INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

REVIEW ARTICLE

Check for updates

OPEN ACCESS

Received: 12-04-2023

Accepted: 01-06-2023

Published: 27-06-2023

Citation: Rashid Y, Iqbal Bhat J (2023) Unlocking the Power of Social Networks with Community Detection Techniques for Isolated and Overlapped Communities: A Review. Indian Journal of Science and Technology 16(25): 1857-1871. https://doi.org/10.17485/IJST/v16i25.841

* **Corresponding author**.

javaid.iqbal@iust.ac.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Unlocking the Power of Social Networks with Community Detection Techniques for Isolated and Overlapped Communities: A Review

Yasir Rashid[1], Javaid Iqbal Bhat[2]*

1 Research Scholar, Department of Computer Science Islamic University of Science & Technology, Kashmir
2 Associate Professor, Department of Computer Science Islamic University of Science & Technology, Kashmir

Abstract

Background: Social Network Analysis is a prominent field of research that captures the attention of numerous data mining expert. Social networks are visualized as network graphs, and identifying communities involves the identification of densely connected nodes. The exploration of community detection in online social networks is an essential field of research. This review paper presents an extensive examination of the latest methodologies and approaches utilized for isolated and overlapped community detection specifically in online social networks. **Objective:** To provide a comprehensive overview of the existing literature on community detection techniques for social networks with isolated and overlapped communities. The review aims to identify the key challenges associated with community detection in such networks and to review the various algorithms and methods that have been proposed to address these challenges. Additionally, this review intends to compare the performance of different community detection techniques on networks with isolated and overlapped communities, and to highlight their strengths and weaknesses. Ultimately, the goal of the review is to provide researchers and practitioners with a better understanding of the current state of the art in community detection for social networks, and to help guide future research in this important area. **Methods:** A comprehensive literature search was conducted on quality databases viz. Scopus, Web of Science, IEEE Xplore, and Science Direct using relevant keywords, and the selected articles were screened for inclusion/ exclusion based on their titles, abstracts, and keywords. The following are some search keywords that can be used for searching community detection papers: Community detection, Graph clustering, Network partitioning, Modularity optimization, Community structure, Graph-based clustering, Network community detection, Overlapping communities, Community detection algorithms, and Evaluation metrics for community detection. Comparative analysis was performed on the chosen articles, considering factors such as algorithms, methodologies, datasets, and evaluation metrics. The results were presented using comparative tables to present the findings. The methodology ensured a comprehensive review of recent literature on community detection, providing valuable insights and trends in the field. **Findings:** The review identified various algorithms and methods that have been proposed to address these challenges, including modularity optimization, spectral clustering, label propagation, and network embeddings. The paper also found that many of these algorithms can be adapted to handle networks with isolated and overlapped communities, but there is no single algorithm that is universally effective. The review highlighted that the performance of community detection techniques can be influenced by several factors, such as network size, density, and community structure. The paper found that some algorithms perform better than others in certain scenarios and that a combination of methods may be necessary to achieve optimal results. The review found that there are several challenges associated with community detection in such networks, including the presence of noise, sparsity, and overlapping communities. **Novelty:** This paper presents several novel contributions to the field of community detection in social networks. One key novelty of the paper is its comprehensive review of the existing literature on community detection techniques for networks with isolated and overlapped communities, which provides a valuable resource for researchers and practitioners working in this area. Another novelty of the paper is its identification of the key challenges associated with community detection in such networks, and its discussion of the various algorithms and methods that have been proposed to address these challenges. The paper also presents a comparison of the performance of different community detection techniques on networks with isolated and overlapped communities, which can help guide future research in this area. Finally, the paper proposes several new approaches for community detection in social networks with complex community structures, which can help advance our understanding of social networks and their potential as powerful tools for communication, collaboration, and information sharing. Overall, the paper's novel contributions make a significant contribution to the field of community detection in social networks.

Keywords: Social network analysis; community detection; influential user; overlapping communities; dynamic networks

1 Introduction

The term "community detection" refers to a technique for locating clusters of network nodes that are more heavily interconnected than the remainder of the network as a whole. Online social networks (OSNs) have emerged as a ubiquitous aspect of modern society, generating vast amounts of data that enable researchers to investigate social behavior and network structure[(1)]. A fundamental task in network analysis is community detection, which involves identifying groups of nodes that are densely connected within themselves but sparsely connected to the rest of the network. Communities offer valuable insights into the structure and function of networks, including the identification of key actors and the detection of information diffusion patterns. As such, community detection has become an important research area in social network analysis. This review paper provides a comprehensive overview of community detection techniques, with a particular focus on their application in online social networks [2]. The paper begins by introducing the concept of community detection and its significance in network analysis. The challenges associated with community detection, such as the resolution limit problem and the detection of overlapping communities, are then discussed. Next, the paper provides a detailed review of community detection methods, including hierarchical clustering, modularity-based approaches, and probabilistic models. The strengths and limitations of each approach are discussed, along with their applications in various fields. Finally, recent advances in community detection, such as the use of deep learning and multi-layer networks, are highlighted, and suggestions for future research directions are provided. Overall, this review paper offers a valuable resource for researchers and practitioners in network analysis and related fields who are interested in community detection in online social networks [3]. Most existing reviews on community detection concentrate on static community detection and relevant methods, which limit the ability to analyze network evolution. Moreover, the entities within communities exhibit cross-substitutional and sequential characteristics. Traditional approaches to community detection, such as spectral clustering and statistical inference, are being overshadowed by deep learning techniques that show remarkable capability in handling high-dimensional graph data with superior performance. Previous reviews of community detection methods have mainly focused on static community detection and discussed related methods. However, these reviews do not cover the latest developments in community detection techniques, such as the use of deep learning and multi-layer networks, which have demonstrated impressive performance in handling high-dimensional graph data [4]. To address these gaps, this review paper aims to provide a comprehensive overview of community detection techniques with a focus on their application in online social networks. We will discuss the challenges associated with community detection, such as the resolution limit problem and the detection of overlapping communities. We will also provide a critical evaluation of existing community detection methods, highlighting their strengths and limitations, and their applications in various fields [5]. Furthermore, we will discuss recent advances in community detection, such as the use of deep learning and multi-layer networks, and their potential to revolutionize community detection in online social networks. The review will conclude with a discussion of future research directions, highlighting the need for developing new methods to handle the evolving challenges in community detection. Overall, this paper aims to serve as a valuable resource for researchers and practitioners interested in community detection in online social networks. Below in Figure 1 organization of community detection techniques is discussed.

