Received: August 16, 2022. Revised: October 24, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

# Exploring Overlapped Communities Based on Optimized Community's Resolution Controlling Evaluation in Social Networks

**Eaint Mon Win[1]*** **Si Si Mar Win[1]**

[1] *University of Computer Studies, Yangon, Myanmar*
* Corresponding author's Email: eaintmonwin@ucsy.edu.mm

**Abstract:** Community structure is one of the main structural features of networks and detecting overlapped community structure is an important field in social network analysis. In recent years, local community detection algorithms which detect overlapped community structure have been developed. However, the most existing algorithms suffer unstable community structure because the influences of parameter for controlling community's resolution of fitness evaluation functions where used in identifying communities. Therefore, this article designs the optimized parameter evaluation formula to avoid the parameter influences and the algorithm is modelled on local expansion strategy. In this work, firstly identifies the seed or core node by using extended jaccard similarity and form initial community via seed. Then local community is detected by expanding the initial community with fitness function based on proposed optimized parameter evaluation and finally overlapped nodes are identified by merging detected local communities. In this article, the algorithm is implemented by using small dataset from network data repository site and large networks from Stanford large network datasets collection. The performance results of algorithm are compared with LFM (Local Fitness Method), OSLOM (Order Statistics Local Optimization Method), DEMON (Democratic Estimate of Modular Organization of a Network), NILPA (Node Important based Label Propagation) and GREESE (Greedy Coupled-seeds Expansion) algorithm. Then, the proposed algorithm proves that it effectively performs and saves execution time.

**Keywords:** Overlapped community, Local expansion, Seed, Optimized fitness function, Jaccard similarity.

## 1. Introduction

In real world, social networks have been rapidly developed with the development of information technology and gradually become an important platform for people to exchange feelings, share experiences, and transmit the information. Not only social networks but also other complex networks such as academic cooperation networks, world wide networks, internet networks and biological networks are ubiquitous in recent years. Therefore, complex network analysis has emerged as one of the most exciting domains of data analysis and mining over the last decade. A network consists of nodes and edges, which connect a pair of vertices. Many of these networks show community structure, which means that they are generally composed of nodes, called communities or clusters. Generally, a community is commonly a sub network of a network that is densely connected internally but is sparsely connected to the rest of the network [1]. Community detection is used in various applications where group decisions are taken, e.g., delivering information within group or recommending products to group such as Target Advertising, Criminology. The most important type of community is disjoint community which means a node belongs to only a community.

At the present, there are many community detection algorithms as well as clustering algorithms. In the last years, most researchers uncovered the community structures among the network by applying clustering algorithms such as Kernighan-Lin [2], K-mean [3]. Girvan and Newman's method [4] is the first algorithm, used edge betweeness as the measurement in a divisive method for detecting communities. That algorithms strictly divide each node into specific communities but cannot find overlapping communities. When in-depth study,

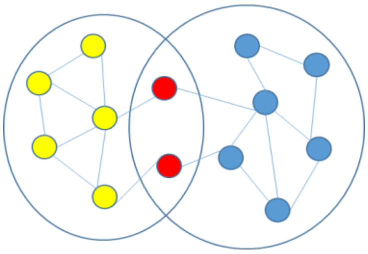

Figure. 1 Overlapped communities with two overlapped vertices

researchers found that many networks have not only disjoint but also overlapping characteristics which are sharing nodes with many communities. The overlapping community is illustrated in figure1. The network is composed of many overlapping and interconnected communities, e.g. in interpersonal networks, everyone belongs to several different communities such as school, family, friends, locations and their works. Therefore, overlapping community detection has become a key issue in network analysis.

The type of algorithm for detecting overlapped structure is divided into four categories: Clique Percolation (CPM), Link partition, Label propagation (LPA) and Local expansion. The search algorithm CFinder [5] based on Clique Percolation Method to find k-clique percolation clusters. A k-clique is a completely connected by k nodes. 3 cliques is a sub graph that completely connected three adjacent nodes. The higher the value of k, the smaller the size of the highly dense groups. But clique based algorithms face the NP hard problems. Gregory [6] and J. Xie et.al [7] developed Label propagation algorithms, COPRA, SLPA respectively. COPRA is a fast algorithm and an iterative method, based on the idea of multi label propagation. It updates the labels according to belonging coefficient of its neighbors until reaches the nodes share the same label into a community. SLPA is adopted on idea of assigning the nodes as speaker and listener for propagating the labels corresponding to listener rules. Finally threshold is used to generate overlapping communities as post processing. In 2012, Coscia and Rossetti [8] proposed local first discovery method by label propagation, called DEMON. It is a democratic approach, where the peer nodes decide if their neighbours should be cluster. The weakness of LPA, it doesn’t know how many iteration number have to be stopped as stop criteria.

