Sparsity-aware Robust Community Detection (SPARCODE)

Aylin Taştan, Student Member, IEEE, Michael Muma, Member, IEEE, and Abdelhak M. Zoubir, Fellow, IEEE

Abstract—Community detection refers to finding densely connected groups of nodes in graphs. In important applications, such as cluster analysis and network modelling, the graph is sparse but outliers and heavy-tailed noise may obscure its structure. We propose a new method for Sparsity-aware Robust Community Detection (SPARCODE). Starting from a densely connected and outlier-corrupted graph, we first extract a preliminary sparsity-improved graph model where we optimize the level of sparsity by mapping the coordinates from different clusters such that the distance of their embedding is maximal. Then, undesired edges are removed and the graph is constructed robustly by detecting the outliers using the connectivity of nodes in the improved graph model. Finally, fast spectral partitioning is performed on the resulting robust sparse graph model. The number of communities is estimated using modularity optimization on the partitioning results. We compare the performance to popular graph and cluster-based community detection approaches on a variety of benchmark network and cluster analysis data sets. Comprehensive experiments demonstrate that our method consistently finds the correct number of communities and outperforms existing methods in terms of detection performance, robustness and modularity score while requiring a reasonable computation time.

Index Terms—sparse graphs, community detection, network analysis, graph-based clustering, cluster enumeration.

I. INTRODUCTION

INFERRING a graph model from empirical observations is a fundamental data science task, and a large number of graph construction algorithms have been proposed, e.g. [1]-[4]. In network modelling, graphs are used to represent the interactions between components of a system [5], and in cluster analysis, the similarity between features can be expressed by a weighted graph [4], [6]. Graph models play a crucial role, for example, in subspace learning [3], [7], manifold learning [8], [9] and semi-supervised learning [10]. In particular, the inference of a graph model forms the basis of graph partitioning [11], [12] and community detection algorithms, which has been a very active area of research in recent years [13]-[19].

Community detection refers to finding densely connected groups of nodes, which helps to deduce the underlying structure and relationships that are inherent to the data. A rather important and typical situation is when the data is corrupted by heavy-tailed noise and outliers [20]-[31]. This may lead to a performance degradation for popular graph-based community detection methods, e.g. [13]-[15], [17] using modularity optimization, which is the most widely used objective function for partitioning [12]. A reason for the performance loss is that these methods apply optimization directly on a graph that is possibly corrupted with undesired edges that may be caused by outliers and noise. Therefore, in real-world settings with large and densely connected graphs, including graphs with a considerable amount of outliers, classical community detection methods based on modularity optimization are not be capable of recovering a graph that well-represents the underlying structure of the clean data. For example, methods that use pair-wise Euclidean distances, such as k-nearest neighbor and ε-ball, are particularly sensitive to noise and outliers [1].

Graph-construction algorithms that use sparse representation may provide a performance gain compared to Euclidean distance-based methods [1], [2], [4], [33] and sparsity-based characterization of locality relations can be valuable for cluster analysis [6], [9], [34]. However, the performance of these clustering methods is sensitive to the level of sparsity which is essential in graphs. Sparsity in graphs has been extensively researched, e.g. in terms of the geometry of graphs and there are several approaches [35], [36] for sparsity approximation. Nevertheless, sparse methods are also affected by outliers, and determining the suitable level of sparsity becomes especially challenging. Furthermore, finding such a sparse embedding is often NP hard [37].

A popular approach to promote sparsity is the LASSO regularization, which brings a relaxation for increasing sparsity on a graph without necessitating dimension reduction [38]. However, the performance of the graphical LASSO critically depends on the selection of the penalty parameter that controls the sparsity of the graph. It is well known that the selection of the penalty parameter is a challenging problem in both semi-supervised [39] and unsupervised settings [40], [41]; and often supervised approaches [42], [38] or neighborhood selection [43], are used. An interesting approach to sparsity control using the penalty parameter for graphical LASSO was made in [44] by utilizing knowledge of the number of connected components of the graph. The approach controls the sparsity based on the a priori knowledge of the number of clusters, which may be difficult to estimate in the presence of outliers. To the best of our knowledge, robust sparsity control for graphical LASSO has not been applied to community detection.
To address the above challenges, we propose a new method for Sparsity-aware Robust Community Detection (SPARCODE). The method begins with a densely connected graph and produces a preliminary sparsity improved graph, obtained via an $\ell_1$-penalized precision matrix estimation. We propose a method to optimize the penalty to provide a mapping of the coordinates from different clusters in such a way that they are embedded as far as possible on the real line. Then, undesired and negligible edges are removed from the sparsity improved graph model and the graph construction is performed in a robust manner by detecting the outliers based on connectivity of nodes in the improved sparse graph model. Finally, fast spectral partitioning is performed on the outlier-free vertices of the robust sparse graph model. The number of communities is estimated using modularity optimization on partitions.

The paper is organized as follows. Section II comprises the basic concepts and problem formulation. The proposed community detection algorithm is detailed in Section III. Section IV demonstrates the community detection performance of SPARCODE in comparison to graph-based and cluster-based community detection algorithms using real-world and simulated data. Finally, conclusions are drawn in Section V.

II. PRELIMINARIES AND PROBLEM STATEMENT

A. Summary of Notations

Lower and upper-case bold letters denote vectors and matrices, respectively; $\mathbb{R}$ denotes the set of real numbers; $\mathbb{Z}_+^+$ denotes the set of positive integers; $\mathbb{S}_{+}^{n\times n}$ denotes the set of symmetric nonnegative definite matrices; $|x|$ denotes the absolute value of $x$; $\|x\|$ denotes the norm of vector $x$, e.g. $\|x\|_1$ is $\ell_1$ norm; $x^{(s)}$ denotes sorted vector $x$ where the sorting is in ascending order; $\text{med}(x)$ denotes the median of vector $x$; $x^\top$ denotes transpose of vector $x$; $I$ denotes a vector of ones; $f(.)$ denotes the probability density function; $\text{mod}(.)$ denotes modulo operation; $x^{(0)}$ denotes the initial value of $x$; $\hat{x}$ denotes the estimate of vector $x$; $\hat{X}$ denotes the estimated matrix of outlier-free samples from matrix $X$.

