# Detecting Hierarchical and Overlapping Community Structures in Networks

Nuwan Ganganath†*, Guanrong Chen‡, and Chi-Tsun Cheng†

†Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong
‡Department of Electronic Engineering, City University of Hong Kong, Hong Kong
*Email: nuwan.marasinghearachchige@connect.polyu.hk

**Abstract**—Community structure can be observed in many natural, biological and social networks. Studies suggest that these communities may have organized in a hierarchical manner while some communities overlap with others. This paper introduces an algorithm to detect such hierarchical and overlapping community structures in networks based on the concept of maximal cliques. It introduces an alternate modularity for evaluating overlapping community structures. Unlike existing algorithms for detecting hierarchical and overlapping community structures, the new algorithm is free of parameter tuning and random seeds. Experiments conducted on two real-world networks show that this algorithm is capable of providing satisfactory and consistent results.

## 1. Introduction

Many real-world systems can be represented in terms of networks, which consist of vertices connected together by edges. A common feature in many networks is the community structure, where the nodes within the same community are more likely to be connected to each other than nodes in between communities. Community structures reveal the organization and interactions of vertices in a network. Uncovering such communities is important to understand the network properties. One of the most popular methods for detecting community structure is proposed by Newman *et al.* using a benefit function, namely *modularity*, to evaluate quality of the community structure [1, 2, 3]. Later, several modularity-based optimization algorithms were developed to find community structures in networks.

While high attentions have been attracted to non-overlapping community structures in networks, many real-world networks consist of overlapping and hierarchical communities [4, 5]. Vertices may belong to more than one community if overlapping communities exist. Community structure is said to be hierarchical if a community can be further divided into sub-communities. Despite the applicability in many real-world systems, due to the complexity of the problem, very few attempts have been reported in the literature which consider both overlapping and hierarchical community structures simultaneously. An early attempt on hierarchical and overlapping community detection on networks was made by Lancichinetti *et al.* [6]. Their algorithm performs a local optimization on a fitness function and the hierarchical organization is uncovered by tuning a parameter to change the resolution. However, their algorithm is not robust and the detection accuracy is not guaranteed due to the random selection of seed vertices. Shen *et al.* proposed an extended modularity measure for evaluating overlapping communities [7]. They also proposed an algorithm called EAGLE to find overlapping and hierarchical community structures in networks. However, their work suffers from two major drawbacks: the extended modularity is unable to distinguish the goodness of the belonging of vertices to each of the overlapping communities, and the threshold for dropping small cliques need to be manually tuned in EAGLE. More recently, Hung *et al.* introduced another algorithm called DenShrink for detecting hierarchical and overlapping community structures [8]. DenShrink uses both density-based clustering and modularity optimization to reveal community structures. However, similar to EAGLE, one needs to manually tune the threshold for detecting micro-communities when using DenShrink.

This paper presents a novel algorithm for detecting overlapping and hierarchical community structures in networks. The proposed agglomerative algorithm consists of two phases: finding *maximal cliques* and constructing a *dendrogram*. This algorithm is capable of providing robust and consistent results as it does not make use of any random seeds. Also, no parameter tuning is necessary for performing the proposed algorithm. Here, a new modularity measure is defined by extending the traditional modularity, which is used to evaluate the quality of the overlapped community decomposition. The new modularity is capable of distinguishing the goodness of the belonging of vertices to each of the overlapping communities.

The rest of the paper is organized as follows. In Section 2, the extension of modularity for overlapping community detection is described. The novel algorithm for detecting overlapping and hierarchical community structures is introduced and discussed in Section 3. Section 4 presents the results of the proposed algorithm on two standard benchmarks. Concluding remarks are given in Section 5.

## 2. Extending the Modularity Measure for Decomposing Overlapping Communities

The basic idea behind Newman's modularity for quantifying communities is the edge density in a subgraph of a network compared to a null-model. Here, the null-model is defined as a subgraph with same number of vertices, same

number of edges, and same degree distribution as the original subgraph, but edges are randomly placed. In such a random graph, the probability of having vertex $i$ connected to vertex $j$ is given by $P_{ij} = k_i k_j / 4m^2$, where $m$ is the total number of edges in the network, and $k_i$ and $k_j$ are the degrees of vertices $i$ and $j$, respectively. However, the same probability for the original subgraph is given by $a_{ij}/2m$, where $a_{ij}$ are the terms in the network adjacency matrix. Therefore, if $V$ is the set of vertices of a graph, Newman's modularity is defined as

$$Q = \frac{1}{2m} \sum_{i,j \in V} \left[ a_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{1}$$

where vertices $i$ and $j$ belong to communities $c_i$ and $c_j$, respectively. If $c_i = c_j$ then $\delta = 1$, otherwise $\delta = 0$. Therefore, the above equality can be rewritten as

$$Q = \frac{1}{2m} \sum_{n} \sum_{i \in c_n, j \in c_n} \left[ a_{ij} - \frac{k_i k_j}{2m} \right]. \tag{2}$$

Strong community structure can be observed if $Q$ is close to 1. When the number of edges within a community gets close to random, $Q$ will tend to 0. Obviously, if all the nodes in a network belong to a single community, then $Q$ =0.

