Fast Community Detection in Complex Networks with a *K*-Depths Classifier

Yahui Tian and Yulia R. Gel

Abstract We introduce a notion of data depth for recovery of community structures in large complex networks. We propose a new data-driven algorithm, *K*-depths, for community detection using the L_1-depth in an unsupervised setting. We evaluate finite sample properties of the *K*-depths method using synthetic networks and illustrate its performance for tracking communities in online social media platform Flickr. The new method significantly outperforms the classical *K*-means and yields comparable results to the regularized *K*-means. Being robust to low-degree vertices, the new *K*-depths method is computationally efficient, requiring up to 400 times less CPU time than the currently adopted regularization procedures based on optimizing the Davis–Kahan bound.

1 Introduction

The explosive growth of online social networking and recent advances on modeling of massive and complex data has led to a skyrocketing interest in analysis of graph-structured data and, particularly, in discovering network communities. Indeed, many real-world networks—from brain connectivity to ecosystems to gang formation and money laundering—exhibit a phenomena where certain features tend to cluster into local cohesive groups. Community detection has been extensively studied in statistics, computer science, social sciences and domain knowledge disciplines and nowadays still remains one of the most hottest research areas in network analysis (for overview of algorithms, see, e.g., [5, 11, 13, 21, 22, 37, 42, 46, 50, 68], and the references therein).

The current paper is motivated by three overarching questions. First, there exists no unique and agreed upon definition of *network community*, typically a community is thought of a cohesive set of vertices that have stronger or better internal connections within the set than with external vertices [37, 44]. Second, community discovery is further aggravated in a presence of (usually multiple) outliers, and until recently the two tightly woven problems of outlier detection

Y. Tian • Y.R. Gel (✉)
University of Texas at Dallas, 800 W Campbell Rd, Richardson TX 75080, USA
e-mail: yxt120830@utdallas.edu; ygl@utdallas.edu

S.E. Ahmed (ed.), *Big and Complex Data Analysis*, Contributions to Statistics,
DOI 10.1007/978-3-319-41573-4_8

and network clustering have been studied as independent problems [5, 15, 48]. Third, vertices with a low degree (or the so-called parasitic outliers of the spectrum [35]) tend to produce multiple zero eigenvalues of the graph Laplacian, which results in a higher variability of spectral clustering and thus a reduced finite sample performance in community detection. Fourth, most of the currently available methods for community discovery within a spectral clustering framework are based on the Euclidean distance as a measure of "cohesion" or "closeness" among vertices, and thus do not explicably account for the underlying probabilistic geometry of the graph.

We propose to address the above challenges by introducing a concept of *data depth* into the network community detection that allows to integrate ideas on cohesion, centrality, outliers, and community discovery under a one systematized "roof." Data depth is a nonparametric and inherently geometric tool to analyze, classify, and visualize multivariate data without making prior assumptions about underlying probability distributions. A new impetus has been recently given to data depths due to their broad utility in high dimensional and functional data analysis (for overview, see, e.g., [9, 26, 27, 34, 40, 47, 58, 73], and the references therein.). Given a notion of data depth, we can measure the "depth" (or "outlyingness") of a given object or a set of objects with respect to an observed data cloud. A higher value of a data depth implies a deeper location or higher centrality in the data cloud. By plotting such a natural center-outward ordering of depth values that serves as a topological map of the data, the presence of clusters, outliers, and anomalies can be evaluated simultaneously in a quick and visual manner. A notion of data depth is novel to network studies. The only relevant paper on the topic is due to [14] who consider a random sample of graphs following the same probability model on the space of all graphs of a given size. This probabilistic framework, however, is not applicable to analysis of most real-world graph-structured data where the available data consists only of a *single network*. In this paper we primarily focus on utility of L_1-depth as the main tool for unsupervised community detection in a spectral setting. Although there exist numerous other depth alternatives, our choice of a depth function is motivated by simplicity and tractability of L_1-depth and the fact that it can be computed using a fast and monotonically converging algorithm [29, 30, 65]. This makes L_1-depth particularly attractive for community discovery in large complex networks.

The paper is organized as follows. Section 2 provides background on graphs, spectral clustering, and K-means algorithm. We introduce the new K-depths method based on the L_1-depth and discuss its properties in Sect. 3. simulation studies are presented in Sect. 4. Section 5 illustrates application of the K-depths method to tracking communities in online social media platform Flickr.

2 Preliminaries and Background

Graph Notations Consider an undirected and loopless graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with a vertex set $\mathcal{V}$ of cardinality n and an edge set $\mathcal{E}$. We assume that $\mathcal{G}$ consists of K non-overlapping communities and K is given. Let A be an $n \times n$-symmetric empirical adjacency matrix, i.e.,

$$A = \begin{cases} 1, & \text{if } (i,j) \in \mathcal{E} \\ 0, & \text{otherwise.} \end{cases}$$

The population counterpart of A is denoted by P. Let D be a diagonal matrix of degrees, i.e., $D_{ii} = \sum_{j=1}^{n} A_{ij}$. Then, the graph Laplacian is defined as

$$L = D^{-1/2} A D^{-1/2}. \tag{1}$$

Spectral Clustering For smaller networks, communities can be identified via optimizing various goodness of partition measures, for instance, Ratio Cut [19], Normalized Cut [60], and Modularity [45], which involve a search for optimal split over all possible partitions of vertices. However, such discrete optimization problems are typically NP-hard and thus are not feasible for larger networks. The computational challenges can be circumvented using spectral clustering (SC) that yields a continuous approximation to discrete optimization [67]. Hence, SC is now one of the most widely popular procedures for tracking communities in large complex networks [66].

The key idea of SC is to embed a graph $\mathcal{G}$ into a collection of multivariate sample points. Given K communities, we identify orthogonal eigenvectors $\mathbf{v}_{\cdot j}, j = 1, \ldots, K$ of the Laplacian L (or adjacency matrix A) that correspond to the largest K eigenvalues, and construct the $n \times K$-matrix $V = [\mathbf{v}_{\cdot 1}, \ldots, \mathbf{v}_{\cdot K}]$. Each row of V, $\mathbf{v}_i \equiv \mathbf{v}_{i\cdot}$, provides a representation in $\mathbb{R}^K$ of a vertex in $\mathcal{V}$. Given this embedding, we can now employ any appropriate classifier to cluster this multivariate data set into K communities, and the most conventional choice is a method of K-means [1, 31, 36, 51].

