Review of Community Detection over Social Media: Graph Prospective

Pranita Jain[1], Deepak Singh Tomar[2]
Department of Computer Science
Maulana Azad National Institute of Technology
Bhopal, India 462001

***Abstract*—Community over the social media is the group of globally distributed end users having similar attitude towards a particular topic or product. Community detection algorithm is used to identify the social atoms that are more densely interconnected relatively to the rest over the social media platform. Recently researchers focused on group-based algorithm and member-based algorithm for community detection over social media. This paper presents comprehensive overview of community detection technique based on recent research and subsequently explores graphical prospective of social media mining and social theory (Balance theory, status theory, correlation theory) over community detection. Along with that this paper presents a comparative analysis of three different state of art community detection algorithm available on I-Graph package on python i.e. walk trap, edge betweenness and fast greedy over six different social media data set. That yield intersecting facts about the capabilities and deficiency of community analysis methods.**

Keywords*—Community detection; social media; social media mining; homophily; influence; confounding; social theory; community detection algorithm

I. Introduction

The Emergence of Social networking Site (SNS) like Face-book, Twitter, LinkedIn, MySpace, etc. open a new perspective for sharing, discussing, organizing and finding the information, experiences, contacts and contents. A SNS can be modeled as a graph G= (V, E), where V is a set of nodes and E is a set of edges that represent the interaction between the nodes as shown in Fig. 1. The propensity of end user towards specific tastes, preferences, and inclination to get associated in a social network leads to the formation of friend and community recommendation system to enhance web life.

Community over SNS can be defined as a group of nodes that have more edges among themselves than those vertices outside the group. Social networks show strong community relationships and reveals useful information about structural and functional attributes. Recently Community detection over SNS can be beneficial for locating a common research area in collaboration networks for traffic management [1], finding a set of likeminded users for profile Investigation [2], [3], marketing [4], [5], recommendations system [6], [7], political belonging [8], and detecting spammers on social networks [9].

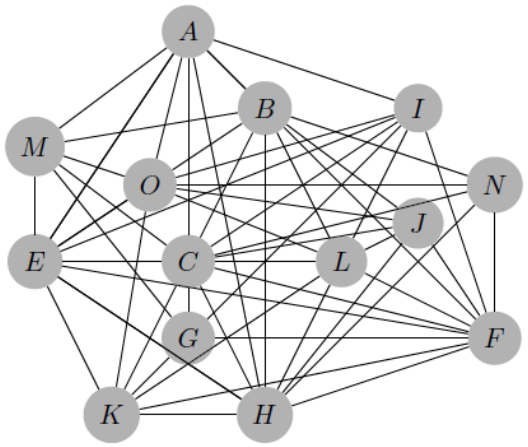

Fig 1. Social Media Network.

Aim of Community detection is to form group of homogenous nodes and figure out a strongly linked subgraphs from heterogeneous network. In strongly linked sub- graphs (Community structure) nodes have more internal links than external. Detecting communities in heterogeneous networks is same as, the graph partition problem in modern graph theory [10], [11], [12], as well as the graph clustering [13], [14] or dense sub graph discovery problem [15] in the graph mining area.

This paper summarized the influence of social theory for community detection over social media and presents a comparative analysis of recent community detection technique over six different social media data set. The rest of the paper is organized as follows: Section II presents overview of social media and their data inconsistency problem for community detection; Section III covers social media mining procedure for community detection and III(A)-III(C) explain social theory for deanonymized social relationship between social atom in social media data set. Section IV explains procedure of community detection over SNS; Section V covers recent research on community detection over social media. Section VI cover description of social media data set and evaluate the performance for benchmark algorithm for community detection over these data sets. Section VII include possible research gap in community detection over SNS and finally, Sect. VIII concludes the paper and outlines the founding.

II. Social Media

With the fast pace of the information age, the average access to the Internet only through computers is a thing of the past. Any individual associated with Internet diversely, is visualized to be substituted by other associated with Internet by hundreds of things. Similarly, there will be more things connected to the Internet than the people who are connected. Internet of thing (IoT) is one of the most emerging technologies on the Internet. Lot of interesting works has been done in the field of IT and its implementation [13], [11]. Another area drawing interest of lot of researchers is Social Networking sites (SNS). SNS facilitates end users to being connect and interact with each other without any geographical boundaries. SNS can be viewed graphically as world of social atoms (i.e., individuals), entities (e.g., content, sites, networks, etc.), and visuals among them.

Social Network provides a platform to extracting and mining multidimensional, multisource, and multisite data to identify individual behavior. Social media data encompasses user profile information and generated content. Besides degree, dimension and versatility, social media data having following inconsistent problem with rich of social ethics such as friendships and followers, etc.

- **Data Inconsistency:** The versatility of social media data that aggregate multidimensional, multisource, and multi- site data, lead statistical inconsistency in data set.
- **Data deficiency:** Due to the privacy preservation norms, SNS API release sanitized version of anonymized data. Where user identity and relationships are replaced by random attributes that lead to compute virtual user behavior.
- **Noise:** In social media there is not any mechanism to control irrelevance in user generated content, which lead noise in social media data set.
- **Evaluation Predicament:** For any supervised learning approach, ground truth is needed the pattern evaluating. Where training data can be used in learning and test data serves as ground truth for testing. Whereas in case of Social media data set, ground truth is often not available for mining process so deprived of trustworthy valuation, the legitimacy of the patterns is doubtful.
- **Missing Values:** Any individuals may avoid fill non-essential profile information on social media sites, such as their date of birth, location, Job profile, Alma mater detail, relationship detail and hobbies which lead inconsistency in behavior analysis.
- **Data Redundancy:** Data redundancy occurs over social media when multiple instances have exactly same feature values. Duplicate blog posts, carbon copy tweets, or fake profiles on social media with original information responsible for data redundancy.

The unpredictable degree, dimension and versatility of social media data need an interdisciplinary computational data analysis approach that encapsulate social theories (Balance theory, Status theory, and Social correlation) with data mining techniques as social media mining.

III. Social Media Mining

Social media mining (SMM), mine the information about social atoms, entities, and their interactions to extract meaningful behavioral patterns of social atoms from social media data set. SMM encapsulate interdisciplinary concepts, theories, fundamental principles, and data mining algorithms to develop computational algorithms for handle user generated content with social theories. For determining the consistency among social atoms, SMM applied Social Balance, Status, and Correlation theory over social media data set.

A. Balance Theory

Social balance theory evaluates relational structural consistency among social atoms. For instance, if two social atoms interact with positive sign edge then they are friends else if interact with negative sign edge then enemy. Social norms for social balance theory state that “Friend of Friend is Friend” and “Enemy of Friend is Enemy” and suggest the relationship among unknown social atoms over the Social media. For example consider the graph (V, E) having six vertex V1, V2, V3, V4, V5 and V6. Where (V1, V2), (V3, V4) and (V2, V6) are connected by positive sign edge, (V2, V3) and (V2, V5) are connected by negative sign edge as shown in Fig. 2(A). Then the social norm of balance theory reflects the negative relationship between (V1, V4) vertices and positive relationship between (V1, V6) vertices as shown in Fig. 2(B).

