Int. J. Appl. Math. Comput. Sci., 2024, Vol. 34, No. 1, 135–147
DOI: 10.61822/amcs-2024-0010

WIDE GAPS AND KLEINBERG'S CLUSTERING AXIOMS FOR *k*–MEANS

MIECZYSŁAW A. KŁOPOTEK [a]

[a] Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland
e-mail: klopotek@ipipan.waw.pl

The widely applied k-means algorithm produces clusterings that violate our expectations with respect to high/low similarity/density within/between clusters and is in conflict with Kleinberg's axiomatic system for distance based clustering algorithms that formalizes those expectations. In particular, k-means violates the consistency axiom. We hypothesise that this clash is due to the unexplained expectation that the data themselves should have the property of being clusterable in order to expect the algorithm clustering them to fit a clustering axiomatic system. To demonstrate this, we introduce two new clusterability properties, i.e., variational k-separability and residual k-separability, and show that then Kleinberg's consistency axiom holds for k-means operating in the Euclidean or non-Euclidean space. Furthermore, we propose extensions of the k-means algorithm that fit approximately Kleinberg's richness axiom, which does not hold for k-means. In this way, we reconcile k-means with Kleinberg's axiomatic framework in Euclidean and non-Euclidean settings. Besides contribution to the theory of axiomatic frameworks of clustering and to clusterability theory, the practical benefit is the possibility to construct datasets for testing purposes of algorithms optimizing the k-means cost function. This includes a method of construction of clusterable data with a global optimum known in advance.

Keywords: clustering theory, clustering axioms, clusterability.

1. Introduction

Clustering is a domain of machine learning with quite vague foundations. The concept of a cluster or clustering is poorly defined. It is associated with high within-cluster similarity and low between-cluster similarity, with high density areas separated with low density areas, with optimizing some cost function, with matching manually assigned labels, with various internal and external clustering scores, etc. (see, e.g., Madhulatha, 2012; Suchy and Siminski, 2023). Also various axiomatic systems have been designed defining clustering related properties, like that of Kleinberg (2002), Ben-David and Ackerman (2009), van Laarhoven and Marchiori (2014), Strazzeri and Sánchez-García (2021), Hopcroft and Kannan (2012), and others. In particular, Kleinberg (2002) proposed three axioms: richness, consistency and scaling invariance (see Axioms 3, 1, 2 in Section 136). Intuitively, scaling invariance states that the results of clustering should be the same independently of the scale of measurement of distances: the same result for kilometers and for miles. Consistency says that, if an algorithm produced a given clustering, then increasing similarities within the same cluster and decreasing similarities between clusters should lead to the same clustering result. Richness means, among others, that the clustering algorithm should itself determine the number of clusters. The conceptual problem with all these notions is that the widely applied k-means algorithm in its base form and derivatives does not care about high/low similarity/density, etc., does not usually determine the number of clusters and thus violates two out of three Kleinberg axioms for distance based clustering algorithms, while checking if the optimum of its cost function is reached would require enumeration of all possible clusterings, and hence is prohibitive in practice.

The special attention that we pay here to Kleinberg's axiomatic system is due to the fact that two of his axioms induce a method for generating new test datasets from existing ones without the need for manual labelling of the new sets. This is important because the development and implementation of new algorithms in the area of machine learning, especially clustering, comparative

studies of such algorithms as well as testing according to software engineering principles require availability of labeled data sets. While standard benchmarks are made available, a broader range of such data sets is necessary in order to avoid the problem of overfitting. In this context, theoretical works on axiomatization of clustering algorithms, especially axioms on clustering preserving transformations like that of Kleinberg (2002), are quite a cheap way to produce labeled data sets from existing ones, given that the respective algorithm to be tested fits the axiomatic framework.

However, the k-means algorithmic family[1] does not fit Kleinberg's axiomatic framework (the richness and consistency axioms are violated). So, what is wrong with this framework? It may be hypothesised that data that have really a clustering structure (are clusterable) will behave according to Kleinberg's intuition, while at the same time we cannot expect such a behavior when the data do not have the clusterability property. In this paper we demonstrate that this hypothesis is accurate with respect to k-means.

We recall some earlier work in Section 2. Then in Section 3 we demonstrate that, if the data have the clusterability property that we call variational k-separability, then Kleinberg's consistency axiom holds for k-means. Furthermore, it is possible to construct a k-range-means algorithm that generalizes k-means by automatic selection of k, for which all three Kleinberg axioms hold for data with variational k-range-separability, when adding the restriction to the consistency axiom that data concentrations within a cluster are not created.

The deficiency of the proposal in Section 3 is that it is applicable to the Euclidean space only, while the so-called kernel-k-means operates de facto in the non-Euclidean space (see, e.g., Girolami, 2002). To overcome this restriction, we propose in Section 4 the clusterability concepts of residual k-separability and residual k-range-separability which imply the restriction of consistency to the realistic case of finite measurement resolution. Section 5 explains how these concepts apply to non-Euclidean spaces.

Section 6 summarizes the results and outlines further research directions.

The main contributions of this paper are proposals of clusterability criteria that reconcile k-means with Kleinberg's axiomatic framework in Euclidean and non-Euclidean settings, and also a method of construction of clusterable data with a global optimum known in advance. Besides the contribution to the theory of axiomatic frameworks of clustering and for clusterability theory, the practical contribution is the possibility to construct datasets for testing purposes of algorithms optimizing the k-means cost function. We propose also a generalization of the k-means algorithm that can self-adjust k when the data are clusterable in the above-mentioned way.

[1] Note that k-means algorithms are used quite commonly on their own and as subroutines of other algorithms, e.g., in the domain of spectral clustering (Lucinska and Wierzchon, 2018).

2. Previous work

We will refer in this paper to the widely used k-means algorithm (k-means++ version, by Arthur and Vassilvitskii (2007)), which belongs to the so-called *k-clustering algorithms*, that is, for a dataset S they return a partition Γ of S into k non-empty groups ($|\Gamma| = k$), where k is a user-defined parameter. Recall that the k-means algorithm was designed to operate primarily in the Euclidean space, that is, we assume an embedding $\mathcal{E} : S \to \mathbb{R}^d$ into a d-dimensional Euclidean space. Then k-means seeks to find a partition Γ of S and positions of cluster centers that minimize the cost (or quality) function

$$Q(\Gamma) = \sum_{C\in\Gamma} \sum_{e\in C} ||\mathcal{E}(e) - \boldsymbol{\mu}(C)||^2, \tag{1}$$

where $\boldsymbol{\mu}(C)$ is the center of cluster C and is known to be equal to

$$\boldsymbol{\mu}(C) = \frac{1}{|C|} \sum_{e\in C} \mathcal{E}(e) \tag{2}$$

in the Euclidean space, which may be reformulated as

$$Q(\Gamma) = \sum_{C\in\Gamma} \frac{1}{2|C|} \sum_{i\in C} \sum_{l\in C} \|\mathcal{E}(i) - \mathcal{E}(l)\|^2. \tag{3}$$

Kleinberg (2002) proposed three seemingly obvious clustering axioms for distance based clustering algorithms: richness, scale-invariance and consistency. Hereby an algorithm is a function $f(S, d) = \Gamma$ producing a partition Γ of S given the (pseudo)distance function $d : S \times S \to \mathbb{R}$ such that $d(i, i) = 0$, $d(i, l) = d(l, i) \geq 0$, where $d(i, j) = 0$ iff $i = j$. [2]

Axiom 1. *(Consistency axiom)* For any clustering function f, if $f(S, d) = \Gamma$ and d' is another (pseudo)distance function such that $d'(i, l) \geq d(i, l)$ iff i, l are from different clusters of Γ and $d'(i, l) \leq d(i, l)$ iff i, l are from the same cluster, then $f(S, d') = \Gamma$.

Axiom 2. *(Scale invariance axiom)* For any clustering function f, if the (pseudo)distance d' has the property that for an $\alpha \in \mathbb{R}^+$, $d'(i, l) = \alpha d(i, l)$, then $f(S, d) = f(S, d')$.

