Deployment of Information Diffusion for Community Detection in Online Social Networks: A Comprehensive Review

Soumita Das and Anupam Biswas

***Abstract*— The flow of information through active users in online social networks (OSNs) plays a major role in forming natural social groups, popularly known as *communities*. Although structural and topological aspects of the network had been central to most of the community detection approaches, incorporation of information flow for community detection has been an emerging topic in the recent past. Often, the flow of information is studied as a traceable process called *information diffusion*. The flow of information in the network affects various factors like temporal characteristics, network attributes, or social attributes. The information diffusion process helps to extract this information including where and when information is generated and in what fashion the dispersion occurs. Thus, it has the potential to aid the community detection process in social networks. In this article, the deployment of the information diffusion process for community detection has been studied extensively. The study is mainly focused on how information flow affects various network properties and social facets and explored the possibility of deployment for community detection. Various information diffusion models and community detection algorithms have been discussed in the context of network properties and social facets. Current challenges, future directions, and modalities for the deployment of information diffusion in community detection have been discussed. In addition, various widely used datasets, evaluation metrics as well as evaluation methods for evaluating community detection algorithms are also detailed.**

***Index Terms*— Community detection, disjoint communities, information diffusion, online social networks (OSNs), overlapping communities.**

I. Introduction

THE online social networking platforms (such as Facebook, Twitter, Whatsapp, and Wechat) plays an important role in peoples' daily life ranging from expressing views/opinions, sharing photos/videos, and spreading information [1], [2]. Earlier people used to receive information passively but with the recent development of communication technologies and online social networks (OSNs), individuals act as early publishers and active communicators [3]. OSNs have facilitated the platform to exchange information among individuals. Thus, information now reaches fast to the people via active communicators in the OSN. This process of spreading information across the network is called information diffusion. Although the active communicators are crucial for information diffusion, it depends on several other aspects of a network like different social facets and properties of the network [4].

Information diffusion problem has been studied extensively and numerous mathematical models have been developed for effective diffusion in the network [5]–[11]. These studies show that the diffusion process reveals several useful details like where and when information are generated and in what fashion the dispersion occurs, how people behave while sharing information, etc. These inherent details have the potential to assist in solving other problems in OSN, such as prediction [12], recommendation systems [13], community evolution [14], and most importantly the community detection, where the objective is to group people in the network [15]. For instance, the distinction of contrasting or similar behavior of people sharing information during the diffusion process is useful in community detection. Conversely, information diffusion is also benefited by the outcome of community detection. The interdependence of both information diffusion and community detection problem is quite evident from different literature [16]–[21].

An in-depth study is necessary to understand and explore the interdependence between *information diffusion* and *community detection*, which are very potent topics of social network analysis. The problem of information diffusion is basically to spread information/ idea/ innovation to the target group of objects [22]. On the other hand, the goal of community detection is to partition the network into groups or clusters or modules called communities such that the vertices within a community are densely connected and vertices among the communities are sparsely connected [23]. Although these topics imply two distinct problems in the social networking domain, recently, research has been going to understand how the essence of one can be used to interpret the other optimally. This survey mainly focuses on the deployment of information diffusion for community detection to understand how it is being utilized for community detection. The following key aspects of deploying information diffusion to detect communities are considered for our study.

1) The flow of information in the network affects various factors like temporal characteristics, network attributes, or social attributes. Therefore, the study mainly focuses on how information flow affects various network properties and social facets and explained how these could possibly be deployed for community detection.
2) The trace of propagated information in the network reveals several useful details like where and when the

Manuscript received July 23, 2020; revised March 30, 2021; accepted April 25, 2021. Date of publication May 17, 2021; date of current version September 30, 2021. *(Corresponding author: Soumita Das.)*

The authors are with the Department of Computer Science and Engineering, National Institute of Technology Silchar, Silchar 788010, India (e-mail: wingsoffire72@gmail.com; anupam@cse.nits.ac.in).

Digital Object Identifier 10.1109/TCSS.2021.3076930

TABLE I

COMPREHENSIVE STUDY OF NETWORK PROPERTIES AND SOCIAL FACETS

OSN Facets	Network Property		Social Facet		Community Detection Approach
	Vertex Related	Edge Related	Vertex Related	Edge Related	
Connection Strength [21]	-	✓	-	-	LPA [19], Thresholding [21]
Value Strength [19]	-	✓	-	-	
Tie strength [24]	-	✓	-	-	
Belonging factor [21]	✓	-	-	-	Thresholding [21]
Degree centrality [25]	✓	-	-	-	K-means clustering [26], Two stage EM [27]
Closeness centrality[28]	✓	-	-	-	
Betweenness centrality[28]	✓	-	-	-	
Clustering coefficient [29]	✓	-	-	-	-
Cosine similarity [28]	-	✓	-	-	Combination of CRF and ICM[20]
Jaccard Coefficient formula [24]	-	✓	-	-	Thresholding [21]
Euclidean distance[28]	-	✓	-	-	Combination of GMM and MAP [18]
Contagion	-	-	✓	-	-
Common neighbors [16]	-	-	-	✓	Thresholding [16]
Homophily [28]	-	-	✓	-	-
Influential spreader [4]	-	-	✓	-	A new follow based community detection algorithm [30]
Topic [31]	-	-	✓	✓	-
Information difference [21]	-	-	✓	-	-
Common interests and traits [21]	-	-	✓	✓	Two stage EM [27], Thresholding [21]
Number of cascade [20]	-	-	✓	✓	Combination of CRF and ICM[20]
Length of cascade [18]	-	-	✓	✓	Combination of GMM and MAP [18]

information is generated and in what fashion the dispersion occurs. Various information diffusion models are discussed and explained how could possibly the trace of these models influence the various network properties and social facets.

3) Different community detection algorithms inherently consider network properties and social facets in the strategy of the algorithm. Therefore, the study explores how the incorporation of information diffusion is likely to influence community detection directly or indirectly affecting network properties and social facets.
4) Evaluation of communities is crucial to ensure accuracy and quality. The study explains various widely used evaluation metrics pertaining to measure both the accuracy and quality. This study also discussed various evaluation methods for evaluating community detection algorithms.

The rest of this article is organized as follows. Section II introduces various network properties and social facets from the viewpoint of information flow, Section III discusses various information diffusion models to understand their use in the community detection, Section IV briefs existing information diffusion approaches for detecting communities in OSN. Section V talks about various evaluation attributes to analyze the detected communities, and Section VI concludes by outlining the future scope of deploying information diffusion for community detection.

II. INFORMATION FLOW DEPLOYMENT

The formation of communities and their detection depends on several facets of OSN like temporal characteristics, network attributes, or social attributes. The flow of information in the network influences these facets at different levels. Thus, it is necessary to understand how information flow affects various facets of the OSN and its impact on community detection. The facets of OSN can be broadly categorized as: 1) network properties and 2) social facets. The network properties measure the structural or topological aspect of the network, while social facets quantify different facets or aspects of social interaction. In this section, various network properties, as well as social facets are discussed from the viewpoint of information flow and explained how these could possibly be deployed for community detection. The key points of the study are listed highlighting the community detection approaches linked to various OSN facets in Table I.

A. Network Properties

1) Edge Strength Measures: The relationship between two vertices sharing an edge is measured by various edge strength measures such as tie strength, connection strength, and value strength. The strength of the relationship is determined by the number and type of vertices affected during the diffusion process. In particular, we have assumed that the information diffusion is based on the similarity or affinity of the vertices under consideration. If the outcome of the diffusion refers to a set of vertices which are the common vertices associated with an edge, exploitation of the number of affected common vertices assists in the computation of edge strength measures such as tie strength and connection strength. It is because the tie strength is measured by Jaccard coefficient [24], [28] which is based on the number of common vertices shared by an edge and it has been discussed in detail in the latter part of this section. Also, the connection strength [21] measure depends on the number of common vertices shared by an edge and is defined as

$$\mathrm{CS}_{vu} = \frac{N(v) \bigcap N(u)}{T(v)}. \tag{1}$$

Equation (1) indicates that tie strength and connection strength of an edge is stronger when larger number of common vertices are affected during the diffusion process and weaker when less number of common vertices are affected during the diffusion process. Now, to understand the effect of information diffusion on the value strength of the inherent edges, let us consider a directed graph where the information diffusion has been incorporated. It is to be added here that vertices sharing an edge in a directed graph are indicated by the source vertex and target vertex. As the information diffusion proceeds, if the common vertices shared by an edge and the neighboring vertices of the target vertex of the edge get affected, the value strength gets influenced. This is because the value strength [19] measure is defined as

$$V_{(u\to v)} = \frac{\mid C_v \setminus C_u \mid}{\mid C_v \mid}. \tag{2}$$

Considering (2), it is evident that the value strength $V_{(u\to v)}$ of an edge is stronger when the diffusion process results in affecting less number of common vertices of the edge and a larger number of neighboring vertices of target vertex. It is inferred from this discussion that the higher the value of connection strength and tie strength between two connected vertices, the higher is the similarity between them. Here, higher value strength signifies dissimilarity between the connected vertices. Utilizing these similarity or dissimilarity information, community detection algorithms namely, label propagation approach (LPA) [19] and thresholding [21] may be applied to detect the inherent communities.

2) Belongingness Measures: The belongingness of a vertex to a certain community is measured by belonging factor or belonging degree. If we consider from the information diffusion point of view, the strength of belongingness of a vertex x to a community c is determined by the type of vertices affected during the diffusion process. Here, we have assumed that the diffusion approach is based on predictive graph-based models that consider similarity or affinity as the diffusion strategy. For example, let us assume that the diffusion process begins at a certain timestamp say $t-1$ and gradually the information proliferates affecting a certain set of vertices with the same information. Ultimately, at the end of the diffusion process, if the affected vertices are the neighboring vertices of x and as they are affected by the same information, we assumed their belongingness to a certain community c to be maximum. Then this information is utilized to determine the belonging factor of x to community c. This is because the belonging factor [19] of x is defined by

$$\mathrm{bf}_t(c,x) = \frac{\sum_{y\in\text{neighbors}\,(x))} \mathrm{bf}_{t-1}(c,y)}{\mid \text{neighbors}(x) \mid}. \tag{3}$$

Equation (3) indicates that, higher the number of neighboring vertices of x belonging to community c at timestamp $t-1$, higher is the belonging factor $\mathrm{bf}_t(c,x)$ of x to community c at timestamp t.

Now, to understand the effect of diffusion on the belonging degree, let us consider that the strategy of diffusion is to affect the vertices based on similarity, and vertices that are affected by the same information indicate their belongingness degree to the same community. To incorporate the diffusion process, both interaction and local topology is considered. At the end of the diffusion process, the number of neighboring vertices of a vertex affected by the same information is used to measure the belonging degree of vertex v. Belonging degree [21] is defined by

$$B(v,c) = \frac{1}{2}(\mathrm{BI}(v,c) + \mathrm{BT}(v,c)). \tag{4}$$

In (4), $\mathrm{BI}(v,c)$ and $\mathrm{BT}(v,c)$ indicate attribution coefficient and belonging coefficient, respectively, which are defined as follows [21]:

$$\mathrm{BI}(v,c) = \frac{\sum_{u\in(N(v)\bigcap c)} I_{\text{sum}(u\leftrightarrow v)}}{\sum_{u\in N(v)} I_{\text{sum}(u\leftrightarrow v)}} \tag{5}$$

$$I_{\text{sum}(u\leftrightarrow v)} = \sum_{t\in L}\left(I_{(u\leftrightarrow v)}(t) - I_{(u\leftrightarrow v)\_}\ \mathrm{cost}(t)\right) \tag{6}$$

$$\mathrm{BT}(v,c) = \frac{\mid N(v)\cap c \mid}{D_v}. \tag{7}$$

Here, as has already been mentioned, the diffusion is dependent on the flow of information throughout the network indicated by $\mathrm{BI}(v,c)$ in (5). It is quantified by the amount of information that propagates between vertices u and v indicated by the term $I_{\text{sum}(u\leftrightarrow v)}$ in (6). Additionally, the dependence of diffusion on the local topology is indicated by $\mathrm{BT}(v,c)$ in (7). Higher the number of neighboring vertices of a vertex say v affected by the same information during diffusion implies a higher probability of the v to be affected by that particular information.

Higher belongingness strength of a vertex to a certain community implies higher probability of the vertex to belong to that community and thereafter, to assign appropriate communities to the remaining inherent vertices, community detection approach namely, thresholding [21] may be utilized.

3) Centrality Measures: Rapid growth in the usage of OSNs has resulted in the availability of several information. But the ideas/information we adopt is highly dependent on the influential entities. These entities are used to activate the inactive vertices by utilizing their influence on the inactive vertices. In this discussion, we have covered three types of influential entity identification measures such as degree centrality, betweenness centrality, and closeness centrality. These entities are utilized in the propagation of information throughout the network to enhance the rate of diffusion. Suppose that a diffusion strategy is incorporated in a network based on pairwise relationship between vertices. If the outcome of the diffusion infers that a given piece of information is propagated to a large fraction of the vertices inherent in the network, then identification of the vertex that possess the maximum number of connections that lead to the propagation of information to those connections gives the degree centrality measure. It finds the most central vertex in terms of number of connections, and the degree centrality [25] is defined by

$$\sigma_D(x) = \sum_{i=1}^{n} a_{ix}. \tag{8}$$

Equation (8) indicates that vertex x optimizes the diffusion if it affects maximum number of its neighboring vertices through influence and this is known as degree centrality. Here, the neighboring vertices of x is represented by adjacency matrix $A = (a_{ix})$. Another diffusion strategy is based on the mean distance of a vertex to all other vertices. This is known as closeness centrality [28] and is defined by

$$C_i = \frac{n}{\sum_j d_{ij}}. \tag{9}$$

Equation (9) indicates that it is easier and faster reach other vertices from vertex i using geodesic path denoted by d_{ij} which gives the closeness centrality score. Next, the diffusion strategy may consider the vertices that lie in the shortest path between two vertices. This is known as betweenness centrality [28] which is defined by

$$x_i = \sum_{\mathrm{st}} n^i_{\mathrm{st}}. \tag{10}$$

Equation (10) indicates that vertex i assists in the propagation of information to the unconnected parts of the network if the number of geodesic paths from vertex s to vertex t that pass through vertex i is high. This is known as betweenness centrality.

It is to be mentioned here that vertices with high centrality form denser subgraphs, and hence, it is utilized by various clustering-based algorithms (e.g., K-means clustering [26]), community detection algorithms, and optimization algorithms [e.g., two-stage expectation maximization (EM) [27]] to predict the inherent communities.

4) Clustering Coefficient: The idea of information diffusion throughout the network is to propagate the information to all the inherent vertices present in the network in the least possible time and exploit some topological characteristics of the network during the diffusion process. One such property is the clustering coefficient. Let us suppose that the diffusion process in a network begins with a vertex say i having the maximum number of neighboring vertices. If the analysis of the outcome of the diffusion pattern indicates that neighboring vertices of i form cliques, it depicts the importance of i among its neighboring vertices and hence in information diffusion. This gives the local clustering [28] coefficient of i and is defined by

$$C = \frac{\text{number of pairs of connected neighbors of } i}{\text{number of pairs of neighbors of } i}. \tag{11}$$

Equation (11) indicates that if the neighboring vertices of i lead to the formation of cliques due to diffusion, it infers the extent of closeness of the neighboring vertices of i and quantifies the importance of i based on clustering coefficient of i. The clustering coefficient highly contributes to the community detection because it measures the extent to which vertices cluster together, and hence, plays a significant role in community detection.

5) Structural Equivalence: OSNs have witnessed tremendous growth recently and hence, there are millions and millions of information out there. But to choose particular information, reliability of the information plays a key role. This is where influence shows its action. Propagating information throughout a social network, if guided by influence is more reliable and reaches a wider set of vertices. This influence is depicted by information diffusion probability which is computed considering user interest, user attributes, and other relevant features. The selection of a certain feature for measuring information diffusion probability is given by similarity and dissimilarity measures. These measures are used to compute the structural equivalence. Our study covers three different types of structural equivalence measures namely cosine similarity, Jaccard coefficient, and Euclidean distance. For instance, if the diffusion requires the spread of information considering user similarity or interest, then the number of common vertices shared by two users indicates the extent of their similarity. Symmetric similarity measures, namely cosine similarity [28] and Jaccard coefficient [24], [28] have been used for similarity computation and is defined by

$$\sigma_{ij} = \cos\,\theta = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{jk}^2}} \tag{12}$$

$$\mathrm{JS}_{ij} = \frac{\mid \Gamma(i) \bigcap \Gamma(j) \mid}{\mid \Gamma(i) \bigcup \Gamma(j) \mid}. \tag{13}$$

Equations (12) and (13) implies that higher the number of common vertices shared by vertices i and j, higher is the diffusion probability value of the edge shared by i and j. These similarity measures used to depict the diffusion probability value are called cosine similarity and the Jaccard coefficient. Also, the proximity of the vertices under consideration can be computed using a dissimilarity measure where the higher the value of dissimilarity, the lower will be the information diffusion probability value. We have chosen a symmetric dissimilarity measure namely, Euclidean distance [28] for calculating dissimilarity between two vertices which is defined by

$$d_{ij} = \sum_k \left(A_{ik} - A_{jk}\right)^2. \tag{14}$$

Equation (14) indicates the number of neighboring vertices that differ between vertices i and j, i.e., the number of neighboring vertices of i that are not the neighboring vertices of j and vice versa. A higher number of dissimilar neighboring vertices indicates a low value of information diffusion probability. Ultimately, these similarity and dissimilarity results are utilized by several community detection approaches for identifying inherent communities. For instance, by exploiting the cosine similarity measure, communities have been detected using a combination of conditional random field (CRF) and iterated conditional mode (ICM) [20]. Similarly, the Jaccard coefficient formula has been utilized for community detection using thresholding [21]. Similarly, Zhang *et al.* [18] discuss the role of Euclidean distance in combination with stochastic measures, namely Gaussian mixture model (GMM) and MAP for predicting communities.

