



# Topological to deep learning era for identifying influencers in online social networks :a systematic review

Yasir Rashid<sup>1</sup> · Javaid Iqbal Bhat<sup>1</sup>

Received: 6 December 2022 / Revised: 20 April 2023 / Accepted: 2 June 2023 /  
Published online: 4 July 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

Influential user detection in social media networks involves identifying users who have a significant impact on the network's dynamics and can shape opinions and behaviours of other users. This paper reviews different topological and deep learning techniques for identifying influencers in online social networks. It examines various methods, such as degree centrality, closeness centrality, betweenness centrality, PageRank, and graph convolutional networks, and compares their strengths and limitations in terms of computational complexity, accuracy, and robustness. The paper aims to provide insights into the state-of-the-art techniques for identifying influencers in online social networks, and to highlight future research directions in this field. The findings of this review paper will be particularly valuable for researchers and practitioners interested in social network analysis.

**Keywords** Prominent user · Influential nodes · Online social networks · Deep learning · Graph convolution networks(GCNs) · Communities

## 1 Introduction

Social media is simply a platform for interaction and communication among users in which they generate, part, access and trade data and opinions thereby creating nets (groups) and networks (net of groups, individuals) (Fig. 1) [12]. In the below Fig. 2 a graph is drawn that depicts the statistics of world wide social media users [12].

The top most Social media platforms among all are Pinterest, Facebook, Reddit, Quora , twitter, Instagram, LinkedIn and YouTube. In the Fig. 3 below is drawn a graph that shows the top most social media platforms used worldwide [29].

When people get compound puzzling recommendations, then obviously they seizure towards Influential Users. In general, we can say influential users are those whose impacts

---

✉ Javaid Iqbal Bhat  
javaid.iqbal@iust.ac.in  
Yasir Rashid  
yasir.rashid@iust.ac.in

<sup>1</sup> Department of Computer Science, Islamic University of Science and Technology, Kashmir, India



Fig. 1 Common social media platforms

are stretched via the wide network. In network science, the problem of identifying prominent users is crucial for a vast range of applications like prediction of essential proteins, detection of financial risks, viral marketing, controlling rumour spreading, information diffusion, predicting career movements, social and political campaigns, control disease/pandemic spreading, community health campaigns, blocking the spread of viruses, diffusion analytics and so on [36].

Influence in the online world has become a crucial factor in determining the success of any digital marketing campaign. With the rapid growth of social media platforms, the number of users and the volume of generated data have increased significantly. In this scenario,

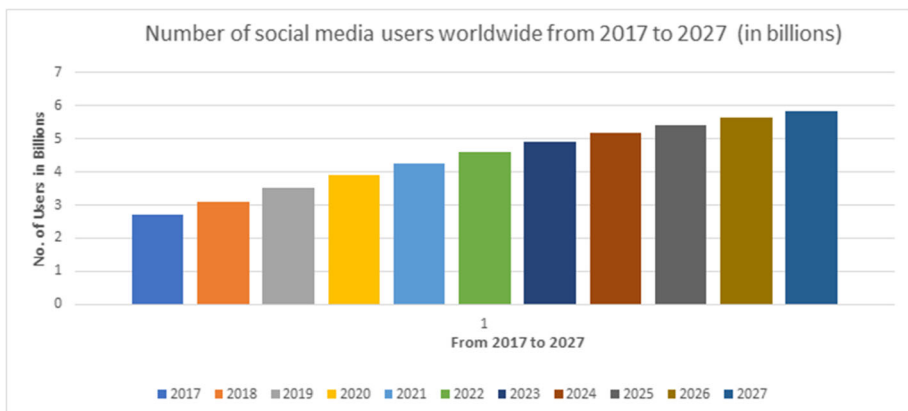
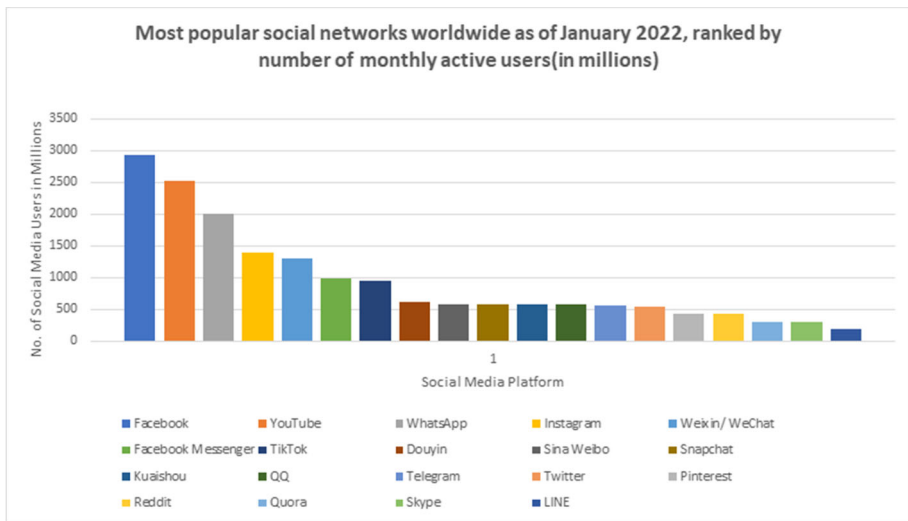


Fig. 2 Global social media users



**Fig. 3** Popular social media platforms

identifying the most influential users among the masses has become a challenging task. The influence of a user can be measured in various ways, including the number of followers, engagement rate, and content quality [29]. In this paper, we present a systematic survey of the existing literature on influential user identification in social media. Our survey covers various approaches and techniques used to identify influential users, including network analysis, machine learning, and content-based methods. This paper aims to provide a comprehensive overview of the state-of-the-art in the field and highlight the strengths and limitations of different methods. By presenting a systematic review of the literature, this paper aims to assist researchers and practitioners in choosing the most appropriate method for their specific needs[50].

This study comprehensively discusses all prevalent traditional methods to identify influential users in online social networks and list their shortcomings that were solved by deep learning based methods. Both structure-based and machine learning-based classical methods for identifying important nodes simply take network structure or node characteristics into consideration when determining the relevance of nodes. This work contributes by proposing an approach that incorporates both structural and behavioural features of online social networks for detecting influential users [11, 65]. This work also attempts to contribute by identifying influential users in online social networks with the focus on community detection (overlapping communities) using Deep learning (DL). The proposal of using Graph convolution Networks to identify overlapping communities and based on those overlapping communities influential users are identified [41].

As the number and size of social networks continue to grow, however, determining a user's social impact poses a number of difficulties. We think there are a number of obstacles in this scenario, and we outline some of them below [19, 26]. It is difficult to characterise the causal relationship of social influence and the uncertainty of social influence because, first, we lack a mathematical definition and measurement of social influence; second, it is difficult to decide the major factors for a specific case to social influence modelling; third, there is no effective ways to properly integrate various factors for influence measurement [19].

In order to find the relevant literature for the study, following Query Strings were searched in quality databases: “Influential user detection in online social networks” “Influential user detection using structural features” “Machine Learning techniques for influential user detection” “Overlapping community detection” “Deep Learning techniques for influential user identification”

The rest of the paper is organized as follows. In Section 2, we provide an introduction on influential node identification, and provide a survey of topological or structure based algorithms for influential user identification in Section 3. In Section 4, we discuss coreness based methods for influential user detection, and machine learning based algorithms for influential user identification are discussed in Section 5. In Section 6, an analytical comparison of various machine learning algorithms for influential user identification is provided, and we present existing earlier machine learning algorithms in Section 7. In Section 8, we discuss Graph convolution Networks and how they can be used for identifying influential users in overlapping communities, and in Section 9, we discuss different datasets. In Section 10 Open Challenges are discussed and in Section 11 Limitations of State-of-the-art are investigated. Finally in Section 12, we discuss conclusion and future scope.

## 2 Identification of impactful nodes in social networks

Initially nodes with plentiful friends and followers were considered to be influential nodes. Later, a variety of network structure and material produced by users-related characteristics, in addition to the amount of followers or friends, were employed to identify prominent nodes [65, 83].

## 3 Influential node determination algorithms in OSN’s

Various factors and techniques are considered for influence measurement. In present section, recent advances on influence measurement and revealing of influential handlers are discussed in context of Online Social Networks as shown in (Fig. 4) [61, 62].

### 3.1 Topology/Physical-structure based methods

In Topology/Physical Structure based methods influence is measured based on topology of social network, which is usually represented as graph. Centrality measures refer to the properties of graph which are used to define the influence pertaining to a node. The most important term used in social network analysis is centrality measure which is quantified for determining the identification of influential users and hence have been clustered into three major types viz: [2]

- Local
- Global
- Semi-Local

In local approach, the node itself and its neighbourhood are employed to determine a node’s effect. In the Online Social networks, context degree depicts the magnitude of an audience of a user, the sum total of social relationships or in other words total number of exchanges. Numerous research may examine degree centrality as a criterion for identifying significant

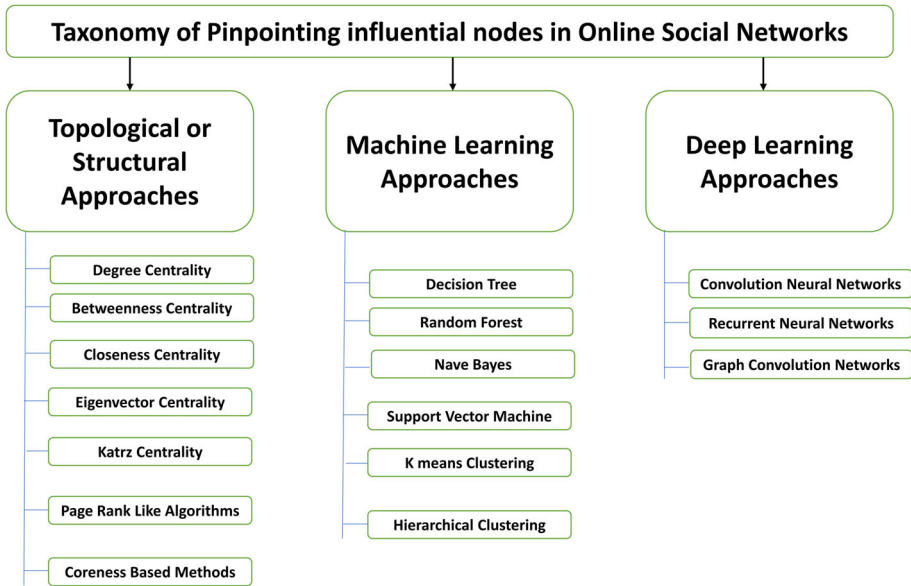


Fig. 4 Methods of influential user detection

nodes in online social networks. However, degree of centrality alone is insufficient to assess the effect of a person; people with a high degree of centrality are not deemed influential [48, 68]. The social network research communal labels a diversity of traditional boundaries for representation of the Influential node detection in a network alike Centrality measure which includes the following as shown in below Fig. 5:

### 3.1.1 Degree centrality

Degree centrality of any user/node in a graph like web may be in simple words referred to the count of edges that are incident with it, or in other terms as the final count of those nodes that are head-to-head to it. Degree centrality is one of the modest measures of node connectivity [54]. Hence, it is sometimes supportive to consider an in-degree (sum total of inbound links) and out-degree (sum total of outbound links) as discrete measures [39].

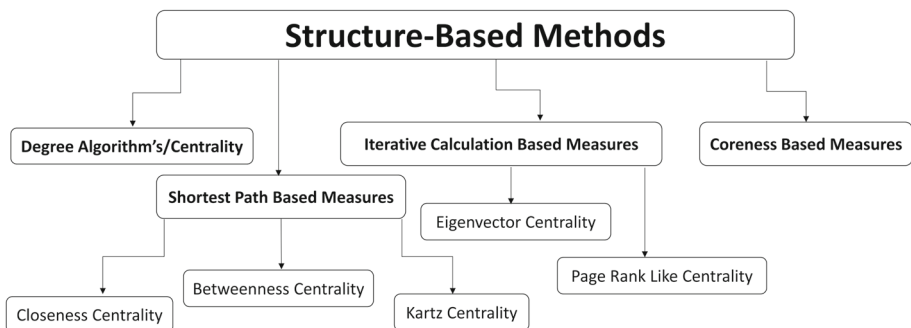


Fig. 5 Topological or Structural Methods for identification of influential users

## 3.2 Shortest Path-Based Measures

### 3.2.1 Closeness centrality

Closeness centrality helps us locate the nodes that are positioned to quickly have an influence on the network as a whole. It examines the node's proximity to other nodes in the network. Therefore closeness centrality defines scores for each and every node on the basis of their 'closeness' prior to rest of nodes in that network. We can use closeness centrality as an extent that recognizes how rapidly can information be sent from a single node to all other nodes, means this centrality assist us to locate better 'broadcasters' [54]. When you are inside of a network that is tightly linked, you will see that complete nodes have a parallel score more often. Because of this, it is more beneficial to utilize closeness centrality to discover influencers inside a single cluster. This section finds the shortest routes connecting every node and provides each node a score based only on the aggregate of all its shortest routes. It is determined by averaging the lengths of the shortest route from each system node to every other system node [93]. It classifies nodes as more central if they are closer to the majority of the graph's other nodes, which is one of proximity centrality's key advantages. When we create a network, a node with a high nearness centrality will typically appear toward the centre of the graph, which substantially influences visual centrality as shown in Fig. 6. The ideas of closeness centrality and farness centrality are meticulously related. The closeness may be defined in such a manner that if a vertex is near to every other vertex, then consequently the value is higher than if the vertex is not near to everything else [25]. Closeness of a vertex may be well-defined in (1)

$$C(v) = \sum_{u \in G} \frac{1}{d(u, v)} \quad (1)$$

The above equation demonstrates that the closeness becomes less when the sum of the distances is more and its vice versa is true. A vertex will be having close association with some other vertices if it has high closeness centrality.

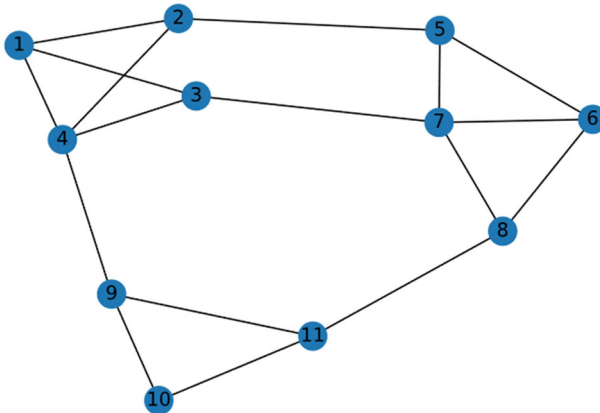


Fig. 6 Closeness centrality

### 3.2.2 Betweenness centrality(BC)

In order to identify the nodes that influence the circulation round a system, Betweenness centrality is worthwhile for analysis of the announcement of dynamics. The frequency with which a node serves as a bridge along the shortest route that is practically possible between any two nodes is denoted by the symbol BC. To calculate BC, first all of the shortest paths have to be located, and then the number of times that each node falls on one of those pathways has to be counted [122]. Betweenness is used to identify significant nodes in social networks online. BC is used as a metric that depicts a individual’s role in allowance of information to regulate. BC of a node can be calculated using (2) (Fig. 7).

$$g(u) = \sum_{s \neq u \neq z} \frac{\sigma_{sz}(u)}{\sigma_{sz}} \tag{2}$$

### 3.2.3 Katz centrality

Katz centrality takes into account all the network paths for determining influence of a node. Katz centrality is referred as measurement of centrality in a network. Established by Leo Katz in 1953, and for a node or actor surrounded by a social network, it is primarily used to determine that node’s relative degree of influence. By counting the nodes that are immediately next to a given node, Katz centrality may be used to assess a node’s relative influence and all other nodes in a social network that are connected to it through immediate neighbors. Katz centrality assesses influence by considering the entire number of links/walks between a pair of players, as opposed to other measures of centrality that simply examine the shortest route (the geodesic) between them [65, 81] (Fig. 8).

