Overlapped Community Detection using Extended Node Similarity by Local Expansion

Eaint Mon Win
University of Computer Studies,Yangon
eaintmonwin@ucsy.edu.mm

May Aye Khine
University of Computer Studies, Yangon
mayayekhine@ucsy.edu.mm

***Abstract*—The study of real networks like social network has been increasingly interested in community research area. With this study, overlapping community detection plays an important role in studying hidden structure of those networks. There are many overlapping community detection algorithms in recent years. In detecting overlapped structure by local expansion strategy, seeds or core nodes are important because communities are formed on chosen seeds. As a result, inappropriate seeds produce low accuracy of community structure. This paper proposes extended jaccard similarity to find appropriate seed. Firstly, identifies seed using extended jaccard similarity fitness. Then, local communities are detected by extending seed according to quality value. The experimental result obtains improved accuracy and performance of algorithm are compared to other local optimization algorithms.**

Keywords—overlapping community, seed, local expansion, jaccard node similarity

I. INTRODUCTION

Nowadays, there are many networks such as social networks, protein interaction networks, scientists' collaboration networks and citation networks in real society. Networks are modeled as graphs. A node represents individual and link represents connection between individual. [1].As an example, social network is shown in figure 1. Social networks often include communities based on common location or common interest. Citation networks have communities according to their research topic [2].

The study of community structure on this networks has attracted to researchers as hot topic. Community detection means decomposition of network into clusters which consists of one or more nodes. The nodes in the same community are densely connected to each other and links between nodes in different communities are sparsely connected corresponding to widely accepted definition. In real world networks, communities which include some nodes share other communities at the same time not within a community [3]. It is known as overlapped communities and is described in figure 2. Community detection algorithms and overlapped community detection algorithms have been developed for uncovering community structure and overlapped structures by many researchers in recent years.

According to the studies, there are many algorithms for uncovering overlapped structures. Clique Percolation Method (CPM) [1] is proposed by Gergely Palla et al. and it detects overlapped structure by creating a set of many k cliques as communities. However, it can occur high complexity problem and depends on choice of parameter. Steve [4] developed Community Overlapped Propagation Algorithm (COPRA) on

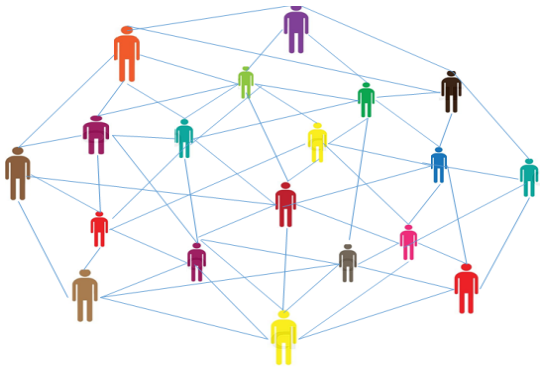

Fig. 1. Social network

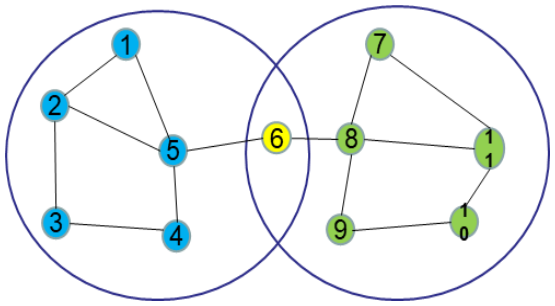

Fig. 2. Overlapped community structure

label propagation idea. It can detect overlapped communities and sets each vertices to unique labels and updates its labels by summing and normalizing the belonging coefficients of vertices in the neighbor set. However, it obtains many meaningless small communities. Bagrow [5] proposed link partition method by decomposing link in original graph. However, above this algorithms require global knowledge of entire network. It is time expensive and space consuming when analyzing large networks. In addition, it is impossible to obtain global information of whole network since network grew rapidly Therefore, local clustering technologies have been studied in overlpped community detection area. It requires local structure information to detect local. This strategy is appropriate for not only static networks but also dynamic networks. In 2011, OSLOM [6] is proposed for detecting overlapped structure based on local optimization of fitness function. The fitness quality is assumed by statistical significant with respect to random flutation. It can perform on large networks not only small networks.

This paper proposes an extended jaccard similarity and overlapping communities are detected. First, identify appropriate seed using this similarity. Then, initial community

is formed by checking fitness quality of neighbor nodes of seed. Second, community is extended by adding neighbor nodes of initial community according to their fitness. If there are no unassigned nodes in any community, overlapped objects are identified by merging obtained local communities.

