Efficient Calculation of Structural Similarity Threshold for the SCAN Network Clustering Algorithm

Vincent Yip
Computer Information Systems
Umpqua Community College
Roseburg, Oregon, USA
Vincent.Yip@umpqua.edu

Sinan Kockara
Department of Computer Science
University of Central Arkansas
Conway, Arkansas, USA
skockara@uca.edu

Chenyi Hu
Department of Computer Science
University of Central Arkansas
Conway, Arkansas, USA
chu@uca.edu

***Abstract*— Community detection algorithms play an important role in discovering knowledge in networks. The Structural Clustering Algorithm for Network (SCAN) is a community detection algorithm which is capable of detecting hubs and outliers, in addition to cluster members. The term hub means node with the ability of collecting and delivering information among clusters while outlier is considered as a noise in the data. Currently, researchers use exhaustive search to determine the structural similarity threshold value (ε) in the SCAN. This paper reports a new approach of using interval ε value to narrow the searching domain for proper ε value for the SCAN. The approach first adopts computational results produced by the Fast Modularity and the Walktrap algorithms to bind the number of clusters of a network and then determine the interval for ε value. For each of our test datasets, the interval prediction reliably finds the true number of clusters. More importantly, the proposed prediction method helps users to eliminate an average of 67.7% of inappropriate ε values used to generate clusters.**

I. Introduction

Networks (or graphs) offer a powerful means of modeling data in different domains ranging from bioinformatics [1][20] to social networks[2]. These networks are commonly used to represent and model real world objects and their relationships. For instance, in a social network, each node describes a person while the edges show the relationships between two members. Community detection (clustering) in complex networks is one of the key interests for pattern recognition. The basic community detection includes properly arranging a network structure by visual inspection. This method is intuitive and can only handle small networks. Thus, automatic community detection techniques are developed over years. Moreover, due to the increase in complexity of data analysis processes, the choices of parameters for clustering algorithms are another crucial point in this analysis process. Without proper selection of parameters, important patterns will be overlooked.

A number of network clustering approaches have been proposed in the literature including the modularity based algorithms [1][2][3], min-max cut (CNM) [4], normalized cut [5], and the Structural Clustering Algorithm for Networks (SCAN) [6] to name a few. Use of these algorithms in biological networks such as the protein-protein interaction networks is quite common. These provide capturing of the biologically meaningful interactions which otherwise is nontrivial. For instance, in a recent study [20], researchers compared SCAN against CNM network clustering methods on the budding yeast (Saccharomyces cerevisiae) protein-protein interaction network.

It is important to notice that SCAN is the only algorithm which is capable of differentiating the role of nodes in a network as hubs or outliers (Figure 1). A hub does not belong to any cluster, while connecting different clusters. An outlier consists of a weak affiliation to a cluster. Therefore, hub plays a crucial role in community detection. However, locating the correct structural similarity (or cosine similarity) threshold (ε) value for the SCAN is nontrivial. Currently, researchers apply exhaustive search to determine the threshold value ε for the SCAN. SCAN was originated from the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [7]. Thus, similar to DBSCAN, SCAN requires a parameter (ε) to qualify each edge to be member of a cluster in a network. Since ε parameter is dataset dependent, different networks have different ε values in order to be clustered correctly.

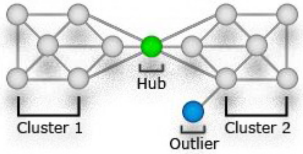


Figure 1. A Network consists of a Hub, an Outlier, and 2 Clusters.

In this study, in order to easily obtain the proper threshold value ε for the SCAN, an interval ε value is used to narrow the searching domain. Our approach first adopts computational results produced by community detection algorithms to bind the number of clusters of a network and then determines the interval ε value.

II. Background

Newman and Girvan [2] are one of the leading pioneers to tackle the automatic community detection problem. They suggested using modularity to qualify the intensity of community structures. The algorithm assumes that members in the same community should be more firmly connected than they would be randomly. To hierarchically divide a given graph into communities, edges with the largest betweenness number of shortest paths passing through the edge are eliminated one after another. This approach has been used for different applications including community structure validation, and as a main function for optimization algorithms to detect communities. Thus, modularity rapidly

978-0-7695-4574-5/11 $26.00 © 2011 IEEE
DOI 10.1109/BIBM.2011.49

became an effective method in the discovery of a community structure [8]. In addition, extended work of Newman et al. [9] and others has proven that clustering with maximized modularity often yields promising community structures in real networks [10].