In the case of overlapping communities, a vertex may belong to more than one community. The strength of their attachment to each community can be different depending on the number of connections they have with each community. Based on this observation, the original definition of modularity for evaluating overlapping communities is extended as

$$\begin{aligned} Q_o &= \frac{1}{2m} \sum_{n} \sum_{i \in c_n, j \in c_n} \left[ \frac{k_i^c k_j^c}{k_i k_j} \right] \left[ a_{ij} - \frac{k_i k_j}{2m} \right], \\ &= \frac{1}{2m} \sum_{n} \sum_{i \in c_n, j \in c_n} k_i^c k_j^c \left[ \frac{a_{ij}}{k_i k_j} - \frac{1}{2m} \right]. \end{aligned} \tag{3}$$

Here, $k_i^c = \sum_{p \in V_{c_n}} a_{ip}$ and $k_j^c = \sum_{q \in V_{c_n}} a_{jq}$, where $V_{c_n}$ is the set of vertices in the community $c_n$. Similarly to the traditional modularity $Q$, the proposed extended modularity $Q_o = 0$ when all the nodes belong to the same community and gets a higher value to indicate a stronger community structure.

## 3. The Algorithm

In this section, a two-phase agglomerative algorithm is introduced to find the hierarchical and overlapping community structure in a network. This algorithm is based on two main concepts: maximal cliques and the extended modularity. A *clique* can be identified as a subset of vertices in a network such that every two vertices in the subset are connected by an edge. A maximal clique is a clique which is not a subset of any other clique.

In the first phase of the algorithm, it finds all the maximal cliques in the network. In many real-world problems, finding maximal cliques is easy due to the sparseness of the networks and many algorithms are proposed for that. Here, the Bron-Kerbosch algorithm which is based on a recursive backtracking procedure [9], is utilized. It provides a set of maximal cliques for a given network. One should note that a single vertex may be included in several maximal cliques, which are referred to as overlapping vertices. Here, not all of the maximal cliques are taken into account. If a maximal clique is made of the vertices from some other maximal cliques, we discard it. (E.g.: If {1, 2, 3, 4, 5}, {5, 6, 7, 8}, and {2, 3, 5, 7} are three maximal cliques in a given network, the proposed algorithm discard {2, 3, 5, 7} as all of its vertices are already included in the first two cliques.) In the implementation, the maximal cliques are stored in a sorted array in descending order based on the number of vertices in each clique. Then, it iterates through the array and discards certain cliques based on the above-mentioned criteria. This first phase considerably reduces the problem size for the second phase of the algorithm.

In the second phase of the algorithm, the maximal cliques generated in the previous phase are considered as the initial communities for detecting the hierarchical community structure of the network. Similarly to the fast algorithm introduced in [2], these initial communities are joined together in pairs such that it results in greatest increase or smallest decrease in the modularity of the network, and a dendrogram is obtained. In contrast to the algorithm explained in [2], here it does not start from sole vertices and it tries to maximize $Q_o$ instead of $Q$. If one community can be represented as a subset of another community which is generated by joining two other communities together, the subset community will also be absorbed into the larger community. (E.g.: If {1, 2, 3, 4, 5, 6} is generated by joining communities {1, 2, 3} and {4, 5, 6} together, and if there is another small community {2, 3, 6}, then the latter will also be absorbed into {1, 2, 3, 4, 5, 6}.) The level of cut of the dendrogram is decided according to the value of $Q_o$ as its maximum value corresponds to the strongest community structure of the network.

## 4. Results and Performance Analysis

This section discusses the results of the proposed algorithm on two real-world networks. The new algorithm was implemented in MATLAB and all the experiments were conducted on a computer with 2.67 GHz processor, 12GB memory, and Windows 7 operating system.

