Journal of Advanced Statistics, Vol. 2, No. 2, June 2017
https://dx.doi.org/10.22606/jas.2017.22002

Network Exploration by Complements of Graphs with Graph Coloring

Tai-Chi Wang[1], Frederick Kin Hing Phoa[2*] and Yuan-Lung Lin[2]

[1]National Center for High-Performance Computing, National Applied Research Laboratories, Taiwan
[2]Institute of Statistical Science, Academia Sinica, Taiwan
Email: fredphoa@stat.sinica.edu.tw

Abstract Network data have become very popular with the growth of technologies and social applications such as Twitter and Facebook. However, few visualization tools have been created for exploring large-scale networks. We propose a simple and quick procedure to explore a network in this study. The algorithm changes the edge representation based on the complement of a simple graph and the partition method of vertex coloring. Furthermore, the colors provide additional information on top of the partitions. Our proposed method is demonstrated in some famous networks.

Keywords: Visualization, Complement of Graph, Greedy Algorithm, Network Partition, N-clique

1 Introduction

Graph theory has been verified as a useful approach to explain complex networks. For example, Erdős and Rényi [1] first provided Poisson random model to explain the relation in a network. Watts and Strogatz [2] used a small-world model to describe the dynamic network system. Barabási and Albert [3] provided a scale-free model to explain the expanded connectivity of networks. There have been huge amount of studies and applications since these models were proposed.

It is difficult to glimpse structures of large-size networks due to their network complexities. Although many useful R-, Python- and C-based softwares such as Cytoscape [4], Gephi [5], and igraph [6] were developed to visualize and analyze network data, many network features were still hindered inside their complicated structures. Few studies proposed simple and efficient methods to visualize networks. Fruchterman and Reingold [7] provided some useful methods to embed proper vertex locations and to make visualization easier. However, once a network is big, the analysis becomes a time-consuming task and the network features are still difficult to clearly visualize.

Among complicated network patterns, recent researches have focused on exploring network patterns by graph partition and network communities. Graph partition is an important issue in network recognition. The main issue of graph partition is how to divide a graph into subgraphs so that the number of edges connecting different subgraphs is minimized. For example, Karypis and Kumar [8] provided a multilevel partition algorithm by considering the coarsening and uncoarsening procedures. Arora et al. [9] developed a $O(\sqrt{\log n})$-approximation algorithm that can quickly partition a graph into two parts by a l_2^2-representation approach. Similar but not exactly equivalent to a graph partition problem that finds the minimum number of cut, a community detection problem focuses on finding subgraphs so that each subgraph is densely connected. Given the reachability of community and the geodesic distance among community members, Wasserman and Faust [10] defined two important quantities: (1) the $n-$clique is a maximal subgraph in which the largest geodesic distance between any two nodes is no greater than n, and (2) the $n-$club restricts the geodesic distance within the subgraph to be no greater than n. These $n-$cliques and $n-$clubs are possible communities under a weaker definition. Furthermore, the modularity-based method [11] considered a modularity measure to evaluate the similarity of groups. A spectral optimization method is used to classify network communities. Bickel and Chen [12] further investigated the theoretical properties of the modularity measure. Bayesian models, which were developed by Handcock et al. [13] and Heard et al. [14], deduced network properties and assigned cluster labels to vertices by posterior samples. Readers who are interested in other community detection approaches and definitions can refer to Wasserman and Faust [10] and Tang and Liu [15].

Although these models and methodologies are helpful in realizing the network patterns, they are commonly limited by their computational efficiencies. In specific, these methodologies might not be helpful and informative if one needs to quickly characterize a complex network within a limited amount of time. For the visualization purpose, an initial step to understand a network is still by raw visual recognitions. In addition, it is common among traditional approaches that they use geodesic distances between pairs of vertices in a network. If we want to quickly glimpse structures of a network according to the interactive distances among vertices, it is essential to develop a simple representation to summarize a network.

In general, we want to construct an algorithm to achieve the following goals:

- In terms of computing efficiency, the algorithm can quickly visualize and explore a network.
- In terms of graph partition, the algorithm can label vertices in a network.
- In terms of cluster pattern visualization, same labels are assigned to close vertices.
- In terms of extended information, the proposed approach can provide additional information on top of network partition.

In order to achieve the above goals, we propose a popular but old-school approach, *graph coloring*, to visualize and to explore networks in a simple way. Graph coloring is one of the most useful tools developed in graph theory. It originally arose from the four-colors problem and has been widely used in other fields such as time tabling and scheduling, register allocation, printed circulated board testing, and so on [16]. In operation research, many researches used this approach to deal with scheduling problems and to make decisions [17, 18]. In compiler optimization, graph coloring is an important tool that used for solving the register allocation problem [19, 20]. The original purpose of graph coloring is to find the minimum chromatic number for labeling a graph. This leads to the partition of a graph into small groups when each group has good properties such as $n-$cliques or $n-$clubs (for recognizing partition and cluster patterns).

Although Graph coloring is a promising approach to feature a network, it suffers some difficulties when applying to network visualization. One of the most difficulties is the meaning of labels of graph coloring. The most important and fundamental property of the graph coloring is "two adjacent vertices must label different colors". The opposite side of this property is "two vertices with the SAME color must be NOT adjacent". Suppose a studied graph is denoted as G. The graph coloring technique on G can shed the light on the dependence of cliques, i.e., the clique members must have different colors. On the other hand, we consider to color the *complement* of a graph, $\overline{G}$, where two vertices of $\overline{G}$ are connected if and only if they are not connected in G. Based on the above "opposite" property, two vertices with the SAME color must be NOT adjacent on the colored $\overline{G}$. Interestingly, this means that the vertices with the same colors are adjacent on the original graph G. Thus, the colors of $\overline{G}$ can be the partition labels. Based on this idea, we can apply the graph coloring technique to partition a network. In this study, we propose the *greedy vertex coloring* approach for coloring a graph. The greedy vertex coloring approach has been verified to run $O(n+m)$ time [21], and this running time is efficient and acceptable in the case of a large scale network. Thus, the graph coloring is a good approach to achieve our goals.

The remaining part of this paper is organized as follows. First, we introduce some preliminary concepts of graph theory and the basic algorithm for coloring in section 2. In section 3, we consider the complements of different edge representations for partitions by graph coloring, and then we use the Karate network [22] to demonstrate our proposed method with an emphasize on aspects such as labeling the complement results, the representation of partitions, and statistics of labels and edges. Section 4 shows the feasibility and efficiency of our proposed approach using other famous networks. The last section provides the discussion and the final conclusion of the paper.

2 Preliminary Definitions and Graph Coloring

In this study, we focus on the undirected network and vertex coloring only. Since an undirected network is simpler to visualize than a directed network, we can directly see the patterns via graph coloring.

2.1 Preliminary Definitions

A simple *graph* is an ordered pair $G = (V, E)$, where $V = \{v_1, \ldots, v_n\}$ is a set of vertices and $E = \{e_1, \ldots, e_m\}$ is a set of edges in which each edge is denoted as (v_i, v_j). In practice, each edge is denoted as

(v_i, v_j), where the vertices v_i and v_j are known as the endpoints of the corresponding edge. Additionally, they are said to be *adjacent* to each other, and are also called *neighbors* to each other. The *adjacency matrix* of G is an n by n square matrix with entries $a_{ij} = 1$ if v_i and v_j are adjacent and $a_{ij} = 0$ otherwise. The *degree* of a vertex v is the number of neighbors of v. A *u-v path* is a sequence of distinct vertices $\{u = v_0, v_1, \ldots, v_{p-1}, v_p = v\}$, and any two consecutive vertices are joined by an edge. The *shortest path* from u to v is the *u-v* path containing minimum number of edges among all paths from u to v, and the *distance* between two vertices is the number of edges in the shortest path from u to v, denoted as $d_G(u, v)$. A graph is *connected* if it has at least one u-v path whenever $u, v \in V$.

A *colored graph* is a graph in which each vertex is assigned to a color. In this study, we use the *proper* vertex coloring, namely two vertices are assigned distinct colors if they are adjacent. The *chromatic number* of a graph G, denoted as $\chi(G)$, is the least number of distinct colors for properly coloring G. The complement or inverse of a simple graph G, denoted as $\overline{G}$, s a graph with the same vertex set of G and any two distinct vertices of $\overline{G}$ are adjacent under the circumstance that they are not adjacent in G.