6) Boundary Vertex: OSN users tend to bond more with those users who are similar to them and are known to them. Hence, information sharing in OSNs whether directed (Twitter) or undirected (Facebook) is driven by similarity and topology. The similarity between users is based on interest, attributes, or other characteristics. Whereas, by topology, we refer to the extent to which neighboring users tend to bond with each other. During the diffusion process, a pair of neighboring vertices may be affected by two different types of information because of the affinity of the vertices to that information. The belongingness of the neighboring vertex/vertices of a vertex to different information is quantified by boundary vertex [21] which is defined by

$$\mathrm{BN}_c = v \in c \mid \Gamma(v) \cap c \neq \phi, \quad \Gamma(v) \nsubseteq c. \tag{15}$$

In (15), $\Gamma(\upsilon)$ indicates the set of neighbors of vertex υ including vertex υ. The outcome of the diffusion process leads to a set of structures where each structure consists of a group of vertices that shares some similarity or topology or a combination of both and hence, gets affected by the same information during the diffusion process. It is obvious that vertices in different structures possess different information and vertices in the same structure possess equal information. Suppose that υ belongs to a certain structure c, then for υ to be a boundary vertex, there exists at least one neighboring vertex of υ that does not belong to structure c. These boundary vertices are exploited by community detection approaches such as thresholding as has been discussed in [21] to detect communities. It demonstrates how the difference in information possessed by the boundary vertices is used for setting a threshold value which ultimately results in community identification.

B. Social Facets

1) Contagion: Following the trend has always been a characteristic of OSN users. A user's activity is influenced by the action of his/her neighbors. During the information diffusion process, the affinity of an unactivated vertex say υ to get activated highly depends on its neighboring vertices. This influence on υ by its neighboring vertices is termed as contagion. This implies that if υ has a higher degree and the maximum number of neighboring vertices of υ are affected by particular information at time $t-1$, then the probability of υ to be influenced by the same information at time t is enhanced. Whereas a low degree of υ implies less dependence on the neighboring vertices of υ for taking a decision.

Let us try to simplify the contagion concept with the following real-world example. Suppose, an individual has a good number of friends and most of them buy a new gadget and each of them boasts about it with the individual, then generally it influences the particular individual also. This influence on an individual by his/her friends is termed as a contagion. It is inferred from this example that the action taken by an individual directly depends on the activity of the maximum number of his/her friends. It is understandable from this discussion that contagion has a high potential to be utilized for community detection purposes. However, our study has not come across any work that exploits the contagion concept for community detection. This concept has been discussed to motivate researchers to utilize this social facet for identifying communities.

2) Common Neighbors: OSNs favor the formation of groups among those users sharing some similarity. These groups are determined by exchanging information throughout the OSN and aggregating similar vertices during the diffusion process. It is assumed that users sharing a high number of common neighbors are strongly related and are very similar. Hence, those users along with their common neighbors have a high affinity to get affected by the same information during the diffusion process. The outcome of the diffusion results in a dense subgraph formation. So, this common neighbor facet has a high probability to be utilized by various community detection algorithms for identifying the inherent communities in a network. An illustration of the same has been shown in [16] which describes how common neighbors and thresholding have been used for community detection purpose.

3) Homophily: Users have an inclination to associate with other like-minded users. By like-minded users, we refer to those users which have a tendency to interact with similar users. This similarity between users is quantified by homophily that is used in OSNs to spread information. To characterize the behavior of users in OSNs, we opted to model the interaction that takes place between similar individuals using the independent cascade (IC) model described in Section III. It is likely that if the diffusion is initiated based on homophily, then as diffusion proliferates through the network, a larger number of homophilic connections are affected.

For instance, let us assume that information diffusion is initiated in a friendship network taken from a university. If the information diffusion is based on the homophily property which considers department of the students as the homophily attribute, then if the outcome of the diffusion is a set of two separate groups, one consisting of a group of friends from computer science and engineering (CSE) department and another group consisting of friends from electronics and communication engineering (ECE) department, it implies that the friendship network is between the students from CSE and ECE department. This is how homophily is utilized by the information diffusion process to generate groups of similar vertices bounded by a common attribute. This homophily property in combination with other community detection approaches has a great potential to detect inherent communities. Our study did not encounter any work that exploits homophily using information diffusion for community detection. So, this facet has been discussed here to motivate researchers to work in this direction.

4) Influential Spreaders: Certain entities in OSNs are exploited to popularize content among the users of the OSN. In this context, content refers to those data that meet the information needs of the users. These entities that are utilized to popularize content possess one to many relationships in the network and are referred t as influential spreaders. These entities play a significant role to spread information throughout the network. Let us suppose that an information diffusion model has been incorporated in a social network and that the diffusion begins from a certain vertex inherent in the network, then the probability of success of the starting vertex to activate the neighboring inactive vertices is based on the influence of the starting vertex on the neighboring inactive vertices. The higher the influence of the starting vertex on its neighboring inactive vertices, the higher is the diffusion probability.

A real-world example has been presented here to further simplify the concept of the influential spreader. For instance, a microblog posted by a celebrity may attract more followers to comment and re-post [32] thereby spreading the content throughout the network. In this case, the celebrity acts as the influential spreader. It is very much prominent that the diffusion process can give rise to many influential vertices. Once these influential vertices have been detected, they attract a good number of adherent vertices [33], [34]. The distance between the influential vertices plays a significant role in effective community detection and the same has been discussed in [4]. Another approach of detecting influential spreaders has been proposed by Yazdani *et al.* [30] where utilization of (25) and (26) resulted in a chain of interconnected vertices where each chain represents a community and this is further accompanied by a new follow-based community detection algorithm.

5) Topic: In OSNs, the spreading pattern may vary with respect to the topic opted and the dynamics of the underlying diffusion network. This dynamics is predicted by observing the diffusion speed of a particular advertisement on an OSN. For instance, if the underlying network is related to a political group, then advertisement about a political compaign will enhance the diffusion speed. Whereas, outdoor equipment promotions would influence the mountaineering group. So, it is inferred from these examples that the diffusion speed of a certain topic in a particular network is used to identify the dynamics of the underlying network, and hence, the affinity for a particular topic in that network is detected. Higher diffusion speed for a particular topic in a certain network indicates that the underlying network has a high affinity for that topic.

Moving further, it is to be mentioned here that some of the vital attributes of a topic are persistence and stickiness. In particular, politically controversial topics have been found to be persistent and using these topics for diffusion results in more exposure. In addition to that, information diffusion approaches can be adopted by various companies to advertise new products, services, and so on to a certain section of visitors. This results in stickiness of the companies with their visitors. These aspects of a topic have a high potential to be utilized by community detection approaches for the extraction of better communities.

6) Information Difference: Analyzing the inherent characteristics of a network by simulating information exchange is a significant task. Here, information exchange between vertices inherent in an OSN has been exploited to predict the implicit affinity of a vertex for particular information. Let us suppose that the diffusion process is initiated by a certain source vertex. Then, the probability of the source vertex infecting its connected vertices is based on the similarity between the source vertex and its neighboring vertices. In this context, similarity indicates people with a common interest or people who are known to each other. Thereafter, tracing the outcome of information diffusion indicates the existence of numerous subgroups of vertices where the vertices in different subgroups possess different information volume and vertices in the same subgroup possess equal information volume. The difference in information between the vertices belonging to different subgroups indicates that they belong to different communities. The use of the information difference for community detection has been discussed elaborately in [21].

7) Social Interaction Channels: With the emergence of OSNs, nowadays people spend a lot of time in OSNs to communicate with their acquaintances or friends. Hence, OSNs are an important platform for information exchange among the users of OSN. For instance, if we consider Facebook, users can comment/reply/like/tag and share posts. Any action taken by the user creates an online diffusion channel among the users. An advantage of these OSN platforms is that it makes social interactions possible even between the users who are not physically close, and thus, encourages a wide range of people to participate in social interactions. Also, social interaction channels assist in more efficacious interactions among the users which results in broader information diffusion in OSN. For instance, if some important message is required to be conveyed in OSN, then the intensity of interaction in a social interaction channel plays a great role. So, it is inferred that the rate of information diffusion is dependent on both the type of message required to be communicated and the social interaction channel used for information propagation.

In this context, our objective is to exploit these social interaction channels for community detection. The role of the social interaction channels for community detection has been discussed in [21]. They illustrated how the results of the information diffusion in social interaction channels and thresholding results in good communities.

8) Common Interests and Traits: Certain characteristics of users are stable across different timestamps and consistent over varied situations and are used to derive the similarity between users in OSN. Understanding these characteristics by observing the behavior of users in information diffusion over social networks has become a research issue. If the certain similarity in the diffusion pattern of a set of users lying in the same interaction-intensive domain is identified by observing users' behavior, it indicates that those users' share some common interests and traits. Also, users' sharing common interest and traits may not share direct contact and still get affected by the same information during the diffusion process because information may reach a user via multiple users' and thus people with similar interests and traits can come in proximity and get jointly affected. Thereafter, the incorporation of an appropriate community detection approach results in the detection of good communities. An illustration of the role of common interests and traits in combination with thresholding for detecting communities has been discussed in [21].

9) Number and Length of Cascades: Information diffusion in OSNs is heavily influenced by its users', hence, identifying the infection order is very important. This infection order can be captured during the diffusion process by the identification of the adoption of a contagion i by user u at a certain time t. This dataset of tuples (u, i, t) is called a cascade. The detection of cascades in information diffusion has attracted researchers recently because of its significant role in the proliferation of information throughout the network. In this context, the objective is to understand the role of contagions in the number and length of cascade. First, the number of cascades generated at the end of diffusion is dependent on the number of contagions adopted for information diffusion. Second, the length of the cascade depends on how contagious are the contagions that have been used for information diffusion and the global transmission time initialized by the diffusion model. Higher global transmission time implies a higher probability of maximum number of vertices to participate in the cascade. However, transmission time between the vertices is inversely proportional to the similarity between the vertices.

During the diffusion process, similar vertices are collected based on the joint participation of the vertices in at least one cascade. Discussion on the role of the number of cascade and length of cascade on the community detection has been explored in [18] and [19], respectively.

III. Unfolding Information Diffusion

Information diffusion is the process by which information spreads from one individual to another through social

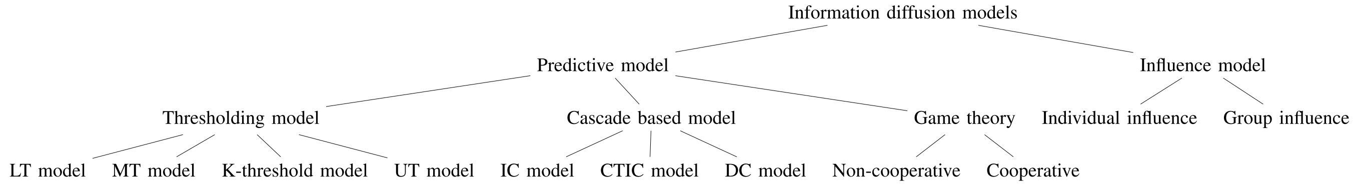


Fig. 1. Classification of information diffusion models.

interactions. Our discussion is centered around the influence that information diffusion has on different network properties and social facets explained in Section II. The objective of information diffusion is to maximize the activated individuals in the network, i.e., those individuals already infected by the information and likely to infect their neighboring inactive individuals at any particular point of time [35], [36]. Several models have been developed to imitate and better understand the diffusion process in real-world social networks, which are broadly categorized as: 1) predictive model; 2) influence model; and 3) optimization model and is shown in Fig. 1. The application of these models gives the trace of the information propagated through a social network. The trace of propagated information includes several useful details like where and when the information is generated and in what fashion dispersion occurs, how people behave while sharing information, the facets affected as a result of the diffusion process etc. In this section, various diffusion models have been discussed from the view point of strategies incorporated for spreading information and their effect on network properties and social facets.

A. Predictive Model

It is used to estimate how information will spread throughout the network in the future. Two categories of predictive models have been developed such as graph-based model and nongraph-based model. In graph-based predictive models, network structure is considered to analyze the dispersion. However, the nongraph-based approach does not assume any graphical structure and is mostly used for modeling epidemiological processes where decisions are taken unconsciously [37], [38]. Since social interaction is central to information diffusion in the network, the study is focused on the graph-based predictive models such as the thresholding model, cascading model, and game-theoretic model [39]. Both the thresholding model and cascading model are probabilistic iterative processes based only on the dynamics of neighboring users without considering the content. However, based on the content, the diffusion patterns may differ which is illustrated by the game-theoretic model.

1) Thresholding Model: The thresholding model is designed based on the assumption that each of the individuals has a threshold value associated with it which indicates the amount of information required by them to be activated [5], [6]. Closer users have a higher probability to propagate more information as compared to regular users. If the propagated information exceeds the threshold of an inactive user, the user gets activated. The setting of the threshold as well as the amount of information propagated and the effect on network properties and social facets gives rise to several variants of the threshold model that have been discussed below.

a) Linear Threshold (LT) Model: Is a graph-based *receiver-centric* diffusion model which is implemented in a directed and weighted graph [40]–[43]. It focuses on the relationship of an inactive vertex with its neighboring vertices and then the summation of all the relationship values with the neighboring vertices are computed and compared with a threshold value to examine whether to activate an inactive vertex or not. It can be said so because this model uses a combination of influence degree and influence threshold to trace the spread of information. The influence degree of a source vertex on a target vertex signifies the relationship strength between the two vertices. The higher the influence degree, the stronger is the relationship strength of the source and target vertices. Additionally, an influence threshold λ_v of a vertex v indicates the threshold value required to activate an inactive vertex by the neighboring vertices. The inclusion of influence degree and influence threshold in this diffusion model shows an inclination of an inactive vertex toward similar vertices and the likelihood to get influenced by its neighboring vertices, so there is a high probability that the network properties and social facets get affected by this approach.

