Community and Outliers Detection in Social Network

Htwe Nu Win[1(✉)] and Khin Thidar Lynn[2]

[1] University of Computer Studies, Mandalay, Mandalay, Myanmar
htwenuwin99@gmail.com
[2] Faculty of Information Science,
University of Computer Studies, Mandalay,
Mandalay, Myanmar
lynnthidar@gmail.com

Abstract. Challenges of detecting communities among users' interactions play the popular role for days of Social Network. The previous authors proposed for detecting communities in different point of view. However, similarity based on edge structure and nodes which cannot group into communities are still motivating. Considering the community detection is motivating from the similarity measurement to detect significant communities which are high tightly connected each other upon the edge structure and outliers which are unnecessary to group into the communities. This paper is proposed the approach of using similarity measure based on neighborhood overlapping of nodes to organize communities and to identify outliers which cannot be grouped into any of the communities based on Edge Structure. The result implies the best quality with modularity measurement which leads to more accurate communities as well as improved their density after removing outliers in the network structure.

Keywords: Social network · Community · Outlier

1 Introduction

Social networks are naturally modeled as graphs, which we sometimes represent it as a social graph. The entities are nodes, and an edge that connects two nodes if the nodes are interacted by the relationship that forms the network. For Facebook friendship graph which we will use in our work, social graphs are undirected and unweighted. The communities of complex networks are groups of nodes which are high tightly connected with nodes of the same group other than with less links connected with nodes of different groups [18]. Communities may be groups of friendship in social networks, sets of web pages concerning with the same topic and groups of cells with similar functions. While identifying the communities in graphs, nodes which cannot groups with any communities and need not be necessary group, will be identified as outliers. The early observer of outlier is Hawkins [2]. He said that "An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". The groups of vertices which are similar to each other is naturally assumed as communities. The similarity between each pair of vertices can be

T. T. Zin and J. C.-W. Lin (Eds.): ICBDL 2018, AISC 744, pp. 58–67, 2019.
https://doi.org/10.1007/978-981-13-0869-7_7

computed with respect to some reference property, local or global, no matter whether they are connected by an edge or not. Vertices which have similarity value with the communities can be group into the corresponding communities. Our proposed approach is adopted from Hawkin's definition, nodes which have no any friendship or there are no common friends, so much deviate and saturated with individual lonely is defined outlier and community based on similarity measure on Edge structure.

This paper explores the use of neighborhood overlapping by using vertex similarity method for outlier and significant community detection. The heart of this approach is to represent the underlying dataset as an undirected graph, where a user refers to each node and friendship between two users represents each edge. Before we measure the similarity among neighborhood overlap, finding seed node by using the degree centrality is necessary which is designed to find nodes that are most "central" to the graph. We operate similarity from the most centrality node and its neighborhood nodes. The values of zero similarity are then used to identify as outliers.

To illustrate, consider graph of Fig. 1, which consists of 18 nodes and 25 edges. Upon applying approach method, three communities are represented by yellow, green and dark orange color. Continually, red color nodes are shown as outliers. It can be seen the significant communities and outliers in this toy example.

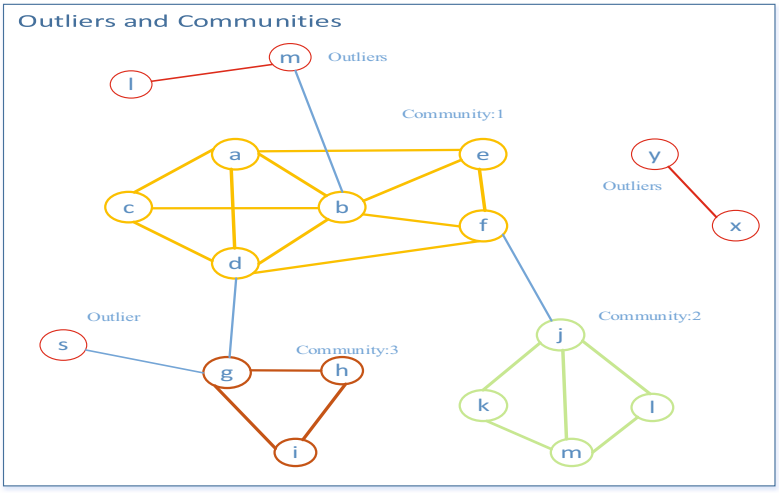

Fig. 1. Example of communities and outliers.