B. Similarity Measures for Graphs

In graph-based clustering, a data set is represented as a weighted graph, where greater weights imply higher similarity between two points. A graph $G = \{V, E, W\}$ is described by a set of vertices $V$, a set of edges $E$, and a weighted adjacency matrix $W$ that is commonly referred to as the affinity (or similarity) matrix. For a data set $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{m \times n}$ with $m$ features and $n$ observations, the associated affinity matrix $W \in \mathbb{R}^{n \times n}$ displays the similarity between pairs of feature vectors. A popular choice, assuming non-negative and symmetrical weight functions is the cosine similarity \[ w_{i,j} = \frac{x_i^\top x_j}{\|x_i\|_2 \|x_j\|_2}, \quad i, j = 1, \ldots, n \tag{1} \]

where $w_{i,j}$ is the similarity coefficient of affinity matrix $W$ for the $i$th and $j$th data vectors $x_i$ and $x_j$, respectively. Alternatively, a graph can be formed, for example, by using Pearson’s linear correlation coefficient, as \[ w_{i,j} = \frac{(\hat{x}_i - \hat{\mu}_i)^\top (\hat{x}_j - \hat{\mu}_j)}{\hat{\sigma}_i \hat{\sigma}_j} \tag{2} \] with associated sample means $\hat{\mu}_i$, $\hat{\mu}_j$, and sample standard deviations $\hat{\sigma}_i$, $\hat{\sigma}_j$, respectively for $i, j = 1, \ldots, n$. In real-world problems with large and noisy data sets, directly inferring clusters from Eq. (1) or (2) may be inefficient, or even infeasible. For such settings, sparse graphs provide a more suitable approach to unveil the underlying data structure. It has been shown that $\ell_1$ graphs can provide a linear sparse representation of the data with respect to an overcomplete dictionary of basis elements \[ \hat{A}. \] If the solution is sufficiently sparse \[ \hat{A}, \] the problem can be generalized over all atoms in $X$, using the data matrix itself as dictionary \[ \hat{A}. \]

C. Problem Statement

Given a data set $X \in \mathbb{R}^{m \times n}$, the aim of this work is to find a label vector $c_K \in \mathbb{R}^{n}$ that partitions $X$ into $K$ independent and mutually exclusive communities. The true community number $K$ is assumed to be unknown. Further, we assume that $X \in \mathbb{R}^{m \times n}$ may be subject to heavy-tailed noise and outliers, which obscure the data structure. Computational efficiency is also of practical interest. Summarizing, the overall aim is to develop a fast and robust clustering algorithm based on computing a sparse graph model.

III. SPARSITY-AWARE ROBUST COMMUNITY DETECTION (SPARCODE)

The main ideas of the proposed method are briefly summarized to provide a general understanding before going into the details. A high-level flow diagram to illustrate its key steps is provided in Fig. I. The community detection problem is addressed as spectral partitioning of sparse graphs. The proposed algorithm, which we call Sparsity-aware Robust Community Detection (SPARCODE), starts with a densely connected weighted graph. Assuming that edges within the communities are more densely connected than the remaining edges, the first step is to increase the sparsity of a given graph by pushing towards zero the similarity coefficients that belong to undesired and negligible edges. The sparsity improved graph model is obtained via an $\ell_1$-penalized precision matrix estimation. The penalty parameter is optimized to provide a mapping of the coordinates from different clusters in such a way that they are embedded as far as possible on the real line. In particular, we split the Fiedler vector and optimize according to what we call the polarization score. Then, undesired and negligible edges are removed from the sparsity improved graph model by applying a threshold, and the graph construction is performed in a robust manner. The outliers, which are represented as red points in Fig. I are detected based on connectivity of nodes in the improved sparse graph model. Finally, fast spectral
graphical LASSO estimates, do generally not satisfy
\( \hat{\Theta}(\hat{\rho}) \) of an undirected Gaussian graph is equivalent to the recover y
\( \Theta \) of the support of the precision matrix
proved graph model. Therefore, a sparsely connected graph
is not relevant for reconstructing edges in the sparsity im-
provement of sparsity in Eq. (4). This provides us with an initial robust
sparse graph model that we will improve in the following step.

According to Eq. (4), the resulting graph structure critically depends on the value of \( \rho \). A key contribution of this work is to provide a simple but effective clustering-oriented answer to how to optimize \( \rho \). The intuition behind our approach is to determine a value for \( \rho \) by mapping the coordinates from different communities in such a way that they lie as far as possible on the real line.

According to \cite{49}, the mapping of vertices onto the real line by minimizing the sum of the edge lengths while maintaining an average unit Euclidean distance between random pairs of points may lead to an excellent partition by cutting the line at a random point. Achieving such a mapping is NP hard, though there are several approximations in literature that achieve the sparsest cut \cite{35}, \cite{37}, \cite{48}. SPARCODE relies on spectral partitioning methods, which use a relaxation to map graph vertices onto the real line, providing connected points stay as close together as possible using squared Euclidean distance \cite{49}. Let \( \rho_k \in \{\rho_{\min}, \ldots, \rho_{\max}\} \) denote the \( k \)th candidate penalty parameter in Eq. (4). Then, the embedding result \( \hat{y}_k \in \mathbb{R}^n \) of the \( k \)th candidate penalty parameter \( \rho_k \) can be approximated by minimizing the following objective function as in \cite{49},

\[
\hat{y}_k = \arg\min_{y_k} \frac{1}{2} \sum_{i,j} \|y_i - y_j\|^2 w_{i,j}(\rho_k) \text{s.t. } y^T D_k y = 1, \quad y^T D_k I = 0, \tag{5}
\]

where \( D_k \in \mathbb{R}^{n \times n} \) is a diagonal weight matrix of \( W(\rho_k) \) for a given \( \rho_k \), with weights \( d_{i,i} = \sum_j w_{i,j}(\rho_k) \) on the diagonal. The vector estimate \( \hat{y}_k = (\hat{y}_1, \ldots, \hat{y}_n)^T \) is known as the Fiedler vector and it shows the algebraic connectivity of a graph \cite{50}. The graph can be partitioned into two subsets by splitting the Fiedler vector, such that \( \hat{y}_i \in \hat{y}_{k,1} \) for \( \hat{y}_i \leq 0 \) and \( \hat{y}_i \in \hat{y}_{k,2} \) otherwise. Here, \( \hat{y}_{k,m} \) denotes the \( m \)th subset of \( \hat{y}_k \) with \( m = 1, 2 \) and \( i = 1, \ldots, n \) \cite{51}.

