}\}$ respectively. From any two communication sequences cs_1 and cs_2, viz., "$a \rightarrow b \rightarrow c$" and "$x \rightarrow y \rightarrow z$" the set of senders S_1 and S_2 and the set of receivers R_1 and R_2 are denoted as:

$S_1 = \{a\}, S_2 = \{x\}$ and $R_1 = \{b, c\}, R_2 = \{y, z\}$

- The server receives the communication sequence cs_i from the client and segregate the set of originating senders and receivers.
- Initially, the list of senders and receivers are empty.
- Iterate each communication sequence and segregate the originator and the list of recipient users.

Step 2:

- From the set of all communication sequences (CS), formulate I_1 and I_2.
- Scan two consecutive subsets from the set of receivers and originating senders I_1 and I_2, and apply Bernstein conditions,

$$R_i \cap S_j = NULL$$
$$S_j \cap S_i = NULL$$
$$R_j \cap S_i = NULL$$

  An additional condition $R_j \cap R_i = NULL$ is introduced to enhance the Bernstein conditions to analyze the dependency in communication sequences.
- If all the above conditions are satisfied, then the communication sequences are said to be independent and can be grouped in to different clusters, P_x and P_y.

Step 3:

- If any one of the Bernstein conditions is not satisfied, then the communication sequences are related to each other i.e., cs_j related to cs_i, represented as *jRi*.
- Scan through the communication sequences for a specified interval and identify the related communication sequences. The identified relations satisfy the conditions of Reflexivity, Symmetricity and Transitivity and hence the set of equivalence relations is constituted as $I_3 = \{3R2, 5R4, \ldots, jRi\}$.

Step 4:

- Prune I_3, scanning the relations from left to right for the elimination of relations which satisfy the properties of symmetricity and Transitivity. i.e., considering *iRj* implies *jRi* and the relations *jRi* and *iRk* imply *jRk*
- Construct the Equivalence Classes, *EC* from the set I_3.
- Initially $EC = \{\ \}$
- The equivalence relations are represented in a Linked List.
- Iterate each Equivalence Relation that is formed at Step 3, to construct the Equivalence Classes.

Step 4.1: If the Equivalence Class is empty, then add the Communication Sequences in the Equivalence Relation, otherwise perform the following steps:

Step 4.2: For each Equivalence Relation, verify the Communication Sequence existence in the dynamically updated Equivalence Class, if both the Communication Sequences exist, then skip that Equivalence Relation otherwise go to Step 4.3

Step 4.3: Check the Communication Sequence in the Dynamic Cluster for a match with the Communication Sequence in the Equivalence Relation and if match is found add it to the Dynamic Cluster, if there is no match then go to Step 4.4

Step 4.4: In case if the Communication Sequences are not matched with any one of the entries in the Dynamic Cluster, it indicates an independent Communication Sequence. So, add it to a new Cluster

Step 4.5: At any point of time if any match is found in different clusters, merge them to form one cluster due to the dependency found across clusters

Step 5:

- Repeat the process from Step 1 by continuously scanning the communication sequences received by the server application for a specified time of interval.

4 Equivalence Class-based Model for Identification of Individual Groups of Users – A Case Study

Bernstein conditions [9] are utilized to analyze the dependencies between groups of users communicating through social media applications.

The algorithm scans the communication sequences considering cs_i as reference and compares cs_j, $(j = i + 1, \ldots n)$ one at a time and the Bernstein conditions have been applied as given in the Step 2 of the algorithm taking in to the consideration of corresponding subsets from the sets of I_1 and I_2.

While considering two communication sequences among group of users i.e., $cs_1 =$ '$a \to b \to c$' and $cs_2 =$ '$d \to a \to e$', the users are interconnected i.e., tightly coupled. Here,

$$R_1 = \{b, c\},\ R_2 = \{a, e\} \quad \text{and}$$
$$S_1 = \{a\}, S_2 = \{d\}$$

One of the Bernstein conditions is not satisfied ($R_2 \cap S_1 \neq NULL$), there is a dependency in these two communication sequences and they are said to be related, $\{2R1\}$.

