Arabian Journal for Science and Engineering (2023) 48:1829–1843
https://doi.org/10.1007/s13369-022-07020-z

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

Seed Community Identification Framework for Community Detection over Social Media

Sumit Kumar Gupta[1] · **Dhirendra Pratap Singh**[1]

Received: 2 February 2022 / Accepted: 5 June 2022 / Published online: 19 July 2022

Abstract
The social media podium offers a communal perspective platform for web marketing, advertisement, political campaign, etc. It structures like-minded end-users over the explicit group as a community. Community structure over social media is the collaborative group of globally spread users having similar interests regarding a communal topic, product or any other axis. In recent years, researchers have widely used clustering techniques of data mining to structure communities over social media. Still, due to a lack of network and implicit communal information, researchers cannot bind mutually robust and modular community structures. The collaborative features of social media are inherent with implicit and explicit end-users. The explicit nature of both active and passive users is easily extracted from the graphical structure of social media. On the other hand, the degree of information inclusion of implicit features depends upon end-users participation. The Implicit features of frequently active users are diversely available, while integrating passive and silent users' implicit features over the community is tedious. This work proposed a social theory based influence maximization (STIM) framework for community detection over social media. It combines user-generated content with profile information, extracts passive social media users through influence maximization, and provides the user space for influencing inactive users. The STIM framework clusters identical nodes over the maximum influencing node axis based on their graphical parameters such as node degree, node similarity, node reachability, modularity, and node density. This framework also provides the structural, relational and mathematical concept for the functional grouping of like-minded people as a community over social media through social theory. Finally, an evaluation has been carried out over six real-time datasets. It analyses that convolution neural network over STIM structure more dense and modular communities via influence maximization. STIM acquired around 93% modularity and 94% Normalized Mutual Information (NMI), resulting in approximately 2.23% and 5.69% improvements in modularity and NMI, respectively, over the best-acquired result of the benchmark approach.

Keywords Social media · Community detection · Influence maximization · Social theory · CNN

1 Introduction

Today social media has become a vital platform for political campaigning, product advertisement and personal or professional publicity. Moreover, social media platforms facilitate the political parties, multinational companies and advertisement teams by offering a dynamic perspective to classify like-minded consumers and voters through community detection.

✉ Sumit Kumar Gupta
sumitgupta888@gmail.com

Dhirendra Pratap Singh
dpsingh.manit@gmail.com

[1] Department of Computer Science, MANIT, Bhopal, India

Social media can be defined as a network of interpersonal relationships of individuals for sharing and participating in many global and local issues. At the same time, a community is a group of individuals having similar desires towards any product, Political policy [1,2], Local [3] or international problem [4].

E-commerce, multinational corporate and political parties have benefited from the ability of social media to identify the structure of communities for policy making. Moreover, intercommunal characteristics of individuals over the communities assist in recognizing groups of crowdturfing users, sockpuppet nodes and fake profiles responsible for spreading fake reviews, rumours and hate speech over social media.

In recent research, Social media mining and clustering techniques have been widely used to identify communities

on social media. Unfortunately, due to a lack of network and implicit communal information, the performance of this research does not yield significant results. However, the selection of communal characteristics and seed nodes affects the intra-communal and inter-communal strength of the community structure.

On the other hand, Communal symptoms are dependent on the implicit and explicit relationships between nodes. The direct connection can be easily deduced from the graphical structure of active social media users. Still, implicit relationships between passive users pose a challenge, as does mining user-generated content via social media platforms such as its 'like' 'dislike', 'follow', 'comment', and 'share'. This concern motivates the development of a mining technique that efficiently extracts both implicit and explicit relationships. The primary concern of this work is to find the impact of implicit relationships between passive users over the structure of communities, i.e. modularities and NMI.

This article aims to develop an influence maximization framework for community detection to bind social media mining (SMM) theories over social media context. Apart from that, this paper introduces the concept of deep learning over the graph perspective of social media to carry high data sparsity and generate communities that include both passive and active users. This concept enables the identification of a trade-off between the similarity of nodes' attributes and the density of connections for Influence Maximization(IM).

The rest of the paper is organized as follows: Section 2 presents structuring of Community over social media; Section 3 covers the recent research on Community structuring ; Section 4 present social theory and influence maximization framework; Section 5 covers the experimental setup Sect. 6 illustrates the performance evaluation of benchmark overlapping community detection algorithm over the proposed framework. and finally, Sect. 7 concludes the paper and outlines the founding and future work.

2 Community Structuring Over Social Media

A community is a collection of human beings bound together by social, geographical, political, economic, and spiritual symmetry in the real world. On social media, communities are a collection of geographically dispersed end-users, share a common social profile and interest in politics, economics, art and culture, research and education, marketing, and other global issues.

Community on social media can be classified as explicit and implicit, on conviction of participation and belongingness of member. If the members of the community are aware of their involvement and belongingness, then it is an explicit community. Whereas, if a cluster of the social entity without knowing their participation and belongingness, share a similar view, opinion and interest over any topic, product, organization and issue are referred to an implicit community.

Recently, researcher focused to structuring efficient implicit dynamic relationship community [5], Learning community [6], knowledge-sharing community [7], identical topical community [8,9].

The algorithm used to detect communities on social media varies in nature depending on the type of parameter used to structure communities. Recently, researchers have concentrated their efforts on isolating either specific member-based communities or firm-based communities. The member-based community detection algorithm is also referred to as a node-centric algorithm because it structures the community of a similar node based on the desired node's specific parameter, such as degree, reachability, and similarity. In contrast, firm-based algorithms structure the community according to group-specific norms such as modularity, balance, density, robustness, and hierarchy.

Graphically, Social media (SM) can be illustrated as a complex relationship network r_n(u, l, r), as shown in Fig. 1. Where 'u' (node) is the set of the end-user having profile over social media podium, 'l' is a set of links between these entities 'u', and r: $u_x \times u_y \rightarrow$ l is a function which assigns relationship status (relative(r),friend(f),colleague(c)) between a pair of end-users u_x and u_y, join through a link $l_{(x,y)}$ as show in Eq. (1).

$$r : u_x \times u_y \rightarrow l^r \begin{cases} 1^r : \text{linked with direct relation } r \\ 0 : \text{Not linked no direct relation} \end{cases} \quad (1)$$

Whereas graphical, the objective of community structuring is to discover higher dense homogeneous sub graphs $sr_n(u_s, l_s, r_s) \subseteq r_n$(u, l, r). Where belonging factor(bf) of node u_s is higher in sr_n as compare to rest of graph ,as shown in the Eq. (2). Then sr_n is the community structure $\in r_n$.

$$bf(u_i, sr_n) > bf(u_i, r_n - sr_n), \forall\, u_i \in sr_n, \exists\, sr_n \in r_n \quad (2)$$

The colour shade of node in Fig. 1 indicate their belongingness over the topical context of the network. Whereas belonging factor bf(u_i, sr_n) is the probability of any node u_i belong to a community sr_n, and its a ratio of number of relationship link of node u_i connect to any node $\in sr_n$ and total number of relationship link of node u_i, as shown in Eq. (3).

$$bf(u_i, sr_n) = \frac{\sum_{v \in sr_n} l_{(v,u_i)}}{l_{u_i}} \quad (3)$$

However, if the relationship graph r_n having two disjoint dense subgraphs sr_{n_1} and sr_{n_2}, having no common node then both are disjoint community else overlapping communities, as shown in Eq. (4).