Given a set of data points $x_i \in \mathbb{R}^d$, for $i = 1, \ldots, n$, the method of K-means [41] aims to group observations into K sets $\mathbf{C} = \{C_1, \ldots, C_K\}$ in such a way that the within-cluster sum of squares is minimized, that is, we minimize

$$\underset{\mathbf{C}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{x \in C_k} ||x - \mu_k||^2, \tag{2}$$

where μ_k is the mean of points in C_k, $||x - \mu_k||^2$ is the squared Euclidean distance between x and k-th group mean μ_k. The optimization (2) is highly computationally intensive. As an alternative, we can employ the Lloyd's algorithm (also known as Voronoi iteration or relaxation) for (2) that is based on iterative refinement and that

allows to quickly identify an optimum (see the outline of the K-means 1). In this paper, the initial centers are chosen randomly from the data set.

Algorithm 1: The K-means algorithm

Input : a set of data points $\mathbf{X} = \{x_1, \ldots, x_n\}$, an initial set of K means $m_1, \ldots, m_K$.
Output: a partition of $\mathbf{X}$.

1 do

2 • Assign points to its nearest cluster in terms of squared Euclidean distance, for $k = 1, \ldots, K$:

$$C_k = \{x_i : ||x_i - m_k||^2 \leq ||x_i - m_j||^2, \forall j, 1 \leq j \leq K\}$$

• Update cluster centers as the mean of points in new clusters, for $k = 1, \ldots, K$:

$$m_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i.$$

3 until *the assignment no longer change*;

Regularization Low-degree vertices tend to produce multiple zero eigenvalues of a Laplacian L, which in turns increases clustering variability and adversely impacts a performance of the K-means algorithm. The problem is closely connected to the concentration of L, that is, the study on how close a sample Laplacian L to its expected value. Sparser networks tend to produce more low-degree vertices and do not concentrate. The idea of regularization in this context is to somehow diminish the impact of such vertices with low degrees, by viewing them as outliers and shrinking them toward the center of spectrum. As a result, regularization leads to a higher concentration. There are a number of regularization procedures ranging from brute-force trimming of outliers to sophisticated methods that are closely connected to regularization of covariance matrices (for more discussion and the most recent literature review see [35]). One of the most popular approaches, by analogy with a ridge regularization of covariance matrices, is to select some positive parameter τ and add τ/n to all entries of the adjacency matrix A [1], that is

$$A_\tau = A + \tau J,$$

where $J = 1/n\mathbf{1}$, $\mathbf{1}$ is $n \times n$-matrix with all elements 1. The resulting regularized Laplacian then takes a form

$$L_\tau = D_\tau^{-1/2} A_\tau D_\tau^{-1/2},$$

where $D_{ii,\tau} = \sum_{j=1}^{n} A_{ij} + \tau$. The optimal regularizer τ can then be selected by minimizing the Davis–Kahan bound, i.e., the bound on the distance between the

sample and population Laplacians (for study on properties of regularized spectral clustering see, [31, 35], and the references therein). However, selecting an optimal regularizer τ is highly computationally expensive. In addition, the impact of small and weak communities on performance of regularized spectral clustering is not clear.

In this light, an interesting question arises on whether we can develop an alternative data-driven and computationally inexpensive method for taming "outliers" with low degrees and bypass the optimization stage of the Davis–Kahan bound? It seems natural to unitize here a statistical methodology that has been developed with a particular focus on analysis of outliers, that is, a notion of data depth.

3 Community Detection Using L_1 Data Depth

In this section, we propose a new unsupervised K-depths algorithm for network community detection based on iterative refinement with L_1 depth.

The L_1 Data Depth In this paper we consider an L_1-data depth of Vardi and Zhang [65]. Consider N distinct observations $x_1, \ldots, x_N$ in $\mathbb{R}^p$ which we need to partition into K clusters, and let $I(k)$ be a set of labels for observations in the k-th cluster. Let each observation x_i be associated with a scalar η_i, $i = 1, \ldots, N$, where η_i are viewed as weights or as "multiplicities" of x_i, and $\eta_i = 1$ if the data set has no ties. The multivariate L_1-median of a k-th cluster, $y_0(k)$, is then defined as

$$y_0(k) = \operatorname{argmin} C(y|k), \tag{3}$$

where $C(y|k)$ is the weighted sum of distances between y and points x_i in the k-th cluster

$$C(y|k) = \sum_{i \in I(k)} \eta_i ||x_i - y|| \qquad \forall k. \tag{4}$$

Here $||u - v||$, $u, v \in \mathbb{R}^p$, is the Euclidean distance in $\mathbb{R}^p$. If $x_1, \ldots, x_N$ are not multicollinear (which is the case of the considered spectral clustering framework), $C(y)$ is positive and strictly convex in $\mathbb{R}^p$. If the set $x_1, \ldots, x_N$ has ties, "multiplicities" η_i can be chosen in such a way that it preserves convexity of $C(y)$ (see [65], for further discussion).

The L_1 depth was proposed by Vardi and Zhang [65], based on the notion of a multivariate L_1-median (3), and the idea has been further extended to clustering and classification in multivariate and functional settings by López-Pintado and Jörnsten [39], Jörnsten [29]. Given a cluster assignment, the L_1 depth of point $x, x \in \mathbb{R}^K$ with respect to a k-th cluster is defined as

$$LD(x|k) = 1 - \max[0, ||\bar{e}(x|k)|| - f(x|k)]. \tag{5}$$

Here $f(x|k) = \eta(x)/\sum_{i\in I(k)}\eta_i$ with $\eta(x) = \sum_{i=1}^{N}\eta_i I(x = x_i)$ and $\bar{e}(x|k)$ is the average of the unit vectors from a point x to all observations in the k-th cluster and is defined as

$$\bar{e}(x|k) = \sum_{i\in I(k), x_i\neq x} \eta_i e_i(x)/\sum_{j\in I(k)}\eta_j,$$

where $e_i(x) = (x_i - x)/||x_i - x||$.

The idea of $1 - LD(x|k)$ is to quantify a minimal additional weight required to assign x so that x becomes the multivariate L_1-median of the k-th cluster $x \cup \{x_i, i \in I(k)\}$ [65]. Hence, L_1 depths as a robust representation of a topological structure of each cluster. Since L_1 is non-zero outside the convex hull of the data cloud, it is a feasible depth choice for comparing multiple clusters [29].