It is obvious when, $R_1 = \{b, c\}$ and $R_2 = \{e, f\}$, and $S_1 = \{a\}$ and $S_2 = \{d\}$, all the Bernstein conditions are satisfied and these communications sequences are said to be independent. In real-time, the above stated equivalence class algorithm is used to identify the independent group of users from the communication sequences as and when they have been received by the server application. If there is a dependency between any two given communication sequences, then they are said to be related. If there is any relation among communication sequences, then those senders and receivers are grouped in to a cluster. In a new communication scenario, considering the following set of communication sequences as given in Equation (1):

$$\begin{aligned} CS = \{&cs_1 : \text{‘}a \to b \to c\text{’}, cs_2 : \text{‘}x \to y \to z\text{’}, cs_3 : \text{‘}s \to x \to l\text{’}, \\ &cs_4 : \text{‘}b \to p \to q\text{’}, cs_5 : \text{‘}a \to p \to q\text{’}, cs_6 : \text{‘}r \to s \to v\text{’}, \\ &cs_7 : \text{‘}n \to a \to m\text{’}, cs_8 : \text{‘}w \to r \to s\text{’}\} \end{aligned} \tag{1}$$

The set of receivers, I_1 is represented as:

$$I_1 = \{\{b, c\}, \{y, z\}, \{x, l\}, \{p, q\}, \{p, q\}, \{s, v\}, \{a, m\}, \{r, s\}\} \text{ and}$$

the corresponding set of senders, I_2 is represented as:

$$I_2 = \{\{a\}, \{x\}, \{s\}, \{b\}, \{a\}, \{r\}, \{n\}, \{w\}\}$$

The set of receivers and senders (I_1 and I_2) have been numbered sequentially from 1 to 8. The first two subsets of I_1 and I_2 satisfy all Bernstein conditions and can be grouped in independent clusters. There is no dependency or relationship among the receivers and senders from the subsets 1 and 2 in I_1 and I_2. The receivers in the communication sequences as given as subsets 2 and 3 in I_1 with the corresponding set of senders in I_2 failed to satisfy the Bernstein condition, $R3 \cap S2 = NULL$. The set of communication sequences 3 and 2 are related since there is a dependency between them. This relationship is symbolically defined by $3R2$, read as 3 relates to 2. Similarly, the dependencies from all the remaining set of communication sequences have been identified and the following set of relations (I_3) formulated as shown in Equation (2).

$$I_3 = \{3R2, 4R1, 5R4, 6R3, 7R5, 8R6\} \tag{2}$$

While scanning through the set of communication sequences with the corresponding subsets in I_1 and I_2, the users in first two communication sequences can be grouped into two clusters namely P_1 and P_2. Since, the communication sequence 3 relates to communication sequence 2, it has also been grouped to P_2. The above set of relations I_3, will satisfy the following conditions:

i. Reflexivity
Each communication sequence has to be referred in the cluster in which it has been allocated. For example, the communication sequence 1 is to be referred in the cluster P_1 only.

ii. Symmetricity
The relation $4R1$ in I_3 indicates that the users in the communication sequences 4 and 1 have to be allotted to the same cluster. Since the communication sequence 1 has already been allocated to cluster P_1, the users in the communication sequence 4 have also been allocated to P_1. After identifying the dependency between communication sequences 1 and 4, the relation $4R1$, will also be same as $1R4$ which indicates the users (senders and receivers) involved in the relation $1R4$ are also grouped in the same cluster P_1. These two related communication sequences do not take part in any inter-cluster communication.

iii. Transitivity
The relations $5R4$ and $4R1$ in I_3 imply that $5R1$. That means, due to dependency or in order to eliminate inter-cluster communication, all the users involved in those relations 1, 4 and 5 have to be grouped in cluster P_1. Since the set of relations (I_3) satisfies all the conditions, it forms equivalence relations.

The procedure to determine equivalence class works in two phases. In the first phase, the equivalence pairs (*i, j*) are read and stored. An equivalence class is denoted by [*EC*] and it is defined to be the set of communication sequences to which [*EC*] is related. In phase two, all pairs of the form (*i, j*) are found. The values *i* and *j* are in the same class. By transitivity, all pairs of the form (*j, k*) imply *k* is also in the same class. Linked list representation is used to hold these pairs. Let *m* and *n* represent the number of related pairs and number of communication sequences respectively. Two arrays have been used. The array SEQ is used to store all the head nodes of *n* lists and the other is Boolean array OUT which tells whether or not a related pair has yet to be printed. The SEQ array is initialized with '*nil*' and OUT array is initialized with '*true*'.

For each relation *iRj*, two nodes are created. The array SEQ[*i*] points to a list of nodes that contains every number which is directly equivalenced to *i* by an input relation. Once a pair is read (*iRj*), *j* is placed on the SEQ[*i*] list and *i* is placed on the SEQ[*j*]. In the same way, all the pairs have been stored. The linked list representation of equivalence relations is shown in Figure 1.