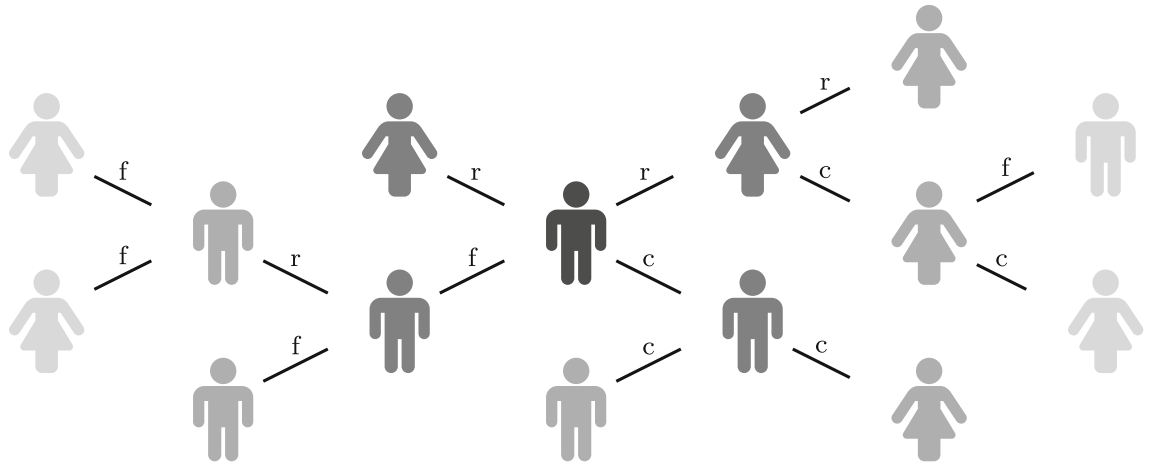

Fig. 1 Relation Sub-Graph over Social Network

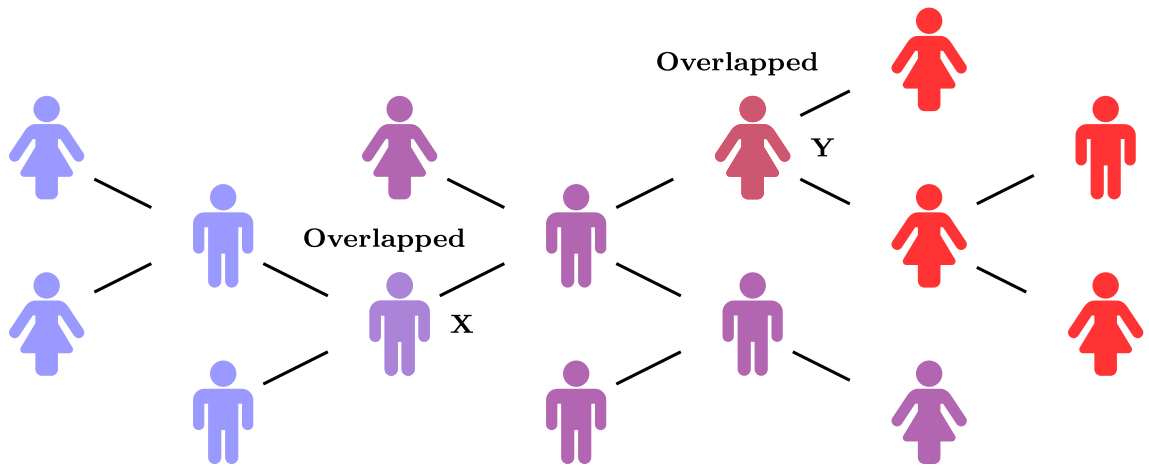

Fig. 2 Communities over Social Network

$$sr_{n_1} \cap sr_{n_2} \begin{cases} = \phi & \text{Disjoint community} \\ \neq \phi & \text{Overlapping community} \end{cases} \quad (4)$$

Apart form that Community detection can be overlapping where node may belong more then one group and If multiple edges containing the same node are assigned to multiple communities, the node is called overlapping one. The Dynamic relationship for overlapped community structure may be build by initializing the mutual knowledge categorization influence nodes and finally correlate overlapped node in shared community structure,as shown in Fig. 2. Where Node (X) and (Y) are basically belong to both blue and red communities with influence of violet communities.

3 Related Work

Social networking has grown in importance due to its unique ability to enable social contact between geographically dispersed users via the internet. A social network can be represented graphically, with nodes representing users and links representing their connections.

Increased interest in communities on social networking has resulted in a resurgence in graph mining algorithms. As a result, numerous techniques have been developed recently to address a broad range of graph mining, web content mining [10], feature extraction [11], and clinical [12] or economical management problems [13].

Saidi et al. [3] discussed an adequate evidential clustering method to identify cyber-terrorist subgroups as community over social media. This method employed probabilistic constrained evidential C-Means (CECM) algorithm to identify must-link and cannot-link constraints to magnify military, finance, and local leaders committees. simultaneously, Li et al. [14] build a correlation theory-based multi-layer network to encapsulate direct and indirect influence relation and drive local community detection. Whereas, Mohotti and Nayak [15] developed a density and content-based cluster-

ing approach for community detection over social media. Mohotti and Naya [16] presents Recurrent Unit (GRU) and Recurrent Neural Networks (RNNs) based approach to structure vulnerable communities responsible for spreading hate speech over social media.

Salehi and Davulcu [17] developed a social interactions and user-generated content based framework to identify detect antagonistic and allied communities over social media. This framework based on a hypothesis that the inter-community perspective of end-users can contemplate inter-community relationships.

Vogiatzis and Dimitrios Keros [18] present a social tagging and density-based community detection algorithm for crisp and overlapping community over hyper-graphs. Whereas, Zhao et al. [19] discussed a latent Dirichlet allocation-based probabilistic link partition (LBLP) model for overlapped community detection. This probabilistic model unified influence and content information of network structure.

Deng et al. [20] developed a probabilistic graph and vector Influence clustering-based community detection algorithm. This method use ITG (information transfer gain) to identify most influence node over communal desire group.Whereas, Hu et al. [21] proposed a Node2vec based spectral clustering algorithm for community detection. This algorithm extracts rich information from low dimension feature vector of identical phantom node.

Tommasel and Godoy [22] discussed a heterogenous community detection algorithm through asymmetric relations, which extract by integrates social and contextual information of end-users over the social media. Whereas, Croitoru et al. [23] present a spatio temporal clustering-based community detection algorithm and perform geospatial analysis of spatial footprint over both physical and cyberspace for information propagation.

Farooq et al. [24] build a graph-based community detection algorithm using influence node centralization. Authors visualizes and evaluates the correlation matrix to extract the most influential node over a communal desire group. Whereas, Katchapakirin et al. [1] present a natural language processing based behavioral information identification model for the social community detection over Facebook.

Abdelsadek et al. [25] discussed knowledge acquisition and interactive visualization dependent two complementary steps for community detection over twitter. However, Kanavos et al. [26] build an influence based graph mining approach for emotional community detection over social media. Whereas Moscato et al. [27] discussed a game theory-based logistics and information streams to identify most influence node on communal group over Twitter.

Hanteer et al. [2] present a hashtag-based topical audience model for multiplexing explicit interactions and conversations between end-users and identify politically influence community over Twitter.

Alduaiji et al. [28] discussed a clique structure and influence propagation-based temporal interaction biased model for community detection. This model grapes temporarily active users-over dense communities through density metric and influence of communal node with the frequency of their interactions with the sibling's node.

Toujani and Akaichi [29] build a genetic algorithm based hybrid hierarchical clustering approach for convergence of locally optimal community detection. Genetic hierarchical clustering algorithm build bottom-up cluster with higher objective function and decompose with lower quality function.

Wang et al. [30] present a homophily based twitter community topic modeling system, where LDA is used to capture the recent topics in the aggregated tweets. Clique algorithm used to verify internal and external topic similarity and extract common interests to build structure-based communities.

Singh et al. [31] discussed probabilistic data structure and quotient filter based storage schema for Community detection over social media. whereas SanchezOro and Duarte [32] proposed a metaheuristic based Greedy algorithm for community detection over social media. Ahmad et al. [33] developed a hybrid influence maximization approach based on dynamic weighted sum and multi-criteria decision-making methods for community detection over social media. Whereas, Guesmi et al. [34] present relational concept analysis based multi-relational community detection over heterogeneous social media.

Raj et al. [35] presents an artificial intelligence based social mining concept and successively employs the idea of granular computing in rough sets for structuring community over social media. whereas Elgazzar et al. [12] states that unsupervised machine learning could be used to keep an eye on health care by using dynamic, evolutionary clustering algorithms like DBSCAN and the Louvain method to find communities in the temporal networks of Covid 19.

Several studies have been done on network structure and partitioning algorithm, rather than node partitioning. An approach for detecting overlapping communities (MOEA/D) using genetics is presented in Wan et al. [36]. MOEA/D use link-based metrics to improve partition density, modularity, and mutual information. Using Dynamic optimization and a Dynamical resource allocation strategy, MOEA/D recognised community structure as a set of temporal overlapped snapshots. Van Lierde et al. [37] suggested an edge-based spectral clustering technique to extract overlapping groups. This approach first computes the graph laplacian eigenvalues. Then use k-means clustering to group the eigenvalue object.