**Calculating Katz centrality** Katz centrality determines the relative importance of a node in a network by calculating the total number of immediate neighbours, also known as nodes of the first degree, and any additional nodes in the network that are typically connected to the additional node under consideration through these neighbours. [65] Therefore, connections established to distant nodes through neighbors are punished by an attenuation factor  $\beta$ . Every link, route, or connection between two nodes is given a weight  $\beta$  proportional to and

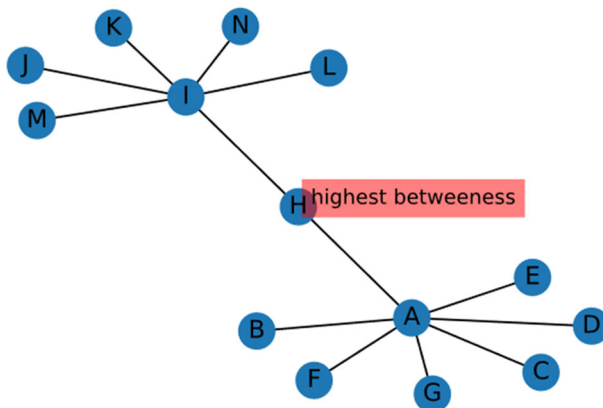


Fig.7 Betweenness Centrality

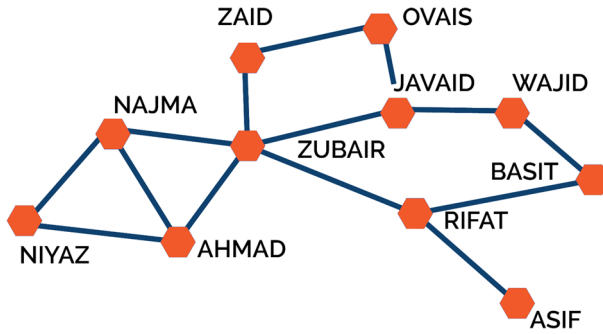


Fig. 8 Katz centrality

the distance between them as measured by  $\beta_d$  [61]. For instance, in the image above, imagine that Basit’s centrality is  $\beta = 0.5$  and is being computed. The weight of the connection linking Basit to his near neighbors Wajid and Rifat will be  $(0.5)^1 = 0.5$ . Since zubair is linked to Basit by indirect link through Rifat, the weight which will be given to this connection (comprises of dual connections) will be  $(0.5)^2 = 0.25$ . The weight that will be applied to the connection that goes via Zaid, Zubair, and Rifat will be  $(0.5)^4$ , which will equal 0.0625, whereas the weight that will be assigned to the connection that goes through Wajid and Javaid will be  $(0.5)^3$ , which will equal 0.125.

Mathematical formula

Consider  $Z$  to be a neighbouring matrix of the social network that is being considered in this instance. The values of the variables that make up  $Z$ ’s elements  $(a_{i,j})$  are either 1 or 0 depending on whether or not the node  $I$  and  $J$  are connected or not. The existence (or lack) of links between these two nodes is determined by the Powers of  $Z$  using their respective intermediates as the connection point. Nodes 2 and 9 in the  $Z_3$  matrix are connected to each other via the first- and second-degree neighbours of node 2, if it is believed that the element  $a_{(2,9)} = 1$  is correct [90]. In terms of statistics, if  $K_{katz}I$  is intended to reflect the centrality (Katz) of node  $I$ , then the following formulation is possible:

$$K_{katz}(i) = \sum_{d=1}^{\infty} \sum_{j=1}^{\infty} \beta^d (Z^d)_{ji} \tag{3}$$

It is essential to take note of the fact that the concept element located in the matrix at the position  $(i, j)$  adjacent to  $Z$  raised to the power  $d$  (abbreviated as  $Z_d$ ) displays the total degree links between nodes  $I$  and  $j$ . It has been decided that the magnitude of a attenuation factor  $\beta$  will be lower than the inverse of the matrix’s eigenvalue’s absolute value that is directly next to  $Z$  that has the highest value. For the sake of determining Katz’s centrality in this circumstance, the following sentence may be utilized:

$$\overrightarrow{K_{katz}} = ((I - \beta Z^T)^{-1} - \overrightarrow{I})I \tag{4}$$

Here  $\overrightarrow{I}$  show the identity matrix and identity vector of size  $m$  (where  $m$  is the number of nodes), both of which are composed of ones.  $Z^T$  represents  $Z$ ’s transposed matrix, whereas  $(I - Z^T)^{-1}$  represents  $(I - Z^T \text{ matrix})$ ’s inversion.



### 3.3 Reiterative Calculation-Based Metric

These methods do not only account directive links or paths mid-nodes but likewise considers all purposeful links or paths of network to compute the user or node stimulus. In a system based on continuous repeated measurements of computation, each node provides his ranking value towards its adjacent nodes and, acquires fresh latest notches from nodes linked in individually after each iteration. This whole cycle is repeated till every node ranges a firm state. Eigenvector centrality and PageRank algorithm are mostly employed for calculation of certain measures in online social networks [97, 119].

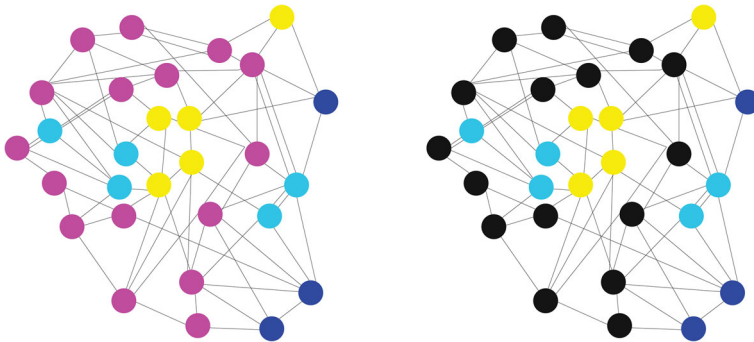
#### 3.3.1 Eigenvector centrality or eigen centrality

In addition to the amount of links that exist between nodes, the effect ratings of connected nodes are taken into account while calculating eigenvector centrality, which is more frequently known as Eigen centrality. Eigenvector Centrality of a node in a network describes the extent to which that node is influential in relation to the number of connections that it keeps open with various other nodes within the network. This is accomplished by giving points or scores to every other node within the network, on the basis that analogous connections to close to the bottom nodes do not contribute any more points or scores, but links to top scoring nodes do generate additional points or scores. One way to look at the relevance of a node is as having a direct proportional relationship to the sum of the impact scores of the other nodes in the network to which it is most closely connected [17]. Eigenvector Centrality is a step ahead of previous centrality metrics since it takes into consideration how well a node is linked, as well as how many linkages their connections have, and so on and so forth across the social network. By calculating Prolonged connections of a node not only those connections which are directly connected to that node, Eigenvector centrality locates nodes as per influence over the complete network. Eigenvector centrality calculates a node's position by taking into attention the value of its neighbour's (Fig. 9). Let us illustrate it with the help of an example, a node which has 100 very widespread friends like Amitabh Bachchan will high eigenvector centrality as compared to a node which have 100 relatively unpopular friends on social platform like Facebook. It is sometimes used to quantify the effect of a node in a network. Using the adjacency matrix, it is quantified by completing a matrix computation to obtain the primary eigenvector. The most important notion is that connections from crucial/vital nodes (as defined by degree centrality) are more important than linkages from less significant nodes. Nodes with high no of edges start gaining importance as the computation progresses but all the nodes start off equal. Their significance spreads to the connecting nodes. After computing again and again, the resulting values stabilize and hence concluding at providing values for eigenvector centrality [18].

A graph  $G(V, E)$  containing  $|u|$  vertices let  $Z = (a_{u,p})$  be the matrix adjacent, i.e.  $a_{(u,p)} = 1$  if vertex  $u$  is connected to other vertex  $p$ , and  $a_{u,p}=0$  otherwise. The following (5) is one possible definition for the relative centrality score of vertex  $u$ :

$$Y_u = \frac{1}{\delta} \sum_{P \in N(u)} y_P = \frac{1}{\delta} \sum_{P \in G} a_{u,p} y_P \quad (5)$$

Whereas the variable  $N(u)$  in the aforementioned equation represents the number of neighbors of  $u$ , and  $\delta$  is a constant value.



**Fig. 9** Degree centrality and Eigenvector centrality

### 3.3.2 PageRank algorithm

Famous algorithm ‘PageRank’ first time invented by Sergey Brin and Larry Page is broadly used for Ranking of web pages. It is named after Larry Page, one of Google’s renowned founders. Irrespective of the content, PageRank set of rules is used for international ranking of internet pages constructed completely at the related edges on a web Graph. PageRank algorithm is used to measures the transitive or directional have an effect on of nodes. Rest of the centrality algorithms mentioned thus far measure the direct have an effect on of a node, while PageRank set of rules in particular offers with the the impact of your neighbour’s and their neighbour’s [28]. If we have a rare influential friend, it’ll raise your PageRank more than just having a variety of low-have an impact on networks. PageRank’s Rules, which are dependent on a node’s main measurements and related link counts, and PageRank scores of all linked hyperlinks, are used in an iterative process to create a node’s value. This value is determined by the node’s PageRank score. Either by randomly traversing the network and keeping track of the number of times each node is struck throughout the course of walks, or by repeatedly spreading a node’s rank among its neighbors, RageRank may be calculated in one of two ways [87]. PageRank algorithm can be employed in diverse systems like it is used to understand vital influential nodes in plentiful practical packages. In order to get a rough assessment of the relevance of a node, The PageRank method employs a computation that takes both the quantity and quality of a node’s connections into consideration. There is much greater probability that the important nodes will get more links from other nodes is the underlying assumption here. As a result of the PageRank algorithm, a person who arbitrarily hits on links will appear at any specific node. PageRank may be measured for collections of social networks of any length. The PageRank computations entail several passes, termed “iterations”, through the gathering to adjust approximately PageRank values so that it will similarly carefully mirror the theoretical actual fee [20].

### 3.3.3 Simplified page rank algorithm

Let us understand this by assuming a small social network of four nodes namely A, B, C, and D. Self-loop links of a node , or in other words many outbound links of one single node to another single node are unnoticed. Initially PageRank is initiated to the equivalent value for all the nodes. Sum of PageRank across all pages was the entire number of web pages in the original version of PageRank, so each node in this example has an opening value of 1.

However Probability distribution concept was used in later versions of PageRank, so value ranges between 0 and 1. Due to this in later versions of PageRank value is 0.33. Upon the next iteration PageRank transferred from a given node to the targets of its outbound links has portioned in equal proportions among all outbound links. PageRank algorithm typically thinks that if several other nodes connect to a certain node, that node must be significant [108]. So, a node that is linked to more would be considered more important than a node that is linked to less. But this approach has a problem that it is easy for someone to artificially inflate importance of itself by creating fake followers or friends that are linked to him. To make myself as vital as I want, I will generate that number of fake followers or friends. In order to overcome this problem and to really define who is important or influential above definition of important or influential node is modified. A node becomes more important linked prior to other important nodes. This definition of significant node seems to be rounded; to compute the significance of a node, we must first determine the significance of other nodes. One way to calculate this leads to Random Surfer Model, in this model imagine someone browsing on web choosing some random node and then they haphazardly choice a link from that node to another node they visit [108]. The Random Surfer keeps recapping this process. The main idea behind this is to preserve score or tally how many times a random surfer visits each node every time. Each time that land on new node we update score of that node. Nodes that have more links to them are more likely to be visited more, leading to have high scores eventually. And because those nodes are more likely to visited, the nodes they are linked to are also likely to be visited more. Here we can say that the link from a more important node will be prioritized than a less important node [20]. After repeating this process several times for a while, we can take a look at resulting scores and thus calculate what percent of total score each node contains. This gives us significant importance of these nodes. There is still one problem with this approach that it is the fact that nodes on the social network might not all be connected to each other. If we are currently visiting one group of nodes, we might not be able to visit another part this social network as they are not reachable. To overcome this problem, we need to rarely reset our Random web surfer, we do this by introducing damping factor. For example, if damping factor is 0.85, it implies that 85% times our Random surfer will visit following the link from the node that currently they are as were doing before and 15% times it will shift to a node chosen completely at Random. Within sometime it will ensure that it explored all parts of the social network and will not get stuck in one particular set of nodes. In this model, now let us assume, take any parts of the social network and then calculate their relative importance of those pages [20]. In the first few steps: the Random Surfer takes the numbers are not particularly accurate, a lot is based on random chance but with the passage of time the Random Surfer continues to explore more and more and the numbers will converge eventually in order to have stable PageRank value for each node and those values can then be employed for regulation of the important or influential nodes. PageRank algorithm passes are illustrated in the below Fig. 10

The PageRank algorithm is expressed using following expression (6) :

$$PR(z) = \frac{1-d}{M} + d \sum_{t \in N(z)} \frac{PR(t)}{Q(t)} \quad (6)$$

M is the number of users inside the network, the sum total of out-degree linkages is Q(t) from user t, N(z) is the Online Social Networks user linking user z, and d is the damping factor.

