RESEARCH ARTICLE OPEN ACCESS

A Comparison Study of Different Community Detection Approaches and Its Potential Applications for Online Networks

Mohamed Farrag*, Laila El Fangary**, Mona Nasr***, Chaimaa Salama****
**(Department of Business Information systems, Canadian International College, Egypt*
Email: Mohamed_farrag@cic-cairo.com)
*** (Department of Information Systems, Faculty of Computers and Information, Helwan University, Egypt*
Email: Lailaelfangary@gmail.com)
**** (Department of Information Systems, Faculty of Computers and Information, Helwan University, Egypt*
Email: M.nasr@helwan.edu.eg)
***** (Department of Information Systems, Faculty of Computers and Information, Helwan University, Egypt*
Email: Chaimaa_salama@yahoo.com)

ABSTRACT
The incredible rising of online networks show that these networks are complex and involving massive data. Giving a very strong interest to set of techniques developed for mining these networks. One of the fundamental applications for it is the community detection. It helps to understand and model the network structure which has been a fundamental problem in several fields. Community detection is one of the fundamental applications that provide a solution discipline where systems are often represented as graphs. Community detection is of great importance in sociology, biology and computer science. This problem is very hard and not yet satisfactorily solved, despite the huge effort of a large interdisciplinary community of scientists working on it over the past few years. It helps to understand and model the network structure, can be useful in various applications such as rating prediction, link and Top-N recommendation, trend analysis and also be the main provider engine for recommender systems. Several community detection approaches have been proposed and applied to many domains in the literature. This paper presents a comparison study of two different existing community detection approaches and also discusses some of its vital applications for online networks.

Keywords - Clique problem, Community detection, Graph mining, Graph theory, Online networks

Date of Submission: 02-10-2017 Date of acceptance: 27-10-2017

I. INTRODUCTION

More and more people today are using online networks. These networks contain massive data. This massive data makes the graphs representing these networks are getting terribly complicated and very complex, millions of nodes are present across many various scientific domains as information retrieval, marketing, bioinformatics, disease classification , pattern recognition and analysis of financial networks. Giving a very strong interest to a set of techniques developed for mining these graphs. Detecting communities in networks is one of the fundamental applications that provide a solution for online networks, disciplines where systems are often represented as graphs. Communities can be considered a summary of the whole network, thus making the network easy to comprehend. The discovery of these communities can increasingly being leveraged as a powerful, low-cost tool for many various scientific domains as information retrieval, marketing plans, bioinformatics, disease classification , pattern recognition and analysis of financial networks. It also help in behavior modeling and prediction, collaborative filtering. It can be useful in various applications such as rating prediction, top-N recommendation, link recommendation and trend analysis and also be the main provider engine for recommender systems. It can also drive business objectives and functions such as enhanced customer interaction, provide a mechanism for executives to assess consumer opinion and use this information to improve products, customer service and perception. Also help in advertising, direct marketing. This paper first presents a fundamental concepts related to community detection. It also presents a comparison study of two different existing community detection approaches. The first approach concerns the relation between community members. While the second approach concerns the overlapping community detection methods divided in two categories. The first category is the clique based methods, while the second category is the non-clique based methods.

Furthermore, this paper also discusses some of essential applications for online networks in different business domain where systems are often represented as graphs. This paper is organized as follows: section two discusses the related concepts of online networks and graph structure. Section three demonstrates the community detection approaches and section four explains the essential application of online networks while conclusion and future work are in the last section.

II. RELATED CONCEPTS

Fundamental concepts related to overlapping community detection are described which have been used throughout different overlapping community detection approaches.

2.1. Online Networks

Some real life examples of online networks that can be analyzed and mining using SNA and graph mining , disciplines where systems are often represented as graphs. That leads to enhance business returns and support different business goals such as marketing strategies, generate leads and reduce support costs, it may help to grow the awareness of business.

Online Communities Built on Public Social Networks, Public social networks, like Facebook, Twitter and LinkedIn, are great for marketing businesses and organizations. Both B2B and B2C companies use them as marketing channels to "become one of us" and reach their markets where they already are. Also like Facebook community's friendship graph or network.

Product co-purchasing networks, like product co-purchasing networks as Amazon. Here the nodes represent products or from other hand represent users, which are really individuals. There is an edge between two nodes if a rating, votes or assign as helpful has been placed between those nodes.

Public Online Communities on a Company-Owned Domain, Features live on a company's domain, rather than Facebook.com or LinkedIn.com. The major benefits to this type of online community are ownership of the data and rules and data analytics of organizations can collect more useful data during the registration process that can be piped directly into a company's CRM system also all of the contents and discussions in the pubic online community give SEO (search engine optimization) power to companies since it all resides on their domain.

Collaboration Networks, Nodes represent individuals who have published research papers. An alternative view of the same data is as a graph in which the nodes are papers. Another sample as Wikipedia where we can look at the people who edit Wikipedia articles and the articles that they edit.

