Signed Graph Clustering Using k-core Decomposition and Signed Motifs: A Comprehensive Approach

Sun, Shurong
Graduate School of Economics, Kyushu University

https://doi.org/10.15017/7325754

出版情報：経済論究. 176, pp.49-72, 2023-11-30. 九州大学大学院経済学会
バージョン：
権利関係：

Signed Graph Clustering Using *k*-core Decomposition and Signed Motifs: A Comprehensive Approach

Shurong Sun †

Abstract

Signed graphs, which encompass both positive and negative relationships, present unique challenges in network analysis. Many real-world networks can be represented by signed graphs with positive and negative links. As the size and complexity of real-world networks continue to grow, the search for efficient clustering methods for signed graphs remains an active and important area of research. In this study, we introduce a comprehensive approach for clustering signed networks, which combines k-core decomposition with signed motifs. To avoid the high computational cost of traditional spectral algorithms for clustering large-scale networks, our proposed method initiates with k-core decomposition that identifies those subgraphs that are able to represent the underlying network structure. Motifs have recently made great improvements in many graph analysis tasks. However, most of the work has focused on unsigned networks. Considering the great potential of using motifs to analyze complex network structures, we combine these two methodologies to explore the signed network clustering. Through a series of numerical experiments on real-world datasets, we demonstrate that our method outperforms the clustering method without k-core decomposition, offering improved accuracy and interpretability. Compared with the algorithm without k-core decomposition, our method significantly reduces the computational time. Furthermore, we also discuss the variation in the accuracy of the clustering for each real-world dataset as the number of clusters increases. Our results show that with respect to choosing the appropriate number of clusters k is strongly influenced by the structure of the network itself, and that finding the right number of clusters is crucial for achieving better clustering accuracy.

Keywords: k-core Decomposition, Signed Motifs, Complex Networks, Signed Graph Clustering

1 Introduction

1.1 Basics of Graph Theory

A graph is a pair $G=(V, E)$, where V is a set whose elements are called vertices, and E is a set of

† Graduate School of Economics, Kyushu University

paired vertices, whose elements are called edges. In this paper, graph and network can be assumed to be synonymous. A weighted graph or a network is a graph in which a number (the weight) is assigned to each edge. Graphs can be weighted or unweighted. Compare to an unsigned graph shown in Figure 1, for a signed graph, edges are assigned $+1$ or -1 weights, as illustrated in Figure 2.

In graph theory, a dense graph is a graph in which the number of edges $|E|$ is close to the maximal number of edges. The opposite, a graph with only a few edges, is a sparse graph. The graph density of simple graphs (i.e., graphs without loops and multiple edges) is defined to be the ratio of the number of edges $|E|$ with respect to the maximum possible edges. For an undirected simple graphs, the graph density D is

$$D=\frac{2|E|}{|V|(|V|-1)} \tag{1}$$

where E is the number of edges and V is the number of vertices in the graph. The maximum number of edges for an undirected graph is $\frac{|V|(|V-1|)}{2}$, so the maximal density is 1 (for complete graphs) and the minimal density is 0 [1].

1.2 Background on Graph Clustering

Clustering in graphs has become a fundamental technique in the study of complex network systems, providing key insights into the potential structure and organizations of large-scale networks. By partitioning a graph into several clusters, where nodes within the same cluster are more tightly connected than nodes outside it, researchers can mine for underlying patterns, identify communities, and reveal hierarchical structures. Clustering methods have garnered widespread attentions and applications in various real-world scenarios. For instance, in social networks, they help communities by identifying groups of users with similar interests or interactions [2]. In biological networks, researchers explore relationships between diseases by examining their underlying biological mechanisms, thereby investigating disease-disease associations at a system-level [3]. Recommenda-

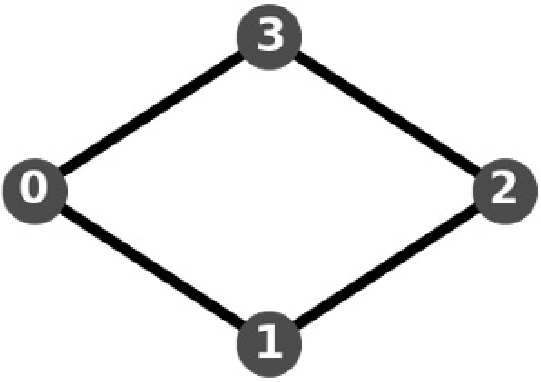


Figure 1: Unsigned Graph

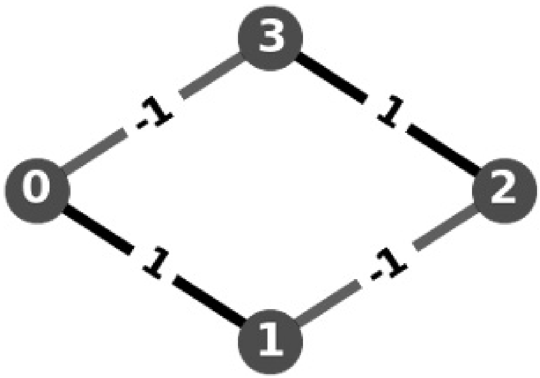


Figure 2: Signed Graph

tion systems employ clustering to group similar products or users, enabling more personalized content suggestions [4]. Furthermore, in transportation networks, clustering can identify densely connected hubs or routes, optimizing logistics and efficiency [5]. These diverse applications emphasis the significance of graph clustering methods in analyzing and understanding complex network systems.

In the past few years, researchers have proposed many methods for graph clustering [6, 7, 8]. Among these state-of-the-art methods, spectral clustering method has been one of the widely used methods for graph clustering task. This approach is based on spectral graph theory, which studies the properties of graphs in relation to characteristics of their matrices (i.e., the Laplacian matrix). The foundational concept of spectral method is proposed by Fiedler in the 1970s [9], he introduced the Fiedler vector or the second smallest eigenvector of the graph Laplacian, which can be used to bipartite a graph. While most of the research has focused on clustering methods on unweighted or non-negative weighted graphs, the task of clustering signed graphs, especially the large-scale signed graphs, remains relatively unexplored and has recently become an increasingly important research topic [10].

1.3 Background on Signed Graph Clustering

A signed graph is a graph in which each edge has a positive or negative sign. Many real-world relationships can be represented by signed graphs with positive and negative links [11], such as like/dislike, trust/distrust, or friends/enemies. Signed graphs bring complexity to network analysis by embedding positive and negative relationships in the graph structure. Unlike traditional graphs, where edges represent only connections, symbolic graphs with positive or negative weighted edges usually contain more information and status.

The presence of positive and negative links brings unique challenges and implications to signed graph clustering. Many clustering algorithms for signed graphs have been proposed in recently years. Traditional methods, like spectral clustering, have been adapted to handle signed graphs. One popular approach is to modify the Laplacian matrix of the graph to account for negative weights, leading to the signed Laplacian [12]. Eigen-decomposition of this matrix then reveals clusters that respect the positive and negative relationships [13]. Another significant direction in signed graph clustering is the utilization of modularity optimization. Researchers have developed signed versions of modularity, aiming to partition the network such that positive links are within clusters and negative links are between them [14]. Various optimization techniques, from greedy algorithms to simulated annealing, have been employed to maximize this signed modularity. Furthermore, with the rise of deep learning, recent studies have attempted to cluster signed graphs using Graph Neural Networks (GNN) [15]. These methods leverage node embeddings that capture the complex interplay of positive and negative relationships, thereby facilitating cluster formations.

Despite these advances, clustering signed graphs remains challenging due to the complex balance of positive and negative relationships. As the size and complexity of real-world networks continue to grow, the search for efficient clustering methods for signed graphs remains an active and important area of research. In this paper, we propose a signed graph clustering algorithm using k-core decomposition and signed motifs, and we implement several numerical experiments on 11 real-world signed networks. We also compare our proposed algorithm with the algorithm without k-core decomposition, and the results show that the proposed algorithm is more efficient. Furthermore, we also discuss the variation in the accuracy of the clustering for each real-world dataset as the number of clusters increases. Our results indicate that the choice of the appropriate number of clusters, k, is strongly influenced by the network's structure. Furthermore, determining the correct number of clusters is vital for achieving optimal clustering accuracy.

This paper is organized as follows. Section 2 summarizes some related works. Section 3 introduces preliminaries and concepts. In section 4, we propose a signed clustering method using k-core decomposition and signed motifs. Section 5 contains numerical experiments on various real-world datasets. Finally, Section 6 summarizes the findings from our experiments along with future research directions.