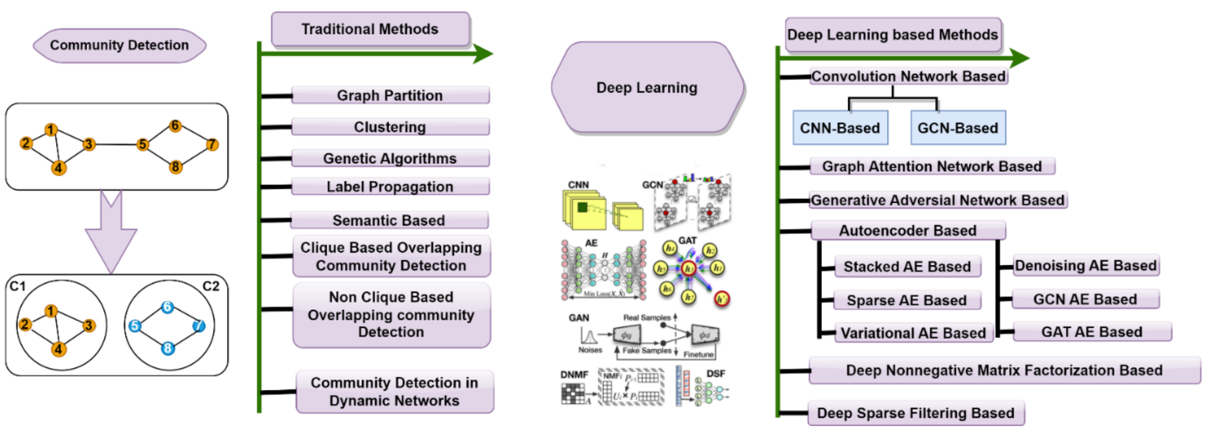

Fig 1. An Organizational Structure for Community Detection Techniques Based on Both Classical and Deep Learning

The purpose of this review on community detection techniques in online social networks is to contribute new insights to the existing literature by focusing on recent advancements in the field. While there have been previous reviews on community detection in social networks, they often lack coverage of the latest techniques and methods that have emerged in recent years [6]. Additionally, we address gaps in the literature by discussing the challenges associated with community detection, such as the resolution limit problem and the detection of overlapping communities, and how these challenges have been addressed with new approaches. Moreover, this review paper provides a comprehensive and current analysis of various community detection techniques, including hierarchical clustering, modularity-based approaches, and probabilistic models [7]. We also highlight recent advancements in the field, such as the application of deep learning and multi-layer networks, and discuss their potential applications in online social networks. Through this review, we aim to provide a valuable resource for researchers and practitioners in network analysis and related fields who are interested in the latest developments and applications of community detection in online social networks. Community recognition has many potential foremost submissions in many fields, which may include: Social network analysis, Biology, Marketing, Cybersecurity, Online communities, Political science, Fraud detection, Environmental science, Recommendation systems, Health care. TimeLine for Community Detection Techniques is discussed below in Figure 2.

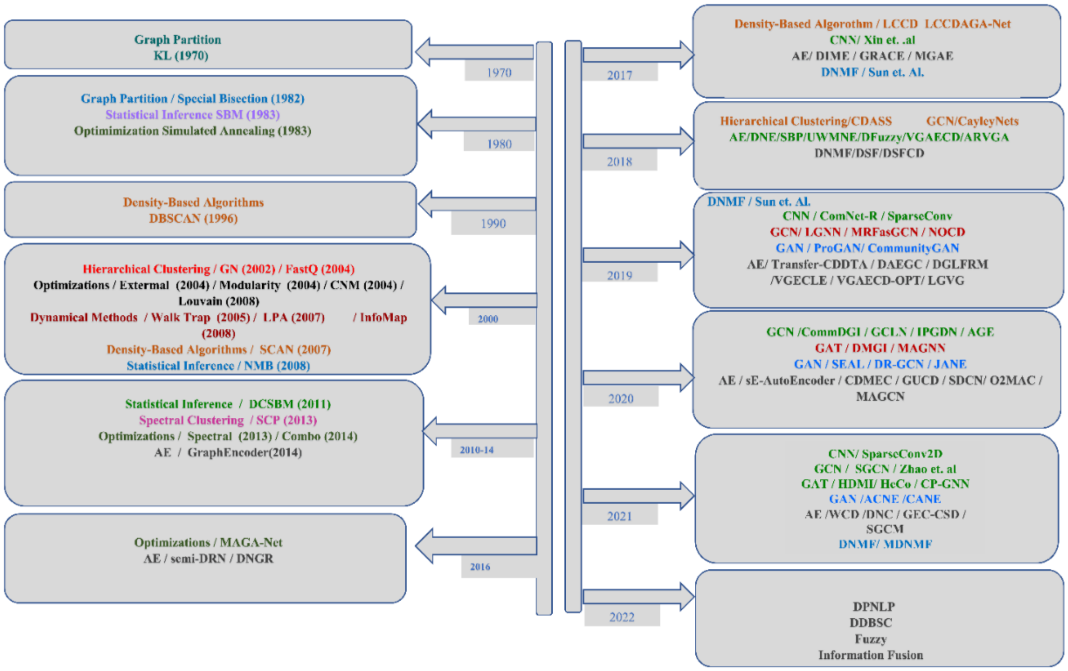

Fig 2. Community Detection Timeline

This paper provides a comprehensive review of community detection techniques for both isolated and overlapped communities by investigating and analyzing the current research status of dynamic community detection. A significant amount of work has been done in community detection using both traditional and deep learning approaches however we found that influential user detection in online social networks in overlapped communities can be further explored for new opportunities by applying deep learning solutions like Graph Convolution Networks (GCNs).

2 Isolated Community Detection Techniques

There are several methods of grouping similar items into communities, depending on the nature of the items and the characteristics that are used to define similarity[7]. Here are a few examples:

2.1 Graph Partitioning-based Community Detection

Graph partitioning is a type of community detection that involves dividing a graph into a set of disjoint subgraphs, or partitions. The goal of graph partitioning is to decrease the count of edges that cross between the different partitions, while still preserving the overall structure of the graph.

There are several algorithms and approaches that can be used for graph partitioning, including:

- **Spectral clustering:** This approach uses eigenvectors from graph Laplacian to identify clusters in the graph, which can then be used to create the partitions.
- **Kernighan-Lin algorithm:** This is a heuristic algorithm that iteratively improves the partitioning of the graph by moving nodes from one partition to another in an attempt to reduce the number of edges crossing between partitions.
  Metis: This is a set of graph partitioning tools that use a variety of methods, including multilevel partitioning and recursive bisection, to partition graphs.
- **Other approaches:** There are many other algorithms and tools available for graph partitioning, including the Fiduccia-Mattheyses algorithm, the Multilevel Recursive Bisection algorithm, and others.