2.2 Greedy Coloring

Graph coloring is an approach in graph theory for labeling a graph. This approach uses "colors" to assign labels to a graph subject to certain constraints. This study only considers the *vertex coloring* as a way of coloring the vertices of a graph so that no two adjacent vertices share the same color. There are many different algorithms to color a graph [23, 24, 25]. Some considered to evenly assign the colors while others considered to sequentially assign the smallest unlabeled colors that are not assigned to the current neighbors. We called the later methods as *greedy coloring*. Since graph coloring is a NP-complete problem, most of these algorithm cannot guarantee to provide the least number of colors for graph coloring. However, the greedy algorithms can provide a quick procedure to assign colors. We refer the readers to Kosowski and Manuszewski [21] for the details of the properties of graph coloring and some other coloring techniques.

Given a graph $G = (V, E)$, let us define the neighborhood of ν, denoted as $N_G(\nu)$, as a neighbor set of vertices adjacent to ν where ν can be a single vertex or a set of vertices. The greedy algorithm is performed in the following procedures (Algorithm 1). According to the algorithm, each color can be indexed by a positive integer, and each uncolored vertex is labeled as 0. In general, the greedy coloring assigns the smallest proper integer to an uncolored vertex first. We use the criterion, "the minimum degree", in step 2 and 4 to be the criterion for selecting nodes in the coloring procedure. To provide a unified and reproducible coloring result, "the smallest ID" criterion in step 4 is suggested when more than one vertices have the same degree. Thus, the ID must be set before applying our approach. Based on the minimum degree criterion, the vertices with lower degrees are labeled to the same color classes (smaller integers). On the other hand, the vertices with higher degrees often contain more important information. We can clearly view these vertices labeled as larger integers. It should be noted that this algorithm is applied to a "connected" graph. For a disconnected graph, this algorithm can be applied to each connected component separately. Please refer to Appendix to check the simple illustration of the algorithm by a toy example.

2.3 Demonstration of Greedy Coloring

In order to simply understand the algorithm and the following discussions, we use the karate network [22] to demonstrate the utility of graph coloring. The karate network is an undirected social network of friendships with 34 members of a karate club and 77 edges at a US university in the 1970s. Fig. 1 shows the network of karate club membership. The edges represent the interaction of the members both during and after the lessons of the club. The labels next to the vertices are the node IDs, and those inside the vertices are the degrees of vertices.

Since the graph coloring technique implies that two vertices are labeled in different colors if they are adjacent, the frequency of labels provides some useful information of a network. If we treat colors as numbers, then the largest number means the possible maximum size of the clique in a network. These cliques are referred to the possible groups or communities of interest. Therefore, we can quickly check the

Algorithm 1 Greedy Coloring

procedure GREEDYCOLORING($G = (V, E)$)
step 1 Initially set the colors of vertices as 0, i.e., $c(v_i) \leftarrow 0 \ \ \forall v_i \in V$.
step 2 Select the vertex with the minimum degree, called v_1.
step 3 Assign $c(v_1) \leftarrow 1$ and set $\nu = \{v_1\}$ to obtain $N_G(\nu)$.
for $j = 2$ to n **do**
step 4 Select the vertex with the minimum degree among the neighbor list $N_G(\nu)$, called v_j. (when there are more than one vertices that have the same degree, "the smallest ID" criterion is suggested.)
step 5 Select the minimum color that is not assigned to $N_G(v_j)$, i.e.,

$$c(v_j) = \min\{k \in \mathbb{N} | k \neq c(w) \ \ \forall w \in N_G(v_j)\}.$$

step 6 Set $\nu = \bigcup_{i=1}^{j} v_i$
end for
step 7 return c ▷ The vector of colors
end procedure

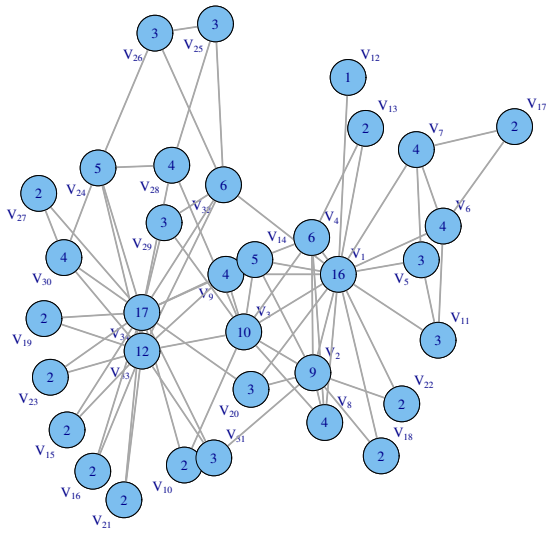


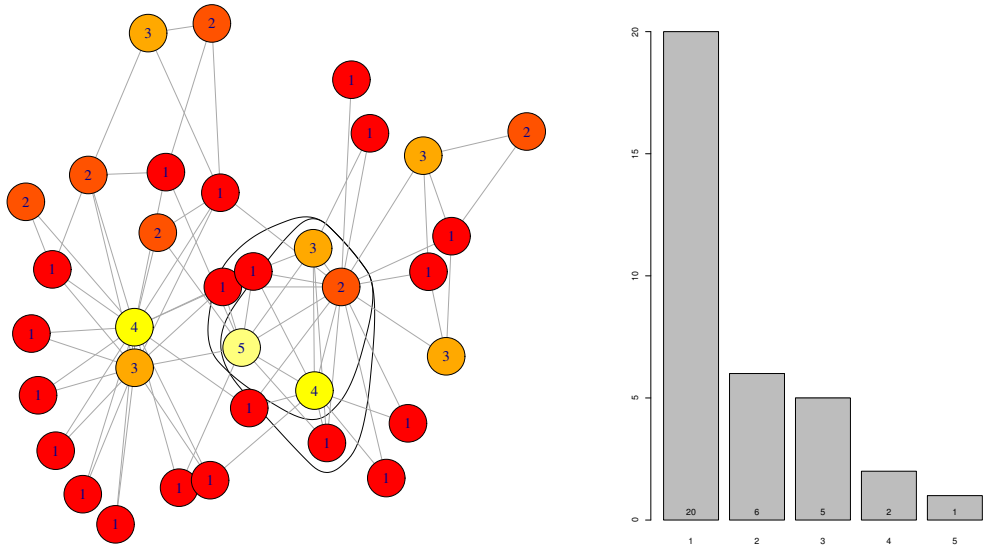
Figure 1: Karate club network

neighbors of these vertices with the maximum number to view the possible community structure of a network. Based on the frequency of labels, we can quickly depict a rough sketch of a network.

Fig. 2 shows the results of graph coloring for the karate network and the frequency of the labels. There is only one vertex denoted as label 5, which means that the possible maximum clique size is 5. To check if the true maximum cliques are with size 5 in this network, we found 2 cliques with size 5 and both of them contain the vertex "v_3" labeled as 5 (the vertices encircled by the black lines shown in Figure 3 (a)).

3 Applications to the Complement of a Simple Graph

Before discussing the coloring results of the complements, there is a need to recall the most important and fundamental property of the graph coloring – two adjacent vertices must label different colors. Like we have mentioned in the introduction section, we can shed some lights by applying the graph coloring technique. This technique cannot include all information of a network. However, some useful information can be provided for quickly viewing a complex network.



(a) Graph Coloring (b) Bar Chart of Labels

Figure 2: Graph coloring of karate network and its bar chart of labels

3.1 Complements of Path k Networks and Graph Partitions

We apply the properties of the complement of a graph G to visualize networks in this section. Based on the graph coloring, two adjacent vertices labeled as different colors show that they are connected with distance 1. Obviously, any two adjacent vertices in G are nonadjacent in $\overline{G}$. This implies that these pairs of vertices are away from distance 1. On the other hand, vertices with the same colors in $\overline{G}$ are adjacent to each other in G. For this reason, graph coloring is also a useful tool for graph partition.

Prior discussing about the complement of a graph, we need to introduce some important theorems that required in finding the bounds of chromatic number $\chi(G)$ of a graph and justify why we color complements of graphs.

Theorem 1. *If the maximum vertex degree of a graph G is Δ, then $\chi(G) \leq \Delta + 1$.*

Theorem 1 provides a preliminary insight on how many colors we use and determines if these colors are helpful in visualizing the network.

Theorem 2. *Given the edge probability p, let $b = 1/(1-p)$. For constant p and constant $\epsilon > 0$, almost every random graph G with p satisfies*

$$(1/2 - \epsilon)n/\log_b n \leq \chi(G) \leq (1/2 + \epsilon)n/\log_b n.$$

Readers can refer to West et al. [26, Chapter 8] for the details of the proofs.