As has been mentioned earlier, the LT model is implementable in a weighted and directed graph defined by $G(V, E, w)$ where V indicates the set of n vertices and $E \sqsubseteq V \times V$ signifies the set of m directed edges. The parameter $w : V \times V \rightarrow [0, 1]$ is a weight function which is expressed in terms of connection strength, value strength, cosine similarity, Jaccard coefficient, homophily, contagion, topic, common interests and traits, and number of common neighbors. Here, w indicates the influence degree and it has to be defined on each edge where $w(u, v) = 0$ implies $(u, v) \notin E$, and $\sum_{u \in V} w(u, v) \leq 1$. Initially, a seed set $S \sqsubseteq V$ is given at step $i = 0$. Thereby, the influence cascades by setting an influence threshold λ_v that lies in the range $[0, 1]$ which is randomly chosen for each vertex v due to lack of information of the individual's true threshold. At step i, an inactive vertex, $v \in V$ is activated when the summation of the influence degree (edge weights) of all its activated neighbors exceeds a threshold λ_v which is defined by [42]:

$$\sum_{u \in \bigcup_{0 \leq j \leq i-1} S_j} w(u, v) \geq \lambda_v. \qquad (16)$$

In (16), the term S_j indicates the set of vertices activated at step j and hence, present in S_i because $j < i - 1$. Thereafter, the influence cascades in discrete steps $i = \{1, 2, 3, \ldots, k\}$. At step k, when no more vertices have a chance to be activated, such that $S_k = \phi$, the process stops. Subsequently, the set of all the activated vertices denoted by $\sigma_L(S)$ is expected to be $|\bigcup_{i \geqslant 0} S_i|$ which indicates the influence spread of the seed set S. The illustration of LT Model is shown in Fig. 2, where initially at time $t = 0$, the vertex v which represents a vertex in the social network is inactive. When most of the neighbors of v which are represented by vertices v_1, v_2, v_3 and v_4 buys a new mp3 player (activated vertices), they try to

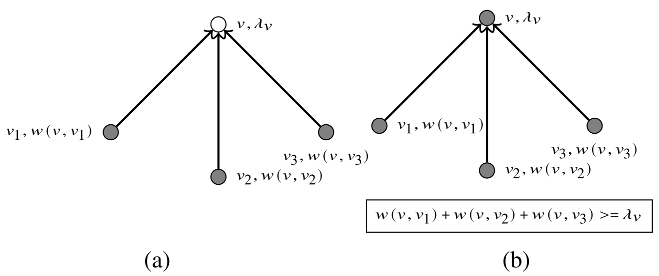


Fig. 2. LT model example, white colored vertices indicate inactive vertices, gray colored vertices indicate active vertices, and black thin lines indicate influence degree of each weighted edge. (a) $t = 0$. (b) $t = 1$.

activate v with influence degree $w(v, v_1)$, $w(v, v_2)$, $w(v, v_3)$, and $w(v, v_4)$, respectively. There is a high probability that v too might get influenced by its neighboring vertices depending on how influential are those vertices on v, and it is represented by

$$w(v, v_1) + w(v, v_2) + w(v, v_3) + w(v, v_4) \geq \lambda_v. \quad (17)$$

At time $t = 1$, vertex v gets activated when the influence degree values of its neighboring vertices denoted by $w(v, v_1)$, $w(v, v_2)$, $w(v, v_3)$, and $w(v, v_4)$ exceeds the threshold value of vertex v denoted by λ_v. Here, the λ_v value indicates the affinity of a vertex to be influenced by its neighboring vertices and so the properties such as belonging factor, topic, common interest and traits, homophily as well as information difference get affected with the change in λ_v value. The change in the parameters like influence degree and influence threshold would alter the diffusion pattern and hence, would affect the network properties and social facets influenced by this model. This model is applicable for spreading rumors and diseases. Motivated by this diffusion model, some other variants of the LT model depending on the network properties and social facets have been discussed as follows.

b) Majority Threshold (MT) Model: Activates its inactive vertex u when the majority of the neighboring vertices of u are active. It can be reduced to the LT model, when value 1 is assigned to the influence weight between connected pairs (u, v) and the threshold for any vertex u is set as $(1/2)D(u)$ where $D(u)$ denotes the degree of vertex u. Hence, the activation of a central vertex with a large degree in a network leads to the activation of a large number of surrounding vertices [43], [44]. This model can be applied in the voting system, distributing computing, etc.

c) k-threshold diffusion Model: Activates its inactivate vertex u when at least k of its neighboring vertices are active. It can be reduced to the LT model when value 1 is assigned to the influence weight between connected vertex pairs (u, v) and the activation threshold for any vertex u is set as k when k neighboring vertices of u have been activated. The k-threshold model will have varied performance depending on the value of k. When $k = 1$, an inactive vertex will be activated when at least one of its neighboring vertex is active. Thus, in a connected component, with initial seed vertices, all the vertices are activated finally. Whereas, when the value of k is greater than the largest vertex degree in the network, e.g., $k > \max_{u \in V} D(u)$, then no vertex will be activated. Moreover, when k is a medium value, some of the vertices will be activated as the information propagates but vertices having less than k neighboring activated vertices will not be activated [22]. Thus, it is inferred that the k-threshold model has the flexibility to set the number of vertices which participates in the activation process of an inactive vertex.

d) Unanimous Threshold (UT): Uses threshold $\lambda_v = D(v)$ which is equal to the degree of vertex v, for activating vertex v. It indicates that UT is the most influence resistant model among other threshold models [45].

2) Cascade-Based Diffusion Model: In this diffusion model, information diffusion takes place in a step-by-step fashion. At each step, active vertices get a chance to activate the inactive vertices and if it succeeds, the state of the inactive vertices will be changed and the propagation continues until no more vertices are left to be activated and this process results in the information cascade. An information cascade [7] occurs when the action of vertices is based on the actions of their neighboring vertices, and it gives us the complete information about the pattern followed by the diffusion process. Different cascade-based diffusion models based on the activation trials and vertices reaction to the activation trials and their effect on the network properties and social facets have been discussed in this section.

a) IC Model: Is a graph-based sender-centric diffusion model which is implemented in a directed graph [37], [40], [46]–[51]. It focuses on the independent behavior of the active vertices on their neighboring inactive vertices to examine the changed state of a vertex. It is because this model uses a diffusion probability to increase the density of activated vertices and hence, it is used to trace the flow of information. The diffusion probability is based on the relationship/frequency of exchange of information between the active and inactive vertices. Hence, it is expected that several network properties and social facets would be affected by this diffusion approach.

As has been mentioned earlier, the IC model is implementable in a directed graph defined by $G(V, E)$ where V indicates the set of n vertices and $E \sqsubseteq V \times V$ signifies the set of m directed edges. When $e \in E$, then $e = (v, w) \in E$ denotes an edge, where $v \neq w$. This model assigns a diffusion probability, $k_{v,w}$ for each edge to activate the inactive vertices where $0 < k_{v,w} < 1$. Initially, the diffusion process starts from a given initial set of active vertices v at timestamp $t+1$ which tries to activate its inactive neighboring vertices w, then the set of parent vertices of w at timestamp $t+1$ denoted by $B(w)$ is defined by [48]

$$B(w) = v : (v, w) \in E. \quad (18)$$

After the diffusion process has begun, it proceeds in discrete timestamps. The set of vertices activated at a certain timestamp t is denoted by $D(t)$. The probability of the vertex w to become activated at timestamp $t + 1$ is based on all the vertices which are activated at $D(t)$ that independently tries to activate w with probability $P_w(t + 1)$ and is defined by [48]

$$P_w(t + 1) = 1 - \prod_{v \in B(w) \cap D(t)} \left(1 - k_{v,w}\right). \quad (19)$$

If w does not belong to $D(t + 1)$, it means that v failed in its attempt to activate w. Although v may not succeed, it will not make any further attempt to activate w in

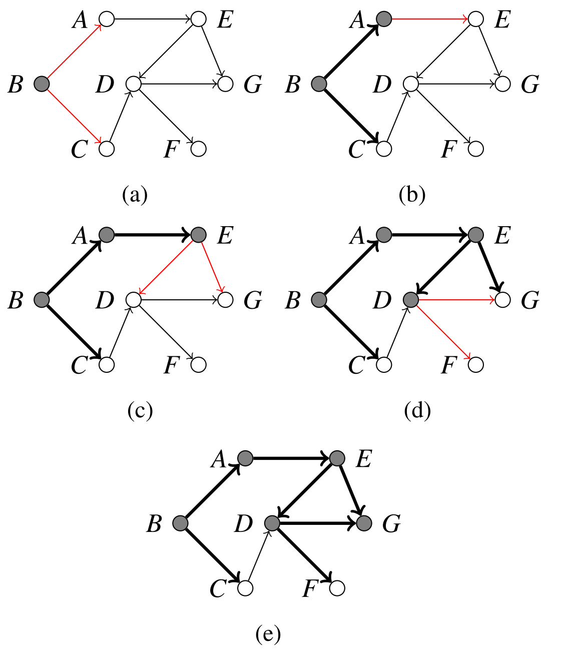


Fig. 3. IC model example. Gray colored vertices indicate active vertices, white colored vertices indicate inactive vertices, thin red arcs indicate edges where active vertices try to activate their inactive neighboring vertices, thick black arcs indicate edges propagated once, and thin black arcs indicate edges not yet propagated. (a) At timestamp $t = 0$. (b) At timestamp $t = 1$. (c) At timestamp $t = 2$. (d) At timestamp $t = 3$. (e) At timestamp $t = 4$.

subsequent rounds. This process is illustrated by the graph shown in Fig. 3. Initially, the process starts from an active vertex B which tries to activate its inactive neighbors A and C with diffusion probability values 0.8 and 0.4, respectively. As the edge $B \rightarrow A$ holds a high diffusion probability value, only vertex A is activated at timestamp $t = 1$. Thereafter, this process continues until there are no more vertices left to be activated and stops at timestamp $t = 4$. As this diffusion strategy highly depends on the diffusion probability value, change in diffusion probability value affects network properties such as connection strength, value strength, cosine similarity, Jaccard coefficient, belonging factor, and social facets namely homophily, contagion, tracking influence, influential spreader, number of cascade, length of cascade, topic, and common interest. Some other variants of the IC model influencing the network properties and social facets have been discussed here.

b) Continuous time delay IC model (CTIC): Is an extended version of the IC model and allows continuous-time delay. In the CTIC model, Saito *et al.* [52] specified for each edge $(u, v) \in E$, real values $r_{u,v}$ and $k_{u,v}$ as time-delay parameter and diffusion parameter, respectively, with $r_{u,v} > 0$ and $0 < k_{u,v} < 1$ in advance. The diffusion process originates from an initial active vertex u at time t, which tries to activate its inactive neighboring vertices using time-delay parameter δ. A time-delay parameter δ is set considering exponential distribution with parameter $r_{u,v}$. So, if vertex v is not activated at time $t + \delta$, then vertex u tries to activate vertex v within that timestamp. However, vertex u is only given a single chance to activate vertex v. If the active neighbors of vertex v also tries to activate vertex v at $t + \delta$, their activations are sequenced in random order. The process stops when no more vertices are left to be activated.

c) Decreasing Cascade (DC) Model: Supports the order independence of vertices. It means that the order in which the vertices try to activate their inactive neighbors does not affect the propensity for activation. It is illustrated by $P(u \rightarrow v \mid S)$ which indicates the probability for vertex u to activate vertex v given S as the set of vertices who failed to activate v. Let S and S' denote two different activation vertex set where $S \subseteq S'$. Then, $P(u \rightarrow v \mid S) = P(u \rightarrow v \mid S')$. The activation probability for a vertex does not increase as the number of activation trials increases. Thus, as vertices of set S fail in the activation trials to activate v, it indicates that probably v is not interested in the information. Thus, as the number of vertices in set S' performs the activation trials, the probability of vertex u to activate vertex v decreases. Hence, $P(u \rightarrow v \mid S) \geq P(u \rightarrow v \mid S')$. It can be inferred from this model that the probability of activating vertex v decreases when more vertices have already attempted in activating v [43], [53].

3) Game Theory Model (GTM): The GT model [8]–[11], [37], [54] is a profit-centric diffusion model which is implementable in a dynamic graph where information is diffused only when profit from information diffusion is more than the cost of information itself. Here, the diffusion takes place in an environment where the vertices are considered agents. Our objective is to discuss the GT model based on the effect that a certain decision has on network properties and social facets. Here, diffusion terminates when profit cannot be further maximized by any change in the diffusion strategy. Considering the strategy incorporated, game theoretic-based diffusion can be of two types: noncooperative and cooperative. In both noncooperative and cooperative game theory-based diffusion, pre-play communication is allowed. However, in noncooperative game theory-based diffusion, each of the agents selfishly tries to maximize his/her profit and the diffusion process stops when the profit of any of the agents cannot be further increased by independently altering the strategies. Besides, cooperative game theoretic-based diffusion considers the joint action of all the agents rather than their individual actions and outcomes optimal results when implemented in connected graphs. After the diffusion, the profit obtained is distributed to all the agents according to their contribution. Finally, the diffusion process stops when the equilibrium state is reached which is obtained by the Shapley value. As this diffusion process is based on the decision taken by the agents which leads to the corresponding actions, the decision taken reveals the relationship strength between the agents participating in the diffusion process and in the process affects some network properties and social facets.

A game is a tuple $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ where N is the set of players and $A := \{a \mid a = (a_i)_{i \in N}, a_i \in A_i, \forall_{i \in N}\}$ is the set of action profiles of the agents and $u_i : A \longrightarrow \mathbb{R}$ is the payoff function of agent i, i.e., $(a_1, \ldots, a_n) \longmapsto u_i(a_1, \ldots, a_n)$. Here, payoff may represent the profit that needs to be maximized or the cost that is needed to be minimized and analysis should be done on when the agents choose to spread information. The influence of the cooperative and noncooperative game theory-based diffusion model on several network properties and social facets have been discussed here.

In noncooperative-based diffusion, an agent makes a decision based on the available information and tries to maximize the profit which is further illustrated by a strategic model, namely prisoner's dilemma shown in Table II.

TABLE II
PRISONER'S DILEMMA

	C	DC
C	3, 3	0, 4
DC	4, 0	1, 1

The symbol *C* denotes cooperation by the agents and symbol DC indicates noncooperation by the agents. The story is of two criminals who were questioned in separate rooms and each were offered the same deal by the police. They were expected to take certain decisions keeping in mind the penalty. The decision taken by each of the agents to take a certain action reveals the edge strength measures, belonging factor between the players involved in the diffusion process. If the adoption of a new information results in a decrease in the penalty, then information difference is disclosed. Also, from Table II, it is evident that in the initial stage, the sentence is one year for each due to lack of evidence. But if one of them confesses the crime of the other, then he is set free if the other does not confess and vice versa. Furthermore, if both confess the crime of each other, then both of them are sentenced to three years of prison each.

Another example has been discussed to demonstrate the cooperative game theory which goes as, suppose a set of friends decides to go to a restaurant and have some meal but one of them decides not to have any food there, then the total bill should be equally divided among the friends who had the meal so that the resulting cost would be fair for all of them. It is inferred that decision taken by the friends in paying the bill is cooperative in nature. It can be said so because, in cooperative game theory, an agent makes a decision based on optimal joint actions and reasonable cost which keeps the coalition stable [55]. Like in noncooperative-based diffusion, the cooperative-based approach also affects edge strength measures, belonging factor, information difference for the same reasons as mentioned earlier. In general, the profit of adopting a particular action may also be associated with the decision taken by the neighboring agents. The decision adopted may be similar if the neighboring agents are highly correlated and this is quantified by structural equivalence measures, homophily. In particular, if the decision of an agent depends on the influence of the neighboring agents, then contagion is affected. The decision taken by an agent may also depend on the intersection of common interests associated with the neighboring agents. The idea of information dissemination using game theory has been applied by various researchers based on different attributes. It is very interesting to see how social facets or network properties or a combination of the two are used to propagate information throughout the network using cooperate and noncooperative game theory diffusion model which has been discussed here.

Hang *et al.* [56] utilized the edge strength measures to uncover the relationship strength between the agents to take a decision accordingly. The stronger the relationship strength, the higher is the probability for information diffusion, Wang *et al.* [57] discussed how social facets, such as a topic and common interest, assists in the diffusion process in an evolutionary network. Their proposed strategy is suitable when decisions are to be taken in a dynamic network for information propagation. Further, Liu *et al.* [58] considered a combination of edge strength measures, common interest and traits, a topic for taking a decision. Their strategy is helpful to determine the future relationships between agents. Li and Shigeno [59] utilize a noncooperative game on weighted graphs for spreading information. Gong *et al.* [60] have discussed the exploitation of fraction of rational vertices and homophily for tracing the influence in information propagation.

B. Influence Model

In this section, we shall introduce the contribution of influence in the spread of information throughout the network. Influence can highly motivate people in the information diffusion process and it is an essential part of the analysis of social networks. It is because analyzing influence gives a solution to some important problems such as community detection in social networks [1]. In this section, our discussion has been focused on individual influence and group influence. In individual influence, the objective is to use various facets such as network properties and social facets for information diffusion to detect the influential users which further helps in faster propagation of information throughout the network.

However, in group influence, the objective is to detect communities using the facets which have high influence in social networks and thereby use the same in the spreading of information [61]–[66]. The objective of the influence maximization problem is to select the k-seeds in the network so as to maximize the influence.

1) Individual Influence: Their objective is to maximize the information diffusion by reducing the time of contagion. To do so, different strategies have been adopted to identify/predict the influential users/opinion leaders. The opinion leaders are those users who act as a bridge in the information diffusion process. The individual influence exploits several network properties or social facets or a combination of the two and utilizes the predictive models discussed in Section III-A to capture the influence spread during the diffusion process [67]–[70]. The outcome of this individual influence is quantified by the increase in the number of users affected during the diffusion. Several variants of the individual influence along with their applications have been briefed below.

