International Journal of Engineering & Technology, 7 (4.19) (2018) 857-863

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Greedy Modularity Graph Clustering for Community Detection of Large Co-Authorship Network

Ahmed F. Al-Mukhtar*[1], Eman S. Al-Shamery[2]

[1, 2]*University of Babylon, College of Information Technology, Department of Software Engineering, Iraq*
**Corresponding Author E-mail: ahmed.almukhtar@uokerbala.edu.iq*

Abstract

Social networks as a domain of complex networks that can be represented as graphs according to the patterns of connections among their elements. Social Communities are a set of nodes with denser connections inside community structures than outside. The goal of graph clustering is to divide the large graph into many clusters depending on multiple similarity criteria. In this work an improved version of the Louvain method is proposed, the *Greedy Modularity Graph Clustering for Community Detection of Large Co-AuthorshipNetwork (GMGC)*which introduces a new concept of weighted edges to enhance the accuracy of the Community Discovery for the large networks. The method is compared with other states of art methods mainly, Vertices Similarity First and Community Mean (VSFCM), and Generalized Louvain method for community detection in large networks (FKCD). Extensive experimental results have been madeon different datasets. The experimental results showed that the proposed method outperforms the other states of arts comparative methods according to the modularity optimization and community partitions evaluations measures.

Keywords: *Graph mining, Graph clustering, Community Detection, Social networks, Complex networks, Collaborative networks.*

1. Introduction

Recently, graph plays crucial roles in many domains of complex networks such as bioinformatics, social networks, sensor networks,and web, that because such kinds of systems can be visualized as connected nodes based on a specific relationship[1], [2]. Studying and analyzing networks structures aim to discover the structure of communities within networks throughout utilizing certain features encoded in these networks. Community structures or "modules" can see as groups of nodes, where nodes inside each group have denser ties than outside, basically, these nodes are arranged together based on similarity criteria. Community detection as an essential domain in social networks aims to distribute vertices based on their locations in the modules, by determining the structural boundaries of each module[3], [4]. Different graph clustering approaches have been developed to solve the community detection problem including spectral clustering, hierarchical clustering, both models cannot scale for large graphs.The fast, simple approach called label propagation in [4], which can execute in linear time. suitable for large networks nevertheless, the method leads to merging many smaller communities into a big cluster. Nowadays, and as social networks data is growing more and more an urgent need emerged to develop fast and more accurate algorithms that can solve problems in a reasonable time. In this paper, a community detection method is proposed for large weighted networks. The suggested method is an improved version of Louvain-method for community discovery. Called Greedy Modularity Graph Clustering for Community Detection of Large Co-Authorship Network (**GMGC**).

New weight concept based on a tuning parameter introduced,the weight function exploits mutual relations among connected nodes besides degree importance of each node. The remaining sections of this paper arrange as follows. Section 2 introduces a review of some most recent related works. Section 3 focuses on clustering techniques. Section 4 describes the proposed method. Section 5 reviews the experimental results. Finally, section 6 concludes the paper.

2. Related works

Ultimately, community detection as graph clustering approach, aims to auto-reveals on hidden clusters in the large network depending on related links among nodes. Zhou et, al in [5] proposes a graph clusteringfor community detection based on node degree and recommendation degree to analyze collaborative networks, using PAM clustering, their method has number of limitations in which, the number of clustered communities must be predetermined in prior in the method, and PAM clustering is suitable for small and medium networks. Boobalan et, al[6] introduced a new graph clustering method K-NAS using k-neighborhood Attribute Structural similarity they identify Local Outlier Factor LOF to find out the components of the dense node, nevertheless their method is based on k-nearest neighbors which needs to many iterations to be converged and also a number of clusters must be determined in prior.Yamazaki et, al in [7]comes with community detection model for large-scale co-purchasing network based on Clique Percolation Method (CPM). the algorithm works in two phases, graph polishing; cleans the original graph and clique enumeration; enumerate all maximal cliques in polished graph, eventually, the method inherits (CPM) limitations in that it focuses on in/out edge counting while ignores links interactions among vertices. k-prototype algorithm ISCD+ an iterative model for fast graph clustering, the authors in[8]introduce a new concept for community detection, the algorithm introduces two factors namely local importance and importance concentration to select nodes with different weights so