Axiom 3. *(Richness axiom)* For any clustering function f, for any S and for any its partition Γ, there exists a (pseudo) distance function d such that $f(S, d) = \Gamma$.

[2] Note that, using Kleinberg's pseudo-distance, the formula (3) can be generalized to (4).

The three axioms proved to be contradictory, that is, no clustering algorithm can fulfil all three requirements at once (see the proof by Kleinberg (2002)).

It is cumbersome for the domain of clustering algorithms if an axiomatic system consisting of, as Kleinberg claimed, "natural axioms" is self-contradictory. A sound axiomatic system is of real importance particularly when explainable methods of AI are demanded. Axioms may be helpful to explain at least partially why a particular clustering was obtained. Therefore numerous efforts have been made to cure such a situation by proposing different axiom sets or modifying Kleinberg's theory.

The first suggestion by Kleinberg was to use only a pair of axioms (as each pair of his axioms is not contradictory). To get rid of the contradiction between the three axioms, Kleinberg introduced the concept of partition Γ' being a refinement of a partition Γ, if for every set $C' \in \Gamma'$ there is a set $C \in \Gamma$ such that $C' \subseteq C$. He defines refinement-consistency, a relaxation of consistency, to require that if distance d' is a consistency transformation of d then $f(S, d')$ should be a refinement of $f(S, d)$ or vice versa. Though there is no clustering function that satisfies scale-invariance, richness, and refinement-consistency, if one defines near-richness as richness without the partition in which each element is in a separate cluster, then there exist clustering functions f that satisfy scale-invariance and refinement-consistency, and near-richness (e.g., single-linkage with the distance-$(\alpha\delta)$ stopping condition, where $\delta = \min_{i,j} d(i, j)$ and $\alpha \geq 1$.)

The refinement consistency does not allow generating labelled new data from existing labelled old data. Regrettably, in spite of these weakenings, the very popular k-means algorithm *is not a clustering algorithm* as it fails on (near)richness and consistency/refinement consistency axioms. Note that k-means is not (near)rich as it returns only such Γ that $k = |\Gamma|$. By weakening Kleinerg's axiom of richness to *k-richness* (richness restricted to partitions Γ with $|\Gamma| = k$), we can get rid of the violation of richness. The violation of the consistency axiom remains, though (see the proof by Kleinberg (2002)).

To overcome Kleinberg's contradictions, Ben-David and Ackerman (2009) proposed to axiomatize the clustering quality function and not the clustering function itself. Regrettably, no requirements are imposed onto the clustering function itself. This means that labelled datasets cannot be derived automatically from existing ones. Further, van Laarhoven and Marchiori (2014) propose to go over to the realm of graphs and develop a set of axioms for graphs. The approach is not applicable to k-means. Ackerman *et al*. (2010) and Meilă (2005) proposed to use the "axioms" not as a requirement to be met by all algorithms, but rather as a way to classify clustering functions. Strazzeri and Sánchez-García (2021) suggests to change the consistency axiom for graphs. Hopcroft and Kannan (2012) propose to seek only clusters with special properties, in this case to cluster the datasets into equal size clusters.

Cohen-Addad *et al*. (2018) suggest to modify the consistency axiom by requiring that Kleinberg's consistency holds only if the optimal number of clusters prior and after his Γ transformation remains the same. Though they show that various algorithms, including k-means, fit this new axiom, the problem is of course that one is usually unable to tell *a priori* the optimal number of clusters, hence usage of such an axiomatic set as a tool for test set generation is pointless.

We have also proposed several approaches to removing the contradictions in Kleinberg's axiomatic system (see, e.g., Kłopotek and Kłopotek, 2022; 2023). All of them are based on the enclosure of clusters into balls and keeping gaps between the balls large. These approaches are valid only for Euclidean spaces.

The proposals in this paper are inspired by the research on the so-called clusterability. As mentioned, Hopcroft and Kannan (2012) made a suggestion that restricting oneself to special data structures can overcome Kleinberg's contradictions. That is, one looks rather at clustering of data that fulfil some properties of clusterability. Though a number of attempts have been made to capture formally the intuition behind clusterability, none of these efforts seems to have been successful, as exhibited by Ben-David (2015) in depth. Ackerman *et al*. (2016) partially eliminate some of these problems, regrettably at the expense of non-intuitive user-defined parameters. As Ben-David (2015) mentioned, the research in the area does not address popular algorithms except for the ϵ-separatedness clusterability criterion related to k-means proposed by Ostrovsky *et al*. (2013). We have made some efforts in this direction (Kłopotek, 2020). This paper also refers to clusterability while clustering via k-means.

The issue of clustering axiomatisation is closely related to the problem of cluster preserving transformations in general. Such transformations are of vital importance because they may be used in the problem of testbed creation for clustering algorithms.

Roth *et al*. (2003) investigated the issue of preservation of clustering when embedding non-Euclidean data into the Euclidean space. They showed that clustering functions, which remain invariant under additive shifts of the pairwise proximities, can be reformulated as clustering problems in Euclidean spaces.

A similar problem was addressed by Kłopotek *et al*. (2020) whereby the issue of interpretation of results of kernel k-means to non-Euclidean data was discussed. A cluster-preserving transformation for this specific problem was proposed via increasing all distances.

Parameswaran and Blough (2005) studied the issue

of cluster preserving transformations from the point of view of privacy preserving. They designed a nearest neighbor data substitution (NeNDS), a new data obfuscation technique with strong privacy-preserving properties while maintaining data clusters. Cluster preserving transformations with the property of privacy preserving focusing on the k-means algorithm are investigated by Ramírez and Auñón (2020). Privacy preserving methods for various k-means variants boosted to large scale data are further elaborated on by Gao and Zhang (2017). Keller *et al*. (2021) investigate such transformations for other types of clustering algorithms. A thorough survey of privacy-preserving clustering for big data is presented by Zhao *et al*. (2020).

Howland and Park (2008) proposed models incorporating prior knowledge about the existing structure and developed for them dimension reduction methods independent of the original term-document matrix dimension. Other, more common dimensionality reduction methods for clustering (including PCA and Laplacian embedding) are reviewed by Ding (2009).

Larsen *et al*. (2019) reformulate the heavy hitter problem of stream mining in terms of a clustering problem and elaborate algorithms fulfilling the requirement of "cluster preserving clustering."

Zhang *et al*. (2019) developed clustering structure preserving transformations for graph streaming data, when there is a need to sample a graph.

3. Variational cluster separation

Let us introduce a couple of useful concepts. First of all, recall the fact that Kleinberg's consistency axiom leads definitely outside of the domain of the Euclidean space. Therefore, to work with the k-means algorithm, we need a reformulation of the k-means cluster quality function. Let us first recall Kleinberg's "distance" concept.

Definition 1. For a given discrete set of points S, the function $d : S \times S \to \mathbb{R}$ will be called a *pseudo-distance function* iff $d(x,x) = 0$, $d(x,y) = d(y,x)$ and $d(x,y) > 0$ for distinct x, y.

Following the spirit of kernel k-means as exposed by Kłopotek *et al*. (2020), let us reformulate the k-means cluster quality function in terms of this pseudo-distance. Define the function $Q(\Gamma, d)$ as follows:

$$Q(\Gamma, d) = \sum_{C\in\Gamma} \frac{1}{2|C|} \sum_{i\in C}\sum_{l\in C} d(i,l)^2, \tag{4}$$

where Γ is a clustering (split into disjoint non-empty subsets of cardinality at least 2) of a dataset S into k clusters, and d is a pseudo-distance function defined over S. $Q(\Gamma, d)$ generalizes $Q(\Gamma)$ from the formula (3) in that it allows non-Euclidean distances.

Let us introduce our concept of well-separatedness.

Definition 2. Let $\Gamma = \{C_1, \ldots, C_k\}$ be a partition of the dataset S and d be a pseudo-distance. Let also

$$d(i,l) > \sqrt{2}\sqrt{Q(\Gamma, d)} \tag{5}$$

for each i, l such that i belongs to a different cluster than l under Γ. Then we say that the set S with distance d is *variationally k-separable* and that this Γ is *variational k-separation* of S. If, furthermore, no cluster of Γ has the property of variational k'-separation for all $k' = 2, \ldots, K+1$ for some integer $K \geq 1$, then Γ is *variational $k+K$-range-separation* of S.