Wang *et al.* [71] used centrality measures and structural holes to identify the opinion leaders in a directed dynamic graph by using individual influence. These opinion leaders are used to predict popular information inherent in the network. While Chen *et al.* [72], Mao *et al.* [73], Wu *et al.* [74], and Ullah and Lee *et al.* [75] used several facets, such as edge strength measures, topic, common interest and traits, mutual information in their diffusion strategy. Bo *et al.* [72] used these facets to identify the opinion leaders and predict popular information. Although, Mao *et al.* [73], Wu *et al.* [74], and Ullah and Lee *et al.* [75] used these facets to predict influence which is employed to identify popular topic, opinion leader, and identification of influential vertices, respectively. Then, Li *et al.* [69] proposed an approach to select the seed vertex based on the topic and benefit of the vertex. Thereafter, the influence is quantified using the IC model. However, Ngugen *et al.* [70] used both benefit function and cost on the topic that have been utilized for seed selection. However, Guo *et al.* [68] proposed a personalized approach where a seed vertex is able to influence only one vertex that is topic-relevant and thereafter uses IC Model for capturing influence spread. Whereas community structure and node coverage gain (CNCG) [76] is a community-based seed selection that considers the role of overlapping community structure and node gain. MLPR (matrix multiplication, linear programming, and randomized rounding) [77] utilizes linear programming to select seed vertices in the LT model. Effective distance-based centrality (EDBC) [78] algorithm have been used to exploit the role of centrality measures in the detection of influential vertices. Olivares *et al.* [79] have proposed a new model to minimize the seed size and maximize the influence spread through the network using optimization. Their approach discusses the efficiency of the particle swarm optimization (PSO) algorithm in solving a multiobjective problem. Heuristic independent path algorithm (HIPA) [80] demonstrated the role of network properties and independent path in the selection of influential vertices for broader information propagation throughout the network. Carnia *et al.* [81] have illustrated how centrality measures are used to find the most influential vertices that can later be exploited for optimal information dissemination. Kumar and Sinha [82] demonstrated the significance of degree centrality in the initiation of information dissemination. Then, their diffusion strategy is composed of two phases. First, the target vertices are allowed to even revert back to the initial state after getting influenced by information. Second, permanence is ensured to the final acceptance of information during diffusion. Namtirtha *et al.* [83] proposed a novel indexing method namely, network global structure-based centrality (NGSC) for seed set selection by considering network properties. Results show that NGSC works excellently on all types of network structures. In order to have a faster diffusion through the network, Kalantari *et al.* [84] introduced a node probing-based overlapping community detection (NPOCD) algorithm to identify the influential vertices and thereby detect the overlapping communities considering edge strength measures.

2) Group Influence: A group is a set of vertices that are densely connected and have common attributes, namely a group of people with common interest, common subject preference, etc. Group influence indicates a community that have high influence in the network and hence, assists in the information propagation. To detect the inherent communities, several facets such as network properties and social facets have been exploited.

Various group influence approaches based on the combination of network properties and social facets have been introduced. For example, Yang *et al.* [61] and Zhou *et al.* [62] proposed PCL-DC model and SA-Cluster-Inc method, respectively, by exploiting centrality measures such as degree centrality, closeness centrality, betweenness centrality, topic, and common interests. Thereafter, Yang *et al.* [61] Zhou *et al.* [62] used two-stage EM algorithm to assign community membership to all the vertices but Zhou *et al.* used a k-means clustering algorithm to cluster the inherent vertices. However, Ruan *et al.* [63] proposed a CODICIL method which uses edge strength measures and similarity measures like cosine similarity, Jaccard coefficient to extract the edge-related information, and thereafter, communities are predicted using a biased edge sampling procedure. Finally, the Metis and Markov clustering algorithm is used for community clustering. Additionally, Yang *et al.* [64] utilized several social facets, such as topic, common interests, and traits along with edge strength measures, Jaccard coefficient to predict communities. The idea of Peng *et al.* [66] is to identify the k-core subgraphs using various centrality measures like degree centrality, closeness centrality, betweenness centrality, and then various community detection algorithms and optimization algorithms are used for detecting communities. Bhattacharya and Sarkar [85] illustrated the impact of information diffusion in social networks influenced by contagion and homophily. Nair *et al.* [86] demonstrated how the position, size, and network properties of a critical mass of persistent minority influences the adoption of a new convention in a network composed of polarized communities.

C. Optimization Model

1) Genetic Algorithm Diffusion Model (GADM): The GADM [87], [88] is a graph-based optimization-centric information diffusion model which is implementable in both directed and undirected graphs. It is based on the assumption that individuals interact with each other to increase their overall information and the diffusion of information takes place by utilizing the amount of information exchange and then measuring the gain or loss associated with it. Here, the information exchange is modeled by a single point crossover operator and gain or loss associated with it is calculated by Holland's hyperplane-defined objective function. Single point crossover is deployed to quantify the new information and it is evaluated by utilizing Holland's hyperplane-defined objective function which is a synthetic objective function. As single point crossover is used to model the information exchange, change in point selection affects the information exchanged between individuals across social interaction channels.

As has been mentioned earlier, the GADM is a graph-based diffusion model implementable on a directed and undirected graph $G(V, E)$. Here, vertices $V = v_1, v_2, \ldots, v_n$ indicates spatially distributed individuals interacting over a period of T discrete timestamps. The interactions result in the diffusion of information throughout the network because each of the individuals holds some information received either from their neighbors or from mainstream media like radio and TV [88]. The information diffusion is measured by comparing the objective score of the state of all the individuals before and after information exchange. Here, the state of an individual is represented by a binary chromosome of length β. After information exchange, the new state is acquired using a single point crossover and the objective score is obtained using Holland's hyperplane-defined objective functions. This function is composed of a set of schemata (short substring with wildcards starting at a specific position). The objective score of the state of the individuals after crossover is represented by the sum of the scores of all the schemata contained in the binary string. If the crossover results in a better objective score of either of the state strings, then the corresponding parent's state string is updated by the following equations [87]:

$$S_u^{t+1} = \underset{x \in S_u^t, S_u^{t+1}, y_1, y_2}{\operatorname{argmax}} \; f(x) \tag{20}$$

$$S_v^{t+1} = \underset{x \in S_v^t, S_v^{t+1}, y_1, y_2}{\operatorname{argmax}} \; f(x). \tag{21}$$

TABLE III
COMPARISON OF PREDICTIVE INFORMATION DIFFUSION MODELS

Model/Proposed By	Network properties	Social facets	Activation condition	Application
IC Model [48]	Edge strength measures, structural equivalence measures, belongingness measures	Topic, common interest, homophily, contagion, tracking influence, influential spreader, length and number of cascade	$P_w(t+1) = 1 - \prod_{v \in B(w) \cap D(t)} (1 - k_{v,w})$	Promote new product, influence maximization
CTIC [52]	Edge strength measures, structural equivalence measures, belongingness measures	Length and number of cascade, topic, common interest	-	Studies time delayed diffusion
DC [53]	Edge strength measures, structural equivalence measures, belongingness measures	Length of cascade, topic, common interest, information difference	$P(u \to v \mid S) = P(u \to v \mid S')$	Collective behavior, information spreading
LT Model [41]	Edge strength measures, structural equivalence measures, belongingness measure	Topic, common interest and traits, homophily, contagion, information difference, number of common neighbors	$w(v, v_1) + w(v, v_2) + w(v, v_3) + w(v, v_4) \geq \lambda_v$	Spreading rumors and diseases, Collective behavior, Influence maximization
k- threshold [22]	-	Topic, Common interest and traits	$k > max_{u \in V} D(u)$	-
UT [45]	-	Topic, Common interest and traits, homophily	$\lambda_v = D(v)$	Network security and vulnerability
MT [43]	-	Topic, common interests and traits, number and length of cascade	$\lambda_v = \frac{1}{2} D(u)$	Voting system, distributing computing etc.
Qifa et al. [56]	Edge strength measures	-	-	Prediction of business strategies
Yanzhuo et al. [57]	-	Topic, common interests and traits	-	
Liu et al. [58]	Edge strength measures	Topic, common interests and traits	-	

From the above discussion, it is clear that the model is governed by information exchange across social interaction channels.

D. Discussion

This section discusses the various diffusion models to quantify the spread of information throughout the social network. The objective of a diffusion model is to maximize the total number of activated vertices throughout the network. Different diffusion models such as predictive models, influence models, and optimization models have been introduced with varying parameters to achieve this goal. However, change in these parameters affects the network properties and social facets which further influence the number of activated vertices in the information diffusion process. Since social interaction is central to information diffusion in the network, hence, our focus has been mainly on the graph-based predictive diffusion models. The predictive models like the LT model and the IC model get only a single chance to activate their inactive neighbors but both of them differ in their diffusion strategies like the LT model focuses on the collective behavior of the vertices. While, on the contrary, the IC model captures the independent behavior of the vertices. But the IC model cannot solve the time-delay problem. To solve this, the CTIC model has been introduced. Furthermore, the GT model is used to simulate the competition and cooperation among the users. It focuses on the profit of the whole network and hence, it is neutral as compared to the LT model and IC model. Thus, it is suitable to be implemented in dynamic networks. We deliberately focused on the predictive models in this section to demonstrate its effect on the network properties and social facets to understand the information diffusion based community detection approaches discussed in Section IV which mostly utilizes the predictive models and so, we have summarized the predictive models in Table III. Also, another aspect to be considered for diffusion is the influence on a vertex by its neighboring vertices and hence, influence models like individual influence, group influence, and influence maximization have been discussed based on network properties or social facets or a combination of network properties and social facets for diffusion purpose. Although the IM approach is used to make the diffusion faster, a disadvantage of IM is that it is NP-hard under most of the diffusion models. Furthermore, motivated by artificial life, GADM was introduced where individuals communicate to increase personal information by exploiting social interaction channels.

IV. INFORMATION DIFFUSION FOR COMMUNITY DETECTION

Network properties and social facets (introduced in Section II) are utilized in several community detection algorithms to detect the underlying communities present in the network. The section above discusses the possibility about how a change in parameters of various diffusion models are likely to influence the network properties and social facets. Hence, prior to detect communities, the information diffusion process is used to influence various network properties and social facets which are likely to affect the detection of communities. Also, these network properties and social facets are used by several community detection approaches to predict communities. Recent literature explored these two alternatives and reported better performance of community detection techniques with the incorporation of information diffusion [16]–[21].

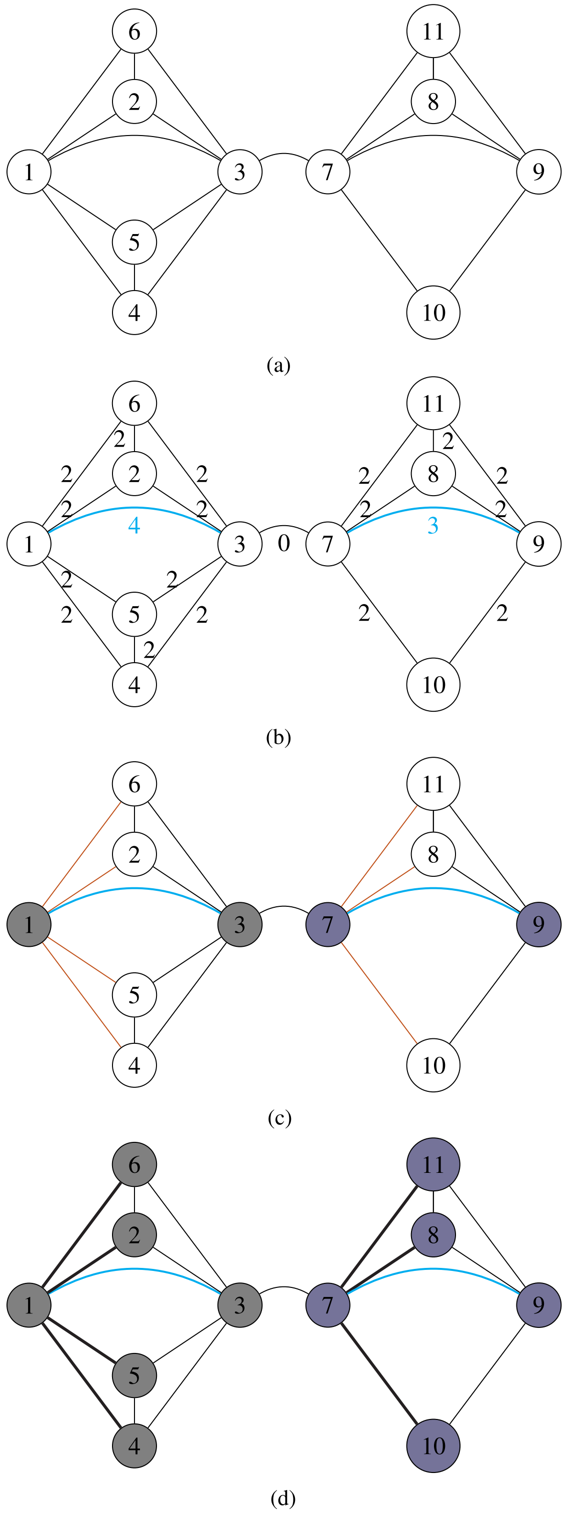
An algorithm proposed by Shen *et al.* [16] has been demonstrated in Fig. 4 to show how communities are detected using a combination of community detection algorithm and information diffusion model. The algorithm is based on the concept that the higher the number of common neighbors shared by two vertices, the greater is the probability for them to generate denser subgraph structures. Hence, they exploited the number of common neighbor information to set the edge weights, and successively, based on a threshold value, the core community members and number of components are predicted at timestamp $t = 1$. Thereafter, these core members are used for the diffusion of membership to all the unlabelled vertices at timestamp $t = 2$ which gives the inherent communities present in the network at timestamp $t = 3$.

This section discusses various community detection algorithms in terms of their utilized network properties and social facets in the underlying strategy of the algorithm. Also, the discussion on the algorithms that use a combination of information diffusion model and community detection algorithms has been carried out.

A. Network-Property-Based Algorithms

Network-property-based algorithms are those algorithms that utilizes network properties for community detection. Various vertex-related and edge-related network properties contribute to the structural aspect of the network and hence, these network properties are adopted by various researchers to detect communities in a topology oblivious network. Communities are detected either through direct usage of these network properties or used indirectly by adopting several strategies. Various network properties have been discussed in Section II and the influence of diffusion on these properties in Section III. This section sheds light on the direct and indirect usage of the network properties for community detection by citing the corresponding algorithms. The motif-aware weighted label propagation (MWLP) algorithm [89] and algorithm proposed by Sein [90] have been considered to analyze the direct utilization of network properties in community detection. To explore the indirect usage of network properties for community detection, CIDLPA algorithm [19], [91] have been introduced.

The network properties utilized by the abovementioned strategies have been discussed here in brief. The MWLP uses number and strength of connections to determine higher-order structure when only lower-order structure is provided and utilizes both higher-order and lower-order structures for detecting communities in a multiscale network and Sein [90] algorithm makes use of degree centrality to identify the seed sets and later incorporates a seed set expansion based community detection algorithm. On the other hand, the CIDLPA considers



(a)

(b)

(c)

(d)

Fig. 4. Demonstration of the information diffusion process to detect communities. In second subfigure, arcs joining core vertices are shown by cyan color and not yet propagated vertices and edges are shown by white vertices and thin black lines respectively. In third subfigure, bittersweet colored arcs indicate the edges joining core vertices with target vertices that are yet to be labeled. Fourth figure shows vertices with two different colors indicating two separate communities, and thick black arcs indicate edges propagated once. (a) Initial graph structure at timestamp $t = 0$. (b) Using number of common neighbors for assignment of weights at timestamp $t = 1$. (c) Core vertices detected by setting threshold as 3 and then diffusion initiated from core vertices at timestamp $t = 2$. (d) Communities at the end of diffusion at timestamp $t = 3$.

an incremental approach which utilizes a voting strategy based on belonging factor and value strength to detect communities in a dynamic network.

Here, we shall discuss CIDLPA, MWLP, and the algorithm proposed by Sein [90] to illustrate the role of vertex-related and edge-related network properties in community detection. Here, both CIDLPA and MWLP algorithms utilize edge strength measures because the stronger connection implies higher relationship strength between two individuals which implies denser subgraph structure exhibited by them, and hence, this aspect is considered in community detection. But, CIDLPA uses the network properties indirectly through the adoption of information diffusion (the influence of information diffusion on network properties have already been discussed in Section II). The advantage of the CIDLPA approach is that it can assign new communities to the vertices which join the network in the current snapshot using a combination of value strength and belonging factor. They incorporated a vote-based strategy that is based on the number of connections, the strength of connections, and belonging factor to assign labels to the vertices. Considering the maximum vote, labels are allotted to the vertices and thereafter, based on the similarity in belonging factor of the labels, communities are assigned to the vertices. The vote function is computed by [19]

$$\text{vote}_j = S_{0j} * \text{bf(sl)} + S_{1j} * \frac{(1 - \text{bf(sl)})}{3}, \quad j \in \text{neighbors}(i). \tag{22}$$

Equation (22) indicates the combined contribution of belonging factor and value strength in assigning labels to vertices. The label that gets maximum vote by the neighboring vertices of a vertex is assigned to the vertex. Here, bf(sl) indicates the belonging factor of vertex for selected label sl. The term S_{0j} and S_{1j} indicates the ineffectiveness and effectiveness of vertex j (the effectiveness of a vertex is defined by the summation of the value strength with its neighboring vertices). For an in-depth understanding of (22), the readers may refer [19].