If a network comes from a random network assumption, the chromatic number is governed by the edge probability p. If p is high, additional colors are needed for graph coloring. On the other hand, if p is low, only a few colors are needed to complete the graph coloring. According to Theorems 1 and 2, if a network is partitioned into few groups for the purpose of easier observation, we need to define a low edge probability for obtaining a low $\chi(G)$.

However, the edge probability is usually invariant. Instead, we consider the kth power of G, which is the simplest graph G^k with the vertex set $V(G)$ and the edge set $\{(v_i, v_j) | d_G(v_i, v_j) \leq k\}$. Let M be the adjacent matrix of G. The kth power of M refers to the number of walks of length k in G. It is expressed as

$$M^k = \underbrace{M \times \ldots \times M}_{k \text{ times}},$$

and its entry M_{ij}^k refers to the number of walks of length k from v_i to v_j. Thus, the adjacent matrix of graph G^k is defined as Σ^k and its entry is

$$[\Sigma]_{ij}^k = \begin{cases} 1 \text{ if } \sum_{d=1}^k [M]_{ij}^d > 0 \ \forall i \neq j \\ 0 \text{ otherwise.} \end{cases}$$

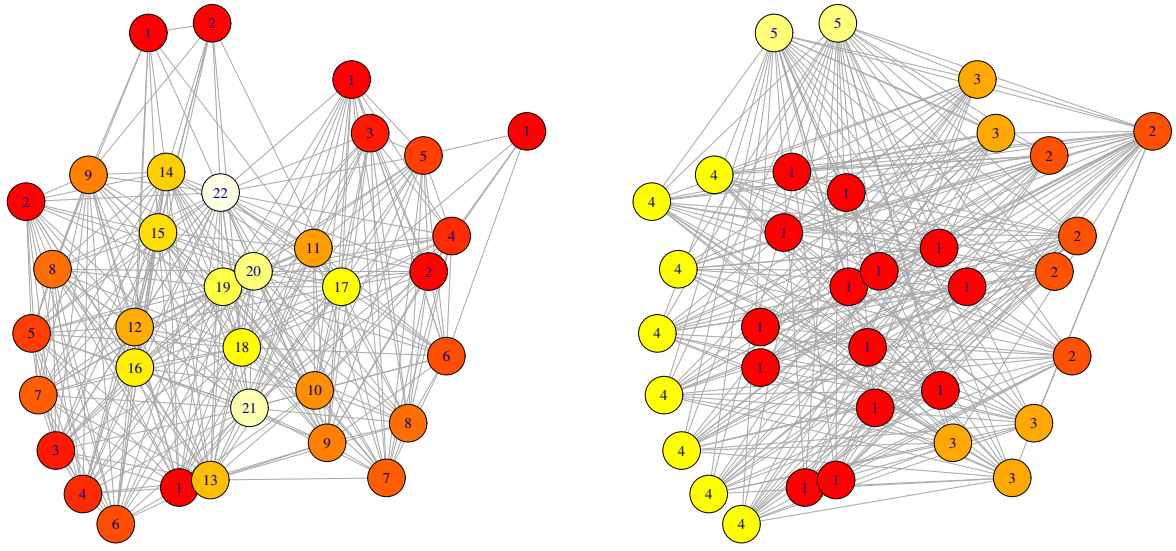
Based on the adjacent matrix of G^k, the G^k graph can be created quickly.

A high edge probability can be obtained based on this manner. This motivates us to use the complement instead of the original graph. If a long distance k is selected, a highly connected network G^k is resulted. On the contrary, the complement graph of G^k, defined as $\overline{G^k}$, is a network with a low edge connection and $\overline{G^k}$ leads to a simple observation. The complement $\overline{G^k}$ can be created by its adjacent matrix Λ, which is obtained by switching the $(0, 1)$ off-diagonal elements in the adjacent matrix of G^k, $[\Sigma]^k$. The element (i, j) of this adjacent matrix of $\overline{G^k}$ is expressed as

$$[\Lambda]_{ij}^k = \begin{cases} 1 \text{ if } [\Sigma]_{ij}^k{=}0, \ \forall i \neq j \\ 0 \text{ otherwise.} \end{cases}$$

Applying the greedy vertex coloring algorithm, two connected vertices in $\overline{G^k}$ represent that they are assigned to different colors and their distance is longer than k. Oppositely, the distance of two vertices with the same color is within k. This is just the definition of $n-$clique with $n = k$. Thus, the partitions obtained by our algorithm are $n-$cliques.

Consider the karate network with distance 2 as an example. Fig. 3 (a) shows the G^2 graph of karate network and Fig. 3 (b) is its complement $\overline{G^2}$. We see little information when the network has a large chromatic number. On the contrary, its complement is clear to see some possible partitions.



(a) Colored Graph G^2 (b) The Complement of G^2

Figure 3: Graph coloring of karate network with distance 2

Although the number of colors is reduced to 5 in $\overline{G^2}$, it is not clear enough to observe the network structure due to the large number of edges. We simplify the network into small partitions, where each one has its vertices with the same color. Only the number of vertices and the number of edges between each pair of colors are recorded. Notice that more edges in the complement suggest that the connection of two partitions is weaker, which is shown in the following table and figure.

Table 1: Information of Partition Graph of $\overline{G^2}$ of karate network

G_1	G_2	$\lvert V(G_1)\rvert$	$\lvert V(G_2)\rvert$	$\lvert E(G_1, G_2)\rvert$
1	2	14	5	37
1	3	14	5	23
1	4	14	8	25
1	5	14	2	15
2	3	5	5	5
2	4	5	8	40
2	5	5	2	10
3	4	5	8	40
3	5	5	2	10
4	5	8	2	13

G_1 and G_2 are the sets of vertices in which all vertices have the same colors/labels in the partition graph. $\lvert V(G_1)\rvert$ and $\lvert V(G_2)\rvert$ are numbers of vertices in G_1 and G_2, respectively. $\lvert E(G_1, G_2)\rvert$ represents the number of edges between these two sets.

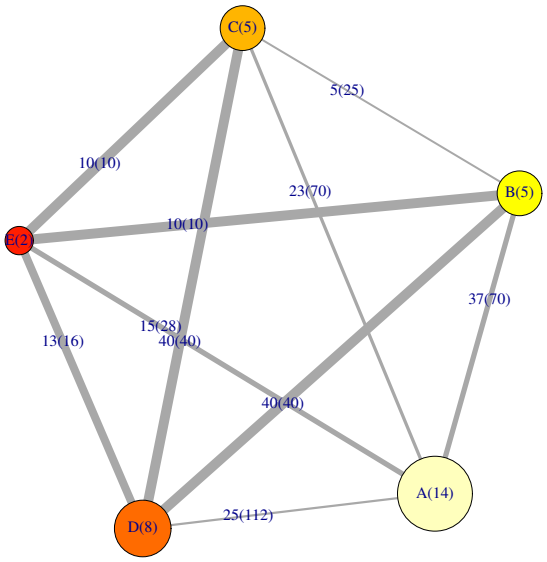
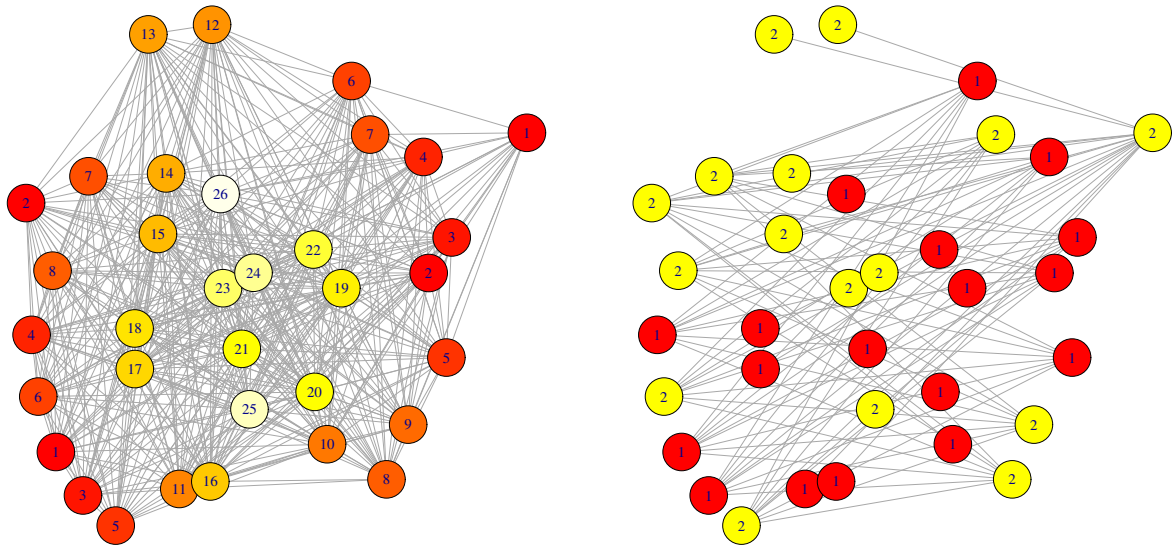