The rest of paper is organized as follows. Section 2 briefly surveys related work. In Sect. 3, we describe the background methodology of our work. And then, we briefly describe about our propose system in Sect. 4. In Sect. 5, we discuss about the experiment and evaluation of our work. Then, we concluded our work in Sect. 6 and talk about our future idea.

2 Related Work

Many approaches of outliers detection algorithm have been proved over years. Each trend has efficient and effective in their ways. After the time of Hawkins [2] definition, many different points of view in outliers were appeared. Graph based outlier detection were also appeared with the flow of researchers. SimRank method [3] identified the proximity measure of graph that quantified the closeness of two nodes in the graph, which based on their relationships with other objects and computed the similarity of the structural context in which the graph objects occur. Several variants of SimRank were also proposed by [5, 17]. On the other hand, [8] proposed the one could use to compute various measures associated with the nodes in the given graph structure, dyads, triads, egonets, communities, as well as the global graph structure. [9] proposed OddBall which is an algorithm to detect anomalous nodes in weighted graphs. Detecting anomalous sub-graphs using variants of the *Minimum Description Length* (MDL) principle was proposed by [1]. Outrank method [4] also used the MDL principle as well as other probabilistic measures to detect several types of anomalies (e.g. unexpected/missing nodes/edges). OutRank and LOADED [13] used similarity graphs of objects to detect outliers. MDL to spot anomalous edges was used by [6]. [7] used proximity and random walks, to assess the normality of nodes in bipartite graphs. [14] proposed the method of detecting outliers with community by using minimum valid size (mvs). If the minimum size is 2, it will be chosen the single node and marked it as outlier. In contrast to the above, we work with *undirected* and unweighted graph data with method of overlapping neighborhood. We explicitly focus on edge structure to detect outliers and significant communities with nodes similarity.

3 Background Theory

3.1 Graph

We consider basic definitions of a given network represented as a directed graph or undirected graph. Facebook friendship network which we will use in our work here, a graph may be undirected, meaning that there is no distinction between the two vertices associated with each edge.

The notation of a graph $G = (V,E)$ consists of two sets V and E. The elements of $V = \{v_1, v_2, \ldots, v_N\}$ are the nodes or vertices of the graph G where each vertex $v_i$is associated with the instance xi from the input data X and the cardinality of $|V|$ is N. The elements of $E = \{e_1, e_2, \ldots, e_M\}$ are links or edges between nodes and the cardinality of $|E|$ is M. An edge connecting the vertices v_i and v_j is denoted by e_{ij} [15]. The example of social network imagine as a graph is shown in Fig. 1.

3.2 Node Degree and Its Neighborhood

In network G, the degree of any node i is the number of nodes adjacent to i. The degree of node v is $d(v)$, that is, the number of edges associated with node v.Generally, the more degree that the node has, the more important it will be [10].

Two vertices v and u are called neighbors, if they are connected by an edge. Let Γ_i be the neighborhood of vertex i in a graph, i.e., the set of vertices that are directly connected to i via an edge. For a given node u, $N(u) = \{v|(u, v) \in E\}$ is a set of containing all neighbors of node u [10].

3.3 Communities

The communities of social network are groups of nodes, with more links connecting nodes of the same group and comparatively less links connecting nodes of different groups. Communities may be groups of related individuals in social networks. Identifying communities in a network can provide valuable information about the structural properties of the network, the interactions among nodes in the communities, and the role of the nodes in each community [18].