Business Blogs, Businesses of any size, ranging from individuals and consultants to large companies, can use publicly viewable blogs to create an open community of prospects, customers, fans, and partners.

Email Networks, The nodes represent email addresses, which are again individuals. An edge represents the fact that there was at least one email in at least one direction between the two addresses. Another approach is to label edges as weak or strong. Strong edges represent communication in both directions, while weak edges indicate that the communication was in one direction only.

Telephone Networks, the nodes represent phone numbers, which are really individuals. There is an edge between two nodes if a call has been placed between those phones. The edges could be weighted by the number of calls made between these phones during the period.

Other environments for online communities, Information networks (documents, web graphs, patents), infrastructure networks (roads, planes, water pipes, power grids), biological networks (genes, proteins, food-webs of animals eating each other), as well as other types, Also Internet traffic routing Nodes are routers while edges are the connections, Hash tag diffusion, Transportation networks, Organization People interaction in organization communities network.

2.2. Graph Structure

Some main aspects concerning the nature and the structure of on-line networks that these network modeled by a graph that are the foremost usually used abstract data structures within the field of computer science, they enable a more complicated and wide-ranging presentation of data compared to link tables and tree structures. [1], [16], [17] A network is usually presented as a graph $G(V, E)$, where V is set of n nodes and E is set of m edges. Graph G consisting of n number of nodes denoting n individuals or the participants in the network. The connection between node i and node j is represented by the edge e_{ij} of the graph. The graph are often represented by an adjacency matrix A in which A_{ij} = 1 in case there is an edge between i and j else A_{ij} = 0 [1], [3], [5].

2.3. Cliques

Another aspect is Cliques, is defined as "a set of vertices in which every pair of vertices is connected by an edge".[5] Clique is a complete sub graph of G or in different words "is a maximum complete sub graph in which all nodes are adjacent to each other" , in a clique of size k, each node maintains degree≥ $k-1$. Normally use cliques as a core or a seed to seek out larger communities. [5], [6], [12] another formal definition "A clique is a fully connected sub graph a set of nodes all of which are connected to each other. " [7], [15] K-cliques are often outlined as complete graph with k vertices. [12] K-cliques are main structures in complicated networks, and a good way to seek out community structure [7], [12]. A clique sample is shown in Fig. 1.

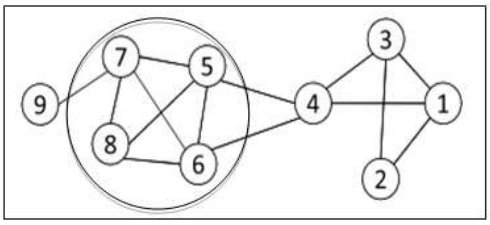


Fig. 1. A clique sample.

2.4. Maximal Clique

It is defined as "a clique that's contained in no larger clique". [7], [12], [14]. Every maximal clique is a clique, by definition, however the other doesn't hold. Therefore there are always more cliques than maximal cliques. In different words "a clique is said to be maximal if it not contained in any other clique." [7]

2.5. Adjacent K-Cliques

We are able to define adjacent K-cliques by two k-cliques that share $k-1$ nodes. [7], [12] K-clique community (cluster or component), outlined as "a union of all k-cliques that may be reached from each other through a series of adjacent k-cliques." or "It is the union of all k-cliques that are k-Clique-connected to a particular k-clique." [8], [12]

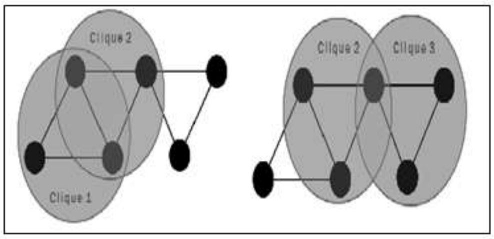


Fig. 2. Adjacent K-cliques share K-1 nodes.

2.6. Community Detection

Real world complicated systems may be represented within the form of networks. To comprehend the in-depth structure of those systems, it's necessary to review and analyze the networks. A trivial property of those networks is community structure obtained by splitting the network into many parts, within which connection between nodes are more dense than the remainder of the network. The sets of this sort of grouping are commonly referred as communities, however additionally called clusters, cohesive groups, or modules as there is no globally accepted unique definition. One among the restrictions of graph partitioning methods is that they typically need the user to specify the number of partitions, which cannot be identified before. One solution proposed to this problem is to use goodness metric as modularity to evaluate the partition of the graph at every step. However, this is often computationally expensive and might be infeasible for massive graphs. however just in case of community detection, it's not known that how many communities are present in the network and it is not at all obligatory for them to be of same size. The community detection approach assumes that almost all of real world networks, divide naturally into groups of nodes (community) with dense connections internally and sparser connections between groups, and therefore the experimenter's job is only to find these already formed groups. The number of partitions and size of them are settled by the network itself and not set by the experimenter. So community detection is "the technique which aims to discover natural divisions of networks into groups based on strength of connection between vertices." [1], [13] No formal definition of community is universally accepted, communities will have numerous properties, and community detection has been approached from many alternative views. Community detection is one among the foremost wide researched issues. Straightforward definition "A community is a densely connected group of nodes that is sparsely connected to the rest of the network" [18], Generally spoken community as "a module or cluster is typically thought of as a group of nodes with more and/or better interactions amongst its members than between its members and the remainder of the network". [16]

