STUDIA INFORMATICA
Nr 2 (27) Systems and information technology 2022

Mieczysław A. KŁOPOTEK[1]
ORCID: 0000-0003-4685-7045

[1] Institute of Computer Science,
Polish Academy of Sciences, Warsaw, Poland
mieczyslaw.klopotek@ipipan.waw.pl

K-means is probabilistically poor

DOI: 10.34739/si.2022.27.01

Abstract. Kleinberg introduced the concept of k-richness as a requirement for an algorithm to be a clustering algorithm. The most popular algorithm k means dos not fit this definition because of its probabilistic nature. Hence Ackerman et al. proposed the notion of probabilistic k-richness claiming without proof that k-means has this property. It is proven in this paper, by example, that the version of k-means with random initialization does not have the property probabilistic k-richness, just rebuking Ackeman's claim.

Keywords. k-means, clustering, probabilistic k-richness

1. Introduction

Kleinberg [8] proposed a new axiomatic system for clustering, initiating a long discussion on what kind of properties clustering algorithms should have and have not. In particular, he coined the term of k-richness of distance-based clustering algorithms, meaning the possibility to partition a set of objects into any k non-empty (disjoint) subsets via modifying the distances between these objects. However, there exist non-deterministic, probabilistic algorithms which do not fit this characterization because of non-deterministic behaviour. Therefore Ackerman at el [1, Definition 3 (k-Richness)] introduce the concept of probabilistic k-richness stating that

for any $\epsilon > 0$ and any predefined partition, a distance function can be found such that the clustering function returns this partition with probability at least $1 - \epsilon$.

They postulate in their Fig.2 (omitting the proof) that probabilistic k-richness in probabilistic sense is possessed by version of the k-means[1] algorithm with random initialization, which will be called here k-means-random, as well as by the version with k-means++ initialization.

The property of k-richness is a quite important one for in studying theoretical properties of clustering algorithms [6, 1, 2, 4, 5, 7, 12, 10, 14, 3, 11] in particular for constructing non-contradictory axiomatic systems. The existence of probabilistic k-richness of k means is assumed e.g. [15].

Kłopotek and Kłopotek [9, Theorem 1] have proven that k-means++ has in fact this property, while the issue is questionable in k-means-random case. They demonstrated that [9, Theorem 2].

Theorem 1. *In one-dimensional space, for $k \geq 3$, when distances between cluster centres exceed 6 times the largest enclosing radius r, k-means-random is not probabilistically k-rich.*

and also [9, Theorem 3].

Theorem 2. *For $k \geq mV_{ball\,,m,R}/V_{simplex\,,m,R-4r}$ where $V_{ball\,,m,R} = \frac{\pi^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}+1\right)} R^m$, $V_{simplex,\,m,R-4r} = \frac{\sqrt{n+1}}{n!\sqrt{2^n}}(R-4r)^m$, where m is the dimension of space, $S_m(R) = \frac{2\pi^{(m+1)/2}}{\Gamma\left(\frac{m+1}{2}\right)} R^m$, when distances between cluster centres exceed 10 times the largest enclosing radius r, and $R = 14r$, k-means-random is not probabilistically k-rich.*

Note that these theorems are not quite the denial of Ackerman's claim (the distances between clusters can be smaller), but from the rational point of view the fact that clusters with wide gaps between them cannot be detected, is quite disturbing.

In this paper, we show that also for small distances between clusters the Ackerman's et al. claim is not valid. The demonstration is based on a one-dimensional example.

We claim that:

---

[1] Various versions of k-means algorithm are described e.g. in [13].

Theorem 3. *k-means-random algorithm is not probabilistically k-rich for $k \geq 4$.*

Proof. The Theorem follows directly from Lemma 1 and Lemma 2 mentioned below. The logic is as follows: There is only a limited number of distinct initializations for a given dataset to be clustered. They are picked randomly according to some distribution, in case of k-means-random independent of the distances between objects. If, for each distance function between the objects, there exists at least one initialization for which the expected clustering cannot be found, then the error ϵ cannot take on any positive value (close to zero).