In an undirected graph $G(V, E)$, where the total number of node, $|V| = n$ and total number of edges,$|E| = m$ are defined. We can identify set of communities such that $Coms = \{V1', V2', \ldots\ldots, Vcn'\}$ where $\cup_{i=1}^{cn} V_i' \subseteq V_i$ and cn is the total number of communities Coms should satisfy, $Vi' \cap Vj' = \phi$[14].

3.4 Vertex Centrality

The importance of a node is determined by the number of nodes adjacent to it. The larger the degree of node, the more important the node is. Those high-degree nodes naturally have more impact are considered to be more important. The degree centrality is defined as

$$C_D(v_i) = d_i = \sum_j A_{ij} \tag{1}$$

When one needs to compare two nodes in different networks, a normalized degree centrality should be used,

$$C_D'(v_i) = d_i/(n-1). \tag{2}$$

Here, n is the number of nodes in a network. It is the proportion of nodes that are adjacent to node v_i[10].

3.5 Outliers

According to Hawkins [2], outliers can be defined as follows: "An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". Most outlier detection schemes adopt Hawkin's definition of outliers and thus assume in our system is that outliers are nodes which have zero values of similarity measure in graph. Each node in a graph cannot be grouped into any of the communities is called outlier. As such, these outliers can be easily detected by existing distance or density based algorithms. However, in this paper we focus on outliers that might be concentrated in certain regions. We take edge structure based approach in graph to solve this problem. Outliers and communities can be defined by:

$$Outs = \{v | v \in V, \neg \exists Vi' \in Coms \wedge v \in Vi'\} = V - \bigcup_{i=1}^{cn} Vi' \tag{3}$$

Outliers directly identified by getting nodes from small communities [14, 19]:

$$Coms \bigcup Outs = v$$

$$Coms \bigcap Outs = 0$$

Where, *Coms* is the community and *Outs* is indicated as outlier.

3.6 Vertex Similarity

It can be assumed that communities are groups of vertices similar to each other. We can compute the similarity between each pair of vertices after searching seed nodes. Most existing similarity method are based on the measurement of distance called Euclidean, Manhattan and etc., Although, to considered the similarity between selected node and is neighborhood, Jaccard Similarity is more convenient in this work which we will measure the similarity based on the overlapped neighbor of seed nodes. Let N_i denote the neighbors of node v_i. Given a link (v_i, v_j), the neighborhood overlap is defined as

$$\begin{aligned} Overlap(vi, vj) &= \frac{number\ of\ shared\ friends\ of\ both\ v_i\ and\ v_j}{number\ of\ friends\ who\ are\ adjacent\ to\ atleast\ v_i\ and\ v_j} \\ &= \frac{\left|N_i \cap N_j\right|}{\left|N_i \cup N_j\right| - 2} \end{aligned} \tag{4}$$

We have -2 in the denominator just to exclude $v_i\ and\ v_j$ from the set $N_i \cup N_j$. If there are no overlap vertices in any two $N_i and N_j$ means $\left|N_i \cap N_j\right| = \emptyset$, it can be identified N_j as outliers of N_i. Assuming like that, our work identify outliers are appeared with among separated communities [10].

4 Proposed Approach

Network can be represented as graph for detecting outliers and significant communities. In this paper, we use concept of undirected graph in which two vertices have common node, if they share one or more of the same vertices. Let Γ_i be the neighborhood of vertex i in a network, i.e., the set of vertices that are directly connected to vi via an edge. Then the number of common friends of vi and vj is $|\Gamma_i \cap \Gamma_i|$. Contrary, nodes which have no any other node in common or sometime they can be isolated, we can identify them as outliers, can be given as $|\Gamma_i \cap \Gamma_i| = 0$. In our work, we propose the system to identify significant community with detecting outlier by using node similarity depend on this concept. Node similarity measurement can be considered in detecting communities, groups of vertices are similar to each other. The detected communities are defined by the computation of the similarity values of vertices and the

corresponding seed node. This paper used the most common used method in considering of the similarity between selected node and its neighborhood is Jaccard Similarity. It is more convenient in this work which will be measured the similarity based on the neighborhood overlap of seed nodes. Let N_i denote the neighbors of node v_i.