Lancichinetti [9] firstly introduced a local expansion algorithm, called LFM to identify overlapped objects and hierarchical structure in network. Initially, it selects random seed and then extended this seed on f fitness function. It is implemented on benchmark graphs. In 2011, OSLOM [10] based on the local optimization of fitness function is introduced to detect community structure in large networks. Its main characteristic is that it based on fitness measure which express the statistical significance of clusters. Those algorithms depend on parameter of the fitness evaluation functions when extend the community. Therefore, they meet influences of various parameter values. In recent year, as well as community expansion algorithms, the seed expansion algorithms have been introduced to choose appropriate seeds because initial seed is selected randomly according to the local expansion strategy. Joyce [11] identified seed by using Graclus Centres and Spread Hubs. When expand the seeds, this algorithm applied personalized PageRank Clustering that optimizes the conductance score. However, it identifies the number of seed to produce number of communities as prior knowledge. Belvin [12] developed a blended strategy with various centrality measures to find superior seed set. It can compare running time over the overlapped algorithms. But it worked on only small datasets. This paper finds seed by using extended jaccard similarity and builds initial community with neighbors of seed. Then, expands this community based on optimized fitness function with parameter evaluation. The main contributions of this work are:

(1) Optimized parameter evaluation function is proposed to decide whether a node should be added or not. It can avoid influences on parameter of fitness functions.

(2) Extend jaccard similarity is used to identify appropriate seed because seed selection process is also important case.

(3) The overlapped detection algorithm is designed and it saves running time.

The rest of paper is organized as follows: In section 2, related works are described. Basic notations as preliminaries are defined in section 3. Next section explains the propose algorithms and system design. Section 5 implements the algorithm and compares the performance results of algorithms. As final section, summarizes the paper.

## 2. Related works

Many researchers have been done local community detection algorithms instead of global due to the time consuming to get global information of the whole network. H. You et.al [13] modified LFM algorithm by expanding local community instead of seed expansion of LFM after choosing the node with maximum degree. In 2017, Xiaobo and Chuxiang [14] modeled an algorithm, called LFMs, to avoid instability of LFM due to random seed selection. LFMs which improved LFM algorithm, considered weighted information. Firstly, it used random walk method to select seed nodes. Then, with cosine similarity to calculate vertex similarity, weight information in network was fully used. This work redefined f fitness function based on similarity matrix. It identified seed as core node for each community rather than randomly seed selection. However, cannot identify good seeds when same maximum degree of nodes. In that work, parameter value of fitness function is set from 0.8 to 1.4. In this case, various implementation results are occurred by tuning parameter setting of fitness function.

To find the good seeds, new overlapped detection algorithm [15] was proposed by calculating node weighting. It calculates the node's weight according to neighbourhood overlap and locates the node with highest weight as the core node. Neighbourhood overlap is denoted as

$$Sim(u,v) = \frac{n_{uv}}{k_u + k_v - n_{uv} - 2} \qquad (1)$$

$n_{uv}$ represents the number of common neighbors, $k_u$ and $k_v$ are the degree of nodes u and v. Then expands this core node to obtain local community by adding the neighbour nodes that satisfy improved community quality metric. As final stage, overlapped objects are detected by merging local communities with overlapping score. This work was done on small real world networks. Although avoid selecting random seed and influences of parameter on results with multiple implementations, it cannot handle large complex networks.

In order to solve the excessive overlap and multiple runs with various parameter, node membership degree was proposed in [16]. It constructs initial community with highest important node and its neighbour nodes. The important node is evaluated on clustering coefficient and degree. Then the initial community is extended by adding nodes that satisfies of node attribution degree. To improve the quality of community structure, the added nodes into the community is accessed by quality assessment function. It was done on only small real world networks.

In 2019, Clique-based overlapping community detection [17] was developed to detect overlapped objects with satisfactory time efficiency. It selects the maximum density node as the core node after the extracts k cliques sub graph from the network. Then initial community is formed with this core node and k clique is added to the community instead of single node is added according to the quality of f fitness function. This algorithm set the parameter of function as default 1. Therefore, it can avoid the instability of community structure due to various parameter setting. However, clique-based methods is inconvenient for sparely connected graphs.

In 2020 [18], Two Expansion Seed (TES) proposed gravitational degree force where used in seed selection for robust discovery. As first time expanded, the selected seeds are expanded by greedy strategy based on f fitness function. After that, first discovered communities are expanded by gravitational degree in second time. It set $\alpha$ parameter from 0.8 to 1.5 and found results of various implementation. Guo et.al [19] uncovered the community structures to address the problem of quality and stability deficiencies on large scale networks. This work detected core members as central nodes or seed set by using local degree central nodes and jaccard coefficient. It takes an initial community for each seed. In extension stage, the community is extended by combining internal force between nodes with fitness function. This method improves the accuracy in complex networks but cannot detect overlapped communities and the implementation is iteratively tested by tuning parameter. Therefore, the instability problem of detected communities still appears.

NI-LAP [20] detected overlapped communities and improved label propagation algorithm. It has three phases: initializing, propagation and filtering process. In initialization, each node is initialized with unique label and node importance or seed is defined by using clustering coefficient and maximum degree. In propagation, this process is based on its neighbours' label to update label with its belonging coefficient. After sum belonging coefficient of the same label, remove the pairs with coefficient lower than threshold as filtering process. This method decides the communities by threshold to delete low coefficient even same label and takes several times to run large scale network.

Wang [21] designed node relevant centrality, is node important evaluation method to select core nodes. This node is selected as seed if score is greater than threshold according to node relevant centrality. In community expanding, neighbours of seed are added to community as candidate nodes by using adaptive function based on fitness function.