2.7. Disjoint Vs Overlapping Communities

Primarily, community may be divided into two types; disjoint communities and overlapping communities. In disjoint communities nodes can be part of only a single community. A non-overlapping community structure or disjoint community structure may be outlined as "set of communities such that all vertices are included in exactly one community." [2] However in overlapping communities partitions aren't essentially disjoint. There might be nodes that belong to more than one community [4], [18]. Usually in any on-line network a node may be part of more than one different group or community, thus for on-line networks, overlapping community detection technique ought to be thought of disjoint community detection technique. A number of community detection algorithms and methods have been proposed and deployed for the identification of communities in literature. There have also been modifications and revisions to many methods and algorithms already proposed.

III. COMMUNITY DETECTION APPROACHES

A number of community-detection algorithms and methods have been proposed and deployed for the identification of communities in literature. There have also been modifications and revisions to many methods and algorithms already proposed. We can divide the previous work in literature in two approaches

3.1. First Approach (Based On The Relation Between Community Members)

This approach divided into four categories. First category is node centric community detection, Second category is group centric community detection (Density-Based Groups), and third category is network centric community detection, while the fourth category is hierarchy centric community detection.

3.1.1. Node centric community detection

Wherever nodes satisfy different properties as complete mutuality which implies cliques; another property is reachability of members as k-clique [13].

3.1.2. Group centric community detection (Density-Based Groups)

It needs the total group to satisfy an explicit condition for instance the group density greater than or equal a given threshold and take away nodes with degree under the typical average degree.

3.1.3. Network centric community detection

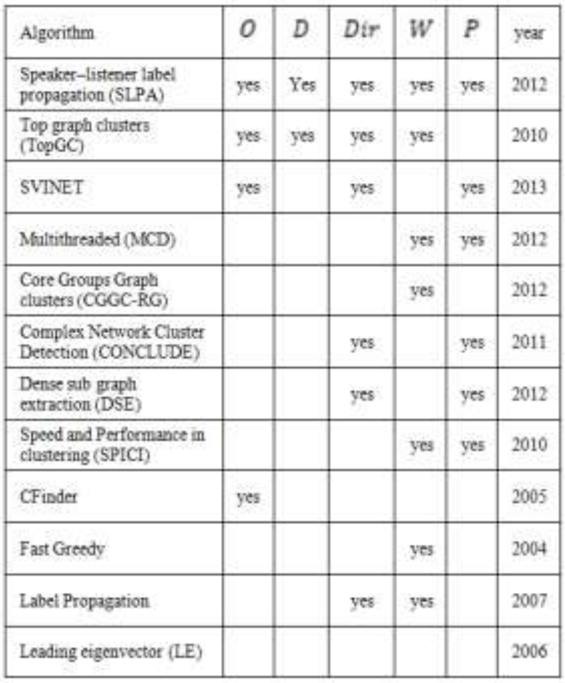
It takes into account connections within a network globally. Its goal to partition nodes of a network into disjoint sets, five different perspectives utilized in network-centric community detection. **First** is the clustering supported vertex similarity using Jaccard similarity and Cosine similarity. **Second** is Latent space models supported k-means clustering. **Third** is the block model approximation based on exchangeable graph models. **Fourth** is spectral clustering are using minimum cut problem that the number of edges between the two sets is reduced. **The fifth** is modularity maximization by measures the strength of a community partition by taking into consideration the degree distribution.

3.1.4. Hierarchy centric community detection

Aims to build a hierarchical structure of communities supported network topology to permit the analysis of a network at different resolutions two representative methods. **First** divisive hierarchical clustering (top-down) and **the second** is agglomerative hierarchical clustering (bottom-up) [3], [13]. The strength of a tie is measured by edge betweenness that is the number of shortest paths that pass along with the edge.

A summary for the most algorithms have been used in this approach for community detection are shown in Table I indicate if the algorithm detect overlapping or disjoint communities. Appropriate for directed and weighted graph.