Figure 4: Partition plot of $\overline{G^2}$ of karate network

The number of edges in the complement graph is monotone decreasing as the distance increases, which is an important property of the complement of a graph. Again, consider the karate network with distance 3 as an example. Fig. 5 (a) shows the G^3 graph of karate network and Fig. 5 (b) is its complement. Fig. 5 is comparably more informative than Fig. 3. In this example, the complement graph $\overline{G^3}$ of the karate network is a bipartite graph, in which the vertices are partitioned into two disjoint parts. That is, the distance of any two adjacent vertices in G^3 is at most 3, and the distance of any two adjacent vertices in $\overline{G^3}$ is higher than 3. Clearly, it is easy to point out which vertices are distanced higher than 3.

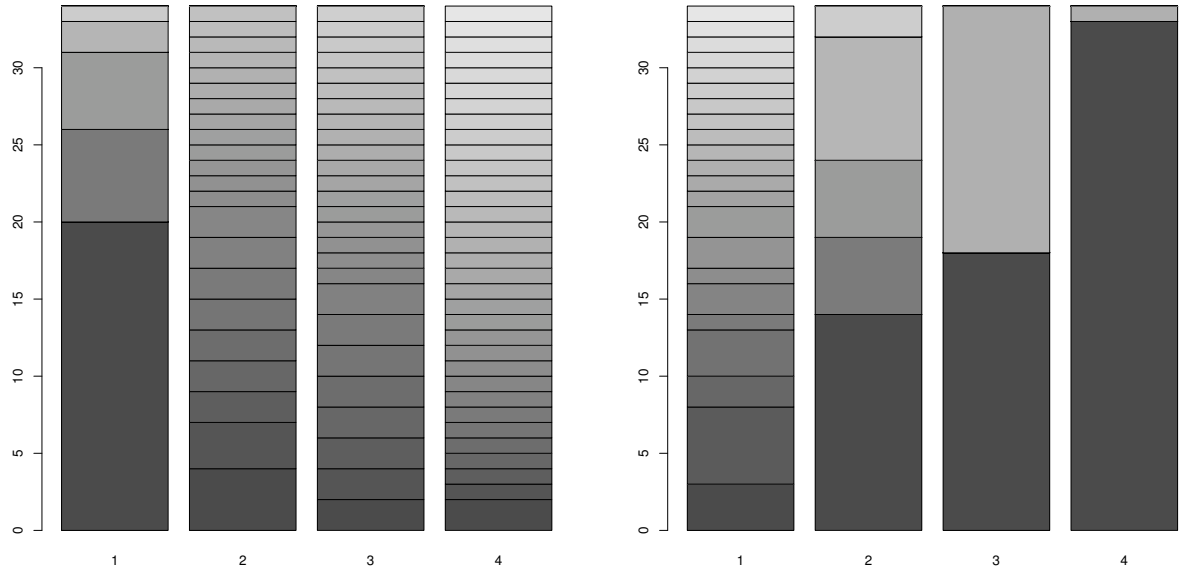


(a) Graph Coloring of G^3 (b) Complement Graph of $\overline{G^3}$

Figure 5: Graph coloring of karate network with distance 3

By considering the distances from 1 to 4, we collect the information of all coloring labels of both original graphs and their complements. These graphs are summarized as stacked bar plots (Fig. 6) in which each segment represents the number of vertices in each color group. According to the changes of

stacks with different distances, it is helpful to visualize the move of the partition patterns. Apparently, we can see the changes according to the increasing distances and the numbers of vertices in the graphs and their complements. These results correspond to Theorems 1 and 2. The edge probability gets higher when the distance increases. Thus, strong connections lead to more colors for graph coloring in the original graphs. On the contrary, only few colors are needed for coloring the complements with higher distances.



(a) Label Frequencies of G^ks (b) Label Frequencies of $\overline{G^k}$s

Figure 6: Graph coloring results of karate network by bar charts with distances from 1 to 4, where each segment represents the number of vertices in each color group.

3.2 Statistics of Labels

The labels of graph coloring imply much information, but this information is implicit unless these labels undergo some transformations. We focus on the cluster pattern in this study. In practice, the clustering coefficient [2] is an important quantity that measures the cluster relation of neighbors in networks. The original clustering coefficient of a vertex v_i is defined as

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_G(v_i), e_{jk} \in E(G)\}|}{|N_G(v_i)|(|N_G(v_i)| - 1)}, \tag{1}$$

where $N_G(v_i)$ is the set of neighbors of vertex v_i, and $|N_G(v_i)|$ is the cardinality of $N_G(v_i)$. This quantity measures which nodes tend to cluster together by discussing the edges between their neighbors. Since our approach also provides the information of clustering, we are interested in knowing whether the labels can deliver this information by considering the relations of vertices instead of edges.

Although we cannot restrict the type of graph coloring, labeling different colors on two adjacent vertices always holds. Therefore, the labels of neighbors of each vertex can be discussed. Since an edge leads to two vertices with two different colors with respect to coloring the original graph, the different colors are considered to represent a kind of relation between two vertices. We call the different colors of vertices the "*variation*". The small variation of labels indicates that the neighbors of vertex have similar colors and also implies that the connection between neighbor is weak. Two vertices with different colors do not guarantee their adjacencies. However, a latent mechanism might exist and label them in different colors according to the greedy coloring algorithm. Thus, the variation of labels is just a comparative quantity that tells the relations among vertices.

In order to compute the variation, we use the variation of categorical data provided in Light and Margolin [27]. The variation of categorical data is expressed as

$$\rho = \frac{1}{2n} \sum_i \sum_j d_{ij},$$

where $d_{ij} = d(X_i, X_j)$ is defined as

$$d_{ij} = \begin{cases} 1 & \text{if } X_i \text{ and } X_j \text{ are labeled as different categories} \\ 0 & \text{if } X_i \text{ and } X_j \text{ are labeled as the same category.} \end{cases}$$

We adjust this equation to network expression as

$$V_i = \frac{1}{\binom{|N_G(v_i)|}{2}} \sum_{\substack{j \neq k; \\ \forall v_j, v_k \in N_G(v_i)}} d_{jk}, \tag{2}$$

where d_{jk} is defined as

$$d_{jk} = \begin{cases} 1 & \text{if } v_j \text{ and } v_k \text{ are labeled as different colors in the original graph } G \\ 0 & \text{if } v_j \text{ and } v_k \text{ are labeled as the same color in the original graph } G. \end{cases}$$

On the other hand, when the coloring result of the graph complement is discussed, the same color obviously represents that two vertices belong to the same group. We consider the similarity of the neighbors by adjusting the variation measure. We define the similarity of labels as

$$S_i = \frac{1}{\binom{|N_G(v_i)|}{2}} \sum_{\substack{j \neq k; \\ \forall v_j, v_k \in N_G(v_i)}} s_{jk}, \tag{3}$$