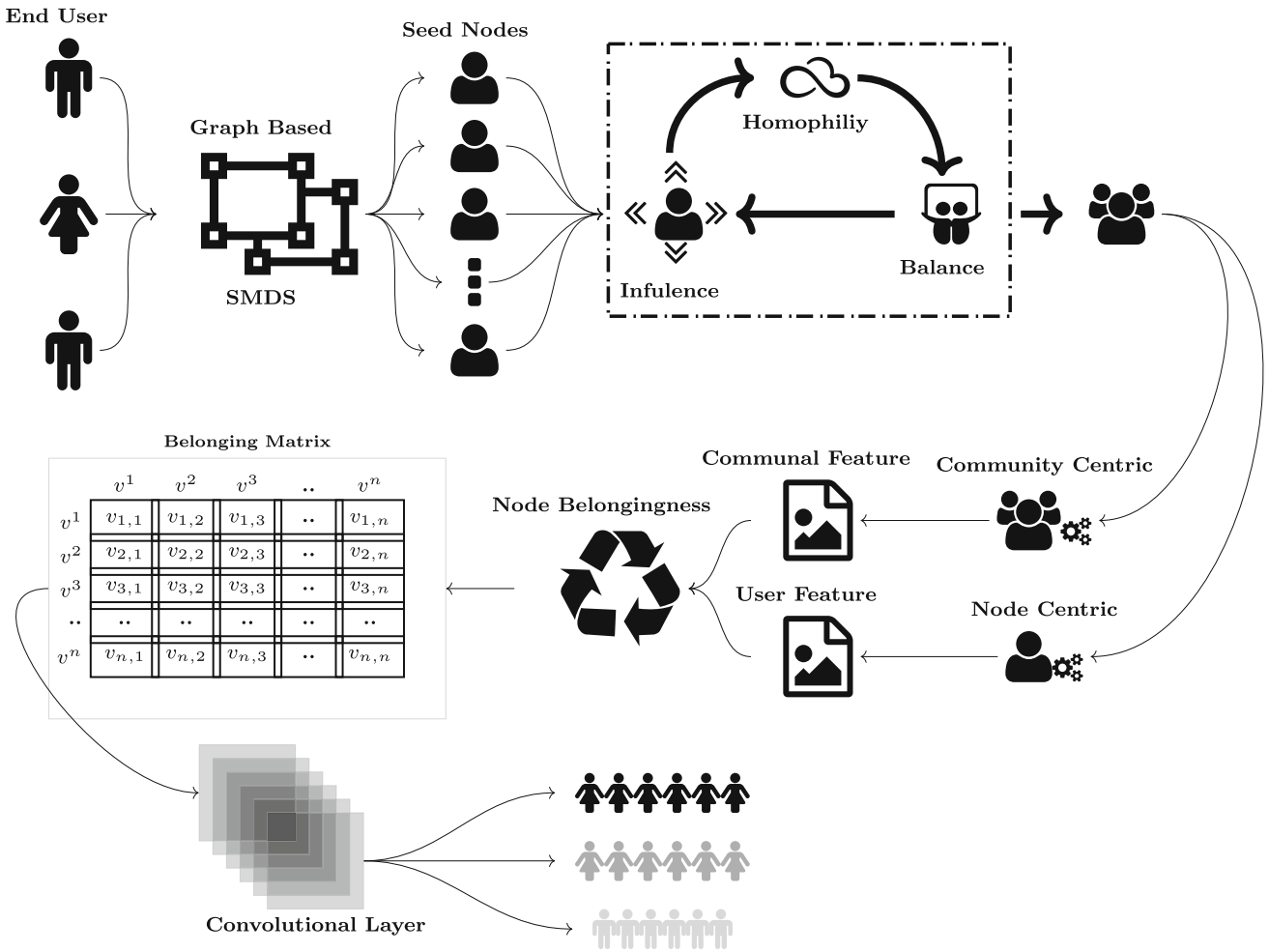

Fig. 3 Seed Communities identification framework for Overlapped Community detection

The existing approaches use user profiles and user-generated data to label end-users to a particular community in recent literature. However, the strength and modularity of these communities are fragile due to the lower participation of the passive, like-minded users. The existing approaches forgive inactive users who do not actively participate in any conversation but have similar ideologies over any products, politics, global and local issues, etc. So a social theory-based explicit community detection framework is needed to extract an influential community from social media platforms, encapsulating passive like-minded users over the community by applying graphs and social theories for influence maximization in the social network. Furthermore, as it is known that most influential users increase the flow of influence in the community, one more issue of community detection is taken, i.e. scalability in an extensive network.

4 Proposed STIM Framework

Community detection over social media is an essential and thought-provoking task. It is observed from the literature that the existing approaches use the users profile and user-generated data for the labelling end users to a particular community. However, Strength and modularity of these community are weak due to lower participation of the passive users.

To extract an effective community form social media platform, a social theory based explicit community detection framework is needed which encapsulate passive like-minded users over the community by applying graph and social theories for influence maximization in the social network. As it is known that most influential users increases the flow of influence in the community with this one more issue of community detection is taken, i.e. scalability in a large network.

To include passive users over the communities, a graphical and social theory based influence maximization (STIM)

framework is developed, as shown in Fig. 3. This framework combines user-generated content with profile information, extracts passive social media users through influence maximization, and provides the user space to influence inactive users.

STIM framework provides abstract, logical, and structural meaning to develop a mathematical model for grouping like-minded people as a community over social media with social theory. It's structuring the identical nodes that come in the aura of the influential nodes via node belongingness over node centric feature (NCF) and communal centric features (CCF) (node degree, node similarity, node reachability, modularity, and density of node) as a community.

4.1 Pattern Analysis

STIM framework analyses and extracts a proper pattern from the social media data set (SMDS). SMDS abstract both user-generated and profile data of end-user. This approach encapsulates social media with data mining concepts, theories, and algorithms to build a practical method for extracting communal intersection from user-generated and profile data. For instance, consider g(v,r_e) as a social media graph having 15 users nodes and 14 relationship edges with three different labels, as shown in Fig. 1. In g(v,r_e) labelling of relationship is build up by the amalgamation of user-generated and profile content over social media.

4.2 Seed Nodes Extraction

After extracting relationship edges and their belongingness over the topical context of desire pattern. STIM employed node centric feature (n_c^f) to discover clique structure as the seed nodes, as shown in algorithm 1. Algorithm 1 may return multiple seed nodes (u_s) having maximum connectivity over $R_n(u, l, r)$.

Algorithm 1 Seed Nodes $R_n(u, l, r)$

Input: $R_n(u, l, r)$
Output: Seed Node (u_s)

1: **function** SEED NODES($R_n(u, l, r)$)
2: $n = \sum_{j=1}^{n} u_i$;
3: **return** Number of nodes(n)$\in R_n(u, l, r)$
4: **for** $i \leftarrow 1$ to n **do**
5: $d\{u_i\} = \sum_{j=1}^{n} r_j[u_i - u_j]$;
6: **return** degree of node u_i
7: **end for**
8: **return** array d$[u_n] \leftarrow$ degree $\forall\, u_i \in R_n(u, l, r)$
9: $u_s = \underset{i<=\sum_{i=1}^{n} r_i}{argmax}\ \{d\{u_i\}\}$
10: **end function**
11: **return** u_s as seed node

4.3 Influence Maximization

After identifying the u_s, SMM used balance theory to derive an implicit relationship between another non-seed node and the seed node. Consider $R_n(u, l, r)$ as a social media graph with fifteen users nodes and fourteen relationship edges, as illustrated in Fig. 1. After applying SMM's balance theory, five hidden implicit relationships edges are extracted over the graph $R_n(u, l, r)$, as indicated by the blue line in Fig. 4.

Following the extraction of the hidden relationship, nodes are classified hierarchically according to their implicit status as determined by influence theory. After applying the influence theory, the clique's node colours are altered. As illustrated in Fig. 4, the brightness of the node's colour revealed its hidden implicit statuses across the clique. Following the discovery of the seed nodes, Influence Maximization employs balancing theory to deduce implicit connections between the passive and active users (seed node).

Simultaneously, the graph transmits effects as explicit characteristics derived from Influence and Homophily correlation theory. Through the Influence theory, the higher status communal node alters the lower status node's belongingness to their respective community. Whereas homophily establishes the belongingness of nodes with similar characteristics within the same community, as illustrated by the similar colour share in Fig. 4.