that to represent communities. In [9]the authors proposed an enhanced method of a median evidential c-mean algorithm for community discovery, they constraint on the prototypes-selection scheme, modularity is used to attain the optimal number of communities, however, as a drawback, the multicenter scheme is not used in their method which may reduce the accuracy of initial seeds.Moosavi et al.in [10], introduced a new approach to detect communities in social networks websites. Their strategy combines two concepts, the similarity of the structures and frequent pattern mining of nodes contents. The implemented algorithm has four steps, preprocessing, computing the frequent-pattern mining to obtain homogeneous groups, extending similar groups into small communities, finally small communities expansion. However, the method suffers from many disadvantages such as the trial and error which is used to determine the appropriate parameters and time complexity resulted from input parameters.Greeshma et al. in [11] proposed an overlapped community detection based on personalized page rank and one seed node, the method considers the well-connected components around the nodes, this can be achieved by applying random walk based page rank-nipple strategy, conductance technique is used to computes the communities, however the method is implemented for unweighted networks.

3. Graph Clustering Techniques

3.1 Community Detection Techniques

Several techniques have been utilized to investigate community structures of networks [4],[5]during last years. The matter of locating communities in a network is intended as a data clustering problem. It might be solved by assigning every node of the network to a cluster, in a meaningful manner. Mainly, there are two prospects to be investigated, Spectral clustering based techniques,and network modularity optimizations techniques. The former depends on cutting optimization process for the given graph network[12], where the problem of graph cut minimization has proven as NP-hard. So that, different approximation techniques have beeninvestigated. An example of using spectral clustering [4] for the moderated sparse matrix, by exploiting network eigenvectors of the Laplacian Matrix L, for the given components $L_{ij} = k_i\delta(i,j) - A_{ij}$, where k_i is the degree of a node i . $\delta(i,j) = 1$ *if and only if* $i = j$ and A_{ij} is the adjacency matrix for the connected graph. The latter relies on network modularity maximization, where the proposed method in this paper exploited this strategy, which is defined as a density measure to the fraction of links that is fall inside the community as compared with the remaining of links outside the community. The modularity canbe formulated as[13][14].

$$k_i = \sum_j A_{ij} \tag{1}$$

Where A_{ij} are the edges weights between i and j . k_i represents all edges that are attached to the vertex i .

$$m = \frac{1}{2}\sum_{ij} A_{ij} \tag{2}$$

m representsaweighted links in the network.

$$\omega(C_i, C_j) = \begin{cases} 1, & \text{if the Vertices } i \text{ and } j \text{ falls in the same} \\ & \text{module.} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\omega(C_i, C_j) \tag{4}$$

Modularity optimization was used by different methods to compare the quality of the partition. The high value of Q ,the detected communities have dense connections among its objects and sparsely connected to the others communities. Accordingly, the issue of maximizing the network modularity has been proved NP-hard complete[12], for this reason, several heuristic strategies are proposed to solve this problem. The well-known *Girvan-Newman strategy*, this approach based on betweenness centrality as a measure of significance, and Q value as evaluation of goodness[13]. The following stepssummarize the method:

i. Computing betweenness for all edges in the network.
ii. Remove the edge with the highest betweenness.
iii. Re-computing betweenness for all affected edges by removal.
iv. Repeat the steps ii), iii) until no edges remain within the network.

The definition of the partition of "goodness" is obtained by maximum modularity, as in Eq. (4).

However, the classical method has some resolution limits[12], means that it tends to detect super communities, subsequently it considerably slow down the algorithm.Finally, it has high computational complexity (i.e., $O(n^3)$, n is the number of nodes in the community) this make it inappropriate for a large network. Years later Clauset, Newman,and Moore (CNM) [14]introduce a faster algorithm, which can be run much faster on the sparse graph. However, it has proven that the method proposed by Clauset et al. is not scalable for networks size larger than 500,000 nodes [15].

The Louvain Method suggested by Blondel et al. in[16]. Considered as a greedy modularity optimization algorithm based on a local strategy that can implement on weighted networks. LM performs in two steps. Initially, each node considered as a community of itself, so there are $|V|$ communities, this maximized network modularity Q , then every node i moved into the community S thatsatisfies the most significantpositive integer gain, this canbe calculated in Eq. (5).

$$\Delta Q = \left[\frac{\sum_S + k_i^S}{2m} - \left(\frac{\sum_{S'} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_S}{2m} - \left(\frac{\sum_{S'}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)\right] \tag{5}$$