Asmi et.al [22] developed WLC to detect overlapping communities. In that research, new weighted belonging degree is defined by focusing the clustering coefficient and common neighbour similarity. It constructed initial community with a node which has largest node strength and then community is expanded by using weighted belonging degree. In 2021 [23], they proposed coupled-seeds expansion method (GREESE) instead of single seed selection. GREESE is composed of four phases: seed selection, extending, propagation and merging phase. For selecting couple seed, it applied common neighbour similarity. It is defined as follows:

$$C(v_i, v_j) = |N(v_i) \cap N(v_j)| \quad (2)$$

This coupled seed with highest similarity is considered as the centre of community and then extended the seed by maximizing the local fitness function that is designed on coupled seed. In extension phase, if fitness of each neighbour of couple seed reaches the threshold, that neighbour node is joined to community. This process is repeated until all neighbors of community meet threshold. In propagation, isolation nodes are assigned to the suitable elementary groups that contain the largest number of their neighbors. Finally, the local communities are merged by considering threshold that is one third of the node existing in smallest communities between discovered and real ones. This work adopted suitable time efficiency in complex networks with no parameter tuning.

Most researches developed many methods to avoid the parameter influences of fitness function f. However, they still need threshold as prior knowledge when extends the local community even the existing algorithms address the problem which lack of stable and ineffective local community. That why, this article focusses on solving the instability of community structure due to different experiments by defining optimized parameter evaluation function. The proposed function finds appropriate parameter value to control community's resolution and provides in considering the fitness quality of nodes to add them into community. Moreover, the proposed algorithm locates suitable seeds for each community and also performs well on both small networks and large scale networks on little time.

## 3. Preliminaries

A graph is usually modelled as G (V, E) and social network is represented as graph model. Let G (V, E) be an undirected and unweighted graph. V represents set of vertices from G and E represents the set of Edges. Denote by $C = \{C1, C2, \ldots\ldots, Cn)$ the network community structure, i.e. a collection of n sub graphs where $Ci \epsilon C$, $Ci$ is the set of V. The notation descriptions of this section are listed in Table 1.

### 3.1 Jaccard similarity

Jaccard similarity or Jaccard coefficient [24] between two vertices is

$$Sim(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|} \quad (3)$$

This equation measures the similarity between two vertices. The larger the value, the more similar between them. However, this measure emphasizes the common neighbours of vertices u and v. Therefore, it produces the zero similarity for pair of vertices that has no common neighbour nodes.

#### 3.1.1. Extended jaccard similarity

Extended jaccard similarity is defined to avoid zero similarity by adding one in numerator.

$$Sim(u, v) = \frac{|N(u) \cap N(v)| + 1}{|N(u) \cup N(v)|} \quad (4)$$

The traditional jaccard can produce the zero if there are no common neighbors of pairs of vertices.

However, the seed is the centre of a community and it depends on node strength (i.e. seed is a core node which closely related to the many other nodes.) According to the Eq. (3), if no common neighbors between two vertices, zero similarity is accepted although some nodes have number of links. For that reason, use above Eq. (4) when finding the seed in this paper.

### 3.2 Weight evaluation

This weighted evaluation is based on the extended jaccard similarity. It is defined as follows:

$$W_u = \sum_{u \in V} Sim(u, v) \quad (5)$$

This evaluation is sum of the similarity between each node and its neighbors. In this way, the core node or seed is assigned by weighted evaluation. Generally, the vertices with high degree in the network do not actually have high weighting. Therefore, weight of node isn't considered on the degree of each vertex in this work and it is considered by using Eq. (5).

### 3.3 Fitness evaluation function

Lancichinetti et.al [9] proposed a fitness evaluation function f to measure tightly connected to the internal nodes of a community. This function is identified as follows:

Received: August 16, 2022. Revised: October 24, 2022.

Table 1. Formal notation

| Notation | Description |
|---|---|
| $G(V,E)$ | Graph consisting vertices and edges |
| $N(u) \cap N(v)$ | Common neighbour of vertex u and v |
| $N(u) \cup N(v)$ | Total number of neighbours of vertex u and v |
| $k_{in}^{C}$ | Number of internal links within community |
| $k_{in}^{C} + k_{out}^{C}$ | Total number of internal and external links of community |
| $\alpha$ | Community's resolution controlling parameter |
| $\rho$ | Density of graph |
| $k_{ext}$ | Number of external links of initial community |
| $k_{int}$ | Number of internal links within initial community |
| $d_{avg}^{G}$ | Average degree of graph |

$$f_C = \frac{k_{in}^{C}}{(k_{in}^{C}+k_{out}^{C})^{\alpha}} \qquad (6)$$

$\alpha$ is a positive real value parameter which is used to adjust the community scale. This quality evaluation function effectively measures the densely connected nodes within communities. Lancichinetti experimented his research with range [0.6, 1.6] and assumed 1 as default. Therefore, various implementation results absolutely rely on the parameter and the best result is occurred by changing different parameter values.

### 3.4 Optimized parameter evaluation function

In Eq. (6), parameter fluctuation occurs various implementation with different results. Therefore, community's size is unstable and don't know when the best result will be meet exactly. To control the community scale in this work, parameter evaluation formula is defined as follows:

$$\alpha = \frac{log(k_{int}-d_{avg}^{G})+\rho}{log(k_{ext})} \qquad (7)$$

This evaluation is worked according to the initial cluster, density and average degree from the whole graph.

## 4. Proposed system

In this section, local expansion algorithm to detect overlapping communities is described and the system design is shown in Fig. 2. This system discovers overlapping objects based on local expansion strategy by optimizing f fitness function for solving parameter independence. Firstly, seed is identified by using extended jaccard similarity and then extends the community by optimized f fitness function. Finally, local communities are merged to detect overlapped nodes.