2 Related Work

2.1 Spectral Methods

Kunegis et al. [16] presented an in-depth study of a social network with negative edges, specifically focusing on a dataset called the Slashdot Zoo. They explored the eigenvalues and eigenvectors of the network's Laplacian to uncover patterns in the distribution of positive and negative edges. Mercado et al. [17] extended spectral clustering to signed graphs via the one-parameter family of signed power mean Laplacians. Moreover, They provided a thorough analysis of the proposed approach in the setting of a general stochastic block model. Esmailian and Jalili [18] embarked on a comprehensive examination of community detection in signed networks, particularly emphasizing the impact of negative ties across varying scales. Their research highlighted a profound influence of negative edges on community structures. By exploring signed modularity, they found as the number of negative ties increases, the density of positive ties is neglected more. These results shed lights on the community structure of signed networks. Chiang et al. [19] extended the spectral clustering algorithm to the signed network setting by considering the signed graph Laplacian, and they developed a multilevel clustering framework called balanced normalized cut (BNC) by formulating a new k-way objective for signed networks. Jiang [20] proposed an extension of the stochastic block model (SSBM) to study the mesoscopic structural patterns in signed networks. Cucuringu et al. [21] [22] proposed a principled spectral algorithm (SPONGE) for clustering signed graphs, that amounts to solving a generalized eigenvalue problem for constrained clustering. Bonchi et al. [23] illustrated the problem of discovering two polarized communities in the signed graph as a Rayleigh quotient problem and refer to the objective function of problem as polarity. Tzeng et al. [24] showed that the proposed objective by Bonchi et al. can be interpreted in terms of the Laplacian of a complete graph, and rely on the spectral properties of this matrix to derive a novel optimization framework, spectral conflicting group detection (SCG).

In summary, the application of spectral methods to signed graphs continues to evolve, with researchers utilizing these techniques to gain insight into the nature of positive and negative relationships. The above studies represent only a small fraction of the emerging interest in this field. However, they emphasize the potential and promise of spectral methods in exploring the complexity of signed networks.

2.2 *k*-core Decomposition

k-core decomposition consists of the process of grouping the nodes in a graph into groups based on the degree sequence and topology of the graph. The term "i-core" refers to the largest subgraph of the original graph such that each node in that subgraph has degree at least i. The standard algorithm for k-core decomposition iteratively removes the nodes with the lowest degree until the graph is empty. When a node is removed from the graph, all its relations are removed and the degree of its neighbors is reduced by one. By this method, different core groups can be discovered one by one. The concept of k-core decomposition has long been a staple of complex network analysis, helping to mine the hierarchical structure and backbone of large-scale networks. Seidman [25] has made foundational contributions to the in-depth study of network structure and its relationship to node degree. Although this study does not focus specifically on k-core decomposition, it lays an important foundation by exploring the effects of node connectivity and inherent structural properties of networks. Batagelj and Zaversnik [26] presented their efficient algorithm for k-core decomposition of networks. The complexity of their method is $O(m)$, where m is the number of edges, allowing for a faster and more scalable decomposition. The effectiveness of the algorithm makes it a staple for subsequent network analysis, especially for large-scale graphs where computational efficiency is critical. Alvarez-Hamelin et al. [27] introduced an innovative approach for visualizing and fingerprinting large-scale networks. They develop a general visualization algorithm that can be used to compare the structural properties of various networks and highlight their hierarchical structure. The low computational complexity of the algorithm, $O(n+m)$, where n is the number of vertices, makes it suitable for the visualization of very large sparse networks. Peng et al. [28] proposed a framework that accelerates community detection by applying an expensive algorithm to the k-core and then using an inexpensive heuristic to infer community labels for the remaining nodes. Giatsidis et al. [29] introduced CoreCluster, an efficient graph clustering framework leveraging graph degeneracy and k-core decomposition, allowing hierarchical processing of graphs. Their results show that CoreCluster can accelerate any clustering algorithm's performance while either preserving or enhancing the quality of the clustering structure. Batagelj and Zaversnik [30] presented an efficient algorithm for k-core decomposition of networks with a complexity of $O(m)$. Additionally, they extend the classical concept of k-core by employing a vertex property function, enabling the determination of generalized cores for local monotone vertex property functions, illustrated through the analysis of a collaboration network in computational geometry.

2.3 Network Motifs

Milo et al. [31] were among the pioneers to delve deep into the world of network motifs, identifying them as simple building blocks of complex networks. Their study highlighted the significance of subgraph patterns in determining network functionality and behavior. They defined "network motifs" and emphasized that their approach may uncover the basic building blocks of most networks. Paranjape et al. [32] introduced a concept of temporal network motifs, offering a methodology to count these motifs in temporal networks and presenting fast algorithms for the task, which are shown to be considerably faster than baseline methods. By analyzing real-world datasets, the research reveals domain-specific motif frequencies and highlights that assessing motifs at different time scales can uncover varied network behaviors. Benson et al. [33] developed a generalized framework for clustering networks on the basis of higher-order connectivity patterns, revealing complex organizational structures in various networks, they emphasized the significance of network motifs as foundational building blocks and proposed an efficient motif-based spectral clustering method for large networks. Huang et al. [15] proposed a unified framework to effectively model signed directed networks based on motifs, called Signed Graph Attention Networks (SiGAT), which is an extension of the Graph Attention Networks (GAT) framework and is specifically designed to address signed networks containing both positive and negative links. Mei et al. [34] introduced KCoreMotif, an efficient graph clustering algorithm designed for large networks that combines k-core decomposition with motif-based spectral clustering. Comparative analysis of real-world datasets demonstrated the effectiveness of KCoreMotif in accurately clustering large networks, highlighting its potential for intra-cluster and inter-cluster trust evaluation of such networks.

Moreover, Motif-based analysis has been applied in different fields (e.g., Biological networks [31], [35], Neuro- science [36], Transcriptional regulation network [37]), to reveal different connection patterns in complex networks. Applications in various fields offer us unique insights into the structure, dynamics, and behavior of complex systems.

3 Preliminaries and Concepts

3.1 k-core Decomposition

The notion of a core was introduced by Seidman [25]. Let $n=|V|$ and $m=|E|$, the degree $d(v)$ of vertex v equals to the number of lines having vertex v as their endpoint. Let k be an integer. A subgraph $\mathcal{H}_k=(C, E|C)$ induced by the subset of vertices $C \subseteq V$ is called a *k-core* or a core of order k iff $d_{\mathcal{H}_k}(v) \geq k$, for all $v \in C$. $E|C=\{(u, v) \in E | u \in C, v \in C\}$, i.e., $E|C$ is the set of edges from E that have both endpoints in C, and $\mathcal{H}_k$ is a maximum subgraph with this property, $d_{\mathcal{H}_k}(v)$ gives the number of edges in the subgraph $\mathcal{H}_k$ that are connected to the vertex v. The core number core(v) of vertex v is the highest order of a core that contains this vertex. The core of maximum order is also called the main core [30]. Its core number is denoted by core(G). The algorithm for determining the cores hierarchy is based on the following property by Batagelj et al. [38]:

If from a given graph $G=(V, E)$ we recursively delete all vertices of degree less than k, and lines incident with them, the remaining graph is the k-core of G.

Figure 3 illustrates an example of the decomposition of a given graph. In simple terms, the 0-core includes all nodes of the graph because every node has a degree of at least 0. 1-core includes all nodes with a degree of at least 1. In other words, isolated nodes will be removed. 2-core is a subgraph where all nodes have at least 2 neighbors, nodes with only 1 connection will be removed, and so on.