Where $\sum_S$ represents a weighted edge inside the community S . $\sum_{S'}$ is the weighted edges incident to nodes inside the community S . k_i is the weighted edges incident to the node i . k_i^S is the weighted edges that connect the node i to the nodes of the S structure. Finally m is the sum of all weighted links in the network. The second step starts with building a new network consisting of all nodes that grouped into communities from the first step, finally,both steps are iterated until no more improvements occur and maximal modularity is attained. Because of locality nature of LM, the modularity resolution limit has solved, since the resolution limit is caused by counting global network modularity[15], LM utilizing localized modularity measure to obtain network maximum modularity.There are other exciting techniques is discussed in this section, for instance, the Vertices Similarity First and Community Mean (VSFCM), Combines the characteristics of network topology with agglomerative clustering to detect communities in social networks. The method in [17] describes social network as identical-discrepancy-contrary, which utilizes the features of topological network and combines these features with

agglomerative clustering. A weighted Clustering coefficient connection Degree (WCCD) (such as vertex degree, clustering coefficient, connection degree, and path) is introduced as a weighted measure. The WCCD first considers the influence of discrepancy relation i on the transformation of discrepancy relation F into identical relation S. Secondly considers vertex cluster coefficient as i value based on density links among vertices. The hierarchical clustering is then used to calculate the similarity among communities and find max similarity that satisfies $\max\{Sim(C_i, C_j)\}$;the communities C_i, C_j are then merged to produce new community C_{new}. However, the method requires an extensive update of similarity operations. Thus it is not scalable for large networks.

Generalized Louvain method for community detection in large networks (FKCD), the authors in [18] introduce a generalized strategy of LM, which utilizes both local and global information of network topology to discover community structures. Calculating the centrality ranking for each edge (*k-path centrality*) lead to obtaining the global features, this done by computing the pairwise distance among nodes in the network. Once thisis done, theLM method is applied to find out community structures within the network. The FKCD method is suitable for relatively large networks. However, it tends to overrate estimating the number of discovered communities, and the value of $Q_{\max}$ is of the discovered communities are lower as compared with the original LM.

3.2 Database and Logic Programming

University of Trier web server every year publish an updated version of computer science bibliographic records which consist of computer science researchers. Mainly, the DBLP-bibliography consists of three main types of records as explained in [19], the first known as "Article-proceeding", which consists of the well-known proceeding journals, second one is called "Journal" contains Article published journals by famous publishers, finally "Author pages" which is used to track authors names in case they publish their articles with different names. Creatinga co-authorship network file from raw data requires a preprocessing step. Howeverthere are some constraints to overcome; mainly, DBLP-file comes in large size (more than 1.5 GB). Therefore it cannot parse the entire file in the memory at once.As a result, the raw data is divided into chunks to fit into the main memory; thenthese chunksare parsed one by one. A specific strategy is considered to filter the raw data. First, all journal records that have two or more authors are selected, then unify all authors names that written in different ways for each published article. For example, John Smith Swanson can be described in one article as JS Swanson, whereas in other paper as John S. Swanson, Author pages' can be helpful in this case.

4. The Proposed Method

4.1 Block Diagram

The proposed method architecture design and analysis of GMGC blog diagram as illustrated in Fig. ***1***, there are four main logical components namely the preprocessing, Graph creation, Community detection, and Evaluation. The first component, preprocessing, is considered as the first component of the block diagram where the raw data is processed and divided into chunks. This component consists of three Steps

- Data selection, where the journal's recordsare selected based on specific constraints.
- Authors Names Unification, this is a necessary step because many authors write their names in different ways in their papers.
- Network File Building, the last step of the preprocessing, the data is organized to put in the shape of a network file.

The next logical component, graph creation, which is a part of Graph mining, at this step the nodes are created and connected through edges to form connected graph, also extracting the essential topological features that are required by the next stage.

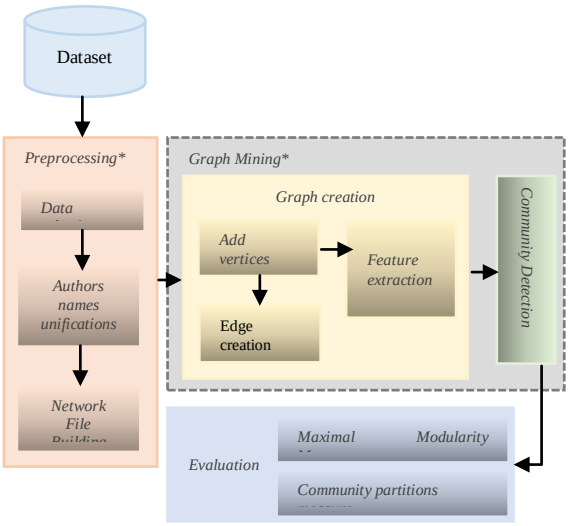

Fig. 1: GMGC Architecture

The third component, Community detection, devoted for extracting structures of communities based on a similarity measure using modularity function. Finally, the last component, Evaluation, applied to monitor and verify the quality of the system results, where system module displays all information simulated using the GMGC system.