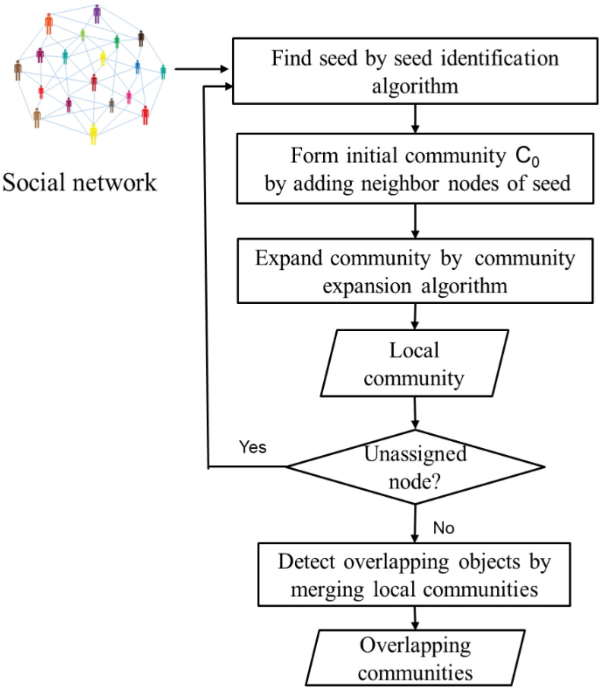
Social network
Find seed by seed identification algorithm
Form initial community $C_0$ by adding neighbor nodes of seed
Expand community by community expansion algorithm
Local community
Unassigned node?
Yes
No
Detect overlapping objects by merging local communities
Overlapping communities

Figure. 2 Proposed system design

### 4.1 Seed identification

The seed is the important node among all of nodes to form local community. Initially, computes the similarity of all pairs of nodes from the whole network using Eq. (4) and then calculates the weight of each node according to Eq. (5). The node with the highest weight is selected as seed and this seed is assigned to initial community. After this process, initial community is discovered with neighbour nodes surrounding seed. Therefore, community expansion process will be continued. After expand the community and obtain a local community, check if there are unassigned nodes to any communities. If that, a node with highest weight from remaining nodes is selected as next seed node to form next local community. Algorithm 1 describes the pseudo code for seed identification.

### 4.2 Expanding community

After accepting seed, initial community is formed with neighbour nodes of seed. Fitness quality of initial community is calculated by using Eq. (6) and then fitness of all members within initial community is also found to decide whether member is removed or retained in the initial community. If fitness of initial community is greater than fitness of member node then it is removed from initial. If not, it is

Algorithm 1. Seed identification algorithm

```
Input: G (V, E)
Output: seed
seed = 0; weight=0;
for each v ∈ V
    calculate similarity Sim (u, v) of each node by using Eq. (4); // u ∈ neighbor of v
    weight (v) = ∑ Sim (v, u);
end for
seed= max {weight (v)};
return seed;
```

retained in initial. Then neighbours of community are considered to extend the community. Therefore, fitness of each of neighbour nodes are calculated and node with higher fitness is added to the community. This process continues until the all neighbours of community satisfy the fitness value. After this process, a local community is obtained. This process is described in Algorithm 2.

Algorithm 2. Community expansion algorithm

```
Input: Initial community $C_0$ , G (V, E)
Output: Local community C
C=∅;
for each v ∈ $C_0$
    calculate fitness value F according to f fitness function;
    if F ($C_0$\v) >= F($C_0$ ) then C= $C_0$ – v;
    end if
end for
for each u ∈ $N(C)$ // neighbor nodes of C
    calculate fitness value F according to f fitness function;
    if F(C ∪ $u$) > F(C) then C=C ∪ u;
    end if
end for
return C;
```

If there are still remain nodes from graph, finds second seed node and detects next local community surrounding second seed. The seed identification and expansion process stop if all of the nodes from graph have been assigned to the corresponding communities.

Finally, each local community of each seed is obtained and all of these local communities are merged. In that case, some nodes share one or more communities and not exactly a community. It leads to the overlapped vertices.

## 5. Experiment and results

This part consists of experimental results of the proposed system. It is tested by using three evaluation criteria on different datasets. As baseline methods, LFM, OSLOM, DEMON and GREESE are selected. This experiment is implemented in java with RAM 8G and processor i5, CPU 2.5GHz and window operating platform. In this experiment, the performance of proposed system is measured with ENMI, F1 and $Q_{ov}$ to know the quality of overlapped communities. According to the literature, ENMI, F1 can be measured with ground truth information and $Q_{ov}$ are measured with the absence of ground truth. The real world datasets used in this implementation are described in section 5.1. The performance results and execution times are also compared to other state of the art algorithms.

### 5.1 Datasets

The popular real world networks applied in the most overlapping community detections algorithms as researchers. Among them, some datasets are used in this experiment. The proposed algorithm performs experiments on nine network datasets and network descriptions are described in Table 2. The applied real world networks are followings:

Zachary's karate [25]: is a very well-known network mostly used by many researchers. Wayne Zachary collected data from the 34 members of a university karate club. He constructed a network of friendships between members of the club. Each node represents a member of the club, and each edge represents a tie between two members of the club.