It is important to note that graph partitioning is an NP-hard problem, that is computationally difficult to find the optimal solution for large graphs. As a result, most graph partitioning algorithms are heuristics that aim to find good, but not necessarily optimal, solutions.

2.2 Clustering-based detection of Communities

Clustering, a subset of community detection, is the process of organizing the nodes of a graph into subgroups called communities based on the relationships between the nodes [8]. There are several clustering algorithms that can be used for community detection :

- **K-means clustering:** This technique clusters observations around a central point, or centroid, and then divides those observations into k subsets. At first, each data point is placed in the cluster that contains its centroid, and then the algorithm repeatedly updates the centroid such that in a cluster it reflects the average of all the data points.
- **Hierarchical clustering:** In this method, clusters are organized into a hierarchy, with each level indicating a more granular breakdown of the data [9]. Agglomerative hierarchical clustering places each data point in its own cluster and then merges them together; divisive hierarchical clustering places all the data in one cluster and then separates it into smaller groups.
- **DBSCAN:** This is a density-based algorithm that identifies clusters by looking for regions of high density in the data. It works by starting at a point with a high density of points and expanding outwards to include all points that are within a specified distance, or Eps, of the starting point.
- **Other approaches:** There are many other clustering algorithms available, including spectral clustering, affinity propagation, and more.

It is important to note that the choice of clustering algorithm will be subject to the features of the data and the specific goals of the analysis. Some algorithms are better suited to certain types of data or certain types of clustering tasks, so it is important to carefully consider which algorithm is the most appropriate for a given situation.

Table 1 summarizes some of the high-tech clustering algorithms for community finding.

Table 1. Clustering algorithms for community detection

Author Name and Algorithm used	Method	Features	Obtainability of code
Newman & Girvan	Alienating Clustering	Betweenness (Edge-Betweenness)	Yes
Newman	Principal method which is known as Modularity-maximization.	eigenvalue and eigenvector , Modularity	Yes Yes Yes
Clauset e al	Modularity of greedy optimization	Boundaries, apexes and Modularity.	Yes
Blondel et al(Louvain Method)	Method of clustering as hierarchical.	Users/Nodes, edges/Links, Modularity	Yes
Guimera et al, Zhou et al	Simulated Annealing	total sum of links, probability of linking , sum of modules, total quantity of dividers, Modularity quantity of edges, factor within and factor outside within , and Modularity.	No
Duch et al	Extremal optimization via modular optimization.	Node numeral, edges, Modularity and degree.	Yes
Ye et al (AdClust)	Method of Clustering as agglometrive.	Pinnacles, Force and Modularity.	No
Wahl and Sheppard	clustering by means of hierarchal fuzzy spectral.	Modularity of fuzzy Similarity of jaccard.	No
Falkowski et al (DENGRAPH)	Clustering through density-based method.	Functions pertaining to distance.	No
Dongen et al (MCL)	Type of clustering Via Markovian procedure.	Total number of nodes.	Yes
Nikolaev et al	clustering primarily based on centrality of entropy.	The probability of transition matrix related to the Markov process.	No
Steinhauser et al	Common methods include simple clustering and a walk known as Random walk.	Likeness matrix and also span of random walks.	No

2.3 The identification of communities using genetic algorithms

The principles of natural evolution serve as the base for a specific class of optimization algorithms known as genetic algorithms (GAs). They are often employed in place of more conventional techniques to handle optimization problems of extreme scale or complexity. For the purpose of network community discovery, GAs may be used to determine the optimal node-clustering scheme for maximising some criterion. To apply a GA to the task of community discovery, an objective function must be defined to measure how well one particular method of node clustering performs. This may depend on the number of inter-cluster connections or the density of connections inside each cluster(10). The GA would produce a population of potential solutions (partitioning's of the nodes) and use evolutionary operators like cross-over and mutation to create new solutions from the current population in order to seek for the partitioning of the nodes which optimises this objective-function. The GA would keep doing this until it either discovers a good answer or reaches a stopping point. An benefit of using GAs is that they can identify high-quality solutions even if the structure of the communities is not well specified, especially when dealing with big, complicated networks. However, based on the network size and the complexity of the goal function, they may be quite computationally costly and may take a long time to operate(11). The aim of the heuristic search technique acknowledged as a genetic algorithm (GA) is to find the best possible solution under certain parameters. When developing a genetic algorithm, the fitness function is first established for a pool of candidate solutions, or chromosomes. If the best solution has been determined, the procedure finishes; otherwise, the current set of solutions is subjected to crossover and mutation operators to build a new set. Detecting communities may be seen as an optimization issue, with the goal being to maximise neutral function that best arrests the intuitive sense that a community has more internal connections than outward ones. GA has been used in the study and analysis of communities in a number of recent studies. These are covered briefly below. Table 2 provides an overview of GA-based community identification methods available in the literature.

Table 2. Genetic algorithms for community detection

Name of Author along with algorithm used	Method	Features Employed for analysis	Availability of code
Pizzuti (GA-Net)	The community-score as function of aptness.	Score of community.	Yes
Pizzuti (MOGA-Net)	optimization of multiple objectives.	Major techniques known as Community-score and Community-fitness.	Yes
Hafez et al	Sole-objective, many optimization objectives.	Gene count, Crossover operators and mutation.	No
Mazur et al	Range of Community and fitness functions as modularity.	Appropriateness function.	No
Liu et al	Clustering and also procedure of genetic algorithm.	Population size, maximal number of generations, sum of the communities and maximum figured generations used for the unimproved fittest chromosome fraction in mined-hubs.	No
Tasgin et al	Basic technique known as Optimization of modularity.	Chromosome count, population size and modularity	No
Zadeh	Cultural algorithm of multiple population.	BS_ average, BSN	No

3 Localization of communities via the spread of labels

Label Propagation involves propagating labels (or "votes") between nodes in a network. It works by initializing the nodes with a unique label (such as a randomly assigned integer), and then iteratively updating the labels of the nodes based on the labels of their neighbors. The algorithm stops when the labels of the nodes have converged, at which point the nodes with the same marker are considered to be part of the matching community. Label propagation has the advantage of being simple to implement and relatively fast, especially for large networks. However, it can be sensitive to the initialization of the labels and may not always find the optimal partitioning of the nodes into communities. It is also limited in the types of community structures that it can detect, as it tends to find communities that are connected and relatively homogeneous in terms of their connectivity patterns(12). In a network, label propagation occurs when a label is broadcast to several nodes. In this method, a node takes on the label of its nearest neighbors. Community discovery methods that rely on the dissemination of labels are discussed here. The methods listed in Table 3 are addressed in depth later in the section.