Pons and Latapy [11] proposed the Walktrap algorithm for automatic community detection. The algorithm adopted the idea of random walk through a network for community detection. The main intuition of this approach was that densely connected portion of a community would tend to trap random walkers. Orman et al. [12] compared different community detection algorithms (Label Propagation, Eigenvector, Walktrap, etc.) with networks generated with a model from Lancichinetti et al. [13]. Normalized mutual information measure was used to access the performance of those algorithms. Walktrap has been one of the best algorithms that generate excellent results by successfully identifying communities even for high mixing coefficient values [12].

III. Methodology

As we discussed earlier, Modularity and the Walktrap algorithms are capable of determining clusters in a network. Thus, they could be used to identify the proper number of clusters in a network. In turn, the ε value of SCAN is calculated with the known number of clusters. Therefore, the determined ε value is associated to the particular dataset. This assists researchers to obtain the ε value more quickly.

Three real-world networks are selected as our test dataset, which is discussed in section 4. These datasets are a standard benchmark for community detection algorithms and are listed as follows: (1) The Zachary karate club [14], (2) A network of NCAA Division-IA football programs [15], (3) Books about US politics [16]. However, by applying the Fast Modularity algorithm (a modularity maximization approach [3]) and the Walktrap algorithm on the three datasets, both algorithms fail to determine the target number of clusters for all three networks (Table I). Note that the target number of clusters represents the true number of communities a particular dataset consist of. Therefore, the number of cluster intervasl from both the Fast Modularity and the Walktrap algorithms are obtained by extracting the coresponding number of clusters in the top three modularity value of the dataset. In turn, a proper interval ε value of SCAN is determined. The purpose of this interval ε value acts as a guideline for SCAN users to identify the ε value more quickly.

TABLE I. Number of cluster prediction with Fast Modularity and Walktrap

Dataset	Algorithm	Predicted # Clusters	Target # Clusters
Zachary	Fast Mod.	3	2
	Walktrap	3	
NCAA	Fast Mod.	7	11*
	Walktrap	10	
US P. Book	Fast Mod.	4	3
	Walktrap	4	

*SCAN algorithm indicates the value as a good result

Originally, the number of communities in a network is determined by maximizing the modularity of the dataset. In this study, the interval that binds the number of clusters is obtained based on the top three modularity values. There are at least three ways to select the number of cluster intervals. Table II shows the three scenarios. In this study, the scenarios 1 and 2 are not chosen because they are considered uncertain. Since the maximum modularity value is proven to be unreliable, there is no way to tell whether the true number of clusters falls in interval [3,4] or interval [4,5]. Therefore, scenario 3, the combination of scenarios 1 and 2, is chosen.

TABLE II. Scenarios of Choosing Number of Cluster Interval

Scenarios	# cluster interval	Modularity Interval
1	[3,4]	Top 1 and 3
2	[4,5]	Top 1 and 2
3	[3,5]	Top 1, 2, and 3

The Fast Modularity algorithm and the Walktrap algorithm are employed to determine the number of clusters interval of each dataset. The corresponding number of cluster interval is extracted based on the top three modularity values (Figure 2). Three dataset mentioned earlier are processed by both Fast Modularity and Walktrap algorithms in order to obtain the number of clusters. As a result, three set of intervals are created for each dataset using: (a) the Fast Modularity algorithm [i,j], (b) the Walktrap algorithm [v,w], and (c) the interval combination (Interval C) of (a) and (b) based on equation 1.

$$Interval\ C = \begin{cases} [\min(i,i,v,w)\,,\max(i,j,v,w)], & if\ intervals\ overlap \\ [i,j],[v,w], & otherwise \end{cases} \quad (1)$$