The advantage of the MWLP algorithm is that it considers numerous relations inherent in the network to detect communities by incorporating higher-order structures. They determined the higher-order structures by utilizing motifs. Based on the identified motif, the number of connections between two vertices is computed which is used to derive a motif adjacency matrix. A larger number of connections between two vertices implies numerous relationships exhibited by them and hence, results in greater weight value in the motif adjacency matrix that implies closer higher-order connections, and hence, this is considered for community detection. However, to avoid any loss of information, MWLP considers both the motif adjacency matrix and vertex adjacency matrix to derive a novel re-weighted matrix which signifies the strength of each of the connections. Thereafter, considering these re-weighted matrix, label propagation proceeds over the network. Here, the label propagation is defined by a vote score which depends not only on the majority of the connections but also on the strength of connections. Then, considering the vote score of each of the vertices, communities are determined. The vote score is defined by [89]

$$S_{\text{vot}}(i) = \lambda \mid \Gamma^{C_i}(v) \mid + (1 - \lambda) W(v, i). \tag{23}$$

Equation (23) indicates that the vote of a vertex v is a combination of the number of neighbors of vertex v that has label i denoted by $\mid \Gamma^{C_i}(v) \mid$ and the motif intimacy between a pair v and i where i is the neighbor of vertex v denoted by $W(v, i)$. The tradeoff between the number of labels of the neighboring vertices and higher-order intimacy of the vertices under consideration is denoted by λ.

To illustrate the role of degree centrality in community detection, the algorithm proposed by Sein *et al.* [90] has been included here. They performed community detection based on the identification of a central vertex/seed vertex. Identifying a seed vertex is very significant because it possesses a larger number of connections and hence, communities grow from these seed sets. For the seed set identification, at first, they incorporated a graph partitioning method to obtain disjoint communities, and thereafter, degree centrality of all the vertices present in each community are calculated to obtain a set of vertices from each community having maximum centrality called seed vertices. Successively, the seed vertices are expanded using random walk, and thereafter, final communities are obtained from the set of vertices that have minimum conductance score.

B. Social-Facet-Based Algorithms

These are those algorithms that use social facets for community detection. Various social facets related to both vertices and edges have been discussed in Section II. However, these social facets are used either directly or indirectly for community detection. Indirect usage is incorporated by adopting a model which is used to exploit the hidden social facets in an unidentified network. For instance, we have discussed the influence of information diffusion models on these social facets. In this section, the direct and indirect utilization of some of these social facets for community detection has been explored. In particular, the algorithms proposed by Red *et al.* [92] and Yazdani *et al.* [30] examined how social facets are used directly for community detection. Moreover, the GID algorithm and the algorithm by Shen *et al.* [16] show the indirect usage of the social facets.

Social facets used by the abovementioned algorithms and their community detection strategies are introduced briefly in this paragraph. For example, Red *et al.* [92] proposed an algorithm that utilizes vertex-related social facets and statistical similarity strategy to determine communities. But this approach fails to detect optimal communities in an incomplete network. For optimal community identification, network connections are required to be considered. While a local leader identification-based community detection approach has been proposed by Yazdani *et al.* [30]. Successively, the GID algorithm proposed by Alvari *et al.* [17] and algorithm proposed by Shen *et al.* [16] demonstrate the indirect utilization of the social facets. The GID algorithm is a game-theoretic based approach that uses social interaction channels to incorporate genetic algorithm based diffusion and converges only after reaching the Nash equilibrium of the game, the algorithm by Shen *et al.* [16] uses the number of common neighbors to determine a threshold value which is used to predict the community structure and thereafter, uses LT diffusion model for community detection.

Out of these algorithms, a detailed discussion about GID, algorithm by Shen *et al.* [16] and algorithm by Yazdani *et al.* [30] have been carried out. GID algorithm exploits the role of information exchange across social interaction channels to take a profit-oriented decision in an environment with multiple agents (where each vertex of the underlying graph is assigned to an agent). The decision may be either to join, stay or leave a community based on maximum utility, i.e., received information. Here, the information exchange is quantified using an optimization diffusion model namely, extended EGADM which is an extended version of the GADM model discussed in Section III with the inclusion of a mutation operator. Ultimately, the communities inherent in the network are uncovered after reaching the local Nash equilibrium (NP-hard) [93] of the game. Besides, Shen *et al.* [16] initially identified the community structure followed by diffusion of membership to predict the inherent communities. Community structure identification is based on a threshold value that is set considering the edge weights. Here, the edge weight is determined by the number of common neighbors shared by the connected vertices. Hence, a high number of common neighbors shared by two connected vertices indicates a larger weight value and implies a stronger relationship exhibited by them. This information is very helpful for community structure identification because vertices possessing higher edge weight value are densely connected. Thus, based on this idea, the threshold value is set to obtain the connected components and core members of communities. The core members are those vertices having edge weight greater than or equal to the threshold value and the community labels of core members are set based on the number of connected components. Thereafter, the core members begin the diffusion of membership to its neighboring unlabelled vertices which is defined as [16]

$$Z_m(u) > \frac{1}{2} Z(u). \tag{24}$$

Here, (24) implies that a vertex u is assigned with a community label m when the weight exhibited by vertex u in community m denoted by $Z_m(u)$ is greater than half the weight of vertex u which is indicated by $Z(u)$ [for further understanding of (24), you may refer [16]]. The vertices that are assigned with community labels using (24) are the strong members of their communities and diffusion continues until there are no more strong members. Ultimately, at the end of diffusion, inherent communities are obtained. However, a disadvantage of this approach is the time complexity related to setting an appropriate threshold value because the communities obtained is largely dependent on the threshold value.

The community detection algorithm proposed by Yazdani *et al.* [30] illustrates the role of local leader (influential spreader) in community detection. As local leaders are permeable to all the neighboring vertices, they are followed by their corresponding neighbors. Hence, identifying the local leaders for all the inherent vertices present in the network gives a chain of interconnected vertices where each of the chain represents a community. However, the local leader is calculated based on the tendency of vertex i to its adjacent vertex j which is defined by [30]

$$R(i, j) = \begin{cases} \dfrac{w_{ij}}{\sum_{k \in N_i} w_{ik}}, & \text{if } i \text{ and } j \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

$$R^*(i, j) = \begin{cases} \prod_{k=1}^{K} R(v_k, v_{k+1}), & K < t \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

Equations (25) and (26) indicate the tendency calculated based on the direct link and common mutual friends respectively (for further understanding of the equations, you may refer [30]).

C. Combined Algorithms

Combined algorithms consider both network properties as well as social facets for identifying communities. These facets are utilized either directly or indirectly by several community detection algorithms as the underlying strategy for detecting communities. Indirect usage indicates the utilization of a model to exploit these facets and then assimilating this information to detect communities. In particular, we have discussed information diffusion models to understand the influence that it has on the network properties and social facets (discussed in Section III). The idea of indirect usage of these facets for community detection has come with the prerequisite to gather network structure information (when complete information about network structure is not available) and accumulate inherent characteristics of the network. Hence, several information diffusion-based algorithms have been explored here to show the indirect usage of these facets in community detection. In addition to this, we have picked some community detection algorithms to analyze the direct usage of some of the facets discussed in Section II. For instance, algorithm by Bhih *et al.* [94], who illustrated the direct use of facets and algorithms, namely community diffusion (CoDi) [20], COmmunity-preserving SocIal Network Embeddings (COSINE) [18], overlapping community detection based on information dynamics (OCDID) [21], and dynamic community detection based on information dynamics (DCDID) [95], demonstrates the indirect usage of the facets.

The usage of the OSN facets by these algorithms in addition to their proposed strategies is discussed here. Bhih *et al.* [94] proposed an algorithm using a similarity matrix which is build using the influence of clustering coefficient, topic, common interests, etc. on community structure and further detected communities using a novel clustering algorithm. However, CoDi and COSINE algorithms are dependent on the exploitation of cascade-based information and structural equivalence measure for community detection. By employing these facets, the CoDi algorithm predicted the community structure using a graphical model and then community labels are assigned using a local iterative method. Also, COSINE exploits the temporal dynamics and community preserving embeddings with the inclusion of these facets for community detection. However, OCDID utilizes information exchange of a vertex with its neighbors to exploit belonging factor, connection strength, boundary vertex, Jaccard coefficient, information difference, and social interaction channels to detect intrinsic communities.

DCDID introduces an efficient community detection approach that utilizes edge strength measures and information to uncover dynamic communities incrementally. In this approach, the efficiency is enhanced by filtering out the unchanged graph.

Out of these algorithms, we shall discuss CoDi, COSINE, and community detection algorithm by Bhih *et al.* [94] to illustrate combined algorithms. CoDi and COSINE are information diffusion-based algorithms that have been picked to explain how communities are detected when only cascade information is provided. Both of these algorithms use the given cascade information and structural equivalence measure on different stochastic frameworks for the community detection purposes. For example, CoDi algorithm is motivated by the propagation pattern of a contagion. As is likely for a contagion to spread in a community, hence, participation of vertices in cascades is considered for community detection. Also, the network information is not available initially, so a larger number of cascades helps to gather intriguing information about the network that aids in detecting communities. Furthermore, this information is used to exploit the similarity between the vertices sharing an edge based on the similarity in their diffusion behaviors which is quantified by cosine similarity. This information is particularly useful because the higher the similarity between two vertices, the higher is the probability for them to belong to the same community, and thereafter, appropriate community labels are assigned to the vertices using a stochastic procedure [conditional random field (CRF) and iterated conditional mode (ICM)]. To realize this method, the cascade information is utilized to derive a weighted undirected graph $G(U, R)$ where U indicates the activated vertices in a cascade, edge between two vertices indicate their joint participation in a cascade and weight indicates similarity between vertices sharing an edge. For community detection purpose, only the vertices $U = \{v_j \mid (v_j \in V) \Lambda(\forall X_i \in O : (t_j(i) < \infty))\}$ is considered. Here, the problem is to assign the vertices in G with appropriate community labels $(Y(u) = j \mid j \in 1, 2, \ldots, k))$ (where Y is a community assignment vector with g elements and k is the number of assumed communities denoted as $C = \{C_1, C_2, \ldots, C_k\}$ that maximizes $P(Y \mid O)$ [20] under the constraints $\cup_{C_i \in C} C_i = U$ and $\cap_{C_i \in C} C_i = \phi$. The maximization of community assignment vector is defined by [20]

$$\hat{Y} = \underset{Y}{\text{argmax}} P(y \mid O). \tag{27}$$

The COSINE algorithm [18] illustrates the role of CTIC, GMM, and MAP estimation for community detection. It utilized the continuous-time cascade information diffusion to preserve the local community structures. A cascade contains information on the activated vertices and their global timestamps. The timestamps are computed using a squared Euclidean distance which helps to derive the dissimilarity of the connected vertices. Higher dissimilarity implies a lesser chance of the vertices to be in the same community. It is to be mentioned here that the global timestamps are considered instead of local timestamps because the CTIC model only considers the earliest possible time of activation. This global transmission information is used to model the temporal dynamics and community preserving embeddings. However, to improve the temporal dynamics information, the timestamps of the survival users with missing timestamps have been computed using a statistical method that implies a larger length of cascade. Thereafter, the temporal dynamics and user's embedding have been used to predict the community structure using GMM. Finally, the community membership for each user have been assigned using the MAP equation [18]

$$\widehat{z_u} = \text{arg max}_k r_{uk} = \text{arg max}_k P(z_u = k \mid x_u). \tag{28}$$

In (28), z_u indicates the latent community membership of user u, r_{uk} indicates the conditional probability of user u in cluster k and x_u indicates embedding (for detailed understanding of in (28), refer [18]).

Bhih *et al.* [94] proposed a community detection algorithm motivated from the homophily property which says that vertices sharing links have common attributes. However, different attributes have a varied effect on the community structure based on the community detection algorithm used. Hence, a hybrid similarity matrix is used to weigh the attributes according to their influence on the community structure. This similarity matrix is a weighted combination of attribute information, shared neighbors, and connectivity information between the vertices. After the attributes have been weighted using this similarity matrix, a state-of-the-art clustering algorithm is incorporated on the weighted graph to obtain optimal communities.

D. Discussion

In the present day scenario, information about the network structure is not available due to several reasons, such as privacy, confidentiality, or may be due to a large amount of data. The ability of the information diffusion process to exploit several structural aspects and social aspects has attracted its usage in community detection in a topology oblivious network. Hence, we have performed a comparative analysis on the various information diffusion-based community detection approaches to demonstrate the influence of information diffusion on the network properties or social facets or a combination of the two and its effect on community detection and it has been summarized in Table IV. The information diffusion-based community detection approach is basically composed of two steps. First, the community structure is to be predicted, and second, the assignment of vertices to specific communities is processed. However, these two steps can be carried by two different approaches such as

1) The influence of information diffusion on the network properties or social facets or a combination of the two is used to predict the community structure, followed by incorporation of community detection algorithm or statistical inference methods for community assignment.
2) The effect of the network properties or social facets or a combination of the two on the community structure formation is analyzed using community detection algorithms or statistical inference methods and successively, for the diffusion of membership, information diffusion model is adopted.

Accordingly, our discussion sheds light on several algorithms that coincide with our objective and their time complexities. In addition to this, some of the most frequently used network properties and social facets for community detection have also been highlighted. For example, algorithms, such as CIDLPA, CoDi, COSINE, GID, OCDID, and DCDID, use the first approach. However, the algorithm by Shen *et al.* [16] utilized

TABLE IV

COMPARATIVE STUDY OF INFORMATION DIFFUSION BASED COMMUNITY DETECTION TECHNIQUES. HERE, CT: COMMUNITY TYPE, OC: OVERLAPPING COMMUNITY, DC: DISJOINT COMMUNITY, CID: CASCADE BASED INFORMATION DIFFUSION AND TH: THRESHOLDING

Ref.&Year	CT	Diffusion Type	Community Detection Approach	Network Properties	Social Facets	Algorithm Name	Time Complexity
Shen et al. (2010) [16]	DC	MT	TH	-	Common neighbors	-	$O((M+N)D_{avg}\log W_{max})$
Hajibagheri et al. (2012) [17]	DC	EGADM	Game-theoretic approach	-	Social interaction channels, mutual information	GID	-
Sattari et al. (2018) [19]	OC	CID	LPA	Belonging factor, value strength	-	CIDLPA	O(n+m)
Ramezani et al. (2018) [20]	DC	CID	CRF and ICM	Cosine similarity	Number of cascade	CoDi	-
Sun et al. (2018) [21]	OC	Information dynamics model	TH	Belonging degree, connection strength, boundary vertex, Jaccard coefficient	Information difference, Social interaction channels	OCDID	$O(k.n+L.n.k+\|c\|.n)$
Zhang et al. (2018) [18]	DC	CTIC	GMM and MAP	Euclidean distance	Length of cascade, social interaction channel	COSINE	$O(I_{max}\|C\|+KN)$
Sun et al. (2020) [95]	DC	Information dynamics model	TH	Structural equivalence measures	Information	DCDID	$O(L.n.k+(\|\Delta E_t\|+L.\|\Delta V_t\|.k_t))$

the second approach. However, each of these algorithms incorporates a different strategy. In particular, CIDLPA has the advantage of allotting new communities to the recently joined vertices in the current snapshot with linear time complexity, CoDi and COSINE show the contribution of cascade properties like a number of cascade and length of cascade, respectively, for community detection which decides their time complexities, GID algorithm shows the game-theoretic approach of community detection. It depends on group decision of all the agents and hence, holds a high time complexity, OCDID algorithm identifies the communities in a natural manner by utilizing information exchange using information dynamics model and hence, outputs natural communities and the time complexity is based on the total number of interactions, DCDID is an extension of OCDID and introduces an incremental approach to detect dynamic communities and the time complexity is based on a combination of initial community partition and incremental community detection, the algorithm by Shen *et al.* [16] does not get a partition when no community structure is found. Here, the community structure is found by setting of a threshold value which is very critical and takes maximum time. However, their approach eradicates the absurd complexity of implementing the second step when no community structure is found. Also, it has been observed that certain network properties and social facets have been used more frequently as compared to others by these algorithms. In particular, during the community formation step, belonging factor, structural equivalence measures, and edge strength measures have been mostly exploited. Additionally, cascade information (such as the number of cascade and length of the cascade), social interaction channels, clustering coefficient have also been used. However, for executing the second step, network properties, such as belonging factor, edge strength measures, have been mostly used. In addition, information difference has also been used by some of the algorithms. It has been observed that the algorithms that utilize more than one network property detect overlapping communities and algorithms using social facets identify disjoint communities. As the strategies acquired by each of these algorithms are different, the running time gets affected accordingly. Out of these algorithms, the CIDLPA algorithm takes the least time because of the application of the LPA strategy which takes linear time complexity and the GID algorithm is of the highest time complexity because of the application of Nash Equilibrium which is NP-hard. In addition to these algorithms, several community detection approaches based on network properties or social facets, or a combination of the two have been discussed. However, the indirect use of these facets by utilizing the information diffusion model yields faster and better results as compared to the direct incorporation of these facets for community detection.