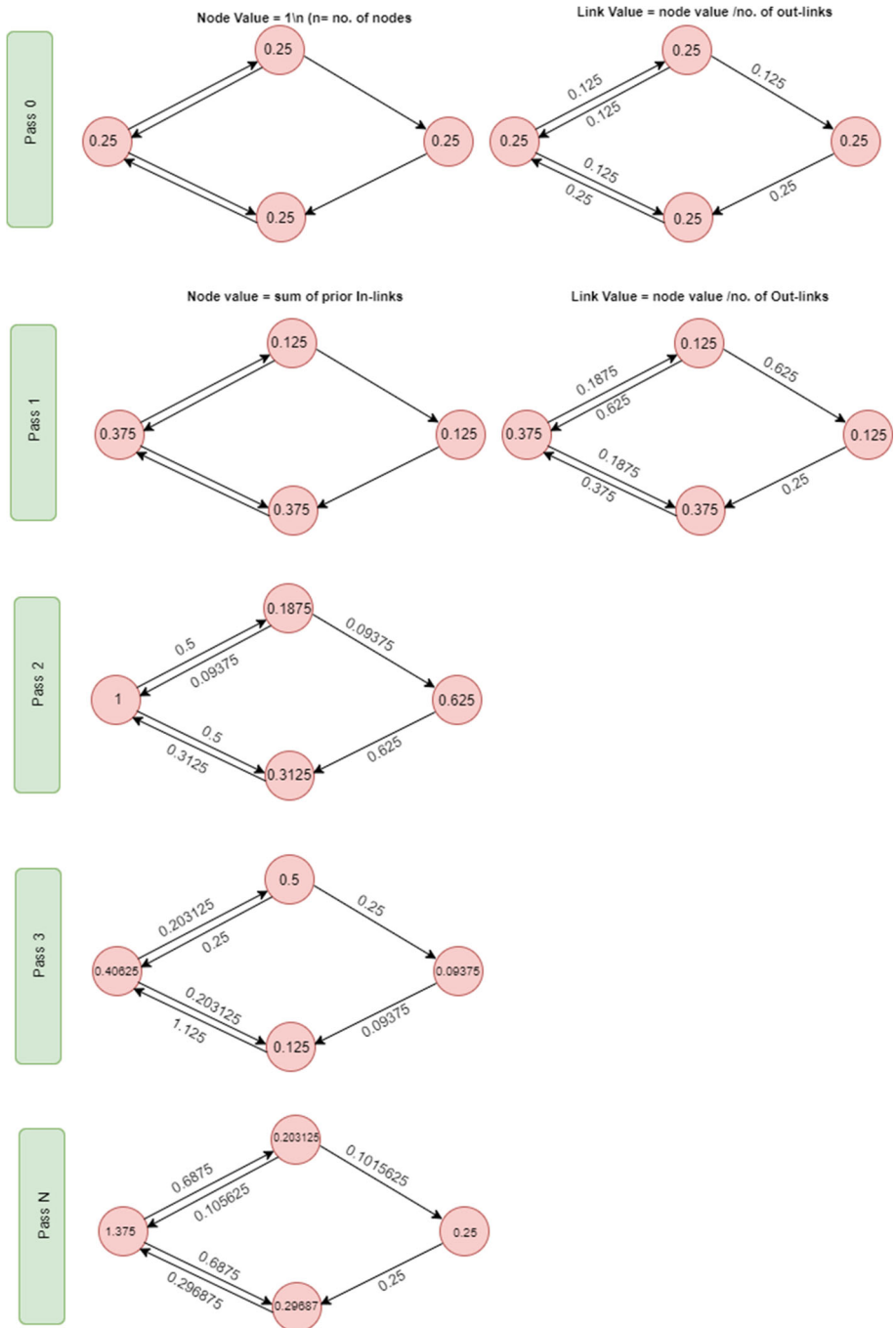


Fig. 10 Passes in pagerank algorithm

### 3.3.4 TunkRank

[66]: At now, Twitter is among the most widely used social networking platforms. TunkRank is an extension of PageRank. In light of the fact that Google takes Twitter's social signals into consideration, one of the things to keep an eye on is the measurement of Twitter impact. It is thus used for measurement of the influence on Twitter. Two basic ideas behind TunkRank for their influence metric are given below as follows:

1. The amount of attention a node can give is spread out among all those nodes a node follows. The more a node follows, the less attention a node can give each node.
2. A node's Influence is proportional to the amount of attention its followers may provide it.

TunkRank score may be measured with the aid of how much quantity of attention followers of a node can both immediately supply a node and what kind of quantity of attention they bring for a node from their community of fans. A feed's influence relies upon now not simplest the quantity of followers it has, but on the amount of attention all of its fans can give. The amount of attention a follower can give is unfold out amongst all of the feeds they comply with. Therefore, it depicts that the greater feeds someone follows, the much less attention they can provide one after the other to one another. TunkRank assumes that the have an effect on of a node is frivolously allotted among its network of buddies. Probability of a node studying any of its follower's tweet is same to one out of the whole wide variety of nodes is in its follower list. This concept can be mathematically represented as given below in following (7):

$$Influence(x) = \sum_{y \in followers(x)} \frac{1 + P \times Influence(y)}{||following(y)||} \quad (7)$$

Where P here in above equation indicates the probability that a node(x) is retweeting node (y)'s tweets, and  $|following(y)|$  represents in the same equation the sum total of nodes followed by node y.

### 3.3.5 TURank

[3]: TURank(Twitter user Rank) is majority used in studying the relationship among nodes and their posts. It also models the flow of information and then makes the calculations regarding the ranking scores of the nodes. TURank (Twitter User Rank) also does the evaluation of nodes score authority in Twitter which is entirely based on analysis of links as it takes the information flow in the twitter into consideration. In TURank, a user-tweet graph is used to represent users and tweets which mock-ups the flow of information, and ObjectRank is employed to calculate scores of user's authorities. The proposed algorithm outperformed over the existing algorithms. Some of the important observations to be considered were:

1. A Prominent node is tailed/ followed by many other prominent nodes.
2. A tweet is considered a valuable tweet if it is retweeted by many influential nodes.
3. A node is more like to be an influential node if it posts many valuable tweets.

Twitter Social graph which assumes best comply with dating for link evaluation had many negatives as when everybody on twitter follows someone, it observed lower back simply as formal courtesy and just a few percentage speak in comply with dating. Therefore, this current set of rules isn't sufficient to evaluate influential human beings. The TURank algorithm considers both the Twitter social network and the actual diffusion of tweets among

customers. To address this trouble TURank used retweet RT as most effective the ones tweets are retweeted extra which contains useful statistics which depicts consumer is influential [107]. To update this User-Tweet Employed is a graph consisting of nodes, like as user profiles and tweets, and edges, such as follow and retweet connections. A person-tweet graph is dynamic since it is rebuilt each time a new retweet is discovered. By thinking about handiest range of retweets is not enough for depicting whether a node is influential or not as retweets are of different categories some are conversational retweets which might be very less and a few are records retweets which might be very big in variety. Considering sorts of retweets, a tweet may have a propensity to be broadly distributed if the purpose of the RT is to flow into relevant data, if the purpose is to communicate with a limited number of clients, it may not be extensively distributed.. So, it isn't wise to rank users based totally on the range of RT's as conversational RTs spreads handiest for members worried inside the communication and recommending RTs spread widely [107]. Tweets are specifically retweeted via influential customers are likely to be extra beneficial than those who retweeted by using non-influential customers. To solve above troubles, TURank set of rules takes into attention the link structure of the tweet go with the flow with the aid of RTs. ObjectRank is a prolonged model of PageRank because it accomplishes analysis of link determine to measure the ranking of objects within the database. To accommodate many types of nodes and edges, it takes edge types and node types into consideration. It creates the authority transfer schema graph, which shapes the discourse domain. In the graph Link form evaluation, the rating is conveyed together with the outgoing links for each node, based on the edge weights specified by the authorization transition schema graph, the total propagated rankings are derived via a weighted average [3]. To compare the final score, below given (8) may be used:

$$O_r = PZO_r + \frac{(1 - P)}{|v|}e \quad (8)$$

where  $O_r$  represents the ObjectRank rank vector,  $P$  is the chance of an arbitrary hop, and  $Z$  represents the test's transition matrix. Note that the item  $a_{ij}$  in the transformation matrix  $Z$  represents the weight of the edge from node  $i$  to node  $j$  if the edge exists; otherwise,  $a_{ij} = 0$ .  $UG = (K, E)$  represents the authorization transference data graph in ObjectRank, which was constructed from the user-tweet schema graph. Thus,  $K$  is the collection of nodes including the overwhelming majority of tweet and user nodes in the data collection, and  $E$  is the set of links containing all existing edges in the collected data, such as post, posted, follow, followed, RT, and  $RT_{ed}$ . The value of the weight  $T(e)$  that was placed on the edge  $e \in E$  by the node  $v \in K$ , which is computed using the (9) shown below.

$$T(e) = \frac{T(e_s)}{OD(v, e_s)} \quad (9)$$

where  $e_s \in S$  represents an links of the identical type as  $e$  and  $OD(v, e_s)$  represents the number of edges of type  $e_s$  that depart node  $v$ .

### 3.3.6 TwitterRank

[106]: On twitter to rank user influence, TwitterRank algorithm make use of concepts like homophily in addition to network topology. An estimate of 72.5% of the users on twitter tailed more than 81% of their followers and 81% of the users have 80.5% of the users they are following them back. TwitterRank algorithm largely takes into account concept of both topical similarity and link structure as it is an extension of PageRank algorithm. The framework of this approach is explained in the below diagram as given below Fig. 11:

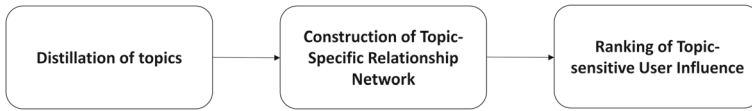


Fig. 11 Flow of twitterrank algorithm

Using LDA (Latent Dirichlet allocation) which is primarily based on the contents of user’s topic, TwitterRank is calculated. Therefore, topic distribution is employed to construct subject similarity ratings between each user pair. TwitterRank on a topic eventually measures user influence iteratively as follows given in below (10):

$$\vec{PK}_t = \mu b_t \times PK_t + (1 - \mu)M_t \tag{10}$$

The likelihood that a user will discover another user is given by the teleportation vector  $M_t$ , where  $PK_t$  represents the impact on subject  $t$  and the edges of the user-to-user connection network are disregarded. The damping factor  $\mu$ , whose value ranges from 0 to 1, controls teleportation probability, whereas transition probability is represented by  $b_t$ . In addition, the chances of transition probability  $b_t$  in TwitterRank is explained and evaluated as shown in the (11) below:

$$b_t^{(I,k)} = \frac{T.tweetsk}{\sum_{f:f} T.tweetsf} \times t^{(i,k)} \tag{11}$$

where  $T.tweetsk$  represents the total number of tweets by user  $k$ ,  $\sum_{f:f}$  specifies the number of tweets by all of user  $I$ ’s friends, and  $t^{(I,K)}$  represents the topic-related similarity between users  $I$  and  $k$ .

### 3.3.7 LeaderRank

[116]:LeaderRank is among the variants of recent proposed position algorithms for defining the influential nodes. Identification of influential nodes is vital for controlling and proper understanding of spreading process. Identifying influential spreaders effectively and competently is however a challenging task. A number centrality-based measures have been proposed till now but they are not as feasible as they are time consuming and their time and space complexities increases along with the network size. The LeaderRank algorithm, a variation of PageRank, is used to identify important spreaders in directed social networks. The idea of Ground Node has been established by the LeaderRank algorithm. Random walk is used to find important spreaders, and each node is linked to the Ground node through a bi-directional connection. Except for the Ground Node, one unit of prestige is given to each node in the ranking process, and this prestige is dispersed evenly to neighboring nodes through directed linkages. This procedure continues until a stable state is achieved. LeaderRank outperformed over PageRank algorithms as it many have advantages over PageRank algorithm.

- Since network is strongly connected, hence the convergence rate is higher in leaderRank prior to page rank algorithm.
- In comparison to PageRank, influential node pinpointing by LeaderRank is faster and efficient.

- Noisy data LeaderRank has higher tolerance as compared to PageRank.
- In comparison to PageRank, LeaderRank is more prone towards robustness of spammers.

$$Z_i(X + 1) = \sum_{u=1}^{N+1} \frac{a(u, v)}{k_u^{out}} Z_u(X + 1) \tag{12}$$

Here value of (N+1)-dimensional adjacency matrix is equal to 1, if there is a direct link between nodes u and v, otherwise it is 0. The elements of this (N+1) dimensional adjacency matrix is represented by  $a_{uv}$ .  $k_u^{out}$  represent outdegree of the node u. Here, the procedure starts by initializing from a point where all nodes have a score of 1 and will quickly converge to a unique steady state, which is primarily defined as  $S_v^\infty = (v=1, 2, N, N+1)$ .  $S_v^\infty$  in LeaderRank is to for positioning of all the nodes, and those nodes which are having high final score are hence more influential. Weighted LeaderRank algorithm is an enhanced version of LeaderRank algorithm as it employed the concept of Biased Random Walk. In Biased Random walk the initial node is more biased towards the nodes having more fans and consequently, a node with high number of fans will get a higher score from the initial node as shown in the Fig. 12.

Weighted LeaderRank is more resistant to spammers and tolerant of noisy data than the original LeaderRank. Consider a Social Network of of nodes, when there is a directed relationship from u to v showing that u would get knowledge from v, then we can clearly depict that u is a fan of v and, therefore, node v would receive scores from u. If node v’s admirers are very influential, then node v itself is likewise highly influential. Consequently, the in-degree parameter is an essential indication of influence spreading. Based on their LeaderRank, nodes with differing in-degrees get various points from the node that initiated the link. (N+1)-dimensional weighted adjacency matrix Z characterizes the Social Network:

- if  $a_{u,v} = 0$  then  $w_{u,v} = 0$ .
- In case of any normal Node u and in case of ground node g,  $w_{g,u} = (k_u^{in})^\infty$  and  $w_{u,g} = 1$ , free parameter is represented by  $\infty$ .
- In all other cases  $w_{u,v} = 1$

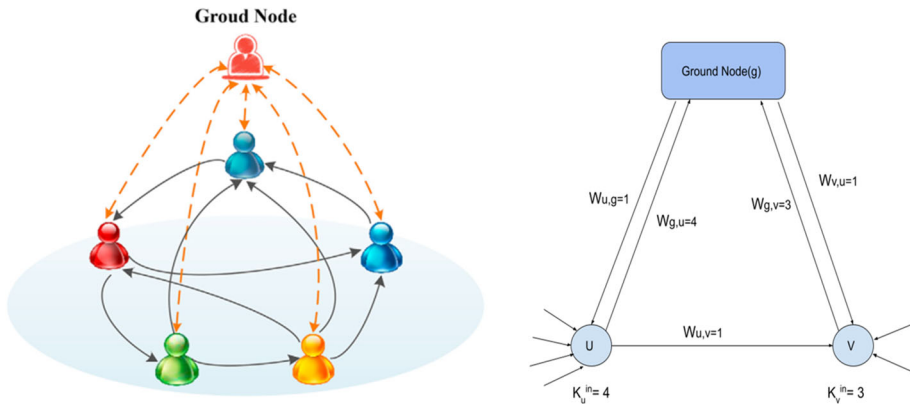


Fig. 12 Leader rank algorithm using ground node



The score from  $v$  to  $u$  is proportional to the weight  $w_{vu}$  after weight of each and every link is determined a biased random is being followed

$$Z_i(X + 1) = \sum_{v=1}^{N+1} \frac{w(v, u)}{\sum_{m=1}^{N+1} w_{vm}} Z_v(X) \tag{13}$$

### 3.3.8 InfluenceRank

[22]: Two models are being used in InfluenceRank technique. Comparative influence in case of InfluenceRank algorithm is being calculated using first model and to determine network-wide effect of node second model is used. Based on three factors as mentioned below:

1. Tweet quality.
2. Ratio of tweets
3. Homophily or Topic similarity.

The nodes relative influence model is calculated. This model is mathematically represented in terms of below given (14):

$$RI(V_k, V_l) = Z_{(vk)} + P(V_k, V_l) + sm(V_k, V_l) \tag{14}$$

Relative influence  $RI(v_k, v_l)$  is defined as the influence on user  $v_l$  by used  $v_k$  and  $Z_{vk}$  is tweet quality and it is measured as below (15):

$$Z_{(vk)} = \frac{RT_{(vk)} + \text{Number of comments}(vk)}{\text{Total tweets}(vk)} \tag{15}$$

Where  $RT(v_k)$  represents the number of retweets. The Ratio  $P(v_k, v_l)$  can be calculated with the below (16):

$$P(V_k V_l) = \frac{RT(V_k, v_l)}{RT(vk)} \tag{16}$$

$sm(v_k, v_l)$  is homophily or topic similarity between users  $v_k$  and  $v_l$ . The influence is calculated mathematically as below (17):

$$\text{Influence}(k) = (1 - \delta) \frac{|\text{Followers}(k)|}{M} + \delta \times \sum_{v1 \in \text{Followers}(v_k)} \frac{RI(V_k, V1) \times \text{influence}(v1)}{\text{following}(v1)} \tag{17}$$

Where  $\delta$  is the damping factor. InfluenceRank varies from the PageRank algorithm in that it employs a biased random walk and does not distribute score value evenly across followers.