and s_{jk} has the contrary definition compared with d_{jk}, i.e.,

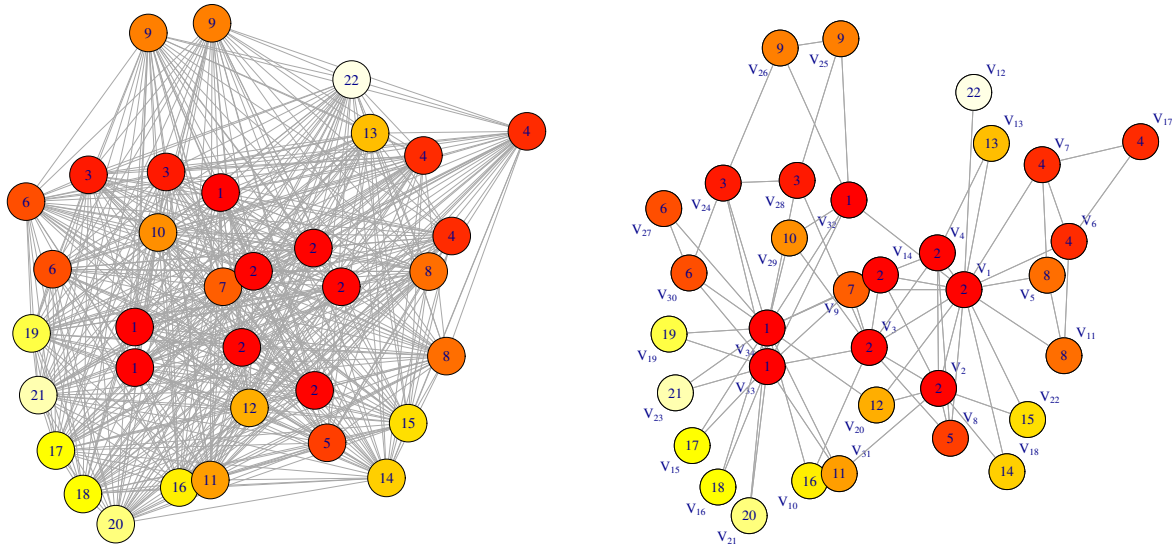
$$s_{jk} = \begin{cases} 1 & \text{if } v_j \text{ and } v_k \text{ are labeled as the same color in } \overline{G^1} \\ 0 & \text{if } v_j \text{ and } v_k \text{ are labeled as different colors in } \overline{G^1}. \end{cases}$$

The $N_G(v_i)$ is obtained under G instead of $\overline{G^1}$ although the coloring result is obtained under $\overline{G^1}$. Since the clustering pattern of the original graph is our focus, we only discuss about how the different representations of vertices affect the measures of clustering pattern.

Compared with the variation of labels, which is a conflicting measure to define the relation of neighbors, the similarity is very conservative. Recall that variation considers the vertices with different colors as 1, but such vertices possibly have no edges. The similarity considers the vertices with the same colors as 1. However, the vertices with different color possibly have edges. These differences lead to the value of clustering coefficient locating between similarity and variation, i.e., $S_i \leq C_i \leq V_i$. From this aspect, we can not only visualize networks by graph coloring, but also compute the information of the coloring labels. Since the graph coloring can produce flexible outputs, more measures besides the variation and similarity can be created if other coloring procedures are applied.

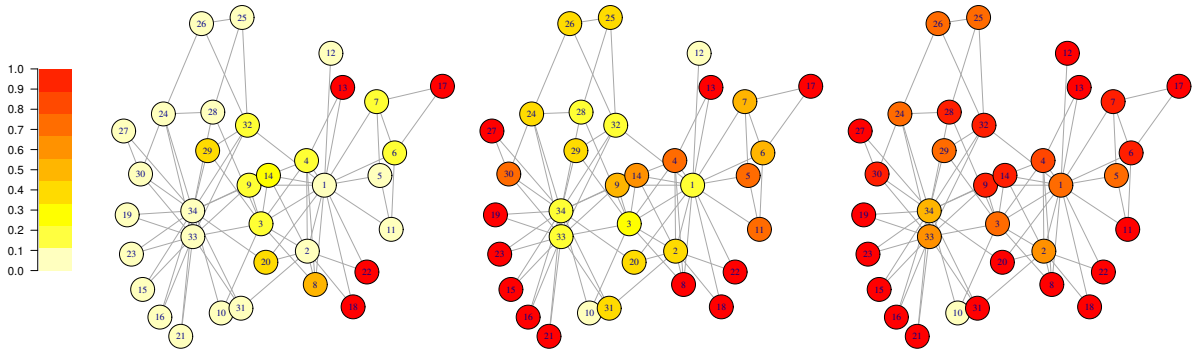
The karate network is applied to compute these statistics, including similarity, variation, and clustering coefficient. $\overline{G^1}$ of karate network is provided in Fig. 7. Since the original karate network contains few edges, its complement shows an opposite pattern that needs to use many labels for coloring. We expect to get low similarities because there are few members in each label.

We present different patterns of similarity, variation, and clustering coefficient in Fig. 8. Most peripheral vertices have high values and center vertices have lower values respectively in clustering coefficient and variation. On the other hand, most vertices of the left part of the network are 0 in similarity, which implies that the left part seems to have no cluster pattern in this conservative criterion, but the middle vertices have some relations with the other vertices.



(a) Graph Coloring of $\overline{G^1}$ (b) Labels of $\overline{G^1}$ in G^1 layout

Figure 7: $\overline{G^1}$ of karate network with graph coloring and the coloring result represented in the G^1 layout



(a) Similarity (b) Clustering Coefficient (c) Variation

Figure 8: Statistics of Labels of karate network

3.3 Statistics of Edges

We consider the number of edges as an additional usage of the complement of a graph. In each complement $\overline{G^k}$, the edges represent the distance between pairs of vertices that is larger than k. Based on these graphs, the number of edges is collected to verify the *random model assumption*.

If the distances of the shortest paths are of interest, Blondel et al. [28] provided the density function of the distance of path by assuming the hidden variables follow the Poisson random graphs. If a network is a Poisson random graph with homogeneous connection probability p, the probability of the shortest path between any pair of vertices greater than distance k is expressed as

$$S(k) = Prob(D > k) = \exp[-\frac{(np)^k}{n}], \tag{4}$$

where n is the number of vertices.

Each edge in $\overline{G^k}$ represents that two vertices are distanced higher than k. Thus, the density of the shortest paths can be estimated by the empirical density, i.e.,

$$\hat{S}(k) = \frac{2|E(\overline{G^k})|}{|V(G)|(|V(G)|-1)}. \tag{5}$$

Similar to the Q-Q plot provided for goodness of fit for normal assumption, the scatter plot is drawn with $\hat{S}(k)$ vs. $S(k)$. If the scatter plot tends to fall on a straight (45-degree) line, a network is likely to be a Poisson random graph.

Consider karate network as an example. We have to create all graphs with different distances and then generate their complements. Based on the above equations, the theoretical densities and the empirical densities of the shortest paths are obtained. The results are given in Table 2 and Figure 9. The similarity between the estimated and theoretical values suggests that this network is possibly a Poisson random graph. Although we cannot provide a statistical testing for the goodness of fit for the results, we can glimpse the possibility of randomness from the provided information.

Table 2: Information of Density Estimation

Distance k	$\hat{S}(k)$	$S(k)$
1.00	0.86	0.87
2.00	0.39	0.53
3.00	0.14	0.05
4.00	0.01	0.00

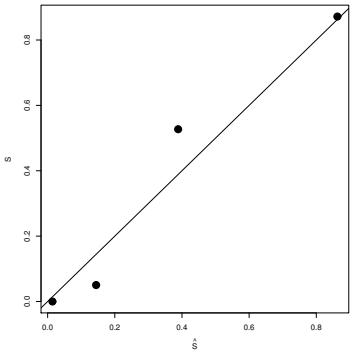


Figure 9: Density of Karate network

4 Partitions of Other Networks

In this section, we provide other classical examples of networks by applying our proposed method to the partition purpose. That is, coloring the complements of graphs with different distances. In order to provide a unified output, a different framework from Fig. 4 for the karate network is applied to visualize the graph partitions. Since we are interested in the relationship between all partitions, the partition graph is represented by a circular embedding. Thus, we can check all the relations among different partitions. In addition, the sizes of vertices and widths of edges are set to be equal but the labels of vertices represent the number of each group and the types of edges show the strength of connection.

There are four types of edges representing the group connections according to the inter-grouping probability. Suppose a complement of a graph is colored into several groups in which each vertex has the same color. Given a pair of groups G_s and G_t, the inter-grouping probability is defined as

$$p_{st} = \frac{|E(G_s, G_t)|}{|V(G_s)||V(G_t)|}, \tag{6}$$

where $|V(G_s)|$ and $|V(G_t)|$ are the number of vertices in G_s and G_t respectively, and where $|E(G_s, G_t)|$ is the number of edges between these two groups. Note that the solid dark line (—) represents a strong distance between two groups and its probability belongs to the interval (0.75-1); whereas the dash dark line (– –) represents a less strong distance between two groups and its probability belongs to the interval (0.5 -0.75); and the solid gray line (—) represents a weak distance between two groups and its probability belongs to the interval (0.25-0.5); and finally the dash gray line (– –) represents a very weak distance between two groups and its probability belongs to the interval (0.0-0.25). Here, a strong distance means that the two groups cannot be reached easily one from another.