4.2 GMGC Method

In this section, the community detection of the proposed method is described, which considered as an optimization problem. Each vertex in social networks is connected to one or more vertices with relation throughout some edges. Mathematically, the graph model is described as a set of the undirected weighted graph $G = \{V, E, W\}$. V is a set of vertices $(v_1, v_2, ..., v_n)$. E is a set of edges $(e_1, e_2, e_3, ..., e_n)$, each edge e_k connects two vertices (v_i, v_j), where W_k maps the edge e_k to a specific weight value. Also, each vertex v_i is a unique author in any published paper, and each edge e_i stands for co-authorship relation between two vertices. The relation described in Error! Reference source not found., the vertices a_i represent authors, and the edges e_j represent papers, the edge e_j connects two authors if they have shared the same paper. Table 1illustrates the co-authorship relations among vertices, where column two represents the number of different papers with respects to each author.

Table 1: The published papers concerning each author

Author (a_i)	Published Papers (P_i)
a_1	P_1
a_2	P_1, P_3
a_3	P_1, P_2
a_4	P_2
a_5	P_2
a_6	P_2
a_7	P_3
a_8	P_3

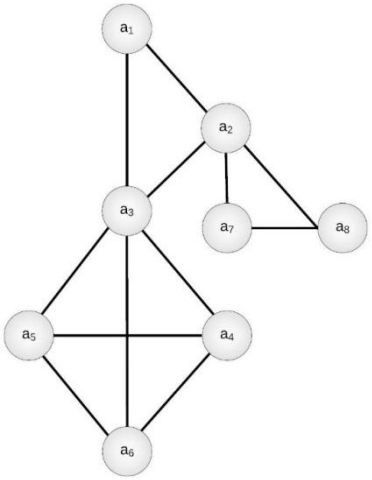

Fig. 2: Sample of co-authorship network

The proposed methodaims to group the similar vertices into one group by applying modularity optimization. Accordingly, it can be implemented in two main stages:

- Preprocessing stage;
- Graph mining stage;

The preprocessing stage is essential to extract network dataset from huge raw data, in this stage extracting all authors in journals based on certain constraints, the input to this stage DBLP-file, the output is the co-authorship network file, algorithm(1) describes the preprocessing step.

Algorithm 1: Preprocessing step

Input: DBLP-file
Output: co-authorship network file

```
For not end of file
    Divide file into chunks
End for
While not end of chunks
    For each Journal-record satisfy constraints
        Vector-file ←extract authors from Journal-record
    End for
End while
For each author in Vector-file
    Unified-name ←Check person page-record to unify authors name
    Update Vector-file
End for
For all authors in Vector-file
    Build co-authorship network file
End for
Return network
```

Graph mining stage concludes two stages, graph creation step; where the network $G=\{V,E,W\}$ is built based on the adjacency list. Community detection step; which is an improved version of LM. The proposed algorithm is contributed with a new weight function, to measure the relations among vertices. This measurement is describedbelow.

Definition 1 Vertex neighbors set. In the undirected weighted graph G, let $\forall v_m \in V$, $N(v_m)$ is denoted for all directly connected vertices to vertex v_m, as shown in Eq. (6).

$$N(v_m)=\{v_n \mid (v_m,v_n)\in E, \forall v_n \in V\} \quad (6)$$

In the Eq. (1), $\forall v_n$ represent all vertices $\{v_1,v_2,\ldots,v_n\}$ that are directly connectedto v_m. Eq. (6) represents the degree of each vertex. Therefore it quantifies the local importance of each vertex within its community

Definition 2, common neighbors set. Let $N(v_m), N(v_n)$ is the set of neighbors for v_m and v_n respectively, where $v_m, v_n \in V$. $\delta(v_m,v_n)$ is denoted as the common vertices between $N(v_m)$ and $N(v_n)$, as shown in Eq. (7).

$$\ell(v_m,v_n)=N(v_m)\bigcap N(v_n) \quad (7)$$

Eq. (7) shows the link reputation, this is done by selecting the common edges for each pair of connected vertices.