Based on the Fiedler vector, we propose to measure the polarization score for each candidate \( \rho_k \) by evaluating

\[
P_{sc}(\rho_k) = \text{med}(\hat{y}_{k,1}) - \text{med}(\hat{y}_{k,2}), \tag{6}
\]

where \( \text{med}(\hat{y}_{k,m}) \) denotes the median of the \( m \)th subset for \( m = 1, 2 \). The median is used as a robust location estimate \cite{21}, \cite{26} for each subset \( \hat{y}_{k,m} \). Given a set of candidate penalty parameters \( \rho_k \in \{\rho_{\min}, \ldots, \rho_{\max}\} \), we estimate \( \rho \) by maximizing the polarization score as

\[
\hat{\rho} = \arg\max_{\rho=\rho_{\min}, \ldots, \rho_{\max}} S(\rho), \tag{7}
\]

where \( \hat{\rho} \) denotes the estimated penalty parameter that provides maximum polarization in the given space. To reduce computational complexity, we first evaluate Eq. (7) on a coarse grid and then use cubic spline interpolation between \( \hat{\rho} \) and its neighboring samples to find the final value \( \hat{\rho} \). By means of Eq. (7), we propose a problem-dependent tuning of the level of sparsity in Eq. (4). This provides us with an initial robust sparse graph model that we will improve in the following step.
B. Compute a Robust Sparse Graph Model

By maximizing the polarization score via Eq. (7), we estimated a graph model \( \hat{G} = \{ V, \tilde{E}, \tilde{W}(\hat{\rho}) \} \), for which the edges between different communities are large, while the edges within the community are small. However, the true graph structure may still be hidden due to noise and outliers. It is intuitively clear that sparse outlying entries and noise have fewer non-zero coefficients in the affinity matrix compared to typical data points, and the value of these coefficients is negligibly small for noise [3]. SPARCODE builds upon this graph property to detect outliers and noisy feature vectors by analyzing the connectivity of graph nodes.

Let \( \hat{W}^{(s)}(\hat{\rho}) = [\hat{w}_1^{(s)}(\hat{\rho}), \ldots, \hat{w}_n^{(s)}(\hat{\rho})] \in \mathbb{R}^{n \times n} \) be the set of column-wise sorted similarity vectors whose \( j \)th element denotes the sorted \( j \)th similarity vector of the estimated affinity matrix \( \hat{W}(\hat{\rho}) \). Then, we obtain the aggregated set of similarity coefficients \( u = \{ u_1, u_2, \ldots, u_n \} \in \mathbb{R}^n \) by computing

\[
u_i = \frac{1}{n} \sum_{j=1}^{n} \hat{w}_{i,j}^{(s)}(\hat{\rho}), \quad i = 1, \ldots, n, \tag{8}\]

where \( u_i \) denotes the \( i \)th element of \( u \) and \( \hat{w}_{i,j}^{(s)}(\hat{\rho}) \) is the element of \( \hat{W}^{(s)}(\hat{\rho}) \) that belongs to the \( i \)th row and \( j \)th column. From Eq. (8), a two-mode Gaussian mixture distribution of the \( i \)th similarity coefficient \( u_i \) can be written as

\[
f(u_i|\lambda) = \sum_{l=1}^{2} \tau_l g(u_i; \mu_l, \Sigma_l) \tag{9}\]

where \( \lambda = \{ \mu_l, \Sigma_l \} \) denotes the parameter set of the model for \( l = 1, 2 \) and \( g(u_i; \mu_l, \Sigma_l) \) is the \( r \)-variate Gaussian probability density function with mean \( \mu_l \) and covariance \( \Sigma_l \), and \( \tau_l \) is the mixing coefficient of the \( l \)th subset. The probability estimate that \( u_i \) belongs to the \( i \)th subset, with \( i = 1, 2, \ldots, n \) and \( l = 1, 2 \), can be calculated as

\[
\tilde{v}_{i,l} = \frac{\tau_l g(u_i; \mu_l, \Sigma_l)}{\sum_{j=1}^{2} \tau_j g(u_i; \mu_j, \Sigma_j)} \tag{10}\]

where \( \tilde{v}_{i,l} \) denotes the probability estimate of \( u_i \) that belongs to the \( l \)th subset. Then, the threshold can be simply evaluated as the interconnection point of the two distributions, i.e.

\[
T = \arg \min_{u_i} |\tilde{v}_{i,2} - \tilde{v}_{i,1}|. \tag{11}\]

The effect of applying the threshold defined in Eq. (11), is to remove the edges from the estimated graph model \( \hat{G} = \{ V, \tilde{E}, \tilde{W}(\hat{\rho}) \} \) so that the feature vectors in data matrix \( X \) which have fewer nonzero coefficients with negligibly small values will have a zero degree in the graph. In this way, outliers and noisy feature vectors are detected and removed, resulting in a cleaned data matrix \( \hat{X} \in \mathbb{R}^n \) and a graph \( \hat{G} = \{ V, \hat{E}, \hat{W}(\hat{\rho}) \} \).

C. A Fast Spectral Partition based on the Robust and Sparse Graph

Assuming that \( \hat{G} = \{ V, \hat{E}, \hat{W}(\hat{\rho}) \} \) is cleaned from outliers and that it is sufficiently sparse, the graph can be partitioned into \( K \) communities based on the Fiedler vector with a fast partitioning method. In practice, however, \( K \) is unknown and must be estimated. We therefore present an approach to estimate the range \( K_{\min} \leq K \leq K_{\max} \) using typical degrees in the graph.