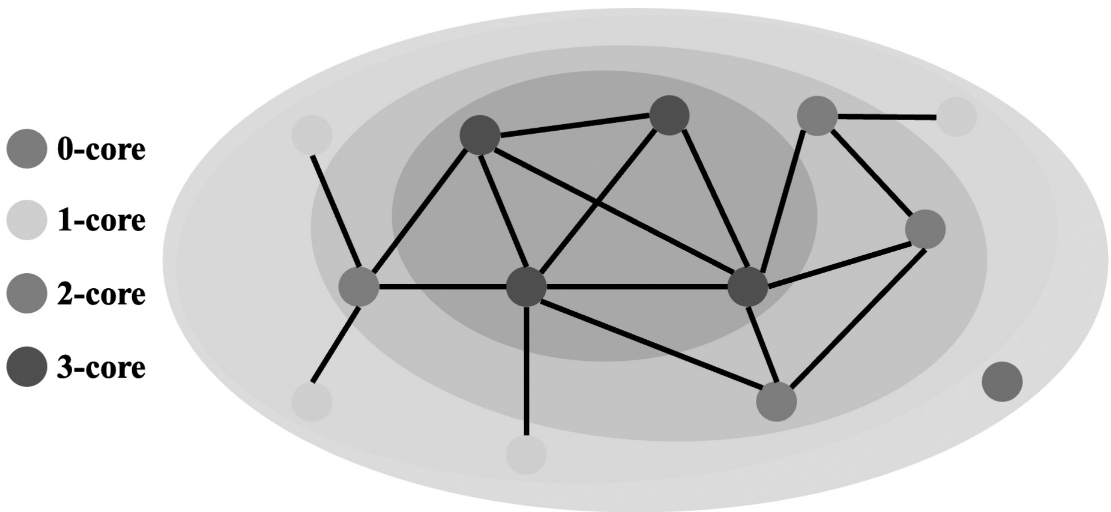


Figure 3: k-core Decomposition of Graph

3.2 Network Motifs

Consider motifs to be a pattern of edges on a small number of nodes. The definitions and formulas in this section are derived from Benson's original work [33]. As defined in [33], a motif on k nodes is represented by a tuple $(B, \mathcal{A})$, where B is a $k \times k$ binary matrix and $\mathcal{A} \subset \{1, 2, \ldots, k\}$ is a set of anchor nodes. The matrix B encodes the edge pattern between the k nodes, and $\mathcal{A}$ labels a relevant subset of nodes for defining motif conductance. In many cases, $\mathcal{A} = \{1, 2, \ldots, k\}$. Let $\chi_{\mathcal{A}}$ be a selection function that takes the subset of a k-tuple indexed by $\mathcal{A}$, and let set$(\cdot)$ be the operator that takes an (ordered) tuple to an (unordered) set. Specifically,

$$\mathrm{set}((v_1, v_2, \ldots, v_k)) = \{v_1, v_2, \ldots, v_k\}. \tag{2}$$

The set of motifs in an unweighted (possibly directed) graph with adjacency matrix A, denoted $M(B, \mathcal{A})$, is defined by

$$M(B, \mathcal{A}) = \{(\mathrm{set}(\boldsymbol{v}), \mathrm{set}(\chi_{\mathcal{A}}(\boldsymbol{v}))) \mid \boldsymbol{v} \in V^k, v_1, \ldots, v_k \text{ distinct}, A_{\boldsymbol{v}} = B\}, \tag{3}$$

where $A_{\boldsymbol{v}}$ is the $k \times k$ adjacency matrix on the subgraph induced by the k nodes of the ordered vector $\boldsymbol{v}$.

For directed signed graphs, the direction also contains some information, e.g., status theory. Status theory is an important social psychological theory for signed graph analysis. In a signed graph, a positive sign "+" or negative sign "−" means target node has a higher or lower status than source node [15]. If individual α perceives themselves as higher status than individual β, then α might offer positive feedback to β, while β may simultaneously view α negatively due to perceived status disparity. Based on status theory, it can be found that "opinion leaders" have higher social status than ordinary users.

To extend the method for signed graphs, we generalize Eq. 3 by allowing the adjacency matrix B to be signed [33].

The set of motifs in a signed (possibly directed) graph with adjacency matrix A, denoted $M(B, \mathcal{A})$, is defined by

$$M(B, \mathcal{A}) = \{(\mathrm{set}(\boldsymbol{v}), \mathrm{set}(\chi_{\mathcal{A}}(\boldsymbol{v}))) \mid \boldsymbol{v} \in V^k, v_1, \ldots, v_k \text{ distinct}, A_{\boldsymbol{v}} = B\}, \tag{4}$$

where $A_{\boldsymbol{v}}$ is the $k \times k$ signed adjacency matrix on the subgraph induced by the k nodes of the ordered vector $\boldsymbol{v}$.

3.2.1 Motif adjacency matrix and motif Laplacian

Given an unweighted, directed graph and a motif set M, the motif adjacency matrix is defined as

$$(W_M)_{ij} = \sum_{(\boldsymbol{v}, \chi_{\mathcal{A}}(\boldsymbol{v})) \in M} \mathbf{1}(\{i, j\} \subset \chi_{\mathcal{A}}(\boldsymbol{v})) \tag{5}$$

for $i \neq j$. Conceptually, the motif adjacency matrix is number of motif instances in M where i and j participate in the motif. Note that weight is added to $(W_M)_{ij}$ only if i and j appear in the anchor set.

The motif diagonal degree matrix is defined by $(D_M)_{ii} = \sum_{j=1}^{n} (W_M)_{ij}$ and the motif Laplacian as

$L_M = D_M - W_M$. The normalized motif Laplacian is defined by $\mathcal{L}_M = D_M^{-1/2} L_M D_M^{-1/2} = I - D_M^{-1/2} W_M D_M^{-1/2}$.

3.2.2 Weighted (Signed) motifs

Suppose the clustering is based on motif sets $M_1, \ldots, M_q$ for q different motifs. Let W_{Mj} be the weighted adjacency matrix for motif M_j, $j=1, \ldots, q$, and let $\alpha_j \geq 0$ be the weight of motif M_j, then the weighted adjacency matrix can be formed as

$$W_M = \sum_{j=1}^{q} \alpha_j W_{Mj} \tag{6}$$

Each motif has an associated nonnegative weight, which is a weighted motif. Let $\omega(\boldsymbol{v}, \chi_{\mathcal{A}}(\boldsymbol{v}))$ be the weight of a motif instance. the motif adjacency matrix as follows:

$$(W_M)_{ij} = \sum_{(\boldsymbol{v}, \chi_{\mathcal{A}}(\boldsymbol{v})) \in M} \omega_{(\boldsymbol{v}, \chi_{\mathcal{A}}(\boldsymbol{v}))} \mathbf{1}(\{i, j\} \subset \chi_A(\boldsymbol{v})) \tag{7}$$

4 Methodology

4.1 Signed Motifs

A motif refers to a recurring and statistically significant subgraph or pattern within a larger graph. Essentially, it is a small connected subgraph that appears in a network more frequently than one would expect in a random network. Motifs can be thought of as the "building blocks" of the network, offering insights into its underlying structure and functionality. Their important role in graph clustering is noteworthy. Since motifs package specific topological patterns, they can be used to identify nodes that participate in similar structural roles, making them an effective tool for clustering.

Benson et al. [33] give 13 connected three node directed motifs from the original 13 types of three-node connected subgraphs by [31]. Among them, M1-M7 are called triangular motifs, which are crucial for social networks. Therefore, we will focus on triangular motifs in this paper. From the study of Huang et al. [15], we list all 32 triangular signed motifs in Figure 4. Due to the time limitation, we only selected the signed motif in the dotted box on the right in Figure 4 as our target signed motif M in the following section. We expect to explore more other motifs as target motifs in our future research.

4.2 Proposed Algorithm

In this paper, we propose an efficient signed graph clustering algorithm for large-scale networks using *k*-core decomposition and signed motifs. The essential idea behind the proposed clustering algorithm is inspired by [33] and [34]. Motifs have recently made great improvements in many graph analysis tasks. However, most of the work has focused on unsigned networks. We have observed that signed motifs are pivotal in clustering large-scale signed networks. Therefore, we aim to investigate the clustering results of this structure in real-world networks. To avoid the high

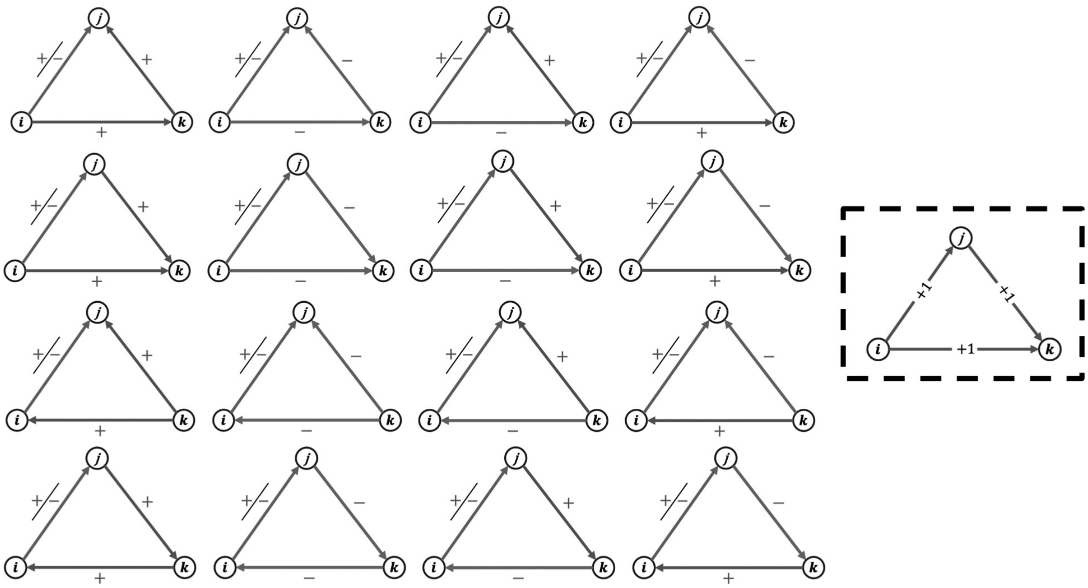


Figure 4: Signed Motifs

computational costs associated with traditional spectral algorithms for clustering large-scale networks, our proposed method begins with k-core decomposition to identify the top-k cores.