In each case, the bar chart of different distances is performed with the modularity measure [11] at top, which is defined as

$$Q = \frac{1}{2|E(G)|} \sum (e_{ij} - \frac{k_i k_j}{2|E(G)|}) \delta(v_i, v_j), \tag{7}$$

where k_i and k_j are the degrees of v_i and v_j, e_{ij} is 1 if v_i and v_j are adjacent, and 0 otherwise; and $\delta(v_i, v_j)$ is 1 if v_i and v_j are labeled as the same color, and 0 otherwise. In order to provide a clear partition, we select the partition with maximum modularity except distance 1. Although the partitions have a high modularity with distance 1 in some cases, we do not choose them because too many labels are not helpful for clear visualization. Our result shows an uncertainty on the goodness of the suggested group size, but we suggest this method as an initial stage to view the structure of network if the distance between vertices is the main concern.

4.1 Dolphin Network

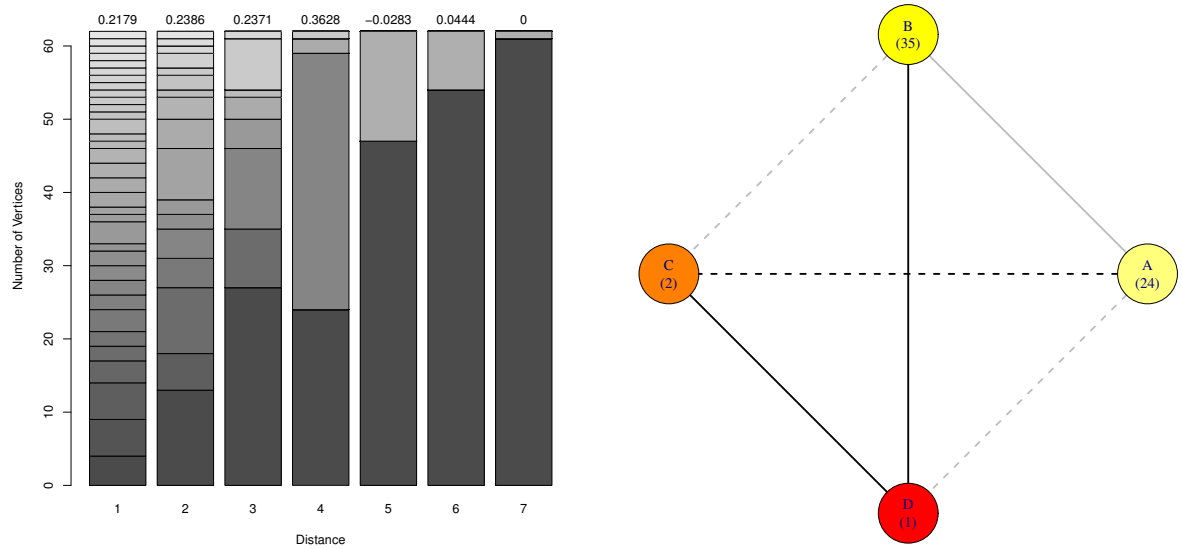
The dolphin network compiled by Lusseau et al. [29] and Lusseau [30] is an undirected social network of frequent associations between dolphins in a community living off Doubtful Sound, New Zealand. This network contains 62 vertices (dolphins) and 159 edges, which means that dolphin pairs were observed to have statistically significant frequent association. The partition results are shown in Fig. 10. One particular feature of this network is that the maximum distance (*diameter*) between dolphin pairs is 7. Since this network only contains 62 vertices, the diameter of this network is high and implicit, which means that the connections between dolphins are low. When the distance 4 ($Q = 0.3628$) is chosen, we obtain the maximum modularity and 4 partitions. The maximum group size is 24 only and this network is mainly partitioned into 2 major groups (A and B in Fig. 10 (b)) that have small distances (gray solid line).

4.2 Les Misérables Network

Les Misérables Network compiled by Knuth et al. [31] contains the weighted network of co-appearances of characters in Victor Hugo's novel "Les Misérables". 77 vertices (characters), and 254 edges are shown to represent the co-appearance of two characters in one or more scenes. The partition result is shown in Fig. 11. The modularity from the bar chart shows a big drop from distance 1 ($Q = 0.3039$) to distance 2 ($Q = 0.1764$) , so is the number of groups. When the distance 2 is selected, this network is partitioned into 11 groups, and all groups have strong distances (black solid line). This result matches our knowledge of this famous novel. That is, the characters belonged to different scenes and eras.

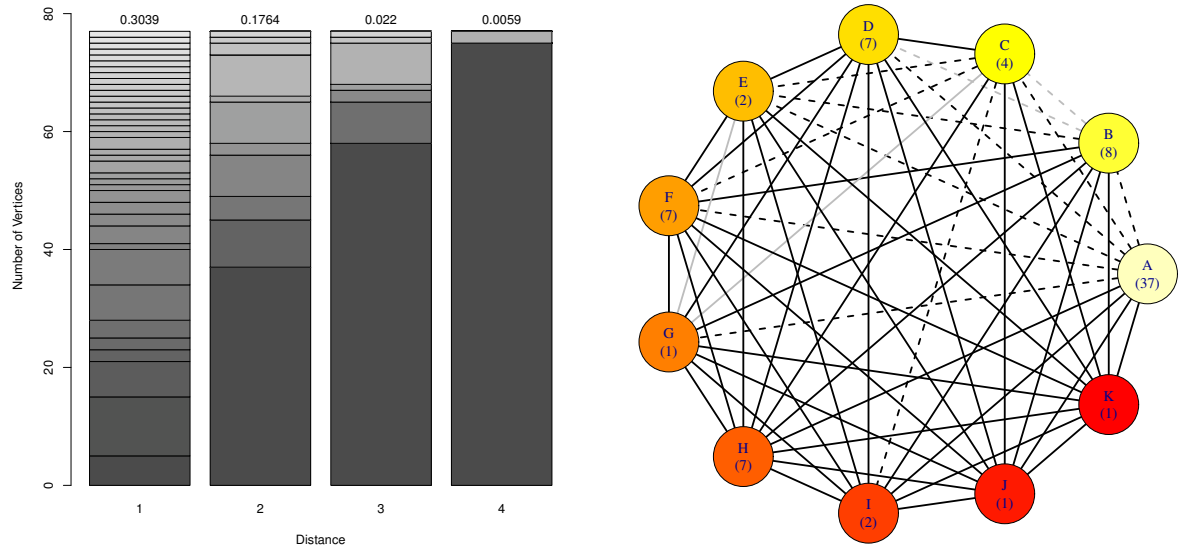
4.3 Football Network

The football network compiled by Girvan and Newman [32] contains the network of American football games between Division I colleges during regular season in Fall 2000. There are 115 vertices (teams)



(a) Bar Chart of Labels (b) The Suggested Partitions with Distance 4

Figure 10: Visualization of dolphin network by the coloring of complements of graphs. (a) The stacked bar chart where each segment represents the number of vertices in each color group. (b) The partition with distance 4.

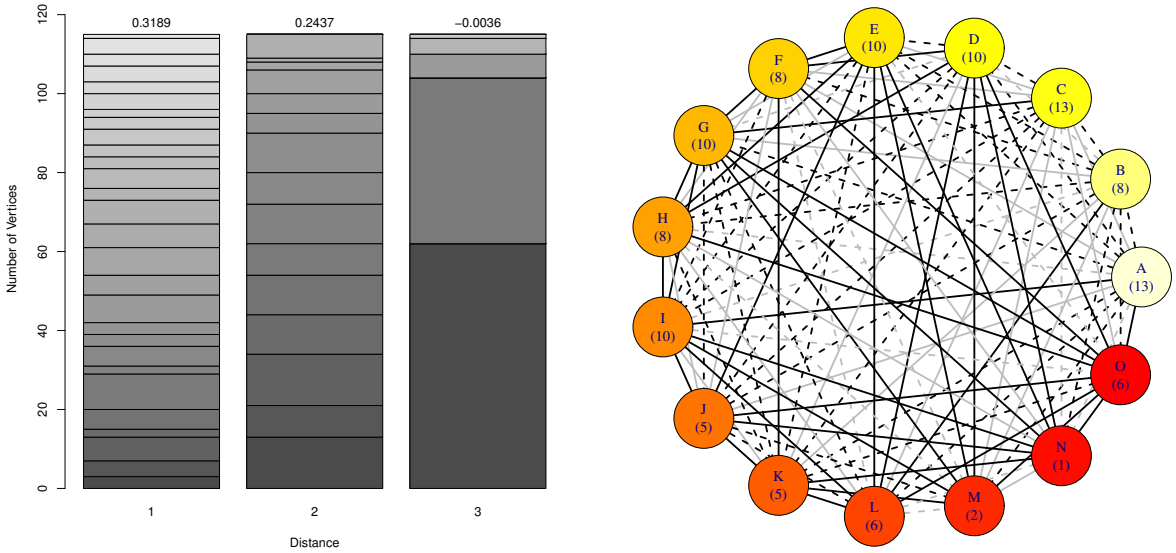


(a) Bar Chart of Labels (b) The Suggested Partitions with Distance 2

Figure 11: Visualization of Les Misérables network by the coloring of complements of graphs. (a) The stacked bar chart where each segment represents the number of vertices in each color group. (b) The partition with distance 2.

that belong to 12 different conferences, and 616 edges show that some teams have one or more games during this regular season. The partition result is shown in Fig. 12. Unlike the bar chart in Fig. 11, the modularity decreases slightly from distance 1 ($Q = 0.3189$) to distance 2 ($Q = 0.2467$), and the number of groups slightly drops at the same time. This implies that the structures between the Les Misérables network and the football network are different, but our plots cannot show the details of the differences.