The sequential array is scanned starting with the first *i*, $1 <= i <= n$ such that OUT[*i*] = *true*. Equivalent element in SEQ[*i*] is printed and OUT[*i*] is changed to false. In order to process the remaining lists which, by transitivity, belong to the same i, a stack of these nodes is created. The elements are popped out from this stack and the equivalence classes are printed. As a result

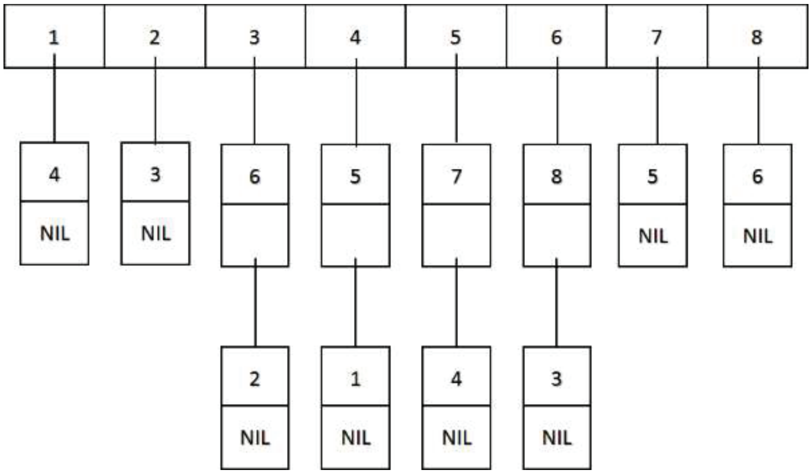

Figure 1 Linked list representation of equivalence relations {3R2, 4R1, 5R4, 6R3, 7R5, 8R6}.

of the reflexivity, symmetricity and transitivity of the relations, the given set of eight communication sequences has been partitioned into 2 equivalence classes: $\{1, 4, 5, 7\}$ and $\{2, 3, 6, 8\}$.

Two clusters have been effectively utilised for the eight communication sequences. The number of equivalence classes is the number of clusters to be allocated and the users of the same equivalence class are allocated to the same cluster.

5 Results and Discussion

A University network is considered as part of a social network where several groups of users are involved in sharing the information in terms of text, images and video messages. The users may have different interests in terms of the kind of messages that are part of information propagation and to whom the messages have to be sent. The proposed model is implemented based on socket programming in Python, where the server program will be listening to various clients' requests for the propagation of information. As and when the communication occurs across clients, the server application listens continuously to all the clients and whenever a communication occurs, keeps track of the communication sequence, extracts the sender and receiver, and updates the list of communication sequences.

In real-time, the number of users connected to any network are large in number and volume of communication sequences are mammoth. The graphical representation of users and their interactions in clusters P_1 and P_2 is given in Figure 2. To evaluate the approach of identifying the equivalence classes, two scenarios are considered and the proposed algorithm is applied and discussed as follows:

Scenario 1:

No. of Users: 15

Communication Sequences (Equation (1)**):**

$$\begin{aligned} CS = \{&cs_1 : \text{`}a \to b \to c\text{'}, cs_2 : \text{`}x \to y \to z\text{'}, cs_3 : \text{`}s \to x \to l\text{'},\\ &cs_4 : \text{`}b \to p \to q\text{'}, cs_5 : \text{`}a \to p \to q\text{'}, cs_6 : \text{`}r \to s \to v\text{'},\\ &cs_7 : \text{`}n \to a \to m\text{'}, cs_8 : \text{`}w \to r \to s\text{'}\} \end{aligned}$$

Equivalence Relations Identified:

$$\{\text{`}4R1\text{'}, \text{`}5R1\text{'}, \text{`}7R1\text{'}, \text{`}3R2\text{'}, \text{`}6R3\text{'}, \text{`}8R3\text{'}, \text{`}7R5\text{'}, \text{`}8R6\text{'}\} \tag{3}$$