Dolphin [25]: is an undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand. The dataset contains a list of all of links, where a link represents frequent associations between dolphins. Football [25]: contains the network of American football games between Division IA colleges during regular season fall 2000. The nodes indicate team to

Table 2. Real world networks

| Data set | No of node | No of edge | Density | Average degree | Ground truth |
|---|---|---|---|---|---|
| Karate | 34 | 72 | 0.14 | 4 | Y |
| Dol_phin | 62 | 159 | 0.08 | 5 | Y |
| Foot_ball | 115 | 613 | 0.09 | 10 | Y |
| Politic_al | 105 | 441 | 0.08 | 8 | Y |
| Risk map | 42 | 83 | 0.096 | 3 | Y |
| Face book | 2888 | 2981 | 0.00007 | 2 | N |
| Net science | 22963 | 48436 | 0.00257 | 3 | N |
| Ama_zon | 334863 | 925872 | 0.004 | 5 | Y |
| Dblp | 317080 | 1049866 | 0.00119 | 6 | Y |
| Astro_Ph | 16046 | 121251 | 0.0009 | 15 | N |
| Cond_Mat | 39577 | 175692 | 0.00022 | 8 | N |
| Tech-pgp | 10680 | 24316 | 0.00043 | 4 | N |

which conferences they belong. The links represent game between teams when they play together. The teams are divided into conferences containing around 8–12 teams each.

Political book [25]: A network of books about US politics published around the time of the 2004 presidential election and sold by the online bookseller Amazon.com. Edges between books represent frequent co purchasing of books by the same buyers, as indicated by the "customers who bought this book also bought these other books" feature on Amazon.

Collaboration network [25]: Net science, AstroPh, CondMat and Dblp are scientists' paper collaboration networks. In these networks, node represents author and relationship between them indicates their paper collaboration.

Risk map [26]: is a map of the popular strategy board game. It is a political map of the Earth, divided into 42 territories, which are grouped into six continents.

Facebook [27]: contains Facebook user–user friendships. A node represents a user. An edge indicates friend relationships of user to user.

Amazon [27]: is a network where nodes represent products and an edge between products i and j signifies that product i was frequently co purchased with product j.

### 5.2 Evaluation criteria

To evaluate the performance of proposed algorithm, the system is measured with three evaluation metrics on both ground truth and no ground truth information. The mostly used metrics to measure quality of overlapping communities are ENMI, F1 and they are evaluated with ground truth. The overlap modularity $Q_{ov}$ is measured with no ground truth. Extended normalized mutual information (ENMI) [9] is defined to measure the performance of overlapped structure for ground truth datasets. That metric is defined as:

$$ENMI(X|Y) = 1 - \frac{1}{2}[H(X|Y) + H(Y|X)] \quad (8)$$

H (X|Y) is noted as normalized conditional entropy for cluster X with respect to cluster Y. X and Y are random variables related to partition C and C cover. These results are occurred with different datasets in Fig. 3.

F1 [28] measure is also widely used in performance evaluation of overlapped communities at community level. To obtain F1 between discovered community and ground truth, precision and recall are defined as follows:

$$precision = \frac{OPN}{|C_{true}|} \quad (9)$$

$$recall = \frac{OPN}{|C_{found}|} \quad (10)$$

$$OPN = \frac{(a \cap b)^2}{a * b}\{\mathrm{a} \in C_{true}\ , \mathrm{b} \in C_{found}\ \} \quad (11)$$

$C_{true}$ and $C_{found}$ represent real community and discovered community respectively. After defining precision and recall, F1 is

$$F\_score = \frac{2 * precison * recall}{precison + recall} \quad (12)$$

F1 results are shown by comparing baseline algorithms on seven datasets in Fig. 4. Many researchers extended modularity function in measurements of quality of overlapping communities for datasets which have no ground truth. $Q_{ov}$ [29] is one of the mostly used evaluation metrics among the extended modularity Q functions. It is defined as the following:

$$Q_{ov} = \frac{1}{2m}\sum_c \sum_{i,j\in c}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\frac{1}{O_i O_j} \qquad (13)$$

Where, m is the total number of edges in the network. $O_i$ and $O_j$ is the number of communities of the vertex i and j belong to respectively. $k_i$ and $k_j$ is the degree of i and j respectively and $A_{ij}$ is the element of adjacency matrix of the network. This performance is compared with other overlapping community detection algorithms on nine datasets without ground truth.

### 5.3 Implementation

In this section, the proposed system is compared with other overlapping community detection algorithms on different datasets and the running time is also described. These evaluations exist interval [0, 1] and the larger the value of these measurements is, the better the overlapped community's quality is.

According to the ENMI of Fig. 3, the proposed system outperforms the other algorithms in karate, amazon and dblp networks. In political book, the proposed system is slightly better than the others. NILPA performs well on the small and sparsely connected network like risk-map but it isn't partitioned into communities for football network. Thus, only one community is occurred in that network and its measurement is zero. By looking result, it doesn't effectively perform on medium and large networks. ENMI value of OSLOM is greater than all algorithms on football network and GREESE is significantly better than the proposed algorithm on dolphin because ground truth communities of these datasets do not rely on their link and depend on their features or opinion. The football and political book datasets have their node features. For example, ground truth of football network is formed according to conferences belong to teams.

Also in the F1 result of proposed system from Fig. 4, the large networks such as amazon and dblp have slightly better results and all algorithms except LFM have the same result in karate and risk map networks. There are no significantly differences. However, GREESE and OSLOM have better performance on dolphin and football, respectively, like ENMI measurement. As NILPA, F1 evaluation is zero for football network because there are no communities.