Definition 3 Edge weight. Let $v_n, v_n \in V$. The edge $(v_m,v_n)\in E$. Then $W(v_m,v_n)$ is denoted as weight measurement to the relation between v_m and v_n, as shown in Eq. (8).

$$W(v_m,v_n)=\eta+\left(\frac{1}{N(v_m)}+\frac{1}{N(v_n)}\right)\ell(v_m,v_n) \quad (8)$$

Where $N(v_m), N(v_n)$ count for all adjacent vertices to v_m and v_n respectively. $W(v_m,v_n)$ quantifies the reputation of the relationshipwithin the community, and η is the tuning parameter, it can be in the range [0, 1]. Specifying the value of η depends on network behavior which relies on the sparsity degree of the network. If the network is too sparse, then the value of η approaches to 1, in contrast, if the network tends to have less sparsity then the value of η approaches to 0.

4.3 Community Detection Algorithm

Algorithm 2: GMGC $\langle Graph\ G=(V,E,W)\rangle$

Input: undirected Graph G, η

Output: $k-$Communities $C_1,C_2,\ldots,C_k$

```
Initialization: vertices {v_i}_{i=1}^n = 0
Graph creation
    Add vertex
        If new node not in {v_i}
            Vertices {v_i} ← new node
        End if
    Add edge
        If v_i, v_j in vertices
            Create (v_i, v_j) link
        End if
Weight Computation*
    For each link (v_i, v_j), i ≠ j
        Compute W(v_i, v_j), Eq. (8)
    End for
Community Detection
    While improvement in Q-Modularity Do
        P ← Partition, Eq. (5)
        Q ← Network Modularity (P)
    End while
Return Communities, Q
```

Initially, GMGC community detection algorithm requires to determine the η factor in prior.At the graph creation step, the

vertices are added to the network. Also, all edges that connect these vertices are created.Next, all network edges must be weighted according to Eq. (8). The community detection step consists of two main phases: the first, the local modularity computations where each vertex is added to its neighborthat satisfies the highest modularity gain, iterations continue until local maximum modularity is gained. Phase two is applied to merge nodes from local communities into supper nodes, then the weights of these nodes are added up. Finally, both phasesarerepeated until no improvement in maximal modularityoccurred.

5. Experimental Results and Discussions

The results are obtained by analyzing several datasets using the GMGC method other stats of art methods mainly, Clause, Moore and Newman (CNM) based on modularity, Vertices Similarity First and Communities Mean (VSFCM) based on WCCD and agglomerative clustering, Louvain method (LM), and the Fast κ-path Community Detection (FKCD) a generalized LM based. The purpose of evaluating a specific algorithm is to ensure the considered algorithm with a particular goal can solve a definite problem in the field.Among the primary challenges in evaluating methods of detecting structures of communities are that there is no such clear determination of community structures in real-world networks, for this reason, each algorithm computes the similarity among vertices regarding its purpose. One of the main evaluating measures is the Maximal Modularity function, which is introduced by Newman and Girvan, as shown in Eq. (4). It has been widely used in the domain of social networks as a quantitative evaluation metric to measure the strength or weakness of community structure partition.

The experimentations of the proposed method have been conducted on real-world social networks, whose dataset available online. All experiments are performed on PC with Windows 10 Pro 64 bit, an i7-6700 HG CPU (260 GHz, and 16 GB RAM. The programming environment is Python 3.6.2 [MSC v.1900 32 bit (Intel)].

Dataset

Ninesocial networks datasets have been utilized to assess the proposed method; theyare based on structure social and communication networks, all datasets are described below,

Zachary's karate club(ZKC), is a social network friendship among 34 members of karate club it is a pairwise link among members who interact among themselves outside the club[20].

Dolphin social network (Dol.), which represents an undirected network of various associations, where the objects represent 62 dolphins[21].

Email network URV (Em.), undirected communication network, nodes represent a network of communication among 1,133 university members[22].

Jazz musicians network (Jaz), is a collaborative network among Jazz musicians each node represents Jazz musicians, edges is a relation between two musicians played together in a band[23].

AmericanCollege Football (FB), American football games network, collected in season fall 2000[24].