To conduct the algorithm, we generate k-core subgraphs in the first step using k-core decomposition of the given signed, which is an important step in reducing computational complexity. Subsequently, identify the specific signed motif (i.e., The motif on right side of Figure 4) on the given graph, conduct k-means spectral clustering on the corresponding eigenvectors of the Laplacian matrix to obtain the clusters on the top k-core subgraph. Finally, for the remaining nodes after the k-core decomposition, we use the same approach as in [34], i.e., loop through all the neighbors of node A and check whether the maximum percentage of neighbors located in the same cluster to the total number of neighbors is more than 50%, and if it is more than that, then the target node A is assigned to the cluster where the majority of its neighbors are located; otherwise, a new cluster is assigned and the target node A is assigned to the new cluster. After implementing the above steps, we can assign clusters to all the nodes.

Pseudocode for identifying signed motifs from the input graph is given as Algorithm 1. The complexity of this algorithm is $O(n^3)$, where n is the number of vertices in the input network, $N(v)$ represents the set of neighboring vertices of v in graph G. Pseudocode for k-core decomposition [13] is given as Algorithm 2, where $d(v)$ is the degree of vertex v in graph G, $N(v)$ represents the set of neighboring vertices of v in graph G. The complexity of k-core decomposition is $O(n+m)$. In the next step of implementing our proposed signed motif based clustering methods on top k-cores subgraph, we use a publicly available implementation [33] to conduct this approach.

Algorithm 1 Identify Specific Signed Motifs M

Input: $G(V, E)$: A signed graph with nodes V and edges E with signs.
Output: $G_M(V_M, E_M)$: A subgraph of G containing only edges that satisfy the signed motif.

$E_M \leftarrow \emptyset$
for each $v_i \in V$ **do**
    for each $v_j \in N(v_i)$ **do**
        if $s(v_i, v_j)=1$ **then**
            for each $v_k \in N(v_i)$ **do**
                if $v_k \neq v_j$ and $s(v_i, v_k)=1$ and $s(v_k, v_j)=1$ **then**
                    $E_M \leftarrow E_M \cup \{(v_i, v_j), (v_k, v_i), (v_k, v_j)\}$
                end if
            end for
        end if
    end for
end for
$G_M(V_M, E_M) \leftarrow$ Create a directed graph with edges from E_M with sign 1
return G_M

Algorithm 2 *k*-core Decomposition

Input: Graph: $G=(V, E)$
Output: core: $V \rightarrow \mathbb{N}$

for v in V **do**
    Compute $d(v)$
end for
Order V as V' such that $v_1, v_2, \ldots, v_n$ where $d(v_1) \leq d(v_2) \leq \ldots \leq d(v_n)$
for $i=1$ to n **do**
    core$[v_i] := d(v_i)$
    for u in $N(v_i)$ **do**
        if $d(u) > d(v_i)$ **then**
            $d(u) := d(u) - 1$
            Update the ordering of V'
        end if
    end for
end for
return core

Algorithm 3 Motif-based Clustering

Input: Graph $G=(V, E)$, motif M, number of clusters k
Output: k disjoint motif-based clusters

```
for each (i, j) in V × V do
    (W_M)_ij ← Number of instances of M that contain nodes i and j.
end for
for each i in V do
    (D_M)_ii ← Σ_{j∈V} (W_M)_ij
end for
L_M ← I − D_M^{-1/2} W_M D_M^{-1/2}
Let z_1, ..., z_k be the eigenvectors corresponding to the k smallest eigenvalues of L_M.
for each (i, j) in V × {1, 2, ..., k} do
    Y_ij ← z_ij / sqrt(Σ_{j=1}^{k} z_ij^2)
end for
for each i in V do
    Embed node i into R^k using the i^th row of matrix Y.
end for
Apply k-means clustering on the embedded nodes to obtain k disjoint motif-based clusters.
```

Pseudocode for motif-based clustering algorithm [33] is given as Algorithm 3. V is the set of vertices in a graph, $V \times V$ represents all possible ordered pairs of vertices. The eigenvector can be computed in $O((m+n)(\log n)^{O(1)})$ time using fast Laplacian solvers [39], it takes $O(n \log n)$ to sort the indices given the eigenvector using a standard sorting algorithm. For a motif on k nodes, W_M can be computed in $\Theta(n^k)$ time by checking each k-tuple of nodes [33].

5 Numerical Experiments

To evaluate the effectiveness of our proposed algorithm, we implemented the algorithm on some real-world networks in this section. Besides, we compared the algorithm with the same approach but without k-core decomposition. All experiments are executed by Python 3 on a 12-core Apple M2 MAX-Chip personal computer at 3.6 GHz with 32GB RAM. The datasets we have used are all publicly available and the detailed information can be found in the Section 5.

5.1 Real-World Datasets

In this section, we describe the real-world datasets used in the numerical experiments.

Table 1: Datasets Description

Name	$\lvert V \rvert$	$\lvert E \rvert$	$\lvert E\_ \rvert / \lvert E \rvert$
Highland	16	58	0.50
Sampson	18	189	0.46
ConVote	219	764	0.20
WoW-EP8	790	116009	0.18
Bitcoin	5881	21492	0.15
WikiVote	7115	100693	0.22
Referendum	10884	251406	0.05
Slashdot	82140	500481	0.24
WikiCon	116717	2026646	0.63
Epinions	131580	711210	0.17
WikiPol	138587	715883	0.12

Highland This is a clusterable signed graph illustrated by Hage & Harary using the Gahuku-Gama system described by Read for the Eastern Central Highlands of New Guinea [40], [41]. Nodes represent tribes, and edges represent friendships.

Sampson The Sampson dataset records social interactions between a group of monks residing as visual experimenters containing monks associated with a crisis in a New England (USA) cloister (or monastery) and collects a number of sociometric rankings [42], with the crisis leading to the departure of several monks. Vertices represent monks, and edges represent ratings.

ConVote This dataset builds from congressional speech data [43], vertices represent politicians speaking in the United States Congress, and edges denote that a speaker mentions another speaker.

WoW-EP8 This is a dataset of legislative amendments collected from the European Parliament website [44]. The dataset spans the 8th legislature (EP8) from 2014 to 2019. Vertices represent parliament representatives (MEPs, for Member of the European Parliament), and edges represent attitude (accepted or rejected).

Bitcoin It is a who-trusts-whom network of people who use Bitcoin to transact on a platform called Bitcoin OTC. Since Bitcoin traders are anonymous, there is a need to maintain a record of the user's reputation in order to prevent transactions with fraudulent and risky users [45]. Vertices represent Bitcoin traders, Positive edges denote trust, and negative edges denote distrust.

WikiVote In order for a user to become an administrator, a Request for adminship (RfA) is issued and the Wikipedia community via a public discussion or a vote decides who to promote to adminship [45]. The vertices in this network represent Wikipedia users, and edges represent the positive and negative votes for electing Wikipedia admins.

Referendum In the past few years, when people focus on the field of online political debate, interest in automatically categorizing users' positions on a given entity, such as a controversial topic or a politician, has grown significantly in polarized debates. The dataset has been collected from a corpus of Italian tweets where users were discussing a highly polarized debate in Italy, i.e. the 2016 referendum on the reform of the Italian Constitution [46]. Edge signs indicate if two users (vertices) are classified to have the same stance or not.

Slashdot This is a signed social network for users of the technology news site Slashdot (slashdot.org), connected by directed "friend" and "foe" relationships. The "friend" and "foe" tags are used on Slashdot to mark users and influence the scores as seen by each user. Positive edges denote the two users (nodes) are friends, negative edges denote the two users (nodes) are foes [45].

WikiCon Wikipedia conflict collects positive and negative conflicts between users editing the English Wikipedia [47]. Vertices represent users and edges represent conflict between two users, with the edge sign representing positive and negative interactions.

Epinions This is a online social network of who trusts whom for the general consumer review site Epinions.com. The network consists of individual users who are connected by directed trust and distrust edges. Members of the site can decide whether to "trust" each other or not. All trust relationships interact and form a trust network, which is then combined with review ratings to determine which reviews are shown to users [45].

WikiPol The network contains interpretive interactions between English-language Wikipedia users who have edited pages about politics. Each interaction, such as text editing, reverting, restoring, and voting, is assigned a positive or negative value [48]. The results can be interpreted as a web of trust. Edges represent interactions and vertices represent users.