The rest of paper is organized as follows. In section II, related work is described and basic notations and definitions used in this paper are presented in section III. In section IV, description of algorithm design is explained with the figure and section V gives experimental result to verify performance of algorithm. Finally, conclusion and future work is described as last section.

II. RELATED WORK

The algorithms based on seed expansion generally select a seed to form initial community and then extend it by continuously checking the neighbor nodes according to their fitness quality. The first local discovery method, Democratic Estimate of the Modular Organization of a Network (DEMON) [7] is first discovery method for detecting overlapped communities. It performed on the concept of label propagation approach and overlapped communities are detected by merging local communities according to the threshold. The well-known algorithm, Local Fitness Method (LFM) [8] is proposed by Lancichinetti et al. It identifies overlapping communities on local optimization. However, this algorithm picks a seed node randomly as initial community. Therefore communities lead to instability. Yan Xing et al. [9] proposed an algorithm by partitioning the network into small local communities according to fitness quality to decrease time complexity of global algorithms and then merges these communities to the final overlapping community structure. Xiabo [10] improved LFM algorithm to avoid instability of LFM by using random walk method in selecting seed. Greedy Clique Expansion (GCE) [11] is proposed to perform well on synthetic data. It identifies cliques as seed and then greedily expands seed by optimization local fitness.

Chen et al. [12] calculated the weight of each vertices based on jaccard similarity to identify seed and proposed a fitness quality metric based on node weighting to detect overlapped communities. JIAN [13] modeled novel algorithm on idea that cliques are core of communities to improve search efficiency. This algorithm adopts a single node with maximum density to derive initial community. Then this node is extended by adding k clique rather than single node. Liu [14] designed overlapping community discovery algorithm based on local optimal expansion idea. They proposed a method of calculating the node degree of membership for expansion community. In 2018 [15], a local approach is proposed based on detecting the core node and expansion that nodes. The node with the highest similarity is selected as core node based on similarity between nodes. Then, the expansion of these nodes are considered by using the concept of node's membership based on strong community. However, this algorithm cannot detect overlap and only for local communities. Most of these local expansion or seed expansion algorithms have been proposed on optimization strategies by focusing community's fitness quality in extending community. They haven't been emphasized to achieve appropriate seed.

III. PRELIMINARIES

A. Basic Notation

A network is usually modeled as G= (V, E). V is the set of nodes: V= $\{v_1, v_2 \ldots v_n\}$ and E is the set of edges: E= $\{e_1, e_2 \ldots e_n\}$ in the network. This paper considers undirected and unweighted graphs. The aim of detection of overlapping communities is to find overlapped communities C_1, C_2, $C_3 \ldots C_n$.

B. Jaccard Similarity

The jaccard similarity, also called jaccard index gives a value that represents the similarity of the neighborhoods of two vertices [16]. It is defined as follow:

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

This similarity is defined to measure the similarity between two nodes by intersecting neighborhoods of vertices u and v. N (u), N (v) represents neighbors of vertex u and v, respectively. $N(u) \cap N(v)$ denotes the number of common neighbor nodes of u and v. $N(u) \cup N(v)$ is the number of neighbor nodes which has adjacent at least one vertex u and v. The larger the value, more similar the two nodes.

C. Extended Jaccard Similarity

The jaccard index method considers similarity for all pairs of nodes by emphasizing only neighborhoods of two vertices and ignores links. Therefore, Berahmand considered pairs of nodes which have links by multiplying adjacency matrix.

$$\text{Similarity}(i,j)=A_{ij} * \frac{|N(u)\cap N(v)|}{|N(u)\cup N(v)|} \quad (2)$$

However, it can produce inappropriate seeds when identifies seed based on this similarity. Traditional jaccard can give zero similarity when computes similarity between pairs because that pair of nodes has no similarity if there are no common neighbors between two nodes. Seed means core node with many relations to other nodes. It does not significantly concern if pairs of nodes have common neighbor nodes or not. Therefore, the following similarity measure is defined to avoid zero in this paper.

$$\text{Sim}(u,v)=A_{uv} * \frac{|N(u)\cap N(v)|+1}{|N(u)\cup N(v)|} \quad (3)$$

A_{uv} means adjacency matrix. This similarity can avoid zero similarity between nodes due to impact of common neighborhoods although there are many relations between them.

D. Similarity based Weight

The weight of each node is defined by summing similarity to choose as seed or core node. The node with highest weight is assigned as important node or seed.

$$W_u=\sum_{v\in N(u)}^{k} \text{Sim}(u,v) \quad (4)$$

N (u) represents neighbor set of vertex u. Sim (u, v) is a similarity matrix, which entries are extended similarity between u and v.