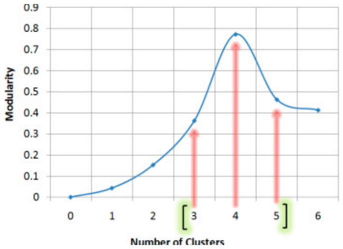
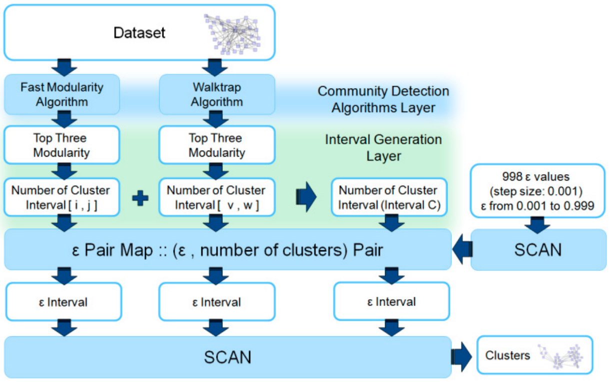


Figure 2. Formation of the Number of Cluster Interval Based on the Modularity Interval.



Dataset
Fast Modularity Algorithm
Walktrap Algorithm
Community Detection Algorithms Layer
Top Three Modularity
Top Three Modularity
Interval Generation Layer
Number of Cluster Interval [i , j]
Number of Cluster Interval [v , w]
Number of Cluster Interval (Interval C)
998 ε values (step size: 0.001) ε from 0.001 to 0.999
ε Pair Map :: (ε , number of clusters) Pair
SCAN
ε Interval
ε Interval
ε Interval
SCAN
Clusters

Figure 3. Workflow of Determining Structural Similarity Threshold Interval.

In this study, the number of community interval of each dataset is determined based on the Fast Modularity and the Walktrap algorithms. SCAN is then applied to determine the structural similarity threshold interval value. Figure 3 illustrates the process of how the number of clusters intervals is generated. After the number of cluster interval is found, the interval is passed to SCAN to locate ε interval and a ε pair map is generated. Therefore, for each dataset, the number of groups in a network that are generated by the SCAN algorithm with different structural similarity values are pre-calculated from ε = 0.001 to ε = 0.999 with the step size of 0.001. A total number of 998 of ε values are used.

A. *Modularity*

When nodes are partitioned randomly or all nodes are clustered into one cluster, the modularity measure yields zero (Q_n=0). In addition, the maximum modularity measure (Q_n=1) suggests that the optimal clustering is reached. Maximizing the modularity is the key to obtain the best quality partitions of a network. Since modularity optimization is an NP-complete problem [17], several computational heuristics have been created [18].

B. *Fast Modularity Algorithm*

The Fast Modularity algorithm is proposed by Clauset et al. [3] and it employs the modularity maximization approach. Since computing all possible partitions in a network by directly optimizing modularity is infeasible [19], Clauset et al. [3] apply modularity to direct their greedy hierarchical agglomeration procedure. The complexity of this algorithm is $O(nlog^2n)$.

C. *Walktrap Algorithm*

The Walktrap algorithm is developed by Pons and Latapy [11]. The algorithm applies the idea of random walk through a network for community detection. The core intuition of this method is that densely connected portion of a community tends to trap random walkers. The complexity of this algorithm is $O(n^2logn)$.

D. *The SCAN Algorithm*

The algorithm starts with examining if a node v is a core $(C\varepsilon\mu(v))$. A node is a core iff this node has at least μ number of neighbor nodes and each neighbor node has structural similarity (ss) greater or equal to ε $(N\varepsilon(v))$. μ and ε are threshold values representing the minimum cluster size and minimum ss value respectively where ss is as follows.

$$ss(v,w) = \frac{|\Gamma(v)\cap\Gamma(w)|}{\sqrt{|\Gamma(v)||\Gamma(w)|}} \qquad (2)$$

If the examined node is not a core node, this node will be labeled as non-member. However, if the node (v) is a core, a new cluster identifier (clusterID) is generated. A queue (Q) which composes of nodes of $N\varepsilon(v)$ is created. For every point in Q (y_i), the direct reachable nodes (x_i) are located (R). A node needs to be the neighbor of a core node with their ss) greater than or equal to ε. The current clusterID is assigned to x_i if it is a non-member or unclassified. After each iteration, a node is eliminated from Q until Q is empty. After every member in Q is processed, each non-member nodes will be identified as either a hub or an outlier. In order to be a hub, a nonmember node is required to have neighbors from two or more different clusters. Otherwise, this nonmember node is an outlier. The complexity of the SCAN algorithm [6] is $O(n)$.