The *K*-Depths Method It is well known that K-means clustering algorithm is non-robust to outliers [16, 59, 69]. This partially is due to the fact that the K-means algorithm is based on a squared Euclidean norm as the measure of "distance" and only captures the information between a pair of points, i.e., a *candidate* center and another point (see Fig. 1a). Also to identify a cluster, the K-means algorithm uses a presumptive cluster center defined by a cluster mean, which makes it sensitive to anomalies and outliers. Although we update centers and clusters until the assignments no longer change, there is no guarantee that the global optimum for (2) can be found [49, 54].

Our idea is motivated by the two overarching questions. Is there an alternative "cohesion" measure to a squared Euclidean norm? Does such a measure allow to

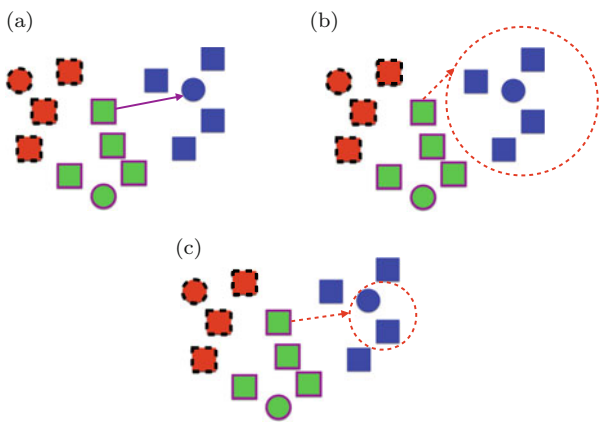

Fig. 1 Comparing K-means and K-depths algorithms. *Circles* denote cluster centers. Each cluster is identified by *colors* and *border* around points. (**a**) K-means. (**b**) K-depths. (**c**) Generalized K-depths

achieve a higher accuracy and stability by taking advantage of more information between clusters and points?

Indeed, such a "cohesion" measure exists, and it can be based on a data depth notion. As discussed earlier, a depth function evaluates how "deep" (or "central") a point is with respect to a group of data (i.e., a cluster). Hence, depth functions allow for more informative and robust "cohesion" (or "distance") measures than a squared Euclidean norm (Fig. 1b).

Our proposed approach is then to use a data depth (particularly, the L_1 depth) to find "nearest" clusters as a part of iterative refinement, and we call the new method "K-depths" clustering algorithm. That is, following the spectral clustering setting, we embed a graph into a collection of multivariate sample points. Then, given K communities, we identify orthogonal eigenvectors of the Laplacian L that correspond to the K largest eigenvalues of L, and construct an $n \times K$-matrix V that is formed by eigenvectors of L. We view each row of V as a representation of a network vertex in $\mathbb{R}^K$, and thus, we get n sample points in K-dimensional space. Clustering of these multivariate points using the K-depths yields a partition of networks into K communities. (The K-depths method is outlined in Algorithm 2. Note that we still use a squared Euclidean norm to initialize the K-depths iterative refinement.) Note that instead of Laplacian spectral embedding, we can also consider adjacency spectral embedding (see [36] and references therein).

Algorithm 2: Spectral clustering K-depths algorithm

Input : network $\mathcal{G}$; number of communities K, depth function LD.
Output: a partition of $\mathcal{G}$.

compute L using (1) ; `// Spectral Clustering`
construct V by combining the leading K eigenvectors of L;
view each row of V as a multivariate representation of each vertex in $\mathcal{V}$;
randomly select K points as initial centers $m_1^0, m_2^0, \ldots, m_K^0$; `// K-depths`
define initial clusters: $C_k^0 = \{x_i : ||x_i - m_k||^2 \leq ||x_i - m_j||^2, \forall j, 1 \leq j \leq K\}$;
do
  extract inner p percent vertices: $I_k = \{x_i : LD(i|k) \geq LD(.|k)_{(p*n_k)}\}$ for $k = 1, \ldots, K$ update clusters: $C_k = \{x_i : LD(x_i|k) \geq LD(x_i|j), \forall j, 1 \leq j \leq K\}$;
until *the assignment no longer change*;

The K-depths algorithm presented above is closely related to the modified Weiszfeld algorithm of [65]. The idea of the K-depths is to evaluate "centrality" of any given point in respect to all points within a cluster (see Fig.1b), that is, in respect to points located inside the cluster and points located close to a cluster borderline. However, points that fall in-between clusters may provide redundant or noisy information, which leads a higher variability of the clustering algorithm. We therefore introduce the generalized K-depths measure which only accounts for the inner parts of a cluster to calculate depth values (Fig. 1c). Given a contour plot of a cluster, we compute L_1 depth values using points which are within an arbitrary percentage contour $p \in [0, 1]$. For instance, Fig. 1b is a special case of Fig. 1c where

the *locality* parameter p is 1, i.e., all 100 % of available data are used in the K-depths algorithm. We can also view the locality parameter p as a trade-off of bias (i.e., detection accuracy) and variance (i.e., detection variability).

Remark The optimal choice of p, similarly to selection of optimal trimming, largely depends on a definition of outlier, types of anomalous behavior, proportion of contamination, and structure of the data. Conventionally, trimming and other robustifying parameters are chosen using various types of resampling, including V-fold crossvalidation, jackknife, and bootstrap (see [2, 24, 25]). Under the network setting, the problem is further aggravated by the lack of an agreed-upon definition of outliers and network anomalies and their dependence on the underlying network model structure (for overviews, see [3–5, 17, 18]). For instance, [5] discuss at least four kinds of outliers: mixed membership, hubs, small clusters, independent neutral nodes. Although selecting p using crossvalidation is likely to be affected by the presence of outliers in an observed network, we believe that one of the resampling ideas such as crossvalidation or bootstrap [12, 63] is still arguably the most feasible approach that allows to minimize parametric assumptions about the network model.

3.1 Properties of Spectral Clustering K-Depths Algorithm

Asymptotic properties of spectral clustering and, particularly, the K-means/ medians algorithms have been widely studied both in probability and statistics (for the most recent overviews, see, e.g., [28, 36, 52, 53], and the references therein). While most of the results focus on denser networks, most recently [36] derive an upper error bound for spectral clustering under moderately sparse stochastic block model with a maximum expected degree of order $\log n$ or higher.

The key result behind deriving all asymptotic properties of the K-means/ medians algorithms is to show that there exists a sequence $\epsilon_n, \epsilon_n \geq 0$ such that $\lim_{n\to\infty} \epsilon_n = 0$ and

$$z^{A_k} \leq (1 + \epsilon_n) z^*(\mathcal{G}), \ n \in Z^+ \tag{6}$$

where $z^{A_k}(\mathcal{G})$ is the approximate polynomial time solution from the K-means/medians algorithms and $z^*(\mathcal{G})$ is the optimal solution. If such a sequence ϵ_n exists, then [10] define asymptotic optimality of the K-medians algorithm.