Table 3. Community detection using label propagation

Name of the Author(Algorithm used)	Methods	Applications of these algorithms	Features/Parameters	Availability of the Code
Raghavan et al (LPA)	Label propagation using Iterative method	SLPA WLPA COPRA LabelRank BMPLA	Labels and Nodes Threshold and Nodes Similarity and Labels Nodes	Yes Yes Yes
Xie et al (LabelRank)	Propagation, Inflation, cut- off, conditional-update.	LabelRankT	Coefficient of Belongingness, and threshold, nodes	No
Wu et al (BMLPA)	Methods using Label Propagation and those communities who are Overlapped	_	No of Nodes, labels to which nodes belong, Average-degrees	No

4 Semantics based community detection

Semantics-based community detection refers to the use of semantic information (such as the meaning of words or the relationships between concepts) to recognize clusters of nodes that are related to each other in some way within a network. This can be useful for finding communities of nodes that share a common theme or topic. There are several approaches to semantics-based community detection, liable on the type of data being analyzed and the explicit areas of the analysis[13]. Some common methods include:

- **Latent semantic analysis (LSA):** This is a method of analyzing the relationships between words in a document based on their co-occurrence patterns. It can be used to identify groups of words that are semantically similar, and therefore potentially belong to the same community.
- **Latent Dirichlet allocation (LDA):** This is a probabilistic prototypical that may use to discover the topics underlying in the documents collection. It can be used to identify communities of documents that are related to each other by the topics that they cover.
- **Conceptual network analysis:** This is a method of analyzing the relationships between concepts in a domain based on their co-occurrence patterns. It can be used to identify groups of concepts that are semantically similar, and therefore potentially belong to the same community.

There are many other approaches to semantics-based community detection, and the prime of system will hang on the exact characteristics of the facts and the goals of the analysis.

5 Techniques for Detection of Overlapped Community

The existence of overlapping communities in networks may be identified using a number of techniques. In order to optimize the modularity score, one of the most common strategy is to divide the network into smaller communities. The modularity score compares how many edges there are between communities to how many edges there are within them.

Overlap community detection algorithms: For overlapping community detection in a network, there are a number of methods developed. Three such algorithms are the OSLOM, COPRA, and SLPA.

Clique percolation method: Clique identification and merger to generate communities is the core of this approach. Cliques are completely linked subgraphs in a network. Size-distinct communities may be discovered by adjusting the number of members in each clique.

Role-based community detection: In this approach, communities are defined by the responsibilities that individual nodes perform in the network (such as "broker," "gatekeeper," etc.). Nodes in a network may be clustered using methods like k-means clustering, and then allowed to belong to numerous clusters at once via a process called "overlap node clustering" (i.e., overlapping communities).

Mixed-membership stochastic block model: This model is a probabilistic method for finding overlapping groups in networks. For this method to function, a statistical model must be fitted to the network topology, with each node corresponding to a certain probability distribution across the group of communities. The unique qualities of the network and the nature of the inquiry dictate which technique is best suited for a given application.

5.1 Detection of overlapping communities based on Cliques

Clique percolation is a technique for discovery lapping communities in networks that relies on cliques. Locating and combining network cliques (completely interconnected subgraphs) yields new communities [14]. Using the clique percolation technique, overlapping communities may be identified by doing the following:

1. The first step is to map out every group dynamic in the system. Cliques are completely linked subgraphs in which all of the nodes have links to every other node in the graph.
2. Combine groups of people that share at least one node. As a result, the overlapping communities in the network are represented by a new set of merged cliques.
3. Keep going until there are no more cliques to combine.

You may discover communities of varying sizes by adjusting the size of the cliques you take into account. You may, for instance, just look for 3-node cliques to identify tiny communities, or you might look for bigger cliques to identify huge communities. Clique percolation has the benefit of being easy to put into practise. For networks with a huge quantity of nodes or a immense number of communities, it may not perform as well as some of the other approaches for discovering overlapping communities. Assembling smaller, fully connected (complete) subgraphs that have common nodes may be thought of as a community. A fully allied subgraph of size k is called a "k-clique." Together, all k-cliques that may be stretched from any other k-clique over a connected set of surrounding k-cliques form a k-clique community. Several researchers have utilized cliques as a tool for studying overlapping groups.

The technique of Clique Percolation (CPM) by Palla et al. to identify touching groups. Using an algorithm developed by Everett et al. in 1983, this method determines communities via component breakdown of clique-clique intersection matrices [15]. Time complexity for CPM is O(exp(n)). However, the CPM developed by Palla et al. failed to recognise the underlying hierarchical structure and overlapping nature of the network. Lancichinetti et al. devised a method to address this issue. Each node's community is determined by a local probe. The method may be repeated as many as necessary to examine the nodes again. The foremost aim was to treasure utmost in a function of fitness. The CFinder84 software for detecting overlapping communities was developed using CPM. To discover overlapping communities by means of maximal cliques, Du et al. created ComTector (Community DeTector). First, we locate all most cliques in the web; they will serve as the hubs for our future community. The remaining vertices are then aggregated into their closest kernels using an iterative agglomerative method. Clusters are attuned via combining pairs of tiny communities to increase the network's modularity. Where C is the over-all sum of found communities and T is the total number of triangles in the network, (C T2) is the total amount of time required to complete the procedure. EAGLE was suggested by Shen et al. and is based on agglomerative hierarchical clustering. Those cliques with a maximum size below a certain cutoff are weeded out in the first phase [16]. Minimalist Maximal Cliques Below. The Community Level Are Ignored (also the subordinate vertices). Similar communities are constantly fused together on the foundation of their defined similarities. This process is recurrent until only ace community leftovers. As Evans et al. argued, identifying overlapping communities in a network requires first dividing its links. Using weighted line graphs, Evans et al.78 broadened the scope of this investigation. As part of a different article, Evans used clique graphs to locate clusters of similar individuals in extensive social networks. Network cliques are identified and then expanded upon using GCE (Greedy Clique Expansion). The procedure by Girvan-split-betweenness Newman served as the foundation for this method. The runtime of procedure is O(m3). The CONGO81 (CONGA Optimized) approach, originally reported in a different publication, improved upon the original CONGA's O(nlog) complexity by using a local betweenness measure. When looking for overlapping communities, Gregory82 suggests a two-stage Peacock approach that makes use of several methods for identifying groups. The author's earlier proposed concept of split betweenness was implemented at the first stage of network transformation. Second, we apply a disjoint community detection method to the modified network and transform the resulting communities into the original network's overlapping communities.