Table 1 shows the number of vertices, edges and negative edges proportion of each real-world signed network.

5.2 Statistical Results

The first column of Table 2 displays the number of target signed motifs for each dataset studied in this paper. The second column presents the number of unsigned motifs (i.e., M5 in [33]) following the same pattern in the studied dataset. Additionally, the final column indicates the proportion of signed motifs relative to unsigned motifs. Furthermore, Figure 5 shows the computation time required to identify signed motifs in the real-world datasets. As anticipated, due to the complexity of WikiCon, the longest time required to identify the signed motif was approximately 380 seconds. As a network expands with more nodes and edges, the computation time predictably increases.

Additionally, we provide an visualization of **Highland**'s signed motifs in Figure 5. The nodes and edges of the detected signed motif in the figure are shown in red and blue respectively, and the edges and nodes that do not belong to the motif are marked in light gray.

Table 2: Statistical Results

	# of signed motifs	# of motifs (M5)	Proportion (%)
Highland	19	68	27.94%
Sampson	17	157	10.83%
ConVote	91	155	58.71%
WoW-EP8	5482933	8716683	62.90%
Bitcoin	22859	33493	68.25%
WikiVote	400303	607279	65.92%
Referendum	2927424	3120811	93.80%
Slashdot	417000	579565	71.95%
WikiCon	3929593	13831236	28.41%
Epinions	3939584	4910076	80.23%
WikiPol	2563542	2978026	86.08%

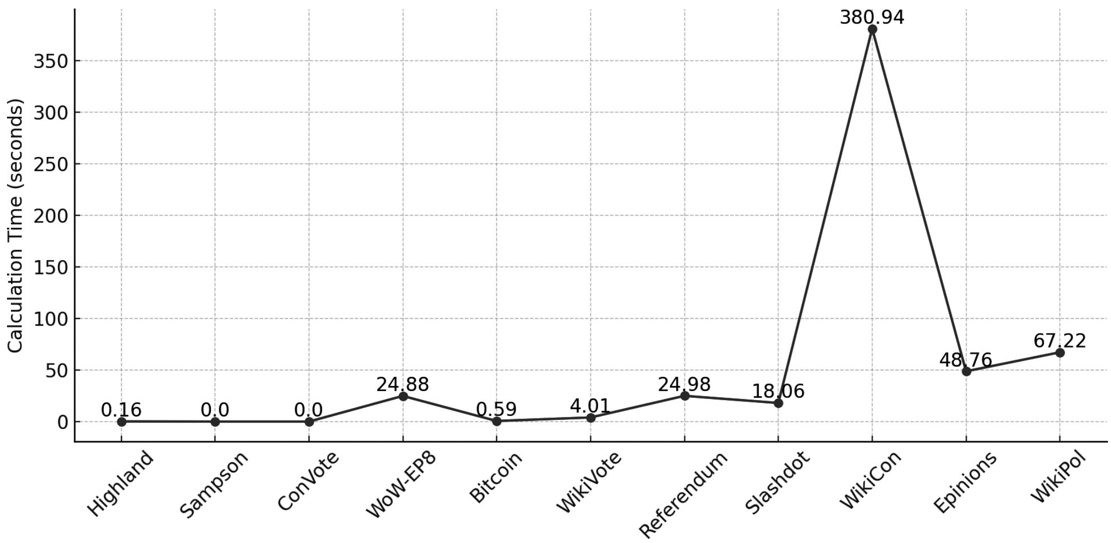


Figure 5: Calculation Time to Identify Signed Motifs for the Studied Datasets

5.3 Evaluation Metrics

Modularity is typically used to measure the strength of graph partitioning, which is a key evaluation metric for clustering methods [49]. For a given graph, the modularity Q is calculated using the formula:

$$Q=\frac{1}{2m}\sum_{ij}\left(A_{ij}-\frac{k_i k_j}{2m}\right)\delta(c_i,\ c_j) \tag{8}$$

where A_{ij} is the element of the adjacency matrix (1 if there's an edge between node i and j, and 0

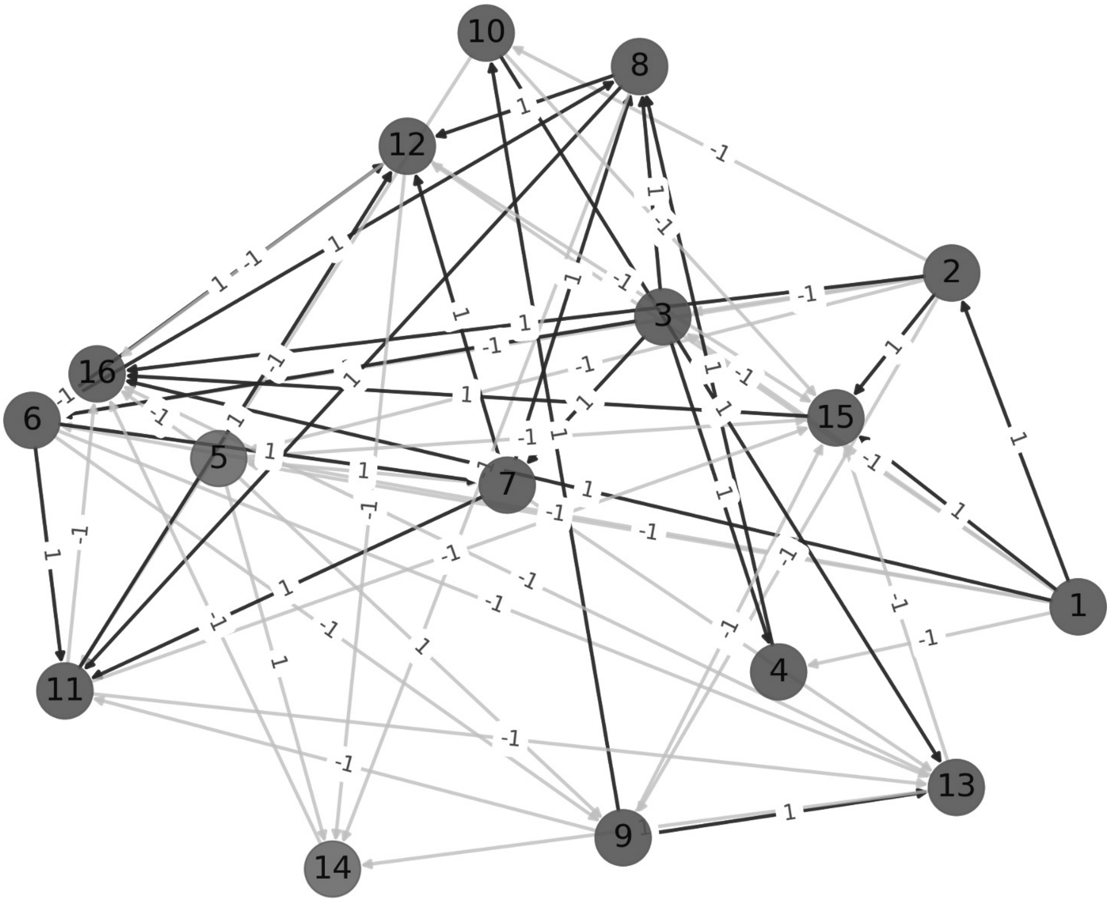


Figure 6: Signed Motifs in Highland

otherwise). k_i and k_j are the degrees of nodes i and j respectively. m is the total number of edges in the graph. c_i and c_j are the communities of nodes i and j. δ is the Kronecker delta function, which equals 1 if $c_i = c_j$ and 0 otherwise. The modularity value Q can range between -1 and 1, a high positive modularity indicates a good community structure.