Defining $z(\mathcal{G})$ in (6) in terms of a Frobenius norm of a distance between the K largest eigenvectors $U_1, \ldots, U_K$ of a population adjacency matrix P and their respective counterparts $\hat{U}_1, \ldots, \hat{U}_K$ from an empirical adjacency matrix A, [33] show that there exists an approximate polynomial time solution to the K-means

algorithm with an error bound

$$||\hat{\Theta}\hat{X} - \hat{U}||_F^2 \leq (1+\epsilon) \min_{\substack{\Theta \in \mathbb{M}_{n,K} \\ X \in \mathbb{R}_{K\times K}}} ||U - \hat{U}||_F^2, \quad \hat{U},\ U \in \mathbb{R}^{n\times K},$$

where $U = [U_1, \ldots, U_K]$ and $\hat{U} = [\hat{U}_1, \ldots, \hat{U}_K]$. Here Θ is a true membership matrix such that Θ_{ig_i} is 1 where $g_i \in \{1, \ldots, K\}$ is the community membership of vertex i, and $\mathbb{M}_{n,K}$ is a collection of all $n \times K$-matrices where each row has exactly one 1 and the remaining $K-1$ entries are 0. For discussion on analogous results on existence of $(1+\epsilon)$-approximate solution for a k-medians algorithm in network applications see, for instance, [36].

Since statistical properties of median and L_1-depth are closed related (see [65]), we state the following conjecture about the error bound of the K-depths algorithm under the L_1 depth and adjacency spectral embedding.

Conjecture 1 There exists an Ω-approximate polynomial time solution to the K-depths method under adjacency spectral embedding which attains

$$||\hat{\Theta}\hat{X} - \hat{U}||_F^2 \leq \Omega \min_{\substack{\Theta \in \mathbb{M}_{n,K} \\ X \in \mathbb{R}_{K\times K}}} ||U - \hat{U}||_F^2, \tag{7}$$

where Ω is a positive constant and $(\hat{\Theta}, \hat{X}) \in \mathbb{M}_{n,K} \times \mathbb{R}_{K\times K}$ is the output of Ω-approximate K-depths algorithm.

Armed with (7), an upper bound on network community detection error of the K-depths algorithm 2 under adjacency spectral embedding can be derived for a stochastic block model (SBM), following derivations of [36, 52, 65]. This error bound for the K-depths increases with an increasing network sparsity and with the growing number of communities. In addition, assuming existence of a Σ-approximate solution to the K-depths algorithm, analogous error bounds can be derived under Laplacian spectral embedding [52, 53].

4 Simulations

In this section we evaluate a finite sample performance of the unsupervised K-depths classifier for detecting network communities and primarily focus on a case of two communities. To measure a goodness of clustering, we employ such standard criteria as misclassification rate and normalized mutual information (NMI). We define misclassification rate as the total percentage of mislabeled vertices, i.e.,

$$\gamma = \frac{1}{n} \sum_{i=1}^{K} |S_i|,$$

where $|S_i|$ is the number of misclassified vertices in the i-th community.

Given the two sets of clusters with a total of n vertices: $\mathbb{R} = \{r_1, \ldots, r_K\}$ and $\mathbb{C} = \{c_1, \ldots, c_J\}$, the NMI is given by Manning et al. [43]:

$$\text{NMI}(\mathbb{R}, \mathbb{C}) = \frac{I(\mathbb{R}; \mathbb{C})}{[H(\mathbb{R}) + H(\mathbb{C})]/2}.$$

Here I is mutual information

$$\begin{aligned} I(\mathbb{R}; \mathbb{C}) &= \sum_k \sum_j P(r_k \bigcap c_j) \log \frac{P(r_k \bigcap c_j)}{P(r_k)P(c_j))} \\ &= \sum_k \sum_j \frac{|r_k \bigcap c_j|}{n} \log \frac{n|r_k \bigcap c_j|}{|r_k||c_j|} \end{aligned}$$

where $P(r_k)$, $P(c_j)$, and $P(r_k \bigcap c_j)$ are the probabilities of a vertex being in cluster r_k, c_j and in the intersection of r_k and c_j, respectively, and H is entropy defined by

$$H(\mathbb{R}) = -\sum_k P(r_k) \log P(r_k) = -\sum \frac{|r_k|}{n} \log \frac{|r_k|}{n}.$$

NMI takes values between 0 and 1, and we prefer a clustering partition with a higher NMI.

4.1 Network Clustering with Two Groups

Here we use a benchmark simulation framework based on a 2-block stochastic block model (SBM)[61, 71]. SBM is a particular case of an inhomogeneous Erdös–Renyi model in which edges are formed independently and probability of an edge between two vertices is determined by group membership of vertices [23].

Following a simulation setting of Joseph and Yu [31], we generate 100 networks of order 3000 from an SBM with a block probability matrix

$$B = \begin{bmatrix} 0.01 & 0.0025 \\ 0.0025 & 0.003 \end{bmatrix}, \tag{8}$$

and assume that the connections within the k-th community follow an independent Bernoulli distribution with probability B_{kk}, $k = 1, 2$.

Table 1(a) summarizes clustering performance of the K-means and K-depths algorithms in terms of misclassification rate and NMI. We find that the K-depths method noticeably outperforms the K-means algorithm, delivering 36 % lower misclassification rate and more than four times higher NMI, although with a

Table 1 Performance of the K-means and K-depths algorithms in respect to misclassification rate γ and NMI, with standard deviation in (), under (a) SBM (8) and (b) Generalized SBM (GSBM). The locality parameter p for the K-depths algorithm is 0.1

Method	γ	NMI
(a)		
K-means	0.44	0.05
	(0.08)	(0.09)
K-depths	0.28	0.23
	(0.13)	(0.19)
(b)		
K-means	0.62	0.24
	(0.21)	(0.08)
K-depths	0.55	0.43
	(0.25)	(0.07)