In Fig. 5, quality of overlapped community is measured by $Q_{ov}$ with no ground truth and this evaluation is described for several datasets by not considering ground-truth. The quality of proposed system is occurred that it outperforms the other algorithms on all datasets. LFM decreases accuracy on all measurements when compare the baseline methods including proposed algorithm.

In occurrence of overlapping rates of Fig. 6, the proposed system has acceptable overlapping size. That rate is performed on percentage by ratio of number of overlapped nodes of discovered communities and total nodes of graph. With respect to this rate, if excessive overlapped nodes, it produces the hierarchical communities and many highly hierarchical structures are led. Demon has overlapping fraction excessively and occurred isolation nodes. NILPA also has high overlapped rate for other networks except football network. This method cannot explore communities and overlapped nodes for that network. OSLOM and LFM cannot identify overlapped node on football and karate

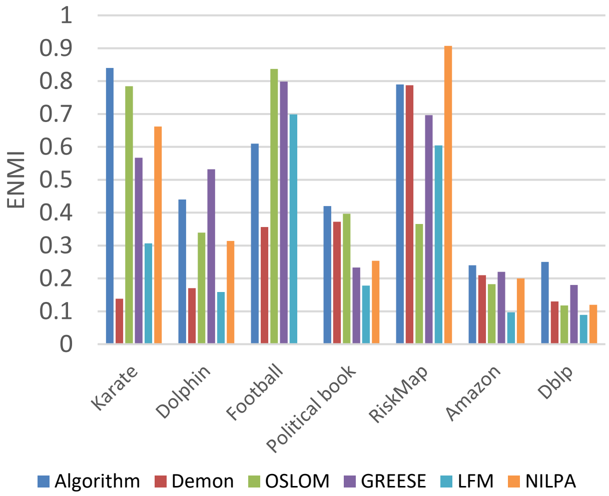

Figure. 3 ENMI result of different algorithms

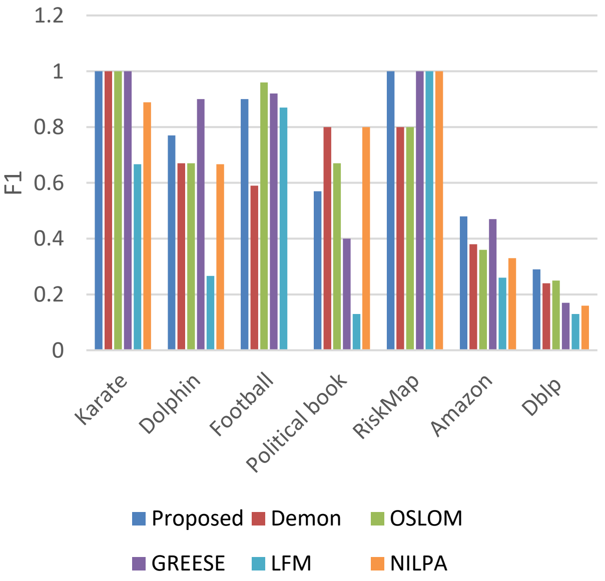

Figure. 4 F1 result of different algorithms

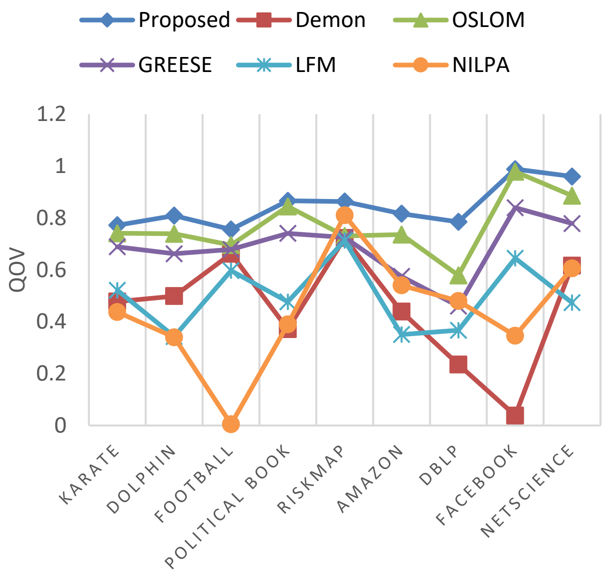

Figure. 5 Comparison of Qov without ground truth

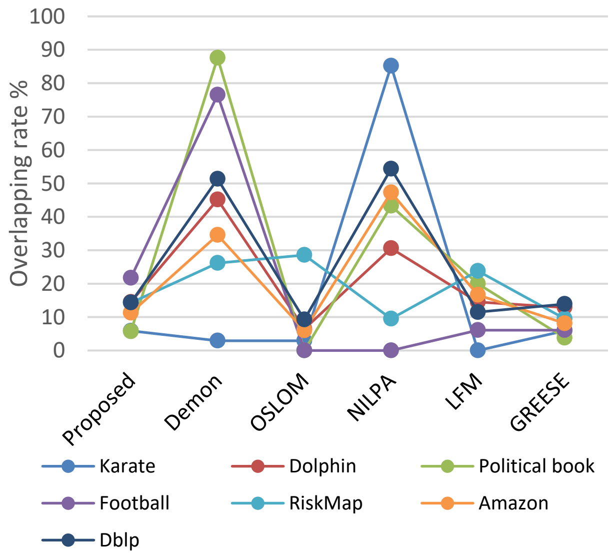

Figure. 6 Overlapping rate of algorithms

network, respectively although they have no highly overlapping ratio on some networks. To control accessing the isolation nodes and excessive overlapped size, GREESE refined the community structure. The proposed system doesn't occur outlier and can detect, in addition to, all nodes from the network to the corresponding communities.