Lemma 1. *In a one-dimensional Euclidean space, given 8 points to cluster into 4 clusters each of two neighbouring data points, for each set of distances between data points, there exists a k-means-random initialization such that the desired clustering is not achieved.*

Proof. See Section 3.

Lemma 2. *In a one-dimensional Euclidean space, given $2k$ points to cluster into k clusters ($k > 4$) each of two neighbouring data points, for each set of distances between data points, there exists a k-means-random initialization such that the desired clustering is not achieved.*

Proof. See Section 4.

2. A Brief Introduction to *k*-means and Its Ricness Problem

Let us define the k-means-ideal algorithm as one that produces a clustering Γ_{opt} attaining the minimum of the cost function $Q(\Gamma)$.

$$Q(\Gamma) = \sum_{i=1}^{m} \sum_{j=1}^{k} u_{ij} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 = \sum_{j=1}^{k} \frac{1}{n_j} \sum_{\mathbf{x}_i, \mathbf{x}_l \in C_j} \|\mathbf{x}_i - \mathbf{x}_l\|^2 \quad (1)$$

where $\mathbf{X}$ is the clustered dataset, Γ is its partition into the predefined number k of non-empty clusters, and u_{ij} is an indicator of the membership of data point $\mathbf{x}_i$ in the cluster C_j having the centre at $\boldsymbol{\mu}_j$. As k-means-ideal is NP-hard, the following algorithm is used in practice:

1 Initialize k cluster centres $\boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_k^{(0)}$. Set $t := 0$.

2 Assign each data element $\mathbf{x}_i$ to the cluster $C_j^{(t)}$ identified by the closest $\boldsymbol{\mu}_j^{(t)}$

3 Update $t := t + 1$. Compute a new $\boldsymbol{\mu}_j^{(t)}$ of each cluster as the gravity centre of the data elements in $C_j^{(t-1)}$.

4 Repeat steps 2 and 3 until reaching a stop criterion (no change of cluster membership, or no sufficient improvement of the objective function, or exceeding some maximum number of iterations, or some other criterion).

If step 1 is performed as random uniform sampling from the set of data points (without replacement), then we will speak about k-means-random algorithm.

Kleinberg [8] introduced an axiomatic system for clustering functions, including the so-called richness axiom/property:

Property 1. *Let Range (f) denote the set of all partitions Γ such that $f(d) = \Gamma$ for some distance function d. If Range (f) is equal to the set of all partitions of $\boldsymbol{X}$, then f has the richness property.*

As Kleinberg's system is contradictory, and a number of attempts failed to produce a reasonable axiomatic system to which vast majority of clustering algorithms would adhere, publications like [1] talk about various "properties" that some clustering algorithms have and other do not, instead of talking about axioms.

In this paper we are interested in the above-mentioned richness axiom of Kleinberg and its variants.

k-means clustering algorithms does not possess this property, as it splits data in (exactly) k clusters. Nor other k-clustering methods do. Therefore, for the purpose of studying this and other so-called k-cluster algorithms, a modified property was proposed, called k-richness:

Property 2 (see e.g. Zadeh and Ben-David [14]). *If for any partition Γ of the set $\boldsymbol{X}$ consisting of exactly k (nonempty) clusters there exists such a distance function d that the clustering function $f(d)$ returns this partition Γ, then f has the k-richness property.*

Only k-means-ideal is k-rich, as shown in [9]. k-richness is problematic for randomized algorithms, like the k-means-random, as their output is not deterministic. Therefore Ackerman et al. [1, Definition 3 (k-Richness)] introduced the concept of *probabilistic k-richness*.

Property 3. *If for any partition Γ of the set $\boldsymbol{X}$ into exactly k clusters and every $\epsilon > 0$ there exists such a distance function d that the clustering function $f(d)$ returns this partition Γ with probability exceeding $1 - \epsilon$, then f has the probabilistic k-richness property.*

They postulated (omitting the proof) that probabilistic k-richness is possessed by k-means-random algorithm (see their Fig.2).

As this property is questionable, [9] introduced another concept, that of weak probabilistic k-richness.