B. Status Theory

Social status theory evaluates relational reputational consistency among social atoms related to its neighbors. For instance, if any social atoms A having lower status then atoms B and subsequently same relationship is between B and C. Then status theory implies that status of A is lag behind C. In directed graphs, status of node depends upon sign and head of directed edge. Positive sign edge reflects higher status to head node whereas negative sign edge reflects lower status to head node with respect to tailed node. For example consider the graph shown in Fig. 3 (A), positive labeled edge shows head node V_2, V_6 and V_5 has higher status than its tailed node V_1 and V_2 respectively. Whereas Negative labeled edge show head node V_3 and V_4 has lower status then its tailed node V_2 and V_3 respectively. Whereas Social norm of status theory evaluate status of all the remaining pair of node as shown in Fig. 3(B).

C. Social Correlation

Social correlation theory is used to evaluate the individual behavior of social atoms with help of Influence, Homophily and Confounding social parameter. Influence connects individuals’ characteristics with social relation; homophily connect social relation with individuals’ characteristics whereas Confounding create a platform to connect similar characteristics individuals.

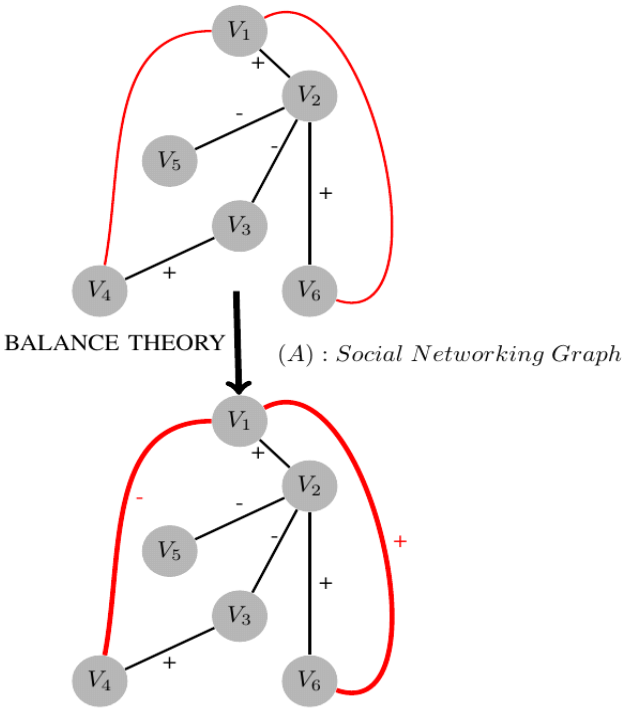

Fig 2. Social Balance Theory.

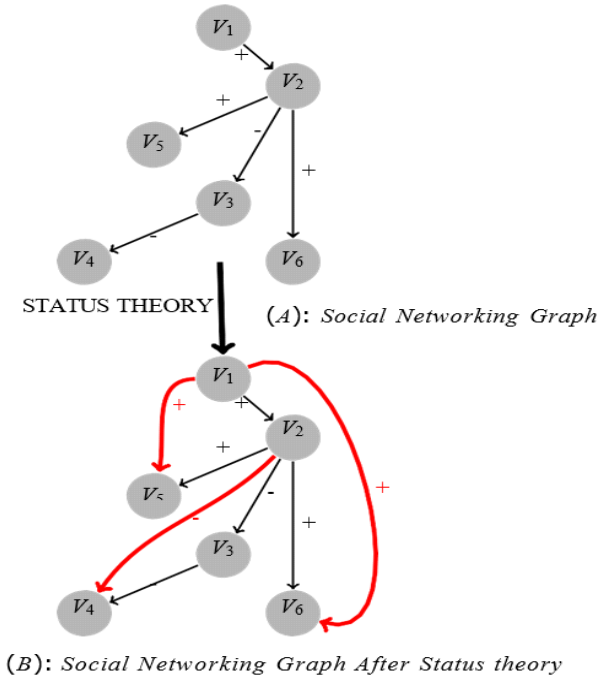

Fig 3. Social Status Theory.

For instance, consider the social graph shown in Fig. 4(A) where red color nodes are the follower of Republican political party and green color nodes are politically neutral. Due to influence correlation theory, post and status message of red color node get influence green color node to become follower of Republican political party as shown in Fig. 4(B). Whereas homophily, group the social atoms (nodes) behalf of their color notation as shown in Fig. 4(c) whereas Confounding state environments effect to make individuals similar. Two individuals living in the same city are more likely to become friends than two random individuals.

IV. Community Detection Over SNS

Social networking Site (SNS) can be represented as a graph G (P, R, W). Where P is set of peoples (vertices) belong to SNS, R is a set of links or relationship between two elements of P, and W: p × p → R is a function which assigns a weight to a couple (Pi,Pj) of vertices Pi and Pj, for instance if W: pi × pj → 1 then their exists an link between Pi and Pj. Whereas if W: pi × pj → 0 then there is no link between Pi and Pj. Social networking sites do not publish real Social network datasets. Before publishing user's data, social networking sites owners anonymized social networks data using conventional anonymized processes (like; k-anonymity [16], i-diversity [17], t-closeness [18]). Anonymized social networks data can be represented with the adjacency matrix AP*P and value of Aij determine the type of network. If Aij=Ajiie AP * P is symmetric matrix then SNS is undirected network.

In the real world, the community is a collection of people having similar social, political and spiritual view, who lives in a similar geographical area. Whereas in SNS, community are the collection of similar thinking social atoms without any geographical boundaries and having similar view on social, political, economic and global issue on social media platform. Aim of community detection is to find out group of vertices (sub graphs) having a high density of links within the group, and lower density of link outside of group. Structure of community can be represented as a set of N community in case of overlapping communities.

Community on SNS can be explicit or implicit. In explicit community, members are well-known about their membership and widely interact with each other. Whereas, whenever group of social atoms silently interact with each other within an unacknowledged group and obscure membership is refer to implicit community. For instance, alumni group of any educational institute over social media is refer to explicit group, where every alumni is known about its group prospective whereas any marketing agency interested to find the group of lady as implicit community having similar choices for certain beauty product for advertisement.