### 3.3.9 InfRank

[46] : InfRank is used to identify influential nodes , it is like PageRank algorithm. Nodes having high number of retweets means they have very high information spreading ability which is used to measure the influence and it also helps in identifying influential nodes in retweet list. Edges that connect different users represent tweets and nodes are used to represent users in case of InfRank algorithm. If atleast one of the tweets of  $v_k$  is retweeted by  $v_k$ , an edge is made between users  $v_k$  and  $v_l$ . User-tweet graph has pros and cons both compared with user-follower graph as user-tweet graph depicts a much stronger social connection because user on twitter can follow another user on twitter despite not retweeting his tweets and can also retweet despite not following the original owner of that tweet. User-follower graph is

dense relatively as compared to user-tweet graph. In terms of weights InfRank distributes Ranking scores. For calculation of weighed edge below (18) is used:

$$K(V_k, V_l) = \frac{|Total\ tweets\ by\ v_l \cap Total\ tweets\ by\ v_k|}{|Total\ tweets\ by\ v_k|} \quad (18)$$

### 3.3.10 SpreadRank

[28]: SpreadRank develops user-tweet network similar to that used in InfRank. SpreadRank is a variant of PageRank. The unique retweet weight is used to weight the edges of the user-retweet network in SpreadRank, and If a division operation is performed in which total number of Retweets is divided by number of tweets, edge weight will be evaluated. In SpreadRank, the time gap between retweets is crucial because the quicker tweets are repeated, the greater the diffusion rate. In SpreadRank, the position of nodes is used to determine the teleport vector, thus if a location is closer to the root node, i.e. closer to the primary tweets, it will get more points. The change of influence from users  $v_k$  to  $v_l$  is assessed using the following (19):

$$P(V_l, V_k) = \frac{\sum_{rkl} f(t_{kl})}{\log \log N_L} \quad (19)$$

Total number of Tweets that are done by user  $V_j$  is represented by  $N_j$  and  $(t_{kl})$  is used to represent time of Retweets.

### 3.3.11 ProfileRank

[95]: ProfileRank is an offshoot of PageRank since it was inspired by it. ProfileRank is a representing the major of user impact and content suitability in information dissemination. ProfileRank recognizes notable persons by measuring a node's capacity to develop and disseminate relevant information for a substantial portion of the network. This approach produces random walks on a bipartite network of user-generated material.

## 4 Coreness-based methods

Location of a node in k-core decomposing or coreness-based methods is measured to be more decisive than its direction in calculating its diffusion of influence. The K-core algorithm is based on network decomposition. The decomposition process clips the network into K-cores. Those Nodes are continuously deleted whose degree value is evaluated to be less than 1 in case of K-core decomposition. In case of the Original Graph Network most connected subgraph is represented by K-core. This decomposition procedure begins at the node with the degree 1 [15]. Those nodes who have degree=1 is assigned to 1-shell. Initially, all nodes with degree  $k=1$  are severed, and the clipping method is continued until there are no more nodes of degree  $k$ . This procedure is performed on the remaining k-shells until we reach the network's cores. Finding important nodes in complicated networks requires K-shell techniques. Influential network users demonstrate that highly linked nodes relate to one another. Many authors claimed to have proposed improved variant of K-core algorithms [14]. Mixed degree takes into account only residual links and ignores the exhausted degree. Residual links means the remaining links and ignoring the exhausted degree means that those links are not considered which are connected to removed nodes [115]. Many other studies [77, 104] consider weighted

PageRank algorithms	Methodology	Objective	Input parameters	Network type	Weight
TunkRank	As a metric of impact, the number of people who will read a retweet is considered.	Retweet and attention probability are taken into consideration in recursive manner for measuring influence A node that is followed by important nodes is likely to be an influential node that is measured for influence.	Probability that the user will retweet and no. of followers	User-follower Network	Edge weight is the continuous possibility that a person would rebroadcast a tweet.
TURank	Connecting people, their tweets, and their followers with acceptable edges forms a network.		Followers, Tweets, and retweet counts.	User-tweet network	
TwitterRank	In estimating Node impact, both Network topology and topic similarity are taken into account.	Topic similarity between node and network topology is taken into consideration for calculating Node influence.	number of followers and subject matter	User-follower Network	The weight of an edge is determined by the topic similarity of nodes.
LeaderRank	The Ground Node idea is proposed, which connects each network node through bidirectional linkages.	For LeaderRank to suggest algorithms that can accurately measure user influence, it has to be more tolerant to noisy data and resistant to spammers.	Number of Ground Nodes and fans by taking into consideration bidirectional links.	Fan-leader-ground network	Weight is disregarded.
Weighted Leader Rank	TwitterRank determines user influence by taking into account both topic similarity and network structure.	Recursively measuring user influence is thought to be possible through network structure and user topic similarity.	Comparable subjects and follower count	User-follower network	The value of the edge relies on how closely the subjects of the users align.

PageRank algorithms	Methodology	Objective	Input parameters	Network type	Weight
InfluenceRank	The algorithm functions in two stages: in the first stage, the relative impact of a user is evaluated, and in the second stage, the total effect of the user network is determined. Before determining the most influential	Recursively measuring users' influence involves taking into account each user's relative influence.	Follower count, follower calibre, tweet calibre, retweet ratio, and subject similarity	Follower network	The relative impact of a user is used to determine the edge's weight.
InfRank	individuals on the retweet list, the research evaluates each user's potential for information diffusion inside the network. With users as nodes	To find leaders, discussers, and influencers in online social network microblogs, InfRank assesses user influence.	Total amount of retweets and the number of prominent persons on the retweet list.	Retweet network	The retweet ratio weighs the edge.
SpreadRank	a network is build by SpreadRank, and when user $v_i$ retweets at least one tweet from user $v_j$ , an edge is created between them. This algorithm	For recursive influence measurement, both the weights and the frequency of retweets are considered.	number of retweets and the spacing between them.	Retweet network	The retweet ratio weighs the edge.
ProfileRank	assesses users' capacity to produce and spread pertinent material to a sizable section of the community.	Recursively measuring user influence involves taking into account both user impact and content relevancy.	Users and content	User-content bipartite directed network	Edges are weighted according to the unique retweet weight, or the ratio of retweets to tweets. On a bipartite user-content network, random walks occur.

complex network. The weight of a node with a high degree of connectivity is given a high value by these approaches [4]. Consider the interaction between nodes as link weights to enhance the K-core.. In this algorithm communication is considered to be an important element for sprinkling information. It simply examines a single network to describe the link between nodes whose weights are determined by their interactions. But reality is that nodes can interact even if they are not connected on social networks directly but they can communicate through common or mutual nodes. Therefore, a multilayer network is required to depict the relationships between nodes [125]. Non-availability Social Network data decides Online Social Networks from many complex networks as we are not able to measure efficiencies of influence calculating techniques for measuring user influence in determining influential nodes in Online Social Networks applied different measurement techniques and noticed that k-core all other PageRank and degree algorithms [87].

## 5 Machine learning algorithms for influential user detection

### 5.1 Machine algorithms for influential user detection based on different dimensions

Machine learning algorithms for influential user detection can be classified along different dimensions, including:

#### 5.1.1 Supervised vs unsupervised

Supervised machine learning algorithms require labeled data to train the model, whereas unsupervised machine learning algorithms do not require labeled data. In the context of influential user detection, supervised algorithms can be used to train a model on labeled data, where the labels indicate the level of influence of each user, and then use the trained model to predict the influence of new users. Unsupervised algorithms, on the other hand, can be used to identify patterns and relationships in the data without the need for labeled data.

#### 5.1.2 Graph-based vs non-graph based

Graph-based machine learning algorithms take into account the graph structure of the data, i.e. the relationships between users in a social network. These algorithms can be used to identify the most influential users based on their connections with other users in the network. Non-graph based machine learning algorithms do not consider the graph structure of the data and can be used to identify influential users based on features such as the number of followers, the frequency of posts, and the engagement of followers.

#### 5.1.3 Regression vs classification

Regression algorithms are used to predict a continuous value, such as the level of influence of a user. Classification algorithms, on the other hand, are used to predict a categorical label, such as whether a user is influential or not.

#### 5.1.4 Traditional machine learning vs deep learning

Traditional machine learning algorithms, such as decision trees, random forests, and support vector machines, have been widely used in the task of influential user detection. Deep learning

algorithms, such as neural networks and convolutional neural networks, have become increasingly popular in recent years and have shown promising results in the task of influential user detection.

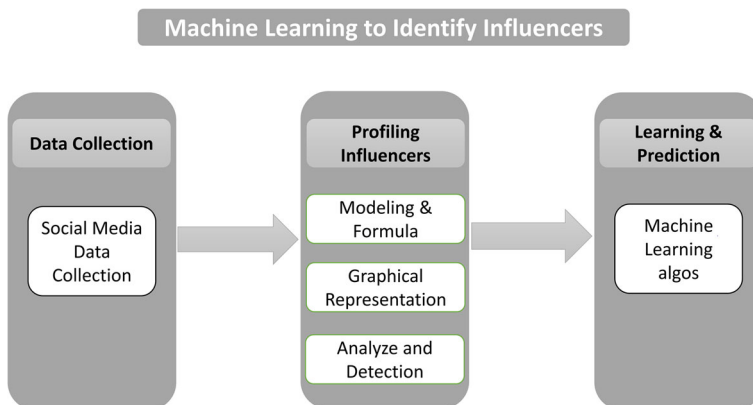
This is a high-level classification of machine learning algorithms for influential user detection based on different dimensions. The choice of algorithm will depend on the specific requirements and the type of data available for analysis.

Machine Learning algorithms is based on node features as shown in Fig. 13 and in these algorithms concept of feature engineering is used.

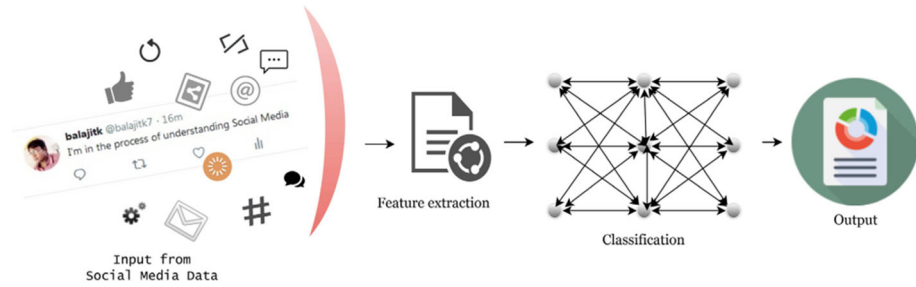
In these machine learning approaches we have to select best features for better performance. These algorithms are based on features that are manually selected [63]. The most prevalent method for predicting prominent people using machine-learning algorithms is supervised learning [42]. Common supervised learning approaches include Nave Bayes, support vector machine (SVM), and decision tree. An effective learning strategy need a strong collection of characteristics with significant discriminative strength to improve the ability to anticipate outcomes. [21, 27] In order to train and evaluate a machine-learning model, learning methodologies need adequate data sets. Instead of proposing particular techniques to enhance the general prediction model, the bulk of research on predicting prominent nodes have recommended key qualities as shown in below Fig. 14 [16, 67].

Predicting influential users in OSNs causes intense disputes about the selection of predictive criteria. A framework based on infection rate and SIR model for influential user detection is illustrated in the below (Fig. 15)

The amount of followers, retweets, and tags are used to predict influential individuals on Twitter. Nonetheless, [67, 93] show that, in addition to direct criteria, additional useful characteristics, like the number of public lists, Quantity of the tweets that are done recently, and the ratio followers vs friends, may be used to forecast powerful persons. In another work [91], many characteristics are retrieved for SVM training. Features use three distinct aggregation methods: score-based, list-based, and SVM-based. Combining user location inside a network, user viewpoint direction, and tweet excellence yields the impact score [16]. Additionally, a logistic regression analysis was used to identify significant drivers of user effect. Four machine-learning methods were trained and evaluated using these characteristics. In [16] the ACQR framework was suggested. This method extracts a set of indicators deemed to be differentiating characteristics for identifying successful consumers in Online environ-

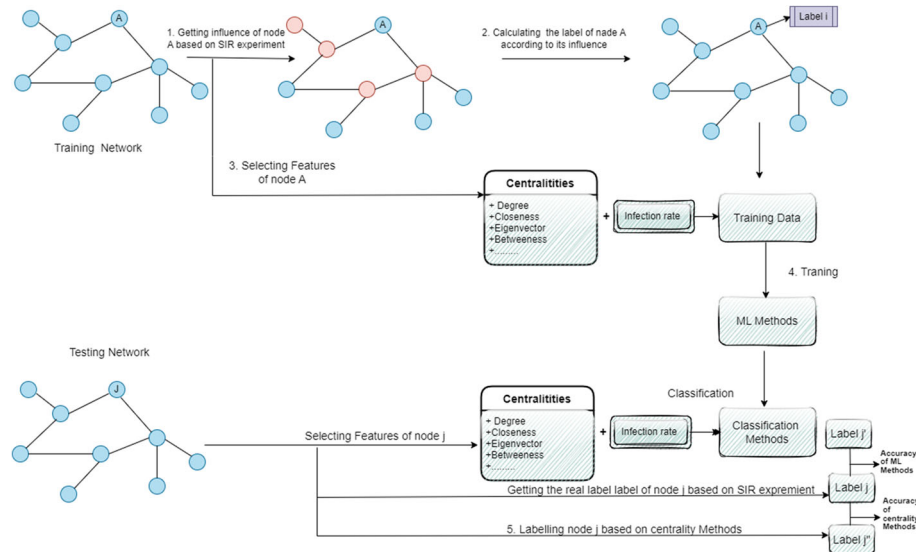


**Fig. 13** Workflow process approach



**Fig. 14** Machine learning for data processing in social media

ments: activity, centrality, post excellence, and prestige. The SVM is then trained using the four characteristics [74]. To identify influential users in OSNs, [27] examine a large variety of traditional criteria, such as user behavior, regional topography, creative features, tweet characteristics, and string of words occurrence weighting. The conclusion of this investigation is that conventional characteristics provide negligible effects. The authors offer a collection of novel performance-enhancing enhancements. However, this study cannot be generalized due to the absence of thorough conventional characteristics from prior research. In addition, this research uses a particular dataset, with conclusions applicable only to that dataset, in order to identify prominent users in OSNs utilizing the numerous variables used to train the learning models. Moreover, following work has also been done towards the detection of influential users by machine learning approaches as follows. For discovering influential nodes, a single framework, such as Latent Interaction Graph Extraction (LIGE), Degree Centrality based Influential Inference (DCII), and Formation Stability Coefficient mostly dependent Re-ranking, is suggested (FSCR). Neural Relationship Inference makes use of LIGE to reconstruct the latent interaction network using the encoder and vibrant track data.