Let \( d \in \mathbb{Z}^{\tilde{n}} \) denote a degree vector whose \( i \)th element corresponds to the degree of the \( i \)th feature vector in \( \hat{X} \in \mathbb{R}^n \). Further, let \( p \in \mathbb{R}^n \) be the empirical probability of occurrence of these degrees in \( d \). Finally, let \( h \in \mathbb{R}^{\tilde{n}/2} \) denotes the vector of degrees, whose probability is greater than the median of probabilities, i.e. \( \text{med}(p) \). Now, the minimum and maximum number of communities \( \{ K_{\min}, K_{\max} \} \in \mathbb{Z}^+ \) can be estimated as

\[
K_{\min} \approx \frac{\tilde{n}}{h_{\max} + 1} \quad \text{and} \quad K_{\max} \approx \frac{\tilde{n}}{h_{\min} + 1} \tag{12}\]

where \( h_{\min} + 1 \) and \( h_{\max} + 1 \) represent the minimum and maximum number of nodes in each community based on \( h_{\min} = \min \{ h \} \) and \( h_{\max} = \max \{ h \} \), respectively. The intuition underlying Eq. (12) is to define a range for a candidate number of communities from the typical connectedness of the graph.

Now, for a set of estimated candidate communities \( K_{\text{cand}} \in \{ K_{\min}, \ldots, K_{\max} \} \), the graph can be partitioned using the sorted Fiedler vector \( \tilde{v}^{(s)} \in \mathbb{R}^n \) of \( \hat{G} = \{ V, \hat{E}, \hat{W}(\hat{\rho}) \} \) as follows. To split the graph into an initial balanced partitioning, where each community has the same number of members, we first compute \( q_{\text{cand}} = \min(h, K_{\text{cand}} \}. \) We identify a number \( q_{\text{cand}} \) of mapping results that we ignore in the initial balanced partitioning. These are the ones for which the standard deviation of the set containing its two immediate neighbors in the sorted Fiedler vector is largest. Then, using the remaining mapping results, the initial partition can be realized by splitting the remaining mapping results into \( K_{\text{cand}} \) equal communities. Finally, the residual \( q_{\text{cand}} \) mapping results can be assigned by minimizing the distance to the center of the initially defined communities on the line. Now, the estimated label vector of a given candidate number of community \( \tilde{c}_{K_{\text{cand}}} \in \mathbb{R}^n \) can be computed for all feature vectors of \( \hat{X} \).

Finally, the number of communities \( K \) can be estimated by comparing in terms of their modularity, the quality of different partitions of the robust graph model i.e.

\[
K = \underset{K_{\text{cand}} = K_{\min}, \ldots, K_{\max}}{\text{arg max}} \{ Q_{K_{\text{cand}}} \}, \tag{13}\]

where

\[
Q_{K_{\text{cand}}} = \frac{1}{2g} \sum_{j,m} \left[ \tilde{w}_{j,m} - \tilde{v}_j \tilde{v}_m \right] \delta(\tilde{c}_j, \tilde{c}_m) \tag{14}\]

is the modularity score [22], [33] for the \( i \)th candidate number of communities, \( \tilde{w}_{j,m} \) denotes the weight of the edge between the \( j \)th and \( m \)th feature vector of \( \hat{X} \), \( \tilde{c}_j = \sum_m \tilde{w}_{j,m} \) is the sum of the weights of edges attached to vertex \( j \), \( \tilde{c}_m \) is the estimated community label associated to vertex \( j \), \( g = \frac{1}{2} \sum_{j,m} \tilde{w}_{j,m} \) and the function \( \delta(\tilde{c}_j, \tilde{c}_m) \) equals 1 if \( \tilde{c}_j = \tilde{c}_m \) and is zero otherwise.
SP ARCODE is benchmarked against state-of-the-art community detection algorithms. We consider a variety of clustering and graph partitioning data sets to demonstrate the applicability of SP ARCODE in both of these community detection settings. We select as competitors the following community detection algorithms: Newman’s Greedy Algorithm (NGA) [13], Le Martelot [14], Newman’s eigenvector method [15], singular value decomposition-based community detection (SVD) [16], the Louvain method [17], the Bayesian approach (BC) [18] and the Combo method [19]. The SVD method is applicable only on bipartite graphs, for details see [16]. Bayesian cluster enumeration (BCE) [53], dip-means and kernel dip-means (K. dip-means) [55], x-means [56], Gaussian k-means (g-means) [57] and density based spatial clustering (DBSCAN) [58] are used as cluster-based competing approaches. The performance measures are evaluated both on synthetic and real-world data sets.

All SP ARCODE implementations use the same default parameters as follows: \( \rho_{\text{min}}^{(0)} = 0.1, \rho_{\text{max}}^{(0)} = 0.99, \ N_{\rho}^{(0)} = N_{\rho} = 5 \). A MATLAB code for SP ARCODE is available at: https://github/SPARCODE

\section*{A. Performance Measures}

The empirical probability of detection \( p_{\text{det}} \), the modularity score \( Q \) and the computation time \( t \) are used as performance measures. The empirical probability of detection is estimated as

\[
p_{\text{det}} = \frac{1}{N_E} \sum_{i=1}^{N_E} \mathbb{I}_{\{K=K\}},
\]

where \( N_E \) denotes the total number of performed experiments, \( K \) is the estimated number of communities and \( \mathbb{I}_{\{K=K\}} \) is the indicator function that is defined as

\[
\mathbb{I}_{\{K=K\}} = \begin{cases} 1, & \text{if } K = K \\ 0, & \text{otherwise} \end{cases}
\]

The modularity score of the estimated community number \( Q_K \) can be calculated using Eq. (14). The modularity score of SP ARCODE is computed based on the affinity matrix of the sparsity improved robust graph model.