4.4 Seed Community Extraction

After increasing the participation of passive user over explicit communal characteristics. STIM use node and community centric feature for structuring seed community. Node Centric feature incorporate node degree, node similarity, and node reachability. Whereas, Modularity and node density of the network is the group centric feature. The graphical features set, select modular, dense, and robust cliques as a seed community and discover the desired community structure by clustering a similar node via social theory.

The structural equivalence between every pair of nodes is calculated through cosine similarity, as shown in Eq. (5) and select clique $R_n'(u', l', r')$ as seed community.

$$\sigma_{Cosine}(u_i, u_j) = \frac{|N(u_i) \cap N(u_j)|}{\sqrt{|N(u_i)||N(u_j)|}} \tag{5}$$

Whereas, k-Club is used to evaluate reachability of selected clique and proceed to refine it, as shown in Eq. (6).

$$min^{d}{}^{gp}(u_i, u_j) \leq k \left\{ maxi \left[\text{cohensive} \left[R_n'(u', l', r') \right] \right] \right\} \Rightarrow k-clan \tag{6}$$

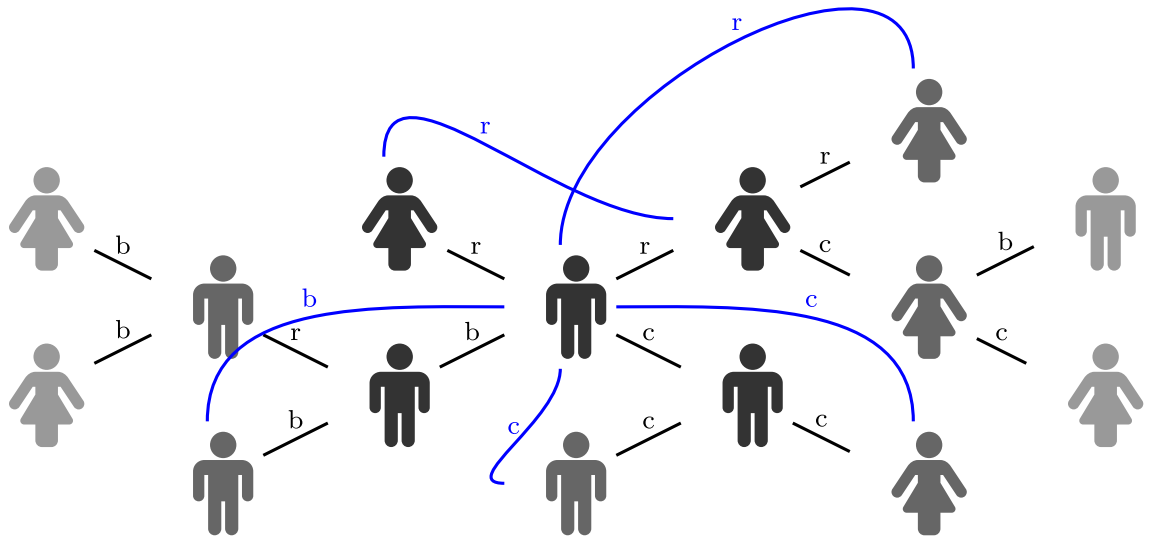

Fig. 4 Social Network after Influence and Balance Theory

Where,

$$(u_i, u_j) \in R_n^{'}(u^{'}, l^{'}, r^{'})$$
$$R_n^{'}(u^{'}, l^{'}, r^{'}) \subseteq R_n(u, l, r)$$

However, Ratio Cut (R_c(p)) and Normalized Cut (N_c(p)) over the vertex u, normalized the desired partition by volume of vertex or degree, as shown in Eqs. (7) and (8).

$$R_c(p) = \frac{1}{n}\sum_{i=1}^{n}\frac{cut(p_i, \bar{p_i})}{|p_i|} \tag{7}$$

$$N_c(p) = \frac{1}{n}\sum_{i=1}^{n}\frac{cut(p_i, \bar{p_i})}{v(p_i)} \tag{8}$$

Where,

$$\bar{p_i} = v - p_i \text{ (complement of cutset)}$$
$$\text{cut}(p_i, \bar{p_i}) = \text{Size of cut}$$

and

$$v(p_i) = \sum_{v \in p_i} d_v \text{ (summation of degree of vertex)}$$

The feature set of cut ratio returns more balance structure as seed community.

Subsequently, to increase the robustness of the resultant seed community structure, a k-connect feature of desire clique graph $R_n^{'}(u^{'}, l^{'}, r^{'})$ set minimum threshold (less than equal to k) of edge to maintain connectivity of desired seed community as shown in Eq. (9).

$$k - \text{connect}[R_n^{'}(u^{'}, l^{'}, r^{'})] = \begin{cases} \leq k \ : \ R_n^{'}(u^{'}, l^{'}, r^{'}) \text{ is connected} \\ > k \ : \ R_n^{'}(u^{'}, l^{'}, r^{'}) \\ is \text{ disconnected} \end{cases} \tag{9}$$

Whereas, Modularity set the minimum distance required among seed community for discrimination, as shown in Eq. (10).

$$d(p_i, p_j) = \sum_{u_i \in p_i, u_j \in p_j} l_{i,j} - \frac{d_i d_j}{2\sum l} \tag{10}$$

where d_i and d_j is degree of vertex u_i and u_j respectively. Distance can be generalized over all the partition of $R_n(u, l, r)$ define in Eq. (11).

$$\sum_{i=1}^{n}\sum_{u_i \in p_i, u_j \in p_j} l_{i,j} - \frac{d_i d_j}{2\sum l} \tag{11}$$

Whereas, after Compilation over all the link l belong to $R_n(u, l, r)$, distance define modularity as shown in Eq. (12).

$$m = \frac{1}{2\sum l}\left(\sum_{i=1}^{n}\sum_{u_i \in p_i, u_j \in p_j} l_{i,j} - \frac{d_i d_j}{2\sum l}\right) \tag{12}$$

However, Graph density evaluate the degree of completeness of clique that chose as a seed community, as shown in Eq.

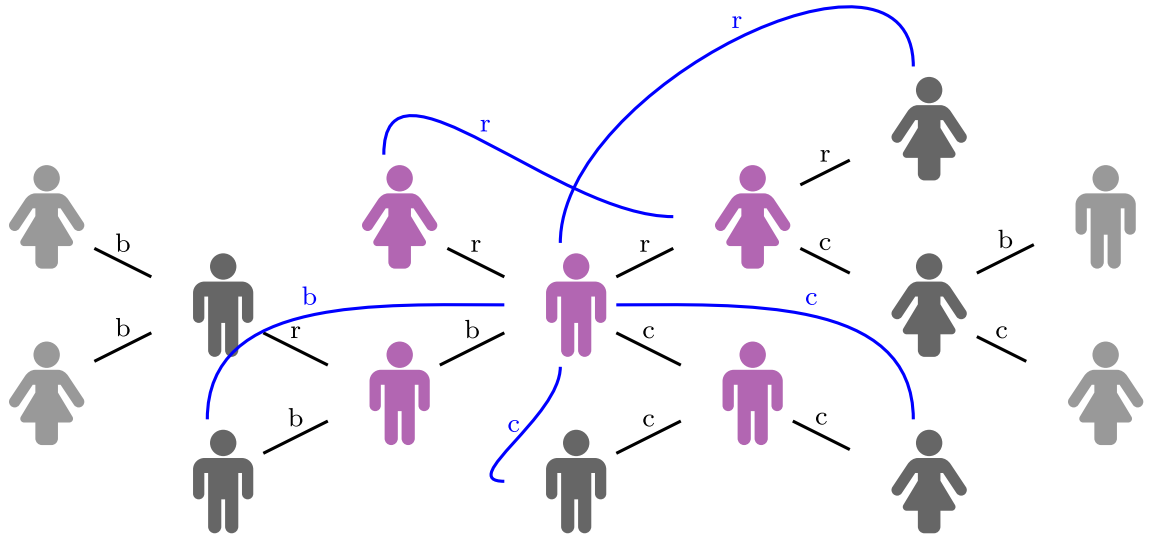

Fig. 5 Social Network after Homophily

(13).

$$\gamma(\text{Seed}_{\text{community}}) = \frac{l}{\frac{|e+1|}{2}} \tag{13}$$

where graph density maximized the density of desire clique that selected as seed community. Finally, Edge Betweenness minimize overlap edge among selected clique as seed community, as shown in Eq. (14).

$$l_i = \sum l_i \in \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{shortest}_{\text{path}}(u_i, u_j) \tag{14}$$

At the last integration of graphical feature, return clique structure as a seed community represented by a group of the Violet colour nodes shown in Fig. 5.