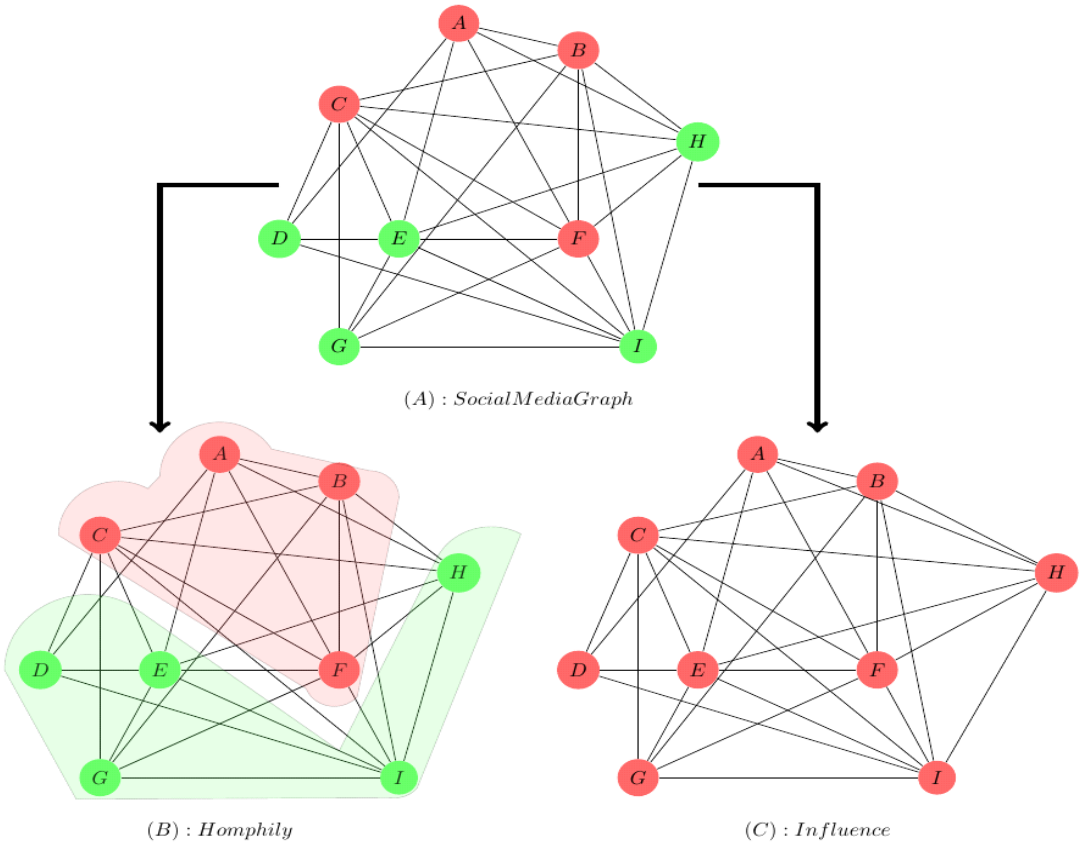

Fig 4. Social Correlation Theory.

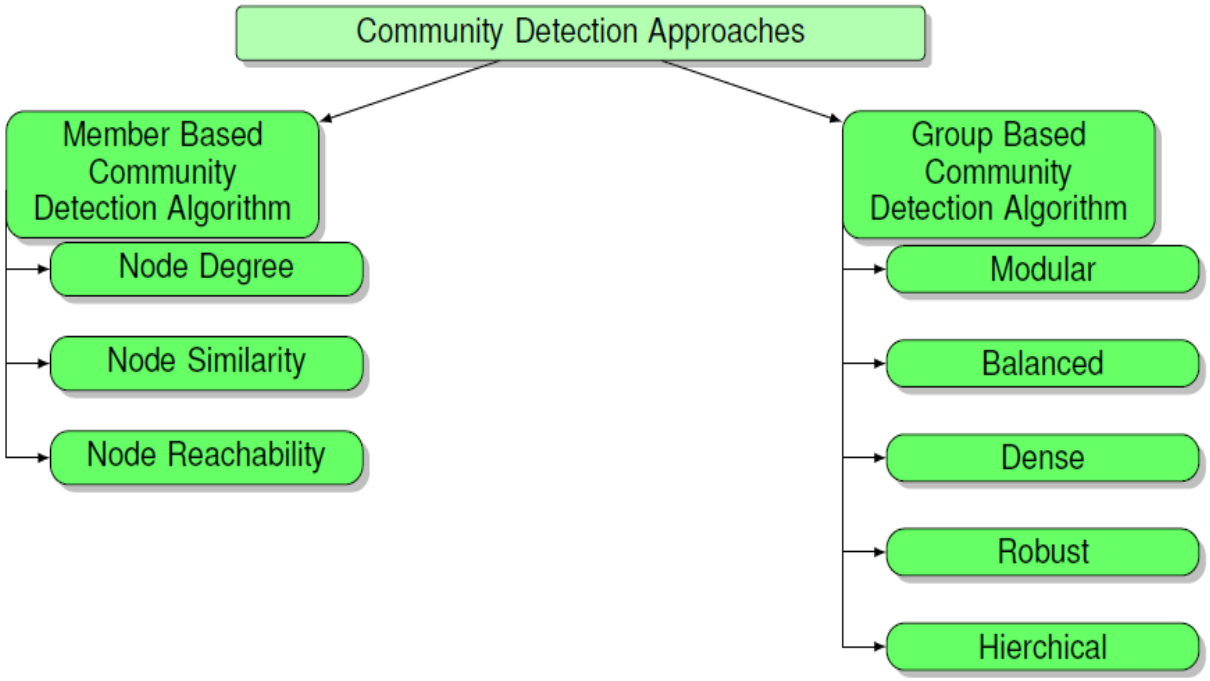

Fig 5. Community Detection Algorithm Hierarchy.

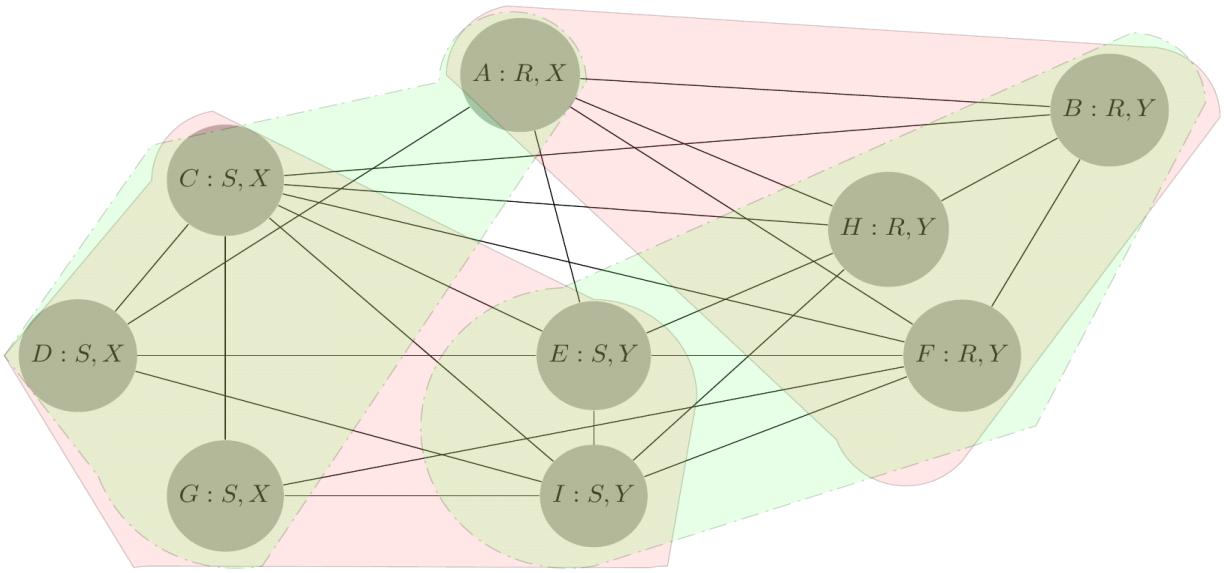

Fig 6. Community Over Social Media.