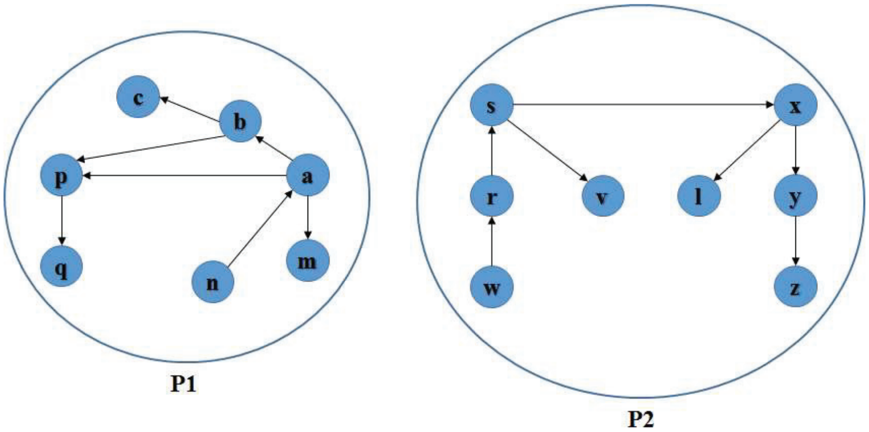

Figure 2 Graphical representation of users and their interactions in clusters P_1 and P_2 – proposed model – scenario 1.

Equivalence Classes Identified:

$$\{\{\text{‘4’}, \text{‘1’}, \text{‘5’}, \text{‘7’}\}, \{\text{‘3’}, \text{‘2’}, \text{‘6’}, \text{‘8’}\}\} \tag{4}$$

From the Equation (4), it is observed that there are two equivalence classes i.e., two independent clusters of communication sequences $\{\{$‘4’, ‘1’, ‘5’, ‘7’$\}$, $\{$‘3’, ‘2’, ‘6’, ‘8’$\}\}$, which indicate that the users’ interactions in the respective communication sequences are absolutely independent and are grouped in clusters P_1 and P_2. The set of users and their communication sequences in clusters $P_1 = \{a, b, c, p, q, m, n\}$ and $P_2 = \{x, y, z, l, s, r, v\}$ are shown in Figure 2. It is also observed the intra-cluster communication among the users of the clusters P_1 and P_2 are not dependent with each other at the time of analysis.

Scenario 2:

The graphical representation of users and their interactions in clusters P_1 and P_2 is shown in Figure 3. Further, to evaluate the dynamic nature of the real-time communication scenario, a new set of communication sequences is added and the observations made from the equivalence class approach are discussed as follows:

Four new communication sequences ‘$y \to a \to b$’, ‘$t \to u \to d$’, ‘$d \to i \to j$’ and ‘$k \to h \to d$’ are appended to the existing list as used in Scenario 1.

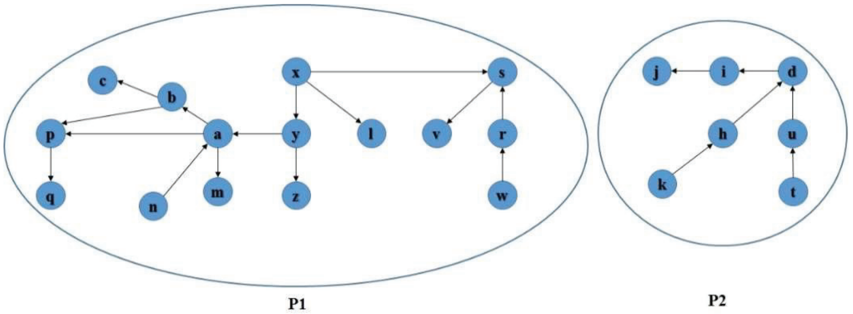

Figure 3 Graphical representation of users and their interactions in clusters P_1 and P_2 – proposed model – scenario 2.

Communication Sequences:

$$\begin{aligned}CS = \{&cs_1 : \text{‘}a \to b \to c\text{’}, cs_2 : \text{‘}x \to y \to z\text{’}, cs_3 : \text{‘}s \to x \to l\text{’},\\ &cs_4 : \text{‘}b \to p \to q\text{’}, cs_5 : \text{‘}a \to p \to q\text{’}, cs_6 : \text{‘}r \to s \to v\text{’},\\ &cs_7 : \text{‘}n \to a \to m\text{’}, cs_8 : \text{‘}w \to r \to s\text{’}, cs_9 : \text{‘}y \to a \to b\text{’},\\ &cs_{10} : \text{‘}t \to u \to d\text{’}, cs_{11} : \text{‘}d \to i \to j\text{’}, cs_{12} : \text{‘}k \to h \to d\text{’}\}\end{aligned} \tag{5}$$