**Fig. 15** Framework for identifying influential users in complex networks based on machine learning

DCII measures the relevance of nodes using the conventional graph model. The use of the formation balancing coefficient, which is part of the re-ranking module known as FSCR, is used to evaluate the stability of the swarm formation and to re-rank the candidates. Our method successfully determines which nodes within a drone swarm are the most influential, as shown by the results of experiments performed on datasets including drone swarms. Because of the very lengthy timeframes required for LIGE extraction, it might be challenging to identify the core node within the allotted time. By adjusting the encoder shape, it is possible to reduce the extraction time to at least twice the unique value[8]. Identifying significant nodes in complex networks is of critical importance for their extended packages. Although centrality measures have traditionally been derived from the topological systems of networks, novel centrality approaches keep in mind other, structural elements that are of practical importance. However, in many cases, the relevance of a node and its several functions-such as local area, global region, etc.-are intertwined in a complicated and nonlinear way, which cannot be described by a single centrality. [47]This research proposes a machine-learning-based method for determining which nodes are important in a given spreading circumstance. This framework first labels each node with its precise transmission capacity obtained from simulated transmission experiments based on the SIR model, and then uses seven device mastering algorithms to analyze the complex court, all based on the prevalent centrality strategies that can reflect nodes' unique topological structures and the contamination rate that is an essential part of the transmission scenarios. The results of experiments on real-world networks show that the model based on system learning has higher classification accuracy than the conventional centrality approaches relying on a single topology. [100]. Using Twitter's API, this study extracts data from the social networking website Twitter. The retrieved data is manually categorized into propaganda and non-propaganda classifications. Three distinct textual characteristics (TF/IDF, bag of words, and tweet length) are being combined to accomplish hybrid feature engineering. According to the findings, propagandistic material was longer than non-propagandistic content. Using machine learning algorithms, tweets are classified as either propaganda or non-propaganda. With an accuracy of 98.5%, a precision of 0.99, a recall of 0.99, and an F1-Score of 0.99, the decision tree classifier outperformed all other machine-learning techniques.

## 5.2 Weighted correlation influence

The Weighted Correlated Impact (WCI) technique was presented in this work. It takes into account the relative contribution based on time and trend-specific online user attributes as shown in Fig. 16. In order to give users of online platforms with an influence score that is reflective of the combined effect, the method in question involves merging the profile activity and the underlying network structure. Twitter's #CoronavirusPandemic hashtag is utilized to measure the performance of the suggested strategy. Additionally, the findings have been confirmed for a second social media trend. The two datasets were constructed using hashtags that are trending internationally. [33]The experimental findings indicate that the suggested WCI outperforms existing techniques in terms of accuracy, recall, and F1-measure validation [93].The base for influential user detection is formed by taking multiple parameters into consideration. As the problem of finding influential users is hard and is in infancy stage, more research can be done for effectively finding prominent users from Social Networks [6].

In this research, a local influence assessment heuristic is used to estimate the local effect in the two-hope region. Using this heuristic, a standard diffusion model's predicted diffusion



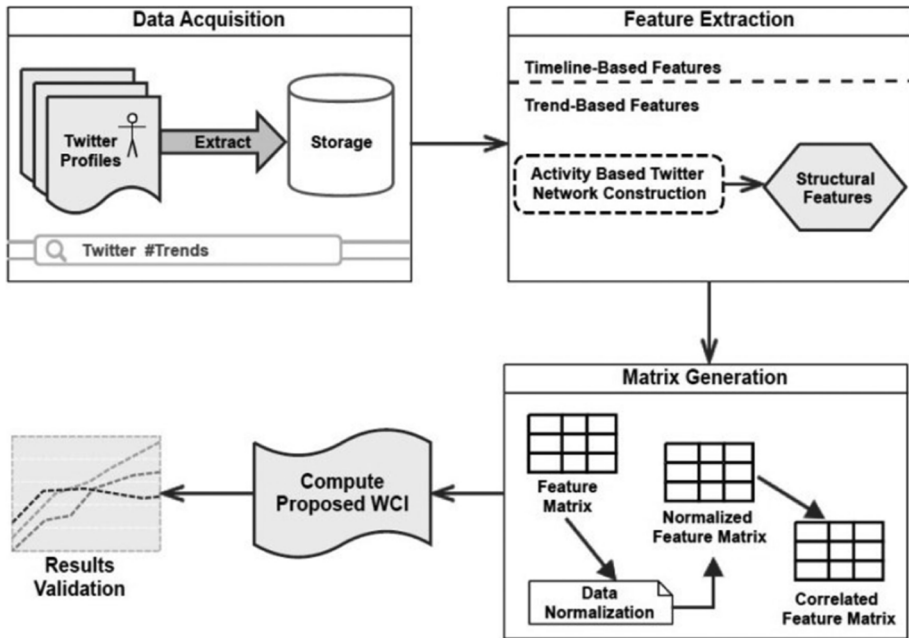


Fig. 16 Proposed framework for WCI

value is calculated. A approach for boosting influence based on ant colony optimization (ACO-IM) is provided in order to optimize the local influence assessment heuristic. ACO-IM modifies the representation and rule of pheromones and heuristic information deposited by ants [30]. The technique used the probabilistic setting to prevent premature convergence. Finally, the experimental results demonstrate the algorithm’s superiority. [78] In addition, statistical tests were conducted to differentiate the suggested strategy from current practices [96].

Two new effective algorithms, called “MinCDegKatz d-hops” and “MaxCDegKatz d-hops”, were created as a direct outcome of this study. Due to their proven effectiveness and performance in terms of influence distribution and reduced execution complication, such algorithms rely on a mix of centralization methods. The proposed approaches combine degree centrality as a neighborhood metric with Katz centrality as a global metric to determine the most important nodes in a network with predetermined weight edges that are larger than some threshold [54]. For this reason, a hop count that changes with the radius of the graph is used to ensure that the different seed sets selected remain unique from one another. The Independent Cascade (IC) and Linear Threshold (LT) models were then used to examine the impact spread of the two proposed techniques[45]. Extensive tests were conducted on a large-scale network in this paper to demonstrate the efficacy of the suggested algorithms vs competing methods [6]. The tests were conducted to demonstrate the efficacy of the suggested algorithms in terms of their ability for propagation and their computational complexity.

This inquiry attempts to locate a collection of dispersed nodes with significant propagation capabilities. We present a social conformity-based node ranking mechanism and a Vote Rank-based community feature as a reaction to the interaction between people and groups in sociological and complex networks [51]. This approach analyzes the effect capacity of a node from both the individual and group viewpoints. According to the conformance princi-

ple, it calculates, from the individual's viewpoint, the attractiveness of the nodes based on the characteristics of their neighbors [50]. Nodes of the exact same degree and pattern may be distinguished [?]. Another perspective on beginning power considers the network's size and the node's position within it [99]. To further address the problem of duplication, a node selection technique is provided that takes into account both resource covering and neighborhood density [71]. Boosting Vote Rank by a combination of node attractiveness, beginning power, and the node suggested technique is possible. The usefulness of the suggested strategies is shown by operational network outcomes. The statistics further suggest that the strength of the node is a result of the enormous energy of the groups [120].

Using k-shell decomposition and N-gram similarity, the authors of this study developed a de-duplicated k-shell influence estimating (DKIE) model in social networks, which may be used to identify key opinion leaders in any given field [102]. The DKIE approach consists of two fundamental processes, the discovery of both domain-specific and universal important spreaders. In the first step, we utilize the k-shell decomposition method to rank each socially engaged user by their importance in terms of the network's structural connections and hence their potential for spreading information. Despite the presence of duplicate connections, it distinguished between groups resembling the core and the real core and identified the most successful spreaders. In the second step, N-gram similarity assessment was utilized to categorize the best spreaders within each area by considering the topics they most often discussed. The results prove that the DKIE technique is useful [10].

The suggested method takes into account local data and decreases computational complexity. Shannon's entropy gives a more exact representation of the information contained in boxes (number of nodes). Specifically, a node's influence becomes proportionally bigger as its local information dimension grows. To prove the usefulness of the proposed method, five different centrality measures were used to rank significant nodes in six different complex real-world networks [79]. An SI model and Kendall's tau coefficient were also used to demonstrate the interconnectedness of these methods. The experimental findings demonstrate that the suggested approach is preferable [105].

This study structured the maximizing of impact issue as an optimization problem, with the inferentiality and distance of the nodes serving as cost functions. Increasing the distance between the seed nodes ensures that all network nodes have access. The issue has been solved using the grey wolf optimization technique [101]. The experimental findings on three actual networks demonstrated that the proposed technique outperforms current influence maximization methods. In addition, its computational time is less than that of other meta-heuristic approaches [43].

The authors of this piece used a comprehensive multilayer network model to shed light on different types of social networks, and as a result, they provided the neighborhood-based node impact indicator. Extensive tests were run on 21 different real-world datasets to prove the suggested method's improved performance [32]. The ability to show how social networks evolve is crucial, therefore we're hoping our study may help pave the way for future scholars in this area [93].

This research offered an organization of cutting-edge methodologies into model-based network comparisons. Due to the structure of network models and the difficulty of recognizing important nodes, this survey will aid researchers in determining the method set that best matches their needs. The recommended category might also aid researchers in determining the best direction for future research [38, 81].

This research suggested a novel Social Network-based Whale Optimization Algorithm (SNWOA) for determining the top-N opinion leaders by evaluating the user's reputation using a variety of traditional network optimization functions. As the number of network

users increased, the algorithm found the optimal solution; hence, the overall complexity of the technique remained constant. A similarity metric that includes neighbor similarity and clustering coefficient was also offered as a fresh approach to discovering communities [123]. First, they estimated the target function of each user based on their centralities, and then they identified local and worldwide opinion leaders using the provided technique with multiple optimization functions. The recommended approach was applied to real and simulated datasets, and the results were compared based on precision, recall, F1-score, and accuracy [85]. In comparison to other current Social Network Analysis (SNA) measures, the suggested algorithms demonstrate superior performance. This study shown that the community partitioning strategy is superior to previous community identification algorithms [93] in terms of computing time and other parameters.

This study aims to identify the combination of network data that best forecasts the overall expected information dissemination. The conclusions and portions of the approach may be used to the field of Influence Maximization, in which nodes are prioritized according to their expected influence spread [38]. This study examines how effectively specific network statistics and their combinations may predict the overall anticipated information distribution (influence spread). Using the Barab'asi-Albert random graph generation model, the suggested technique focuses on all non-isomorphic networks with  $N \in \{9, 25, 50, 100, 200, 400\}$  nodes. Each network's unique seed nodes are then given the Independent Cascade (IC) and Weighted Cascade (WC) spread models to replicate the original network's dissemination [93]. Finally, the overall information distribution was predicted using Machine Learning approaches such Random Forest Regression and k-Nearest Neighbors regression [90].

The study's goal was to create a two-stage mining approach (GAUP) for finding the most significant nodes in a network related to a certain topic. Given a set of user-generated texts annotated with topic labels, GAUP first calculates user preferences using a latent feature model based on SVD or a model based on vector space [43]. GAUP uses a greedy algorithm in its second phase to guarantee a solution that is 63% as good as the optimal one. In the expert finding evaluation, GAUP performs better than the present state-of-the-art greedy algorithm, SVD-based collaborative filtering, and HITS [124].

This study discusses the characteristics, models, and central nodes of virtual social networks. Techniques for locating highly consequential aggregation points are discussed [1]. Both the Diffusion model and the Centrality-based model have been widely employed to locate key hubs. The Shapley value technique, the Preference model, the Web structure and page ranking method, the Random walk model, and the Path diversity model have also been investigated [127]. This research also compares and contrasts the various models used in the process of identifying significant nodes [90].

This study introduced a novel method for identifying prominent members in online social networks. This unique strategy is based on the structural qualities and behavior of social network members [24]. Topological characteristics and connections between nodes inside social boundaries are structural attributes. The behavior of a node in Social Networks is represented by user behaviors and public data [94].

In this research, a convex network inference model was developed and the method was applied to the social network using prior information. In addition to this method, the proposed framework included two modifications: the self-exploring model, which enables user information discovery without knowledge propagation, and the rating model, which integrated rating data. The credit distribution model for influence maximization was used in comparison. The new approach yielded the greatest results in identifying a group of very important people on the network [72]. Additionally, we simulated assaults on the

social network that altered 0.55 percent to 55 percent of the social network. The findings shown that the overall weight of the network may be altered by up to 80%, although the impact per review for the 100 most important people is altered by just 18%. This percentage might increase to 26.6% [26] if attackers target people with a significant number of reviews.

By evaluating user interaction data, this study tackles the problem of pinpointing social network leaders and followers. The study developed a new model, the Longitudinal User-Centered Influence (LUCI) model, based on user interaction data, that classifies people into four types: introvert leaders, extrovert leaders, followers, and neutrals. The model was tested by using data scraped from the popular social networking website Everything [29]. According to the results of the tests conducted, the LUCI model was able to achieve a classification accuracy of up to 90.3% when classifying users into leaders and followers based on their responsibilities. [73] Second, the model was applied to a data set consisting of interactions between more than three million Facebook users over the course of one year and collected from Facebook [56]. Lastly, it has a graph, and the observed attributes match those expected by the general description of the four roles [?].

This research presented a Content-Based Influence Measurement, or CIM, to predict important members in social networks by taking both content and context into consideration. [97] We use Context-Based Influence Measurements, which assessed a number of crucial topological aspects of the friendship network for various user groups [11]. Consider static social network structure, such as centrality analysis, to assess the robustness of our technique. [9] Our experimental findings indicate that various user categories display distinct topological features in the friendship network Influence has been described as the total number of social interactions a user obtains via his/her posts, which not only reflects the user's popularity but also the interaction strength between influential users and their followers and the quality of the influential user's postings [100]. The findings on 7394 Flickr users demonstrate that the suggested method is better to the alternatives. [110] It also demonstrates that the interaction intensity between followers and influential users varies, and that content alone cannot predict influential users [50].

Understanding the social dynamics of a Facebook-based activist organization was the goal of this article, which used methods from the field of Social Network Analysis (SNA). It is also of interest to us to determine who is the most influential in the group and to infer the strength of the bonds between them [26]. We have successfully used sentiment analysis to determine which users have had positive and negative impacts on the whole; this may contribute to a better knowledge of the group [84]. The researchers conducted a temporal study to establish a correlation between online actions and pertinent real-world occurrences. The experimental findings indicate that applying such data analysis methods to the online behavior of users is an effective method for predicting the impact and connection strength of group members [100]. Finally, findings were checked against reality and it was determined that the suggested technique is quite promising for attaining its objectives [76].

This paper proposes a system based on frequent pattern mining to identify influential users and the optimal timing to disseminate information. The 24-hour span of a day is divisible into subsequent time parts [109]. In each time segment, an influence transaction including supporters' influence power will be defined. After many days of collecting transaction data, recurring patterns may be identified to determine the optimal moment for influence users to disseminate information [49].

## 6 Analytical comparison of techniques used

This article describes the majority of methods that have been utilized to identify prominent members in online social media platforms. [113] Each method's mechanism, pros and cons are addressed in its respective area.

### 6.1 Associative rule learning method

The Associative Rule is used to identify significant members within the context of online social networks [75]. This approach employs the following metrics: degree of conviction, self-assurance meter, guide metric, and lift metric. This strategy produced pretty excellent results for small datasets, but for huge datasets it may be disastrous [7].