V. COMMUNITY EVALUATION ATTRIBUTES

Evaluation of communities is necessary to determine the performance of community detection algorithms. Three main attributes for evaluating the communities are: test-bed, metrics, and method. Communities identified with any community detection algorithm are, in general, evaluated in two perspectives: accuracy and quality. Information diffusion-based community detection algorithm is also evaluated on the same line. It has already been discussed in the sections above that how direct or indirect involvement of diffusion process affects various network properties and social facets viz-a-viz detection of communities. Both accuracy and quality are crucial for evaluating communities. To measure accuracy only vertex labels are considered, while predominantly edges are used to measure quality. Due to this fundamental difference between accuracy and quality, often the evaluation process faces tradeoff between the two [96], [97]. Some literature works have also reported bias while evaluating communities [98], [99]. Different methodologies are incorporated to deal with the tradeoff between accuracy and quality as well as bias. Moreover, to measure accuracy requires groundtruth communities but quality does not require it. Thus, specific datasets are necessary to evaluate depending on what perspective communities are to be evaluated.

In this section, considering the above brief insights, various widely used evaluation metrics pertaining to measure both the accuracy and quality of identified communities are explained with their internal details. Different methodologies adopted to evaluate communities are also discussed. Popularly used datasets and repositories to evaluate communities are also listed.

A. Evaluation Test-Bed

Benchmark graphs comprising both artificially generated synthetic networks as well as real-world networks are used as evaluation test-bed for evaluating community detection algorithms. A synthetic network that have been widely used for evaluating community detection algorithms is the **GN benchmark** designed by Girvan and Newman that follows the statistical properties of networked systems for generating networks having community structures of the same size (Reference number 100). Another synthetic network generator called **LFR benchmark** graphs [101], [102] named after Lancichinetti–Fortunato–Radicchi is the most widely used synthetic networks for community detection. The LFR benchmark graph generator is a generalized version of the GN benchmark graph, which considers the power law distribution of degree and community size and is very similar to real-world networks. The LFR benchmark model handles the following properties for generating a community structure of the network such as the number of vertices n, the desired average degree and maximum degree of vertices as k and $k_{\max}$, respectively, the degree distribution γ , the community size distribution β, and a mixing parameter μ which indicates the desired average proportion of links between a vertex and the vertices external to the current community. Modification of the values of n and μ results in different network structures. The rest of the parameters are used as default values [101]. Real-world OSNs are dynamic in nature. There are basically four evolution events such as expansion and contraction, merging and splitting, birth and death, and switching vertices in a dynamic network [20], [103]. The LFR model has been further improved to generate **dynamic synthetic network** using the method and parameters of LFR graphs [101], [104]. A wide variety of real-world networks are publicly available online in various repositories such as SNAP [105], networkrepsitories [106], networkdata [107], kaggle [108], LINQS [109], datasets for social network analysis [110], KONECT [111].

B. Evaluation Metrics

Communities are evaluated in terms of different metrics, which are categorized as accuracy and quality and a summary of the evaluation metrics have been shown in Table V. Generally, two types of communities are detected, namely disjoint community and overlapping community. In the disjoint community, vertices can have membership in the community. On the other hand, vertices can have membership in multiple communities. Different accuracy metrics, as well as quality measures, have been defined to evaluate both kinds of communities. A detailed survey on different types of metrics can be found here [3], while we have considered only popular and recently developed metrics in this subsection.

TABLE V
SUMMARY OF EVALUATION METRICS

Metrics Type	Metrics name	Community type
Accuracy	ARI	Disjoint Community
	Purity	
	NMI	
	FNMI	
	F-measure	
	F1-score	Overlapping Community
	ENMI	
Quality	Modularity	Disjoint Community & Overlapping Community
	f-Modularity	
	Conductance	Disjoint Community
	Permanence	
	ANUI	Overlapping Community

1) Accuracy Metrics: Accuracy metrics are used to evaluate the correctness of the communities identified by community detection algorithms based on comparing the detected communities with the ground truth communities [112]. For a given network, $G(V, E)$, consider that $C = \{c_1, c_2, \ldots, c_j\}$ contains the set of detected communities and $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$ contains the set of ground-truth communities. Based on detected communities C and groundtruth communities Ω different accuracy metrics are defined as follows.

a) Accuracy metrics for disjoint community: Normalized mutual information **(NMI)** measure compares the similarity between detected communities with groundtruth communities. It gives higher value when the similarity is higher [19], [113]–[116]. NMI is defined as follows:

$$\mathrm{NMI}(\Omega, C) = \frac{I(\Omega, C)}{[H(\Omega) + H(C)]/2} \tag{29}$$

where I indicates the mutual information, which is defined as follows:

$$I(\Omega, C) = \sum_k \sum_j \frac{\mid \omega_k \cap c_j \mid}{N} \log \frac{N \mid \omega_k \cap c_j \mid}{\mid \omega_k \mid\mid c_j \mid} \tag{30}$$

and H signifies the entropy which is given as

$$H(\Omega) = -\sum_{k} \frac{|\omega_k|}{N} \log \frac{\omega_k}{N}. \tag{31}$$

Amelio and Pizzuti [117] have reported unfairness of NMI while evaluating communities due to the uses of a number of communities. They have proposed an improved version of NMI referred as fair NMI (FNMI), which does not consider the number of communities.

Adjusted random index *(ARI)* is less sensitive to the number of communities. It is the best suited for comparing the number of partitions when the true structure of the network is known [3], [118]–[121]. It considers chance-correction, which is defined as

$$M_c = \frac{M - E(M)}{M_{\max} - E(M)} \tag{32}$$

where M_c indicates the chance corrected measure, $E(M)$ indicates the value expected for some null model and $M_{\max}$ indicates the maximum value of M. The following formula gives the expected value of the intersection between ω_i and c_j:

$$E\binom{N_{\omega_i c_j}}{2} = \binom{N_{\omega_i}}{2}\binom{N_{c_j}}{2} \binom{N}{2}. \tag{33}$$

After simplifying (32), the formula for ARI is given by

$$\begin{aligned}&\mathrm{ARI}(\Omega, C)\\ &= \frac{\sum_{ij}\binom{N_{\omega_i c_j}}{2} - \sum_i \binom{N_{\omega_i}}{2}\sum_j \binom{N_{c_j}}{2}\binom{N}{2}N}{\frac{1}{2}\left(\sum_i \binom{N_{\omega_i}}{2} + \sum_j \binom{N_{c_j}}{2}\right) - \sum_i \binom{N_{\omega_i}}{2}\sum_j \binom{N_{c_j}}{2}/\binom{N}{2}}.\end{aligned} \tag{34}$$

The upper bound of ARI is 1, which indicates that both partitions are similar and ARI is 0 when it is chance-corrected. The value 0 or below 0 indicates that the relationship between Ω and C is equal to or less than that expected from random partitions.

b) Purity [122]: Assigns the detected community to the ground truth label which is most frequent for the community. It is zero when the detected and ground-truth communities are not similar. It is defined by the following equation:

$$\mathrm{Purity}(\Omega, C) = \frac{1}{N}\sum_{k}\max_{j}|\omega_k, c_j| \tag{35}$$

where the upper bound is one which signifies the similarity between the detected and ground-truth communities.

c) F-measure: Is used when the detected community is very less as the majority of vertices in the estimated community belongs to actual communities. F-measure is the harmonic mean of both versions of purity and is defined as follows [3]:

$$\text{F - Measure} = \frac{2.\mathrm{Purity}(\Omega, C).\mathrm{Purity}(C, \Omega)}{\mathrm{Purity}(\Omega, C) + \mathrm{Purity}(C, \Omega)}. \tag{36}$$

F-Measure can also be defined in terms of the harmonic average of precision and recall [123]. Recall measures the fraction of vertex pairs that belong to the same community in the referent benchmark graph while also belonging to the same community in the resulting partition. Whereas, precision is a measure of the fraction of vertex pairs that belong to the same community in the resulting partition and are also members of the referent benchmark network.

d) Accuracy metrics for overlapping community: **F1-score** measures the similarity between two overlapping partitions. It considers the vertex level for evaluation of overlapping communities and only considers the binary values, i.e., 0 and 1 [3]. It is given by the following formula:

$$F1 = \frac{1}{2}\left(\frac{1}{|\Psi|}\sum_{\psi_i\in\Psi} F1\left(\psi_i, C_{g(i)} + \frac{1}{|C|}\sum_{c_i\in C} F1\big(\Psi_{g'}(i), c_i\big)\right)\right). \tag{37}$$

Equation (37) gives the average of the F1-score of the best matching detected communities and ground-truth communities and F1-score of the ground truth communities and best matching detected communities.

Extended normalized mutual information **ENMI** [21] is another extended version of NMI designed to evaluate overlapping community. ENMI is defined as follows:

$$\mathrm{ENMI}(X \mid Y) = 1 - [H(X \mid Y) + H(Y \mid X)]/2 \tag{38}$$

where $H(X \mid Y)$ is given by the following equation:

$$H(X \mid Y) = 1 - \frac{1}{|C'|}\sum_{k}\frac{H(X_k \mid Y)}{H(X_k)}. \tag{39}$$

In (38) and (39), variables X and Y indicate the random variables associated with partitions C and C', respectively. $H(X \mid Y)$ is the normalized conditional entropy of cluster X given cluster Y. The value of ENMI is ranged between 0 and 1. The value 0 indicates that the partition does not match with the groundtruth, whereas the value 1 indicates that the partition and groundtruth completely matches.

2) Quality Metrics: Unlike accuracy metrics, ground truth is not required in quality metrics. Since accuracy metrics do account for connectivity within the network in any way, the structural feasibility of communities cannot be ensured [112]. On the other hand, quality metrics consider the structural feasibility of a community, which is defined in terms of connectivity within the community and the connectivity outside the community. If vertices within communities are densely connected in comparison to the connections to the rest of the network then identified communities are referred to as structurally feasible or interpreted as the quality of communities is high. Quality is measured to ensure the structural feasibility of identified communities. For a given network, $G(V, E)$, say, $C = \{c_1, c_2, \ldots, c_j\}$ contains the set of detected communities. Based on the predicted communities C different quality metrics are explained below.

a) Modularity: It is widely used to evaluate the quality of both disjoint and overlapping communities. The goodness of detected communities is evaluated by considering that the number of edges internal to a community should be greater than that in a random graph with similar degree distribution [3], [124]. Girvan and Newman's modularity function is used to evaluate the quality of detected communities, which is defined for disjoint community evaluation as follows:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\sigma\left(C_i, C_j\right)\right) \tag{40}$$

where m denotes the total number of edges in the network, i and j are two separate vertices where $i \neq j$, A_{ij} denotes the link between vertices i and j which is 1 if there exists a link between the two vertices, otherwise it is 0. d_i and d_j indicates the degree of vertices i and j respectively, and $\sigma\ (C_i, C_j) = 1$ if i and j belongs to the same community. If the community is modular, Q will have a high value.

The modularity has been modified to evaluate overlapping communities by defining a belonging coefficient [125], $\alpha_{i\psi}$ for each vertex i to measure its belongingness to a fuzzy overlapping community such that the following conditions hold:

$$0 \leq \alpha_{i\psi} \leq 1 \quad \forall i \in V,\ \ \psi \in \Psi \text{ and } \sum_{\psi \in \Psi} \alpha_{i\psi} = 1. \tag{41}$$

Modified modularity for overlapping communities is defined as follows [125]:

$$Q_{\text{ov}} = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \alpha_{i\psi} \alpha_{j\psi}. \tag{42}$$

b) f-Modularity: Utilizes information-theoretical perspective to detect both disjoint and overlapping communities. Its value is directly dependent on the characteristics of a community. In particular, the value of f-Modularity decreases as communities are contracted and vanishes when there is no community structure. Moreover, it is based on dual f-mutual information and hence, contains the mutual information properties also [126] and is defined as follows:

$$\begin{aligned} &\text{Mod}^f (G) \\ &\quad := \max_{D \in C} \sum_{u,v} \left(\partial f \left(D_{u,v} \right) F_{u,v} - f^* \left(\partial f \left(D_{u,v} \right) \right) J_{u,v} \right). \end{aligned} \tag{43}$$

In (43), Mod^f indicates f-Modularity, D is a matrix which acts as a distinguisher between frequency matrix F and random matrix J. To represent different instances of f-modularity, different convex functions f and constraint sets C are used.

c) Conductance: It is also used to evaluate disjoint communities. In contrast to the modularity, it considers the external connections as well for the evaluation purpose. It is based on the concept that the set of vertices that are isolated from the rest of the graph are considered to form prominent communities. It considers the structural cohesiveness of the communities [3] for the evaluation purpose. Conductance of predicted communities C is defined as follows:

$$f(w) = \frac{\left| E_w^{\text{out}} \right|}{2\left| E_w^{\text{in}} \right| + \left| E_w^{\text{out}} \right|} \tag{44}$$

where E_w^{out} signifies number of edges that are outside the community and E_w^{out} indicates number of edges that are inside the community. Low conductance and high number of internal connections as compared to the number of external connections yields better communities.

d) Permanence [127]: It is recently developed measure for disjoint communities. In contrast to the conductance, it not only considers the idea that the number of internal connections must be more than the number of external connections but this approach also considers the strength of internal connections which indicates the viability of the set of vertices of a community to remain in its assigned community.

Permanence is defined based on the inclination of a vertex to remain in its assigned community. While computing permanence, only the pull from the neighboring community having maximum number of edges, $E_{\max}$ have been considered and the strength of internal connectivity of a vertex is measured by internal clustering coefficient, $c_{\text{in}}(v)$. Permanence value of a vertex ranges from $+1$ (indicates that the vertex is strongly connected to the assigned community) to -1 (vertex is weakly and perhaps wrongly assigned to the current community). If the value of permanence is 0, it means that the vertex is equally externally connected by all its neighboring communities. The two criteria mentioned above has been aggregated and finally the permanence of a vertex v is given by the following formula:

$$\text{Perm}(v) = \left[\frac{I(v)}{E_{\max}(v)} \times \frac{1}{D(v)} \right] - [1 - c_{\text{in}}(v)] \tag{45}$$

where $I(v)$ indicates the internal connections, $D(v)$ indicates the total degree of vertex v. Permanence of the network can be computed by summation of the permanence of all the vertices and then normalize it by dividing it by the total number of vertices. It signifies the extent to which the vertices of a network are bound to their respective communities [127].

Here, the internal clustering coefficient for a single vertex say v which is termed as internal clustering coefficient is defined by [28]

$$c_{\text{in}}(v) = \frac{N_v}{V_n} \tag{46}$$

where N_v is number of pairs of neighbors of v that are connected and V_n is number of pairs of neighbors of v.

Average normalized unifiability and isolability *(ANUI) [96]* is used to analyze the quality of overlapping communities. It is preferable for sparsely connected networks. ANUI assigns equal weightage to both unifiability and isolability in the evaluation process of the communities. The strategy is to restrict a cluster from unifying with other communities and in addition to that the communities should isolate themselves from the rest of the network. For a graph G that comprises k communities C, the ANUI is defined as follows:

$$\text{ANUI}(G, C) = \frac{Q_{\text{AUI}}(G, C)}{2} \tag{47}$$

$$= \frac{Q_{\text{AVI}}(G, C)}{1 + Q_{\text{AVU}}(G, C) \times Q_{\text{AVI}}(G, C)}. \tag{48}$$

In (48), Q_{AVU} is given by

$$Q_{\text{AVU}}(G, C) = \frac{1}{k} \sum_{i=1}^{k} \text{Unifiability } (C_i). \tag{49}$$

In (49), the term Unifiability of community C_i to another community C_j is given by

$$\text{Unifiability } (C_i) = \sum_{i=1}^{k} \text{Unifiability } \left(C_i, C_j \right). \tag{50}$$

Also, in (48), Q_{AVI} is given by

$$Q_{\text{AVI}} = \frac{1}{k} \sum_{i=1}^{k} \text{Isolability } (C_i). \tag{51}$$

In (51), the term Isolability (C_i) is given by

$$\text{Isolability } (C_i) = \frac{\left\{ (u, v) \mid u \in_{C_i} v \right\}}{\left\{ \{ (u, v); (u, w) \} \mid u \in_{C_i} v \ \& w \notin C_i \right\}}. \tag{52}$$

The nNumerator in (52) indicates connections within the community C_i and denominator is the total number of connections of the community.

C. Evaluation Methods

The evaluation method is crucial for evaluating community detection algorithms. Though there is no definite approach to follow, a value-based method is prevalent for evaluating community detection algorithms. In a value-based method, simply different metrics (explained in the previous subsection) are computed to determine the accuracy and quality of the communities. However, since each of the metrics measures communities from different perspectives, with value-based method poses difficulty when multiple metrics are there. Therefore, several other alternative methods have been developed. In the following subsections, various methods which are developed recently to evaluated community detection algorithms are discussed.