No. of Users: 22

Equivalence Relations Identified:

$$\text{Existing: } \{\text{‘}4R1\text{’}, \text{‘}5R1\text{’}, \text{‘}7R1\text{’}, \text{‘}3R2\text{’}, \text{‘}6R3\text{’}, \text{‘}8R3\text{’}, \text{‘}7R5\text{’}, \text{‘}8R6\text{’}\} \tag{6}$$

$$\begin{aligned}\text{Updated: } \{&\text{‘}4R1\text{’}, \text{‘}5R1\text{’}, \text{‘}7R\text{’}, \text{‘}9R1\text{’}, \text{‘}3R2\text{’}, \text{‘}9R2\text{’}, \text{‘}6R3\text{’}, \text{‘}8R3\text{’},\\ &\text{‘}9R4\text{’}, \text{‘}7R5\text{’}, \text{‘}9R5\text{’}, \text{‘}8R6\text{’}, \text{‘}11R10\text{’}, \text{‘}12R11\text{’}\}\end{aligned} \tag{7}$$

Equivalence Classes Identified:

$$\{\{\text{‘4’}, \text{‘1’}, \text{‘5’}, \text{‘7’}, \text{‘9’}, \text{‘3’}, \text{‘2’}, \text{‘6’}, \text{‘8’}\}, \{\text{‘11’}, \text{‘10’}, \text{‘12’}\}\} \tag{8}$$

From the Equation (8), it is observed that the two independent equivalence classes of communication sequences {{‘4’, ‘1’, ‘5’, ‘7’}, {‘3’, ‘2’, ‘6’, ‘8’}} obtained in Scenario 1 as shown in Equation (4) are now merged in a single equivalence class {‘4’, ‘1’, ‘5’, ‘7’, ‘9’, ‘3’, ‘2’, ‘6’, ‘8’} due to the occurrence of inter-cluster communication in Scenario 2. The newly added ninth communication sequence $cs_9 =$ ‘$y \to a \to b$’ to the previous scenario causes dependency among clusters since the user 'y' in cluster P_2 initiates an interaction with user 'a' in cluster P_1. Further, the same message is forwarded from user 'a' to user 'b' as part of intra-cluster communication and hence new relations are formed, *9R1* and *9R2*. The communication sequence cs_9 has caused merging of clusters P_1 and P_2 (as shown in Figure 2) into a single cluster P_1 (as shown in Figure 3). It is also observed from Equation (8), the cluster P_2 is now updated with entirely new set of communication sequences $\{\text{'11'}, \text{'10'}, \text{'12'}\}$, which are independent of the communication sequences in P_1. The newly formed set of users and their communication sequences in clusters $P_1 = \{a, b, c, p, q, m, n, x, y, z, l, s, r, v, w\}$ and $P_2 = \{d, i, j, k, h, u, t\}$ are shown in Figure 3.

The equivalence class approach efficiently handles the dynamic nature of communication sequences and the clusters are formed accordingly. The outcome of the equivalence class approach is the formation of clusters that are independent to each other based on the dependency among the communication sequences at the given interval. The proposed model is applicable to the voluminous set of communication sequences in real-time and the absolute set of independent equivalence classes shall be obtained dynamically with respect to the communication sequences.

5.1 Community Detection in Social Networks

In any social network, based on the density of the interactions between the users, groups of users can be formed which are popularly known as communities [11, 12]. During the process of assignment of nodes to any community, the metric '*modularity*' [11] is used to evaluate how densely the users are connected with each other within the cluster. The Louvain's algorithm is the most popular algorithm in the field of detecting communities in large social networks based on discovering maximum modularity. The working principle of Louvain's method is to maximize the number and strength of the nodes in the same cluster and minimizing the number and strength of the nodes across clusters. The Louvain's procedure for community detection [11, 12] is described below:

Step 1
Initially assume the number of clusters is equal to the number of nodes.

Step 2
If a cluster experiences highest improvement in modularity, then the node is reassigned to it, else the node remains in the cluster to which it is assigned.

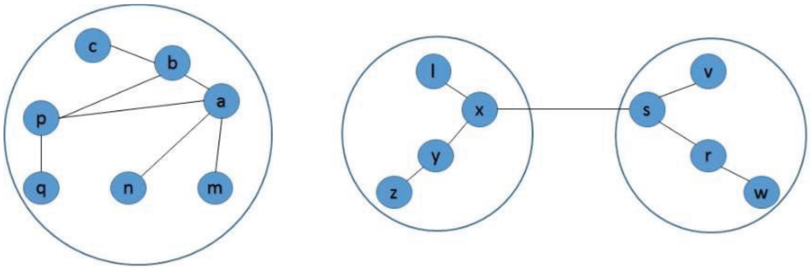

Figure 4 Graphical representation of users and their interactions in clusters P_1, P_2, and P_3 – Louvain's method – scenario 1.

Step 3
Represent each cluster found in Step 2 as a single node, and consolidating former inter-cluster edges into weighted edges that connect the new nodes.