The first experiment was conducted on the Zachary karate club network [10] (see Figure 1), which is commonly used as a benchmark for community detection methods. It consists of 34 vertices and 78 edges. After completing the first phase of the algorithm on the network under test, the network is reduced to 23 communities, which are given in the bottom layer of the dendrogram shown in Fig-

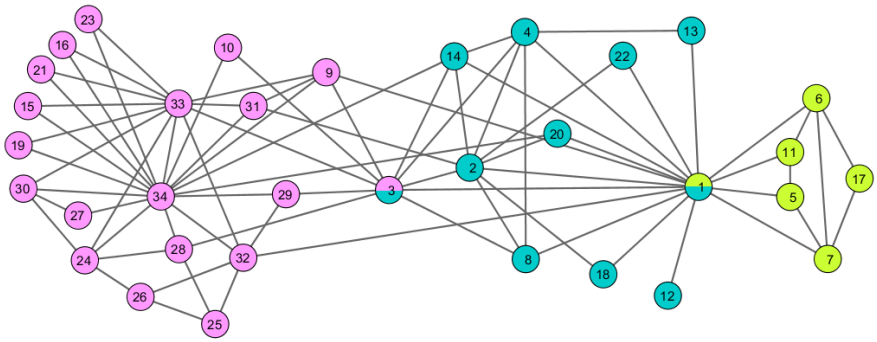


Figure 1: The overlapping community structure detected by the proposed algorithm on the Zachary karate club network

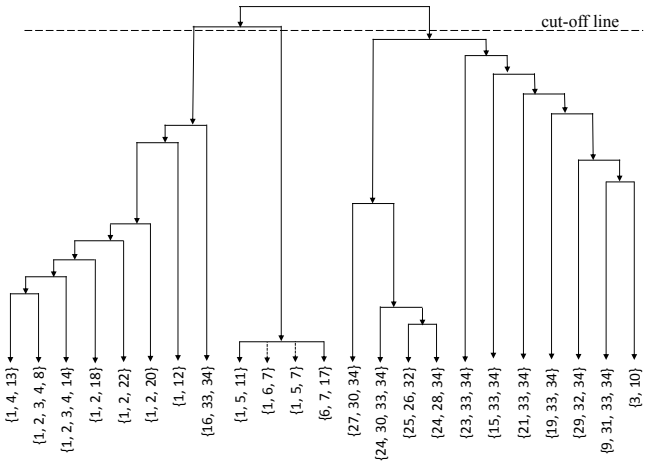


Figure 2: A dendrogram illustrating the hierarchical community structure of the Zachary karate club network

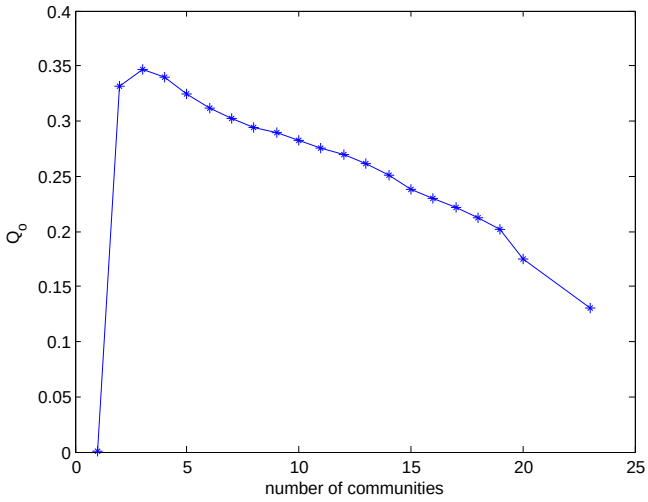


Figure 3: The extended modularity for the hierarchical community structure detected by the proposed algorithm on the Zachary karate club network

ure 2. Second phase of the algorithm reveals the hierarchical community structure by optimizing $Q_o$. In the first step of the second phase, it combines communities {1, 5, 11} and {6, 7, 17} together, resulting in {1, 5, 6, 7, 11, 17}. Since {1, 6, 7} and {1, 5, 17} are subsets of the resulted community, they are absorbed into the same larger community, as illustrated by dashed lines in the dendrogram. Therefore, at the end of the first step in creating dendrogram, the number of communities reduces from 23 to 20. This process continues until a single community remains. The value of $Q_o$ corresponding to each step is recorded. The change of $Q_o$ against the number of communities is shown in Figure 3. The cut-off line of the dendrogram is decided based on the highest $Q_o$ for this network. The strongest community structure for the given network is detected when $Q_o = 0.34662$ which corresponds to three overlapping communities with two overlapping nodes. This is illustrated in Figure 1. By investigating further into the hierarchical community structure of the network, it can be observed that the community {1, 5, 6, 7, 11, 17} (represented in yellow color) joins with its neighboring community (represented in cyan color) at a cost of $\Delta Q_o = 0.0152$, reducing the network to only two overlapping communities with only vertex {3} remaining in overlapping between these two communities.