TABLE I. Summary of most algorithms used in community detection, indicates that the algorithm identifies O: overlapping communities or D: disjoint communities, can be run on Dir: directed or W: weighted graphs and if it requires input p: parameter. [18]



Algorithm	O	D	Dir	W	P	year
Speaker–listener label propagation (SLPA)	yes	Yes	yes	yes	yes	2012
Top graph clusters (TopGC)	yes	yes	yes	yes		2010
SVINET	yes		yes		yes	2013
Multithreaded (MCD)				yes	yes	2012
Core Groups Graph clusters (CGGC-RG)				yes		2012
Complex Network Cluster Detection (CONCLUDE)			yes		yes	2011
Dense sub graph extraction (DSE)			yes		yes	2012
Speed and Performance in clustering (SPICI)				yes	yes	2010
CFinder	yes					2005
Fast Greedy				yes		2004
Label Propagation			yes	yes		2007
Leading eigenvector (LE)						2006

3.2. Second Approach (Overlapping community detection methods)

Common clustering methods lead to a partitioning in which a node can belong to a single community. For example, the node may be a member of a specific community but not, at the same time, be found as a member of another one. While every node in overlapping community can belong to multiple communities. Fig. 3 shows a sample of overlapping communities.

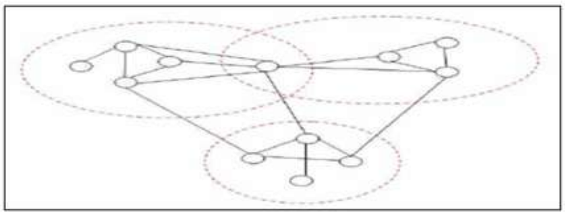


Fig. 3. Overlapping communities

The second approach divide the methods for detecting overlapping communities in two categories. First category is clique based methods and the second is non-clique based methods.

3.2.1. Clique based methods

In clique based methods for overlapping community detection a community may be consider as “a union of smaller, complete (fully connected) sub graphs that share nodes”.[1] A k-clique is a fully connected sub graph consisting of k nodes[12]. Also we can consider “A k-clique community can be defined as a union of all k-cliques that can be reached from each other through a series of adjacent k-cliques”. [1]

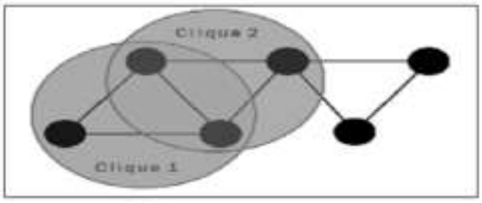


Fig. 4. Two adjacent cliques 1 and 2 perform a community

A summary of clique based methods for overlapping community detection are listed in Table II Show the algorithm, the approach have been used and the parameters for this approach.

Table II. Summary of clique based methods for overlapping community detection [1]

Author (Algorithm)	Approach	Parameters
Palla et al. (CPM) ,	Clique percolation method Nodes	threshold weight
Lancichinetti et al.	Fitness function	Fitness function
Du et al. (ComTector)	Kernels-based clustering	Set of all kernels
Shen at al (EAGLE)	Agglomerative hierarchical clustering	Similarity between two communities
Evans et al.	Line graph, clique graph	Links, partition
Lee et al. (GCE)	Cliques-based expansion	Fitness function
Gregory et al. (CONGA, CONGO, Peacock algorithm)	Split betweenness vertex	Short paths ratio of max. Edge betweenness and max. split betweenness

One of the most widely used techniques to discover overlapping communities is the clique percolation method (CPM). [9], [18] CPM is an efficient algorithm in discovering overlapping communities; it has a wide range of application in social networks and biological networks. The basic idea of this method is that the idea of a k-clique community that was outlined as the union of all k-cliques (complete sub graphs of size k) that can be reached from each other through a series of adjacent k-cliques (where adjacency means sharing $k-1$ vertices). The k-clique community may be thought of a usual module (community, cluster or complex) because of its dense internal links and sparse external linkage with other part of the whole network. Then build the overlap matrix of those k-cliques. Finally, a number of k-clique communities are detected by analysis the overlap matrix. [9], [10] CPM algorithm first detects all complete sub graphs of the network that are not parts of greater complete sub graphs. These maximal complete sub graphs are known as cliques, after the cliques are settled, the clique-clique overlap matrix is ready. In this symmetric matrix every row (and column) represents a clique and the matrix elements are equal to the number of shared vertices between the corresponding two cliques, and the diagonal entries are equal to the size of the clique. The k-clique-communities for a given value of k are equal to such connected clique components in which the neighboring cliques are linked to each other by at least $k-1$ shared nodes. These components can be found by removing each off-diagonal entry. [1], [8],

[12] a straightforward illustration for the extraction of the k-clique communities at $k = 4$ using the clique clique overlap matrix. We tend to use a sample network consists of eighteen nodes and thirty six edges as shown in Table III.