Example 1. Examine the following set of points in the Euclidean space in one dimension: $S = \{1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6\}$. A clustering with k-means into $k = 2$ clusters will yield the clustering $\Gamma = \{\{1.1, 1.2, 1.3, 1.4, 1.5, 1.6\}, \{3.1, 3.2, 3.3, 3.4, 3.5, 3.6\}\}$ with $Q(\Gamma, d) = 0.35$; hence $\sqrt{2}\sqrt{Q(\Gamma, d)} < 0.84$. The distance between elements of distinct clusters is at least 1.5, therefore Γ is a variational 2-separation of S. If we split $A \in \Gamma$ into $k' = 2$ clusters Γ_A, then $Q(\Gamma_A, d) > 0.04$ so that $\sqrt{2}\sqrt{Q(\Gamma_A, d)} > 0.28$, while the distance between elements of distinct clusters can be as small as 0.1 so that Γ_A is not a variational 2-separation of A. Therefore Γ is a variational $2+1$-range-separation of S. ♦

The following is then easily implied.

Theorem 1. *If the pseudo-distance d fulfills the condition (5) under the clustering Γ of S, then Γ is the optimal k-clustering of S with d under kernel k-means.*

Proof. Assume, contrary to our claim, that not Γ but Γ' different from it is the optimal k-clustering of S. Γ' would then contain at least one cluster C' with at least two data points P, R ($|C'| = n' \geq 2$) such that both stem from distinct clusters of Γ. Hence their distance amounts to at least $\sqrt{2}\sqrt{Q(\Gamma, d)}$. All the other $n'-2$ elements of C' fall into three categories: those belonging under Γ to the same cluster as P (n_P elements), those belonging under Γ to the same cluster as R (n_R elements), and the remaining ones (n_s elements). Then, $n_P + n_R + n_s = n' - 2$. Accordingly, within the cluster C' there are at least $(n_P+1)\cdot(n_R+1) + n_s \cdot (n_R + n_P + 2)$ pairs of data points with the distance at least $\sqrt{2}\sqrt{Q(\Gamma, d)}$. Hence the contribution of C' to the quality function amounts to

$$\begin{aligned} Q(\{C'\}, d) &= \frac{1}{2n'} \sum_{i\in C'}\sum_{l\in C'} d(i,l)^2 \\ &\geq \frac{1}{n'}\left((n_P+1)\cdot(n_R+1)\right. \\ &\quad \left. + n_s \cdot (n_R + n_P + 2)\right) \cdot 2Q(\Gamma, d) \\ &\geq \frac{1}{n'}(n'-1)\cdot 2Q(\Gamma, d). \end{aligned} \tag{6}$$

As $Q(\{C'\}, d) \geq Q(\Gamma, d)$, we have $Q(\Gamma', d) \geq Q(\Gamma, d)$ as claimed in this theorem. Γ is in fact optimal. ■

As the theorem holds for the pseudo-distance, it holds also for the Euclidean distance.

Theorem 2. *If the clustering Γ of S under the pseudo-distance d fulfills the condition (5), then there exists no other Γ' of S that fulfills the condition (5).*

Proof. As already shown in Theorem 1, Γ is optimal. Therefore, Γ' would have to be optimal but different from Γ. Yet this is impossible as putting two elements from distinct clusters would significantly increase the quality function value, as seen in the proof of the previous theorem. ■

Definition 3. We say that a clustering function $f(S,d)$ returns a *variational k-clustering* of S if S is variationally k-separable under d and $f(S,d)$ returns the clustering Γ being a variational k-separation of S.

Example 2. As the set S from Example 1 is variationally 2-separable, the function f should return the clustering Γ from that example, according to Theorem 1. ♦

Theorem 3. *A variational k-clustering Γ will remain a variational k-clustering after a consistency transform. In other words, the consistency transform preserves clustering by a function detecting the variational k-clustering.*

Proof. The increase in inter-cluster distances does not violate the variational k-separation because the distances between clusters will be larger than prescribed by the variational minimal distance from the formula (5). The decrease in intra-cluster separation does not violate variational k-separation because the variational minimal distance will be smaller, so that the distances between clusters will fit better this minimal distance. ■

Example 3. After a consistency transformation with respect to Γ, the set S from Example 1 may change coordinates:

$$S' = \{1.1, 1.15, 1.2, 1.25, 1.3, 1.35, 3.3,$$
$$3.36, 3.42, 3.48, 3.54, 3.6\}.$$

A clustering with k-means into $k = 2$ clusters will yield the clustering

$$\Gamma' = \{\{1.1, 1.15, 1.2, 1.25, 1.3, 1.35\},$$
$$\{3.3, 3.36, 3.42, 3.48, 3.54, 3.6\}\}$$

with $Q(\Gamma', d) < 0.11$ hence $\sqrt{2}\sqrt{Q(\Gamma', d)} < 0.48$. The distance between elements of distinct clusters is at least 1.95, therefore Γ' is a variational 2-separation of S', and both Γ and Γ' are identical. ♦

Let us concentrate on the Euclidean distances only for a moment. Let us ask the question how difficult it would be to discover the optimal clustering. Let us apply for this purpose the k-means++ algorithm, developed by Arthur and Vassilvitskii (2007), or more precisely the derivation of the initial clustering. The pseudo-code of k-means++ is presented as Algorithm 1. Recall that wide gaps between clusters guarantee that, after hitting each cluster during the initialization stage, an optimum clustering is achieved. Let us consider a step when i seeds have hit i distinct clusters. Then the probability of hitting an unhit cluster in the next step amounts to

$$\mathrm{P}_{\mathrm{HUH}} = \frac{\mathrm{SSD}_{\mathrm{unhit}}}{\mathrm{SSD}_{\mathrm{unhit}} + \mathrm{SSD}_{\mathrm{hit}}} = 1 - \frac{\mathrm{SSD}_{\mathrm{hit}}}{\mathrm{SSD}_{\mathrm{unhit}} + \mathrm{SSD}_{\mathrm{hit}}}, \tag{7}$$

where $\mathrm{SSD}_{\mathrm{hit}}$ is the sum of the squared distances to the closest seed from the elements of hit clusters $C_1, C_2, \ldots, C_i \in \Gamma_{\mathrm{hit}}$, and $\mathrm{SSD}_{\mathrm{unhit}}$ is the sum of the squared distances to the closest seed from the elements of unhit clusters $C_{i+1}, \ldots, C_k$. Let C be a hit cluster. Then $Q(\{C\}, d)$ will be the upper bound for the squared

Algorithm 1. Pseudo-code for k-means++ algorithms; the termination condition of the while loop may be a fixed number of loop runs or no change in clustering within a loop run.