When the distance 2 is selected, this network is partitioned into 15 groups, and most groups have similar group sizes around 10 except some groups with few members. In fact, this network is comprised by 12 college football leagues, which is close to our finding.



(a) Bar Chart of Labels (b) The Suggested Partitions with Distance 2

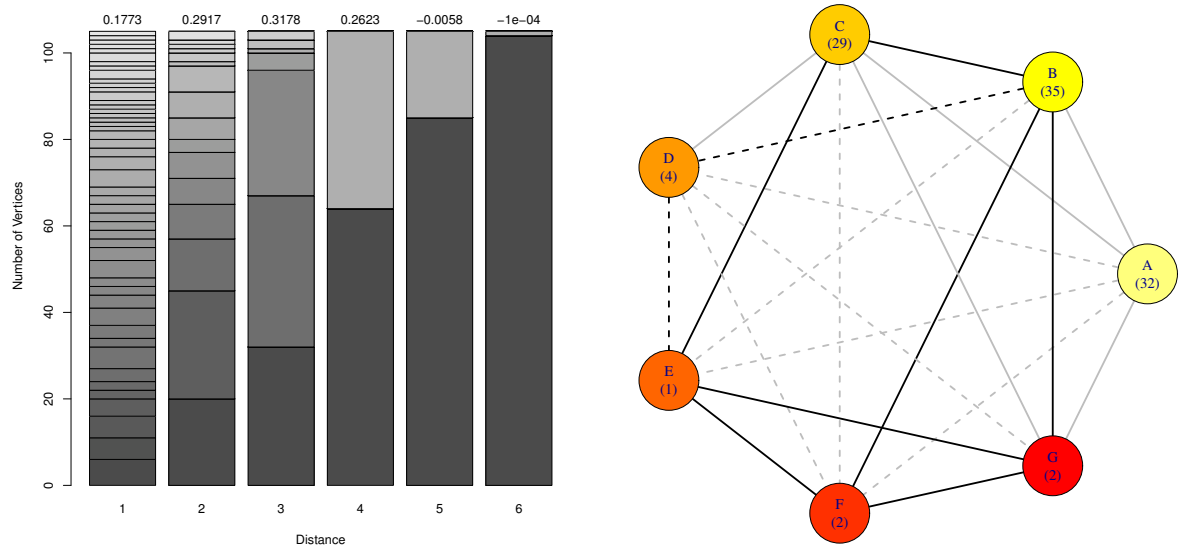
Figure 12: Visualization of football network by the coloring of complements of graphs. (a) The stacked bar chart where each segment represents the number of vertices in each color group. (b) The partition with distance 2.

4.4 Political Book Network

The political book network contains the network of book sale about US politics published around the time of the 2004 presidential election and sold by the online bookseller. This network contains 105 vertices (books), and 441 edges are shown to represent that two books are bought at the same time at least once. The network was compiled by V. Krebs but is unpublished, yet can be found on Krebs' website http://www.orgnet.com/. The partition result is shown in Fig. 13. A great improvement is found in both modularity and group size with distance 3 ($Q = 0.3178$). Only 7 groups are recorded and are mainly composed by three groups (A, B, and C) which are similar to the number of original categories of these books (liberal, conservative, and neutral).

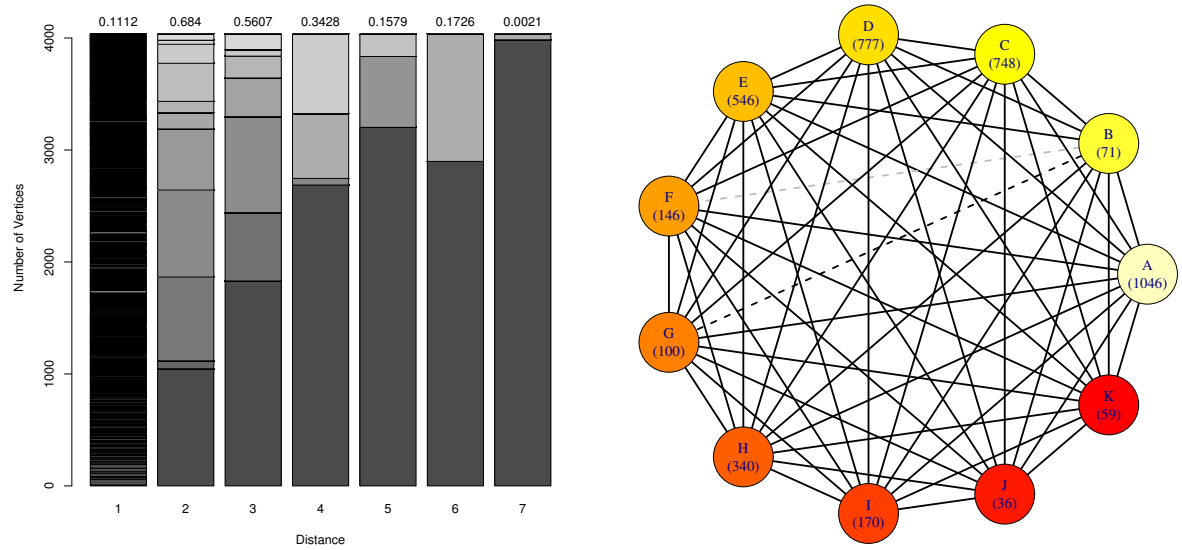
4.5 Facebook Network

The Facebook network complied by Leskovec and Mcauley [33] is obtained from 10 ego-networks, consisting of 193 circles and 4,039 users. This network comprises 4,039 vertices (people), and 88,234 edges are shown to illustrate that two friends of an ego are also friends to each other. The partition result is shown in Fig. 14. This example demonstrates the utility of our proposed method. Despite the fact that this is not a big network, one cannot look at its whole structure quickly. By applying our method, this network is quickly partitioned into 11 groups with distance 2 and got a high modularity 0.684. In fact, this network is compiled by combining 10 ego-networks. Our result is very close to the real one. Note that the first group in the bar chart does not show one black color only. Instead, too many labels result in out of capacity for gray colors. Besides, it only takes around 2 minutes to execute the coloring of complements of graphs with all 7 different distances in our personal computer (Intel Core i7-4770 CPU 3.40GHz).



(a) Bar Chart of Labels (b) The Suggested Partitions with Distance 3

Figure 13: Visualization of book network by the coloring of complements of graphs. (a) The stacked bar chart where each segment represents the number of vertices in each color group. (b) The partition with distance 3.



(a) Bar Chart of Labels (b) The Suggested Partitions with Distance 2

Figure 14: Visualization of Facebook network by the coloring of complements of graphs. (a) The stacked bar chart where each segment represents the number of vertices in each color group. (b) The partition with distance 2.

5 Discussion and Conclusion

It is difficult to quickly glimpse a large-scale network by a limited plot region within a short limit of time. In this study, we summarize a large network into few partitions by the greedy vertex coloring approach, which is an efficient algorithm and has been verified to run $O(n+m)$ time according to their interactive distances. The basic notion of graph coloring is to force two connected vertices to be assigned to different

colors/labels. By considering the complement, two connected vertices become unconnected and close vertices have high probabilities to be assigned to the same colors in this regime. We can treat different colors as different groups to glimpse the structure of a network. In general, a network can be partitioned into n-cliques by the graph coloring approach. Thus, we can recognize graph partition and cluster patterns in a simple way. Based on the labels of vertices after coloring, we get further information by summarizing these labels. Some famous networks are demonstrated to show the abilities of our provided method. If the distances between pairs of vertices are the main concern, our method is a useful tool to summarize a large network.