5.4 Experimental Results

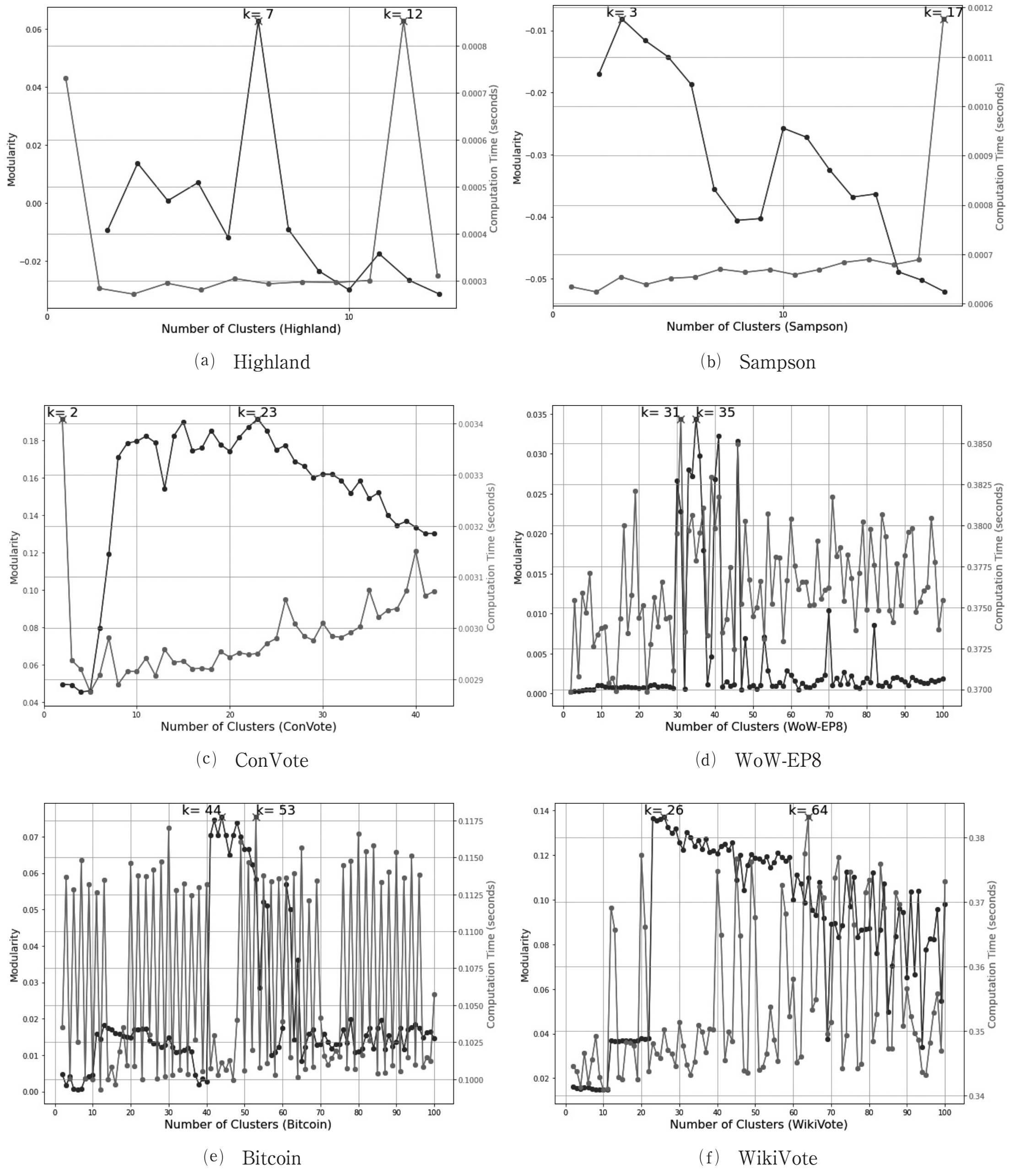
In this section, we evaluate the efficiency of the proposed signed graph clustering algorithm by comparing it with the conventional spectral clustering algorithm without k-core decomposition.

To explore the effect of k value on clustering, we iterate k from 2 to 100 for all studied datasets and calculate the corresponding modularity and computation time. The results are shown in 7. In order to intuitively show the reduction in computational complexity after k-core decomposition, we also conduct the same analysis without k-core decomposition, and the results are shown in Figure 8. Note that due to the large size of the data, clustering is very inefficient for large-scale networks without some special algorithms. We have only given the results for large-scale networks with k from 2 to 10, where results of **WikiCon** are not given due to memory limitation, but it is possible to get a glimpse of the other large-scale networks, whose efficiency are quite low.

Figure 7 show the movement of the modularity value and the corresponding computational time required as k increases for the relatively small-scale network and the relatively large-scale network by conducting the proposed signed graph clustering algorithm. Figure 8 show the movement of the modularity value and the corresponding computational time required as k increases for the relatively small-scale network and the relatively large-scale network by conducting the same procedure without k-core decomposition.

By analyzing Figure 7, 8, we can find that:

- Overall, as the k of clusters increases, the accuracy (i.e., Modularity) increases first. Then, when k reaches a certain value, the modularity value will decrease.
- It is not the case that better clustering is obtained as the number of clusters detected increases; on the contrary, modularity performs better at smaller cluster on relatively large-scale networks.
- **Sampson** and **WikiCon** have negative modularity in most cases, which means that the clustering is basically invalid. After looking at the specific data structure, we found that **Sampson** has many self-loop subgraphs, while **WikiCon** has negative edges accounting for 63% of all edges, and the signed motif we used in the algorithm has no negative edges, which also causes the ineffectiveness of clustering on **WikiCon**.
- Although the overall modularity values are relatively small, our results are still reasonable compared to the work of [34].
- The computation time remains relatively consistent regardless of the number of clusters. It fluctuates slightly but remains within a narrow range. Similar to the small-scale datasets, the computation time for the large-scale datasets remain slight fluctuations across different numbers of clusters. In contrast, as shown in Figure 8, the computational time of clustering methods that do not use k-core decomposition increases significantly as k increases.
- The value of modularity performs best for the number of clusters detected is also different under the two methods, but we can find that the value of modularity on all datasets overall performs better with the algorithm that conducts k-core decomposition, and the computation time is also obviously much shorter.
- On networks with a larger proportion of signed motifs (shown in Table 2), we cannot draw clear conclusions, but we did find that when the proportion of signed motifs is smaller (e.g., **Sampson** and **WikiCon**), the accuracy of clustering will decrease.
- Regarding the appropriate value of k, which is greatly influenced by the structure of a network itself, it is crucial to find the appropriate number of clusters to achieve better clustering accuracy. This is also where we will focus on our future work.

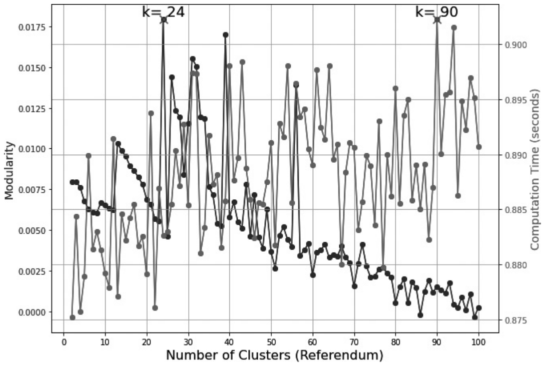


(a) Highland　(b) Sampson

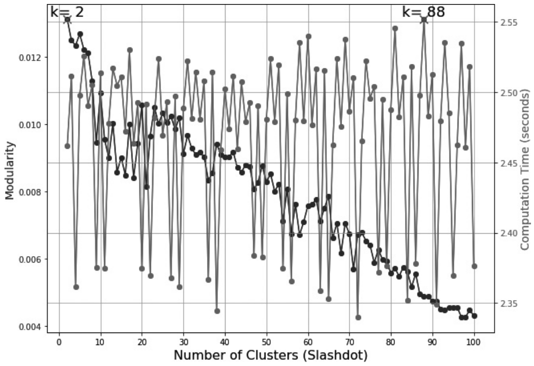
(c) ConVote　(d) WoW-EP8

(e) Bitcoin　(f) WikiVote

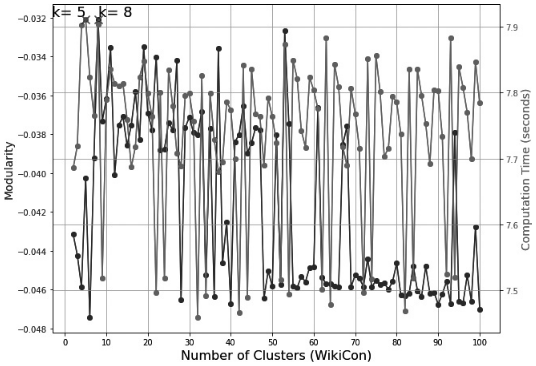
Figure 7: Modularity and Computation Time of the Proposed Algorithm vs. Number of Clusters