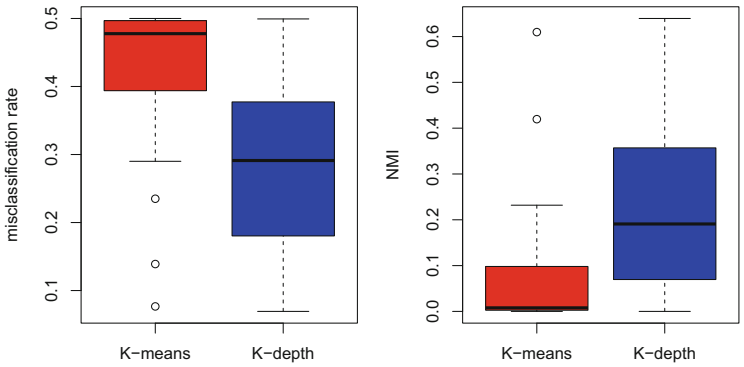

Fig. 2 Boxplots of clustering performance of the K-means and K-depths in terms of misclassification rate and NMI for the SBM (8)

somewhat higher variability. Remarkably, the boxplot for misclassification rate and NMI (see the left panel of Fig. 2) indicates that despite a higher variability, the lower quartile of the misclassification rates delivered by the K-depths algorithm is smaller than the upper quartile of the misclassification rates yielded by the K-means algorithm. A similar dynamics is also observed for NMI (see the right panel of Fig. 2).

We find that regularization of both K-means and K-depths where an optimal regularizer τ is selected using optimizing the Davis–Kahan bound as per [31] improves community discovery. That is, the regularized K-means outperforms the regularized K-depths in terms of misclassification rates, i.e., 0.16 vs. 0.22; and the regularized K-depths outperforms the regularized K-means in terms of NMI, i.e., 0.44 vs. 0.40. However, regularization turns out to be highly computationally

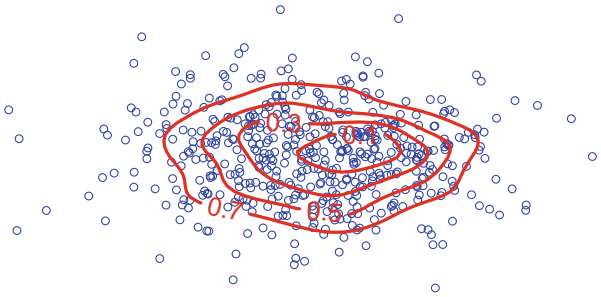

Fig. 3 Contour plots based on the L_1-data depth and varying data proportions p, i.e., p is 0.1, 0.3, 0.5, and 0.7

expensive, that is, finding an optimal regularization for a single network of 3000 vertices under SBM (8) requires 1800 s (with 1 additional sec for the K-means algorithm itself). In contrast, the unregularized K-depths algorithm takes only 4 s. (The elapsed time is assessed in R on an OS X 64 bit laptop with 1.4 GHz Intel Core i5 processor and 4 GB 1600 MHz DDR3 memory.)

Thus, being intrinsically robust to low-degree vertices, the new K-depths method provides a simple and computationally efficient alternative to the currently adopted regularization procedures based on optimizing the Davis–Kahan bound.

Choice of a Locality Parameter Let us explore the impact of a locality parameter p, $p \in [0, 1]$, on a clustering performance of the K-depths algorithm. Note that p controls how many points are selected to form the "deepest" sub-clusters which other points are compared with. Figure 3 visualizes sub-clusters and the respective contour plots based on the L_1-depth, corresponding to $p = (0.1,\ 0.3,\ 0.5,\ 0.7)$. If p is 1, the whole data cloud is used, while lower values of p lead to a higher concentration of points around the cluster center and aim to minimize the impact of outlying points or noise. Hence, a locality parameter p can be viewed as a trade-off between bias and variance. Figure 4 shows the performance of the K-depths algorithm in respect to varying p and the SBM (8). We find that in general both mean and variance of misclassification rates and NMI are stable and comparable for p of less than 0.5. As expected, higher values of p lead to a better performance in terms of average misclassification rates and NMI but also result in a substantially higher variability. In general, an optimal p can be selected via crossvalidation, and choice of p is likely to be linked with a sparsity of an observed network. However, given the stability of the K-depths performance, as a rule of thumb we suggest to use a p of 0.5 or less.

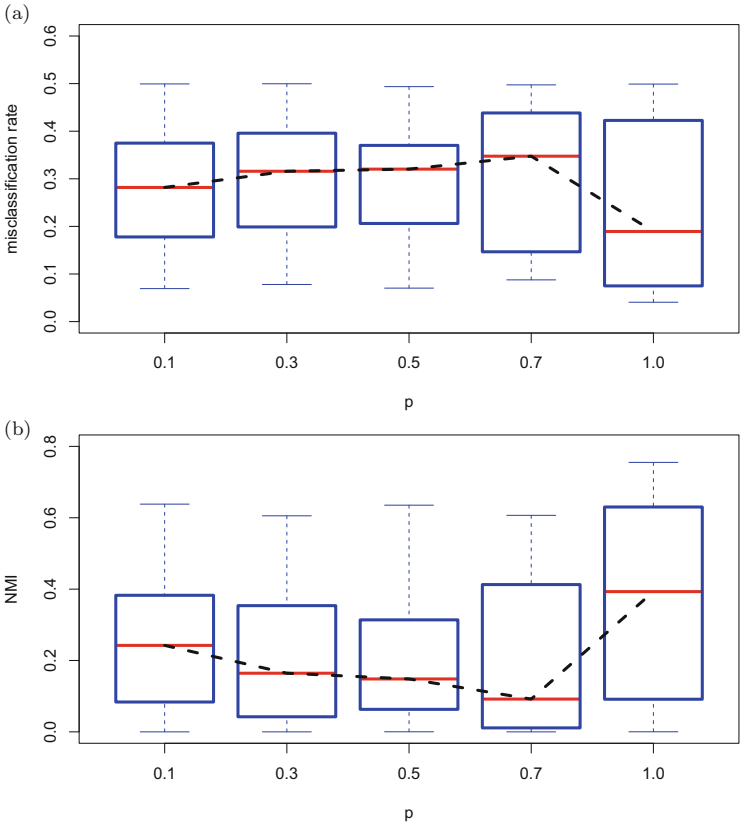

Fig. 4 Boxplots of misclassification rates (**a**) NMI (**b**) with various choices of locality parameter p. The dashed line connects medians for resulting misclassification rates and NMI for various locality parameters p, in plots (**a**) and (**b**) respectively

4.2 Network Clustering with Outliers

Now we evaluate the performance of the K-depths algorithm in respect to a network with outliers. In particular, we consider the so-called *Generalized Stochastic Block Model (GSBM)* of Cai and Li [5] which is based on incorporating small and weak communities (outliers) into a conventional SBM structure. More specifically, consider an undirected and loopless graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $N = n + m$ vertices, where n is the number of "inliers" which follow the standard SBM framework and m is the number of "outliers" which connect with other vertices in random. Each inlier vertex is assigned to one of the two communities, while all outliers are placed into the 3rd community. An example of GSBM is shown in Fig. 5, two strong communities are colored by red and green within solid circles, the outliers (one weak and small community) are colored by blue within a dashed circle.