5.2 Detection of Overlapping Communities based on Non clique methods

Methods that don't rely on cliques are used to identify overlapped communities. This includes, but is not limited to:

In command to optimize modularity score, a mutual policy is to divide the network into smaller communities. The modularity score compares how many edges there are between communities to how many edges there are within them. Algorithms for Overlapped community detection: To recognize overlapped communities in a grid, there are a number of methods developed. Three such algorithms are the OSLOM, COPRA, and SLPA [16]. Detecting communities depend on the functions that nodes perform inside a network (such as "broker," "gatekeeper," etc.) is known as role-based community detection. Nodes in a network may be clustered using methods like k-means clustering, and then allowed to belong to numerous clusters at once via a process called "overlap node clustering" (i.e., overlapping communities). A probabilistic model that may be used to identify overlapping communities in networks is the mixed-membership stochastic block model. For this method to function, a statistical model must be fitted to the network topology, with each node corresponding to a certain probability distribution across the group of communities. The unique qualities of the network and the nature of the inquiry dictate which technique is best suited for a given application.

6 Community Detection for Dynamic Networks

Community discovery refers to the problem of finding groups of nodes over time that are more closely connected to one another than to the rest of the network. It is possible to recognize communities in dynamic networks using a variety of techniques, such as:

- **Dynamic extension of static community detection methods:** Some approaches for detecting communities in static networks may be adapted for use in dynamic ones by applying the procedures at numerous time periods and taking the communities' development over time into account. By applying it at different times in time and keeping track of nodes' mobility across communities over time, the Louvain method, for instance, may be adapted to operate on dynamic networks.
- **Temporal modularity maximization:** The goal of this strategy is to maximize the temporal modularity score by dividing the network into communities at each time instant. The temporal modularity score evaluates the community's consistency over the time by quantifying the number of edges both inside and between them at each time point.
- **Link streaming algorithms:** These algorithms transmit the network's edges as they arrive and update the communities instantly. Both the Temporal Clustering technique and the Dynamic Dendrogram algorithm are examples of link streaming algorithms.
- **Dynamic mixed-membership stochastic block models:** It is possible to utilize these probabilistic models to spot communities in active networks. All they do is assign a probability distribution across the community set to each node inside network at each time point, fitting a statistical model to the network topology.

The unique properties of the dynamic network and the nature of the investigation dictate which technique is best suited to a given application.

7 Deep Learning Based Community Detection

In order to organise this review, we came up with a taxonomy for deep community identification techniques based on the most recognisable features of the deep learning models used. Convolutional networks, Graph attention networks(GAT), Generative adversarial networks(GAN), Autoencoders(AE), Deep nonnegative matrix factorization(DNMF), and Deep sparse filtering(DSF) are the six categories summed up in the taxonomy (DSF). There are two types of convolutional networks in this classification system: CNNs and graph convolutional networks (GCN). To represent latent characteristics for community detection, both make convolutions. GATs are noteworthy because they place a premium on picking up cues from the local community. By using an adversarial training procedure between the input graph and the fabricated samples, the GAN model has been effectively deployed for community identification. Graph convolutional AEs, graph attention AEs, and variational AEs are only few of the subcategories that may be found under the overarching AE framework (VAEs)[1].

7.1 Convolution Network Based Community Detection

Two models that have proven useful for community detection are convolutional neural networks (CNNs) and graphical convolutional networks (GCNs). To reduce computational costs and keep the CNN robust to feature representations, convolutional neural networks (CNNs) are a subtype of feedforward DNNs designed for grid-like topological data, such as image data. As their name implies, graph-centric networks (GCNs) are optimised for graph data structures. They are constructed using convolutional neural networks and can approximate spectrum filters at the first order. When it comes to a GCN, the layer-by-layer propagation rule is given in below equation I:

$$H^{(l+1)} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}w^{(l)}\right) \tag{1}$$

7.2 CNN Based Community Detection

Currently available approaches for community identification based on convolutional neural networks (CNNs) use CNN models with limited input data. That's why it's important to convert the graph data into a labelled picture format beforehand (see Figure 3). Table I of the Supplemental Material summarises the problems that are addressed by the methods described here, all of which are related to community detection. As an unsupervised learning problem, traditional community identification is well-suited to modern deep learning models. However, inadequate topological structures might have an impact on the neighbourhood analysis and thus lower the accuracy of community detection. Unfortunately, there is a lack of information on network architecture in the real world. Hence, the study presented the first supervised CNN-based community finding method for non-perfect networks (TINs). In order to represent the network and identify communities, a deep neural network (DNN) with two convolutional layers built using max-pooling operators is utilised. Intact latent characteristics are progressively recovered from simple inputs using the CNN architecture. At each node, convolutional kernels represent the local characteristics from a variety of perspectives, making up the convolutional layers.

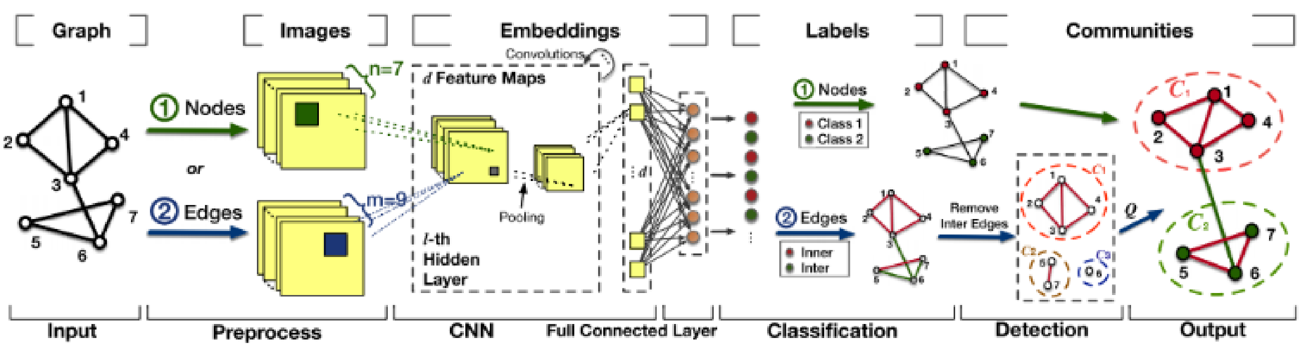

Fig 3. General framework for CNN-based community detection

7.3 GCN Based Community Detection

In order to universally capture complex properties for community identification, GCNs combine information about node neighbor's in deep graph convolutional layers (Figure 4). Communities may be identified using one of two distinct types of GCN-based techniques: 1) community clustering using unsupervised network representation, and 2) community categorization using supervised/semi-supervised methods. The absence of labels in the actual world is one of the main challenges for community categorization algorithms. Whereas methods like matrix reconstructions and objective optimizations are limited in their ability to group communities, network representations are more malleable in this regard. Common community discovery strategies may be used in GCNs as deep graph operators. One such model that enhances SBMs with superior community identification ability at lower computational cost is the line graph neural network (LGNN)[17]. LGNN learns the characteristics represented by nodes in directed networks by combining a non-backtracking operator with the message-passing principles of belief propagation.