Property 4. *A clustering method is said to have weak probabilistic k-richness property if there exists a function $pr\ (k) > 0 (k \in \mathbb{N})$ independent of the sample size and distance that for any partition Γ of the set **X** consisting of exactly k clusters, then there exists such a distance function d that the clustering function returns this partition Γ with probability exceeding $pr\ (k)$.*

$pr(k)$ is a minimum probability that the algorithm returns the required partition. It depends only on k and not on the structure of the clustered data set.

We discuss in this paper only the property of probabilistic k-richness.

3. Proof of Lemma 1

Let us investigate the case when $k = 4$.

The proof will consist in investigating relations between node distances and showing that under some special initial seeding (step 1 of k-means) there is no chance that a clustering of 8 nodes into 4 pairs can occur. We will consider the following mutually excluding cases: So consider a set of $n = 8$ nodes $n_1, \dots, n_8$ arranged in this order on a horizontal straight line from left to right with distances between them denoted as follows:

$d(n_1, n_2) = a_1, d(n_2, n_3) = p_{12}, d(n_3, n_4) = a_2, d(n_4, n_5) = p_{23}, d(n_5, n_6) = a_3, d(n_6, n_7) = p_{34}, d(n_7, n_8) = a_4$.

This is illustrated symbolically in the figure below.

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- 0 --p12-- 0 --a2-- 0 --p23-- 0 --a3-- 0 --p34-- 0 --a4-- 0]
```

The clustering, that we want to show is impossible under the selected seeding, is the following $\Gamma_0 = \{\{n_1, n_2\}, \{n_3, n_4\}, \{n_5, n_6\}, \{n_7, n_8\}\}$.

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- 0]--p12--[0 --a2-- 0]--p23--[0 --a3-- 0]--p34--[0 --a4-- 0]
```

By convention, the clusters are delimited with square brackets [] in the figures.

It is obvious that splitting a data set into 4 clusters of two elements can be performed only this way.

We will prove that whatever distances we take, there exists always the possibility of an initial seeding such that k-means-random will not find the clustering Γ_0 we want.

Note that if the clustering Γ should exist at all, the following must hold: $|a_1 - a_2| < 2p_{12}, |a_2 - a_3| < 2p_{23}, |a_3 - a_4| < 2p_{34}$,

because otherwise the clusters will take over elements of the neighboring ones. We will consider sharp inequalities only because k-means makes a random choice of tiers, hence the probability of a failure seeding is only reduced by a fixed factor and cannot got arbitrarily close to 0 in case of tiers.

So let us proceed case by case.

3.1. Case $a_1 < p_{12} < a_2$

Let us investigate the case when $k = 4 \text{AND} a_1 < p_{12} < a_2$. Let us choose the seeds (step 1 of k-means) $s_1 = n_2, s_2 = n_4, s_3 = n_5, s_4 = n_7$. After step 2 , the clusters will form: either

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- * --p12-- 0]--a2--[*]--p23--[*]--a3--[0 --p34-- * --a4-- 0]
```

or

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- * --p12-- 0]--a2--[*]--p23--[* --a3-- 0]--p34--[* --a4-- 0]
```

(the asterisks illustrate the seeds). A cluster $\{n_1, n_2, n_3\}$ will form around s_1 and the center of this cluster will eventually lie to the right of n_2. Hence the next cluster to the right of it will have no possibility to gain control over n_3 because it is closer to n_2 than to n_4. Hence the relation $a_1 < p_{12} < a_2$ under appropriate seeding prohibits emerging of Γ_0, the thesis of the Lemma holds in this case.

By symmetry, it holds also for $a_4 < p_{34} < a_3$.