Although we leave an open question whether the partition reaches the chromatic number via the greedy coloring procedure, the partitions have provided useful clues to future researches. For example, if finding the best communities is of interest, one can start from the suggested partitions obtained by our algorithm and create some add or delete algorithms to modify the groups. Even though other coloring methods can provide better coloring results, most of them take more time on coloring a graph. Using time-consuming algorithms contradicts our initial purpose of computing efficiency. Therefore, we look forward to constructing and searching for better methods to execute graph coloring.

On the other hand, during the demonstration to the best partitions, the modularity is suggested as the selection criterion. However, this criterion cannot provide the information of the number of groups. If we think parsimony partitions are a good representation, a modified modularity is needed to take this information into account. A modified criterion is considered in our future work. Furthermore, if we care about other information rather than community, other criteria are also needed.

Note that the methods and statistics provided in this study contain no statistical inferences. However, it does not imply that no statistical inference can be constructed directly. For example, the number of edge in different complements of graphs can be used to test the goodness of fit of random model. The number of edge between two different groups can also be a test statistic to verify if two groups are significantly split. It takes more effort to improve this method in the future.

Acknowledgments. The authors would thank Ms. Wendy Tzu-Wen Kao and Dr. Martin Tshishimbi Wa Lukusa for their helps in improving English of this paper. This work was supported by (a) Career Development Award of Academia Sinica (Taiwan) grant number 103-CDA-M04, (b) Ministry of Science and Technology (Taiwan) grant numbers 104-2118-M-001-016-MY2 and 105-2118-M-001-007-MY2, and (c) Thematic Research Program of Academia Sinica (Taiwan) grant number AS-103-TP-C03.

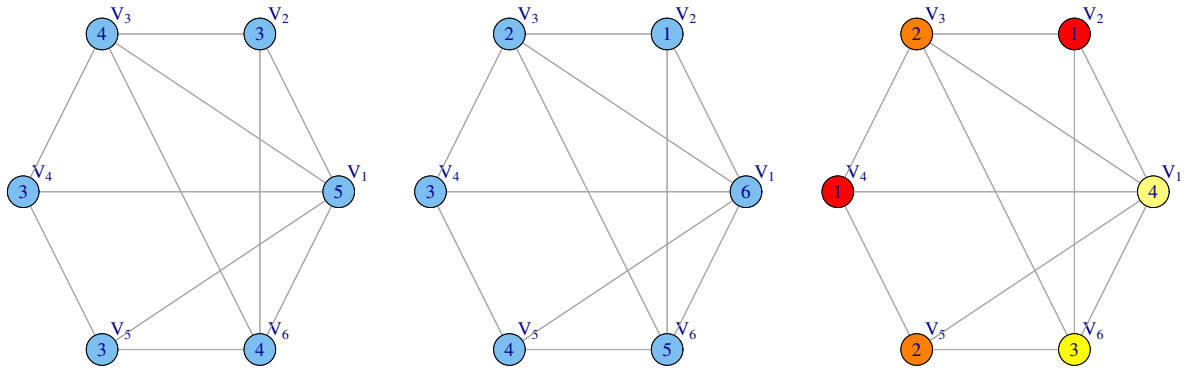
References

1. P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.
2. D. J. Watts and S. H. Strogatz, "Collective dynamics of ąěsmall-worldąęnetworks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.
3. A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.
4. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.
5. M. Bastian, S. Heymann, M. Jacomy *et al.*, "Gephi: an open source software for exploring and manipulating networks." *ICWSM*, vol. 8, pp. 361–362, 2009.
6. G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal, Complex Systems*, vol. 1695, no. 5, 2006.
7. T. M. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and experience*, vol. 21, no. 11, pp. 1129–1164, 1991.
8. G. Karypis and V. Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs," *SIAM Journal on scientific Computing*, vol. 20, no. 1, pp. 359–392, 1998.
9. S. Arora, S. Rao, and U. Vazirani, "Expander flows, geometric embeddings and graph partitioning," *Journal of the ACM (JACM)*, vol. 56, no. 2, p. 5, 2009.

10. S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.
11. M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.
12. P. J. Bickel and A. Chen, "A nonparametric view of network models and newman–girvan and other modularities," *Proceedings of the National Academy of Sciences*, vol. 106, no. 50, pp. 21 068–21 073, 2009.
13. M. S. Handcock, A. E. Raftery, and J. M. Tantrum, "Model-based clustering for social networks," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 170, no. 2, pp. 301–354, 2007.
14. N. A. Heard, D. J. Weston, K. Platanioti, and D. J. Hand, "Bayesian anomaly detection methods for social networks," *The Annals of Applied Statistics*, vol. 4, no. 2, pp. 645–662, 2010.
15. L. Tang and H. Liu, "Community detection and mining in social media," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 1–137, 2010.
16. D.-Z. Du and P. M. Pardalos, *Handbook of combinatorial optimization: supplement*. Springer Science & Business Media, 1999, vol. 1.
17. F. T. Leighton, "A graph coloring algorithm for large scheduling problems," *Journal of research of the national bureau of standards*, vol. 84, no. 6, pp. 489–506, 1979.
18. P. Hell and J. Nešetřil, "On the complexity of< i> h</i>-coloring," *Journal of Combinatorial Theory, Series B*, vol. 48, no. 1, pp. 92–110, 1990.
19. P. Briggs, K. D. Cooper, and L. Torczon, "Improvements to graph coloring register allocation," *ACM Transactions on Programming Languages and Systems (TOPLAS)*, vol. 16, no. 3, pp. 428–455, 1994.
20. M. Thorup, "All structured programs have small tree width and good register allocation," *Information and Computation*, vol. 142, no. 2, pp. 159–181, 1998.
21. A. Kosowski and K. Manuszewski, "Classical coloring of graphs," *Contemporary Mathematics*, vol. 352, pp. 1–20, 2004.
22. W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, pp. 452–473, 1977.
23. A. Hertz and D. de Werra, "Using tabu search techniques for graph coloring," *Computing*, vol. 39, no. 4, pp. 345–351, 1987.
24. D. S. Johnson, C. R. Aragon, L. A. McGeoch, and C. Schevon, "Optimization by simulated annealing: an experimental evaluation; part ii, graph coloring and number partitioning," *Operations research*, vol. 39, no. 3, pp. 378–406, 1991.
25. P. Galinier and J.-K. Hao, "Hybrid evolutionary algorithms for graph coloring," *Journal of combinatorial optimization*, vol. 3, no. 4, pp. 379–397, 1999.
26. D. B. West *et al.*, *Introduction to graph theory*. Prentice hall Upper Saddle River, 2001, vol. 2.
27. R. J. Light and B. H. Margolin, "An analysis of variance for categorical data," *Journal of the American Statistical Association*, vol. 66, no. 335, pp. 534–544, 1971.
28. V. D. Blondel, J.-L. Guillaume, J. M. Hendrickx, and R. M. Jungers, "Distance distribution in random graphs and application to network exploration," *Physical Review E*, vol. 76, no. 6, p. 066101, 2007.
29. 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.
30. D. Lusseau, "The emergent properties of a dolphin social network," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. Suppl 2, pp. S186–S188, 2003.
31. D. E. Knuth, D. E. Knuth, and D. E. Knuth, *The Stanford GraphBase: a platform for combinatorial computing*. Addison-Wesley Reading, 1993, vol. 37.
32. M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.
33. J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in neural information processing systems*, 2012, pp. 539–547.

Appendix: Toy Example of Greedy Coloring Algorithm

Here is a toy example to illustrate the greedy coloring approach and its procedures based on Algorithm 1. We construct a synthetic network with 6 vertices and 11 edges. Fig. 15 (a) shows the degrees of this network. In our algorithm, we start the coloring at the vertex having the smallest degree, and then sequentially color the graph from this vertex. Fig. 15 (b) shows the order for coloring vertices based on our provided algorithm. We select the node with the smallest vertex ID if there are more than one adjacent vertices that have the same smallest degree, so V_2 is the first node to be assigned a label. According to the order, this network can be successively colored and the coloring result is shown in Fig. 15 (c). Besides, nodes $\{V_1, V_2, V_5, V_6\}$ form a clique. Thus, the chromatic number of this graph is 4 and our algorithm reaches the chromatic number in this example.



(a) Degree (b) Coloring Order (c) Labels of Colors

Figure 15: Graph coloring for the synthetic network where the labels inside the vertices are the node degrees.