TABLE III. Sample network edge list

Network Edges					
N1,N2	N2,N10	N3,N9	N6,N8	N8,N9	N12,N13
N1,N3	N2,N18	N3,N10	N6,N9	N8,N10	N12,N14
N1,N4	N3,N4	N4,N5	N6,N10	N9,N10	N12,N15
N2,N3	N3,N6	N4,N6	N7,N8	N11,N12	N13,N14
N2,N4	N3,N7	N5,N6	N17,N11	N11,N13	N14,N15
N2,N9	N3,N8	N6,N7	N17,N12	N11,N14	N15,N16

- **Uncover the maximal cliques**

CPM finds all cliques using brute-force algorithm starting by set **A** initially contains vertex v, Set **B** contains neighbors of v. then transfer one vertex w from B to A and remove vertices that are not neighbors of w from B. Then repeat until a reaches desired size if fail, step back and try other possibilities. [8], [12] The seven maximal cliques detected by CPM for the sample network are *{N3, N6, N8, N9, N10}*, *{N3, N6, N7, N8}*, *{N2, N3, N9, N10}*, *{N3, N4, N6}*, *{N11, N12, N17}*, *{N11, N12, N13, N14}*, *{N12, N14, N15}.*

- **Maximal Cliques to k-Clique Communities**

A straightforward illustration Table IV

shows the steps of extraction of K-cliques communities at $K = 4$ using the clique overlap matrix. [12]

TABLE IV. CPM steps to extract communities at $K = 4$ [8], [12], [28]

M	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**M**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	5							**1**	5	3	3	2	0	0	0
2		4						**2**	3	4	1	2	0	0	0
3			4					**3**	3	1	4	1	0	0	0
4				3				**4**	2	2	1	3	0	0	0
5					3			**5**	0	0	0	0	3	2	1
6						4		**6**	0	0	0	0	2	4	1
7							3	**7**	0	0	0	0	1	1	3
Step1: count nodes for each maximal cliques								**Step2:** Calculate the intersection nodes between each two maximal cliques							
M	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**M**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	5	3	3	2	0	0	0	**1**	5	3	3	0	0	0	0
2	3	4	1	2	0	0	0	**2**	3	4	0	0	0	0	0
3	3	1	4	1	0	0	0	**3**	3	0	4	0	0	0	0
4	2	2	1	0	0	0	0	**4**	0	0	0	0	0	0	0
5	0	0	0	0	0	2	1	**5**	0	0	0	0	0	0	0
6	0	0	0	0	2	4	1	**6**	0	0	0	0	0	4	0
7	0	0	0	0	1	1	0	**7**	0	0	0	0	0	0	0
Step3: K=4 Change to zero any maximal clique less than 4								**Step4: K=4** Change to zero any intersection between maximal clique less than K-1 =3							
M	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Results**: (M1,M1)-(M1,M2)-(M1,M3)-(M6,M6) Community1:(M1,M2,M3) Community2:(M6)							
1	1	1	1	0	0	0	0								
2	1	1	0	0	0	0	0								
3	1	0	1	0	0	0	0								
4	0	0	0	0	0	0	0								
5	0	0	0	0	0	0	0								
6	0	0	0	0	0	1	0								
7	0	0	0	0	0	0	0								
Step5: K=4 Change all non-zero to 1															

The final K-cliques communities at $K = 4$ using CPM. Consists of two communities. C_1 =*{N2, N3, N6, N7, N8, N9, N10}* and C_2 =*{N11, N12, N13, N14}.*

Another techniques are employed in literature as a clique based overlapping communities like the algorithm projected by Lancichinetti et al.[19] It performs a local exploration in order to find the community for each of the nodes. During this method, the nodes could also be revisited any number of times. The target was to detect local maximal based on a fitness function. Also CFinder software system was developed supporting CPM for overlapping community detection. Then Du et al. [20] proposed Comtector to detect the overlapping communities using maximal cliques. At first, all maximal cliques within the network that form the kernels of a possible community are detect. Then, the agglomerative procedure is iteratively used to add the vertices left to their nearest kernels. The obtained clusters are adjusted by merging a combine of fractional communities to optimize the modularity of the network. EAGLE is another work using agglomerative hierarchical clustering based on maximal clique algorithm has been projected by Shen et al. [21] Firstly, maximal cliques are detected, and those smaller than a threshold are neglected. Then Subordinate maximal cliques are neglected, and the remaining give the initial communities. The similarity is found between these communities, and communities are repeatedly integrated along on the premise of this similarity. This is often used until one community remains at the end. Evans et al. [22] proposed that by partitioning the links of a network, the overlapping communities is also discovered. In another work, Evans et al. [23] used weighted line graphs. In another work, Evans [24] used clique graphs to discover the overlapping communities in real-world networks. Also Greedy clique expansion [25] at the first detect cliques in a network. These cliques act as seeds for expansion along with the greedy optimization of a fitness function. A community is discovered by expanding the selected seed and performing its greedy optimization via the fitness function proposed by Lancichinetti et al. [19] another work is Cluster-overlap Newman Girvan algorithm (CONGA) was proposed by Gregory. This algorithm was based on the split- betweenness algorithm of Girvan–Newman. CONGO optimized the proposed algorithm [26], which used a local betweenness measure, giving an improved complexity. A two-phase Peacock algorithm for overlapping community detection is proposed in Gregory [27] using disjoint community-detection methods. In the first phase, the network transformation was performed using the split betweenness concept proposed earlier by the author. Within the second phase, the remodeled network is processed by a disjoint community detection algorithm, and the discovered communities were transformed back to overlapping communities of the original network. [1]

3.2.2. Non-Clique Based Methods for Overlapping Community Detection

Non-clique based methods are used to detect the overlapping communities are shown in Table V these methods have been explained in this section.