The running time comparison is shown in Fig. 7. In this figure, OSLOM and NILPA aren't illustrated because it takes long time on medium sized datasets about thousand in seconds. Therefore, only running time of four algorithms are described with medium network scale. Although LFM algorithm is occurred

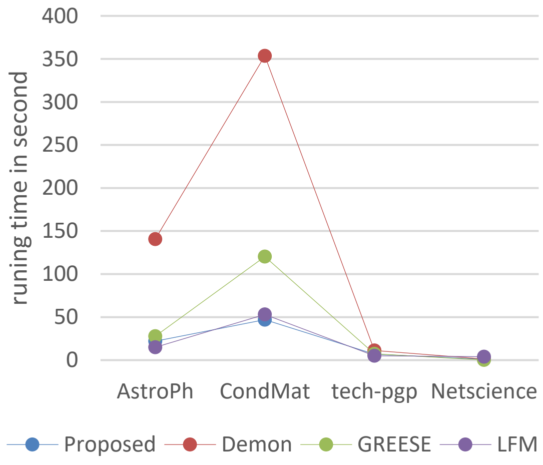

Figure. 7 Comparison of running time (seconds)

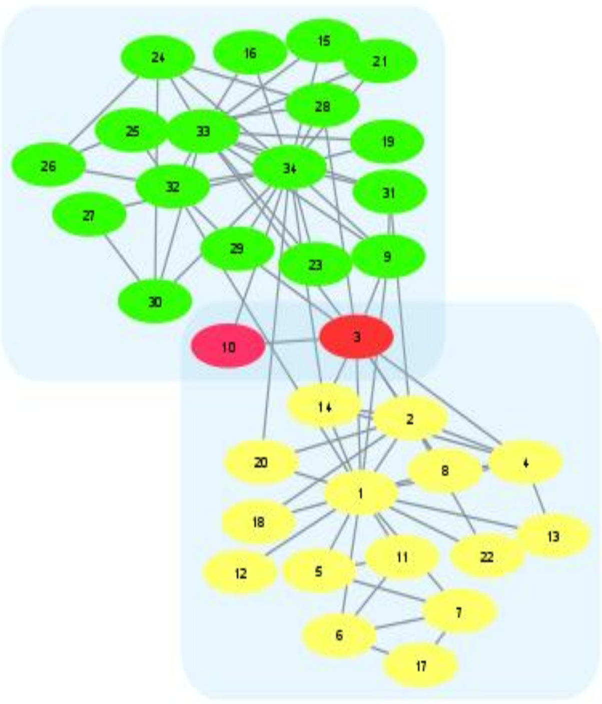

Figure. 8 Overlapped discovery result of proposed algorithm on karate

in lower accuracy, running time fast. Therefore, it has no significantly difference with proposed algorithm on all datasets. When compare all algorithms including OSLOM, the proposed algorithm saves the running time except LFM. DEMON and GREESE run for a while, the larger the data size, the longer the time.

### 5.4 Case study

To further clarify the performance of proposed method, detected overlapped structure from karate dataset is illustrated in Fig. 8. It shows the discovered two communities from proposed algorithm for karate network with green and yellow color. Community 1 consists of {1 2 3 4 5 6 7 8 11 12 13 14 18 20 22 10 17} and community 2 consists of {9 10 15 16 19 21 23 24 27 28 29 30 31 32 33 3 26 25}. Therefore, 3 and 10 are identified as overlapped nodes and described by red node. These two communities which include overlapped nodes are overlapped communities.

The ground truth data are formed with two communities such as {1 2 3 4 5 6 7 8 9 10 11 12 13 14 17 18 20 22} and {9 10 15 16 19 21 23 24 25 26 27 28 29 30 31 32 33 34}. This figure can see the detected two communities of proposed algorithm while GREESE, LFM and NILPA found three communities, four and five communities, respectively. LFM cannot detected overlapped node for this network. GREESE, NILPA discovered one and eleven overlapped nodes, respectively. OSLOM discovered two communities: {3 9 10 15 16 19 21 23 24 25 26 27 28 29 30 31 32 33 34} and {1 2 3 4 5 6 7 8 11 12 13 14 17 18 20 22}. DEMON detected also only two communities: {1 5 6 7 11 17} and 1 2 3 4 8 9 13 14 15 16 18 19 20 21 22 23 24 27 28 29 30 31 32 33 34}.

## 6. Conclusion

This work presents an overlapping community detection algorithm by local community expansion strategy. The most overlapped detection algorithms applied local expansion strategy that occurred instability of community structures because fluctuation of fitness evaluation function they used. The detected community structure depends on parameter which control community's resolution. Therefore, this work provides a parameter evaluation formula by emphasizing on solving the instability of community structure to avoid parameter influences. Instead of global community discovery for overlapped detection to save time consuming, local community expansion algorithm is designed by optimizing f fitness evaluation function based on parameter evaluation formula. Moreover, the extended jaccard similarity is used to locate seed because core node or seed is important to identify center of local community.