E. Quality Evaluation Function

Lancichinetti et al. defined a community fitness function f_G to obtain community identified by maximizing fitness of node. It measures the tightness of internal nodes of the community.

$$f_G = \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^{\propto}} \quad (5)$$

Where $k_{in}^G + k_{out}^G$ represent total internal and external degrees of the nodes of community G. k_{in}^G is internal degree of nodes within a community and k_{out}^G is external nodes' degree of a community. α is a controlling parameter to control communities' size, also known as resolution parameter. Smaller α value, larger the community size is. Default value is 1.The fitness function describes the interior and exterior link of the community. Larger fitness means more obvious community structure.

IV. Algorithm Design

A. Identifying Initial Community

The steps of finding initial community by using extended jaccard similarity are described in the following.

Step1: Similarity is calculated by using equation 3 among nodes from network.

Step2: Computes the weight of each node using equation 4.

Step3: The node with highest weight is selected as seed.

Step4: After identifying seed, forms initial community by adding neighbor nodes of seed.

Step5: Finds fitness values of nodes within initial community to measure quality of community using equation 5.

Step6: If fitness values does not decrease when remove each node within initial community, remove these nodes from community.

Step7: If the value decreases, remains the node in initial community

B. Expanding Community

In this phase, initial community is expanded by adding neighbor nodes according to fitness quality values.

Step1: Computes fitness of neighbor nodes of community using equation 5.

Step2: If the value increases, node with larger fitness is added to the community to identify local community.

Step3: Go to step3 of identifying initial community phase for selecting the seed with highest weight among unassigned nodes.

After that, next local community is obtained. The process continues until all nodes in network are assigned in corresponding communities.

Finally, local communities are merged to identify overlapped nodes. In figure 3, the flow of algorithm is designed on local expansion strategy by selecting appropriate seed.

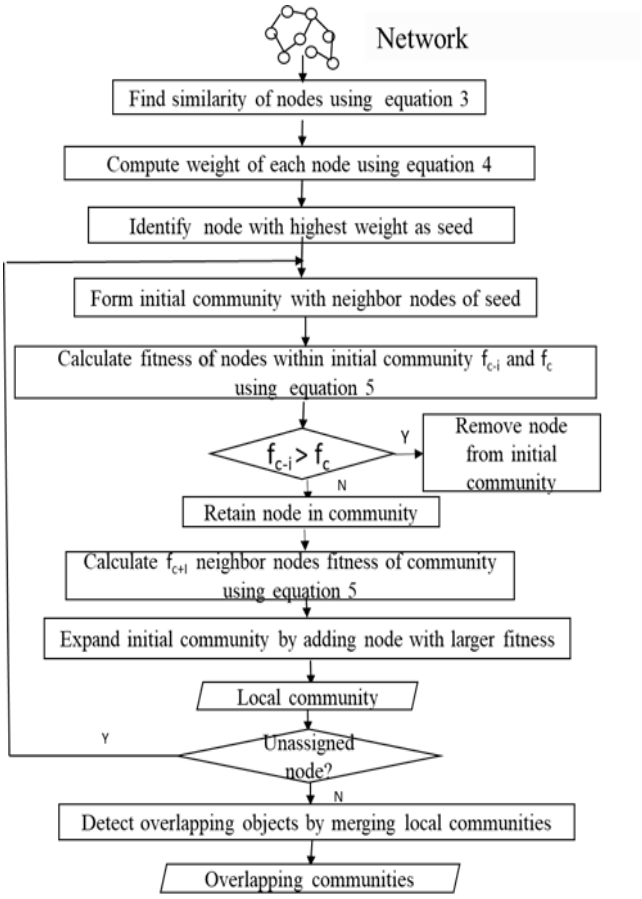

Fig. 3. Proposed algorithm

V. Experimental Result

In experiment, algorithm is implemented by extended normalized mutual information (ENMI) to measure accuracy of algorithm and compare the performance with Order Static Local Optimization Method (OSLOM), LFM, DEMON on different datasets. This paper uses Zachary Karate Club, Dolphin and Political books and American Football networks. These datasets are mostly used and well-known network datasets. The number of nodes, edges and ground truth communities are described in table 1. In this paper, parameter α is set between 0.7 and 1.5.