4.5 Node Belongingness

After extracting implicit seed communities through influence maximization, NCF generates vertex degree vector and reachability matrix, as shown in Eqs. (15) and (16).

$$n_d^v = \left\{n_d^1, n_d^2, n_d^3, ...n_d^m\right\} \quad \forall\, m \le n-1 \tag{15}$$

where n_d^v is represent node degree vector and n_d^i is the number of node having degree i in desire clique structure. Whereas, node_{rm} represent node rechability square matrix having n*n dimension and r_{u_i,u_j} is the modular distance between node u_i and u_j

$$\text{node}_{rm} = \left[r_{u_i,u_j}\right]_{n*n} \tag{16}$$

After extracting the node feature vector and matrix, multiplying the vertex degree vector and the node reachability matrix returns ui, j as the node with the greatest influence. Simultaneously, the K-means algorithm constructs a community of similar nodes using the Jaccard coefficient as the similarity index over the initial point $u_{i,j}$.

4.6 Convolution Neural Network

With Convolution Neural Network(CNN), this framework defines belonging matrix are modelled as user similarity vectors . A convolution neural network is then used to discover latent features from user similarity vector . Finally, aggregate the extracted community matrix as shown in Fig. 6 and Eq.(17).

$$r^c = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}\frac{cut(r(\theta)_i * r(\bar{c})_2)}{|m(c)|} \\ \frac{1}{n}\sum_{i=2}^{n}\frac{cut(r(\theta)_i * r(\bar{c})_2)}{|m(c)|} \\ \frac{1}{n}\sum_{i=n}^{n}\frac{cut(r(\theta)_i * r(\bar{c})_n)}{|m(c)|} \end{cases} \tag{17}$$

Where

1. $\chi_{r(c)_1}$ represents first community profile
2. $\chi_{r(c)_2}$ represent second community profile
3. $\chi_{r(c)_n}$ represent nth community profile

CNN uses the belonging matrix as the kernel matrix and uses the matrix's dimension as filter size. However, it implies convolution, max-pooling, fully connected, and softmax layer to define the probability distribution of each node over K communities. Where ReLU activation layer can employed

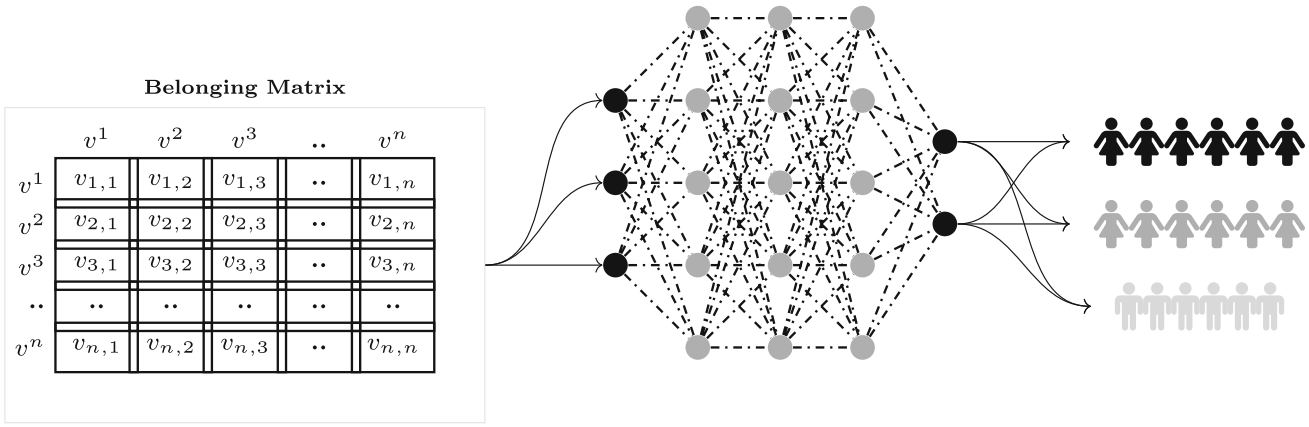

Fig. 6 Convolution Neural Network Model for Community Identification

Table 1 Modularity of structured community via CNN Algorithm over STIM

Data Set	Balance Theory (BT)	Social Homophily (SH)	Social Influence (SI)	Influence Maximization (IM)
Zachary Karate Club	84.72	84.12	86.24	89.72
American College football	80.62	84.22	87.48	90.24
Dolphin Social Network	77.26	79.84	82.72	87.08
Books about US Politics	80.12	83.42	86.24	90.19
Les Miserables	84.92	83.72	86.56	91.94
Word Adjacencies	79.12	78.68	81.18	84.34

Table 2 Normalized mutual information of structured community via CNN Algorithm over STIM

Data Set	Balance Theory (BT)	Social Homophily (SH)	Social Influence (SI)	Influence Maximization (IM)
Zachary Karate Club	90.07	91.12	93.78	95.12
American College football	89.28	90.52	94.12	96.79
Dolphin Social Network	86.34	89.72	92.22	94.12
Books about US Politics	89.42	90.89	92.34	96.11
Les Miserables	89.51	90.92	93.51	97.72
Word Adjacencies	86.62	85.08	87.72	90.26

gradient non-linearity functionality over CNN, as given in Eq. (18).

$$CNN_{Relu} = \xi\left(\sum_{x=1}^{n_{m-1}} conv\left(k_{i,j}^{m}, f_{j}^{m-1}\right) + w_i^m\right) \tag{18}$$

$$max\ p_i^t = max\left(max\ p_i^{t'} : t \le t' < t + pw(s * st)\right) \tag{19}$$

where f_j^{m-1} represent the j^{th} feature map at the 'm-1' layer; f_i^m represents i^{th} feature map at 'm' layer; $k_{i,j}^m$ represent kernel size of convolutional layer; n_{m-1} indicate the number of feature maps in the $(m-1)^{th}$ layer and w_i^m represent influence of i^{th} feature map in m layer. Dropout layer to reduces the dimension of feature map. The Max Pooling layer reduces two-unit areas on these feature maps to their maximum value, as shown in Eq. (19).

5 Environment Setup

The comprehensive experimental, computational environment is built for effectively analyzing the performance of proposed STIM community detection technique. The i^{11} CPU with 8.0 GB RAM, Integrated graphics card and a 1TB hard drive in windows 10 is used to perform these experiments. Further, the functionality of open source software node4j (version 0.9.1) and RStudio (version 3.3.0) are explored. The modules are developed in the python programming language.

The performance of the structuring community is evaluated using modularity(M) and normalised mutual information(NMI) parameters. Modularity is a network structural measure that assesses the strength of subgraphs (groups, clusters, or communities) within the network to extract community structure Newman [38]. In a network, a group of nodes with higher modularity is relatively dense next to one another, resulting in the formation of communities, as shown in Eq. (20).

$$M = \frac{1}{2|E|}\sum_{xy}\left[e_{xy} - \frac{w_x w_y}{2|E|}\right]\delta(c_x, c_y) = \sum_{i=1}^{n}\left(f_{ii} - f_i'^2\right) \tag{20}$$

Where e_{xy} represents the edge from node x to node y, W_x represent the summation of the weights of the edges linked to node x, c_x is the belonging community structure of node x,(c_x , c_y) is a probabilistic function that equals to 1 if both the respective node x and y belong to same community structure, otherwise 0. f_{ii} represent the edge in the community i and F_i' is the belonging probability of random edge to the community i that attached to vertices in the community i.

Whereas, Normalized mutual information (NMI) is a normalization of intra-community mutual information score to scale the similarity between intracommunity node discused in Eq. (21):

$$NMI(x, c) = \begin{cases} 0 \text{ node are totally dissimilar} \\ 1 \quad \text{node are totally similar} \end{cases} \tag{21}$$

and mutual information is calculated as shown in Eq. (22):

$$NMI(x, c) = \frac{2 * i(x, c_i)}{e(x) + e(c)} \tag{22}$$

where x is the class label, c is the community structure, e is the entropy and $i(x;c)$ is the information gain for element c_i for class label x.

6 Result Analysis

To evaluate the performance of structuring community, six different experimental campaigns are carried out over social media based real network data sets, namely Word adjacencies (WA), Zachary karate club (ZKC)[39], Dolphin social network(DCN) [40], Les Miserables (LM), Books about US politics(BP) and American College football (ACF) [41].