US Air97 network (UA), agent-agent social network, among 332 agents. It can be downloadedfrom[1]

SNAP/ca-HepTh (Hep), is undirected collaboration network of high energy physics, consists of 9877 nodes[25].

SNAP/ca-GrQc (GrQ), collaboration network, which is based on scientific collaborations among authors specifically in Quantum Cosmology category, which consists of 5242 nodes[25].

SNAP ca-CondMat (Con), collaboration network among authors papers, conains23,133 nodes[25].

In this paper all datasets are based on undirected networks,

Table *2*, represents the experimental results on relatively small datasets, the first column represents real networks. The second column is the number of vertices/edges statistics. The third column shows the statistical experimental results of CNM method; each cell represents the value of $Q_{\max}$ concerning numbers of communities. The fourth column is the VSFCM statistical experimental resultswhere the cellsrepresent the value $Q_{\max}$ concerning numbers of communities. The

Table 2: The experimental results of the community discovery algorithms for small datasets

N/W	**Ver./Edg.**	**LM**	**VSFCM**	**GMGC**	
ZK	34/78	0.381/3	0.419/4	0.493/4	η :0.5
Dol.	62/159	0.496/4	0.515/4	0.689/6	η :1.0
Em.	1133/5451	0.504/10	0.474/20	0.781/19	η :1.0
Jaz	198/2742	0.439/4	0.365/4	0.479/5	η :0.5
UA	332/2126	0.319/7	0.328/13	0.346/69	η :0.0
FB	115/613	0.548/6	0.578/7	0.901/12	η :0.5

From experience viewpoint, the experimental results show that the tuning parameter mainly can be one of the values which are either 0, 0.5 or 1 depending on network behavior.

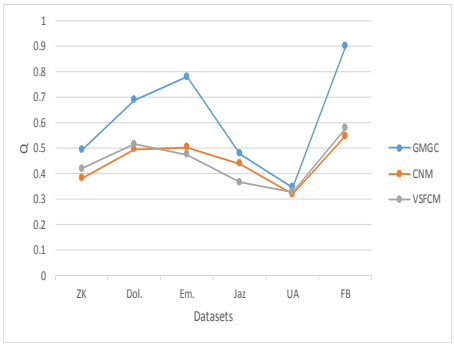

Fig. 3: Modularity results for six datasets

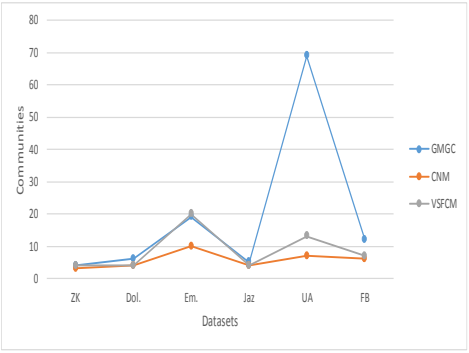

Fig. 4: number of partitions for six datasets

As illustrated in

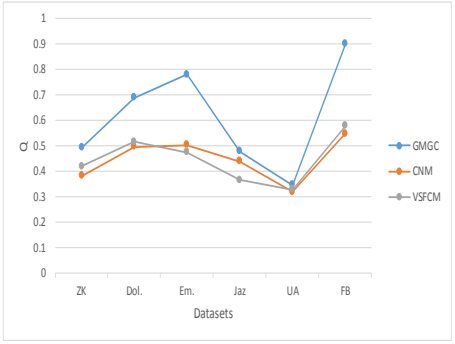

Fig. 3, and from the modularity point views the curve of the experimental results show that the GMGC method has the lead concerning other methods, this means the proposed algorithm has

achieved better community structure. Fig. ***4***, and from the perspective of community partition measure, GMGC method shows that the results converge to the number of communities in the real world especially in the case of FB and ZK, in contrast, and concerning the UA dataset GMGC gives a high number of partitions as compared with the other methods.Fig. ***5***and Fig. ***6*** review the conducted GMGC experimental results on ZK dataset, the curve results in Fig. ***5***represents Q_{max} concerning the tuning parameter η, as it can be seen Q_{max} shows the highest value when $\eta = 0.5$ and the lowest value of Q_{max} can be observed at both ends of the curve. Fig. ***6*** shows the experimental results number of community partitions concerning the tuning parameter η, when $\eta = 0$ the number of partitioned communities equals to five, and when $\eta = 1$ the networkwas partitioned into three communities. As for the rest of η values, the number of partitions settles to 4.