IV. Dataset

All datasets used for experiment and evaluation in this study are publicly available.

A. *The Zachary Karate Club*

The famous Zachary's karate club dataset [14] describes the mutual relationship of 34 people with 78 edges. The dataset was carefully examined by Zachary over a period of two years. The graph is separated into two clusters because of the contrasts between an administrator and a teacher of the club. Based on the original study, the division of the club was caused by the disagreement of the teacher and the administrator. Hence two groups were formed with the leader being the teacher and the administrator. The dataset is freely available online.

B. *A Network of NCAA Division-IA Football Programs*

The dataset is an un-weighted network that was extracted from the schedule of 115 National Collegiate Athletic (NCAA) Division I-A football teams in 2000 [15]. 115 schools are divided into eleven conferences with addition of four independent schools. Each vertex of the network describes a college football team while every edge shows the relationship where two football teams play against each other. Conferences that correspond to clusters are identifiable in the network because inter-conference matches happen more frequently than the intra-conference matches.

C. *Books about US Politics*

The books about US politics dataset are compiled by Valdis Krebs [16]. Each vertex of the network describes US politics books sold by Amazon.com while every edge shows the frequent co-purchasing of books by the same buyers.

V. Experiments and Results

Three set of intervals are created for each dataset using: (a) the Fast Modularity algorithm, (b) the Walktrap algorithm, and (c) the interval combination of (a) and (b) based on equation 1. As presented in Table III, all intervals generated successfully covered the target number of clusters in all datasets. As discussed in section 3, the ε intervals are extracted from the pre-calculated ε values from SCAN using the ε pair map. Once again, the predicted ε interval values correctly include ε values which the SCAN paper [6] proposed in all datasets (Table IV). Note that the Zachary dataset is not examined in the original SCAN paper. Table V illustrates the percentage of steps saved by using the proposed prediction method versus the original visualization approach. In other words, suppose a user attempts to locate the best cluster structure of the Zachary dataset using the SCAN algorithm with the ε step size of 0.001. In the worst case, a user requires to view 998 instances of community structures (for ε from 0.001 to 0.999). However, with the proposed prediction method, only 189 instances of cluster structures are required to be examined. In Table V, the "interval width" and the "percentage steps saved" are determined as follows: (e.g. for the Zachary dataset).

Interval Width = (0.387 – 0.199) * 1000 + 1 = 189
Steps Saved = ((998 - Interval Width) / 998)*100 = 81.1%

TABLE III. Number of Cluster Interval Prediction with Fast Modularity and Walktrap Algorithm

Dataset	Algorithm	Predicted # of Clusters Interval	Target # of Clusters
Zachary	Fast Mod.	[2,4]	2
	Walktrap	[2,4]	
	Both	[2,4]	
NCAA	Fast Mod.	[6,8]	11*
	Walktrap	[9,11]	
	Both	[6,11]	
US P. Book	Fast Mod.	[3,5]	3
	Walktrap	[3,5]	
	Both	[3,5]	

* SCAN algorithm indicates the value as a good result.

TABLE IV. Predicted Structural Similarity E Interval based on Fast Modularity and Walktrap

Dataset	Algorithm	Predicted *ε* Interval	SCAN proposed *ε*
Zachary	Fast Mod.	[0.199,0.387]	NA
	Walktrap	[0.199,0.387]	
	Both	[0.199,0.387]	
NCAA	Fast Mod.	[0.205,0.6]	0.5
	Walktrap	[0.273,0.572]	
	Both	[0.205,0.6]	
US P. Book	Fast Mod.	[0.209,0.612]	0.35
	Walktrap	[0.209,0.612]	
	Both	[0.209,0.612]	

TABLE V. Percentage of Steps Saved by Using the Interval E Values

Dataset	Algorithm	Interval Width	% Steps Saved
Zachary	Fast Mod.	189	81.1%
	Walktrap	189	
	Both	189	
NCAA	Fast Mod.	396	60.3%
	Walktrap	300	69.9%
	Both	396	60.3%
US P. Book	Fast Mod.	404	59.5%
	Walktrap	404	
	Both	404	
Average			**67.7%**

The "% Steps Saved" shows that the proposed prediction method helps users to eliminate an average of 67.7% of instances where inappropriate ε values are using to generate clusters.