\section*{B. Synthetic Graph Model 1: Correlated Communities Study with \( K=7 \)}

An undirected weighted graph model is created, consisting of seven communities with different sample sizes for each community. The first six communities inhibit correlation between communities in addition to correlations within the community. In contrast, the seventh community only has correlations within the community. Moreover, the density of the seventh community is assumed to be higher compared to that of the other six. For illustration purposes, the graph for the designed affinity matrix is shown in Fig. 2a. All parameters to generate the data set are given in Appendix A.

Tab. I summarizes the community detection performance results for Scenario 1 in terms of the estimated community number, the modularity score and the computation time. The considered setting is challenging for most algorithms, with the exception of that in [18], and all benchmark community detection algorithms underestimate the number of clusters due to the correlations between communities. SP ARCODE correctly finds the number of communities and outperforms the method in [18] both in terms of the modularity score and the computation time. As shown in Fig. 2b, the estimated sparsity improved robust graph model clearly partitions the network into seven communities.
C. Synthetic Graph Model 2: Robustness Study with $K = 3$

An undirected weighted graph consisting of three communities in the presence of outliers is created. The communities are correlated with each other in addition to exhibiting strong correlations within. The outliers correlate equally with all communities with negligibly small correlation coefficients. The graph for the designed affinity matrix of Scenario 2 is shown in Fig. 3a, all parameters to generate the data set given in Appendix B. While the NE, NGA and Bayesian community detection algorithms overestimate the community number because of outliers, the remaining competitor community detection methods estimate the community number correctly. However, for the competing methods, the outliers cause a noticeable drop in modularity. By contrast, the SPARCODE algorithm, overcomes this problem by detecting these outliers in the graph model and providing robustness by minimizing the effect of outliers on the quality of the partition. Fig. 3b shows based on the robust sparse graph model, that a separation of the communities is possible and outliers can easily be distinguished from other data points.

D. Executing Time

The executing time is reported a function of the number of nodes in the network. All experiments are performed based on Scenario 2 that was explained in the previous section. All implementations are in MATLAB using the default parameters given by the authors, except for the Combo algorithm, for which we use the available C implementation. The results are summarized in Fig. 4. For sample sizes up to 1000 nodes, SPARCODE is comparable to Combo (see Fig. 4a), though for large networks, the executing time of SPARCODE grows quicker (see Fig. 4b).

E. Real-World Benchmark Graph Models

The performance is benchmarked on the following seven real-world networks:

- **Zachary’s Karate Club (Karate):** The network is a social bipartite network that consists of friendship between 34 members of a karate club [59].
- **Dolphins:** The network is a social network that consists of social interactions of 62 dolphins with $K = 2$ communities based on the reaction after a dolphin left from the group [60].
- **American College Football (Football):** The network represents 115 US college teams and the games that they played [52]. The ground truth for the community number is 12.
- **Political Blogs (P. Blogs):** The network consists of blogs about US politics with 1490 nodes and $K = 2$ communities as “liberal” and “conservative” [61].
- **Jazz Musicians (Jazz):** The network consists of jazz musicians which can be divided into $K = 4$ communities based on cities where bands are recorded [62].
- **Carpinteria Salt Marsh (C.S.M.):** The network is a type of food web which can be divided into $K = 2$ based on species as parasites and free-living or $K = 4$ subwebs based on links, e.g. parasites-parasites, predator-parasites [63].
- **Caenorhabditis Elegans (C. Elegans):** The biological network examines the neuronal layout of C. Elegans for 279 neurons that can be partitioned into $K = 3$ communities [64].

The cosine similarity, defined in Eq. (1), is used to obtain the affinity matrix $W$ for all networks except for the political blogs network, which is examined with Pearson’s linear correlation coefficients as designed in Eq. (2) because none of the algorithms estimated the community number correctly with the cosine similarity.
TABLE III: Performance of graph-based approaches on well-known networks. The results whose computation take more than 12 hours and nontarget networks for SVD method denoted as "-".

| Network  | KGA | Martelot | NE  | SVD | Louvain | Bayesian | Combo | SPARCODE | K |
|----------|-----|----------|-----|-----|---------|----------|-------|----------|----|
| Karate   | 2   | 2        | 3   | 2   | 2       | 3        | 2     | 2        | 2  |
| Dolphins | 2   | 3        | 6   | 2   | 4       | 2        | 4     | 2        | 2  |
| Football | 7   | 4        | 8   | -   | 7       | 5        | 7     | 8        | 12 |
| P. Blogs | -   | 541      | 1   | 1   | 505     | 7        | 4     | 2        | 2  |
| Jazz     | 20  | 21       | 7   | -   | 21      | 10       | 3     | 4        | 4  |
| C.S.M.   | 10  | 12       | 8   | 1   | 13      | 10       | 5     | 2        | 2,4|
| C. Elegans| 81  | 3        | 10  | -   | 3       | 5        | 3     | 3        | 3  |

The community detection results are summarized in Tab. III. The results include the estimated community number for all cases for which the computation time was lower than 12 hours. The community detection results of the SVD approach are given for bipartite networks, only. As can be seen, the SPARCODE algorithm shows the best overall performance, and estimates the community number correctly for all networks, except for the Football network, for which it nevertheless provides the value that is closest to the true community number.

F. Real-World Radar-Based Gait Analysis Data Sets:

The performance is benchmarked on the following two radar data sets:

**Human Gait Data Set:** As detailed in [65], the experimental data was collected in an office environment at Technische Universität Darmstadt using a 24 GHz radar system [66]. The data can be grouped into $K = 5$ different object communities, which are normal walk (‘Normal’), limping with one leg (Limping one), limping with two legs (‘Limping two’), walking with a cane (‘Cane’) and walking with a cane out of synchronisation (‘Cane out of sync.’). The data contains 16 measurements per subject (8 towards, 8 away from the radar) and ten subjects. The duration of each measurement is equal to six seconds. In total, 800 measurements for five different gait communities of ten subjects were used in our experiments. As an illustration, a scatter plot of three important features of the radar-based human gait data is shown in Fig. 5. It can be seen that the outlying observations corresponding to the ‘Cane’ community overlap with the true observations of ‘Normal’. Moreover, the outliers of ‘Limping two’ have a considerable sample size and strong correlations with ‘Normal’, which makes it a challenging scenario for any community detection algorithm.