Table V. Summary for non-clique-based methods for overlapping community detection [1]

Author (Algorithm)	**Approach**	**Parameters**	**Code Availability**
Nicosia et al.[29]	Modularity for overlapping communities genetic algorithm approach	in degree, out degree, belongingness coefficient	No
Pizzuti(GA-NET+) [30]	GA based	Community score	http://staff.icar.cnr.it/pizzuti/codes.html
Lancichinetti et al.(OSLOM)[31]	Edge direction	weights, hierarchy N vertices, E edges, degree of sub graph, internal and	http://www.oslom.org/

		external degree of subgraph	
Baumes et al. [32]	Clusters of overlapping vertices	Internal edge intensity external edge intensity internal edge probability, edge ratiointensity ratio	No
Chen et al. [33]	Game theory based	Set of communities Gain function, Loss function	No
Alvarietal. [34]	Game theory based Set of snapshots	with V vertices and E edges	https://github.com/hamidalvari/D-GT
Shi et al.(GaoCD) [35]	Objective function	partition density Size of population, running generation ratio of cross over ratio of mutation	No
Xing et al.(OCDLCE) [36]	Community detection	merging and refining Nodes, edges, neighbors of node	No
Zhang et al. [37]	Preference-based non-negative matrix factorization	Number of nodes, edges, communities	No
Kozdoba et al. [38]	Cluster aggregation	Probability measures, number of components, threshold parameter	No
Whang et al. [39]	Seed expansion	Nodes, edges, number of seeds, Page rank link following parameter, biconnected cores	https://www.cs.utexas.edu/-joyce/codes/cikm2013/nise.html
Rees et al. [40]	Egonets, Friendship groups	Number of nodes	No

Nicosia et al. [29] perform an extension of Newman’s modularity for directed graphs and overlapping communities. In this work a defined belongingness coefficient, a genetic approach has been used in this work for the optimization of modularity function. Pizzuti [30] produce one more algorithm that uses the genetic approach for overlapping community detection is GA-NET+. It discover overlapping communities using community score. The order statistics local optimization method (OSLOM) [31] discover clusters in networks and show many kinds of graph properties like edge direction, edge weights, overlapping communities, hierarchy, and network dynamics. This algorithm based on a fitness function stating the statistical significance of clusters with respect to random fluctuations, which is estimated with tools of extreme and order statistics. Baumes et al. [32] in another work reflected a community as a subset of nodes that induces a locally optimal sub graph with respect to a density function. Two subsets with remarkable overlap can be locally optimal, that helps to find overlapping communities. Chen et al. [33] produce an algorithm based on a game-theoretic approach to detect the overlapping communities. Every node is assumed to be an agent trying to improve the utility by joining or not the community. The community of the nodes is used to create the result of the algorithm. To detect the overlapping communities’ agent is formulated as the combination of a gain and a loss function. Every agent is permitted to select multiple communities. Alvari et al. [34] produce an algorithm based on two methods, PSGAME centered on the Pearson correlation and NGGAME based on the neighborhood similarity measure. GaoCD, produce by Shi et al. [35] discover overlapping communities. It detects clusters of links with the same structures as links ordinarily characterize distinctive relations between the nodes. Hence, nodes fit into many communities. The overlapping community detection by local community expansion algorithm [36] to detect overlapping communities was centered on community expansion. The algorithm has three phases, community detection, community merging, and community refining. Zhang et al. [37] have integrated the idea of implicit link preferences in this approach. It centered on preference-based non-negative matrix factorization (PNMF). They have considered the fact that for any node, preferences of the friends’ (neighbor) nodes are higher than any other non-neighbor nodes. Kozdoba et al. [38] produce an algorithm CLAGO to detect the overlapping communities the algorithm is divided into two steps. The first step, the disjoint communities are found, and then, a random walk through them helps to detect the overlapping communities. Whang et al. [39] produce neighborhood-inflated seed expansion, a seed based expansion method to detect communities. To target the main seeds, two approaches, ‘Graclus centers’ and ‘Spread hubs’ have been produced. For seed detecting, the algorithm of PageRank Clustering is used. Rees et al. [40] have used the idea of egonets and friendship groups. Egonet is the viewpoint from any node. It consists of the nodes that are adjacent to the central ego node and the edges between those nodes. From the egonets, the friendship groups are detected. The communities are then discovered from these friendship groups.