VI. Conclusion

In this study, the Fast Modularity algorithm and the Walktrap algorithm are employed to predict the target number of communities of a dataset. An interval value is obtained by extracting the corresponding number of clusters in the top three modularity values of the dataset using the aforementioned algorithms. For each dataset, the number of groups in a network that are generated by the SCAN algorithm with different structural similarity values are pre-calculated from ε = 0.001 to ε = 0.999 with the step size of 0.001. A total number of 998 of ε values are used. Since the number of communities interval of a network has been predicted, the range of ε values is determined based on the pre-calculated values from SCAN. For each of our test datasets, the interval prediction reliably binds the true number of clusters. More importantly, the proposed prediction method helps users to eliminate an average of 67.7% of inappropriate ε values used to generate clusters.

References

[1] R. Guimera and L. A. N. Amaral, "Functional cartography of complex metabolic networks." Nature 433, 895–900 (2005).

[2] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks", Phys. Rev. E 69, 026113 (2004).

[3] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community in very large networks", Physical Review E 70,066111 (2004).

[4] C. Ding, X. He, H. Zha, M. Gu, and H. Simon, "A min-max cut algorithm for graph partitioning and data clustering", Proc. of ICDM 2001.

[5] J. Shi and J. Malik, "Normalized cuts and image segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 22, No. 8, 2000.

[6] Xu, X., Yuruk, N., Feng, Z., Schweiger, T.A.J.: SCAN: a Structural Clustering Algorithm for Networks. In: Proc. of SIGKDD 2007, pp. 824–833. ACM Press, New York (2007).

[7] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96), Portland, OR, pages 291-316. AAAI Press, 1996.

[8] Newman, M. E. J., "Modularity and community structure in networks", Proc. Natl. Acad. Sci. USA in press (2006).

[9] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. Physical Review E, 72, 2005.

[10] A. Clauset. Finding local community structure in networks. Physical Review E, 72, 2005.

[11] Pascal Pons, Matthieu Latapy. "Computing Communities in Large Networks Using Random Walks." In ISCIS. pp. 284-293 2005.

[12] Orman, Günce Keziban and Vincent Labatut, A Comparison of Community Detection Algorithms on Artificial Networks, J. Gama et al., LNAI 5808, Berlin Heidelberg, Springer-Verlag, 242–56, 2009.

[13] Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E Stat. Nonlin Soft. Matter Phys. 78, 46110 (2008.)

[14] Zachary W W J Anthropol Res 33:452–473, 1977.

[15] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks. Preprint cond-mat/0308217 (2003).

[16] http://www.orgnet.com/

[17] Brandes, Ulrik, Daniel Delling, Marco Gaertler, Robert Goerke, Martin Hoefer, Zoran Nikoloski & Dorothea Wagner. "On modularity clustering." IEEE Transactions on Knowledge and Data Engineering 20(2):172–188 2008.

[18] Danon, Leon, Albert Diaz-Guilera, Jordi Duch & Alex Arenas. "Comparing community structure identification." Journal of Statistical Mechanics: Theory and Experiment, 2005.

[19] Brandes, U., Delling, D., Gaertler, M., et al.. On finding graph clusterings with maximum modularity. LNCS, vol. 4769. Springer, Berlin, Heidelberg. pp. 121–132, 2007.

[20] Mutlu Mete, Fusheng Tang, Xiaowei Xu, Nurcan Yuruk, A Structural Approach for Finding Functional Modules from Large Biological Networks, BMC Bioinformatics, vol. 9, S19, 2008.