Recent research focuses to build efficient community detection algorithms to find implicit communities accurately. On the basis of community kernels, the community identification algorithms for social media comes with two different versions, namely member-based and group-based community identification ion algorithms. Member based community detection algorithm [19] is employed to extract the community around any specific social atoms' specification such as similarity, degree, and reach ability whereas graph-based algorithm is used to extract the community with certain group specification or norms such as modular, balanced, dense, robust, and hierarchical as shown in Fig. 5.

In member-based algorithm, if degree of node is used as a feature for community detection then it selects maximum clique over social media graph as community. Node degree-based algorithm suffer from NP hard problem i.e. not able to verify extracted clique as community contain every node of graph or not whereas in Node similarity-based community detection algorithm, similarity function such as Jaccard coefficient, sim function, sign and cosine function are used to form group of likelihood node as community. However, node reachability based community detection algorithm forms a group of nodes as community on behalf of member reachability factor i.e. two nodes belong to same community if there is a path available between these two nodes for communication.

In Group Based Community Detection algorithm use normalized and ratio cut partitioning algorithm to divide the graph into different community as balanced community detection scheme whereas, Robust community detection algorithm use k-vertex connected graph-based approach to find sub-graph as community that robust enough and not lose their node connectivity even after removing same edge and vertices. In modular community detection approach, modularity matrix is used to partitioned graph into k sub graph as community. In dense community detection approach, high dense clique are consider as community. Whereas Hierarchical group Based Community Detection algorithm is use to generates community hierarchies. Initially all node are consider to be in one community after that gradual aggregation and division split large community into desired sub-community.

For understanding graphical prospective of community detection algorithm, consider the example of two research group R (A, B, H, F) and S (C, D, E, G, I) mutually lives in two different city X (A, C, D, G) and Y (B, H, F, E, I) as shown in Fig. 6. Where researcher label with their name (A), research group (R) and city (X) as (A: R X). If foundation of community is characterized by specific social atoms such as 'A' with their geographical area specification then member based algorithm is used and shown by red color group. However if foundation of community is characterized by research group membership specification then graph based algorithm and shown by green color group in Fig. 6.

V. Related Work

Social networking has become an increasingly important application in recent years, because of its unique ability to enable social contact over the internet for geographically dispersed users. A social network can be represented as a graph, in which nodes represent users, and links represent the connections between users. An increased level of interest in the field of social networking has also resulted in a revival of graph mining algorithms. Therefore, a number of techniques have recently been designed for a wide variety of graph mining and management problems [11]. In recent years, some attempts tried to show that community structures are one of the significant characteristics in the most complex networks such as social networks due to numerous trends of human being to forming groups or communities. Due to the significant applications of community detection, several community detection approaches have been presented in literature which can be classified into six categories: spectral and clustering methods [20], [21], [15], [22], hierarchical algorithms [23], modularity-based methods [24], [25], evolutionary model-based methods [26], [27], local community detection methods, and feature- based assisted methods [11].

TABLE I. ARTICLE SUMMARY

R	Y	Task	M	Algorithm	Data Set	Merit	Future Scope
20	2015	Overlapped Community detection	G	Fuzzy C-Means	Zachary's Karate Club data	Improve Precision	Community Detection over multiple centers.
21	2015	Character co-appearances Community	M	Entropy centrality	Zachary's karate club, dolphin network	Minimized Iteration	Overlapping character co-appearances communities
15	2015	Overlapped community detection	M	Semantic link weight (SLW) based link-field-topic (LFT)	Qlsp , Krebs polbooks, Dolphins network	Significant Semantic modularity	Dynamic community-topic Relationship.
28	2015	Underlying community Detection	G	Pair counting method, Generalized linear preference	Facebook API, Twitter API	Multiple center community detection	Extract ground-truth for Underlying community Detection
29	2015	Parameter-free community detection	G	Page Rank , k-means	LFR networks, GN networks, Zachary's karate club	No need to initialized initial seeds and the number of communities	Optimal number of communities
30	2015	Overlapped community detection	G	Fuzzy Membership function	PCM model. Co-authorship network,	Dual center community detection	Optimal community center
31	2015	Disjoint community detection	M	Backbone degree algorithm	Zacharys Karate Club, DBLP collaboration network	Use biological And sociological model	Use biological and sociological model for detecting overlapping communities.
23	2015	Hierarchical structure of community members	M	Random Walk and Linear Regression	Karate Club, Dolphins network	Multi-resolution of community detection	Seed selection for Optimal number of communities
24	2015	Tightness greedy optimization for Community detection	G	Memetic algorithm (MA) based on genetic algorithm	Zachary's karate club, dolphin network, American College football, Books about US politics	Local structural information of networks to improve the diversity of the population	Overlapping dynamic community detection and cost minimization
25	2015	Biogeography based Optimized Community detection	M	Modularity and normalized mutual information	Synthetic datasets, Football dataset	Community detection over Dynamic network	Bio-geographical optimization over Large scale networks in real life
32	2017	Correlation analysis for community structure detection	M	Modularity function, Greedy and the fast unfolding search.	Karate Club and College Football	Average correlation degree get enhanced	Heuristic method for each different objective function.
33	2017	Evolutionary optimization for community detection	M	GA and fuzzy	Dolphin, Email, Football, Jazz, Karate, lesmis , polbooks , Sawmill, Strike, Words	Linear regression and quintile plots	Quality and convergence rate
40	2017	Join the method for overlapping and non-overlapping community detection	G	AGM, MMSB, IEDC	Football, Polbooks, Polblogs, caltech, Rice	NMI, F1 score and conductance measure enhance	Probabilistic method.
41	2017	Detect overlapping communities	M	Density based link clustering algorithm, DBLC algorithm, CPM algorithms	Karate club, dolphin, Books, football, Netscience, Email	Overlapping nodes	Communities in Multi-Mode Networks
42	2017	Detection of communities in topologically incomplete networks	G	Structured deep convolutional neural network (CNN)	Football, livejournal, youtube	Better robustness	Shared Community Structure in Multi-Dimensional Networks
43	2016	Solution of Imbalance problem in community detection	M	Normalized mutual information (NMI) Claculation	Zachary network, The college football network, The dolphin network, The Les Miserables network	Communities can be distinguished correctly	Heterogeneity helps reduce the noise

Along with that total sixteen articles (published in 2015 to 2017) presented in this survey are summarized in Table 1 that contains eight columns. The main task of the articles is illustrated in the third column. Column fourth illustrates method used i.e. either group or member-based analysis whereas G and M is used to represent Group based and Member based, respectively. Column fifth illustrates method and algorithm used for community detection in different application whereas sixth column describes the name of data set and its source that has been used for evaluating different methodology.