A. Evaluation Metric

The normalized mutual information is widely used to measure the quality of disjoint community detection when the ground truth is known, where it originated from information theory. It ranges from 0 to 1 by normalization, and a higher value represents a better quality. Extended normalized mutual information (ENMI) [17] is proposed by Lancichinettiet to evaluate the quality of overlapping community detection. It is suggested that only a small amount of additional information is needed to infer one partition from the other if the two partitions are similar. It is defined as follows:

$$ENMI(X|Y) = 1 - \frac{1}{2}[H(X|Y) + H(Y|X)] \quad (6)$$

X and Y are random variables related to partitions C and C', respectively, and H (X|Y) means the normalized conditional entropy for cluster X with respect to cluster Y.

TABLE I. DESCRIPTION OF REAL DATASET

Dataset	Node	Edge	Average degree	Number of cluster
Karate	34	78	4	2
Dolphin	62	159	5	2
Polbooks	105	441	8	3
Football	115	613	10	12

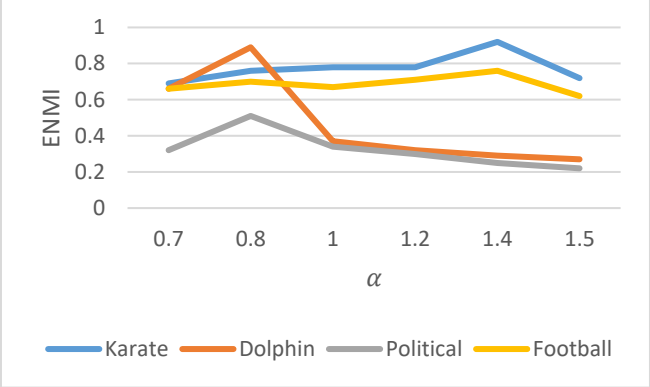

Fig. 4. ENMI of algorithm on different parameters

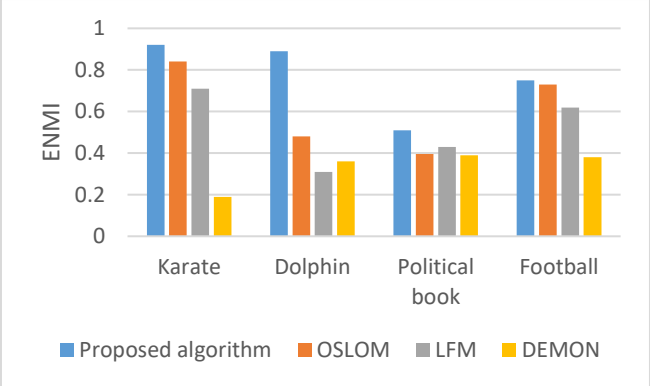

Fig. 5. Performance result of four algorithms on different datasets

B. Evaluation Performance on Real World Networks

Zachary Karate Club [18] network is well known social network. The data was collected from the members of a karate club in an American University. Each node represents a member of the club, and each edge represents a tie between two members of the club. According to the experiment, there are two communities structure are obtained at α =1 but the overlapping nodes are too much occurred at that value. However, overlapping nodes are not significantly different even α reaches 1.2. Then, high ENMI value is occurred in parameter value 1.4 on this network. Moreover, algorithm outperforms when compare accuracy of other algorithms such as LFM, OSLOM, DEMON which are 0.71, 0.84 and 0.19, respectively. The performance of algorithm is shown in figure 4 corresponding to different parameter and performance comparison of four algorithms is described in figure 5 according to different networks.

The next dataset, Dolphin network [19], is a directed social network of bottlenose dolphins. The nodes are the bottlenose dolphins (genus *Tursiops*) of a bottlenose dolphin community living off Doubtful Sound). An edge indicates a frequent association. In this dataset, high accuracy is occurred at α=0.8 and three overlapping nodes are uncovered but many communities are detected at α=1 as an inaccurate result. Therefore, accuracy can decrease if value is high due to many numbers of communities are produced (i.e. Many communities are composed of small number of nodes). According to the experiment, accuracy is better than other three algorithms on this network.

The other miscellaneous network is Polbooks used in the paper [20]. It is a network about US politics published around 2004 presidential election. Nodes are books about US politics sold by the online bookseller Amazom.com. Edges 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. The average degree of this network is large but network density is lower than density of karate social network. As the result, many overlapping nodes and number of communities are too much occurred at α =1 and above due to communities' size is small. However, improved accuracy is occurred at small parameter value 0.8 than large value. Moreover, ENMI of algorithms is slightly different with each other.