1) Value Based Analysis: A value-based method is quite simple and an intuitive approach, which is being widely used for evaluating communities identified by community detection algorithms. Simply any metric value is computed to evaluate the accuracy or quality of the communities. However, the problem arises when multiple metrics are there and compared the performance of one algorithm with others. This is because higher values of some metrics indicate better performance of the algorithm, while lower values of some metrics may indicate better performance of the algorithm. Biswas and Biswas [128] have introduced the notion of qualitative accuracy measure of the community detection algorithms. Both accuracy and quality of the community detection algorithms are analyzed considering their respective metrics. They have noted that a higher value of accuracy of a given community detection algorithm does not guarantee a higher value of quality of the given algorithm and vice versa. Moreover, some literature works have also reported bias while evaluating communities [98], [99]. Hence, alternative approaches are developed to incorporate the indications of different metrics.

2) Multiple Criterion Decision Making (MCDM): This community detection evaluation method makes a decision based on multiple criteria, i.e., multiple metrics. It combines indications of different metrics to generate a single score. Values obtained for multiple metrics with multiple algorithms are considered to generate a score. A decision matrix, say $M_{m\times n}$ is considered, where m and n represent the number of criterion and alternative, respectively. Each entry M_{ij} indicates the score for jth alternative or algorithm in terms of ith criteria or evaluation metric. To predict the best alternative, a relative score of all the alternatives is computed using technique for order preference by similarity to ideal selection (TOPSIS) [129], [130]. TOPSIS first assigns a customized weight to each of the criteria, then the decision matrix, $M_{m\times n}$ is normalized for each criteria by

$$\overline{X_{ij}} = \frac{X_{ij}}{\sqrt{\sum_{j=1}^{n} X_{ij}^2}}. \tag{53}$$

Next, the weights associated with each criteria have been multiplied with their respective normalized score to obtain the weighted normalized decision matrix. Then, TOPSIS selects the ideal best score, V_i^+ and ideal worst score, V_i^- by selecting the maximum and minimum score, respectively, for beneficial criteria from the weighted normalized decision matrix. But V_i^+ and V_i^- should be the minimum and maximum score for nonbeneficial criteria. Thus, to predict the best alternative, a relative performance score P_j on the ideal best and ideal worst values using Euclidean distance is given as

$$P_j = \frac{S_j^-}{S_j^+ + S_j^-}. \tag{54}$$

S_j^+ and S_j^- in (54) is given by

$$S_j^+ = \sqrt{\left[\sum_{i=1}^{m}\left(V_{ij} - V_i^+\right)^2\right]} \tag{55}$$

$$S_j^- = \sqrt{\left[\sum_{i=1}^{m}\left(V_{ij} - V_i^-\right)^2\right]}. \tag{56}$$

The separation with ideal best and ideal worst should be minimum and maximum, respectively, for the best results. The best algorithm gives the highest score.

3) Relative Inclination Toward Accuracy (RITA): MCDM techniques are very useful for comparing the performance of multiple community detection algorithms in terms of multiple metrics. However, community detection algorithms are evaluated on two aspects: accuracy and quality. Different metrics have been developed to ensure both aspects. MCDM technique only generates an overall relative score for the algorithm-based multiple metrics considered. The overall score generated with the MCDM technique cannot express the accuracy or quality of communities explicitly when both kinds of metrics are considered. RITA framework [112] overcomes the drawback of MCDM techniques. With RITA analysis inclination of the algorithm toward accuracy or quality can be easily determined. It uses both accuracy and quality metrics as the MCDM evaluation method discussed in Section V-C2 in the evaluation process. First, a relative percentage weight has been assigned to each category. Relative percentage weight means that if the accuracy metrics have been assigned with w% weight, then the other metrics get $(100 - w)$% weight. If there are m accuracy metrics, the weight is divided equally among all the metrics as given by the following formula:

$$w_A = \frac{1}{m}(w) \tag{57}$$

and if there are n quality metrics, then the weight is distributed as

$$w_Q = \frac{1}{n}(w). \tag{58}$$

The weights of each of the metrics w_A and w_Q are updated with each iteration till the total ending weight say e is reached. The relative score of performance of each of the community detection algorithm has been obtained with respect to relative weight assignment to different metrics using TOPSIS method [129].

4) Visual Analysis: The visual analysis method [131] initially assumes a set of community detection algorithms, a set of evaluation metrics, and a set of datasets. Evaluation metrics from both categories, i.e., quality and accuracy are considered to analyze community detection algorithms through one-to-many quality comparison and one-to-many accuracy comparison, respectively. At first, t trials are considered for each of the algorithms and optimized results of the t trials are used for both the actor algorithm and each of the competitor algorithms. Then, using quantile–quantile plot and simple linear regression (SLR), the performance of the actor and competitor algorithms are analyzed. Quantile–quantile plot is used in the evaluation process to ensure the involvement of each data. Regression line (RL) is used to evaluate the dominance of the algorithms by observing the angle between RL and neutral line (NL) as well as the intersection position of RL and NL. If a higher value implies higher dominance, then RL below NL implies actor dominance over competitor dominance.

VI. Conclusion

This article surveyed the emergence of deploying information diffusion for community detection in OSN. Dissemination of information plays an important role in the formation of communities implicitly in OSN. This survey analyzed various factors affecting communities and their dependence on information flow. Primarily two categories of factors, network properties and social facets are studied extensively from the viewpoint of information flow and how influencing those factors imply a proportionate effect on the communities. The network properties measure the structural or topological aspect of the network, while social facets quantify different facets or aspects of social interaction. An in-depth study of information diffusion models is carried out to explore the possibility of utilizing those to influence different network properties and social facets.

The study reveals that all the diffusion models regardless of whether it is receiver-centric, sender-centric, or profit-centric utilizes mostly the network properties (e.g., belongingness measure, structural equivalence, and edge strength measures) while taking any decision. Therefore, the deployment of diffusion models is more prevalent to influence the network properties for community detection by incorporating the effects on those properties. In the same way, those diffusion models that utilize social facets are also incorporated in the community detection process. However, the selection of an appropriate diffusion model plays an important role in the success of deploying information diffusion for community detection. Inappropriate cascades may lead to inaccurate communities while cascade-based diffusion models are deployed for community detection. On the other hand, discovering the Nash equilibrium poses additional difficulty in the game theoretic-based community detection. Since discovering the Nash equilibrium is an NP-hard problem, the complexity of the game theoretic approach is high as group decision of all the agents is needed to be considered. It has also been observed that utilization of more than one network properties in diffusion-based algorithms is beneficial in detecting overlapping communities and social facets are better suited for disjoint communities. Evaluation of identified communities is necessary irrespective of whether communities are identified incorporating information diffusion in the community detection process or not. The survey includes recently developed evaluation methodologies, widely used evaluation metrics and repositories of datasets to evaluate community detection algorithms.

The proliferation of information across the network is the key to the deployment of information diffusion for community detection. However, the dynamics of OSN consequently lead to a major challenge for the detection of communities using information diffusion. In fact, analyzing information diffusion in dynamic networks itself remains one of the challenging tasks. Moreover, present day OSN exhibits multimodal, multirelational, multidimensional networks. Handling such networks in the context of deploying information diffusion for community detection can be an interesting future direction of research. Another future direction could be handling heterogeneity within network as well as among the network from the perspective of information diffusion deployment in community detection.

References

[1] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 1994.

[2] J. Scott, *Social Network Analysis*. Newbury Park, CA, USA: Sage, 2017.

[3] T. Chakraborty, A. Dalmia, A. Mukherjee, and N. Ganguly, "Metrics for community analysis: A survey," *ACM Compt. Surv.*, vol. 50, no. 4, pp. 1–37, Aug. 2017.

[4] M. Wani and M. Ahmad, "Information diffusion modelling and social network parameters (A survey)," in *Proc. Int. Conf. Adv. Comput., Commun. Electron. Eng.*, Kashmir, India, Mar. 2015, pp. 87–91.

[5] M. Granovetter, "Threshold models of collective behavior," *Amer. J. Sociol.*, vol. 83, no. 6, pp. 1420–1443, May 1978.

[6] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Washington, DC, USA, 2003, pp. 137–146.

[7] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Lett.*, vol. 12, no. 3, pp. 211–223, Aug. 2001.

[8] C. Jiang, Y. Chen, and K. J. R. Liu, "Graphical evolutionary game for information diffusion over social networks," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 4, pp. 524–536, Aug. 2014.

[9] C. Jiang, Y. Chen, and K. J. R. Liu, "Evolutionary dynamics of information diffusion over social networks," *IEEE Trans. Signal Process.*, vol. 62, no. 17, pp. 4573–4586, Sep. 2014.

[10] W. Luo, W. P. Tay, and M. Leng, "Infection spreading and source identification: A hide and seek game," *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4228–4243, Aug. 2016.

[11] J. Ok, Y. Jin, J. Shin, and Y. Yi, "On maximizing diffusion speed over social networks with strategic users," *IEEE/ACM Trans. Netw.*, vol. 24, no. 6, pp. 3798–3811, Dec. 2016.

[12] Z. Wang, Y. Wu, Q. Li, F. Jin, and W. Xiong, "Link prediction based on hyperbolic mapping with community structure for complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 450, pp. 609–623, May 2016.

[13] P. Moradi, S. Ahmadian, and F. Akhlaghian, "An effective trust-based recommendation method using a novel graph clustering algorithm," *Phys. A, Stat. Mech. Appl.*, vol. 436, pp. 462–481, Oct. 2015.

[14] Z. Wang, Z. Li, G. Yuan, Y. Sun, X. Rui, and X. Xiang, "Tracking the evolution of overlapping communities in dynamic social networks," *Knowl.-Based Syst.*, vol. 157, pp. 81–97, Oct. 2018.

[15] M. Plantié and M. Crampes, "Survey on social community detection," in *Social Media Retrieval*. London, U.K.: Springer, 2013, pp. 65–85.

[16] K. Shen, L. Song, X. Yang, and W. Zhang, "A hierarchical diffusion algorithm for community detection in social networks," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery*, Huangshan, China, Oct. 2010, pp. 276–283.

[17] H. Alvari, A. Hajibagheri, and G. Sukthankar, "Community detection in dynamic social networks: A game-theoretic approach," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Beijing, China, Aug. 2014, pp. 101–107.

[18] Y. Zhang, T. Lyu, and Y. Zhang, "Cosine: Community-preserving social network embedding from information diffusion cascades," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, New Orleans, LA, USA, Feb. 2017, pp. 2620–2627.

[19] M. Sattari and K. Zamanifar, "A cascade information diffusion based label propagation algorithm for community detection in dynamic social networks," *J. Comput. Sci.*, vol. 25, pp. 122–133, Mar. 2018.

[20] M. Ramezani, A. Khodadadi, and H. R. Rabiee, "Community detection using diffusion information," *ACM Trans. Knowl. Discovery from Data*, vol. 12, no. 2, pp. 1–22, Mar. 2018.

[21] Z. Sun, B. Wang, J. Sheng, Z. Yu, and J. Shao, "Overlapping community detection based on information dynamics," *IEEE Access*, vol. 6, pp. 70919–70934, Nov. 2018.

[22] J. Zhang and S. Y. Philip, "Information diffusion," in *Broad Learning Through Fusions*. Cham, Switzerland: Springer, 2019, pp. 315–349.

[23] A. Biswas and B. Biswas, "FuzAg: Fuzzy agglomerative community detection by exploring the notion of self-membership," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 2568–2577, Oct. 2018.

[24] M. S. Granovetter, "The strength of weak ties–American journal of sociology," *Amer. J. Sociol.*, vol. 78, no. 6, pp. 1360–1380, May 1973.

[25] A. Landherr, B. Friedl, and J. Heidemann, "A critical review of centrality measures in social networks," *Bus. Inf. Syst. Eng.*, vol. 2, no. 6, pp. 371–385, Oct. 2010.

[26] L. Tang and H. Liu, "Community detection and mining in social media," *Synth. Lectures Data Mining Knowl. Discovery*, vol. 2, no. 1, pp. 1–137, Jan. 2010.

[27] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: A discriminative approach," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Paris, France, Jun. 2009, pp. 927–936.

[28] M. Newman, *Networks: An Introduction*. Oxford, U.K.: Oxford Univ. Press, 2010.

[29] K. Steinhaeuser and N. V. Chawla, "Community detection in a large real-world social network," in *Social Computing, Behavioral Modeling, and Prediction*. Boston, MA, USA: Springer, 2008, pp. 168–175.

[30] M. Yazdani, A. Moeini, M. Mazoochi, F. Rahmani, and L. Rabiei, "A new follow based community detection algorithm," in *Proc. 6th Int. Conf. Web Res. (ICWR)*, Tehran, Iran, Apr. 2020, pp. 197–202.

[31] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Selecting information diffusion models over social networks for behavioral analysis," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Berlin, Heidelberg, Sep. 2010, pp. 180–195.

[32] Q. Ma, X. Luo, and H. Zhuge, "Finding influential users of Web event in social media," *Concurrency Comput., Pract. Exper.*, vol. 31, no. 3, p. e5029, 2019.

[33] M. Kimura, K. Saito, and R. Nakano, "Extracting influential nodes for information diffusion on a social network," in *Proc. AAAI*, Vancouver, BC, Canada, Jul. 2007, pp. 1371–1376.

[34] M. Kitsak *et al.*, "Identification of influential spreaders in complex networks," *Nature Phys.*, vol. 6, no. 11, pp. 888–893, Aug. 2010.

[35] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proc. 13th Conf. World Wide Web (WWW)*, New York, NY, USA, 004, pp. 491–501.

[36] R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: Introduction*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[37] M. Li, X. Wang, K. Gao, and S. Zhang, "A survey on information diffusion in online social networks: Models and methods," *Information*, vol. 8, no. 4, p. 118, Sep. 2017.

[38] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *ACM SIGMOD Rec.*, vol. 42, no. 2, pp. 17–28, May 2013.

[39] W. Chen, L. V. S. Lakshmanan, and C. Castillo, "Information and influence propagation in social networks," *Synth. Lectures Data Manage.*, vol. 5, no. 4, pp. 1–177, Oct. 2013.

[40] S. Shelke and V. Attar, "Source detection of rumor in social network–A review," *Online Social Netw. Media*, vol. 9, pp. 30–42, Jan. 2019.

[41] A. Guille and H. Hacid, "A predictive model for the temporal dynamics of information diffusion in online social networks," in *Proc. 21st Int. Conf. companion World Wide Web (WWW Companion)*, Lyon, France, 2012, pp. 1145–1152.

[42] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Proc. IEEE Int. Conf. Data Mining*, Sydney, NSW, Australia, Dec. 2010, pp. 88–97.

[43] S. S. Singh, K. Singh, A. Kumar, H. K. Shakya, and B. Biswas, "A survey on information diffusion models in social networks," in *Proc. Int. Conf. Adv. Inform. Comput. Res.*, Shimla, India, Jun. 2018, pp. 426–439.

[44] D. Peleg, "Local majority voting, small coalitions and controlling monopolies in graphs: A review," in *Proc. 3rd Colloq. Struct. Inf. Commun. Complex.*, Siena, Italy, Oct. 1996, pp. 152–169.

[45] H. Zhang, S. Mishra, M. T. Thai, J. Wu, and Y. Wang, "Recent advances in information diffusion and influence maximization in complex social networks," in *Opportunistic Mobile Social Networks*. Boca Raton, FL, USA: CRC Press, 2014, pp. 1–37.

[46] B. Abrahao, F. Chierichetti, R. Kleinberg, and A. Panconesi, "Trace complexity of network inference," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Chicago, IL, USA, Aug. 2013, pp. 491–499.

[47] J. Leskovec, A. Singh, and J. Kleinberg, "Patterns of influence in a recommendation network," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, Singapore, Apr. 9–12, 2006, pp. 380–389.

[48] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, Zagreb, Croatia, Sep. 2008, pp. 67–75.

[49] A. Kumari and S. N. Singh, "Online influence maximization using rapid continuous time independent cascade model," in *Proc. 7th Int. Conf. Cloud Comput., Data Sci. Eng. (Confluence)*, Noida, India, Jan. 2017, pp. 356–361.

[50] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proc. 23rd Int. Conf. World Wide Web (WWW)*, Seoul, South Korea, 2014, pp. 925–936.

[51] X. Yu and T. Chu, "Learning the structure of influence diffusion in the independent cascade model," in *Proc. 36th Chin. Control Conf. (CCC)*, Dalian, China, Jul. 2017, pp. 5647–5651.

[52] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Learning continuous-time information diffusion model for social behavioral data analysis," in *Proc. Asian Conf. Mach. Learn.*, Nanjing, China, Nov. 2009, pp. 322–337.

[53] D. Kempe, J. Kleinberg, and É. Tardos, "Influential nodes in a diffusion model for social networks," in *Proc. Int. Colloq. Automata, Lang., Program.*, Berlin, Germany, Jul. 2005, pp. 1127–1138.

[54] N. Alon, M. Feldman, A. D. Procaccia, and M. Tennenholtz, "A note on competitive diffusion through social networks," *Inf. Process. Lett.*, vol. 110, no. 6, pp. 221–225, Feb. 2010.