### 6.2 TLFM (Temporal and Locational Fellowship Model)

TLFM (Temporal and Locational Fellowship Model) has been used for identifying prominent people in an online social network [122]. TLFM employs both geographical and temporal criteria to identify important users. This method's limitation is that it only responds to offline inquiries, making it primarily an offline system [80]. The following functionalities are utilized:

- Probabilistic model is used for pinpointing influential users.
- Spatial based affect, which is a pinpoint location in context of geo spatial.
- Utilizes a strategy dependent on time to identify significant customers. Thus, uses a probabilistic model that is temporally and spatially based to identify important customers.

### 6.3 3) LUCI (Longitudinal User Centred Influence) model

The Longitudinal User Centred Influence (LUCI) version of a model for identifying leaders (influential users) in an Internet-based social network has been described [118]. It employs both the ego and community coefficients. Based on these two coefficients, LUCI applies a kernel K approach rule set and presents correct findings [94]. Include the following characteristics:

- Based on fundamental Internet social network interactions.
- Utilizes both ego and network coefficients.
- Ego coefficient assesses the relationship between the past and future interactions of a person.

#### 6.3.1 Influence Ranking (INFR) model

In order to identify important people in Online Social Networks, the Influence Ranking (INFR) method has been presented for Online Speech-Based Networks (SBN). [5] This method use the betweenness coefficient to identify influential clients [92]. This strategy employs both greedy and heuristic effect maximizing strategies for identifying K influential consumers. This method might be extremely Swift [52]. Features include the following:

- This strategy is used to identify K prominent clients in speech-based internet networks (SBN).

- For calculation centrality coefficient is used to have an impact.
- K is taken as hard coded; the variety of influential nodes.

## 7 Earlier ML techniques

Proposed is a broad approach for identifying prominent users based on graph traversal [92]. Multiple metrics are used to measure the significance of every node in a social network from various viewpoints. Degree centrality and H-index are examples of neighborhood-based methodologies for assessing centrality [64]. Different centrality-based approaches, such as Closeness centrality, Betweenness centrality, information centrality, and Load centrality, are path-based methods that account for global topological characteristics. [111] PageRank and Eigenvector centrality evaluate the influence based on iterative refinement manner [89]. Despite decades of study on influential user identification based on these centrality-based methodologies, there is no agreement on the appropriate centrality metric to utilize inside different kinds of networks in diverse fields for identifying important individuals [70].

This paper has proposed a different method of selecting new centrality based method having below mentioned characteristics:

1. It must be general so that different other existing centrality based methods can be incorporated.
2. Based on different principles that have remained unexplored till date must be used to solve node centrality issue.
3. It must outperform other centrality methods used on different networks.

This paper proposes influential node identification method TARank, which is based on graph traversal framework. By treating the target node as the root node and building Breadth First Search, the effect of each node is assessed (BFS) [126]. This BFS tree is generating a curve, where the X-axis represents the level and the Y-axis represents the cumulative score of all visited nodes to date. [51] The area under the curve is utilized to calculate the final influence score of the target node. This method is superior as compared to earlier used centrality based measures for identifying influential users [64].

For pinpointing influential nodes in social networks like Instagram a new method is proposed in which User Generated Content (UGC) is taken into consideration. [55] This approach combines multiple sorts of characteristics collected from photographs uploaded on visual social networks without understanding the network's architecture or determining whether a user is important or not [35]. One of the most important contributions of this study is that it does not depend just on the amount of interactions, such as the number of posts, likes, and follows [7]. This research of identifying influential users is also not dependent on structure of their connectivity [12]. This research focuses on both low and high level visual feature extraction from different images and then developing a classifier by combining informative features [34]. This classifier is then used to differentiate between photographs submitted by prominent users and those shared by regular users. [59] Basically there are three ways of identifying influential users:

1. In the case of topological approaches, impact is computed in a network graph based on the position of each node.
2. Among users by incorporating interactions influence of a user is calculated in case of user interaction-based methods.

3. Influence of a user is calculated in case of content based methods based on the content of users published on the visual social networks.

With the proliferation of social bots and phony accounts that can automatically amass followers and create messages, interaction-based techniques have failed to reliably identify significant individuals [31]. While potentially offering more accuracy, the more complex algorithms demand a lot of computer power to calculate the more intricate features. Content based approaches, make an effort to employ a third information source that was disregarded by the first two types of solutions, all the information contained on user postings (or User Generated Contents) [7, 8]. To properly assess the influence of users, these methodologies provide several integration strategies for this data into earlier approaches. [37] Detection of influential posts and pinpointing influential users in case of online social networks are interconnected [41]. As a result, recognising influential posts can be used to study identifying influencers. [53] Influential posts have elements and traits that can effectively influence the followers. By gathering local and global features from data, this study uses supervised machine learning to identify influential posters [7]. Three sections make up the suggested methods:

1. Data gathering.
2. Extraction of characteristics and identification of prominent posts that emphasize the visual representation of pictures, as well as classification of visual material as influential or unimportant.
3. Development of a classifier and identification of significant users, using many classification techniques to construct a final classifier that can identify influential posts and noteworthy individuals.

## 8 Graph convolution networks for influential user detection

Neural Networks have gained a lot of success in the last decade. Neural Networks work mostly on Euclidean data or structured data but most of the real world data is graphical data which is unstructured or Non-Euclidean. [121] This non-regularity in data gives birth to Graph Neural Networks [123]. In the past few years different variants of Graph Neural Networks were developed like Graph Attention Networks (GAT), Graph Convolution Networks (GCN) etc [40]. There are two main types of GCN algorithms; spatial graph convolutional networks (GCNs) and spectral graph convolutional networks (GCNs) [103]. The concept of “convolution” in GCNs is equivalent to that of “convolution layers” in conventional neural networks. Multiplying the input neurons by a set of weights, sometimes called a filter or kernel, is what this term refers to. When applied to a whole picture, these filters function like a sliding window, allowing convolutional neural networks to pick up characteristics from nearby cells. When a filter is shared across nodes in the same layer, it is applied uniformly over the whole picture [60, 69]. Cat’s nose and ears, for instance, are detected using the same filter in the same layer when CNN is used to distinguish between cat and non-cat pictures [82]. GCNs also carry out operations in which the model learns the characteristics by examining the surrounding nodes [40, 88]. When comparing convolutional neural networks (CNNs) and generalized convolutional neural networks (GNNs), it’s important to note that CNNs are optimized for use with regular (Euclidean) structured data, while GNNs are the generalized version of CNNs that can handle data with varying connection numbers and unordered nodes. [23] In case of GCNs feature selection is automatic as compared to Machine Learning methods where feature selection is manual as shown in the below Figs. 17 and 18

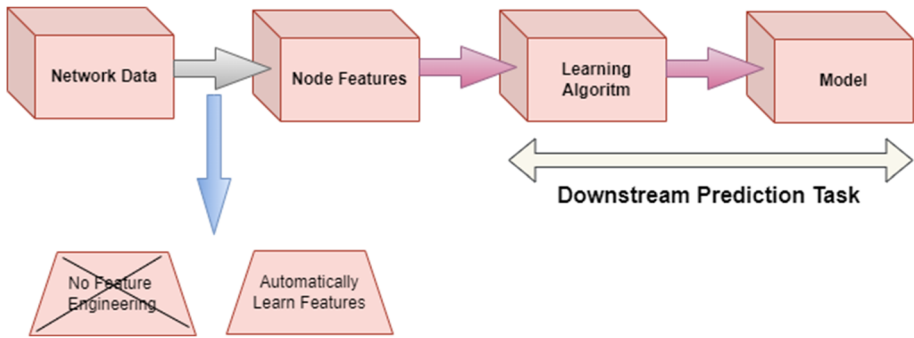


Fig. 17 Automatic Feature Selection in GCN

### 8.1 Correlation matrix

A correlation matrix is a table that displays the coefficients of correlation between a set of variables. The values in the matrix range from -1 to 1, with 1 representing an ideal positive correlation, 0 representing no connection, and -1 representing an ideal negative correlation [57]. The result of a correlation matrix can be utilised to get insight into the relationships between variables [117]. To gain insights into the relationship between various input features and influential user detection, we performed a correlation analysis on a subset of the datasets included in our systematic survey [13]. Specifically, we analyzed the following features for their correlation with the ground truth labels of influential or non-influential users:

- Number of followers
- Number of retweets
- Number of mentions
- Number of replies
- User activity level (defined as the total number of tweets)

We used Pearson’s correlation coefficient to measure the strength and direction of the correlation between each of these features and the ground truth labels [35]. The results of our analysis are summarized in Table below. Table : Correlation between input features and ground truth labels of influential or non-influential users.

As shown in the above table , the number of followers has the strongest positive correlation with the ground truth labels, with a correlation coefficient of 0.71. This indicates that users with a larger number of followers are more likely to be influential [58]. The number of retweets and user activity level also show moderate positive correlations with the ground truth labels, while the number of mentions and replies have weaker positive correlations [44].

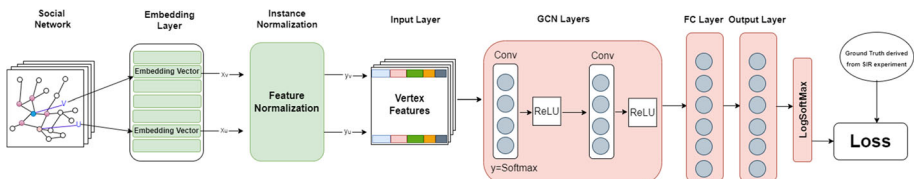


Fig. 18 Framework for identifying influential users using graph convolution networks



Input Feature	Correlation Coefficient
No of Followers	0.71
No of retweets	0.52
No of mentions	0.32
No of replies	0.22
User activity level	0.58

Our correlation analysis provides insights into the importance of different input features for influential user detection, and can inform the development of more accurate and effective detection models [98]. These results suggest that features such as the number of followers, retweets, and user activity level may be particularly useful for identifying influential users, while features such as mentions and replies may be less important [114].

### 8.1.1 Strong and weak correlation

The degree of link between two variables is referred to as their correlation. When two variables are connected, their movements tend to be predictable. A correlation coefficient ranging from -1 to 1 can be used to quantify the correlation's strength.

- **Strong positive correlation:** A strong positive correlation is shown by a correlation coefficient of 0.9 or more. For example, there is a strong link between how tall someone is and how much they weigh. People who are taller tend to weigh more than people who are shorter.
- **Weak positive correlation:** If the correlation coefficient is between 0.3 and 0.7, the correlation is weak. For instance, there is a weak link between a person's level of education and their income. Most of the time, people with more education make more money, but there are a lot of exceptions.
- **Strong negative correlation:** A correlation coefficient of -0.9 or less means that the two things don't go together well. For example, there is a strong link between a person's age and how long it takes them to react. As people get older, they tend to take longer to act.
- **Weak negative correlation:** If the correlation coefficient is between -0.3 and -0.7, then the correlation is weakly negative. For example, there is a weak negative link between a person's height and their blood pressure. Blood pressure tends to be lower in taller people, but the link is not very strong.

## 8.2 Variables for analysis of influential user detection

There are several variables that can be taken into account for analysis in the task of influential user detection:

- **Number of followers:** The number of followers is a key indicator of the reach and potential influence of a user. A user with a large number of followers is likely to have more influence over their followers [86].
- **Number of followers:** The number of followers is a key indicator of the reach and potential influence of a user. A user with a large number of followers is likely to have more influence over their followers [86].
- **Frequency of posts:** The frequency of posts can indicate the level of engagement and activity of a user. A user who posts frequently is likely to have a larger audience and more opportunities to influence their followers [86].

- **Engagement of followers:** The engagement of followers can indicate the level of interest and influence that a user has over their followers. A user with highly engaged followers is likely to be more influential than a user with followers who are less engaged [86].
  - **Network centrality:** Network centrality measures the centrality or importance of a node in a network. In the context of influential user detection, network centrality measures can be used to identify users who are central to the network and have a high degree of influence over other users [86].
  - **Content quality:** The quality of the content produced by a user can also be a factor in their level of influence. A user with high-quality content is likely to be more influential than a user with low quality content [86].
  - **Sentiment analysis:** Sentiment analysis can be used to determine the sentiment of the content produced by a user. A user with positive sentiment is likely to be more influential than a user with negative sentiment [86].
  - **Demographic information:** Demographic information, such as age, gender, location, and education level, can also be taken into account in the analysis of influential users. Different demographic groups may be influenced by different types of users and content [86].
- These are some of the variables that can be taken into account for analysis in influential user detection. The choice of variables will depend on the specific requirements and the type of data available for analysis

### 8.3 Evaluation techniques

- **Accuracy, precision, hit count:** Accuracy in influential user detection refers to the degree to which the identified influential users truly have a substantial impact on the network, and the extent to which they accurately represent the true influential users within the network [93].  
Precision in social network analysis, precision in influential user detection refers to the proportion of identified influential users that are actually influential, as opposed to those that are falsely identified as influential [112]. In other words, it measures the accuracy of the positive predictions made by the influential user detection algorithm. High precision means that the identified influential users are more likely to be truly influential, and thus, the results of the analysis are more reliable [6]. Low precision, on the other hand, means that there are more false positives, which can lead to inaccurate or misleading results.  
Hit count typically refers to the number of times a user has been mentioned or referenced by other users in a given dataset, such as in a social network or online community. In other words, the hit count is a measure of how frequently a user has been “hit” by others in the form of likes, comments, shares, or other types of engagement. The underlying assumption is that users who receive a high hit count are more likely to be influential than those who receive a low hit count, as their actions and content are more visible and have a greater impact on others.
- **Recall** is a measure of the proportion of actual influential users that the model correctly identifies as influential. It is calculated as the ratio of true positive (TP) predictions (i.e., the number of influential users correctly identified by the model) to the total number of actual influential users in the dataset (TP + false negative (FN) predictions).
- **F1 score** is a weighted average of precision and recall, where precision is the proportion of predicted influential users that are actually influential. It is calculated as  $2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$ . F1 score combines both precision and recall into a single metric, and provides a way to compare models based on their ability to balance

between correctly identifying influential users (recall) and avoiding falsely identifying non-influential users (precision).