The second experiment was conducted on Lusseau's social network of dolphins [11] (see Figure 4), which is also considered as a benchmark for community detection in networks. It consists of 62 vertices and 159 edges. After completing the first phase of the algorithm on the network under test, the network is reduced to 43 communities. Second phase of the algorithm discovers the hierarchical community structure by optimizing $Q_o$. The variation of $Q_o$ against the number of communities is shown in Figure 5. The strongest community structure for Lusseau's social network of dolphins is detected when $Q_o = 0.4210$, which comprehends three overlapping communities with seven overlapping nodes as illustrated in Figure 4. By moving one step upward in the hierarchical community structure of the network, the communities represented in pink and yel-

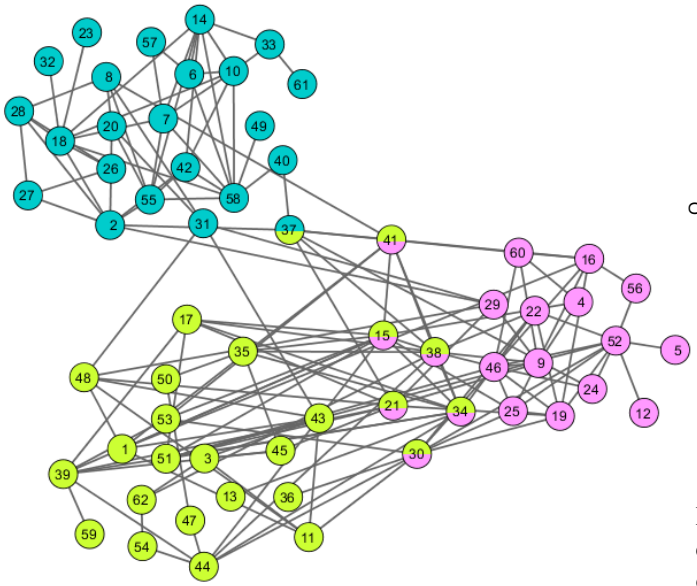


Figure 4: The overlapping community structure detected by the proposed algorithm on Lusseau's social network of dolphins

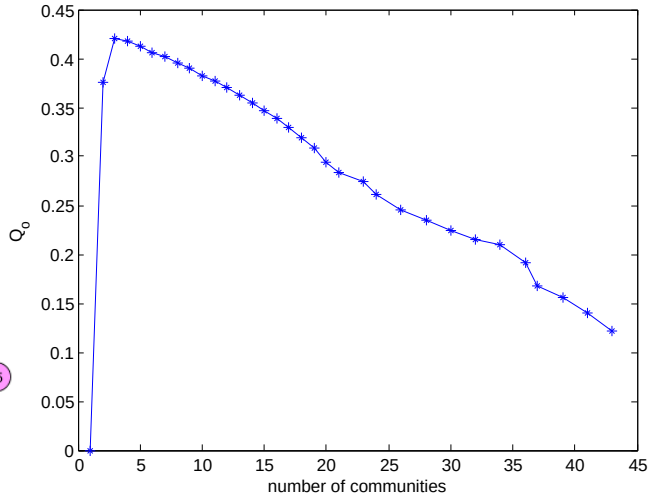


Figure 5: The extended modularity for the hierarchical community structure detected by the proposed algorithm on Lusseau's social network of dolphins

low join together at a cost of $\Delta Q_o = 0.0455$, reducing the network to two overlapping communities with only vertex {37} remaining in overlapping between these two communities.

## 5. Conclusions

In this paper, an algorithm for uncovering hierarchical and overlapping community structures in networks is proposed. This algorithm consists of two phases. In the first phase, it detects the maximal cliques in the network. Maximal cliques that are not made of the vertices from some other larger maximal cliques are proceeded to the second phase. In the second phase, the algorithm creates a dendrogram to represent the hierarchical community structure of the network. A new modularity metric is introduced for evaluating the overlapping community structure in each level of the dendrogram. The experimental results show that the proposed algorithm is capable of detecting hierarchical and overlapping community structures in networks. Uncovering such a structure may help in understanding network behaviors and dynamics.

## References


[1] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, p. 026113, 2004.

[2] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, p. 066133, 2004.

[3] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, p. 066111, 2004.

[4] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[5] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[6] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, p. 033015, 2009.

[7] H. Shen, X. Cheng, K. Cai, and M.-B. Hu, "Detect overlapping and hierarchical community structure in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 8, pp. 1706 – 1712, 2009.

[8] J. Huang, H. Sun, J. Han, and B. Feng, "Density-based shrinkage for revealing hierarchical and overlapping community structure in networks," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 11, pp. 2160 – 2171, 2011.

[9] C. Bron and J. Kerbosch, "Algorithm 457: Finding all cliques of an undirected graph," *Commun. ACM*, vol. 16, no. 9, pp. 575–577, Sep 1973.

[10] W. Zachary, "An information flow modelfor conflict and fission in small groups1," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[11] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, 2003.