Require: D: a set of objects embedded in the Euclidean space (that is, for each $i \in D$ there exists its representation $\boldsymbol{x}_i$ in the Euclidean space), to be clustered
k: the number of clusters to be returned
{Initialize the set M of k cluster centers as follows:}
Pick one element e of D at random and initialize the set M with $\boldsymbol{\mu}_1 = \mathcal{E}(e)$.
for $j \leftarrow 2$ **to** k **by** 1 **do**
  Assign to each $e \in D$ a weight $w_e = \min_{\boldsymbol{\mu} \in M} ||\mathcal{E}(e) - \boldsymbol{\mu}||^2$ and a probability $p_e = w_e / \sum_{e' \in D} w_{e'}$.
  Sample one element $e \in D$ according to the above-mentioned probability p_e.
  $M \leftarrow M \cup \{\boldsymbol{\mu}_j = \mathcal{E}(e)\}$
end for
{This ends the initialization of M}
while termination not reached **do**
  Let $\Gamma = \{C_1, \ldots, C_k\}$, where each $C_j = \emptyset$.
  for each $e \in D$ **do**
    $C_j = C_j \cup \{e\}$, where $j = \arg\min_{j'=1,\ldots,k} ||\mathcal{E}(e) - \boldsymbol{\mu}_{j'}||$
  end for
  for $j \leftarrow 1$ **to** k **by** 1 **do**
    $\boldsymbol{\mu}_j = \boldsymbol{\mu}(C_j)$, according to (2)
  end for
end while
return Γ { Γ: the clustering of D into k clusters }

distance between any element of C and the cluster center. Hence $2Q(\{C\}, d)$ will be the upper bound of the sums of the squared distances between a seed from C and its other elements. Therefore,

$$\mathrm{SSD}_{\mathrm{hit}} \leq 2Q(\Gamma_{\mathrm{hit}}, d) \leq 2Q(\Gamma, d). \tag{8}$$

On the other hand,

$$\mathrm{SSD}_{\mathrm{unhit}} \geq 2Q(\Gamma, d) \sum_{j=i+1}^{k} n_j, \tag{9}$$

where $n_j = |C_j|$. Hence,

$$\begin{aligned} \mathrm{P}_{\mathrm{HUH}} &= \frac{\mathrm{SSD}_{\mathrm{unhit}}}{\mathrm{SSD}_{\mathrm{unhit}} + \mathrm{SSD}_{\mathrm{hit}}} = \frac{1}{1 + \frac{\mathrm{SSD}_{\mathrm{hit}}}{\mathrm{SSD}_{\mathrm{unhit}}}} \\ &\geq \frac{1}{1 + \frac{2Q(\Gamma,d)}{2Q(\Gamma,d)\sum_{j=i+1}^{k} n_j}} = \frac{1}{1 + \frac{1}{\sum_{j=i+1}^{k} n_j}} \\ &= \frac{\sum_{j=i+1}^{k} n_j}{\sum_{j=i+1}^{k} n_j + 1} = 1 - \frac{1}{\sum_{j=i+1}^{k} n_j + 1}. \end{aligned} \tag{10}$$

If we assume that the cardinality of all clusters is the same and equals m, then we have

$$\mathrm{P}_{\mathrm{HUH}} \geq 1 - \frac{1}{m(k-i)+1}, \tag{11}$$

so that the overall expected probability of hitting all clusters during initialization amounts to at least

$$\mathrm{P}_{\mathrm{hitAll}} \geq \prod_{i=1}^{k-1} \left(1 - \frac{1}{m(k-i)+1}\right). \tag{12}$$

If m exceeds k, then this probability is very close to one (assuming $m > 50$). If not all clusters are of the same cardinality, but m is its lower bound, then the above formula gives a lower bound to the this probability.

Theorem 4. *There exists a function detecting a variational k-clustering, with high probability, that has the property of scale-invariance, consistency and k-richness, given that the function operates in the Euclidean space and the consistency transformation is performed in the Euclidean space, too.*

Proof. We have just shown that k-means++ can be used to detect, with high probability, a variational k-clustering if the data lie in the Euclidean space. It is known to have the property of scale-invariance. Then k-richness is easily shown: formulate a k-clustering Γ, set distances between points within each cluster to values such that each cluster fits the Euclidean space, and then move the clusters in the Euclidean space in such a way that the condition (5) is matched and complete the distance definition. The consistency property holds because of Theorem 3. ■

We return to investigating pseudo-distances. Let us go beyond k-richness, expanding our discussion towards the concept of richness. Already Kleinberg showed that full richness does not make sense and restricted himself to near-richness. Below we restrict the concept of near-richness to range-k_x-richness.

Definition 4. A clustering function f has the range-k_x-richness property if, for any dataset S for each $\Gamma \in 2^S$ consisting of non-empty subsets of at least two elements such that $|\Gamma| \leq k_x$, there exists a distance function d such that $f(S, d) = \Gamma$.

Note that near richness imposes the restriction $|\Gamma| \leq |S| - 1$. It allows also for clusters with one element only which we forbid in range-k_x-richness.

Definition 5. We say that a clustering function $f(S, d)$ returns a *variational range-k_x clustering* of S if S is variationally k-separable under d for some $1 \leq k \leq k_x$, and for $\Gamma = f(S, d)$ for no cluster $C \in \Gamma$ there exists k', $2 \leq k' \leq k_x - k + 1$ that C is variationally k'-separable. The maximal k with this property shall be called the *level of variational range-k_x clustering*.

Example 4. If we apply k-means with k-ranging from 1 to 3 to the set S from Example 1, then it will return a *variational range-3 clustering* of S, because S is variationally 2-separable under the Euclidean distance, and for $\Gamma = f(S, d)$ for no cluster $C \in \Gamma$ there exists k', $2 \leq k' \leq 2$ that C is variationally k'-separable. ♦

What will happen when performing Kleinberg's consistency operation? A cluster that is not variationally k' separable may turn to a variationally k' separable one if we apply a consistency transformation. Therefore we need to restrict the consistency transformation. We suggest to replace it with the relative consistency transformation, defined as follows:

Definition 6. Consider a dataset S and a distance function $d : S \times S \to \mathbb{R}$ and a clustering function f. Let $f(S, d) = \Gamma$. Define a different distance function d' such that for any cluster $C \in \Gamma$: (1) for $i, j, l \in C$, $d'(i, j) \leq d(i, j)$ and if $d(i, j) \leq d(i, l)$ then $d'(i, j) \leq d'(i, l)$, and $\frac{d'(i,l)}{d'(i,j)} \leq \frac{d(i,l)}{d(i,j)}$, (2) for $i \in C$ and $l \notin C$ $d'(i, l) \geq d(i, l)$. This transformation from d to d' with be called the *relative consistency transformation*.

Example 5. Assume the following distance matrix within a cluster:

$$\begin{bmatrix} 0.0 & 0.1 & 0.2 & 0.3 & 0.4 & 0.5 \\ 0.1 & 0.0 & 0.1 & 0.2 & 0.3 & 0.4 \\ 0.2 & 0.1 & 0.0 & 0.1 & 0.2 & 0.3 \\ 0.3 & 0.2 & 0.1 & 0.0 & 0.1 & 0.2 \\ 0.4 & 0.3 & 0.2 & 0.1 & 0.0 & 0.1 \\ 0.5 & 0.4 & 0.3 & 0.2 & 0.1 & 0.0 \end{bmatrix}.$$

A sample *relative consistency transformation* would yield the following:

$$\begin{bmatrix} 0.000000 & 0.09000 & 0.1620 \\ 0.090000 & 0.00000 & 0.0900 \\ 0.162000 & 0.09000 & 0.0000 \\ 0.226800 & 0.16200 & 0.0900 \\ 0.287280 & 0.22680 & 0.1620 \\ 0.344736 & 0.28728 & 0.2268 \end{bmatrix}$$

$$\begin{bmatrix} 0.2268 & 0.28728 & 0.344736 \\ 0.1620 & 0.22680 & 0.287280 \\ 0.0900 & 0.16200 & 0.226800 \\ 0.0000 & 0.09000 & 0.162000 \\ 0.0900 & 0.00000 & 0.090000 \\ 0.1620 & 0.09000 & 0.000000 \end{bmatrix}$$

♦

The relative consistency transformation defined above differs from Kleinerg's consistency transformation in the following way: (i) it preserves the ordering of distances within a cluster, (ii) it prevents the emergence of more dense areas within a cluster. In this way, no new clusters emerge within a cluster after this transformation, contrary to Kleinberg's definition. This new definition removes a crucial deficiency of Kleinberg's axiomatic system.

Definition 7. If the clustering function f for each data set S and each distance function d and each of its relative consistency transforms d' have the property that $f(S,d) = f(S,d')$, then we shall say that f has the property of *relative consistency*.