Our work is based on the following intuitive properties:

- Outliers are defined by nodes which have no any intersect value to its related ones.
- Seed node is determined by degree centrality method which has the largest degree.
- Communities are calculated by Vertex Similarity among nodes that are linked to seed node.
- Nodes that belong to the same communities are likely to be more links connected to each other.
- Each community is likely to be fewer links outside the rest of the graph.
- Nodes in the same community are likely to share common neighborhood node.

The heart of this work has three main processes, firstly finding the intersect value of related node to detect outliers, then use the method of the degree centrality to determine seed nodes and finally use the method of neighborhood overlap based on vertex Similarity for detecting the communities.

5 Experiments

5.1 Description of Datasets

In this paper, a real undirected network, the popular studies in social network analysis, Zachary Karate Club Dataset [20] is used. In this Dataset statistics, "nodes" represents the number of friends; "Edges" represents the number of friendship in the network. There are 34 member nodes and 78 edges as shown in Fig. 2.

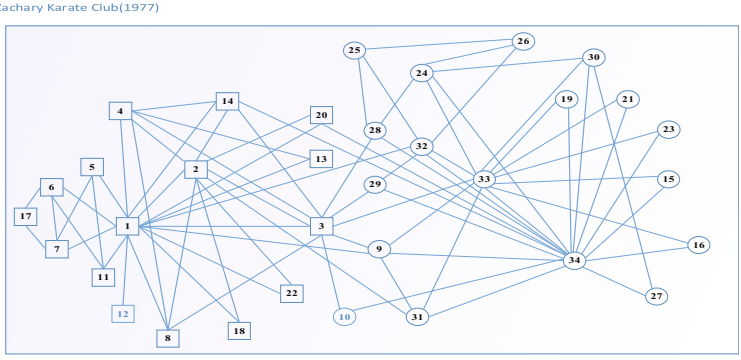

Fig. 2. Original network of Zachary Karate Club.

5.2 Evaluation Method

Generally, in thinking about the evaluation method of how good of a community, a set of nodes which have no ground truth community in undirected and unweighted graph, there are two criteria of interest. The first is the number of edges between the members of the cluster, and the second is the number of edges between the members of the cluster and the remainder of the network. The two groups are Multi-criterion scores and Single-criterion scores. Firstly, it is represented as Multi-criterion scores, combines both criteria (number of edges inside and the number of edges crossing) into a single objective function; the next criterion is the objective functions employs only a single of the two criteria (*e.g.*, volume of the cluster or the number of edges cut).

Multi-criterion Scores. Let $G(V, E)$ be an undirected graph with $n = |V|$ nodes and $m = |E|$ edges. Let S be the set of nodes in the cluster, where n_s is the number of nodes in S, $n_s = |S|$; m_s is the number of edges in S, $m_s = |\{(u, v) : u \in S, v \in S\}|$; and c_s, the number of edges on the boundary of, $c_s = |\{(u, v) : u \in S, v \notin S\}|$; and $d(u)$ is the degree of u.

It is considered the metrics $f(S)$ that capture the notion of a quality of the cluster. Lower value of score $f(S)$ signifies a more community-like set of nodes [16].

- Conductance: $f(S) = \frac{c_s}{2m_s + c_s}$ measures the fraction of total edge volume that points outside the community. The smaller the value of Conductance is, the better the community quality is [16].
- Expansion: $f(S) = \frac{c_s}{n_s}$ measures the number of edges per node that point outside the community. The smaller the value of Expansion, the better the community quality is [16].