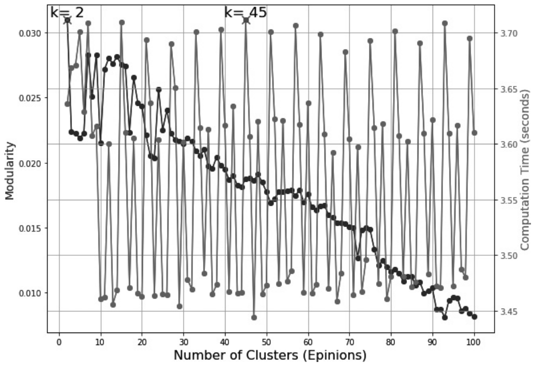
(g) Referendum



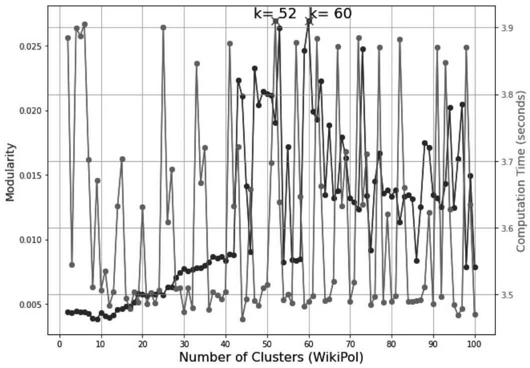
(h) Slashdot



(i) WikiCon

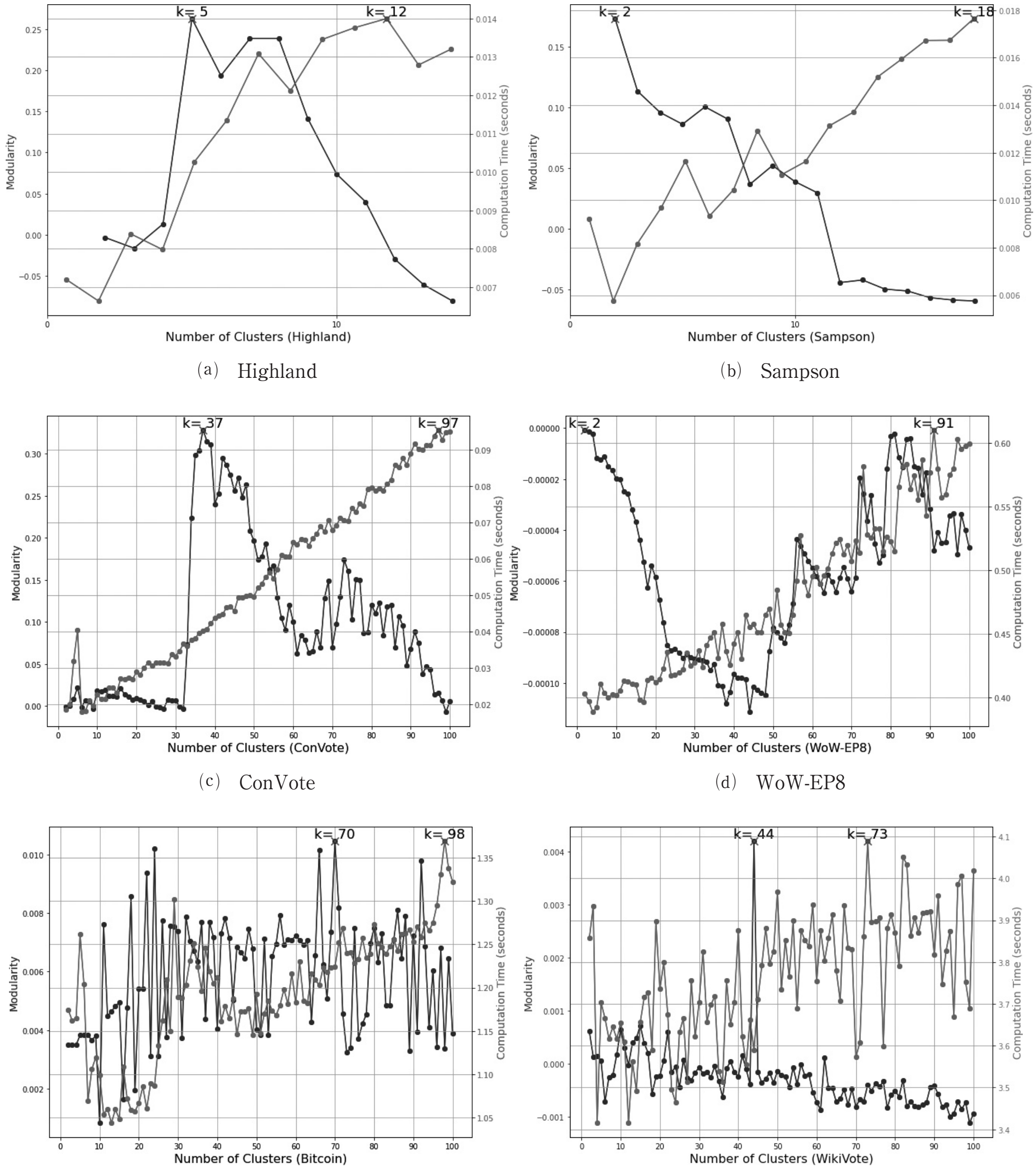


(j) Epinions



(k) WikiPol

Figure 7: Modularity and Computation Time of the Proposed Algorithm vs. Number of Clusters

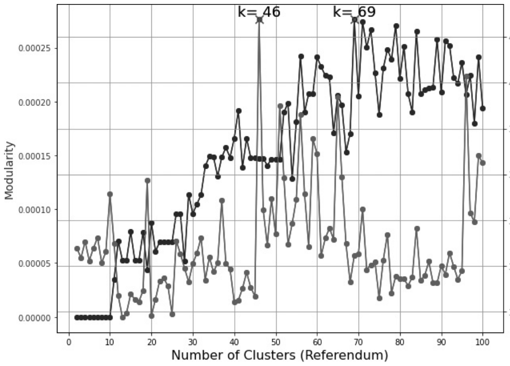


(a) Highland　　(b) Sampson

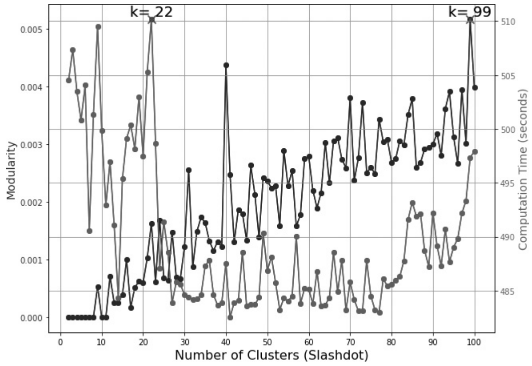
(c) ConVote　　(d) WoW-EP8

(e) Bitcoin　　(f) WikiVote

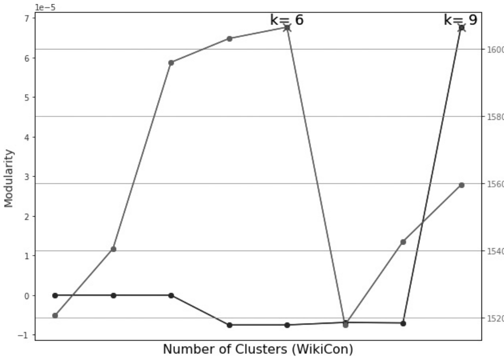
Figure 8: Modularity and Computation Time of the Algorithm without *k*-core Decomposition vs. Number of Clusters



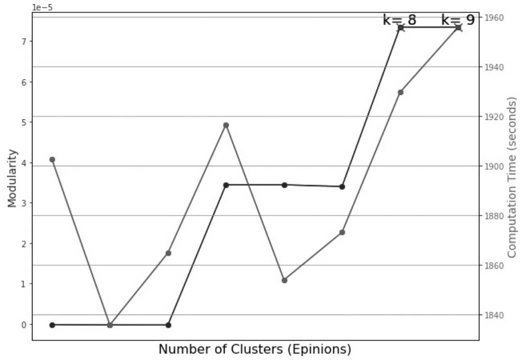
(g) Referendum



(h) Slashdot



(i) WikiCon



(j) Epinions

Figure 8: Modularity and Computation Time of the Algorithm without *k*-core Decomposition vs. Number of Clusters

6 Conclusion and Future Work

In this paper, we propose a signed graph clustering method that leverages *k*-core decomposition and signed motifs. While motifs have significantly advanced various graph analysis tasks, the majority of existing research has centered on unsigned networks. Recognizing the great potential of motifs in mining complex network structures, we combine two cutting-edge methods to explore pivotal aspects of signed network clustering. Our numerical experiments, conducted on 11 real-world networks, accompanied by a thorough discussion, affirm the superior efficacy of our algorithm. As motifs play important roles in more and more fields, especially through their ability to explore potential information existing in the complex structures of large networks, it greatly motivates us to continue exploring other potentially applicable motifs, not limited to triangular motifs. In this paper, we selected a single motif and presented experimental results based on it. We will try other patterns of motifs and compare their effects on the clustering efficiency of signed graphs in our future work.

Furthermore, when implementing the spectral clustering algorithm, we used the traditional *k*-means method. However, in recent years, many other state-of-the-art spectral clustering algorithms have been developed. We will integrate these algorithms in our follow-up research to develop more efficient clustering methods for large-scale signed graphs. Furthermore, as we begin to understand more about the underlying properties of signed graphs, there is potential for interdisciplinary applications, research fields like sociology, biology, and physics, ensuring that signed graph clustering methods remains at the forefront of graph analysis research.