Theorem 5. *A variational range-k_x clustering at the level k will remain a variational range-k_x clustering at the level k after the relative consistency transform. In other words, the relative consistency transform preserves clustering by a function detecting a variational range-k_x clustering.*

Proof. An increase in the inter-cluster distances does not violate the variational k-separation because the distances between clusters will be larger than prescribed by the variational minimal distance. The decrease of the intra-cluster separation does not violate the variational k-separation because the variational minimal distance will be smaller; therefore distances between clusters will fit better this minimal distance. Furthermore, a decrease of the intra-cluster separation does not turn a non-variationally separable set into a separable set for the following reason: assume S_1 and S_2 are two subclusters of a cluster S which we treat as candidates for being variationally separated after the transformation. This implies that the distances between elements of S_1 and S_2 were larger than within S_1 and within S_2 after the operation, and so were they before the operation. But if they were larger before the operation, then they are more strongly shortened than those within S_1 and S_2. Yet this means that the decrease in the variational minimal distance is smaller than the decrease in the distances between S_1 and S_2. Hence the variational separation cannot occur. ■

We need, however, an algorithm that would actually perform the above-mentioned clustering. To identify one, we have to return to the Euclidean distance. Consider the following algorithm f, discovering a variational range-k_x clustering of a dataset (Algorithm 2): try out all $k = k_x$ to 2 if there exists a variational k-clustering; and if so, then check each sub-cluster on no variational k' separability. Obviously, k-means++ would be a suitable sub-algorithm for checking if there exists a variational k-clustering.

Let us estimate the complexity of Algorithm 2. The computation of distances is the most expensive step here, The complexity of k-means++ (Algorithm 1), to which Algorithm 2 is referring, consists of two parts: initializing and iteration. During initialization in the j-th step of k steps, $(j-1)n$ distances will be computed, where n is the number of data samples. Thus the complexity amounts to $O(nk(k-1)/2)$. If the initialization is successful (each cluster is seeded), then in the case considered in this paper at most two iteration steps are needed, in each kn distance computations. Hence we are left with $O(nk^3)$ complexity. Algorithm 2 calls k-means++ at most $k_x!$ times, so its complexity is $O(nk_x^3 k_x!)$. But in practice, with no strange data distribution, only $2k$ calls are needed. Thus the expected complexity is $O(nk_x^4)$

Theorem 5 implies the following.

Theorem 6. *The clustering function described by Algorithm 2, detecting a variational range-k_x clustering with high probability, has the property of scale-invariance, relative consistency and range-k_x richness if operating in the Euclidean space.*

However, we will have a problem with the relative consistency transformation of a distance d to a distance d'. In the general case, even if d is an Euclidean distance, d' does not need to be one. As shown by Kłopotek *et al.* (2020), a distance function d' being non-Euclidean can be turned into a Euclidean one d'' by adding an appropriate constant δ^2 to each squared distance $d'(i,j)^2$ of distinct elements and the clustering with (kernel) k-means will preserve the k-clustering of S. However, it is possible that the property of variational k separability will be lost via such an addition operation. But our goal is to find a separability criterion for which there exists a clustering function, operating in the Euclidean space, such that it fits axioms. In other words, if the axiomatic transformations lead outside the Euclidean space, then application of Euclidization should transform the data so

Algorithm 2. Discovery of variational range-k_x clustering of a dataset.

Require: S: a set of objects embedded in the Euclidean space
k_x: the maximal number of clusters to be obtained
if $k_x < 2$ **then**
  return $k = 1, \Gamma = \{S\}$
end if
for $k \leftarrow k_x$ **to** 2 **by** -1 **do**
  Cluster S using k-means++ (Algorithm 1) getting Γ
  if Γ ensures that according to Definition 2 S is variationally k-separable **then**
    OK = TRUE
    for $S' \in \Gamma$ **do**
      Apply this algorithm to S' with $k'_x = k_x - k + 1$ obtaining k' and Γ'
      if $k' > 1$ **then**
        OK=FALSE
      end if
    end for
  end if
  if OK **then**
    return k, Γ
  end if
end for
return $k = 1, \Gamma = \{S\}$

that the clustering function returns a clustering that fits the axiomatic requirements.

4. Residual cluster separation

Assume that $\sigma(d)$ is the lowest distance d over the set S.

Theorem 7. *Let Γ be a clustering of the set S, and let $n = |S|$. Then*

$$Q(\Gamma, d) \geq (n-k)\frac{\sigma(d)^2}{2}. \tag{13}$$

Proof. We have

$$\begin{aligned} Q(\Gamma, d) &\geq \sum_{C\in\Gamma} \frac{1}{2|C|} \sum_{i\in C} \sum_{l\in C; l\neq i} \sigma(d)^2 \\ &= \sum_{C\in\Gamma} \frac{|C|-1}{2}\sigma(d)^2 = (|S|-k)\frac{\sigma(d)^2}{2}. \end{aligned} \tag{14}$$

■

Define the function

$$\beta(\Gamma, d) = 2\left(Q(\Gamma, d) - (n-k-1)\frac{\sigma(d)^2}{2}\right). \tag{15}$$

Let us introduce our next concept of well-separatedness.

Definition 8. Let $\Gamma = \{C_1, \ldots, C_k\}$ be a partition of the dataset S and d be a pseudo-distance. Let also

$$d(i, l) > \sqrt{\beta(\Gamma, d)} \tag{16}$$

for each i, l such that i belongs to a different cluster than l under Γ. Then we say that the set S with distance d is *residually k-separable*, and Γ is the *residual k-separation of S*. If, furthermore, no cluster of Γ has the property of residual k'-separation for all $k' = 2, \ldots, K+1$, then Γ is a *residual $k + K$-range-separation* of S.

Example 6. Consider the following set of points in the Euclidean space in one dimension:

$$S = \{1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9\}.$$

A clustering with k-means into $k = 2$ clusters will yield the clustering

$$\Gamma = \{\{1.1, 1.2, 1.3, 1.4, 1.5, 1.6\}, \{2.4, 2.5, 2.6, 2.7, 2.8, 2.9\}\}$$

with $Q(\Gamma, d) = 0.35$; therefore $\sqrt{2}\sqrt{Q(\Gamma, d)} > 0.83$. The distance between the elements of distinct clusters is at least 0.8; therefore, Γ is *not* variational 2-separation of S, but $\sqrt{\beta(\Gamma, d)} < 0.79$. Therefore, Γ is a residual 2-separation of S. If we split $A \in \Gamma$ into $k' = 2$ clusters Γ_A, then $Q(\Gamma_A, d) > 0.04$, so that $\sqrt{\beta(\Gamma_A, d)} > 0.22$, while the distance between the elements of distinct clusters can be as small as 0.1, so that Γ_A is not a residual 2-separation of A. Therefore, Γ is a residual $2+1$-range-separation of S. ♦

Theorem 8. *Assume that the set S with distance d is* residually k-separable. *Then Γ minimizes $Q(\Gamma, d)$ over all clusterings of the dataset S.*

Proof. By analogy to the proof of Theorem 1, we can demonstrate that $\beta(\Gamma', d) \geq \beta(\Gamma, d)$. Hence

$$\begin{aligned} &2\left(Q(\Gamma', d) - (n-k-1)\frac{\sigma(d)^2}{2}\right) \\ &\qquad \geq 2\left(Q(\Gamma, d) - (n-k-1)\frac{\sigma(d)^2}{2}\right). \end{aligned} \tag{17}$$

That is, $Q(\Gamma', d) \geq Q(\Gamma, d)$. ■

Theorem 9. *Assume we have two pseudo-distance functions d_1, d_2 over S such that for any two distinct x, y there holds $d_2^2(x, y) = d_1^2(x, y) + \Delta$ for some constant Δ. Then*

$$\beta(\Gamma, d_2) = \beta(\Gamma, d_1) + \Delta. \tag{18}$$

Proof. We have

$$\begin{aligned} Q(\Gamma, d_2) &= \sum_{C\in\Gamma} \frac{1}{2|C|} \sum_{i\in C}\sum_{l\in C} d_2(i,l)^2 \\ &= \sum_{C\in\Gamma} \frac{1}{2|C|} \sum_{i\in C}\sum_{l\in C; l\neq i} \left(d_1(i,l)^2 + \Delta\right) \\ &= Q(\Gamma, d_1) + \sum_{C\in\Gamma} \frac{1}{2|C|} \sum_{i\in C}\sum_{l\in C; l\neq i} \Delta \\ &= Q(\Gamma, d_1) + (n-k)\frac{\Delta}{2} \\ &= Q(\Gamma, d_1) + \sum_{C\in\Gamma} \frac{1}{2|C|} \sum_{i\in C}\sum_{l\in C; l\neq i} \Delta \\ &= Q(\Gamma, d_1) + (n-k)\frac{\Delta}{2}. \end{aligned} \tag{19}$$

Furthermore,

$$\begin{aligned} \beta(\Gamma, d_2) &= 2\left(Q(\Gamma, d_2) - (n-k-1)\frac{\sigma(d_2)^2}{2}\right) \\ &= 2Q(\Gamma, d_1) + (n-k)\Delta \\ &- (n-k-1)\sigma(d_1)^2 - (n-k-1)\Delta \\ &= 2Q(\Gamma, d_1) - (n-k-1)\sigma(d_1)^2 + \Delta \\ &= \beta(\Gamma, d_1) + \Delta. \end{aligned} \tag{20}$$

■

The above theorem implies the following.