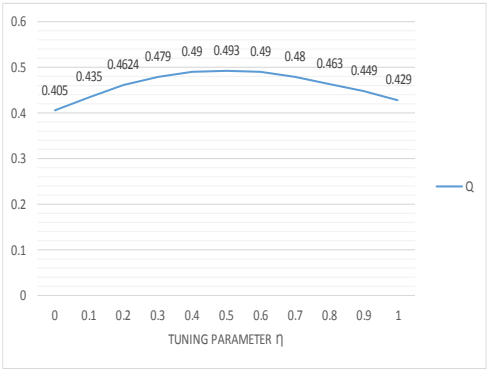

Fig. 5: Change in tuning parameter concerningthe Maximal Modularity for ZK dataset

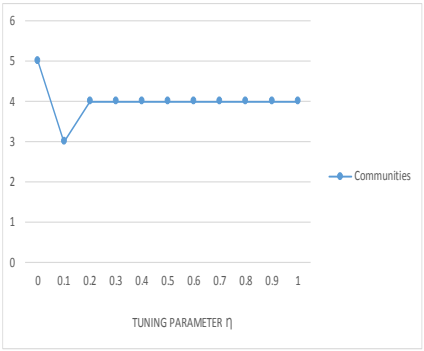

Fig. 6: Change in Tuning parameter concerningthe partitions number

For now, the conducted experimental results were based on relatively small datasets where the size is less than 1,150 nodes. Next,inTable 3,the proposed method was compared with the different state of art methods that can work on larger networks. Namely, LM and FKCD, the first column shows the experimental results for three datasets. The second column shows a network statistics of the number of vertices/edges concerning each network. The third column shows the LM method Q_{max} values corresponding to each dataset. The fourth column describes the conducted results of the Q_{max} concerning community partitions for the FKCD method. Where the last column shows the GMGC method, Q_{max} concerning community partitions and corresponding to the tuning parameter η.

Table 3: The experimental results of three community discovery algorithms for large datasets

N/W	**Ver/Edg**	**LM**	**FKCD$_{k=20}$**	**GMGC**	
GrQ	5,242/28,980	0.816	0.786/883	0.883/400	η :0.5
Hep	9,877/51,971	0.768	0.648/1,501	0.94/499	η :0.5
Con	23,133/186,932	0.731	0.599/2,819	0.703/3033	η :0.0

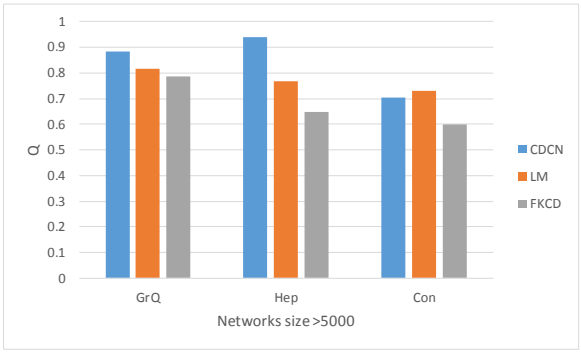

Fig. 7: Modularity results for three datasets

The illustrated results in Fig. ***7*** shows the performance of GMGC, LM, and FKCD according to Q_{max} modularity, it is shown that the GMGC method outperforms the other methods concerning***GrQ*** and ***Hep*** networks, whereas, and according to the Q_{max} results it is almost equal to LM concerning the***Con*** network.
Finally, the Co-authorship dataset network was implemented by the GMGC method. The illustrated experimental results inTable 4show that the best-achieved results concerning Q_{max} and the community partitions when setting tuning parameter η to 1.

Table 4: Statistical results of Co-authorship

Statistics	**results**
No. of nodes	98816
No. of edges	351368
Chosen (η)	1.0
modularity Q_{max}	0.840
Community partitions	7461

6. Conclusion

Nowadays, and with the rapid incremental of the social network's data new computational researches have become required, since that, the community detection has become playing the crucial role in the field of social networks analysis. The suggested work, Greedy Modularity Graph Clustering for Community Detection (GMGC) that can handle large network data. The proposed method introduces a new weight concept based on the tuning parameter. The weight function exploits mutual relations among connected vertices besides the reputation degree of each vertex. The tuning parameter has a significant benefit since it controls the shape, and community partitions, this is clearly shown by observing the increase in Q modularity. Extensive experimental results have been made using different datasets, involves comparing the proposed method with many states of art methods such as*CNM, LM, VSFCM,*and *FKCD*. All experimental results showed that GMGC was better than the other methods according to Q modularity and community partitions measures. For future work to put forward, it is recommended to adopt a heuristic strategy to facilitate detecting the tuning parameter automatically.