Single-criterion Scores. Next it is also considered community scores that is a single criteria. Here it is considered the following two notions of a quality of the community that are based on using one or the other of the two criteria of the previous subsection:

- Volume: $\sum_{u \in S} d(u)$ is sum of degrees of nodes in S. The larger the value of this metric is, the better the community quality is [16].
- Edges cut: c_s is number of edges needed to be removed to disconnect nodes in S from the rest of the network. The smaller it is, the better the community quality is [16].

5.3 Results and Evaluation

The communities and outliers result of our evaluation are shown in Figs. 3, 4 and 5. Firstly, the outliers are defined with the zero value by calculating the method of intersection between their corresponding nodes which are based on the edge structure. In this paper, there are two outliers (node 10 and node 12) that had no friendship of common neighbors between different nodes. Then, seed nodes were determined by using vertex centrality. After determining seed node, the similarity measurement was processed the computation of overlapped value of common node between the correspondent seed node and its neighborhood. As shown in Fig. 3, there are two detected communities and two outliers by using the proposed approach.

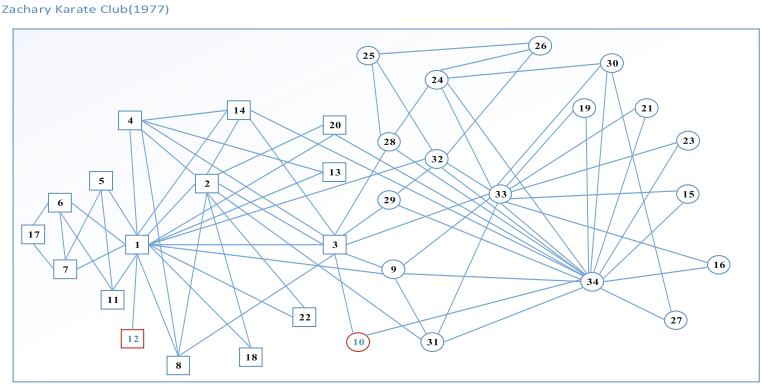

Fig. 3. Two Communities and Outliers by using proposed approach.

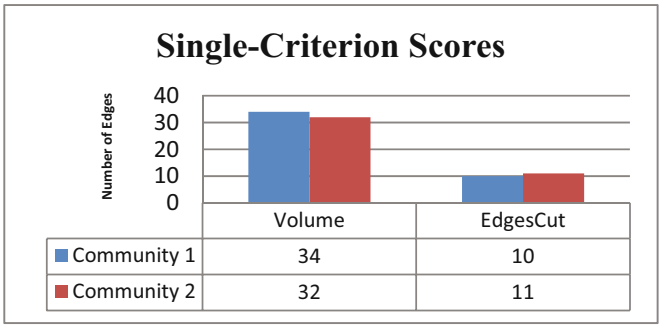

Fig. 4. Single-criterion scores

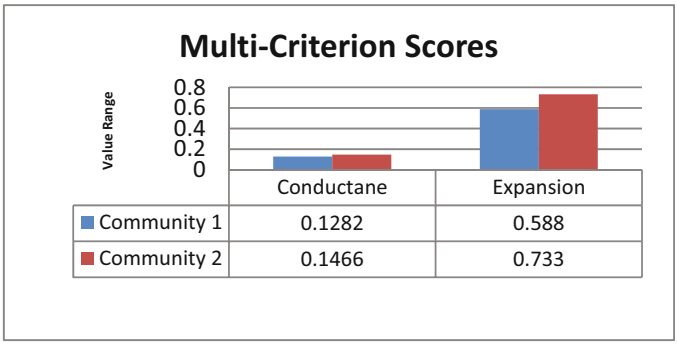

Fig. 5. Multi-criterion scores

This paper is measured by using two different quality of community as described in Sect. 5.2. Figure 4 showed the significant result based on the first two connections. In the last connectivity, the amount of the expansion value is very few in the expected number of links between the communities, therefore, it can be seen that the smaller conductance and expansion values are significant as shown in Fig. 5. Therefore, it can be decided that considering detecting outlier is effective to get the significant community in social network.