Theorem 10. *Assume we have two pseudo-distance functions d_1, d_2 over S such that for any two distinct x, y there holds $d_2^2(x,y) = d_1^2(x,y) + \Delta$ for some constant Δ. Then the set S with pseudo-distance d_1 is* residually k-separable *iff the set S with pseudo-distance d_2 is* residually k-separable.

Proof. As $\beta(\Gamma, d_2) = \beta(\Gamma, d_1) + \Delta$, then it would be sufficient for *residual k-separability* of S under d_2 that the squared pseudo-distance between the elements of distinct clusters is increased by Δ which is the case by definition of d_2. Therefore, an increase in the distances from d_1 to d_2 preserves the residual k-separation. On the other hand, if Γ with $|\Gamma| = k$ is not a residual k-separation under d_1, then there exist two elements i, l from distinct clusters such that $d_1(i,l)^2 \leq \beta(\Gamma, d_1)$. Therefore $d_2(i,l)^2 = d_1(i,l)^2 + \Delta \leq \beta(\Gamma, d_1) + \Delta = \beta(\Gamma, d_2)$. ■

Definition 9. We say that a clustering function $f(S,d)$ returns *a residual k-clustering* of S if S is residually k-separable under d and $f(S,d)$ returns the clustering Γ being a residual k-separation of S.

Theorem 11. *A residual k-clustering Γ will remain a residual k-clustering after the consistency transform, given that no pseudo-distance gets shorter than the shortest distance at the beginning (*lower-bounded consistency*). In other words, the consistency transform preserves clustering by a function detecting a variational k-clustering.*

Proof. An increase in the inter-cluster distances does not violate the residual k-separation because the distances between clusters will be larger than prescribed by the residual minimal distance from the formula (16). A decrease in the intra-cluster separation does not violate the residual k-separation because the residual minimal distance (16) will decrease so that the distances between clusters will fit better this minimal distance. ■

The concept of lower-bounded consistency may appear somehow awkward from the mathematical point of view, but it is not so if we look at technical reality. Any distance measurement is restricted by some resolution factor of the measuring device. Therefore, if two points are too close, they may be indistinguishable. Thus, assuming a minimal distance between distinct data points makes technically sense.

Let us concentrate for a moment on Euclidean distances. Obviously, k-means++ is no more suitable for discovering residual k-clustering. We propose to create a modification of k-means++, called *res-k-means++* (Algorithm 3). Instead of taking the squared distances to the closest seed in Line 4 in Algorithm 1, assign to each element $e \in D$ a probability p_e, proportional to the difference between the squared distance of e to the closest element of M and the squared smallest distance.

Let us ask the question how difficult it would be to discover an optimal clustering. Let us study the res-k-means++ algorithm, or more precisely, the derivation of the initial clustering, whereby Γ is the true clustering. Consider a step when i seeds have hit i distinct true clusters $\mathcal{H}$. For a hit cluster C, let $h(C)$ be the hit seed of this cluster. Then the probability of hitting an unhit cluster in the next step amounts to

$$\mathrm{P}_{\mathrm{HUH}}^{(M)} = \frac{\mathrm{SSDM}_{\mathrm{unhit}}}{\mathrm{SSDM}_{\mathrm{unhit}} + \mathrm{SSDM}_{\mathrm{hit}}}, \tag{21}$$

where $\mathrm{SSDM}_{\mathrm{unhit}}$ is the sum of the squared distances to the closest seed minus $\delta(d)^2$ from the elements of unhit clusters, and $\mathrm{SSDM}_{\mathrm{hit}}$ is the sum of the squared distances minus $\delta(d)^2$ to the closest seed from elements of hit clusters,

$$\mathrm{SSDM}_{\mathrm{hit}} = \sum_{C\in\mathcal{H}} \sum_{e\in C; e\neq h(C)} \left(d^2(e, h(C)) - \sigma(d)^2\right). \tag{22}$$

However, for any $l \in C$,

$$\begin{aligned} &\sum_{e\in C; e\neq l} (d^2(e,j) - \sigma(d)^2) \\ &\qquad \leq 2Q(\{C\}, d) - (|C|-1)\sigma(d)^2). \end{aligned} \tag{23}$$

Algorithm 3. Pseudo-code for res-k-means++; the termination condition of the while loop may be a fixed number of loop runs or no change of clustering within a loop run.

Require: D: a set of objects embedded in the Euclidean space (that is, for each $i \in D$ there exists its representation $\boldsymbol{x}_i$ in the Euclidean space), to be clustered
k: the number of clusters to be returned
{ Initialize the set M of k cluster centers as follows: }
Pick one element e of D at random and initialize the set M with $\boldsymbol{\mu}_1 = \mathcal{E}(e)$.
for $j \leftarrow 2$ **to** k **by** 1 **do**
  Assign to each $e \in D$ a weight $w_e = \min_{\boldsymbol{\mu} \in M} ||\mathcal{E}(e) - \boldsymbol{\mu}||^2 - \sigma(d)^2$ and a probability $p_e = w_e / \sum_{e' \in D} w_{e'}$.
  Sample one element $e \in D$ according to the above-mentioned probability p_e.
  $M \leftarrow M \cup \{\boldsymbol{\mu}_j = \mathcal{E}(e)\}$
end for
{This ends the initialization of M}
while termination not reached **do**
  Let $\Gamma = \{C_1, \ldots, C_k\}$, where each $C_j = \emptyset$
  for each $e \in D$ **do**
    $C_j = C_j \cup \{e\}$, where $j = \arg\min_{j'=1,\ldots,k} ||\mathcal{E}(e) - \boldsymbol{\mu}_{j'}||$
  end for
  for $j \leftarrow 1$ **to** k **by** 1 **do**
    $\boldsymbol{\mu}_j = \boldsymbol{\mu}(C_j)$, according to (2)
  end for
end while
return Γ { Γ: the clustering of D into k clusters }

Thus

$$\begin{aligned}\mathrm{SSDM}_{\mathrm{hit}} &\le \sum_{C \in \mathcal{H}} \left(2Q(\{C\}, d) - (|C| - 1)\sigma(d)^2\right) \\ &\le 2Q(\Gamma, d) - (n - k)\sigma(d)^2) < \beta(\Gamma, d).\end{aligned} \tag{24}$$

On the other hand,

$$\begin{aligned}\mathrm{SSDM}_{\mathrm{unhit}} &\ge \beta(\Gamma, d) \sum_{C \in \Gamma - \mathcal{H}} |C| \\ &= \beta(\Gamma, d) \sum_{j=1}^{k} n_j.\end{aligned} \tag{25}$$

Hence

$$\begin{aligned}\mathrm{P}^{(M)}_{\mathrm{HUH}} &= \frac{\mathrm{SSDM}_{\mathrm{unhit}}}{\mathrm{SSDM}_{\mathrm{unhit}} + \mathrm{SSDM}_{\mathrm{hit}}} \\ &\ge \frac{1}{1 + \frac{\beta(\Gamma,d)}{\beta(\Gamma,d)\sum_{j=i+1}^{k} n_j}} \\ &= \frac{1}{1 + \frac{1}{\sum_{j=i+1}^{k} n_j}} \\ &= \frac{\sum_{j=i+1}^{k} n_j}{\sum_{j=i+1}^{k} n_j + 1} \\ &= 1 - \frac{1}{\sum_{j=i+1}^{k} n_j + 1}.\end{aligned} \tag{26}$$

If we assume that the cardinality of all clusters is the same and equals m, then we have

$$\mathrm{P}^{(M)}_{\mathrm{HUH}} \ge 1 - \frac{1}{m(k-i)+1}. \tag{27}$$

Hence that the overall expected probability of hitting all clusters during initialization amounts to at least (as in Eqn. (12))

$$\mathrm{P}^{(M)}_{\mathrm{hitAll}} \ge \prod_{i=1}^{k-1} \left(1 - \frac{1}{m(k-i)+1}\right). \tag{28}$$

Remarks on high probability and unequal cluster sizes are here the same as with Eqn. (12).