In order to represent the underlying structure, for SPARCODE, similarity measures as it is explained in Eq. (3) and Eq. (2) are used respectively, and the resulting graph is shown in Fig. 6a. Clearly, the separation into communities from such a graph is extremely difficult. In contrast, the sparsity improved graph model shown in Fig. 6b better reveals separated communities.
Fig. 5: Scatter plot for three important features of radar-based human gait data belonging to five object communities.

Fig. 6: Graphical models of human gait data set.

Fig. 7: Scatter plot for the first three principal components (PC) of person identification data over four object communities.

(a) Initial graph model
(b) Estimated sparsity improved graph model

Fig. 8: Graphical models of person identification data set.

(a) Initial graph model
(b) Estimated sparsity improved graph model

**Person Identification Data Set:** As detailed in [67], the experimental data has been collected in an office environment at Technische Universität Darmstadt using the same radar system as explained for the previous data set. The data can be grouped into $K = 4$ object communities, each representing a different person. The data set includes the measurements of four test subjects that are walking slowly and without swinging their arms, towards and away from the radar. The duration of recordings is equal to six seconds and the number of measurements is equal to 13 for each person. In total, 187 stride pairs that are obtained from 52 observations of the four subjects are used in community detection.

The scatter plot of the person identification data set containing $K = 4$ different persons is shown in Fig. 7. Even though, judging based on visual impression, the examination of the person identification data set seems easier in comparison to human gait data, it still has communities that correlate with each other which makes the detection of the community number very difficult. Just like for the previous radar data set, SPARCODE uses similarity measures as defined in Eq. (3) and Eq. (2), respectively. The initial and the sparsity improved robust graph models of the person identification data set are shown in Fig. 8a and Fig. 8b, respectively. As can be seen, the estimated sparsity improved graph model is separable into four communities by eliminating the outliers are of zero degree. Therefore, a simple graph partitioning method is sufficient to partition such a robust sparse graph model into the correct number of communities.

**G. Real-World Cluster Analysis Benchmark Data Sets:**

In this section, we benchmark the performance of SPARCODE on eight well-known data sets from the UCI Machine Learning Repository. These are:

- **Fisher’s Iris Data Set (Iris):** The iris data set includes 150 observations from three species of the Iris flower [68].
- **Ionosphere Data Set:** The data set includes 351 radar returns from the ionosphere in order to define quality for further analysis [69]. The subspace number is equal to two.
- **Parkinson Acoustic Data Set:** The data set consists of replicated acoustic features of Parkinson’s disease with 240 instances from two communities which are “healthy” and “patient” [70].
- **Diabetic Retinopathy Debrecen Data Set (D. Retinopathy):** The data set includes image-based features of diabetic retinopathy with 1151 observations from two object communities [71].
- **Connectionist Bench Data Set (Sonar):** The data set includes 208 observations of $K = 2$ communities based on sonar
9

Table IV: Performance of graph-based approaches on clustering data sets. The results whose computation take more than 12 hours and nontarget networks for SVD method denoted as "-".

| Data Set                  | NGA | Martelot | NE | SVD | Louvain | Bayesian Combo | SP ARCODE | K |
|---------------------------|-----|----------|----|-----|---------|----------------|-----------|---|
| Human Gait [65]           | -   | 3        | 1  | -   | 3       | 5              | 3         | 5 |
| Person Identification [67]| 9   | 11       | 1  | -   | 8       | 8              | 5         | 4 |
| Iris [68]                 | 1   | 2        | 2  | -   | 2       | 2              | 3         | 3 |
| Ionosphere [69]           | 109 | 106      | 1  | 1   | 101     | 10             | 6         | 2 |
| Parkinson Acoustic [70]   | 1   | 1        | 55 | 1   | 1       | 2              | 1         | 2 |
| D. Retinopathy [71]       | -   | 2        | 2  | 1   | 2       | 2              | 2         | 2 |
| Sonar [72]                | 1   | 2        | 28 | 1   | 3       | 2              | 2         | 2 |
| QSAR Bioconcentration [73]| -   | 3        | 1  | -   | 3       | 7              | 3         | 3 |
| Cardiotocography [74]     | -   | 2        | 206| -   | 2       | 15             | 2         | 3 |
| Divorce Predictors [75]   | 3   | 3        | 1  | 1   | 3       | 7              | 2         | 2 |

returns collected from a metal cylinder and a cylindrically shaped rock positioned on a sandy ocean floor [72].

QSAR Bioconcentration Classes Data Set (QSAR Bioconcentration): The data set consists of the bioconcentration factor of 779 chemicals to determine mechanism of bioconcentration [73]. The data set can be partitioned into three communities.

Cardiotocography Data Set: The Cardiotocography data set consists of 2126 observations of fetal cardiotocograms which can be partitioned into three communities in terms of fetal state or ten communities based on morphologic pattern [74].

Divorce Predictors Data Set: The data set consists of 170 observations from two object communities using divorce predictors scale [75].

1) Comparisons with Graph-based Approaches: The graphs for all clustering data sets are designed using Pearson’s linear correlation coefficients as in Eq. (2), except for the Ionosphere and Cardiotocography data sets, where the graphs are designed as the same procedure with human gait and person identification data sets.

The estimated community numbers, for all community detection algorithms whose computation time is less than 12 hours, are summarized in Tab. IV. Again, the SVD is only applicable for bipartite networks. SP ARCODE correctly estimates the number of clusters for all data sets and outperforms all its competitors. None of the competitor community detection algorithms is able to correctly estimate the community number of the person identification data set correctly. These overestimate it in most cases, which can be explained by the considerable number of outliers.

In addition to the accuracy in terms of the estimated number of communities, the results are also evaluated with respect to partitioning quality and computation time. To report the modularity and computation time in the same scale as the probability of detection, the performances of all algorithms are ranked and then averaged over all real data sets. The performance measures of the SVD approach are evaluated on the bipartite graphs, only. All results are summarized in Fig. 9. As can be seen from the figure, SP ARCODE achieves the best performance with 0.94 probability of detection while the nearest competitor follows with 0.35 empirical probability of detection.