The scenarios applied for the evaluation of the equivalence class approach are also considered with Louvain's method for comparison. Initially, the following communication sequences (Equation (1)) are applied to the Louvain's method and the partitions obtained are given in Figure 4.

$$\begin{aligned} CS = \{&cs_1 : \text{‘}a \to b \to c\text{’}, cs_2 : \text{‘}x \to y \to z\text{’}, cs_3 : \text{‘}s \to x \to l\text{’}, \\ &cs_4 : \text{‘}b \to p \to q\text{’}, cs_5 : \text{‘}a \to p \to q\text{’}, cs_6 : \text{‘}r \to s \to v\text{’}, \\ &cs_7 : \text{‘}n \to a \to m\text{’}, cs_8 : \text{‘}w \to r \to s\text{’}\} \end{aligned}$$

Scenario 1 – Louvain's Method:
No. of Users: 15
Partitions Identified: $\{\text{‘}a\text{’} : 0, \text{‘}b\text{’} : 0, \text{‘}c\text{’} : 0, \text{‘}m\text{’} : 0, \text{‘}n\text{’} : 0, \text{‘}p\text{’} : 0, \text{‘}q\text{’} : 0, \text{‘}x\text{’} : 1, \text{‘}y\text{’} : 1, \text{‘}z\text{’} : 1, \text{‘}s\text{’} : 2, \text{‘}l\text{’} : 1, \text{‘}r\text{’} : 2, \text{‘}v\text{’} : 2, \text{‘}w\text{’} : 2\}$

It is observed from Figure 4 that, Louvain's method has grouped the users in the given communication sequences 1 to 8 in to three different partitions. Partition 1 comprises of the users $\{a, b, c, p, q, m, n\}$ whereas partitions 2 and 3 comprises of users $\{x, y, z, l\}$ and $\{s, r, v, w\}$ respectively. It is noticed that among the clusters P_1, P_2, and P_3, the cluster P_1 is independent, and the clusters P_2 and P_3 are interconnected with the edge (x, s). In the case of equivalence class approach due to the dependent relation *3R2*, the communication sequences 2 and 3 and the corresponding users $\{x, y, z, l, s, r, v, w\}$ are grouped in cluster P_2 as shown in Figure 2. Further, to understand the behaviour of the Louvain's method, the observations are made by applying

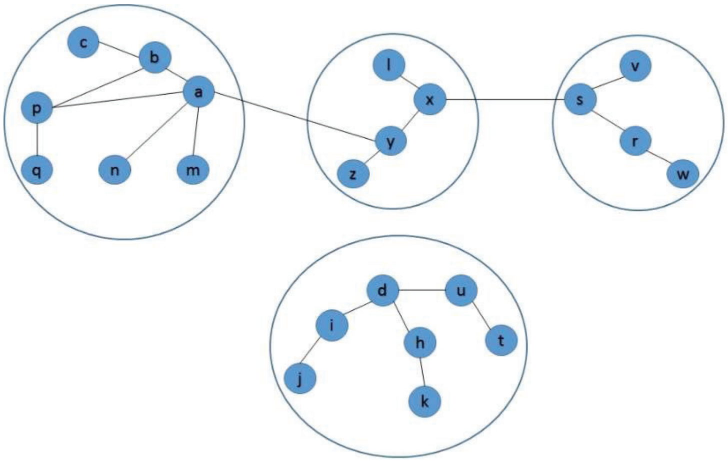

Figure 5 Graphical representation of users and their interactions in clusters P_1, P_2, P_3 and P_4 – Louvain's method – scenario 2.

the communication sequences as considered in the Scenario 2 of equivalence class approach and the results are shown in Figure 5.

Scenario 2 – Louvain's Method:
Communication Sequences (Equation (5)**):**

$$\begin{aligned}CS = \{&cs_1 : `a \to b \to c\text{'}, cs_2 : `x \to y \to z\text{'}, cs_3 : `s \to x \to l\text{'},\\ &cs_4 : `b \to p \to q\text{'}, cs_5 : `a \to p \to q\text{'}, cs_6 : `r \to s \to v\text{'},\\ &cs_7 : `n \to a \to m\text{'}, cs_8 : `w \to r \to s\text{'}, cs_9 : `y \to a \to b\text{'},\\ &cs_{10} : `t \to u \to d\text{'}, cs_{11} : `d \to i \to j\text{'}, cs_{12} : `k \to h \to d\text{'}\}\end{aligned}$$