Acknowledgement

This work was supported by JST SPRING, Grant Number JPMJSP2136.

References

[1] Thomas F Coleman and Jorge J Moré. Estimation of sparse jacobian matrices and graph coloring blems. *SIAM journal on Numerical Analysis*, 20(1): 187-209, 1983.

[2] Itai Himelboim, Marc A Smith, Lee Rainie, Ben Shneiderman, and Camila Espina. Classifying twitter topic-networks using social network analysis. *Social media+ society*, 3(1): 2056305117691545, 2017.

[3] Kai Sun, Joana P Gonçalves, Chris Larminie, and Nataša Pržulj. Predicting disease associations via biological network analysis. *BMC bioinformatics*, 15(1): 1-13, 2014.

[4] Yang Song, Lu Zhang, and C Lee Giles. Automatic tag recommendation algorithms for social recommender systems. *ACM Transactions on the Web (TWEB)*, 5(1): 1-31, 2011.

[5] Yuxuan Ji and Nikolas Geroliminis. On the spatial partitioning of urban transportation networks. *Transportation Research Part B: Methodological*, 46(10): 1639-1656, 2012.

[6] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5): 75-174, 2010.

[7] Hao Yin, Austin R Benson, Jure Leskovec, and David F Gleich. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 555-564, 2017.

[8] HongFang Zhou, Jin Li, JunHuai Li, FaCun Zhang, and YingAn Cui. A graph clustering method for community detection in complex networks. *Physica A: Statistical Mechanics and Its Applications*, 469: 551-562, 2017.

[9] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2): 298-305, 1973.

[10] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1361-1370, 2010.

[11] Fritz Heider. Attitudes and cognitive organization. *The Journal of psychology*, 21(1): 107-112, 1946.

[12] Jérôme Kunegis, Stephan Schmidt, Andreas Lommatzsch, Jürgen Lerner, Ernesto W De Luca, and Sahin Albayrak. Spectral analysis of signed graphs for clustering, prediction and visualization. In *Proceedings of the 2010 SIAM international conference on data mining*, pages 559-570. SIAM, 2010.

[13] Hervé Abdi. The eigen-decomposition: Eigenvalues and eigenvectors. *Encyclopedia of measurement and statistics*, pages 304-308, 2007.

[14] Pranay Anchuri and Malik Magdon-Ismail. Communities and balance in signed networks: A spectral approach. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 235-242. IEEE,

2012.

[15] Junjie Huang, Huawei Shen, Liang Hou, and Xueqi Cheng. Signed graph attention networks. In *Artificial Neural Networks and Machine Learning–ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17-19, 2019, Proceedings 28*, pages 566-577. Springer, 2019.

[16] Jérôme Kunegis, Andreas Lommatzsch, and Christian Bauckhage. The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web*, pages 741-750, 2009.

[17] Pedro Mercado, Francesco Tudisco, and Matthias Hein. Spectral clustering of signed graphs via matrix power means. In *International Conference on Machine Learning*, pages 4526-4536. PMLR, 2019.

[18] Pouya Esmailian and Mahdi Jalili. Community detection in signed networks: the role of negative ties in different scales. *Scientific reports*, 5(1): 14339, 2015.

[19] Kai-Yang Chiang, Joyce Jiyoung Whang, and Inderjit S Dhillon. Scalable clustering of signed networks using balance normalized cut. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 615-624, 2012.

[20] Jonathan Q Jiang. Stochastic block model and exploratory analysis in signed networks. *Physical Review E*, 91(6): 062805, 2015.

[21] Mihai Cucuringu, Ioannis Koutis, Sanjay Chawla, Gary Miller, and Richard Peng. Simple and scalable constrained clustering: a generalized spectral method. In *Artificial Intelligence and Statistics*, pages 445-454. PMLR, 2016.

[22] Mihai Cucuringu, Peter Davies, Aldo Glielmo, and Hemant Tyagi. Sponge: A generalized eigenproblem for clustering signed networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1088-1098. PMLR, 2019.

[23] Francesco Bonchi, Edoardo Galimberti, Aristides Gionis, Bruno Ordozgoiti, and Giancarlo Ruffo. Discovering polarized communities in signed networks. In *Proceedings of the 28th acm international conference on information and knowledge management*, pages 961-970, 2019.

[24] Ruo-Chun Tzeng, Bruno Ordozgoiti, and Aristides Gionis. Discovering conflicting groups in signed networks. *Advances in Neural Information Processing Systems*, 33: 10974-10985, 2020.

[25] Stephen B Seidman. Network structure and minimum degree. *Social networks*, 5(3): 269-287, 1983.

[26] Vladimir Batagelj and Matjaz Zaversnik. An o (m) algorithm for cores decomposition of networks. *arXiv preprint cs/0310049*, 2003.

[27] J Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. Large scale networks fingerprinting and visualization using the k-core decomposition. *Advances in neural information processing systems*, 18, 2005.

[28] Chengbin Peng, Tamara G Kolda, and Ali Pinar. Accelerating community detection by using k-core subgraphs. *arXiv preprint arXiv:1403.2226*, 2014.

[29] Christos Giatsidis, Fragkiskos D Malliaros, Nikolaos Tziortziotis, Charanpal Dhanjal, Emmanouil Kiagias, Dimitrios M Thilikos, and Michalis Vazirgiannis. A k-core decomposition framework for graph clustering. *arXiv preprint arXiv:1607.02096*, 2016.

[30] Vladimir Batagelj and Matjaž Zaveršnik. Fast algorithms for determining (generalized) core groups in social networks. *Advances in Data Analysis and Classification*, 5(2): 129-145, 2011.

[31] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594): 824-827, 2002.

[32] Ashwin Paranjape, Austin R Benson, and Jure Leskovec. Motifs in temporal networks. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 601-610, 2017.

[33] Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353 (6295): 163-166, 2016. https://github.com/arbenson/higher-order-organization-matlab.

[34] Gang Mei, Jingzhi Tu, Lei Xiao, and Francesco Piccialli. An efficient graph clustering algorithm by exploiting k-core decomposition and motifs. *Computers & Electrical Engineering*, 96: 107564, 2021.

[35]　Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6): 450-461, 2007.
[36]　Olaf Sporns and Rolf Kötter. Motifs in brain networks. *PLoS biology*, 2(11): e369, 2004.
[37]　Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1): 64-68, 2002.
[38]　Vladimir Batagelj, Andrej Mrvar, and Matjaž Zaveršnik. Partitioning approach to visualization of large graphs. In *Graph Drawing: 7th International Symposium, GD' 99 Štiřín Castle, Czech Republic September 15-19, 1999 Proceedings 7*, pages 90-97. Springer, 1999.
[39]　Luca Trevisan. Lecture notes on expansion, sparsest cut, and spectral graph theory. *URl: https://people.eecs.berkeley.edu/~luca/books/expanders.pdf*, 2013.
[40]　Per Hage and Frank Harary. Structural models in anthropology. 1983. http://konect.cc/networks/ ucidata-gama/.
[41]　Kenneth E Read. Cultures of the central highlands, new guinea. *Southwestern Journal of Anthropology*, 10(1): 1-43, 1954.
[42]　Samuel F Sampson. A novitiate in a period of change: An experimental and case study of relationships. *Unpublished Ph. D. dissertation, Department of Sociology, Cornell University*, 1968. http://konect.cc/networks/moreno_sampson/.
[43]　Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *arXiv preprint cs/0607062*, 2006. http://konect.cc/networks/convote/.
[44]　Victor Kristof, Matthias Grossglauser, and Patrick Thiran. War of words: The competitive dynamics of legislative processes. In *Proceedings of The Web Conference 2020*, pages 2803-2809, 2020.
[45]　Jure Leskovec and Andrej Krevl. Snap datasets: Stanford large network dataset collection, 2014. https://snap.stanford.edu/data/index.html.
[46]　Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. Stance evolution and twitter interactions in an italian political debate. In *International Conference on Applications of Natural Language to Information Systems*, pages 15-27. Springer, 2018.
[47]　Jérôme Kunegis. Konect: the koblenz network collection. In *Proceedings of the 22nd international conference on world wide web*, pages 1343-1350, 2013.
[48]　Silviu Maniu, Talel Abdessalem, and Bogdan Cautis. Casting a web of trust over wikipedia: an interactionbased approach. In *Proceedings of the 20th international conference companion on World wide web*, pages 87-88, 2011.
[49]　Mingming Chen, Konstantin Kuzmin, and Boleslaw K Szymanski. Community detection via maximization of modularity and its variants. *IEEE Transactions on Computational Social Systems*, 1(1): 46-65, 2014.