Some datasets used in the experiments do not include ground truth information. Therefore, the performance of the proposed algorithm is evaluated by ENMI and F1 for with ground-truth and Qov for without ground-truth datasets. The results show that, the proposed algorithm achieves better accuracy on most datasets and Qov is occurred on all networks as significant improvement and overlapped fraction isn't high. In addition, it saves running time than the others.

In real world, the nature of network structure dynamically changes according to the timestamps. Therefore, evolving communities can be occurred on dynamic networks (e.g. operations are occurred such as communities' growth, contraction, merging, splitting, birth, continue and death). For this changes of community structure, these studies will be done to handle dynamic networks as future work.

## Conflicts of Interest

The authors declare no conflict of interest.

## Author Contributions

The paper conceptualization, methodology, software, validation, formal analysis, investigation, preparation, writing-review and editing, visualization have been done by 1st author. The supervision and project administration have been done by 2nd author.

## References

[1] S. Fortunato, "Community detection in graphs", *Physics Reports*, Vol. 486, pp. 75-174, 2010.

[2] B. W. Kernighan and S. Lin, "An effcient heuristic procedure for partitioning graphs", *Bell Sys. Tech. J.*, pp. 291-308, 1970.

[3] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations", In: *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281-297, 1967.

[4] 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, 2002.

[5] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek, "Cfinder. locating cliques and overlapping modules in biological networks", *Bioinformatics*, Vol. 22, pp. 1021-1023, 2006.

[6] S. Gregory, "Finding overlapping communities in networks by label propagation", *New J. Phys.*, Vol. 12, No. 10, p. 103018, 2010.

[7] J. Xie, B. K. Szymanski, and X. Liu, "Slpa. Uncovering overlapping communities in social networks via a speaker listener interaction dynamic process", In: *Proc. of 2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 344-34, 2011.

[8] M. Coscia, G. Rossetti, F. Giannotti, and D. Pedreschi, "Demon. a local-first discovery method for overlapping communities", In: *Proc.*

*of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 615-623, 2012.

[9] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks", *New Journal of Physics*, *IOP Publishing*, Vol. 11, p. 33015, 2009.

[10] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks", *PloS One*, *Public Library of Science*, Vol. 6, p. e18961, 2011.

[11] J. J. Whang, D. F. Gleich, and I. S. Dhillon, "Overlapping community detection using neighbourhood-inflated seed expansion", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, pp. 1272-1284, 2016.

[12] R. V. Belfin and P. Bródka, "Overlapping community detection using superior seed set selection in social networks", *Computers & Electrical Engineering*, Vol. 70, pp. 1074-1083, 2018.

[13] H. You, X. Zhang, H. Fu, Z. Zhang, M. Li, and X. Fan, "Algorithm of detecting overlapping communities in complex networks", In: *Proc. of IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2014.

[14] Y. Xiaobo, C. Chuxiang, and W. Zhiwan, "Improved LFM algorithm in weighted network based on rand walk", In: *Proc. of 2017 29th Chinese Control and Decision Conference (CCDC)*, pp. 3719-3723, 2017.

[15] X. Chen and J. Li, "Overlapping Community Detection by Node-Weighting", In: *Proc. of the 2nd International Conference on Compute and Data Analysis,* pp. 70-74, 2018.

[16] H. Liu, L. Fen, J. Jian, and L. Chen, "Overlapping community discovery algorithm based on hierarchical agglomerative clustering", *International Journal of Pattern Recognition and Artificial Intelligence*, *World Scientific*, Vol. 32, 2018.

[17] J. Ma and J. Fan, "Local optimization for clique-based overlapping community detection in complex networks", *IEEE Access*, pp. 5091-5103, 2019.

[18] Y. Li, J. He, Y. Wu, and R. Lv, "Overlapping community discovery method based on two expansions of seeds", *Symmetry*, Vol. 13, 2020.

[19] K. Guo, L. He, Y. Chen, W. Guo, and J. Zheng, "A local community detection algorithm based on internal force between nodes", *Applied Intelligence*, pp. 328-340, 2020.

[20] I. B. E. Kouni, W. Karoui, L. B. Romdhane, and Lotfi, "Node importance based label propagation algorithm for overlapping detection in networks", *Expert Systems with Applications*, Vol. 162, 2020.

[21] P. Wang, Y. Huang, F. Tang, H. Liu, and Y. Lu, "Overlapping Community Detection Based on Node Importance and Adjacency Information", *Security and Communication Networks*, Vol. 2021, 2021.

[22] K. Asmi, D. Lotfi, and M. E. Marraki, "A new local algorithm for overlapping community detection based on clustering coefficient and common neighbor similarity", In: *Proc. of the ArabWIC 6th Annual International Conference Research Track*, pp. 1-6, 2019.

[23] K. Asmi, D. Lotfi, and A. Abarda, "The greedy coupled-seeds expansion method for the overlapping community detection in social networks", *Computing*, Vol. 104, pp. 295-313, 2022.

[24] P. M. Kogge, "Jaccard coefficients as a potential graph benchmark", *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW*), pp. 921-928, 2016.

[25] http.//wwwpersonal.umich.edu/~mejn/netdata/

[26] http.//www.orgnet.com

[27] https.//snap.stanford.edu/data/

[28] H. Wu, L. Gao, J. Dong, and X. Yang, "Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein interaction networks", *PloS One*, p. e91856, 2014.

[29] M. Chen and B. K. Szymanski, "Fuzzy overlapping community quality metrics", *Social Network Analysis and Mining*, Vol. 5, 2015.