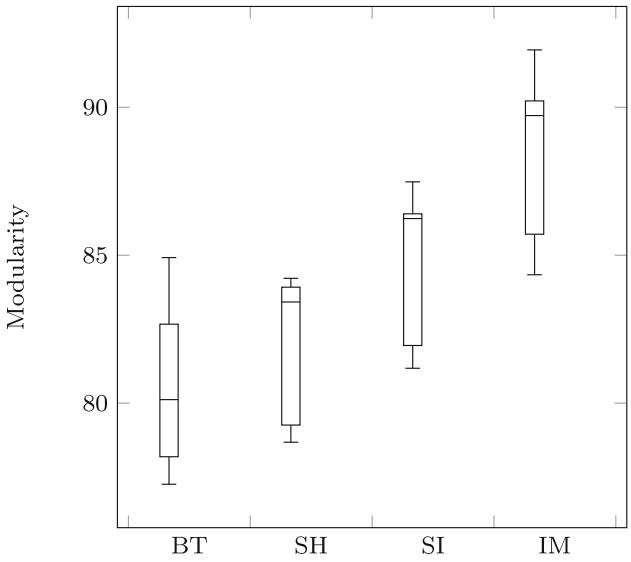

Fig. 7 Statistical Distribution of Modularity of STIM with Data set

6.1 Performance Evaluation of CNN over STIM Framework

Performance evaluation of CNN algorithm over STIM framework for structuring community on different social media based data set, after incorporating social media mining theory is illustrated in Tables 1 and 2 as Modularity and Normalized Mutual information.

CNN algorithm over Proposed STIM framework structuring the community on social media dataset with the modularity of 77.26–84.92%, 78.68–84.22% and 81.18–87.98% after incorporation of balance theory(BT) , social homophily(SH) and social influence(SH), respectively, as shown in Fig. 7.

After incorporating balance theory, it's observed that the CNN algorithm over STIM framework acquire maximum modularity on an unbalanced LM dataset. In contrast, with Homophily and influence theory, the STIM framework acquires maximum modularity on lightly correlated ACF datasets. However, after incorporating all SMM theories as influence maximization(IM), the modularity of the resultant community is significantly increased and structuring 84.34–91.94% modular community.

CNN algorithm over proposed STIM framework after amalgamation of influence maximization leads the highest modularity on the high dense imbalanced networks as LM and ACF datasets.

On the other hand, evaluation of the CNN algorithm over proposed STIM framework with the concern of inter-community normalized mutual information. It observed that resultant structured communities have 86.34–90.07%, 85.08–91.12% and 87.72–94.12% after incorporating bal-

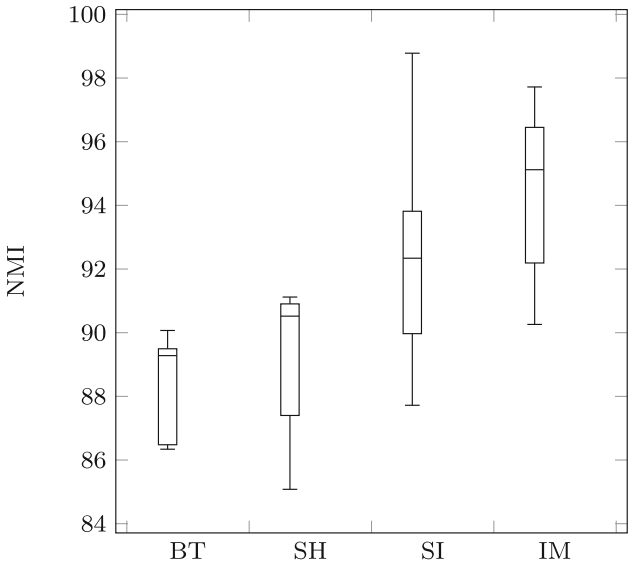

Fig. 8 Statistical distribution of NMI of STIM with data set

ance theory, Homophily and influence theory, respectively, as shown in Fig. 8.

After incorporating influence theory, it's observed that the CNN algorithm over STIM framework acquire maximum NMI on a high dense lightly correlated ACF dataset. In contrast, with Balance and Homophily theory, the STIM framework achieves maximum NMI on high dense imbalanced ZKC datasets. However, after incorporating all SMM theories as influence maximization, the NMI within the resultant community is significantly increased and structuring 90.26–97.72% mutually informative node community.

CNN algorithm over proposed STIM framework, after the amalgamation of influence maximization, structured a highly informative community on the high dense imbalanced networks as LM and ACF datasets.

6.2 Performance Evaluation of Benchmark Algorithm With Influence Maximization

Performance evaluation of benchmark algorithm for structuring community on different social media based data set, after incorporating social media mining theory is illustrated in Tables 3 and 4 as Modularity and Normalized Mutual information.

The modularity of structured community via Walktrap(WT) algorithm on social media data set is statically distributed as 68.25–72.84%, 70.82–78.64% and 75.86–79.64% with the amalgamation of balance, homophily and influence theory. However, Mutual information of structured community is slightly higher than modularity, i.e. statically distributed as 70.26–77.82%, 74.52–82.42% and 78.62–86.24% on balance, homophily and influence theory.

Whereas, after incorporating all SMM theories as influence maximization, the modularity of the resultant community via WT algorithm is significantly increased and structuring 78.36–82.78% modular and 82.72–89.72% informative community, as shown in Tables 3, 4 and Fig. 9.

However, the modularity of structured community via Fast-Greedy(FG) algorithm on social media data set is statically distributed as 69.85–76.26%, 72.24–79.12% and 74.62–80.60% with the amalgamation of balance, homophily and influence theory. However, Mutual information of structured community is slightly higher than modularity, i.e. statically distributed as 74.14–81.52%, 76.18–85.24% and 79.02–89.12% on balance, homophily and influence theory.

Whereas, after incorporating all SMM theories as influence maximization, the modularity of the resultant community via FG algorithm is significantly increased and structuring 80.59–84.56% modular and 84.84–91.80% informative community, as shown in Tables 3, 4 and Fig. 10.

Whereas, Edge Betweenness(EB) algorithm sturcturing community acquired 70.12–77.26%, 74.23–80.18% and 78.24–82.24% statically distributed modularity and 78.08–82.54%, 79.68–86.48% and 82.46–88.98% statically distributed NMI with the amalgamation of balance, homophily and influence theory.

Whereas, after incorporating all SMM theories as influence maximization, the modularity of the resultant community via EB algorithm is significantly increased and structuring 81.22–86.24% modular and 87.67–92.78% informative community, as shown in Fig. 11.

The performance of the baseline community detection algorithm is significantly boosted up after rectifying network information by incorporating influence maximization as social theories. The community detection algorithm, Walk trap, Fast greedy and Edge-betweenness acquire approximate 4–5%, 3–7% and 3– 7% improvement in modularity and 3–5%, 5–10% and 3– 10% improvement in NMI with the stand-alone case of social theories over social media data sets respectively.

6.3 Comparative Result Analysis

Further to validate the proposed STIM community detection framework's a comparison has been made with the state-of-the-art community detection algorithm on the social media-based real network data set, including Walk-trap, fast-greedy and edge betweenness.

The performance of CNN over STIM community detection framework's for structuring modular and informative communities on social media data set, is illustrated in Tables 1 and 2 as influence maximization.At the same time, Modularities and NMI of structuring communities via state of the art algorithms (Walk trap, Fast greedy and Edge-betweenness) are illustrated in Tables 3 and 4.