- **MAP (mean average precision):** measures the average precision of the ranked list of influential users generated by the algorithm, considering all possible levels of recall. To calculate MAP, the algorithm first generates a ranked list of users based on their predicted influence scores. Then, for each level of recall, the algorithm calculates the precision of the top-ranked users up to that point. Precision is calculated as the ratio of true positive predictions to the total number of predicted positive (influential) users at that level of recall. Once the precision is calculated for all levels of recall, the algorithm averages the precision scores to obtain the MAP. MAP provides a comprehensive evaluation of the algorithm's ability to correctly rank influential users, considering both the order and the number of influential users in the ranking.
- **Mean Absolute Error:** the MAE measures the average absolute difference between the predicted influence scores and the actual influence scores for a set of users. For example, suppose we have a dataset of users and their corresponding influence scores, and we want to develop a model that predicts the influence scores for new users. We can use MAE to evaluate the accuracy of our model by comparing the predicted influence scores to the actual influence scores for a set of test users. A low MAE indicates that the model is accurately predicting the influence scores, while a high MAE suggests that the model is not performing well. By minimizing the MAE, we can improve the accuracy of the model and make it more effective at identifying influential users in a social network.
- **Root Mean Square Error (RMSE)** is another commonly used metric in machine learning and data analysis to measure the difference between predicted and actual values. In the context of influential user detection, RMSE can also be used to evaluate the performance of a model that predicts the influence of different users on a social network. Like MAE, RMSE measures the average difference between predicted and actual values, but it takes into account the squared differences between the predicted and actual values, rather than just the absolute differences.
- **AUC (Area Under Curve)** is a useful metric for evaluating the performance of binary classifiers in influential user detection, and a high AUC score indicates that the classifier is effective at distinguishing between influential and non-influential users.
- **Normalized Discounted Cumulative Gain (NDCG)** is a commonly used metric for evaluating the effectiveness of ranked lists, such as those generated in influential user detection. NDCG takes into account both the relevance of the items in the list and their position in the list, and provides a normalized score that ranges from 0 to 1.
- **Pearson correlation coefficient** is a commonly used method for calculating the correlation between two variables. It measures the linear correlation between two variables, with a value ranging from -1 to +1, where -1 indicates a perfectly negative correlation, +1 indicates a perfectly positive correlation, and 0 indicates no correlation. A high positive correlation coefficient indicates that the predicted influence scores are highly correlated with the actual influence scores, which suggests that the model is accurately predicting the influence of users in the social network. On the other hand, a low or negative correlation coefficient suggests that the model is not performing well.

In summary, correlation coefficient is a useful metric for evaluating the performance of models in influential user detection, and Pearson correlation coefficient is a common method for measuring the correlation between predicted and actual influence scores.

## 8.4 Hyperparameters for influential user detection using GCNs

Graph convolutional networks (GCNs) are a type of neural network architecture designed for processing data that is structured as a graph, such as social networks. In the context of influential user detection, GCNs can be used to identify the most important or influential users in a social network based on their connections to other users.

Here are some hyperparameters that are commonly associated with graph convolution networks for influential user detection: Number of convolutional layers: This hyperparameter determines how many layers of convolutions are applied to the graph. More layers may lead to better performance but also increase the risk of overfitting.

- Number of units per layer: This hyperparameter determines the number of units (or neurons) in each convolutional layer. A larger number of units can capture more complex relationships but also increase the risk of overfitting.
- Activation function: The activation function applies a non-linear transformation to the output of each layer. Common activation functions include ReLU, sigmoid, and tanh.
- Dropout rate: Dropout is a regularization technique that randomly drops out a percentage of the units in each layer during training. This helps to prevent overfitting by forcing the network to learn more robust features.
- Learning rate: This hyperparameter determines how quickly the network adjusts its weights during training. A larger learning rate can lead to faster convergence, but may also cause the network to overshoot the optimal weights.
- L2 regularization parameter: L2 regularization is a penalty term added to the loss function to discourage the network from assigning too much importance to any one feature. The L2 regularization parameter determines the strength of the penalty term.
- Initialization method: The initialization method determines how the network initializes its weights. Common initialization methods include random initialization and Xavier initialization.

These are just a few of the hyperparameters that can be tuned in a graph convolutional network for influential user detection. The optimal values for these hyperparameters will depend on the specific task and dataset being used.

## 9 Datasets for influential user detection

There are several datasets available for influential user detection, some the datasets have been discussed below [Table 1](#)

## 10 Open challenges

- As the number and size of social networks continue to grow, however, determining a user's social impact poses a number of difficulties. We think there are a number of obstacles in this scenario, and we outline some of them below. It is difficult to characterise the causal relationship of social influence and the uncertainty of social influence because, first, we lack a mathematical definition and measurement of social influence; second, it is difficult to decide the major factors for a specific case to social influence modelling; third, there is no effective ways to properly integrate various factors for influence measurement.

**Table 1** List of datasets for influential user detection

Dataset Name	No of Nodes	No of Edges	Source
CORA	2708	5429	<a href="http://www.cs.umass.edu/">http://www.cs.umass.edu/</a>
CITeseer	3312	4732	<a href="http://www.cs.umd.edu/sen/">http://www.cs.umd.edu/sen/</a>
PUBMED	19717	44338	–
Reddit	232965	5376619	–
PPI	56944	818716	–
NetHEPT	15233	58891	<a href="http://research.microsoft.com/enus/">http://research.microsoft.com/enus/</a>
NetHPY	137154	196591	<a href="http://research.microsoft.com/enus/">http://research.microsoft.com/enus/</a>
Email-EuAll	265214	420045	<a href="https://snap.stanford.edu/data/">https://snap.stanford.edu/data/</a>
munmun-twitter-social	465017	834797	<a href="http://konect.uni-koblenz.de/">http://konect.uni-koblenz.de/</a>
Sina Weibo	63641	1391718	–
Epinions	131828	841372	–
wikinote	7115	103689	–
DBLP	215222	611542	<a href="https://aminer.org/billboard/">https://aminer.org/billboard/</a>
ACM	468114	1862602	<a href="https://aminer.org/billboard/">https://aminer.org/billboard/</a>
Yelp	1282	30838	<a href="https://github.com/librahu">https://github.com/librahu</a>
Zacharys karate club	34	78	–
Dotphins Social Network	62	159	–
Les Miserables	77	254	–
Books about US Politics	105	441	–
American College Football	115	613	–

- topological network architectures provide the foundation for traditional centrality approaches. Different centrality methods consider various network structures. For example, some methods consider global structures represented by betweenness centrality and closeness centrality, while others consider local patterns represented by degree and semi-local centrality. The significance of a node, however, is not only tied to its local and global structures in many practical scenarios, such as the propagation scenario, but is also related to other factors, such as the significance of its neighbors and the infection rates. Between a node's functional significance and its various characteristics, there is a complicated coupling relationship. The common flaw of traditional centralities is that it is impossible to consider all relevant aspects in a single centrality, which results in real-world performance and flexibility restrictions. These traditional measures weren't found too much useful because of many reasons like they have high computational complexities, they also do global ranking in which they are based on their betweenness centrality select top-k nodes but there is possibility that these top- k nodes might belong to single community due to which the reach or diffusion then isn't too much.
- Another method for identifying users is machine learning which is based on node features and where we use concept of feature engineering . In machine learning we have to select better features for getting better performance. Machine learning techniques like Logistic Regression and Support Vector Machine have been applied to influential user detection in social networks but such techniques rely on features that are manually selected. Machine learning methods depend too much on feature selection of the target objects and ignore the

relationships between the targets. Therefore, a poor selection of these features deteriorates the performance of the technique often yielding unacceptable results.

- To solve this problem, a deep learning model is needed that will consider both network structure and node features.
- Identification of multiple types of influence: Identifying influential users can be challenging because influence can come in different forms. For example, some users may have a broad reach, while others may have a deep impact on a specific community. Algorithms need to be able to identify multiple types of influence accurately.
- Privacy concerns: Identifying influential users can involve collecting and analyzing large amounts of data, which can raise privacy concerns. It is essential to develop algorithms that respect the privacy of users while still providing accurate results.
- Dynamic networks: Social networks are dynamic, and the influence of users can change over time. Algorithms need to be able to adapt to changes in the network structure and identify new influential users as they emerge.
- Cross-platform analysis: Influential users can have an impact across multiple social media platforms. Detecting cross-platform influence is a challenging problem that requires analysing data from multiple sources and identifying common patterns of behaviour.

## 11 Limitations of State-of-the art approaches

Topological approaches consider only structure of the network and are not scalable. However, the real world networks are extremely large and dynamic and these approaches fail to scale up to such levels.

Machine Learning-based approaches require manual feature selection, and improper selection of such attributes degrade the performance of these approaches.

Deep learning-based approaches particularly the GCN does automatic feature selection and is scalable. GCN is also suitable for non-Euclidean data . GCN-based techniques are by far most effective for influential user detection, considering both structure as well as attributes of the network. These models are able to represent the social networks at scale. Although, there are no specific limitations of the GCN with regard to detection of influential users, however, some limitations inherent to the GCN model that may also be the limitations for influential user detection using GCN are:

- Data requirements: Deep learning methods typically require a large amount of labeled data to train effectively. In the case of influential user detection, labeled data may be hard to come by, especially for newer or niche communities.
- Limited interpretability: Deep learning models can be difficult to interpret, meaning that it may be challenging to understand why a particular user is deemed influential. This can limit the model's usefulness in scenarios where transparency and accountability are essential.
- Overfitting: Deep learning models can be prone to overfitting, where the model becomes too complex and learns to fit the noise in the data rather than the underlying patterns. This can lead to poor generalization performance on unseen data.
- Computationally expensive: Deep learning models can require a significant amount of computational resources, making them impractical for some applications.
- Limited graph structures: Deep learning models, including GCNs, are best suited for analysing graphs with a homogeneous structure. However, many real-world graphs have

a heterogeneous structure, which can make them challenging to analyse using deep learning methods.

## 12 Conclusion and future prospects

In conclusion, Identifying influential users is important in various fields including social media, online marketing, and e-commerce. There are various methods for detecting influential users, but there is still room for improvement in terms of generalizability and scalability. Challenges in influential user detection include selecting the right features, dealing with dynamic networks, privacy concerns, distinguishing influence from popularity, dealing with spam and fake accounts, and defining influence. Despite these challenges, influential user detection will remain relevant in the future due to the increasing importance of social media and online communities.

The popularity of online social networks has resulted in a need for identifying significant users, and various techniques have been used, such as structural and machine learning-based approaches. However, these methods suffer from limitations, including the computational complexity of large networks and the need for manually specified criteria. Future research directions include the use of deep learning models that consider both network topology and node characteristics, identifying influential people via user-generated content, identifying significant community members, and using Graph Neural Networks to improve recommender systems. The majority of research conducted so far identifies populations that are disconnected or isolated. However, we may utilize community identification to find groups that overlap, and then extend influential user recognition to these locations. Graph convolution Networks will be utilized to find overlapping communities, and prominent users will be found based on these overlapping groups. This subject may be approached as a classification problem, after which an in-depth model can be developed to identify important individuals. To identify influential users, future approaches should incorporate both structural and behavioural features of online social networks.

**Data Availability Statement** Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

## Compliance with ethical standards

**Conflicts of interest** The authors declare that they have no conflict of interest.

## References

1. Abbruzzese R, Gaeta A, Loia V, Lomasto L, Orciuoli F (2021) Detecting influential news in online communities: an approach based on hexagons of opposition generated by three-way decisions and probabilistic rough sets. *Inf Sci* 578:364–377
2. Albert R, Jeong H, Barabási A (2000) Error and attack tolerance of complex networks. *Nature* 406(6794):378–382
3. Al-Garadi, MA, Varathan, KD, Ravana, SD, Ahmed, E, Mujtaba, G, Shahid Khan, MU, Khan, SU (2018) Analysis of online social network connections for identification of influential users: survey and open research issues. *ACM Comput Surv* 51(1)
4. Al-garadi, MA, Varathan, KD, Ravana, SD (2016) Identification of influential spreaders in online social networks using interaction weighted K-core decomposition method. *Phys A*

5. Ali M, Hassan M, Kifayat K, Kim JY, Hakak S, Khan MK (2023) Social media content classification and community detection using deep learning and graph analytics. *Technol Forecast Soc Change* 188(May 2022)
6. Alshahrani M, Fuxi Z, Sameh A, Mekouar S, Huang S (2020) Efficient algorithms based on centrality measures for identification of top-K influential users in social networks. *Inf Sci* 527:88–107
7. Alwan WH, Fazl-Ersi E, Vahedian A (2020) Identifying influential users on instagram through visual content analysis. *IEEE Access* 8:169594–169603
8. Amati G, Angelini S, Gambosi G, Rossi G, Vocca P (2019) Influential users in Twitter: detection and evolution analysis. *Multimed Tools Appl* 78(3):3395–3407
9. Arora A, Bansal S, Kandpal C, Aswani R, Dwivedi Y (2019) Measuring social media influencer index—insights from Facebook, Twitter and Instagram. *J Retail Consum Serv* 49:86–101
10. Arularasan AN, Suresh A, Seerangan K (2019) Identification and classification of best spreader in the domain of interest over the social networks. *Clust Comput* 22:4035–4045
11. Bahutair M, Al Z, Ibrahim A (2022) NodeRank?: Finding influential nodes in social networks based on interests. *J Supercomput* 78(2):2098–2124
12. Balaji TK, Sekhara C, Annavarapu R, Bablani A (2021) Machine learning algorithms for social media analysis: a survey. *Comput Sci Rev* 40:100395
13. Bao ZK, Liu JG, Zhang HF (2017) Identifying multiple influential spreaders by a heuristic clustering algorithm. *Phys Lett Sect A Gen Atom Solid State Phys* 381(11):976–983
14. Basaras P, Katsaras D, Tassioulas L (2017) Identifying influential spreaders in complex multilayer networks: a centrality perspective 4697(c):1–15
15. Batagelj V, An O ( m ) (2002) Algorithm for cores decomposition of networks. pp 1–9
16. Bigonha, C, Cardoso, TNC, Moro, MM, Gonçalves, MA, Almeida, VAF (2012) Sentiment-based influence detection on Twitter pp 169–183
17. Bonacich P (2007) Some unique properties of eigenvector centrality. *Soc Netw* 29(4):555–564
18. Borgatti SP, Everett MG (2006) A Graph-theoretic perspective on centrality. *Soc Netw* 28(4):466–484
19. Boroujeni RJ, Soleimani S (2022) The role of influential nodes and their influence domain in community detection: an approximate method for maximizing modularity. *Exp Syst Appl* 202(May):117452
20. Brin S, Page L (2012) Reprint of: the anatomy of a large-scale hypertextual web search engine. *Comput Netw* 56(18):3825–3833
21. Chai W, Xu W, Zuo M, Wen X (2013) ACQR: a novel framework to identify and predict influential users in micro-blogging
22. Chen W, Cheng S, He X, Jiang F (2012) InfluenceRank: an efficient social influence measurement for millions of users in microblog
23. Chen X, Li J (2018) Overlapping community detection by node-weighting. *ACM Int Conf Proc Ser (March)*:70–74
24. Chiroque LF, Chiroque LF (2021) A dissertation submitted by in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Mathematical Engineering. (September)
25. *Communication Patterns* (2023) C7. pp 725–730
26. Corsi N, Nguyen D-d, Butaney M, Majdalany SE, Corsi MP, Malchow T, Piontkowski AJ, Trinh Q-d, Loeb S, Abdollah F (2022) Top 100 urology influencers on twitter: is social media influence associated with academic impact? Top 100 urology influencers on twitter: is social media influence associated with academic impact ? *Eur Urol Focus* (October)
27. Cossu JV, Dugue N, Labatut V (2015) Detecting real-world influence through twitter. In: *Proceedings - 2nd European network intelligence conference, ENIC 2015*, pp 83–90
28. Ding Z-y, Jia Y, Zhou B, Han Y, He L, Zhang J-f (2013) Measuring the spreadability of users in microblogs \*. *14(9)*:701–710
29. Elbaghazaoui BE, Amnai M (2022) Data profiling and machine learning to identify influencers from social media platforms. *10*:201–218
30. Gammoudi F, Sendi M, Nazih M (2022) A survey on social media influence environment and influencers identification. *Soc Netw Anal Min*
31. Garza HDL, Maymone MBC, Vashi NA (2021) Impact of social media on skin cancer prevention
32. Ghosal D, Majumder N, Poria S, Chhaya N, Gelbukh A (2019) DialogueGCN: a graph convolutional neural network for emotion recognition in conversation. 2
33. Gong X, Yu H, Yu T (2023) Literature review on the influence of social networks. *SHS Web Conf* 153:01009
34. Gross J, Wangenheim FV (2022) Influencer marketing on instagram: empirical research on social media engagement with sponsored posts influencer marketing on instagram: empirical research on social media. *J Interact Advert* 0(0):1–12