[55] H. Alvari, S. Hashemi, and A. Hamzeh, "Detecting overlapping communities in social networks by game theory and structural equivalence concept," in *Proc. Int. Conf. Artif. Intell. Comput. Intell.*, Jul. 2011, pp. 620–630.

[56] Q. Hang, J. Zhu, B. Song, and N. Zhang, "Game model of information transmission in social networks," *J. Chin. Comput. Syst.*, vol. 35, pp. 473–477, Apr. 2014.

[57] Y. Wang, J. Yu, W. Qu, H. Shen, X. Cheng, and C. Lin, "Evolutionary game model and analysis methods for network group behavior," *Chin. J. Comput.*, vol. 38, no. 2, pp. 282–300, 2015.

[58] D. Liu, Y. Wang, Y. Jia, J. Li, and Z. Yu, "From strangers to neighbors: Link prediction in microblogs using social distance game," in *Proc. Diffusion Netw. Cascade Anal. (WSDM)*, New York, NY, USA, Feb. 2014, pp. 1–4.

[59] T. Li and M. Shigeno, "Nash equilibria for information diffusion games on weighted cycles and paths," *J. Oper. Res. Soc. Jpn.*, vol. 64, no. 1, pp. 1–11, Jan. 2021.

[60] H. Gong, C. Guo, and Y. Liu, "Measuring network rationality and simulating information diffusion based on network structure," *Phys. A, Stat. Mech. Appl.*, vol. 564, Feb. 2021, Art. no. 125501.

[61] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection," in *Encyclopedia of Social Network Analysis and Mining*. New York, NY, USA: Springer, 2018, pp. 301–312.

[62] Y. Zhou, H. Cheng, and J. X. Yu, "Clustering large attributed graphs: An efficient incremental approach," in *Proc. IEEE Int. Conf. Data Mining*, Sydney, NSW, Australia, Dec. 2010, pp. 689–698.

[63] Y. Ruan, D. Fuhry, and S. Parthasarathy, "Efficient community detection in large networks using content and links," in *Proc. 22nd Int. Conf. World Wide Web (WWW)*, 2013, pp. 1089–1098.

[64] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dallas, TX, USA, Dec. 2013, pp. 1151–1156.

[65] B. Yang and S. Manandhar, "Community discovery using social links and author-based sentiment topics," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Beijing, China, Aug. 2014, pp. 580–587.

[66] C. Peng, T. G. Kolda, and A. Pinar, "Accelerating community detection by using K-core subgraphs," 2014, *arXiv:1403.2226*. [Online]. Available: http://arxiv.org/abs/1403.2226

[67] Y. Li, J. Fan, Y. Wang, and K.-L. Tan, "Influence maximization on social graphs: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 10, pp. 1852–1872, Oct. 2018.

[68] J. Guo, P. Zhang, C. Zhou, Y. Cao, and L. Guo, "Personalized influence maximization on social networks," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, San Francisco, CA, USA, 2013, pp. 199–208.

[69] Y. Li, D. Zhang, and K.-L. Tan, "Real-time targeted influence maximization for online advertisements," in *Proc. VLDB Endowment: 41st Int. Conf. VLDB Endowment*, Kohala Coast, HI, USA, Aug. 2015, pp. 1070–1081.

[70] H. T. Nguyen, T. N. Dinh, and M. T. Thaip, "Cost-aware targeted viral marketing in billion-scale networks," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, San Francisco, CA, USA, Apr. 2016, pp. 1–9.

[71] C. Wang, X. Guan, T. Qin, and Y. Zhou, "Modelling on opinion leader–influence in microblog message propagation and its application," *J. Softw.*, vol. 26, pp. 1473–1485, Jun. 2015.

[72] B. Chen, X. Tang, L. Yu, and Y. Liu, "Identifying method for opinion leaders in social network based on competency model," *J. Commun.*, vol. 35, no. 11, pp. 12–22, 2014.

[73] J.-X. Mao, Y.-Q. Liu, M. Zhang, and S.-P. Ma, "Social influence analysis for micro-blog user based on user behavior," *Chin. J. Comput.*, vol. 37, no. 4, pp. 791–800, 2014.

[74] X. Wu, H. Zhang, X. Zhao, B. Li, and C. Yang, "Mining algorithm of microblogging opinion leaders based on user-behavior network," *Appl. Res. Comput*, vol. 32, pp. 2678–2683, May 2015.

[75] F. Ullah and S. Lee, "Identification of influential nodes based on temporal-aware modeling of multi-hop neighbor interactions for influence spread maximization," *Phys. A, Stat. Mech. Appl.*, vol. 486, pp. 968–985, Nov. 2017.

[76] Z. Wang, C. Sun, J. Xi, and X. Li, "Influence maximization in social graphs based on community structure and node coverage gain," *Future Gener. Comput. Syst.*, vol. 118, pp. 327–338, May 2021.

[77] F. Ghayour-Baghbani, M. Asadpour, and H. Faili, "MLPR: Efficient influence maximization in linear threshold propagation model using linear programming," *Social Netw. Anal. Mining*, vol. 11, no. 1, pp. 1–10, Dec. 2021.

[78] A. Ullah, B. Wang, J. Sheng, J. Long, and N. Khan, "Identification of influential nodes via effective distance-based centrality mechanism in complex networks," *Complexity*, vol. 2021, pp. 1–16, Feb. 2021.

[79] R. Olivares, F. Muñoz, and F. Riquelme, "A multi-objective linear threshold influence spread model solved by swarm intelligence-based methods," *Knowl.-Based Syst.*, vol. 212, Jan. 2021, Art. no. 106623.

[80] S. Kianian and M. Rostamnia, "An efficient path-based approach for influence maximization in social networks," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114168.

[81] E. Carnia, B. Fermadona, H. Napitupulu, N. Anggriani, and A. Supriatna, "Implementation of centrality measures in graph represented information spreads with hashtag# bersatulawancorona in Twitter," in *Proc. J. Phys., Conf.*, Sanur-Bali, Indonesia, Jan. 2021, Art. no. 012068.

[82] P. Kumar and A. Sinha, "Information diffusion modeling and analysis for socially interacting networks," *Social Netw. Anal. Mining*, vol. 11, no. 1, pp. 1–18, Jan. 2021.

[83] A. Namtirtha, A. Dutta, B. Dutta, A. Sundararajan, and Y. Simmhan, "Best influential spreaders identification using network global structural properties," *Sci. Rep.*, vol. 11, no. 1, pp. 1–15, Dec. 2021.

[84] H. Kalantari, M. Ghazanfari, M. Fathian, and K. Shahanaghi, "Multi-objective optimization model in a heterogeneous weighted network through key nodes identification in overlapping communities," *Comput. Ind. Eng.*, vol. 144, Jun. 2020, Art. no. 106413.

[85] S. Bhattacharya and D. Sarkar, "Study on information diffusion in online social network," in *Proc. Int. Conf. Frontiers Comput. Syst.*, West Bengal, India, Nov. 2021, pp. 279–288.

[86] S. Nair, K. W. Ng, A. Iamnitchi, and J. Skvoretz, "Diffusion of social conventions across polarized communities: An empirical study," *Social Netw. Anal. Mining*, vol. 11, no. 1, pp. 1–17, Dec. 2021.

[87] M. Lahiri and M. Cebrian, "The genetic algorithm as a general diffusion model for social networks," in *Proc. 24th AAAI Conf. Artif. Intell.*, Atlanta, GA, USA, Jul. 11–15, 2010.

[88] N. Mozafari and A. Hamzeh, "An enriched social behavioural information diffusion model in social networks," *J. Inf. Sci.*, vol. 41, no. 3, pp. 273–283, Jun. 2015.

[89] P.-Z. Li, L. Huang, C.-D. Wang, J.-H. Lai, and D. Huang, "Community detection by motif-aware label propagation," *ACM Trans. Knowl. Discovery Data*, vol. 14, no. 2, pp. 1–19, Mar. 2020.

[90] N. N. Sein, "Overlapping community detection using local seed expansion," *Int. J. Comput.*, vol. 37, no. 1, pp. 27–34, 2020.

[91] Z. Liu, H. Wang, L. Cheng, W. Peng, and X. Li, "Temporal label walk for community detection and tracking in temporal network," *Appl. Sci.*, vol. 9, no. 15, p. 3199, Aug. 2019.

[92] V. Red, E. D. Kelsic, P. J. Mucha, and M. A. Porter, "Comparing community structure to characteristics in online collegiate social networks," *SIAM Rev.*, vol. 53, no. 3, pp. 526–543, Jan. 2011.

[93] L. Blumrosen and N. Nisan, "Combinatorial auctions," *Algorithmic Game theory*, vol. 267, pp. 267–298, Jun. 2007.

[94] A. Bhih, P. Johnson, and M. Randles, "An optimisation tool for robust community detection algorithms using content and topology information," *J. Supercomput.*, vol. 76, no. 1, pp. 226–254, Jan. 2020.

[95] Z. Sun, J. Sheng, B. Wang, A. Ullah, and F. Khawaja, "Identifying communities in dynamic networks using information dynamics," *Entropy*, vol. 22, no. 4, p. 425, Apr. 2020.

[96] A. Biswas and B. Biswas, "Defining quality metrics for graph clustering evaluation," *Expert Syst. Appl.*, vol. 71, pp. 1–17, Apr. 2017.

[97] J. Creusefond, T. Largillier, and S. Peyronnet, "On the evaluation potential of quality functions in community detection for different contexts," in *Proc. Int. Conf. School Netw. Sci.*, Cham, Switzerland, Jan. 2016, pp. 111–125.

[98] H. Almeida, D. Guedes, W. Meira, and M. J. Zaki, "Is there a best quality metric for graph clusters?" in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Athens, Greece, Sep. 2011, pp. 44–59.

[99] G. K. Orman, V. Labatut, and H. Cherifi, "Comparative evaluation of community detection algorithms: A topological approach," *J. Stat. Mech.: Theory Exp.*, vol. 2012, no. 8, Aug. 2012, Art. no. P08001.

[100] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, Apr. 2002.

[101] A. Lancichinetti and S. Fortunato, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, no. 1, Jul. 2009, Art. no. 016118.

[102] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 78, no. 4, Oct. 2008, Art. no. 046110.

[103] L. Weng, F. Menczer, and Y.-Y. Ahn, "Virality prediction and community structure in social networks," *Sci. Rep.*, vol. 3, no. 1, pp. 1–6, Aug. 2013.

[104] D. Greene, D. Doyle, and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in *Proc. Int. Conf. Adv. Social Netw. Anal. Mining*, Odense, Denmark, Aug. 2010, pp. 176–183.

[105] *SNAP Datasets: Stanford Large Network Dataset Collection*. Accessed: Mar. 27, 2021. [Online]. Available: http://snap.stanford.edu/data

[106] *Network Repository. an Interactive Scientific Network Data Repository*. Accessed: Jun. 25, 2020. [Online]. Available: http://networkrepository.com/

[107] *The UCI Network Data Repository is an Effort to Facilitate the Scientific Study of Networks*. Accessed: Jun. 25, 2020. [Online]. Available: https://networkdata.ics.uci.edu/resources.php

[108] *Kaggle Forum*. Accessed: Jun. 25, 2020. [Online]. Available: https://www.kaggle.com/general/2122

[109] *Statistical Relational Learning Group*. Accessed: Jul. 1, 2020. [Online]. Available: https://linqs.soe.ucsc.edu/data

[110] *Datasets for Social Network Analysis*. Accessed: Jul. 1, 2020. [Online]. Available: https://www.aminer.org/data-sna

[111] *The Koblenz Network Collection*. Accessed: Jul. 1, 2020. [Online]. Available: http://konect.uni-koblenz.de/networks/

[112] A. Biswas and B. Biswas, "A framework for analyzing community detection algorithms," in *Proc. IEEE Students' Technol. Symp. (TechSym)*, IIT Kharagpur, India, Sep. 2016, pp. 61–66.

[113] Z. Li, Y. Hu, B. Xu, Z. Di, and Y. Fan, "Detecting the optimal number of communities in complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 391, no. 4, pp. 1770–1776, Feb. 2012.

[114] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," *Natural Lang. Eng.*, vol. 16, no. 1, pp. 100–103, Jan. 2010.

[115] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, Dec. 2002.

[116] S. Fortunato and A. Lancichinetti, "Community detection algorithms: A comparative analysis: Invited presentation, extended abstract," in *Proc. 4th Int. ICST Conf. Perform. Eval. Methodologies Tools*, Pisa, Italy, 2009, p. 27.

[117] A. Amelio and C. Pizzuti, "Is normalized mutual information a fair measure for comparing community detection methods?" in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Paris, France, Aug. 2015, pp. 1584–1585.

[118] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, "Extending the definition of modularity to directed graphs with overlapping communities," *J. Stat. Mech.: Theory Exp.*, vol. 2009, no. 03, Mar. 2009, Art. no. P03024.

[119] M. Hoffman, D. Steinley, and M. J. Brusco, "A note on using the adjusted rand index for link prediction in networks," *Social Netw.*, vol. 42, pp. 72–79, Jul. 2015.

[120] G. W. Milligan and M. C. Cooper, "A study of the comparability of external criteria for hierarchical cluster analysis," *Multivariate Behav. Res.*, vol. 21, no. 4, pp. 441–458, Oct. 1986.

[121] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985.

[122] T. Y. Lin, S. Ohsuga, C.-J. Liau, and X. Hu, *Foundations and Novel Approaches in Data Mining*. Berlin, Germany: Springer-Verlag, 2005.

[123] Y. Xing, F. Meng, Y. Zhou, G. Sun, and Z. Wang, "Overlapping community detection extended from disjoint community structuxingre," *Comput. Informat.*, vol. 38, no. 5, pp. 1091–1110, Feb. 2020.

[124] M. Arasteh and S. Alizadeh, "A fast divisive community detection algorithm based on edge degree betweenness centrality," *Int. J. Speech Technol.*, vol. 49, no. 2, pp. 689–702, Feb. 2019.

[125] H.-W. Shen, X.-Q. Cheng, and J.-F. Guo, "Quantifying and identifying the overlapping community structure in networks," *J. Stat. Mech.: Theory Exp.*, vol. 2009, no. 07, Jul. 2009, Art. no. P07042.

[126] Y. Guo, Z. Huang, Y. Kong, and Q. Wang, "Modularity and mutual information in networks: Two sides of the same coin," 2021, *arXiv:2103.02542*. [Online]. Available: http://arxiv.org/abs/2103.02542

[127] T. Chakraborty, S. Srinivasan, N. Ganguly, A. Mukherjee, and S. Bhowmick, "On the permanence of vertices in network communities," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Aug. 2014, pp. 1396–1405.

[128] A. Biswas and B. Biswas, "Investigating community structure in perspective of ego network," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 6913–6934, Nov. 2015.

[129] G. Kou, Y. Peng, and G. Wang, "Evaluation of clustering algorithms for financial risk analysis using MCDM methods," *Inf. Sci.*, vol. 275, pp. 1–12, Aug. 2014.

[130] C.-L. Hwang and K. Yoon, *Multiple Attribute Decision Making: Methods and Applications a State-of-the-Art Survey*. Berlin, Germany: Springer-Verlag, 2012.

[131] A. Biswas and B. Biswas, "Analyzing evolutionary optimization and community detection algorithms using regression line dominance," *Inf. Sci.*, vol. 396, pp. 185–201, Aug. 2017.

Soumita Das received the B.Tech. degree in information technology from North Eastern Hill University (N.E.H.U), Shillong, India, and the M.Tech. degree in computer science and engineering from Tezpur University. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, National Institute of Technology Silchar (NITs), Silchar, India.

Her research interests include social network analysis.

Anupam Biswas received the B.E. degree in computer science and engineering from the Nehru National Institute of Technology Allahabad, Prayagraj, India, in 2013, the M.Tech. degree in computer science and engineering from the Jorhat Engineering College, Jorhat, India, in 2011, and the Ph.D. degree in computer science and engineering from IIT Varanasi, Varanasi, India, in 2017.

He is currently working as an Assistant Professor with the Department of Computer Science and Engineering, National Institute of Technology Silchar, Silchar, India. He has authored or coauthored several research articles in reputed international journals, conference, and book chapters. His research interests include machine learning, deep learning, computational music, information retrieval, social networks, and evolutionary computation.

Dr. Biswas has served as a Program Chair for the International Conference on Big Data, Machine Learning and Applications (BigDML 2019). He has served as a General Chair for the 25th International Symposium on Frontiers of Research in Speech and Music (FRSM 2020) and co-edited proceedings of FRSM 2020 published as book volume in Springer AISC Series. He has edited three books titled *Health Informatics: A Computational Perspective in Healthcare*, *Principles of Social Networking: The New Horizon and Emerging Challenges*, and *Principles of Big Graph: In-depth Insight* in different Springer and Elsevier book series.