Zhou et al. [20] present probabilistic cluster prototype framework as Median variant of Evidential C-means (MECM) for detecting overlapped community based on belief function theory. Whereas Yu Xin et al. [15] present semantic overlapped community detection algorithm based on link-field-topic (LFT) model for structural transformation, and predict the emotional tendency.

Alexander G. Nikolaev [21] presents network entropy centrality-based community detection algorithm. W. Fan et al. [28] work over underline community detection to after analyzing social and profile interaction information and relationship.

Yafang Li [29] work over rank-based community structure grouping web pages through page pank centrality algorithm. Samira Malek et al. [30] work over fuzzy based duo centric overlapped community detection. YunfengXu et al. [31] work over biological structure to analysis strength and backbone degree of social network for member-based community detection.

Cai-hong mu et al. [24] present a graph based greedy optimized community detection approach that use memetic algorithm (ma) based on genetic algorithm to compute local structural information of networks to improve the diversity of the population but increase computational cost. Xu Zhou [25] proposed an optimized Biogeography based Community detection approach over dynamic network. Biogeography information extracted through Modularity and normalized mutual information of member.

LianDuan et al. [32] present a Correlation analysis for community structure detection by using Modularity function, Greedy and the fast unfolding search exercise. Anupam Biswas [33] present an Evolutionary algorithm based optimized com-munity detection algorithm. The methodology relies simply on linear regression and quintile plots to explain the dominance of one algorithm over another.

VI. Data Set

The data sets used in Community detection are important issues in these fields. The main sources of data are from the web club as show in Tables 1 and 2. Tables 1 and 2 contain detail about variety of data set that has been used in different application. The main sources of data are Social networking sites, which provided their API application like twitter API and face book API to fetch data from social media platform. These data are important to the business holders as they can take business decisions according to the analysis results of users' community about their products. This paper, evaluate the performance of three different state-of-the-art community detection algorithms available in the igraph package [34] such as Walk trap [35], Fast-Greedy [36], and Edge Betweenness [37] for undirected, unweighted graphs with non-overlapping communities, over six different data set shown in Table 2.

Table 2 that contain 3 columns. Network information of the data set (mention in first column) is illustrated in the second column. Where V , E , CC , AD and MD is used to represent number of Vertex , Edge , cluster coefficient , average degree and Maximum degree, respectively .Column Third illustrate modularity of basic stand-alone algorithm used for community detection in different application.

The six bench mark data set namely Word adjacencies, Zachary karate club [38] , Dolphin social network [39], Les Miserables, Books about US politics and American College football [37] is use to evaluate modularity of Walktrap, Fast-Greedy, and Edge Betweenness algorithm over community detection.

- **Word Adjacencies:** Word adjacencies data set is an undirected network data of common noun and adjective adjacencies of a novel "David Copperfield" by 19th century writer Charles Dickens. The dataset included 112 words (vertex), 58 adjectives and 54 nouns included with 425 edges. A vertex represents either a noun or an adjective. An edge connects two words that occur in adjacent positions. The network is not bipartite, i.e., there are edges connecting adjectives with adjectives, nouns with nouns and adjectives with nouns.
- **Zachary Karate Club:** Zachary karate club data was collected from the members of a university karate club by Wayne Zachary in 1977. Each node represents a member of the club, and each edge represents a tie between two members of the club. The network is un-directed. An often-discussed problem using this dataset is to find the two groups of people into which the karate club split after an argument between two teachers
- **Dolphin Social Network:** Dolphin social network [39] is a directed social network of bottlenose dolphins. The nodes are the bottlenose dolphins (genus Tursiops) of a bottlenose dolphin community living off Doubtful Sound, a fjord in New Zealand (spelled fiord in New Zealand). An edge indicates a frequent association. The dolphins were observed between 1994 and 2001.
- **Les Miserables:** Les Miserables is undirected network contains co-occurrences of characters in Victor Hugo's novel 'Les Miserables'. A node represents a character and an edge between two nodes shows that these two characters appeared in the same chapter of the book. The weight of each link indicates how often such a co-appearance occurred.
- **Books about US politics:** Books about US politics is a network of books about US politics published around the online bookseller Amazon.com. Edges between books represent frequent co purchasing of books by the same buyers.

- **American College Football:** Whereas American College football is a network of American football games between Division IA colleges during regular season fall 2000.

Performance evaluation of community detection Algorithm over social media data set is illustrated in Table 2. Modularity is network structural measurement that evaluates the strength of sub graph (groups, clusters or communities) in network for extracting community structure [44]. In a network, group of nodes having higher modularity are relatively dense each other and leads to the appearance of communities in a given network.

VII. Expected Research Avenue

- **Noise Handling:** Redundancy and complementary information of network element is act as Noise over network. A multi-mode network presents correlations between different kinds of objects for e.g., Users of similar interests are likely to have similar tags. Multi-dimensional networks have complementary information at different dimensions for e.g., some users seldom send email to each other, but might comment on each other's photos. Recently researcher take heterogeneity helps reduce the noise [43].
- **Communities in Multi-Mode Networks:** Multi-mode community detection, in particular, has great potential to provide insight into networks that are becoming increasingly complex with the evolution of social media and find out communities of each mode. Multi-mode networks clearly have a significant usefulness when it comes to representing complex social media data and other communication data. The new data demands of increasingly complex social and technical interactions online can be elegantly met by this new network representation that enables and even facilitates analysis. It stands to reason that fields outside of social network analysis can even benefit from using this representation in their techniques. Datasets for detecting communities in multi-mode communities become larger and larger, increasingly sophisticated algorithms are needed to draw meaningful conclusions from that data.
- **Communities in Multi-Dimensional Networks:** In Multi-dimensional networks, multiple connections may exist between a pair of nodes, reflecting various interactions (i.e., dimensions) between them. Multidimensionality in real networks may be expressed by either different types of connections (two persons may be connected because they are friends, colleagues, they play together in a team, and so on), or different quantitative values of one specific relation (co-authorship between two authors may occur in several different years, for example). The main challenge of Multidimensional Community Discovery is to detecting communities of actors in multidimensional networks and characterized the community found.
- **Shared Community Structure in Multi-Dimensional Networks:** Social media users interact at different social media sites. A latent community structure is shared in a multi-dimensional network and a group member sharing similar interests. The main goal is to find out the shared community structure by integrating the network information of different dimensions.

The modularity of community detection algorithm is depend upon network parameter i.e. number of Vertex, Edge, cluster coefficient , average degree and Maximum degree of network data set. Walk trap algorithm gain 0.3532216,0.6029143, 04888454, 0.5069724, 0.5215055 and 0.2162131 modularity over ZKC,ACF, DSN,BUP,LM and WA social network data set as shown in Table 2 and Fig. 7.