Learning node embeddings in GCNs does not prioritise community structures since these networks were not intended for use in community identification. In order to fill this need, researchers at Google developed a semi-supervised GCN community identification model called MRFasGCN that uses an enhanced Markov random field (eMRF) as a new convolutional layer to describe hidden communities. Because of this, MRFasGCN is geared toward the community and can refine the rough GCN output smoothly. In order to identify the structural nodes for unsupervised community discovery, SGCN generates a local label sampling model. SGCN combines the label sampling model with GCN to encode both the network topology and node attributes, allowing for the training of each node's community membership without any prior label knowledge.

Within a framework of probabilistic inference, a generative model that infers the community affiliations of nodes is the solution to the problem of discovering overlapping communities. For instance, the neural overlapping community detection (NOCD) method learns community affiliation vectors by minimizing the Bernoulli-Poisson (BP) probabilistic model's negative log-likelihood. By establishing a threshold, we may keep identifying and eliminating weak connections, ultimately arriving at the communities that will be the basis for our final analysis. All hidden characteristics in a node's surroundings are reflected in a spectral GCN. Laplacian smoothing is continually executed in deep GCN layers until the characteristics of adjacent nodes

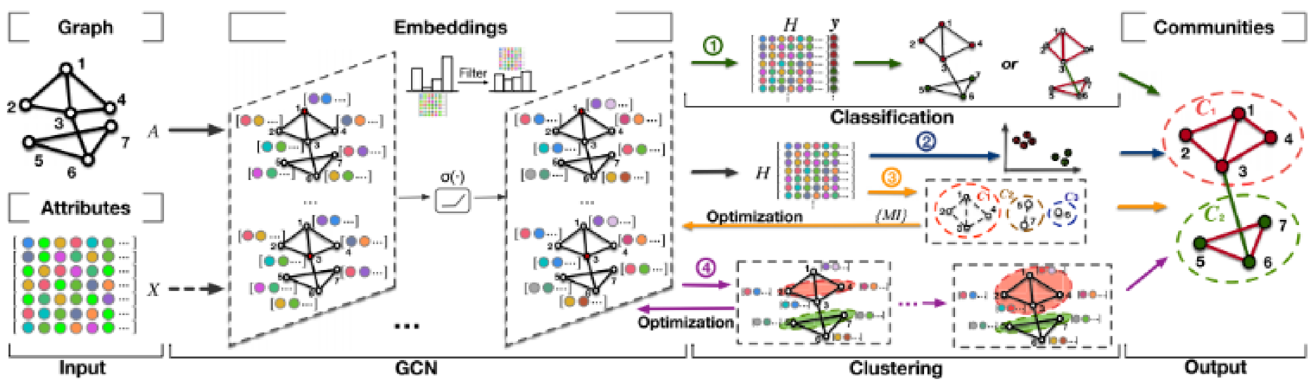

Fig 4. General framework for GCN-based community detection

converge to the same values. Over-smoothing occurs in community detection when using these models, however. Graph convolutional ladder-shape networks (GCLN) are a novel GCN architecture created to lessen the detrimental effects. Using a U-Net in the CNN domain, it provides unsupervised community discovery using k-means clustering [18].

In the GCLN, we construct a contracting route and an expanding way in a same fashion. Observations made along the narrower route are blended with the more granular knowledge gained along the wider one. Since GCNs represent every possible connection type and aggregate characteristics along them, GCNs have redundant representations since various connection types are often handled as simple edges. To automatically uncover the subtleties of a graph's independent latent properties through neighborhood routing, the Independence Promoted Graph Disentangling Network (IPGDN) segments the neighborhood into distinct portions. To improve its community identification capabilities, the model may then learn a more accurate disentangled representation of the target communities. Hilbert-Schmidt independence criteria (HSIC) regularization ensures that all latent embeddings are completely separate from one another When utilizing a GCN for community discovery in attributed graphs, neighboring nodes and nodes with similar characteristics tend to cluster together. Consequently, graph convolutions need to filter out high-frequency sounds when multiplying the aforementioned two graph signals. Adaptive graph convolution (AGC) combines a low-pass filter with spectral clustering through the below equation 2

$$P(\lambda_P) = \left(1 - \frac{1}{2}\right)^k \tag{2}$$

8 Graph Attention Based Community Detection

GAT (Graph Attention Network) is a type of neural network that can be used for community detection in networks. GAT is an extension of GCN (Graph Convolutional Networks) which allows to attend on different node's feature while computing the node representation. Discovering clusters of highly linked nodes, or "communities," in a network is the goal of community detection. Features of the network topology may be extracted using GAT networks, and then utilized to make predictions about the communities to which individual nodes belong. Traditional approaches may have difficulty finding communities in such huge and complicated networks, but our method may help. GATs use attention mechanism to decide the importance of the neighborhood nodes for a given node. The attention mechanism allows GAT to assign different weights to different neighbor's, which can be useful in cases where not all neighbors are equally important for the task at hand. Below in Figure 5 is explained how GAT is used for community detection [19].

9 Generative Adversial Network Based Community Detection

When it comes to identifying network communities, GAN (Generative Adversarial Networks) is an unorthodox approach. Generic Adversarial Networks (GANs) are a kind of neural network used for generating content, most often images and texts. Generally speaking, GANs have two components: a generator network and a discriminator network. The discriminator attempts to tell apart the created data from the genuine data while the generator creates the new data. In adversarial training, the generator network attempts to trick the discriminator network with fabricated data, while the discriminator network seeks to accurately identify the created data. While GANs are not typically used for community detection, it is possible to adapt the

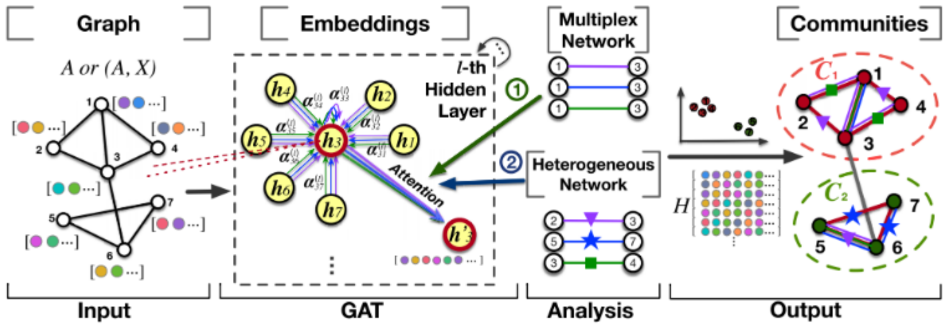

Fig 5. General framework for GAT-based community detection

GAN architecture for this task. For example, one could train a GAN to generate graphs with specific community structures and use the discriminator to identify the communities in the generated graphs. However, this is an active area of research and there is not yet a widely accepted method for using GANs for community detection. Below in Figure 6 shows how GAN's are used for detecting communities[20].