Theorem 12. *There exists a function detecting a residual k-clustering, with high probability, that has the property of scale-invariance, lower-bounded consistency and k richness, given that the function operates in the Euclidean space and the consistency transformation is performed in the Euclidean space, too.*

Proof. We have just shown that res-k-means++ can be used to detect, with high probability, a residual k-clustering, if the data lies in the Euclidean space. Obviously, it has the property of scale-invariance. Here k-richness is easily shown: formulate a k-clustering Γ, set distances between points within each cluster to values such that each cluster fits the Euclidean space, and then move the clusters in the Euclidean space in such a way that the condition (16) is matched and complete the distance definition. The consistency property holds because of Theorem 11. ■

Let us return to pseudo-distances.

Definition 10. We say that a clustering function $f(S, d)$ returns a *residual range-k_x clustering* of S if S is residually k-separable under d for some $1 \le k \le k_x$, and for $\Gamma = f(S, d)$ and for no cluster $C \in \Gamma$ there exists k', $2 \le k' \le k_x - k + 1$ that C is residually k'-separable. The maximal k with this property will be called the level of this residual range-k_x clustering.

Let us ask what will happen when performing Kleinberg's consistency operation. The first problem that we encounter is that the consistency transform performed on a cluster that is not a residually k' separable may turn to a residually k' separable one. Therefore, we need to

restrict the consistency transformation to a relative one. Yet this is not sufficient. As we make use of the concept of the smallest distance in our formulas, we need to add the restriction that the lowest distance will not be decreased.

Theorem 13. *The residual range-k_x clustering at the level k will remain the residual range-k_x clustering at the level k after lower bounded relative consistency transform. In other words, the lower-bounded relative consistency transform preserves a clustering by a function detecting a residual range-k_x clustering.*

Proof. An increase in the inter-cluster distances does not violate the residual k-separation. A decrease in the intra-cluster separation according to the imposed limitations does not turn a non-residually separable set into a separable set. The proof is analogous to that of Theorem 5. ■

So far, we have investigated properties of algorithms discovering residual clusterings. We need, however, an algorithm that would really perform the above-mentioned clustering. To identify one, we have to return again to the Euclidean distance. To discover, with high probability, *a residual range-k_x clustering* in the Euclidean domain, we suggest to use res-k-means++ as a sub-algorithm for the master Algorithm 4: try out all $k = k_x$ to 2 if there exist, residual k-clusterings, and if so, then check each sub-cluster on no residual k' separability.

This implies the following.

Theorem 14. *Algorithm 4 detecting a residual range-k_x clustering with high probability has the property of scale-invariance, lower-bounded relative consistency and range-k_x richness in the Euclidean domain.*

5. From the Euclidean space to Kleinberg's concept of distance

A number of practical settings can give rise to non-Euclidean and, in general, pseudo-distances: imprecise distance measurements, the impact of hydrological networks, the possibility of using different alternative means of transport, measurements based on road networks or on fuel consumption, just to mention a few.

We have demonstrated that the k-means algorithm, designed originally for the Euclidean space, does not need to be in conflict with Kleinberg's consistency axiom if the dataset contains clearly separated clusters. What is more, after a slight adjustment of the consistency axiom to some real world conditions (resolution of data is finite), based on k-means, a clustering algorithm can be constructed matching in practice the three Kleinberg axioms. There is, however, one deficiency in the approach: it assumes that the data are embedded in the Euclidean space, while Kleinberg insisted that his axioms should hold also

Algorithm 4. Residual clustering.

Require: S: a set of objects embedded in the Euclidean space
k_x: the maximal number of clusters to be obtained
if $k_x < 2$ **then**
  return $k = 1, \Gamma = \{S\}$
end if
for $k \leftarrow k_x$ **to** 2 **by** -1 **do**
  Cluster S using res-k-means++ getting Γ
  if Γ ensures that according to Definition 8 S is residually k-separable **then**
    $OK =$ TRUE
    for $S' \in \Gamma$ **do**
      Apply this algorithm to S' with $k'_x = k_x - k + 1$ obtaining k' and Γ'
      if $k' > 1$ **then**
        $OK =$ FALSE
      end if
    end for
  end if
  if OK **then**
    return k, Γ
  end if
end for
return $k = 1, \Gamma = \{S\}$

outside the Euclidean realm. Most of the proofs presented do not depend on the Euclidean embedding. The weak point in going beyond it is the k-means algorithm, which was designed for the Euclidean space. While the development of k-means++ (Arthur and Vassilvitskii, 2007) is not bound to the Euclidean space, the minimum of the k-means quality function is guaranteed on the concrete properties in the Euclidean space. The so-called kernel k-means probably seeks the same minimum as the traditional k-means after the Euclidization proposed by Lingoes (1971),[3] but the problem is that after this Euclidization the variational k-separability may be lost so that there is no guarantee that k-means seeks to optimize by variationally finding k-separation.

We have discussed so far the case of clustering axioms for the Euclidean distance: see Theorem 14. This axiomatic system does not approximate quite what Kleinberg proposed because he used a more relaxed version of distance function, i.e., the pseudo-distance.

Therefore, consider the lower-bounded relative consistency transformation applied to a pseudo-distance d yielding another pseudo-distance d', that is, one outside of the framework of the Euclidean space. The proof of Theorem 14 can be easily converted to the case of pseudo-distances, using insights from Theorem 10.

[3]Note that some variants of center-based algorithms (e.g., Sabo, 2014) would require a different Euclidization method (e.g., Cailliez, 1983).

We need only to adapt accordingly the res-k-means algorithm. Adaptation of algorithms can follow the results of Kłopotek *et al*. (2020). As shown there, a distance function d being non-Euclidean can be turned into a Euclidean one d_E by adding an appropriate constant δ^2 to each squared distance $d(i,j)^2$, and the clustering with k-means under d_E will preserve the k-clustering obtained via kernel k-means with the original distance d. Theorem 10 strengthens that result by stating that the residual k-separation property before and after this transformation is the same. Hence we can use Algorithm 4 for discovery of a residual range-k_x clustering, after transforming to the Euclidean distance in the spirit of Theorem 10.

However, at a closer look, the transformation to the Euclidean distance can be "virtual", that is, no operations are needed at all, because the function β will return same results prior and after the Euclidization. Hence no actual Euclidization is needed in order to apply Algorithm 4.

6. Conclusions

This research showed that one should not throw away Kleinberg's axioms because of their contradiction. We pointed at what was missing in Kleinberg's axiomatic system, that is, the idea that clustering transformation functions make sense only if they are applied to a clustering performed on a clusterable dataset. We showed that, if the dataset is clusterable according to a properly defined separation criterion, then k-means stops to be inconsistent in terms of Kleinberg, and a version of k-means can be created that matches all three clustering axioms. It is also easily seen that single-link algorithms, used by Kleinberg (2002), can be upgraded to match all three Kleinberg axioms with clusterable data.

Acknowledgeably, the gaps between clusters used in this paper are (very) large[4] and therefore further research should seek to lower inter-cluster distances while still keeping the axiomatic system intact. Alternatively, one may investigate the degrees of violation of Kleinberg's axiomatic system given the extent to which the clusterability criteria are violated.