Based on quality of partition, NE achieves the best performance with averaged rank score of 0.72; SP ARCODE, Louvain and Combo follow with 0.64 and 0.61 for both, respectively. Although NE shows the best performance on modularity, it has the worst performance with considerable difference in terms of its probability of detection. NE, Louvain and SVD methods are the best algorithms with respect to computation time where Martelot, Combo and SP ARCODE follow with 0.98, 0.94 and 0.93, respectively.

To summarize the performance of graph-based community detection approaches on real data sets, the detailed performance measures that is explained in Fig. 9 are further aggregated by equally weighting each performance measure. The overall performance of eight competitor methods is summarized in Fig. 10. As can be seen, SP ARCODE achieves the highest overall performance score of 0.84 whereas, Combo is the best competitor with a score of 0.64. Martelot, Louvain and NE follow with an overall performance score of 0.62, 0.61 and 0.59, respectively.
for Different Community Detection Methods

| Data Set                        | BCE | dip-means | K. dip-means | x-means | g-means | DBSCAN | SPARCODE | K |
|--------------------------------|-----|-----------|--------------|---------|---------|--------|----------|---|
| Human Gait [65].               | 7   | 4         | 3            | 176     | 9       | 6      | 5        | 5 |
| Person Identification [67].    | 4   | 4         | 3            | 22      | 10      | 3      | 4        | 4 |
| Iris [68].                     | 3   | 2         | 2            | 31      | 3       | 2      | 3        | 3 |
| Ionosphere [69].               | 2   | 31        | 1            | 32      | 2       | 2      | 2        | 2 |
| Parkinson Acoustic [70].       | 1   | 1         | 1            | 47      | 1       | 1      | 2        | 2 |
| D. Retinopathy [71].           | 1   | 2         | 2            | 199     | 16      | 1      | 2        | 2 |
| Sonar [72].                    | 1   | 1         | 1            | 29      | 1       | 1      | 2        | 2 |
| QSAR Bioconcentration [73].    | 5   | 4         | 6            | 174     | 55      | 1      | 3        | 3 |
| Cardiotocography [74].         | 8   | 4         | 4            | 435     | 51      | 2      | 3        | 3, 10 |
| Divorce Predictors [75].       | 1   | 2         | 2            | 21      | 3       | 1      | 2        | 2 |

TABLE V: Performance of cluster-based approaches on clustering data sets. The results whose computation take more than 12 hours and nontarget networks for SVD method denoted as "-".

Fig. 10: Performance of different algorithms based on equal weights on performance metrics.

2) Comparisons with Cluster-based Approaches: In this section, the SPARCODE algorithm is compared with six well-known clustering methods that estimate the number of communities \( K \). The community detection results of different approaches are summarized in Tab. V. As can be seen, none of the competitor algorithms performs well in the highly contaminated human gait data set. However, generally speaking, cluster-based approaches show better performance compared to competitor graph-based approaches on person identification, iris and ionosphere data sets, except for x-means, which generally largely overestimates the number of communities.

The overall community detection performance of all algorithms is summarized in Fig. 11. The SPARCODE method clearly outperforms existing cluster-based community detection approaches in terms of probability of detection and outperforms the best competitors (BCE and dip-means), which have a performance score of 0.3, by a large margin.

Fig. 11: Performance of different algorithms based on rank in terms of probability of detection.

V. CONCLUSION

We proposed SPARCODE, a community detection method that uses spectral partitioning based on estimating a robust and sparse graph. The level of sparseness is controlled by maximizing the modularity of the graph. SPARCODE includes a graph construction-based outlier detection method to increase robustness. Overall, when compared to both cluster-based and graph-based community detection algorithms on real and synthetic data sets, SPARCODE achieves the highest community detection performance providing a high quality of partition at a reasonable computation time compared to existing graph-based approaches. Judging from its applicability in a large variety of problems, compatibility with different graph-based methods and success in highly contaminated data sets, SPARCODE is a promising new algorithm for performing community detection in a robust manner.
APPENDIX A

The first scenario considers a weighted graph with \( n = 300 \) nodes that belong to seven different communities. The affinity matrix \( W \in \mathbb{R}^{n \times n} \), whose entries vary between zero and one, is symmetric, zero diagonal and nonnegative. The communities are labelled as C1, \ldots, C7, and the parameters of similarity coefficients within the communities and between different communities are summarized in Tab. VI. The number of objects for each community is denoted as \( n_C \). The similarity coefficients are generated as \( w = \mu + r \sigma \) where \( r \) is a random real number distributed as \( U[0, 1) \).

|     | C1   | C2   | C3   | C4   | C5   | C6   | C7   | \( n_C \) |
|-----|------|------|------|------|------|------|------|----------|
| C1  | \( \mu \) | 0.8  | 0.005| 0.35 | 0    | 0    | 0.25 | 45       |
|     | \( \sigma \) | 0.19 | 0.0045| 0.1  | -    | -    | 0.05 |          |
| C2  | \( \mu \) | 0.005| 0.1  | 0    | 0.35 | 0.19 | 0    | 40       |
|     | \( \sigma \) | 0.0045| 0.1  | -    | 0.1  | 0.15 | -    |          |
| C3  | \( \mu \) | 0.35 | 0    | 0.9  | 0    | 0    | 0    | 55       |
|     | \( \sigma \) | 0.1  | 0.09 | -    | -    | -    | -    |          |
| C4  | \( \mu \) | 0    | 0.35 | 0.85 | 0    | 0    | 0    | 42       |
|     | \( \sigma \) | 0.1  | -    | 0.14 | -    | -    | -    |          |
| C5  | \( \mu \) | 0.25 | 0    | 0    | 0    | 0.85 | 0    | 46       |
|     | \( \sigma \) | 0.05 | -    | -    | 0.15 | -    | -    |          |
| C6  | \( \mu \) | 0    | 0    | 0    | 0    | 0    | 0.95 | 35       |
|     | \( \sigma \) | -    | -    | -    | -    | -    | 0.04 |          |