No. of Users: 22
Partitions Identified:

$$\begin{aligned}\{&`a\text{'} : 0, `b\text{'} : 0, `c\text{'} : 0, `m\text{'} : 0, `n\text{'} : 0, `p\text{'} : 0, `q\text{'} : 0, `x\text{'} : 1, `y\text{'} : 1,\\ &`z\text{'} : 1, `s\text{'} : 2, `l\text{'} : 1, `r\text{'} : 2, `v\text{'} : 2, `w\text{'} : 2, `t\text{'} : 3, `u\text{'} : 3,\\ &`d\text{'} : 3, `i\text{'} : 3, `j\text{'} : 3, `k\text{'} : 3, `h\text{'} : 3\}\end{aligned}$$

It is observed from the Figure 5 that the Louvain method has grouped the users in the given communication sequences 1 to 12 in to four different partitions. Partition 1 comprises of the users $\{a, b, c, p, q, m, n\}$ whereas partitions

2, 3 and 4 comprises of users $\{x, y, z, l\}$, $\{s, r, v, w\}$ and $\{d, i, j, k, h, u, t\}$ respectively. It is noticed, in the Louvain method, among the clusters P_1, P_2, P_3 and P_4, the cluster P_4 is independent where as P_1 and P_2 are interconnected with an edge (*a, y*) and the clusters P_2 and P_3 are interconnected with the edge (*s, x*). In the equivalence class approach due to the dependent relations *3R2, 8R3* and *9R1*, the communication sequences cs_1, cs_2, cs_3, cs_8, cs_9 and all the corresponding users are grouped in cluster P_1 as shown in Figure 3 and in the case of independent communication sequences cs_{10}, cs_{11} and cs_{12}, the corresponding users $\{d, i, j, k, h, u, t\}$ are grouped as $P_{2,}$ which is independent to P_1 as shown in Figure 3. It is observed that in case of identifying the independent groups of communication sequences, the equivalence class approach outperforms the Louvain's method. In the proposed model, the number of partitions are exactly same as the number of independent groups of communication sequences.

To summarize, the modularity measure for clustering the groups will not exactly create partitions with respect to the number of independent groups and the users are scattered in various partitions with less strength where as in the proposed model even with a weaker strength across the users of different clusters, they are merged to form a single cluster due to the dependency in the communication sequences. This approach is helpful in identifying the actual originator of a message and its group. Extension of the proposed model to track the originator of any fake or viral messages, which causes potential threats to the society, is considered to be the future work.

6 Conclusion

The number of equivalence classes represent the number of clusters. Each equivalence class has distinct communication sequences as its members and all are related. In this approach, the inter-cluster communication has been eliminated. The concept of equivalence class approach with enhanced Bernstein conditions is applied to identify the absolute dependency in the communication sequences rather than the estimation of strength of relationship among users in each communication sequence or strength of nodes within any cluster or across clusters. This model provides the information of users and their interactions in one cluster, which are completely independent across other clusters. Louvain's algorithm is also applied to formulate the clusters as and when the communication sequences were received by the server application and to compare the results of the proposed equivalence class model to show the variations in the clustering sequences. As a continuation of this model, further investigation is in progress to predict the actual originator of any message by traversing the communication sequences across several users of a cluster. Further, it becomes inherent to identify the interests of members in each group and the exact roles played by each member in the group, thus an indication shall be given to the service providers not to transfer the messages, which are not relevant to that group. It is planned further to classify the behaviour of the users within any cluster for the reduction of propagating irrelevant messages which will aid in the reduction of the data traffic across packet data network and also to the end users' mobile data consumption.

References

[1] Number of social network users worldwide from 2010 to 2021. Statista, The Statistics Portal. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. (accessed 9 September 2021).

[2] Online Video & Entertainment. Statista, The Statistics Portal. https://www.statista.com/markets/424/topic/542/online-video-entertainment/. (accessed 9 September 2021).

[3] Reach & Traffic. Statistics and Market Data on Online Reach & Traffic. Statista, The Statistics Portal. https://www.statista.com/markets/424/topic/539/reach-traffic/. (accessed 9 September 2021).

[4] Global Social Network Penetration Rate as of January 2018. Statista, The Statistics Portal. https://www.statista.com/statistics/269615/social-network-penetration-by-region/. (accessed 9 September 2021).

[5] Most famous social network sites worldwide as on October 2018. Statista, The Statistics Portal. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. (accessed 9 September 2018).

[6] How Social Media Has Changed How We Communicate. Future of Work. https://fowmedia.com/social-media-changed-communicate. (accessed 9 September 2021).