References

Ackerman, M., Adolfsson, A. and Brownstein, N. (2016). An effective and efficient approach for clusterability evaluation, https://arxiv.org/abs/1602.06687.

Ackerman, M., Ben-David, S. and Loker, D. (2010). Towards property-based classification of clustering paradigms, *in* J.D. Lafferty *et al.* (Eds), *Advances in Neural Information Processing Systems 23*, Curran Associates, Inc., San Francisco, pp. 10–18.

Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding, *in* N. Bansal *et al.* (Eds), *11th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007*, SIAM, Philadelphia, pp. 1027–1035.

Ben-David, S. (2015). Computational feasibility of clustering under clusterability assumptions, https://arxiv.org/abs/1501.00437.

Ben-David, S. and Ackerman, M. (2009). Measures of clustering quality: A working set of axioms for clustering, *in* D. Koller *et al.* (Eds), *Advances in Neural Information Processing Systems 21*, Curran Associates, Inc., San Francisco, pp. 121–128.

Cailliez, F. (1983). The analytical solution of the additive constant problem, *Psychometrika* **48**(2): 305–308.

Cohen-Addad, V., Kanade, V. and Mallmann-Trenn, F. (2018). Clustering redemption—Beyond the impossibility of Kleinberg's axioms, *in* S. Bengio *et al.* (Eds), *Advances in Neural Information Processing Systems*, Vol. 31, Curran Associates, Inc., San Francisco.

Ding, C. (2009). Dimension reduction techniques for clustering, *in* L. Liu and M. Oezsu (Eds), *Encyclopedia of Database Systems*, Springer, Boston, p. 846.

Gao, Z. and Zhang, L. (2017). DPHKMS: An efficient hybrid clustering preserving differential privacy in spark, *in* L. Barolli *et al.* (Eds), *Advances in Internetworking, Data & Web Technologies*, Lecture Notes on Data Engineering and Communications Technologies, Vol. 6, Springer, Cham, pp. 367–377.

Girolami, M. (2002). Mercer kernel-based clustering in feature space, *IEEE Transactions on Neural Networks* **13**(3): 780–784.

Hopcroft, J. and Kannan, R. (2012). *Computer Science Theory for the Information Age*, Chapter 8.13.2., p. 272ff, http://www.cs.cmu.edu/~venkatg/teaching/CStheory-infoage/hopcroft-kannan-feb2012.pdf.

Howland, P. and Park, H. (2008). Cluster preserving dimension reduction methods for document classification, *in* M. Berry and M. Castellanos (Eds), *Survey of Text Mining: Clustering, Classification, and Retrieval. Second Edition*, Springer, London, pp. 3–23.

Keller, H., Möllering, H., Schneider, T. and Yalame, H. (2021). Privacy-preserving clustering, *in* S.-L. Gazdag *et al.* (Eds), *Crypto Day Matters 32*, Gesellschaft für Informatik e.V./FG KRYPTO, Bonn.

Kleinberg, J. (2002). An impossibility theorem for clustering, *Proceedings of the 15th International Conference on Neural Information Processing Systems, Montreal, Canada*, pp. 446–453.

[4]The problem can be seen already based on the simple datasets mentioned in Examples 1 and 6. The required minimal distance between elements of distinct clusters is 9 and 8 times as big as the smallest intra-cluster distances. If the number of data elements in each cluster were increased q^2 times, the required gaps between clusters would have to be approximately q times as large. If the number of clusters were increased q^2 times, the required gaps between clusters would have to be approximately q times as large. Note, however, that large gaps are required also in other works on clusterability, e.g., the one by Ostrovsky *et al*. (2013).

Kłopotek, M.A. (2020). An aposteriorical clusterability criterion for k-means++ and simplicity of clustering, *SN Computer Science* **1**(2): 80.

Kłopotek, M.A. and Kłopotek, R.A. (2023). On the discrepancy between Kleinberg's clustering axioms and k-means clustering algorithm behavior, *Machine Learning* **112**(7): 2501–2553.

Kłopotek, M. and Kłopotek, R. (2022). Richness fallacy, *in* G. Manco and Z.W. Raś (Eds), *Foundations of Intelligent Systems*, Lecture Notes in Computer Science, Vol. 13515, Springer, Cham, pp. 262–271.

Kłopotek, R., Kłopotek, M. and Wierzchoń, S. (2020). A feasible k-means kernel trick under non-Euclidean feature space, *International Journal of Applied Mathematics and Computer Science* **30**(4): 703–715, DOI: 10.34768/amcs-2020-0052.

Larsen, K.G., Nelson, J., Nguyundefinedn, H.L. and Thorup, M. (2019). Heavy hitters via cluster-preserving clustering, *Communications of the ACM* **62**(8): 95–100.

Lingoes, J. (1971). Some boundary conditions for a monotone analysis of symmetric matrices, *Psychometrika* **36**: 195–203.

Lucińska, M. and Wierzchoń, S.T. (2018). Clustering based on eigenvectors of the adjacency matrix, *International Journal of Applied Mathematics and Computer Science* **28**(4): 771–786, DOI: 10.2478/amcs-2018-0059.

Madhulatha, T. (2012). An overview on clustering methods, *IOSR Journal of Engineering* **2**(4): 719–725.

Meilă, M. (2005). Comparing clusterings: An axiomatic view, *Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany*, pp. 577–584.

Ostrovsky, R., Rabani, Y., Schulman, L.J. and Swamy, C. (2013). The effectiveness of Lloyd-type methods for the k-means problem, *Journal of the ACM* **59**(6): 28:1–28:22.

Parameswaran, R. and Blough, D.M. (2005). A robust data-obfuscation approach for privacy preservation of clustered data, *Proceedings of the Workshop on Privacy and Security Aspects of Data Mining, Houston, USA*, p. 18–25.

Ramírez, D.H. and Auñón, J.M. (2020). Privacy preserving k-means clustering: A secure multi-party computation approach, `https://arxiv.org./abs/2009.10453`.

Roth, V., Laub, J., Kawanabe, M. and Buhmann, J. (2003). Optimal cluster preserving embedding of nonmetric proximity data, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(12): 1540–1551.

Sabo, K. (2014). Center-based l_1-clustering method, *International Journal of Applied Mathematics and Computer Science* **24**(1): 151–163, DOI: 10.2478/amcs-2014-0012.

Strazzeri, F. and Sánchez-García, R.J. (2021). Possibility results for graph clustering: A novel consistency axiom, `https://arxiv.org/abs/1806.06142`.

Suchy, D. and Siminski, K. (2023). GrDBSCAN: A granular density-based clustering algorithm, *International Journal of Applied Mathematics and Computer Science* **33**(2): 297–312, DOI: 10.34768/amcs-2023-0022.

van Laarhoven, T. and Marchiori, E. (2014). Axioms for graph clustering quality functions, *Journal of Machine Learning Research* **15**: 193–215.

Zhang, J., Zhu, K., Pei, Y., Fletcher, G. and Pechenizkiy, M. (2019). Cluster-preserving sampling from fully-dynamic streaming graphs, *Information Sciences* **482**: 279–300.

Zhao, Y., Tarus, S.K., Yang, L.T., Sun, J., Ge, Y. and Wang, J. (2020). Privacy-preserving clustering for big data in cyber-physical-social systems: Survey and perspectives, *Information Sciences* **515**: 132–155.

Mieczysław A. Kłopotek received his MSc and PhD degrees in computer science from the Dresden University of Technology. He then worked at the Semiconductor Research and Production Center in Warsaw. Thereafter he joined the Institute of Computer Science of the Polish Academy of Sciences, where he obtained his DSc in 1999. In 2009 he was granted the professorial title by the President of Poland. He has worked at IBM Warsaw on parallel statistical software for Netezza appliance, co-authoring five patents. He has also led a research group developing the first large-scale Polish semantic search engine. His current research encompasses data, text and web mining and machine learning.

Received: 18 August 2023
Revised: 3 November 2023
Accepted: 28 November 2023