**TABLE VI**: Similarity coefficients for seven object communities for Scenario 1.

**APPENDIX B**

The second scenario considers a weighted graph with \( n = 300 \) nodes. In contrast, the community number is \( K = 3 \) and we investigate the effect of outliers by creating outliers which have negligibly small correlation with all communities. The communities are labelled as C1, \ldots, C3, and the outliers as O, and the associated parameters of similarity coefficients within and between communities and outliers are summarized in Tab. VII. Both the number of objects of communities and outliers are denoted as \( n_C \). The similarity coefficients are generated as \( w = \mu + r \sigma \) where \( r \) is a random real number distributed as \( U[0, 1) \) as in the first scenario.

|     | C1   | C2   | C3   | \( n_C \) |
|-----|------|------|------|----------|
| C1  | \( \mu \) | 0.7  | 0.01 | 0       | 66       |
|     | \( \sigma \) | 0.29 | 0.29 | -       |          |
| C2  | \( \mu \) | 0.01 | 0.8  | 0.27   | 102      |
|     | \( \sigma \) | 0.29 | 0.19 | 0.1    |          |
| C3  | \( \mu \) | 0    | 0.27 | 0.7    | 90       |
|     | \( \sigma \) | -    | 0.24 | 0.05   |          |
| O   | \( \mu \) | 0.005| 0.005| 0.005  | 42       |
|     | \( \sigma \) | 0.025| 0.025| 0.025  |          |

**TABLE VII**: Similarity coefficients for three object communities and outliers for Scenario 2.

ACKNOWLEDGMENT

The work of A. Taştan is supported by the Republic of Turkey Ministry of National Education. The work of M. Muma has been funded by the LOEWE initiative (Hesse, Germany) within the emergenCYT centre and is supported by the ‘Athene Young Investigator Programme’ of Technische Universität Darmstadt, Hesse, Germany.

REFERENCES

[1] B. Cheng, J. Yang, S. Yan, Y. Fu and T. S. Huang, “Learning with \( \ell^1 \)-graph for image analysis,” *IEEE Trans. Image Process.*, vol. 19, pp. 858-866, 2009.
[2] Q. Liu, Y. Sun, C. Wang, T. Liu and D. Tao, “Elastic net hypergraph learning for image clustering and semi-supervised classification,” *IEEE Trans. Image Process.*, vol. 26, pp. 452-463, 2016.
[3] E. Elhamifar and R. Vidal, “Sparse subspace clustering: Algorithm, theory, and applications,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 2765-2781, 2013.
[4] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, pp. 210-227, 2008.
[5] F. Fages, S. Soliman, and N. Chabrier-Rivier, “Modelling and querying interaction networks in the biochemical abstract machine BIOCHAM,” *J. Biol. Phys. Chem.*, vol. 4, pp. 64-73, 2004.
[6] A. Taştan, M. Muma and A. M. Zoubir, “An unsupervised approach for graph-based robust clustering of human gait signatures,” in *Proc. 2020 IEEE Radar Conf. (accepted)*.
[7] J. M. Hofman and C. H. Wiggins, “Bayesian Approach to Network Modularity,” *Phys. Rev. Lett.*, vol. 83, pp. 016107, 2011.
[8] M. E. J. Newman, “Fast algorithm for detecting community structure in networks,” *Phys. Rev. E*, vol. 69, pp. 066133, 2004.
[9] E. L. Martelot and C. Hankin, “Multi-scale community detection using stability as optimization criterion in a greedy algorithm,” in *Proc. Intl. Conf. Knowl. Discovery and Inf. Retrieval*, pp. 208-217, 2011.
[10] M. E. J. Newman, “Finding community structure in networks using the eigenvectors of matrices,” *Phys. Rev. E*, vol. 74, pp. 036104, 2006.
[11] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 888-905, 2000.
[12] B. Karrer and M. E. J. Newman, “Stochastic blockmodels and community structure in networks,” *Phys. Rev. E*, vol. 83, pp. 016107, 2011.
[13] M. E. J. Newman, “Fast algorithm for detecting community structure in networks,” *Phys. Rev. E*, vol. 69, pp. 066133, 2004.
[14] F. Robust statistics: the approach based on influence functions, *Robust statistics: the approach based on influence functions*, John Wiley & Sons, 2011.
[15] M. E. J. Newman, “Finding community structure in networks using the eigenvectors of matrices,” *Phys. Rev. E*, vol. 74, pp. 036104, 2006.
[16] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 888-905, 2000.
[17] B. Karrer and M. E. J. Newman, “Stochastic blockmodels and community structure in networks,” *Phys. Rev. E*, vol. 83, pp. 016107, 2011.
[18] M. E. J. Newman, “Fast algorithm for detecting community structure in networks,” *Phys. Rev. E*, vol. 69, pp. 066133, 2004.
[19] E. L. Martelot and C. Hankin, “Multi-scale community detection using stability as optimization criterion in a greedy algorithm,” in *Proc. Intl. Conf. Knowl. Discovery and Inf. Retrieval*, pp. 208-217, 2011.
[20] M. E. J. Newman, “Finding community structure in networks using the eigenvectors of matrices,” *Phys. Rev. E*, vol. 74, pp. 036104, 2006.
[21] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 888-905, 2000.
[22] B. Karrer and M. E. J. Newman, “Stochastic blockmodels and community structure in networks,” *Phys. Rev. E*, vol. 83, pp. 016107, 2011.
[23] M. E. J. Newman, “Fast algorithm for detecting community structure in networks,” *Phys. Rev. E*, vol. 69, pp. 066133, 2004.
[24] E. L. Martelot and C. Hankin, “Multi-scale community detection using stability as optimization criterion in a greedy algorithm,” in *Proc. Intl. Conf. Knowl. Discovery and Inf. Retrieval*, pp. 208-217, 2011.
[25] M. E. J. Newman, “Finding community structure in networks using the eigenvectors of matrices,” *Phys. Rev. E*, vol. 74, pp. 036104, 2006.