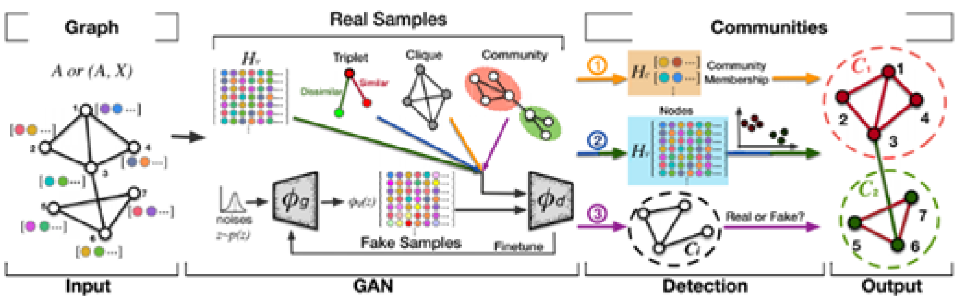

Fig 6. General framework for GAN-based community detection

10 Auto-Encoders for Community Detection

Autoencoders (AEs) are neural networks that is utilized for community detection in networks. Autoencoders are an unsupervised learning approach, which means that it doesn't require labelled data. Learning a lower-dimensional representation of the input data is the goal of an AE, which is an unsupervised learning approach. A lower-dimensional representation of a network's structure may be learned using autoencoders, which can subsequently be utilized for community detection. The idea is that the lower-dimensional representation should group together nodes that belong to the same community, while separating out nodes that belong to different communities. The output of the AE is a feature vector for each node, which can then be used as input for a clustering algorithm to group the nodes into communities. In this approach, the autoencoder is trained to reconstruct the graph structure by encoding it into a lower-dimensional representation, and then decoding it back to the original structure. This process of training the autoencoder forces the network to find the most relevant and salient features of the graph, which can be used for community detection. Note that this approach presupposes that nodes form communities because they have more similarities with one another than with the rest of the network[21].

11 Community Detection Based on Deep Nonnegative Matrix

To better discern the community structure of intricate networks, researchers have developed a deep learning variation of NMF called Deep Nonnegative Matrix Factorization (DNMF). In DNMF, a neural network architecture is used to learn low-dimensional representations of the network, which may be used to locate groups of nodes sharing common characteristics. It has been proven that this method is successful in detecting communities in vast, sparse networks, and it may be used for a broad variety of purposes, such as in social networks, gene regulatory networks, and brain networks[22].

12 Community Detection Based on Deep Sparse Filtering

Using deep neural networks and sparse filtering methods together, Deep Sparse Filtering (DSF) is a method for community discovery in networks. Unlike supervised methods, DSF may be used without the need of labelled inputs. To develop a low-dimensional representation of the network's structure with the help of a deep neural network is the core principle underlying DSF. Then, a sparse filtering method takes this reduced-dimensional representation as input, allowing it to identify communities by clustering together nodes that have more similarities with one another than with the rest of the network. The deep neural network used in DSF is a type of autoencoder, which is trained to reconstruct the network's structure by encoding it into a lower-dimensional representation and then decoding it back to the original structure. This process forces the network to learn the most salient features of the graph, which can be used for community detection. Sparse filtering is a technique for extracting features from the data by selecting a small number of the most informative and representative features. In DSF, sparse filtering is applied on the low-dimensional representation obtained from the autoencoder. The sparse filtering algorithm groups the nodes that have similar features together and identifies them as a community. DSF has been applied to many other kinds of networks, including social networks, biological networks, and citation networks, and has been found to be successful at recognizing communities in these settings [23].

13 Standard Datasets for Community Detection

According to Table 4, the repositories utmost used for tentative investigations in community detection study may be classified into genuine datasets and fake (made) datasets. It is usual practice to test community detection algorithms on a variety of different standard datasets. Some examples include [24]

1. LFR Benchmark Networks: To appraise the efficacy of community identification algorithms, researchers have created artificial networks. The community structures of the networks are well-understood, allowing for fine-grained regulation of factors like the amount of overlap and the range of fluctuation in community sizes.

2. The New York City Social Circle: This is a social network of over 1,500 nodes representing individuals in the New York City area, and was collected by Brian Uzzi and Jarrett Spiro in the 2000s. The network has several communities, corresponding to different social circles.

3. The Enron Email Corpus: This is a large dataset containing over half a million emails from over 150 employees of the Enron Corporation [25]. The dataset has been used to construct a network of communication between the employees, and the network has several communities corresponding to different departments within the company.

4. The Political Blogs Corpus: This is a dataset containing over 12,000 blog posts from over 200 political blogs. The dataset has been used to construct a network of hyperlinks between the blogs, and the network has several communities corresponding to different political viewpoints [26].

These datasets can be useful for testing and comparing the concert related to diverse community detection algorithms, and therefore for evaluating the robustness of the algorithms to various types of network structures.

Below in Table 4 are some the Datasets for community detection.

Table 4. Datasets for community detection.

Type	Dataset name	Link for download
Real dataset	ZCC(Zachary-karate-club)	Yes
	Dataset of Dolphin	Yes
	Dataset of American-college football-Network	Yes
	Dataset of Southern women	Yes
Benchmark dataset	Girvan-Newman	Not Available
	Dataset known as LFR (Lancichinetti, Fortunato,Radicchi)	Yes

14 Conclusion and Future Scope

The topic of community identification offers a fantastic chance for finding new groups in the ever-expanding social networks of today. Finding communities with algorithms in real social networks like Twitter, Facebook, LinkedIn, etc., may provide large amounts of data that can find application in a wide variety of fields. Insights like this might have practical applications in the realms of commerce, learning, and development. This work not only details the datasets used by existing algorithms but also addresses future social network community finding applications. To sum up, detecting communities inside social media platforms is a hot topic in the academic world right now. Traditional strategies as well as newer ways developed to deal with the peculiarities of online social networks have been presented as potential solutions to the issue. While there has been a lot of progress made, there are still a number of obstacles to overcome. These include the fact that large-scale networks are still a problem, as well as the necessity to include new forms of data (such text or location data) and the management of dynamic networks. To further expand our knowledge of community structure in online societal networks and towards creating more effective approaches for recognizing isolated and overlapping groups, future studies will likely continue to concentrate on these and related challenges. Moreover, the detection of overlapping communities can lead us to detection of influential users that not only influence a single community but multiple communities. In this study, we have identified several potential research gaps in the field of community detection in online social networks. These gaps include addressing scalability issues, incorporating temporal dynamics, handling overlapping communities, incorporating multiple types of relationships, addressing bias issues, and incorporating domain knowledge. Future studies could focus on exploring these areas and developing new techniques and methods to overcome these challenges in community detection.