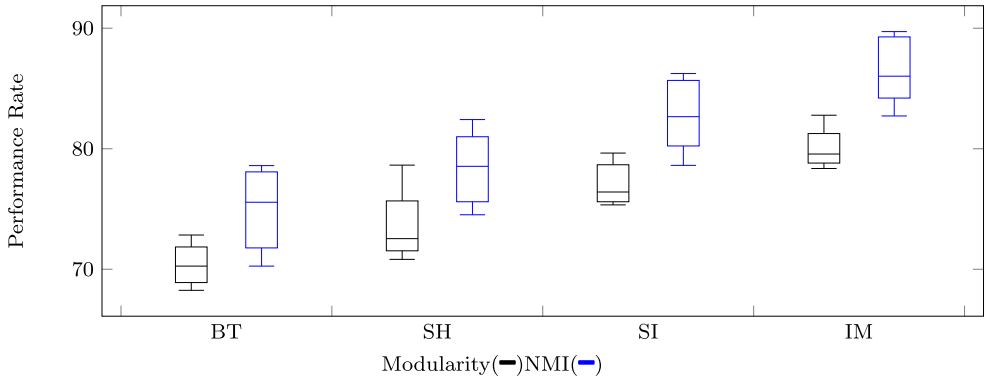

Fig. 9 Statistical distribution of performance of walktrap algorithm with dataset

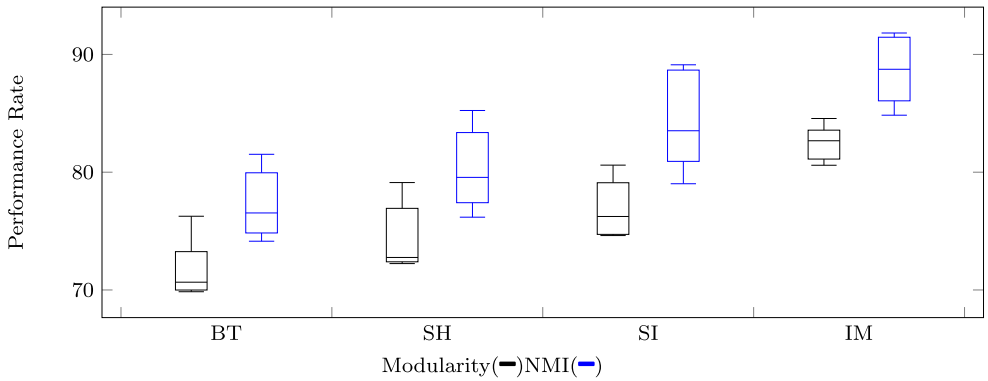

Fig. 10 Statistical Distribution of performance of fast greedy with data set

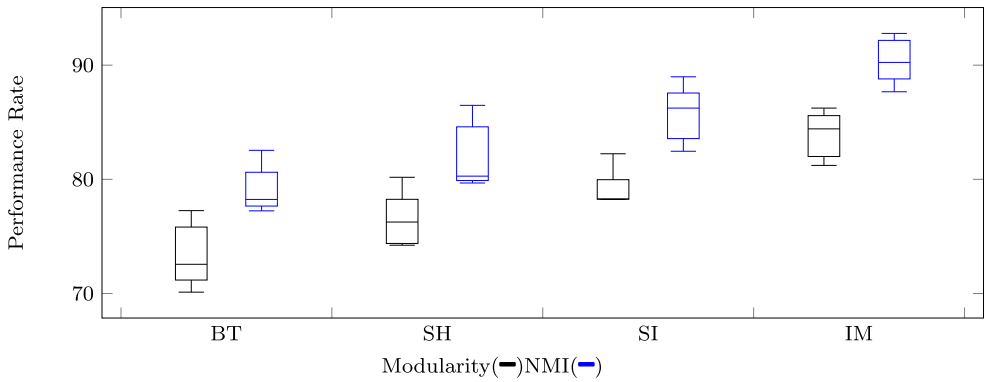

Fig. 11 Statistical distribution of performance of edge betweenness with data set

Table 3 Modularity on existing approach

Data Set	Walktrap (WT)				Fast-Greedy (FG)				Edge Betweenness (EB)			
	BT	SH	SI	IM	BT	SH	SI	IM	BT	SH	SI	IM
Zachary Karate Club	70.26	72.54	76.41	79.26	71.29	75.24	78.27	82.67	72.24	76.26	78.66	84.42
American College football	71.56	74.52	78.52	81.24	70.66	72.76	74.82	82.87	74.52	77.28	78.24	86.24
Dolphin Social Network	69.54	72.24	75.34	79.51	69.85	72.54	76.24	81.64	70.12	74.23	78.28	82.78
Books about US Politics	72.84	78.64	79.64	82.78	76.26	79.12	80.6	84.56	77.12	80.18	81.28	84.98
Les Miserables	72.14	76.82	78.82	81.28	75.22	78.62	79.94	84.27	77.26	79.22	82.24	86.24
Word Adjacencies	68.25	70.82	75.86	78.36	70.14	72.24	74.62	80.59	72.56	74.54	78.28	81.22

Table 4 Normalized mutual information on existing approach

Data Set	Walktrap (WT)				Fast-Greedy (FG)				Edge Betweenness (EB)			
	BT	SH	SI	IM	BT	SH	SI	IM	BT	SH	SI	IM
Zachary Karate Club	75.56	78.54	82.66	86.02	76.54	79.56	83.52	88.74	78.08	80.28	84.68	90.24
American College football	78.6	80.25	86.24	89.72	79.68	82.45	88.88	91.12	80.02	84.22	86.24	92.78
Dolphin Social Network	73.28	76.68	81.84	85.69	75.54	78.64	82.82	87.28	77.24	79.68	84.28	89.92
Books about US Politics	78.34	82.42	85.92	89.48	80.22	84.28	88.46	91.82	81.22	84.98	88.88	92.12
Les Miserables	77.82	81.74	85.42	89.08	81.52	85.24	89.12	91.8	82.54	86.48	88.98	92.22
Word Adjacencies	70.26	74.52	78.62	82.72	74.14	76.18	79.02	84.84	78.24	80.12	82.46	87.67

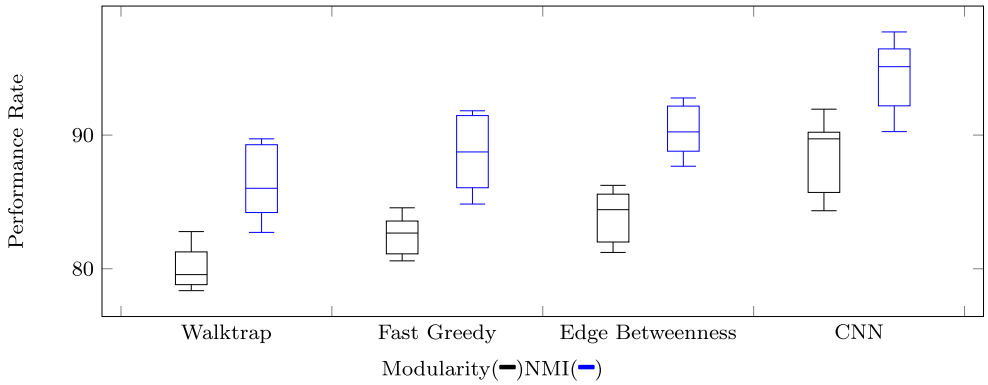

Fig. 12 Statistical distribution of comparative performance

CNN algorithm acquires 84.34–91.94% modular and 90.26–97.72% informative community over STIM framework. Whereas WT algorithm gain 78.36–82.78% modular and 82.72–89.72% informative community, FG algorithm 80.59–84.56% modular and 84.84–91.80% informative community and EB algorithm 81.22–86.24% modular and 87.67–92.78% informative community over different variant of social media based data set, as shown in Fig. 12.

CNN algorithm over STIM framework significantly leads the performance and acquired approximate 1.14%, 3.92%, 3.16%, 2.48%, 5.41% and 5.69% improvement in NMI over the best-acquired result by benchmark technique with the amalgamation of graphical and social theory, over six different real network data sets respectively. The variation of improvement is because of the network parameter of the data set. STIM leads the performance over the high dense network as its have higher implicit relationships between passive users. CNN algorithm acquires mutually dense community structure as its select highest influence node over implicit seed community as the axis of the community through influence maximization.

7 Conclusion

This paper proposed the four-tier influence maximization framework for community detection (STIM) over social media. STIM amalgamates graphical centric feature (GCF) and social theory for extracting implicit clique structure as seed community and proceeds to discover the desired community structure by the axis of influence maximization node.

After assessing community detection performance with the axis of the influence maximization based seed node over different social media data sets, it observed that the STIM significantly structured highly dense and informative community. Acquired approximate 93% modularity and 94% NMI over the community structure over different variants of data sets. In contrast, STIM leads the performance by 2.23% improvement in modularity and 5.69% improvement in NMI over the best-acquired result by benchmark technique with the amalgamation of graphical and social theory over six different real network data sets respectively.