TABLE II. Modularity of Benchmark Algorithm Over Data Sets

Data Set	Network Information					Modularity		
	V	E	CC	AD	MD	Walktrap	Fast-Greedy	Edge Betweenness
Zachary's karate club	34	78	25.6	4.5882	17	0.3532216	0.3806706	0.4012985
American College football	115	615	5.73	10.71	13	0.6029143	0.5497407	0.599629
Dolphin social network	62	159	30.9	5.1290	12	0.4888454	0.4954907	0.5193821
Books about US politics	105	441	-	-	-	0.5069724	0.5019745	0.5168011
Les Miserables	77	254	49.9	6.5974	36	0.5214055	0.5005968	0.5380681
Word adjacencies	112	425	15.7	7.5893	49	0.2162131	0.2946962	0.08053702

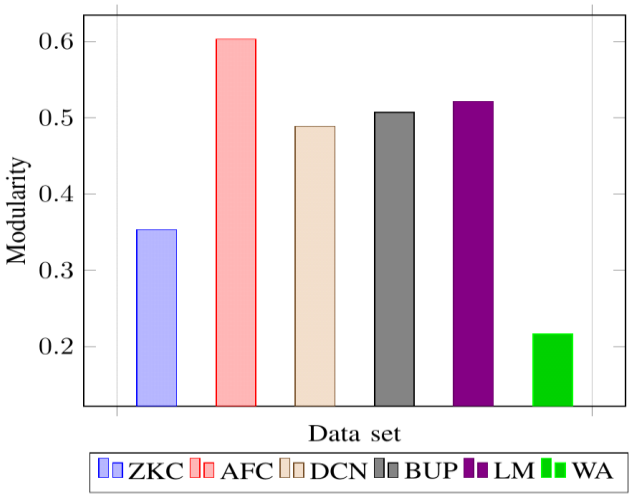

Fig 7. Community Detection with Walktrap Algorithm.

Modularity of Walk trap algorithm is increase with density of node in network i.e. depend upon average degree of network. Walk trap algorithm archive highest modularity over AFC data set, that having highest average degree with respect to other. But there is one exception with WA data set i.e. WA data set having second highest average degree but having lowest modularity. This exception is due to its higher maximum degree. Density of node is mutually depend upon average degree and maximum degree, if average degree is closer to maximum degree then node are highly dense in network.

Whereas in case of Fast Greedy and Edge Betweenness algorithm, modularity over ZKC, ACF, DSN, BUP, LM and WA data set is (0.3806706, 0.5497407, 0.4954907,0.5019745, 0.5005968, 0.294692) and (0.4012985, 0.599629, 0.5193821, 05168011, 0.5380681, 08053702), respectively. Both the algorithm show same pattern of modularity with respect to density as shown in Table 2 and Fig. 8 and 9, respectively.

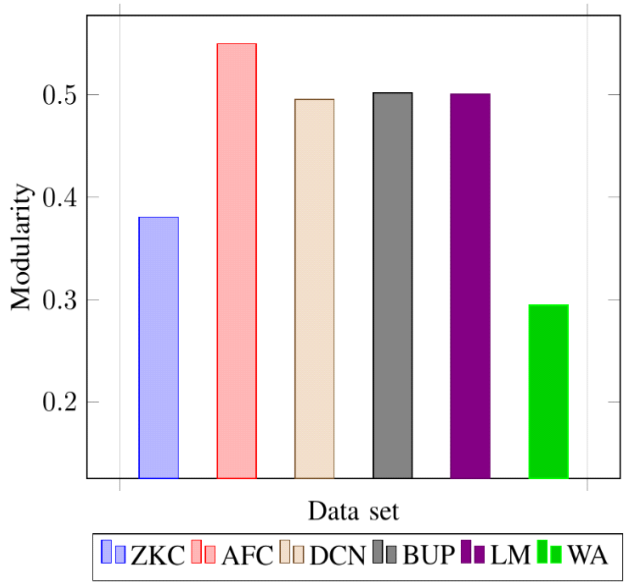

Fig 8. Community Detection with Fast-Greedy Algorithm.

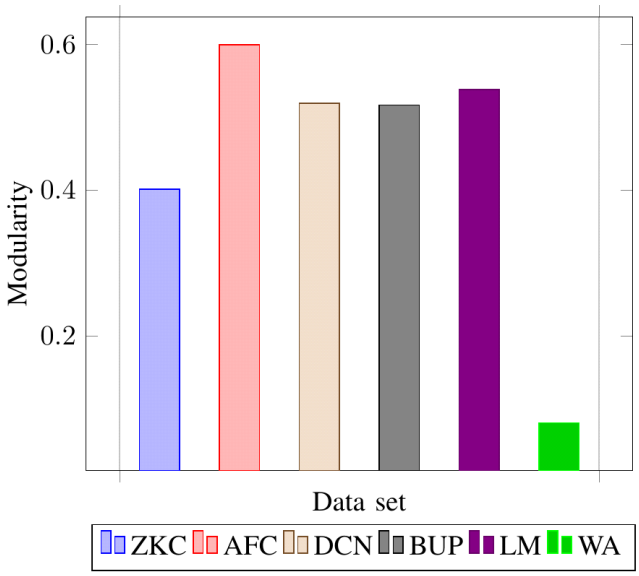

Fig 9. Community Detection with Edge Betweenness Algorithm.

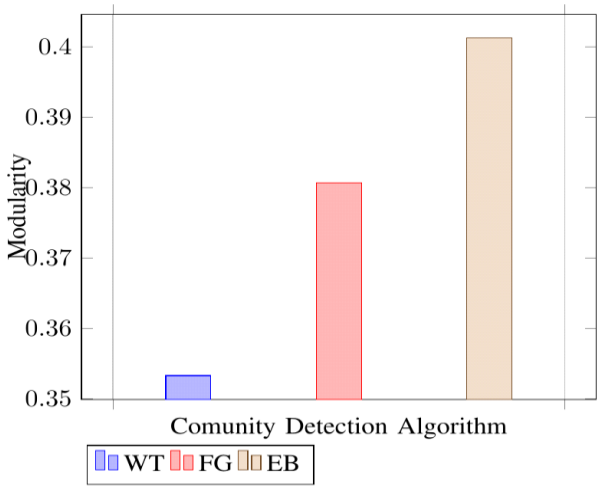

Fig 10. Community Detection over Zachary's karate club data set.

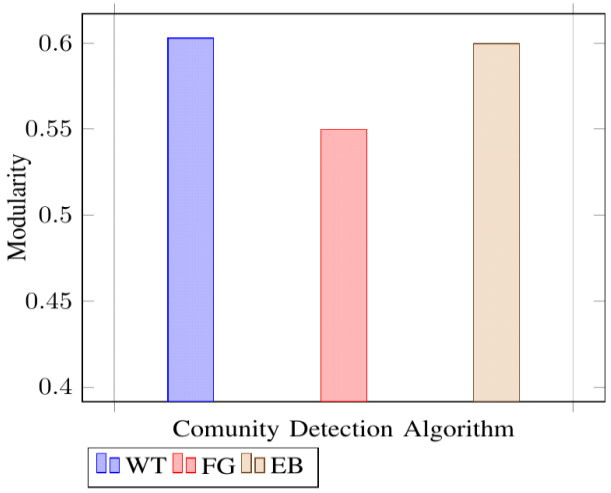

Fig 11. Community Detection over American College football data set.