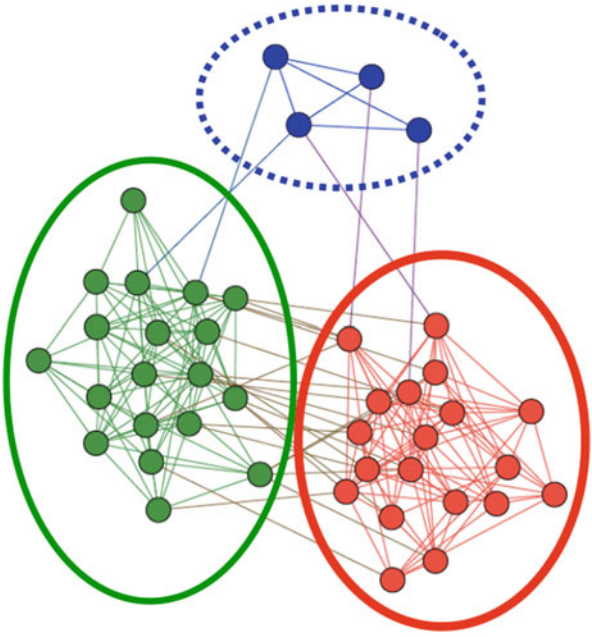

Fig. 5 Network with outliers, or small and weak community, under GSBM

In this section we consider a GSBM of Cai and Li [5] by adding 30 outliers (i.e., one small and weak community) into a standard 2-block SBM (8).

In particular, we set a probability of an edge between outliers to be of 0.01. Connection between inliers and outliers is defined by an arbitrary $(0, 1)$-matrix Z, $Z \in \mathbb{R}^{n\times m}$, such that $\mathbb{E}Z = \boldsymbol{\beta}\mathbf{1}^T = [\boldsymbol{\beta}, \ldots, \boldsymbol{\beta}]$ and the component of $\boldsymbol{\beta}$ are 3000 i.i.d. copies of U^2, where U is a uniform random variable on $[0, 0.0025]$.

Following [5], we define a misclassification rate based only on inliers in the dominant 1st and 2nd communities, i.e.,

$$\gamma = \frac{1}{n}\sum_{k=1}^{2} |S_k|,$$

where $|S_k|$ is a number of misclassified vertices in the k-th community and $k = 1, 2$. Similarly, NMI is defined calculated only on inliers and a number of clusters K are set to 3 for both K-means and K-depths algorithms.

Table 1(b) summarizes the results for misclassification rates and NMI delivered by the K-means and K-depths algorithms. In general, misclassification rates for both methods under the GSBM model are noticeably higher than the analogous rates under a standard SBM. However, the K-depths algorithm still outperforms the K-means method, yielding a 10 % lower misclassification rate. In turn, NMI delivered by the K-depths algorithm is almost twice higher than the corresponding NMI of the K-means method, i.e., 0.43 vs. 0.24, respectively. Remarkably, under

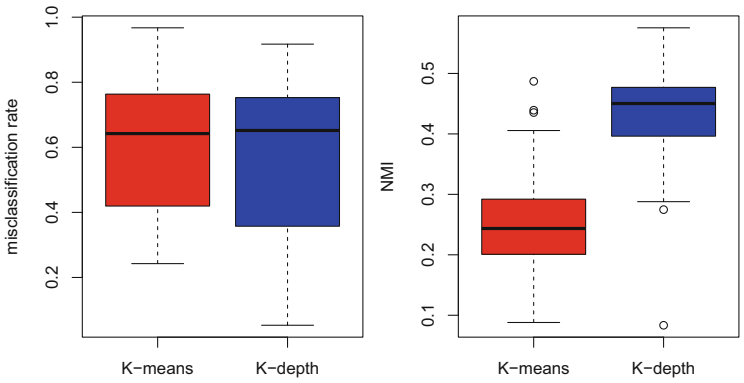

Fig. 6 Boxplots of clustering performance of the *K*-means and *K*-depths in terms of misclassification rate and NMI under the GSBM

the GSBM variability of both methods is very similar, while the upper quartile of NMI for the *K*-means algorithm is lower than almost all values of NMI delivered by the *K*-depths algorithm (see Fig. 6).

5 Application to Flickr Communities

In this section, we illustrate the *K*-depths algorithm to tracking communities in Flickr. Flickr is a popular website for users to share personal photographs and also an online platform. This data set contains the information of 80,513 Flickr bloggers, each blogger is viewed as a vertex, and the friendship between bloggers is represented by undirected edges. The data is available from [70]. Bloggers are divided into 195 groups depending on their interests. As discussed by Tang and Liu [62], the network is very sparse and scale-free (i.e., its degree distribution follows a power law).

In our study, we consider a subnetwork of Flickr by extracting vertices that belong to the second and third communities and edges within and in-between of these communities. Isolated vertices (vertices with no edges) are removed. The resulting data represents an undirected graph with 216 vertices and 996 edges; the second community contains 155 vertices and 753 edges, while the third community contains 61 vertices and 19 edges.

We now apply the *K*-means and *K*-depths algorithms to identify clusters in the Flickr subnetwork (see Table 2). We find that the *K*-depths algorithm delivers a misclassification rate of 0.35, which is more than 26 % lower than the misclassification rate of 0.47 yielded by the *K*-means algorithm. In turn, NMI yielded by the *K*-depths algorithm is comparable with NMI of the *K*-means algorithm.

Table 2 Misclassification rate (γ) and Normalized Mutual Information (NMI) criteria for the K-means and K-depths methods for the Flickr subnetwork

Method	γ	NMI
K-means	0.47	0.07
K-depths	0.35	0.07

The locality parameter p for the K-depths algorithm is 0.5

6 Conclusion and Future Work

In this paper, we introduce a new unsupervised approach to network community detection based on a nonparametric concept of data depth within a spectral clustering framework. In particular, we propose a data-driven K-depths algorithm based on iterative refinement of the L_1 depth. The new method is shown to substantially outperform the classical K-means and to deliver comparable results to the regularized K-means. The K-depths algorithm is simple and computationally efficient, requiring up to 400 times less CPU time than the currently adopted regularization procedures based on optimizing the Davis–Kahan bound. Moreover, the K-depths algorithm is intrinsically robust to low-degree vertices and accounts for the underlying geometrical structure of a graph, thus paving the way for using the L_1 depth and other depth functions as an alternative to computationally expensive selection of optimal regularizers.