[7] Manavik P. Raj, K. J. Joseph, Jesus Milton Rousseau, Corporate Communication & Social Media: A study of its usage pattern, International Journal of Humanities and Social Science Invention, 4(2015).

[8] Cisco Visual Networking Index: Forecast and Methodology, 2016–2021. White Paper Cisco Public. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf (accessed 9 September 2021).

[9] J. P. Tremblay, R. Manohar, Discrete Mathematical Structures with Applications to Computer Science, McGraw Hill Education, 2001.

[10] Kai Hwang, Faye A. Briggs, Computer Architecture and Parallel Processing, TATA McGraw Hill, 1985.

[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment, 2008.

[12] Annalyn Ng, Kenneth Soo, Numsense! Data Science for the layman: No Math Added, Kindle Edition, 2017.

[13] Ahmed Alsayat, Hoda El-Sayed, Social Media Analysis using Optimized K-Means Clustering, Proceedings of the 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), Towson, MD, 2016, pp. 61–66. doi: 10.1109/SERA.2016.7516129.

[14] Xu Yang, Yapeng Wang, Dan Wu, Athen Ma, K-Means Based Clustering on Mobile Usage for Social Network Analysis Purpose, Proceedings of the 2010 6th Conference on Advanced Information Management and Service (IMS), Seoul, 2010, pp. 223–228.

[15] Francis T. O'Donovan, Connie Fournelle, Steve Gaffigan, Oliver Brdiczka, Jianqiang Shen, Juan Liu, Kendra E. Moore, Characterizing User Behavior and Information Propagation on a Social Multimedia Network, Proceedings of the 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), San Jose, CA, 2013, pp. 1–6.

[16] Kuldeep Singh, Harish Kumar Shakya, Bhaskar Biswas, Clustering of People in Social Network based on Textual Similarity, Recent Trends in Engineering and Material Sciences, 2016, pp. 570–573.

[17] Rongjing Xiang, Jennifer Nevellie, Monica Rogati, Modeling Relationship Strength in Online Social Networks, Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26–30, 2010

[18] Keerthana, N., Vinod, V. and Sudhakar, S., A novel method for multidimensional cluster to identify the malicious users on online social networks. Journal of Engineering Science and Technology, 15(6), 2020, pp. 4107–4122.

[19] M. Turèanik, Web Users Clustering by their Behaviour on the Network, Proceedings of the 2020 New Trends in Signal Processing (NTSP), October 14–16, Demanovska dolina, Slovakia, 2020, pp. 1–5, doi: 10.1109/NTSP49686.2020.9229548.

[20] Kun He, Yingru Li, Sucheta Soundarajan, John E. Hopcroft, Hidden community detection in social networks, Information Sciences, Volume 425, 2018, pp. 92–106.

[21] L. Wu, Q. Zhang, C. Chen, K. Guo and D. Wang, "Deep Learning Techniques for Community Detection in Social Networks," in *IEEE Access*, vol. 8, pp. 96016–96026, 2020, doi: 10.1109/ACCESS.2020.2996001.

[22] Debadatta Naik, Dharavath Ramesh, Amir H. Gandomi, Naveen Babu Gorojanam, Parallel and distributed paradigms for community detection in social networks: A methodological review, Expert Systems with Applications, Volume 187, 2022, doi:10.1016/j.eswa.2021.115956.

Biographies

R. Srinivasan received his master's degree in Computer Applications from Alagappa University, Karaikudi and master's degree in Business Administration from Periyar University, Salem and an Executive Diploma from XLRI, Jamshedpur. He is currently a research scholar in Computer Science and Engineering at National Institute of Technology Nagaland. His area of Research work includes Data Analytics, Machine learning and Artificial Intelligence.

V. Ramachandran received his master's degree and philosophy of doctorate degree in Electrical Engineering from College of Engineering Guindy, Anna University, Chennai, India. He has 36 years of teaching experience at various levels in the Department of Information Science and Technology, College of Engineering, Anna University, Chennai. He is currently working as Professor in the Department of Computer Science and Engineering, DMI College of Engineering, Chennai, Tamilnadu, India. His research interest includes Cloud Computing, Web Technologies and Internet of Things.

Nagaraju Baydeti received his bachelor's and master's degrees in the field of Computer Science and Engineering from College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India. He received philosophy of doctorate degree in Computer Science and Engineering from National Institute of Technology Nagaland. He is currently working as an Assistant Professor in the Department of Computer Science and Engineering at National Institute of Technology Nagaland. His research interest includes Next Generation Mobile Networks, Wireless Communication and Networks, and Social Network Analysis.