References

1. Katchapakirin, K.; Wongpatikaseree, K.; P. Yomaboot, P.; Kaewpitakkun, Y.: Facebook social media for depression detection in the thai community. In 2018 15th international joint conference on

computer science and software engineering (jcsse), pp 1–6. (2018) https://doi.org/10.1109/JCSSE.2018.8457362.
2. Hanteer, O.; Rossi, L.; D'Aurelio, D.V.; and Magnani, M.: From interaction to participation: the role of the imagined audience in social media community detection and an application to political communication on twitter. In 2018 IEEE/ACM international conference on advances in social networks analysis and mining (asonam), pp 531–534. (2018) https://doi.org/10.1109/ASONAM.2018.8508575.
3. Saidi, F.; Trabelsi, Z.; and Ghazela, H.B.: A novel approach for terrorist sub-communities detection based on constrained evidential clustering. In 2018 12th international conference on research challenges in information science (rcis), pp 1–8. (2018)
4. Gupta, S.; Singh, D.P.: Recent trends on community detection algorithms: a survey. Mod. Phys. Lett. B **34**(35), 2050408 (2020)
5. Sheng, J.; Wang, K.; Sun, Z.; Wang, B.; Khawaja, F.; Ben, L.; Zhang, J.: Overlapping community detection via preferential learning model. Phys. A: Stat. Mech. Appl. **527**, 121265 (2019)
6. Lei, Y.; Zhou, Y.; Shi, J.: Overlapping communities detection of social network based on hybrid c-means clustering algorithm. Sustain. Cities Soc. **47**, 101436 (2019)
7. Yuan, Y.; Soh, D.W.; Yang, H.H.; Quek, T.Q.S.: Learning overlapping community-based networks. IEEE Trans. Signal Inform. Process over Netw. **5**(4), 684–697 (2019)
8. Reihanian, A.; Feizi-Derakhshi, M.-R.; Aghdasi, H.S.: Overlapping community detection in rating-based social networks through analyzing topics, ratings and links. Pattern Recogn. **81**, 370–387 (2018)
9. Whang, J.J.; Hou, Y.; Gleich, D.F.; Dhillon, I.S.: Non-exhaustive, overlapping clustering. IEEE Trans. Pattern Anal. Mach. Intell. **41**(11), 2644–2659 (2019)
10. Patel, O.P.; Bharill, N.; Tiwari, A.; Patel, V.; Gupta, O.; Cao, J.; Li, J.; Prasad, M.: Advanced quantum based neural network classifier and its application for objectionable web content filtering. IEEE Access **7**, 98069–98082 (2019). https://doi.org/10.1109/ACCESS.2019.2926989
11. Bharti, K.K.; Singh, P.K.: Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. Expert Syst. Appl. **42**(6), 3105–3114 (2015)
12. Elgazzar, H.; Spurlock, K.; Bogart, T.: Evolutionary clustering and community detection algorithms for social media health surveillance. Mach. Learn. Appl. **6**, 100084 (2021)
13. Pizzuti, C.: Evolutionary computation for community detection in networks: A review. IEEE Trans. Evol. Comput. **22**(3), 464–483 (2018)
14. Li, X.; Xu, G.; Lian, W.; Xian, H.; Jiao, L.; Huang, Y.: Multi-layer network local community detection based on influence relation. IEEE Access **7**, 89051–89062 (2019)
15. Mohotti, W.A.; and Nayak, R.: Corpus-based augmented media posts with density-based clustering for community detection. In 2018 IEEE 30th international conference on tools with artificial intelligence (ictai), pp 379–386. (2018)
16. Mossie, Z.; Wang, J.-H.: Vulnerable community identification using hate speech detection on social media. Inform. Process. Manag. **57**, 102087 (2020)
17. Salehi, A.; and Davulcu, H.: Detecting antagonistic and allied communities on social media. In 2018 ieee/acm international conference on advances in social networks analysis and mining (asonam), pp 99–106. (2018)https://doi.org/10.1109/ASONAM.2018.8508297.
18. Vogiatzis, D.; and Dimitrios Keros, A.: A density based algorithm for community detection in hyper-networks. In 2017 12th international workshop on semantic and social media adaptation and personalization (smap), pp 59–64. (2017)
19. Zhao, S.; Yu, L.; Cheng, B.: Probabilistic community using link and content for social networks. IEEE Access **5**, 27189–27202 (2017)
20. Deng, X.; Zhai, J.; Lv, T.; Yin, L.: Efficient vector influence clustering coefficient based directed community detection method. IEEE Access **5**, 17106–17116 (2017)
21. Hu, F.; Liu, J.; Li, L.; Liang, J.: Community detection in complex networks using node2vec with spectral clustering. Stat. Mech. Appl. Phys. A **545**, 12433 (2019)
22. Tommasel, A.; Godoy, D.: Multi-view community detection with heterogeneous information from social media data. Neurocomputing **289**, 195–219 (2018)
23. Croitoru, A.; Wayant, N.; Crooks, A.; Radzikowski, J.; Stefanidis, A.: Linking cyber and physical spaces through community detection and clustering in social media feeds. Comput. Environ. Urban Syst. **53**, 47–64 (2015)
24. Farooq, A.; Joyia, G.J.; Uzair, M.; and Akram, U.:Detection of influential nodes using social networks analysis based on network metrics. In 2018 international conference on computing, mathematics and engineering technologies (icomet), pp 1–6. (2018) https://doi.org/10.1109/ICOMET.2018.8346372.
25. Abdelsadek, Y.; Chelghoum, K.; Herrmann, F.; Kacem, I.; Otjacques, B.: Community extraction and visualization in social networks applied to twitter. Inf. Sci. **424**, 204–223 (2018)
26. Kanavos, A.; Perikos, I.; Hatzilygeroudis, I.; Tsakalidis, A.: Emotional community detection in social networks. Comput. Electr. Eng. **65**, 449–460 (2018)
27. Moscato, V.; Picariello, A.; Sperlaa, G.: Community detection based on game theory. Eng. Appl. Artif. Intell. **85**, 773–782 (2019)
28. Alduaiji, N.; Datta, A.; Li, J.: Influence propagation model for clique-based community detection in social networks. IEEE Trans. Comput. Soc. Syst. **5**(2), 563–575 (2018). https://doi.org/10.1109/TCSS.2018.2831694
29. Toujani, R.; and Akaichi, J.: Ghhp: Genetic hybrid hierarchical partitioning for community structure in social medias networks. In 2018 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), pp 1146–1153. (2018) https://doi.org/10.1109/SmartWorld.2018.00199.
30. Wang, F.; Orton, K.; Wagenseller, P.; Xu, K.: Towards understanding community interests with topic modeling. IEEE Access **6**, 24660–24668 (2018). https://doi.org/10.1109/ACCESS.2018.2815904
31. Singh, A.; Garg, S.; Batra, S.; Kumar, N.: Probabilistic data structure-based community detection and storage scheme in online social networks. Futur. Gener. Comput. Syst. **94**, 173–184 (2019)
32. Sanchez-Oro, J.; Duarte, A.: Iterated greedy algorithm for performing community detection in social networks. Futur. Gener. Comput. Syst. **88**, 785–791 (2018)
33. Ahmad, A.; Ahmad, T.; Bhatt, A.: Hwmscb: a community-based hybrid approach for identifying influential nodes in the social network. Stat. Mech. Appl. Phys. A **545**, 123590 (2019)
34. Guesmi, S.; Trabelsi, C.; Latiri, C.: Community detection in multi-relational social networks based on relational concept analysis. Proced. Comput. Sci. **159**, 291–300 (2019)
35. Raj, E.D.; Manogaran, G.; Srivastava, G.; Yulei, W.: Information granulation-based community detection for social networks. IEEE Trans. Comput. Soc. Syst. **8**(1), 122–133 (2021). https://doi.org/10.1109/TCSS.2019.2963247
36. Wan, X.; Zuo, X.; Song, F.: Solving dynamic overlapping community detection problem by a multiobjective evolutionary algorithm based on decomposition. Swarm Evolut. Comput. **54**, 100668 (2020)
37. Van Lierde, H.; Chow, T.W.S.; Chen, G.: Scalable spectral clustering for overlapping community detection in large-scale networks. IEEE Trans. Knowl. Data Eng. **32**(4), 754–767 (2020)
38. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**(23), 8577–8582 (2006)

39. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**(4), 452–473 (1977)
40. Lusseau, David: The emergent properties of a dolphin social network. Proc. R. Soc. London Ser. B Biol. Sci. **270**, 186–188 (2003)
41. Girvan, M.; Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**, 7821–7826 (2002)