References

[1] M. Panda, S. Dehuri, and G.-N. Wang, Eds., Social Networking, vol. 65. Cham: Springer International Publishing, 2014.
[2] D. F. Nettleton, “Data mining of social networks represented as graphs,” Comput. Sci. Rev., vol. 7, pp. 1–34, Feb. 2013.
[3] S. Fortunato and D. Hric, “Community detection in networks: A user guide,” Phys. Rep., vol. 659, pp. 1–44, Nov. 2016.
[4] S. Fortunato, “Community detection in graphs,” Phys. Rep., vol. 486, no. 3–5, pp. 75–174, Feb. 2010.

[5] H. Zhou, J. Li, J. Li, F. Zhang, and Y. Cui, "A graph clustering method for community detection in complex networks," Phys. Stat. Mech. Its Appl., vol. 469, pp. 551–562, Mar. 2017.
[6] M. P. Boobalan, D. Lopez, and X. Z. Gao, "Graph clustering using k-Neighbourhood Attribute Structural similarity," Appl. Soft Comput., vol. 47, pp. 216–223, Oct. 2016.
[7] T. Yamazaki, N. Shimizu, H. Kobayashi, and S. Yamauchi, "Weighted Micro-Clustering: Application to Community Detection in Large-Scale Co-Purchasing Networks with User Attributes," 2016, pp. 131–132.
[8] L. Bai, X. Cheng, J. Liang, and Y. Guo, "Fast graph clustering with a new description model for community detection," Inf. Sci., vol. 388–389, pp. 37–47, May 2017.
[9] K. Zhou, A. Martin, Q. Pan, and Z. Liu, "Median evidential c-means algorithm and its application to community detection," Knowl.-Based Syst., vol. 74, pp. 69–88, 2015.
[10] S. A. Moosavi, M. Jalali, N. Misaghian, S. Shamshirband, and M. H. Anisi, "Community detection in social networks using user frequent pattern mining," Knowl. Inf. Syst., vol. 51, no. 1, pp. 159–186, Apr. 2017.
[11] V. Greeshma and K. S. Vani, "Community Detection in Networks Using Page Rank Vectors," Int. J. Bioinforma. Biosci., vol. 5, no. 1/2/3/4, pp. 01–07, Dec. 2015.
[12] P. Ambika and M. B. Rajan, "Survey on diverse facets and research issues in social media mining," in Research Advances in Integrated Navigation Systems (RAINS), International Conference on, 2016, pp. 1–6.
[13] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," Phys. Rev. E, vol. 69, no. 2, Feb. 2004.
[14] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," Phys. Rev. E, vol. 70, no. 6, p. 066111, 2004.
[15] M. Chen, K. Kuzmin, and B. K. Szymanski, "Community detection via maximization of modularity and its variants," IEEE Trans. Comput. Soc. Syst., vol. 1, no. 1, pp. 46–65, 2014.
[16] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," J. Stat. Mech. Theory Exp., vol. 2008, no. 10, p. P10008, Oct. 2008.
[17] X. Chen, J.-F. Guo, F.-C. Liu, and C.-Y. Zhang, "Study on similarity based on connection degree in social network," Clust. Comput., vol. 20, no. 1, pp. 167–178, Mar. 2017.
[18] P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Generalized louvain method for community detection in large networks," in Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on, 2011, pp. 88–93.
[19] M. Ley, "DBLP: some lessons learned," Proc. VLDB Endow., vol. 2, no. 2, pp. 1493–1500, 2009.
[20] W. W. Zachary, "An Information Flow Model for Conflict and Fission in Small Groups," J. Anthropol. Res., vol. 33, no. 4, pp. 452–473, 1977.
[21] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," Behav. Ecol. Sociobiol., vol. 54, no. 4, pp. 396–405, Sep. 2003.
[22] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in organisations," Phys. Rev. E, vol. 68, no. 6, Dec. 2003.
[23] P. Gleiser and L. Danon, "Community Structure in Jazz," Adv. Complex Syst., vol. 06, no. 04, pp. 565–573, Dec. 2003.
[24] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proc. Natl. Acad. Sci., vol. 99, no. 12, pp. 7821–7826, Jun. 2002.
[25] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in Pro. 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 631–636.