On other hand with different prospective of analyzing the performance of community detection algorithm over different social media data set. It is observed that over ZKC data set, edge betweenness algorithm lead the performance by gaining 0.4012985 modularity as shown in Fig. 10 whereas walktrap and fast greedy gain 0.3532216 and 0.3806706 modularity, respectively. Over AFC data set, walktrap algorithm leads the performance by gaining 0.6029143 modularity as shown in Fig. 11 whereas fast greedy and edge betweenness gains 0.5497407 and 0.599629 modularity, respectively. Over DHN data set, edge betweenness algorithm leads the performance by gaining 0.5193821 modularity as shown in Fig. 12 whereas walktrap and fast greedy gain 0.4888454 and 0.4954907 modularity, respectively. Over BUP data set, edge betweenness algorithm leads the performance by gaining 0.5168011 modularity as shown in Fig. 13 whereas walktrap and fast greedy gain 0.5069724 and 0.5019745 modularity, respectively. Over LM data set, edge betweenness algorithm leads the performance by gaining 0.5380681 modularity as shown in Fig. 14 whereas walktrap and fast greedy gain 0.5214055 and 0.5005968 modularity, respectively. However, over WA data set, fast greedy algorithm lead the performance by gaining 0.2946962 modularity as shown in Fig. 15 whereas walktrap and edge betweenness gains 0.2162131 and 0.08053702 modularity, respectively.

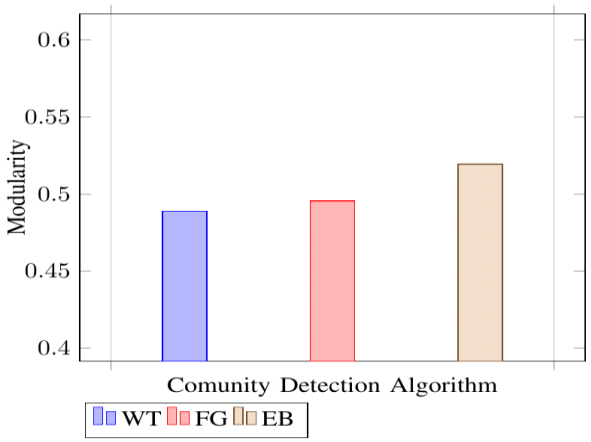

Fig 12. Community Detection over Dolphin social network data set.

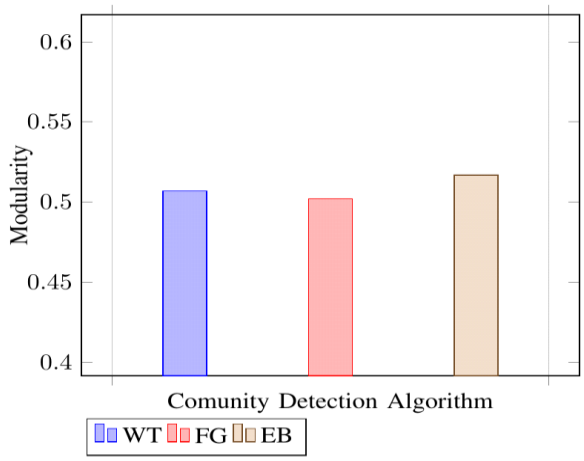

Fig 13. Community Detection over Books about US politics data set.

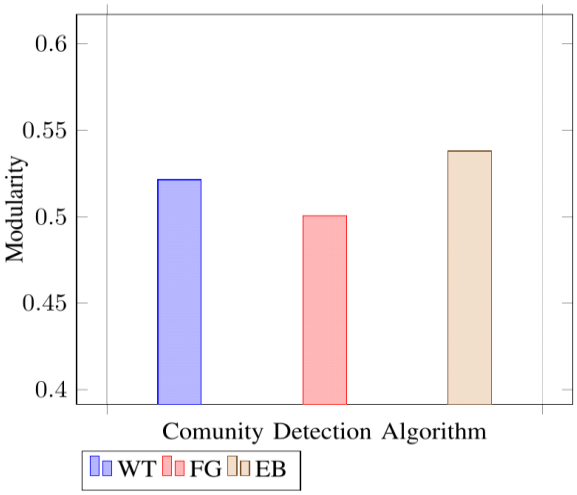

Fig 14. Community Detection over Les Miserables data set.

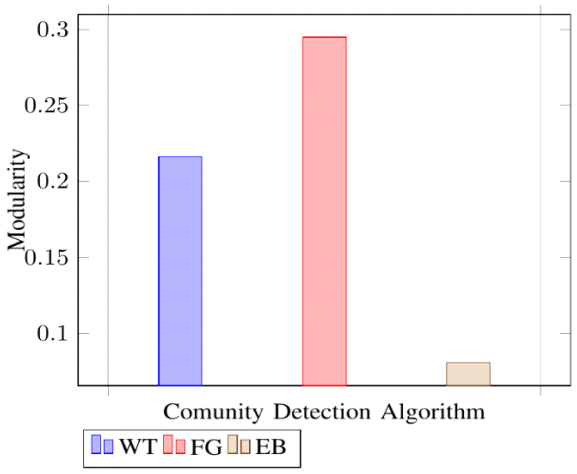

Fig 15. Community Detection over Word Adjacencies data set.

After evaluating the performance of community detection algorithm over different social media data set, it is observed that community detection algorithm gives its best performance over high dense network as AFC and LM data set.

VIII. Conclusion

Community detection is one of the emerging fields of the social media mining. Researcher has done lot of work in community detection. Major issues of community detection are scalability and quality of the community. Some of the algorithm scalable in large network and provides better results as compare to another algorithm. This paper compared the basic stand-alone algorithm such as Walktrap, Fast-Greedy and Edge Betweenness over six different data sets. As result it is proved that algorithms are scalable in the large network as per the evaluation parameter. The unique feature of this paper is to evaluate all the features of the algorithm on the large social network. After evaluating the performance of community detection algorithm over different social media data set, it is observed that community detection algorithm gives its best performance over high dense network as AFC and LM data set. This paper also discusses challenges like Communities in

Multi-Mode, Multi-Dimensional and share Networks and handling Noise over community detection. Along with that there is a problem of influence maximization in the social network that detects influence flow in the community with influence-user of the community. As it is known that most influential user increase the flow influence in the community with this one more issue of community detection is taken i.e. scalability in large network.

REFERENCES

[1] M. Sammarco, M. E. M. Campista, and M. D. de Amorim, "Scalable wireless traffic capture through community detection and trace similarity," *IEEE Transactions on Mobile Computing*, vol. 15, pp. 1757–1769, July 2016.

[2] R. R. Singh and D. S. Tomar, "Approaches for user profile investigation in orkut social network," *CoRR*, vol. abs/0912.1008, 2009.

[3] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of facebook networks," *CoRR*, vol. abs/1102.2166, 2011.

[4] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Towards detecting compromised accounts on social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 14, pp. 447–460, July 2017.