6 Conclusion

This paper proposed the approach to identify outlier and detect the community of social networks using the overlapping of neighborhoods with the method of vertex similarity. It described the detail computation to detect outliers with no common node and community with common node. The previous methods had been considered the structure of community without outlier. Therefore, this approach proposed outlier detecting method by using edge structure to get the effective community. In this paper, we based on edge structure of unweighted and undirected graph to detect the outlier and significant communities without thinking any information of the structure.

We used the standard real network dataset from Zachary Karate club [20]. It was used two quality measurement, single-criterion score and multi-criterion scores [16] to define how much our community is good and significant. By using the method of determining outlier with detecting community the density of our method's result is significant in measurement. The drawback of our method is only based on edge structure without considered any information such as profile of user. So, our future will be used feature of users to detect effective community.

References

1. Noble, C.C., Cook, D.J.: Graph-based anomaly detection. In: KDD, 2003, pp. 631–636 (2003)
2. Hawkins, D.: Identification of Outliers. Chapman and Hall, London (1980)
3. Jeh, G., Widom, J.: SimRank: a measure of structural-context similarity. In: Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Edmonton, Alberta, pp. 538–543 (2002)
4. Moonesinghe, H.D.K., Tan, P.N.: Outrank: a graph-based outlier detection framework using random walk. Int. J. Artif. Intell. Tools **17**(1), 19–36 (2008)
5. Chen, H.-H., Giles, C.L.: ASCOS: an asymmetric network structure context similarity measure. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Niagara Falls, Canada (2013)
6. Chen, J.Y., Zaiane, O.R., Goebel, R.:. Detecting communities in large networks by iterative local expansion. In: International Conference on Computational Aspects of Social Networks (2009)
7. Sun, J., Qu, H., Chakrabarti, D., Faloutsos, C.: Neighborhood formation and anomaly detection in bipartite graph. In: ICDM, 27–30 November 2005

8. Henderson, K., Eliassi-Rad, T., Faloutsos, C., Akoglu, L., Li, L., Maruhashi, K., Prakash, B.A., Tong, H.: Metricforensics: a multi-level approach for mining volatile graphs. In: Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington, DC, pp. 163–172 (2010)
9. Akoglu, L., McGlohon, M., Faloutsos, C.: Anomaly detection in large graphs. In: CMU-CS-09-173, November 2009
10. Tang, L., Liu, H.: Community Detection and Mining in Social Media. A Publication in the Morgan & Claypool Publishers series (2010). ISBN 9781608453559
11. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**(23), 8577–8582 (2006)
12. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
13. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. U.S.A. **99**, 7821–7826 (2002)
14. Wang, M., Wang, C., Yu, J.X., Zhang, J.: Community detection in social networks: an indepth benchmarking study with a procedure oriented framework. In: 41st International Conference on Very Large Data Bases, Proceedings of the VLDB Endowment, 31 August–4 September 2015, Kohala Coast, Hawaii, vol. 8, no. 10 (2015). Copyright 2015 VLDB Endowment 21508097/15/06
15. Plantié, M., Crampes, M.: Survey on social community detection. In: Social Media Retrieval, Computer Communications and Networks, pp. 65–85. Springer (2013). ISBN 978-1-4471-4554-7
16. Chen, M., Nguyen, T., Szymanski, B.K.: On measuring the quality of a network community structure. In: Proceedings of the IEEE Social Computing Conference, Washington DC, 8–14 September, pp. 122–127 (2013)
17. Zhao, P., Han, J., Sun, Y.: P-rank: a comprehensive structural similarity measure over information networks. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, China, pp. 553–562. ACM (2009)
18. Fortunato, S.: Phys. Rep. **486**, 75 (2010)
19. Eberle, W., Holder, L.: Discovering structural anomalies in graph-based data. In: ICDM Workshops, pp. 393–398 (2007)
20. Zachary, W.W.: An information how model for conflict and fission in small groups. J. Anthropol. Res. **33**, 452–473 (1977)