The final tested dataset is College football network. It is a network of American football games between Division IA colleges during regular season fall 2000. Each node represent football team which belongs to specific conference and links are games between teams when two teams play each other. Almost all of the nodes are same degrees and densely connected to each other within a community when identify communities. In the evaluation result, many overlapping communities are occurred and community structures are almost the same at α =0.8 and 1 but ENMI gradually increase when reaches 1.2 and 1.4. When compare OSLOM, it does not obviously improve accuracy on this dataset. But it is a little better than OSLOM and significantly different with other algorithms.

According to the test results, overlapped nodes are too much found in dolphin and political book networks at α=0.8. If α<0.8, there are only one community for dolphin and only two communities for political book. For karate and football network, many small communities are uncovered if α>1.4. Therefore, α value is set between 0.7 and 1.5 in experiment. In contrast, found that large value yields large amount of small communities and small value yields small amount of large communities.

VI. CONCLUSION

This paper proposes an extended jaccard similarity to find the appropriate seeds in identifying initial community. Local expansion algorithms in detecting overlapped communities have been proposed by many researchers. However, they emphasize on community expansion and ignore selecting seeds to obtain good seed that improve the accuracy of algorithm. There are some algorithms to identify seed by using seed selection methods but they can only detect community structure and cannot detect overlapped structure. Therefore, uncovering overlapping community structure using extended jaccard similarity is described in this paper. In experimental results shown on real world datasets, proposed algorithm can accurately detect communities than other algorithms.

As future research, algorithm will be intended to apply on dynamic networks due to dynamic nature of real world networks. Moreover, parameter α in fitness function will be

contributed to control community's size and to avoid multiple implementation.

REFERENCES

[1] S. Fortunato, "Community detection in graphs",Physics reports, pp: 75-174, 2010, Elsevier

[2] V. Chandrashekar, "A Framework for Community Detection from Social Media", 2013 ;International Institute of Information Technology Hyderabad.

[3] A. Karataş and S Şahin, "Application areas of community detection: A review", 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), pp: 65-70, 2018, IEEE

[4] G. Steve, Finding overlapping communities in networks by label propagation. New Journal of Physics12, 2010

[5] A.Yong-Yeol, Bagrow, P.James and S.Lehmann, "Link communities reveal multiscale complexity in networks nature", pp:761-764, 2010, Nature Publishing Group

[6] A. Lan cichinetti, F. Radicchi, Ramasco and S.Fortunato, "Finding statistically significant communities in networks" PloS one, 2011, Public Library of Science

[7] M. Coscia, G.Rossetti, F. Giannotti, D. Pedreschi, "Demon: a local-first discovery method for overlapping communities", Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp:615-623, 2012

[8] A. Lancichinetti; S. Fortunato; J. Kertész; "Detecting the overlapping and hierarchical community structure in complex networks", New journal of physics, 2009, IOP Publishing

[9] Y.Xing, F.Meng, Y.Zhou, R.Zhou and Z.Wang, "Overlapping Community Detection by Local Community Expansion", J. Inf. Sci. Eng, pp:1213-1232, 2015

[10] Y. Xiaobo, C. Chuxiang and W. Zhiwang, "Improved LFM algorithm in weighted network based on rand walk", pp:3719-3723, 2017, IEEE

[11] C. Lee, F. Reid, A. McDaid and N. Hurley, "Detecting highly overlapping community structure by greedy clique expansion", arXiv preprint arXiv:1002.1827,2010

[12] X. Chen and J. Li, "Overlapping Community Detection by Node-Weighting", ICCDA 2018, DeKalb, IL, USA. © 2018 Association for Computing Machinery

[13] M. Jian and F. Jianping, "Local Optimization for Clique-based Overlapping Community Detection in Complex Networks", 2019, IEEE

[14] L. Hongtao, F. Linghu, J. Jie and L.Chen, "Overlapping community discovery algorithm based on hierarchical agglomerative clustering", International Journal of Pattern Recognition and Artificial Intelligence, 2018, World Scientific

[15] K. Berahmand, A. Bouyer and M. Vasighi, "Community detection in complex networks by detecting and expanding core nodes through extended local similarity of nodes" pp:1021-1033, 2018, IEEE

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

[17] X. Jierui, S. Kelley and B.K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study", 2013, ACM New York, NY, USA

[18] W. W. Zachary, "An information flow model for conflict and fission in small groups," J. Anthropol. Res, pp. 452–473, 1977.

[19] D. Lusseau, "The emergent properties of a dolphin social network," Proc. Roy. Soc. London B, Biol. Sci, pp. S186–S188, 2003

[20] Newman, E.J Mark and M. Girvan, "Finding and evaluating community structure in networks", Physical review E, pp:26113, 2004, APS