IV. POTENTIAL APPLICATIONS FOR ONLINE NETWORKS.

The enormous growth of the online networks and their users, the graphs representing these sites are becoming very complex, hence, are difficult to visualize and understand. Communities can be considered a summary of the whole network, thus making the network easy to comprehend. The discovery of these communities in online networks can increasingly being leveraged as a powerful, low-cost tool for enterprises to drive business objectives and functions such as enhanced customer interaction, greater brand recognition and more effective employee recruitment. Provide a mechanism for executives to assess consumer opinion and use this information to improve products, customer service and perception. Value add for additional and probably unique personalized service for the customer, Increase trust and customer loyalty Increase sales, click trough rates, conversion etc. Opportunities for promotion, persuasion Obtain more knowledge about customers. Discover new products and services. Also help in advertising, direct marketing and predict to which products or services a particular customer was likely to respond to, may also help in behavior modeling, prediction and collaborative filtering. Also can be useful in various applications such as **rating prediction,** predict the rating of target user for target item. **Top-N recommendation,** Forecast the top-N highest-rated objects between the items not yet rated by target user. **Link recommendation,** Predict the top-N users to which the target user is most likely to connect. It may help also in **Trend Analysis in Citation Networks,** citation networks are constructed by citation relationships between papers and researchers. Communities in a citation network characterize correlated papers on a single topic or researchers working on the similar topic. Here, the network is grouped into communities where a community is represented as papers on that topic or researchers working on that topic. Detecting these communities in citation networks can offer information about several core topics in which papers are published as well about the researchers working on these topics. May be a mechanism for **Improving Recommender Systems with Community Detection,** Recommender systems (RS) use data of similar users or similar items to generate recommendations. This is analogous to the identification of groups or similar nodes in a graph. Hence, community detection holds an immense potential for recommendation algorithms.

V. CONCLUSION

Community detection in recent incredible rising online networks provide a massive potential information. The main fundamentals of online networks, community structure, and two different approaches are presented in this paper. The first approach based on the relation between community members and the second approach cover the overlapping community detection methods, which categorized in two categories. The first category is clique based methods while the second category is non-clique based methods for overlapping community detection in online networks. Potential application for these different approaches to discover the communities in real networks of Facebook, Twitter, and online shopping networks etc. can provide a massive amount of information for many purposes. Analysis and discovering these communities is used in sociology, biology, computer science and many other branches of science. The discovered information could be valuable for commercial, educational, or developing determinations.

REFERENCES

[1] Bedi, Punam, and Chhavi Sharma. "Community detection in social networks." Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery 6.3 (2016):* 115-135.

[2] Cuvelier, Etienne, and Marie-Aude Aufaure. "Graph mining and communities detection." *Business Intelligence. Springer Berlin Heidelberg*, 2012. 117-138.

[3] Adedoyin-Olowe, Mariam, Mohamed Medhat Gaber, and Frederic Stahl. "A survey of data mining techniques for social media analysis." *arXiv preprint arXiv:1312.4617* (2013).

[4] Afsarmanesh, Nazanin, and Matteo Magnani. "Finding overlapping communities in multiplex networks." *arXiv preprint arXiv:1602.03746 (2016).*

[5] Zafarani, Reza, Mohammad Ali Abbasi, and Huan Liu. Social media mining: an introduction. Cambridge University Press, 2014.

[6] McCreesh, Ciaran, et al. "Clique and constraint models for maximum common (connected) subgraph problems." *International Conference on Principles and Practice of Constraint Programming*. Springer International Publishing, 2016.

[7] Reid, Fergal, Aaron McDaid, and Neil Hurley. "Percolation computation in complex networks." Advances in Social Networks Analysis and Mining (ASONAM), *2012 IEEE/ACM International Conference on. IEEE*, 2012.

[8] Palla, Gergely, et al. "k-clique Percolation and Clustering*." Handbook of Large-Scale Random Networks.* Springer Berlin Heidelberg, 2008. 369-408.

[9] Wang, Jianxin, et al. "Identifying protein complexes from interaction networks based on clique percolation and distance restriction." *BMC genomics 11.2* (2010): S10.

[10] McDaid, Aaron, and Neil Hurley. "Detecting highly overlapping communities with model-based overlapping seed expansion." Advances in Social Networks Analysis and Mining (ASONAM*), 2010 International Conference on. IEEE*, 2010.

[11] Zachary, Wayne W. "An information flow model for conflict and fission in small groups." *Journal of anthropological research 33.4* (1977): 452-473.

[12] Palla, Gergely, et al. "Uncovering the overlapping community structure of complex networks in nature and society." *arXiv preprint physics/0506133 (2005).*

[13] Nandi, G., and A. Das. "A survey on using data mining techniques for online social network analysis." *Int. J. Comput. Sci. Issues (IJCSI)* 10.6 (2013): 162-167.

[14] Pattabiraman, Bharath, et al. "Fast Algorithms for the Maximum Clique Problem on Massive Sparse Graphs." WAW. 2013.

[15] Palsetia, Diana, et al. "Clique guided community detection." Big Data (Big Data)*, 2014 IEEE International Conference* on. IEEE, 2014.