References

1) Huang X, Chen D, Wang D, Ren T. Identifying Influencers in Social Networks. *Entropy*. 2020;22(4):450. Available from: https://doi.org/10.3390/e22040450.
2) Zhang W, Yang J, Ding XY, Zou XM, Han HY, Zhao QC. Groups make nodes powerful: Identifying influential nodes in social networks based on social conformity theory and community features. *Expert Systems with Applications*. 2019;125:249–258. Available from: https://doi.org/10.1016/j.eswa.2019.02.007.
3) Bioglio L, Pensa RG. ScienceDirect Impact of Neighbors on the Privacy of Individuals in Impact of Neighbors the Privacy of Individuals in Online on Social Networks Impact of Neighbors on the Privacy of Individuals in Online Social Networks Online Social Networks. *Procedia Comput Sci*. 2017;108:28–37. Available from: https://doi.org/10.1016/j.procs.2017.05.030.
4) Nouh M, Nurse JRC. Identifying Key-Players in Online Activist Groups on the Facebook Social Network. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. 2015. Available from: https://doi.org/10.1109/ICDMW.2015.88.
5) Sarkar D, Kole DK, Jana P. Survey of Influential Nodes Identification in Online Social Networks. *International Journal of Virtual Communities and Social Networking*. 2016;8(4):57–69. Available from: https://doi.org/10.4018/IJVCSN.2016100104.
6) Nigam S, Agrawal C, Jain A. Determining Influential User(S) In Online Social Networks. 2016. Available from: https://www.researchgate.net/publication/315847435_DETERMINING_INFLUENTIAL_USERS_IN_ONLINE_SOCIAL_NETWORKS.
7) Zareie A, Sheikhahmadi A, Jalili M. Identification of influential users in social networks based on users' interest. *Information Sciences*. 2019;493:217–231. Available from: https://doi.org/10.1016/j.ins.2019.04.033.
8) Leung CK, Cameron JJ. DIFSoN : Discovering Influential Friends from Social Networks. 2012. Available from: https://doi.org/10.1109/CASoN.2012.6412389.
9) Kao LJ, Huang YP. Mining Influential Users in Social Network. *2015 IEEE International Conference on Systems, Man, and Cybernetics*. 2015. Available from: https://doi.org/10.1109/SMC.2015.216.
10) Arularasan AN, Suresh A, Seerangan K. Identification and classification of best spreader in the domain of interest over the social networks. *Cluster Computing*. 2019;22(S2):4035–4045. Available from: https://doi.org/10.1007/s10586-018-2616-y.
11) Alshahrani M, Fuxi Z, Sameh A, Mekouar S, Huang S. Efficient algorithms based on centrality measures for identification of top-K influential users in social networks. *Information Sciences*. 2020;527:88–107. Available from: https://doi.org/10.1016/j.ins.2020.03.060.
12) Hafiene N, Karoui W, Romdhane LB. Influential nodes detection in dynamic social networks: A survey. *Expert Systems with Applications*. 2020;159:113642. Available from: https://doi.org/10.1016/j.eswa.2020.113642.
13) Molaei S, Farahbakhsh R, Salehi M, Crespi N. Identifying Influential Nodes in Heterogeneous Networks. *Expert Systems with Applications*. 2020;p. 113580. Available from: https://doi.org/10.1016/j.eswa.2020.113580.
14) Praas J. Predicting influence spread in Online Social Networks using combinations of node centralities. 2020. Available from: https://purl.utwente.nl/essays/80574.
15) Jain L, Katarya R, Sachdeva S. Opinion leader detection using whale optimization algorithm in online social network. *Expert Systems with Applications*. 2020;142:113016. Available from: https://doi.org/10.1016/j.eswa.2019.113016.
16) Jain S, Sinha A. Identification of influential users on Twitter: A novel weighted correlated influence measure for Covid-19. *Chaos, Solitons & Fractals*. 2020;139:110037. Available from: https://doi.org/10.1016/j.chaos.2020.110037.
17) Zareie A, Sheikhahmadi A, Jalili M. Identification of influential users in social network using gray wolf optimization algorithm. *Expert Systems with Applications*. 2020;142:112971. Available from: https://doi.org/10.1016/j.eswa.2019.112971.
18) Shafiq MZ, Member S, Ilyas MU, Liu AX, Radha H. Identifying Leaders and Followers in Online Social Networks. 2019. Available from: https://web.cs.ucdavis.edu/~zubair/files/jsac_leader_follower_2013.pdf.
19) Fortunato S, Castellano C. Community Structure in Graphs. *Computational Complexity*. 2012;2012:490–512. Available from: https://doi.org/10.1007/978-1-4614-1800-9_33.
20) Plantié M, Crampes M. Survey on Social Community Detection. *Computer Communications and Networks*. 2013;p. 65–85. Available from: https://doi.org/10.1007/978-1-4471-4555-4_4.
21) Chen J, Yuan B. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*. 2006;22(18):2283–2290. Available from: https://doi.org/10.1093/bioinformatics/btl370.
22) Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*. 2009;11(3):033015. Available from: https://doi.org/10.1088/1367-2630/11/3/033015.
23) Evans TS, Lambiotte R. Line graphs, link partitions, and overlapping communities. *Physical Review E*. 2009;80(1). Available from: https://doi.org/10.1103/PhysRevE.80.016105.
24) Alotaibi N, Rhouma D. A review on community structures detection in time evolving social networks. *Journal of King Saud University - Computer and Information Sciences*. 2022;34(8):5646–5662. Available from: https://doi.org/10.1016/j.jksuci.2021.08.016.

25) Sangaiah AK, Rezaei S, Javadpour A, Zhang W. Explainable AI in big data intelligence of community detection for digitalization e-healthcare services. *Applied Soft Computing*. 2023;136:110119. Available from: https://doi.org/10.1016/j.asoc.2023.110119.

26) Jiang L, Shi L, Liu L, Yao J, Ali ME. User interest community detection on social media using collaborative filtering. *Wireless Networks*. 2022;28(3):1169–1175. Available from: https://doi.org/10.1007/s11276-018-01913-4.