35. Gu J, Abrams LC, Broniatowski DA, Evans WD (2022) An investigation of influential users in the promotion and marketing of heated tobacco products on instagram: a social network analysis
36. Guenon N, Hunter DS (2020) Detecting bots and assessing their impact in social networks. [arXiv:1810.12398v5](https://arxiv.org/abs/1810.12398v5) [cs . SI]
37. Guruprasad S, Mathias VL, Dcunha W (2021) Heart disease prediction using machine learning techniques. In: 2021 5th International conference on electrical, electronics, communication, computer technologies and optimization techniques, ICEECCOT 2021 - Proceedings, pp 762–766
38. Hafiene N, Karoui W, Ben Romdhane L (2020) Influential nodes detection in dynamic social networks: a survey. *Exp Syst Appl* 159:113642
39. Han M, Li Y (2018) Influence analysis: a survey of the state-of-the-art. *Math Found Comput* 1(3):201–253
40. He X, Wang X, Wang M (2023) LightGCN: simplifying and powering graph convolution network for recommendation
41. Hidri A (2021) Towards inferring influential facebook users
42. Hinton GE (2006) Communicated by Yann Le Cun A fast learning algorithm for deep belief nets 500 units 500 units. 1554:1527–1554
43. Huang X, Chen D, Wang D, Ren T (2020) Identifying influencers in social networks. *Entropy* 22(4)
44. Ibnoulouafi A, El Haziti M (2018) Density centrality: identifying influential nodes based on area density formula. *Chaos Solitons Fractals* 114:69–80
45. Ignat O, Yu JA, Halevy A (2021) Detecting inspiring content on social media
46. Jabeur LB, Tamine L, Boughanem M (2012) Active microbloggers: identifying influencers, leaders and discussers in microblogging networks. pp 111–117
47. Jeelani Z, Gani G, Qadir F (2023) Linear cellular automata-based impulse noise identification and filtration of degraded images. *Signal Image Vid Proc*
48. Jiang J, Wilson C, Wang X, Sha W, Huang P, Dai Y, Zhao BY (2013) Understanding latent interactions in online social networks. *ACM Trans Web* 7(4)
49. Kao L-j (2015) Mining influential users in social network
50. Karoui W, Hafiene N, Ben L (2022) Machine learning - based method to predict influential nodes in dynamic social networks. *Soc Netw Anal Min*
51. Karoui W, Hafiene N, Ben L (2023) Exploring modified areas for updating influential nodes. *Inf Syst* 112:102132
52. Khanday HA, Ganai AH, Hashmy R (2018) Users in online social networks. In: 2018 International conference on soft-computing and network security (ICSNS), pp 1–6
53. Khanday AMUD, Khan QR, Rabani ST (2021) Identifying propaganda from online social networks during COVID-19 using machine learning techniques. *Int J Inf Technol (Singapore)* 13(1):115–122
54. Kim ES (2009) An analytical way to find influencers on social networks and validate their effects in disseminating social games. 2008(December 2008)
55. Kumar S, Mallik A, Panda BS (2023) Influence maximization in social networks using transfer learning via graph-based LSTM. *Exp Syst Appl* 212(October 2021)
56. Kumar S, Singhla L, Jindal K, Grover K (2021) IM-ELPR: influence maximization in social networks using label propagation based community structure. pp 7647–7665
57. Kumar S, Panda BS (2020) Identifying influential nodes in social networks: neighborhood coreness based voting approach. *Phys A Stat Mech Appl* 553:124215
58. Li J, Cai T, Deng K, Wang X, Sellis T, Xia F (2020) Community-diversified influence maximization in social networks. *Inf Syst* 92(August)
59. Li H, Shang Q, Deng Y (2021) A generalized gravity model for influential spreaders identification in complex networks. *Chaos Solitons Fractals* 143:110456
60. Li W, Li Y, Liu W, Wang C (2022) An influence maximization method based on crowd emotion under an emotion-based attribute social network. *Inf Proc Manag* 59(2):102818
61. Liao H, Mariani MS, Medo M, Zhang YC, Zhou MY (2017) Ranking in evolving complex networks. *Phys Rep* 689:1–54
62. Liu D, Chen X (2011) Rumor propagation in online social networks like twitter - a simulation study. In: Proceedings - 3rd international conference on multimedia information networking and security, MINES 2011, IEEE, pp 278–282
63. Liu N, Li L (2007) Identifying domain-dependent influential microblog users: a post-feature based approach. (1):3122–3123
64. Liu Y, Wei X, Chen W, Hu L, He Z (2021) A graph-traversal approach to identify influential nodes in a network. *Patterns* 2(9):100321
65. Lü L, Chen D, Ren XL, Zhang QM, Zhang YC, Zhou T (2016) Vital nodes identification in complex networks. *Phys Rep* 650:1–63

66. Mao G-j (2016) A pagerank-based mining algorithm for user influences on micro-blogs
67. Mei Y, Zhong Y, Yang J (2015) Finding and analyzing principal features for measuring user influence on twitter. (March 2006):478–486
68. Mislove A, Marcon M, Gummadi KP, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. In: Proceedings of the ACM SIGCOMM internet measurement conference, IMC, ACM Press, New York, pp 29–42
69. Mithe S (2019) Deep learning on graphs using graph convolutional networks
70. Mnasri W, Azaouzi M, Romdhane LB (2021) Parallel social behavior-based algorithm for identification of influential users in social network. *Appl Intell* 51(10):7365–7383
71. Morone F, Makse HA (2015) Influence maximization in complex networks through optimal percolation. *Nature* 524(7563):65–68
72. Nafees L, Cook CM, Nikolov AN, Stoddard JE (2021) Can social media in fl uencer ( SMI ) power in fl uence consumer brand attitudes? The mediating role of perceived SMI credibility. *Digit Bus* 1(2):100008
73. Naga Srinivasu, P, Panigrahi, R, Singh, A, Bhoi, AK (2022) Probabilistic buckshot-driven cluster head identification and accumulative data encryption in WSN. *J Circ Syst Comput* 31(17)
74. Newman MEJ (2007) Course 8 complex networks. *Les houches summer school proceedings* 85:309–342
75. Ngo D-t, Cao C-n, Hoang P-l, Ngo X-b, Nguyen T-d, Tran D-a (2020) Identifying micro-influencers on social media using user graph construction approach
76. Nouh M, Nurse JRCC (2016) Identifying key-players in online activist groups on the facebook social network. In: Proceedings - 15th IEEE international conference on data mining workshop, ICDMW 2015, pp 969–978
77. Open Access (2012) A k-shell decomposition method for weighted networks
78. Panchendrarajan R, Saxena A (2022) Topic-based influential user detection: a survey. *Appl Intell* (May)
79. Peng S, Zhou Y, Cao L, Yu S, Niu J, Jia, W (2018) Influence analysis in social networks: a survey. *J Netw Comput Appl* 106(2017):17–32
80. Pham H, Shahabi C (2016) Spatial influence - measuring followship in the real world. In: 2016 IEEE 32nd international conference on data engineering, ICDE 2016, pp 529–540
81. Praas J (2020) Predicting influence spread in online social networks using combinations of node centralities. pp 1–10
82. Qiu J, Tang J, Ma H, Dong Y, Wang K, Tang J (2023) DeepInf: social influence prediction with deep learning
83. Rübiger S, Spiliopoulou M (2015) A framework for validating the merit of properties that predict the influence of a twitter user. *Exp Syst Appl* 42(5):2824–2834
84. Rani S (2022) Influential node detection and ranking with fusion of heterogeneous social media information. pp 1–23
85. Razis G, Anagnostopoulos I, Zeadally S (2020) Modeling influence with semantics in social networks: a survey. *ACM Comput Surv* 53(1)
86. Rezaie B, Zahedi M, Mashayekhi H (2020) Measuring time-sensitive user influence in Twitter. *Knowl Inf Syst* 62(9):3481–3508
87. Riquelme F, González-Cantergiani P (2016) Measuring user influence on Twitter: a survey. *Inf Proc Manag* 52(5):949–975
88. Saeidi M, Milios E, Zeh N (2021) Graph Convolutional Networks for Categorizing Online Harassment on Twitter. pp 946–951
89. Salve AD, Mori P, Guidi B, Ricci L, Pietro RD (2021) Predicting influential users in online social network groups. *ACM Trans Knowl Discov Data* 15(3):1–50
90. Sarkar D, Kole DK, Jana P (2016) Survey of influential nodes identification in online social networks. *Int J Virt Commun Soc Netw* 8(4):57–69
91. Sarna S, Aditi G, Musfique D, Iqbal A (2022) Query-oriented topical influential users detection for top-k trending topics in twitter. pp 13415–13434
92. Science Nonlinear, Phenomena Complex, Jain S, Sinha A (2020) Chaos, solitons and fractals identification of influential users on Twitter: a novel weighted correlated influence measure for Covid-19. *Chaos Solitons Fractals: Interdisciplinary J Nonlinear Sci Nonequilibrium Complex Phenomena* 139:110037
93. Science Nonlinear, Phenomena Complex, Jain S, Sinha A (2020) Identification of influential users on Twitter: a novel weighted correlated influence measure for Covid-19. *Chaos Solitons Fractals* 139:110037
94. Shafiq MZ, Ilyas MU, Liu AX, Radha H (2013) Identifying leaders and followers in online social networks. *IEEE J Sel Areas Commun* 31(9):618–628
95. Silva A, Guimarães S, Meira Jr, W, Zaki M (2023) ProfileRank: finding relevant content and influential users based on information diffusion
96. Singh SS, Singh K, Kumar A, Biswas B (2019) ACO-IM: maximizing influence in social networks using ant colony optimization. *Soft Computing*

97. Sudar KM, Ashish PD (2022) TFAD: TCP flooding attack detection in software-defined networking using proxy-based and machine learning-based mechanisms. *Clust Comput* :4
98. Talukder A, Alam MGR, Tran NH, Niyato D, Park GH, Hong CS (2019) Threshold estimation models for linear threshold-based influential user mining in social networks. *IEEE Access* 7(July):105441–105461
99. Tiwary A, Kumar M (2022) Determination Of category - wise influential users using information retrieval technique from twitter. 13(3):2313–2327
100. Torino PDI (2022) Micro Influencer Classifier : an academic and economic approach. (April)
101. Wang F, She J, Ohyama Y, Wu M (2019) Deep-learning-based identification of influential spreaders in online social networks. In: *IECON 2019 - 45th Annual Conference of the IEEE industrial electronics society*, vol 1, pp 6854–6858
102. Wang N, Sun Q, Zhou Y, Shen S (2016) A study on influential user identification in online social networks. *Chinese J Electr* 25(3):467–473
103. Wang Q, Zhuang D, Xie H (2021) Identification of influential nodes for drone swarm based on graph neural networks. *Neural Process Lett* 53(6):4073–4096
104. Wei B, Liu J, Wei D, Gao C, Deng Y (2014) Weighted k-shell decomposition for complex networks based on potential edge weights. *Phys A*
105. Wen T, Deng Y (2020) Identification of influencers in complex networks by local information dimensionality. *Inf Sci* 512(xxxx):549–562
106. Weng J, Weng J, Lim E-p, Jiang J (2010) Institutional Knowledge at Singapore Management University TwitterRank: finding topic-sensitive influential Twitterers TwitterRank: finding topic-sensitive influential Twitterers
107. Yamaguchi Y, Takahashi T (2010) TURank: Twitter user ranking based on user-tweet graph analysis. pp 240–253
108. Yin Z, Zhang Y (2012) Measuring pair-wise social influence in microblog. In: *Proceedings - 2012 ASE/IEEE International conference on privacy, security, risk and trust and 2012 ASE/IEEE international conference on social computing, SocialCom/PASSAT 2012*, pp 502–507
109. Yu E, Chen D, Fu Y, Xu Y (2022) Identifying critical nodes in complex networks by graph representation learning
110. Yuan S, Zeng H, Zuo Z, Wang C (2023) Overlapping community detection on complex networks with graph convolutional networks. *Comput Commun* 199(April 2022):62–71
111. Yuan S, Zeng H, Zuo Z, Wang, C (2023) Overlapping community detection on complex networks with graph convolutional networks. *Comput Commun* 199(April 2022):62–71
112. Zareie A, Sheikahmadi A, Jalili M (2019) Identification of influential users in social networks based on users' interest. *Inf Sci* 493:217–231
113. Zareie A, Sheikahmadi A, Jalili M (2020) Identification of influential users in social network using gray wolf optimization algorithm. *Exp Syst Appl* 142:112971
114. Zareie A, Sheikahmadi A, Jalili M, Fasaie MSK (2020) Finding influential nodes in social networks based on neighborhood correlation coefficient. *Knowl-Based Syst* 194:105580
115. Zeng A, Zhang C-j (2013) Ranking spreaders by decomposing complex networks. *Phys Lett A* 377(14):1031–1035
116. Zhang Y-c, Yeung CH, Zhou T, Lu L (2011) Leaders in social networks, the delicious case. 6(6)
117. Zhang Z, Li X, Gan C (2020) Identifying in fl uential nodes in social networks via community structure and in fl uence distribution difference. (April)
118. Zhang Z, Li X, Gan C (2020) Identifying in fl uential nodes in social networks via community structure and in fl uence distribution difference. (March)
119. Zhang C, Li W, Wei D, Liu Y, Li Z (2022) Network dynamic GCN influence maximization algorithm with leader fake labeling mechanism. pp 1–9
120. Zhang W, Yang J, yu Ding X-y, mei Zou X-m, yu Han H-y, chao Zhao Q-c. (2019) Groups make nodes powerful: identifying influential nodes in social networks based on social conformity theory and community features. *Exp Syst Appl* 125:249–258
121. Zhang T, Yang K, Ji S, Ananiadou S (2023) Emotion fusion for mental illness detection from social media: a survey. *Inf Fus* 92(November 2022) :231–246
122. Zhao G, Jia P, Huang C, Zhou A, Fang Y (2020) A machine learning based framework for identifying influential nodes in complex networks. *IEEE Access* 8:65462–65471
123. Zhao G, Jia P, Zhou A, Zhang B (2020) Neurocomputing InfGCN: identifying influential nodes in complex networks with graph convolutional networks. *Neurocomputing* 414:18–26
124. Zhou J, Zhang Y, Cheng J (2012) Preference-based mining of top- K influential nodes in social networks. *Fut Gener Comput Syst*

125. Zhou S, Blazquez M, McCormick H, Barnes L (2021) How social media influencers' narrative strategies benefit cultivating influencer marketing: tackling issues of cultural barriers, commercialised content, and sponsorship disclosure. *J Bus Res* 134(May):122–142
126. Zhu Y, Tang J, Tang X (2020) Pricing influential nodes in online social networks. *Proc VLDB Endowment* 13(10):1614–1627
127. Zhuang Y-b, Li Z-h, Zhuang Y-j (2021) Heliyon Identification of influencers in online social networks: measuring influence considering multidimensional factors exploration. 7(March)

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

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