[16] Leskovec, Jure, Kevin J. Lang, and Michael Mahoney. "Empirical comparison of algorithms for network community detection." *Proceedings of the 19th international conference on World Wide Web. ACM, 2010*.

[17] Yang, Jaewon, Julian McAuley, and Jure Leskovec. "Community detection in networks with node attributes." Data Mining (ICDM), *2013 IEEE 13th international conference on*. IEEE, 2013.

[18] Harenberg, Steve, et al. "Community detection in large‐ scale networks: a survey and empirical evaluation." Wiley Interdisciplinary Reviews: Computational Statistics 6.6 (2014): 426-439.

[19] Lancichinetti, Andrea, Santo Fortunato, and János Kertész. "Detecting the overlapping and hierarchical community structure in complex networks." New Journal of Physics 11.3 (2009): 033015.

[20] Du, Nan, et al. "Community detection in large-scale social networks." *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007* workshop on Web mining and social network analysis. ACM, 2007.

[21] Shen, Huawei, et al. "Detect overlapping and hierarchical community structure in networks." Physica A: Statistical Mechanics and its Applications 388.8 (2009): 1706-1712.

[22] Evans, T. S., and Renaud Lambiotte. "Line graphs, link partitions, and overlapping communities." Physical Review E 80.1 (2009): 016105.

[23] Evans, Tim S., and Renaud Lambiotte. "Line graphs of weighted networks for overlapping communities." The European Physical Journal B-Condensed Matter and Complex Systems 77.2 (2010): 265-272.

[24] Evans, Tim S. "Clique graphs and overlapping communities." Journal of Statistical Mechanics: Theory and Experiment 2010.12 (2010): P12037.

[25] Lee, Conrad, et al. "Detecting highly overlapping community structure by greedy clique expansion." *arXiv preprint* arXiv:1002.1827 (2010).

[26] Gregory, Steve. "A fast algorithm to find overlapping communities in networks." Machine learning and knowledge discovery in databases (2008): 408-423.

[27] Gregory, Steve. "Finding overlapping communities using disjoint community detection algorithms." Complex networks (2009): 47-61.

[28] Adamcsek, Balázs, et al. "CFinder: locating cliques and overlapping modules in biological networks." Bioinformatics 22.8 (2006): 1021-1023.

[29] Nicosia V, Mangioni G, Carchiolo V, Malgeri M. Extending the definition of modularity to directed graphs with overlapping communities. J Stat Mech Theory Exp 2009, 3:P03024. doi:10.1088/17425468/2009/03/P03024.

[30] Pizzuti C. Overlapped community detection in complex networks. In: *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*, ACM, 2009, 859-866. doi:10.1145/ 1569901.1570019

[31] Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S. Finding statistically significant communities in networks. PLoS One 2011, 6:e18961. doi:10.1371/journal.pone.0018961.

[32] Baumes J, Goldberg MK, Krishnamoorthy MS, Magdon-Ismail M, Preston N. Finding communities by clustering a graph into overlapping subgraphs. In: IADIS *International Conference on Applied Computing*, 2005, 97–104.

[33] Chen W, Liu Z, Sun X, Wang Y. A game-theoretic framework to identify overlapping communities in social networks. Data Min Knowl Disc 2010, 21:224–240. doi:10.1007/s10618-010-0186-6.

[34] Alvari H, Hashemi S, Hamzeh A. Detecting overlapping communities in social networks by game theory and structural equivalence concept. In: Artificial Intelligence and Computational Intelligence. Berlin Heidelberg: *Springer-*

Verlag; 2011, LNAI 7003:620–630. doi:10.1007/978-3-642-23887-1_79.

[35] Shi C, Cai Y, Fu D, Dong Y, Wu B. A link clustering based overlapping community detection algorithm. Data Knowl Eng 2013, 87:394–404. doi:10.1016/j. datak.2013.05.004

[36] Xing Y, Meng F, Zhou Y, Zhou R. Overlapping community detection by local community expansion. J Inform Sci Eng 2015, 31:1213–1232.

[37] Zhang H, King I, Lyu MR. Incorporating implicit link preference into overlapping community detection. In: *Twenty-Ninth AAAI Conference on Artificial Intelligence,* 2015, 396–402.

[38] Kozdoba M, Mannor S. Overlapping community detection by online cluster aggregation. 2015, *arXiv preprint arXiv:1504.06798* .

[39] Whang JJ, Gleich DF, Dhillon IS. Overlapping community detection using seed set expansion. In: Proceedings of the 22nd ACM *International Conference on Information & Knowledge Management(CIKM),* ACM, San Francisco, CA, Oct 27-Nov 1 2013, 2099–2108. doi:10.1145/ 2505515.2505535.

[40] Rees BS, Gallagher KB. Overlapping community detection by collective friendship group inference. In: *International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2010,* 375-379, doi:10.1109/ASONAM.2010.28.