In addition to asymptotic analysis of the K-depths clustering, in the future we plan to advance the K-depths approach to other types of depth functions, for example, the classical ones: half-space depth, Mahalanobis depth, random projection depth etc [38, 55–57, 72], and to the most recent such as Monge-Kantorovich depth [6, 20] and to explore utility of the K-depths method as initialization algorithm (for discussion, see [64] and the references therein). Another interesting direction is to investigate the relationship between properties of the K-depths approach and the trimmed K-means algorithms [7, 8, 32], both in networks and general multivariate clustering contexts.

Acknowledgements Authors are grateful to Robert Serfling, Ricardo Fraiman, and Rebecka Jörnsten for motivating discussions at various stages of this paper. Yulia R. Gel is supported in part by the National Science Foundation grant IIS 1633331. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET) of Canada.

References

1. Amini, A.A., Chen, A., Bickel, P.J., Levina, E.: Pseudo-likelihood methods for community detection in large sparse networks. Ann. Stat. **41**, 2097–2122 (2013)
2. Arlot, S., Celisse, A., et al.: A survey of cross-validation procedures for model selection. Stat. Surv. **4**, 40–79 (2010)

3. Baddar, S.A.-H., Merlo, A., Migliardi, M.: Anomaly detection in computer networks: a state-of-the-art review. J. Wirel. Mob. Netw. Ubiquit. Comput. Dependable Appl. **5**(4), 29–64 (2014)
4. Bhuyan, M.H., Bhattacharyya, D.K., Kalita, J.K.: Network anomaly detection: methods, systems and tools. IEEE Commun. Surv. Tutorials **16**(1), 303–336 (2014)
5. Cai, T.T., Li, X.: Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. Ann. Stat. **43**(3), 1027–1059 (2015)
6. Chernozhukov, V., Galichon, A., Hallin, M., Henry, M.: Monge-Kantorovich depth, quantiles, ranks, and signs. arXiv preprint arXiv:1412.8434 (2014)
7. Cuesta-Albertos, J., Gordaliza, A., Matrán, C., et al.: Trimmed *k*-means: An attempt to robustify quantizers. Ann. Stat. **25**(2), 553–576 (1997)
8. Cuesta-Albertos, J.A., Matrán, C., Mayo-Iscar, A.: Trimming and likelihood: robust location and dispersion estimation in the elliptical model. Ann. Stat. **36**(5), 2284–2318 (2008)
9. Cuevas, A., Febrero, M., Fraiman, R.: Robust estimation and classification for functional data via projection-based depth functions. Comput. Stat. **22**, 481–496 (2007)
10. Emelichev, V., Efimchik, N.: Asymptotic approach to the problem of *k*-median of a graph. Cybern. Syst. Anal. **30**(5), 726–732 (1994)
11. Estrada, E., Knight, P.A.: A First Course in Network Theory. Oxford University Press, Oxford (2015)
12. Fallani, F.D.V., Nicosia, V., Latora, V., Chavez, M.: Nonparametric resampling of random walks for spectral network clustering. Phys. Rev. E **89**(1), 012802 (2014)
13. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3), 75–174 (2010)
14. Fraiman, D., Fraiman, F., Fraiman, R.: Statistics of dynamic random networks: a depth function approach. arXiv:1408.3584v3 (2015)
15. Gao, J., Liang, F., Fan, W., Wang, C., Sun, Y., Han, J.: On community outliers and their efficient detection in information networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2010)
16. García-Escudero, L.Á., Gordaliza, A.: Robustness properties of *k* means and trimmed *k* means. J. Am. Stat. Assoc. **94**(447), 956–969 (1999)
17. Gogoi, P., Bhattacharyya, D., Borah, B., Kalita, J.K.: A survey of outlier detection methods in network anomaly identification. Comput. J. **54**(4) (2011)
18. Gupta, M., Gao, J., Han, J.: Community distribution outlier detection in heterogeneous information networks. In: Machine Learning and Knowledge Discovery in Databases, pp. 557–573. Springer, Berlin (2013)
19. Hagen, L., Kahng, A.B.: New spectral methods for ratio cut partitioning and clustering. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **11**(9), 1074–1085 (1992)
20. Hallin, M.: Monge-Kantorovich ranks and signs. GOF DAYS 2015, p. 33 (2015)
21. Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., Padmanabhan, K., Samatova, N.: Community detection in large-scale networks: a survey and empirical evaluation. WIRE Comput. Stat. **6**, 426–439 (2014)
22. Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., Padmanabhan, K., Samatova, N.: Community detection in large-scale networks: a survey and empirical evaluation. Wiley Interdiscip. Rev. Comput. Stat. **6**(6), 426–439 (2014)
23. Holland, P., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: first steps. Soc. Networks **5**(2), 109–137 (1983)
24. Huber, P.J., Ronchetti, E.: Robust Statistics. Wiley, Hoboken vol. 10(1002). doi:9780470434697 (2009)
25. Hubert, M., Rousseeuw, P.J., Van Aelst, S.: High-breakdown robust multivariate methods. Stat. Sci. **23**(1), 92–119 (2008)
26. Hugg, J., Rafalin, E., Seyboth, K., Souvaine, D.: An experimental study of old and new depth measures. In: Proceedings of the Meeting on Algorithm Engineering & Experiments, pp. 51–64. Society for Industrial and Applied Mathematics (2006)
27. Hyndman, R.J., Shang, H.L.: Rainbow plots, bagplots, and boxplots for functional data. J. Comput. Graph. Stat. **19**, 29–45 (2010)
28. Jin, J.: Fast community detection by score. Ann. Stat. **43**(1), 57–89 (2015)