[5] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 7821–7826, June 2002.

[6] W. Fan, K. Yeung, and W. Fan, "Overlapping community structure detection in multi-online social networks," in *2015 18th International Conference on Intelligence in Next Generation Networks*, pp. 239–234, Feb 2015.

[7] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Phys. Rev. E*, vol. 68, p. 065103, Dec 2003.

[8] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 u.s. election: Divided they blog," in *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, (New York, NY, USA), pp. 36–43, ACM, 2005.

[9] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, (New York, NY, USA), pp. 1–9, ACM, 2010.

[10] J. Bonneau, J. Anderson, and G. Danezis, "Prying data out of a social network," in *2009 International Conference on Advances in Social Network Analysis and Mining*, pp. 249–254, July 2009.

[11] C. Pizzuti, "Evolutionary computation for community detection in networks: A review," *IEEE Transactions on Evolutionary Computation*, vol. 22, pp. 464–483, June 2018.

[12] S. Hour and L. Kan, "Structural and regular equivalence of community detection in social networks," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pp. 808–813, Aug 2014.

[13] C. Wang, W. Tang, B. Sun, J. Fang, and Y. Wang, "Review on community detection algorithms in social networks," in *2015 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pp. 551–555, Dec 2015.

[14] S. Bouhali and M. Ellouze, "Community detection in social network: Literature review and research perspectives," in *2015 IEEE International Conference on Service Operations And Logistics, And Informatics (SOLI)*, pp. 139–144, Nov 2015.

[15] X. Yu, J. Yang, and Z.-Q. Xie, "A semantic overlapping community detection algorithm based on field sampling," *Expert Systems with Applications*, vol. 42, no. 1, pp. 366 – 375, 2015.

[16] T.-K. Huang, M. S. Rahman, H. V. Madhyastha, M. Faloutsos, and B. Ribeiro, "An analysis of socware cascades in online social networks," in *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, (New York, NY, USA), pp. 619–630, ACM, 2013.

[17] N. Shrivastava, A. Majumder, and R. Rastogi, "Mining (social) network graphs to detect random link attacks," in *2008 IEEE 24th International Conference on Data Engineering*, pp. 486–495, April 2008.

[18] S. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, *Extraction and Analysis of Facebook Friendship Relations*, pp. 291–324. London: Springer London, 2012.

[19] R. Hosseini and R. Azmi, "Memory-based label propagation algorithm for community detection in social networks," in *2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP)*, pp. 256–260, March 2015.

[20] K. Zhou, A. Martin, Q. Pan, and Z. ga Liu, "Median evidential c-means algorithm and its application to community detection," *Knowledge-Based Systems*, vol. 74, pp. 69 – 88, 2015.

[21] A. G. Nikolaev, R. Razib, and A. Kucheriya, "On efficient use of entropy centrality for social network analysis and community detection," *Social Networks*, vol. 40, pp. 154 – 162, 2015.

[22] A. Croitoru, N. Wayant, A. Crooks, J. Radzikowski, and A. Stefanidis, "Linking cyber and physical spaces through community detection and clustering in social media feeds," *Computers, Environment and Urban Systems*, vol. 53, pp. 47 – 64, 2015. Special Issue on Volunteered Geographic Information.

[23] F. Chen and K. Li, "Detecting hierarchical structure of community members in social networks," *Knowledge-Based Systems*, vol. 87, pp. 3 15, 2015. Computational Intelligence Applications for Data Science.

[24] C.-H. Mu, J. Xie, Y. Liu, F. Chen, Y. Liu, and L.-C. Jiao, "Memetic algorithm with simulated annealing strategy and tightness greedy optimization for community detection in networks," *Applied Soft Computing*, vol. 34, pp. 485 – 501, 2015.

[25] X. Zhou, Y. Liu, B. Li, and G. Sun, "Multiobjective biogeography based optimization algorithm with decomposition for community detection in dynamic networks," *Physica A: Statistical Mechanics and its Applications*, vol. 436, pp. 430 – 442, 2015.

[26] P. M. Zadeh and Z. Kobti, "A multi-population cultural algorithm for community detection in social networks," *Procedia Computer Science*, vol. 52, pp. 342 – 349, 2015. The 6th International Conference on Ambient Systems, Networks and Technologies (ANT-2015), the 5th International Conference on Sustainable Energy Information Technology (SEIT-2015).

[27] X. Niu, W. Si, and C. Q. Wu, "A label-based evolutionary computing approach to dynamic community detection," *Computer Communications*, vol. 108, pp. 110 – 122, 2017

[28] W. Fan and K. Yeung, "Similarity between community structures of different online social networks and its impact on underlying community detection," *Communications in Nonlinear Science and Numerical Simulation*, vol. 20, no. 3, pp. 1015 – 1025, 2015.

[29] Y. Li, C. Jia, and J. Yu, "A parameter-free community detection method based on centrality and dispersion of nodes in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 438, pp. 321-334, 2015.

[30] S. M. M. Golsefid, M. H. F. Zarandi, and S. Bastani, "Fuzzy duocentric community detection model in social networks," *Social Networks*, vol. 43, pp. 177 – 189, 2015.

[31] Y. Xu, H. Xu, and D. Zhang, "A novel disjoint community detection algorithm for social networks based on backbone degree and expansion," *Expert Systems with Applications*, vol. 42, no. 21, pp. 8349 – 8360, 2015.

[32] L. Duan, Y. Liu, W. N. Street, and H. Lu, "Utilizing advances in correlation analysis for community structure detection," *Expert Systems with Applications*, vol. 84, pp. 74 – 91, 2017.

[33] A. Biswas and B. Biswas, "Analyzing evolutionary optimization and community detection algorithms using regression line dominance," *Information Sciences*, vol. 396, pp. 185 – 201, 2017.

[34] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 11 2005.

[35] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Computer and Information Sciences - ISCIS 2005* (p. Yolum, T. Güngör, F. Gürgen, and C. Özturan, eds.), (Berlin, Heidelberg), pp. 284–293, Springer Berlin Heidelberg, 2005.

[36] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, p. 066111, Dec 2004.

[37] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 7821–7826, 2002.

[38] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[39] D. Lusseau, "The emergent properties of a dolphin social network," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. suppl 2, pp. S186–S188, 2003.

[40] M. Hajiabadi, H. Zare, and H. Bobarshad, "Iedc: An integrated approach for overlapping and non-overlapping community detection," *Knowledge-Based Systems*, vol. 123, pp. 188 – 199, 2017.

[41] X. Zhou, Y. Liu, J. Wang, and C. Li, "A density based link clustering algorithm for overlapping community detection in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 486, pp. 65 – 78, 2017.

[42] X. Xin, C. Wang, X. Ying, and B. Wang, "Deep community detection in topologically incomplete networks," *Physica A: Statistical Mechanics and its Applications*, vol. 469, pp. 342 – 352, 2017.

[43] P. G. Sun, "Imbalance problem in community detection," *Physica A: Statistical Mechanics and its Applications*, vol. 457, pp. 364 – 376, 2016.

[44] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.