29. Jörnsten, R.: Clustering and classification based on the L 1 data depth. J. Multivar. Anal. **90**(1), 67–89 (2004)
30. Jörnsten, R., Vardi, Y., Zhange, C.-H.: A robust clustering method and visualization tool based on data depth. In: Dodge, Y. (ed.) Statistics in Industry and Technology: Statistical Data Analysis, pp. 353–366. Birkhäuser, Basel (2002)
31. Joseph, A., Yu, B.: Impact of regularization on spectral clustering. Ann. Stat. **44**(4), 1765–1791 (2016)
32. Kondo, Y., Salibian-Barrera, M., Zamar, R.: A robust and sparse k-means clustering algorithm. arXiv preprint arXiv:1201.6082 (2012)
33. Kumar, A., Sabharwal, Y., Sen, S.: A simple linear time $(1 + \epsilon)$-approximation algorithm for k-means clustering in any dimensions. In: Annual Symposium on Foundations of Computer Science, vol. 45, pp. 454–462. IEEE Computer Society Press, New York (2004)
34. Lange, T., Mosler, K.: Fast nonparametric classification based on data depth. Stat. Pap. **55**, 49–69 (2014)
35. Le, C.M., Vershynin, R.: Concentration and regularization of random graphs. arXiv preprint arXiv:1506.00669 (2015)
36. Lei, J., Rinaldo, A.: Consistency of spectral clustering in stochastic block models. Ann. Stat. **43**(1), 215–237 (2015)
37. Leskovec, J., Lang, K.J., Mahoney, M.: Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th International Conference on World Wide Web, pp. 631–640. ACM, New York (2010)
38. Liu, R.Y., Parelius, J.M., Singh, K.: Special invited paper: multivariate analysis by data depth: descriptive statistica, graphics and inference. Ann. Stat. **27**(3), 783–858 (1999)
39. López-Pintado, S., Jörnsten, R.: Functional analysis via extensions of the band depth. In: Complex Datasets and Inverse Problems: Tomography, Networks and Beyond. Lecture Notes-Monograph Series, pp. 103–120. Beachwood, Ohio, USA (2007)
40. López-Pintado, S., Romo, J.: On the concept of depth for functional data. J. Am. Stat. Assoc. **104**, 718–734 (2009)
41. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1(14), pp. 281–297 (1967)
42. Malliaros, F.D., Vazirgiannis, M.: Clustering and community detection in directed networks: a survey. Phys. Rep. **533**, 95–142 (2013)
43. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, vol. 1. Cambridge University Press, Cambridge (2008)
44. Newman, M., Clauset, A.: Structure and inference in annotated networks. arXiv preprint arXiv:1507.04001 (2015)
45. Newman, M.E.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**(23), 8577–8582 (2006)
46. Newman, M.E.J.: Networks: An Introduction. Oxford University Press, Oxford (2010)
47. Nieto-Reyes, A., Battey, H.: A topologically valid definition of depth for functional data. preprint. Stat. Sci. **31**(1), 61–79 (2016)
48. Ott, L., Pang, L., Ramos, F., Chawla, S.: On integrated clustering and outlier detection. In: Proceedings of NIPS (2014)
49. Pena, J.M., Lozano, J.A., Larranaga, P.: An empirical comparison of four initialization methods for the k-means algorithm. Pattern Recogn. Lett. **20**(10), 1027–1040 (1999)
50. Plantiè, M., Crampes, M.: Survey on social community detection. Social Media Retrieval Computer Communications and Networks (2012)
51. Qin, T., Rohe, K.: Regularized spectral clustering under the degree-corrected stochastic blockmodel. In: NIPS, pp. 3120–3128 (2013)
52. Rohe, K., Chatterjee, S., Yu, B.: Spectral clustering and the high-dimensional stochastic blockmodel. Ann. Stat. **39**, 1878–1915 (2011)
53. Sarkar, P., Bickel, P.: Role of normalization in spectral clustering for stochastic blockmodels. Ann. Stat. **43**, 962–990 (2013)

54. Selim, S.Z., Ismail, M.A.: *k*-means-type algorithms: a generalized convergence theorem and characterization of local optimality. IEEE Trans. Pattern Anal. Mach. Intell. **6**(1), 81–87 (1984)
55. Serfling, R.: Generalized quantile processes based on multivariate depth functions, with applications in nonparametric multivariate analysis. J. Multivar. Anal. **83**, 232–247 (2002)
56. Serfling, R.: Quantile functions for multivariate analysis: approaches and applications. Statistica Neerlandica **56**, 214–232 (2002)
57. Serfling, R.: Depth functions in nonparametric multivariate inference. In: Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 72(1). American Mathematical Society, Providence, RI (2006)
58. Serfling, R., Wijesuriya, U.: Nonparametric description of functional data using the spatial depth approach (2015). Accessible at www.utdallas.edu/~serfling
59. Sharma, S., Yadav, R.L.: Comparative study of *k*-means and robust clustering. Int. J. Adv. Comput. Res. **3**(3), 207 (2013)
60. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 888–905 (2000)
61. Sussman, D.L., Tang, M., Fishkind, D.E., Priebe, C.E.: A consistent adjacency spectral embedding for stochastic blockmodel graphs. J. Am. Stat. Assoc. **107**(499), 1119–1128 (2012)
62. Tang, L., Liu, H.: Relational learning via latent social dimensions. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 817–826 (2009)
63. Thompson, M.E., Ramirez Ramirez, L.L., Lyubchich, V., Gel, Y.R.: Using the bootstrap for statistical inference on random graphs. Can. J. Stat. **44**, 3–24 (2016)
64. Torrente, A., Romo, J.: Refining *k*-means by bootstrap and data depth (2013). https://www.researchgate.net/profile/Juan_Romo/publication/242090768_Reflning_k-means_by_Bootstrap_and_Data_Depth/links/02e7e528daa72dc0a1000000.pdf
65. Vardi, Y., Zhang, C.-H.: The multivariate l1-median and associated data depth. Proc. Natl. Acad. Sci. **97**(4), 1423–1426 (2000)
66. von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**(4), 395–416 (2007)
67. White, S., Smyth, P.: A spectral clustering approach to finding communities in graph. In: SDM, vol. 5, pp. 76–84 (2005)
68. Wilson, J.D., Wang, S., Mucha, P.J., Bhamidi, S., Nobel, A.B.: A testing based extraction algorithm for identifying significant communities in networks. Ann. Appl. Stat. **8**(3), 1853–1891 (2014)
69. Witten, D.M., Tibshirani, R.: A framework for feature selection in clustering. J. Am. Stat. Assoc. **105**(490), 713–726 (2012)
70. Zafarani, R., Liu, H.: Social computing data repository at ASU (2009)
71. Zhang, Y., Levina, E., Zhu, J.: Community detection in networks with node features. arXiv preprint arXiv:1509.01173 (2015)
72. Zhou, W., Serfling, R.: General notions of statistical depth function. Ann. Stat. **28**, 461–482 (2000)
73. Zuo, Y., Serfling, R.: General notions of statistical depth function. Ann. Stat. **28**, 461–482 (2000)