3.2. Case $a_1 > p_{12} > a_2$

Let us investigate the case when $k = 4$ AND $a_1 > p_{12} > a_2$. Assume the following seeding: $s_1 = n_1, s_2 = n_3, s_3 = n_5, s_4 = n_7$. After step 2, one of the following clusterings will emerge:

```
n1      n2      n3      n4      n5      n6      n7      n8
[*]--a1--[0 --p12-- *]--a2--[0 --p23-- * --a3-- 0]--p34--[* --a4-- 0]
```

```
n1      n2      n3      n4      n5      n6      n7      n8
[*]--a1--[0 --p12-- *]--a2--[0 --p23-- *]--a3--[0 --p34-- * --a4-- 0]
```

```
n1      n2      n3      n4      n5      n6      n7      n8
[*]--a1--[0 --p12-- * --a2-- 0]--p23--[* --a3-- 0]--p34--[* --a4-- 0]
```

```
n1      n2      n3      n4      n5      n6      n7      n8
[*]--a1--[0 --p12-- * --a2-- 0]--p23--[*]--a3--[0 --p34-- * --a4-- 0]
```

As visible, the following clusters will form: The first cluster $\{n_1\}$, the second containing at least n_2, n_3 and at largest extent also n_4, the third at least n_5 and the forth at least n_7, n_8.

During subsequent iteration the following occurs: The forth cluster keeps n_7, n_8 forever. Therefore the third cluster center will be either in the middle of $[n_5, n_6]$ or to the left of it and it will be so as long as the second cluster does not get control over n_5. The second cluster center lies to the left of n_3. Therefore the first cluster does not get control over n_2. Note that the distance of the center of the third cluster to n_5 is less than $a_3/2$, and that of the second cluster more than $a_2 + p_{23}$. Therefore in the next step the second cluster will not get n_5 and so its distance will remain above $a_2 + p_{23}$ and it will not change as long as it does not get control over n_5, but it cannot and so this will stay forever so. Under these circumstances the distance of the second cluster center to n_2 will be smaller than that of the first and so it will stay forever.

Therefore a cluster $\{n_1, n_2\} \in \Gamma_0$ cannot form. The thesis of the Lemma holds in this case.

By symmetry same applies to $a_4 > p_{34} > a_3$.

We need to check $a_1 > p_{12} < a_2$ and $a_1 < p_{12} > a_2$

3.3. Case $a_1 > p_{12} < a_2$

Let us investigate the case when $k = 4$ AND $a_1 > p_{12} < a_2$.

3.3.1. Case: $a_1 < (2p_{12} + a_2)/3$

Let us investigate the case when $k = 4 \text{AND} a_1 > p_{12} < a_2$ AND $a_1 < (2p_{12} + a_2)/3$. Consider the seeding $s_1 = n_2, s_2 = n_4, s_3 = n_5, s_4 = n_7$. In step 2 one of the clusterings will occur.

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- * --p12-- 0]--a2--[*]--p23--[* --a3-- 0]--p34--[* --a4-- 0]
```

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- * --p12-- 0]--a2--[*]--p23--[*]--a3--[0 --p34-- * --a4-- 0]
```

The first cluster will consist of n_1, n_2, n_3 and the second only of n_4. In order for the second cluster to gain control over n_3, the following condition needs to hold $a_2 < (2p_{12} + a_1)/3$ because otherwise the second cluster will never get n_3 (as its center will be at n_4 or to the right of it). But $a_2 < (2p_{12} + a_1)/3 < 3a_1/3 = a_1$ implying $a_2 < a_1$. This contradicts our assumption that $a_1 < (2p_{12} + a_2)/3$ and that $p_{12} < a_2$ because if we insert the second into the first we get: $a_1 < (2p_{12} + a_2)/3 < (2a_2 + a_2)/3 = a_2$ that is $a_1 < a_2$. The thesis of the Lemma holds in this case.

3.3.2. Case: $a_1 > (2p_{12} + a_2)/3$

Let us investigate the case when $k = 4 \text{AND} a_1 > p_{12} < a_2 \text{AND} a_1 > (2p_{12} + a_2)/3$. Assume the following seeding: $s_1 = n_1, s_2 = n_3, s_3 = n_5, s_4 = n_7$. Then the following clusters may occur in step 2 of k-means:

```
n1      n2      n3      n4      n5      n6      n7      n8
[*]--a1--[0 --p12-- *]--a2--[0 --p23-- *]--a3--[0 --p34-- * --a4-- 0]
```

```
n1      n2      n3      n4      n5      n6      n7      n8
[*]--a1--[0 --p12-- *]--a2--[0 --p23-- * --a3-- 0]--p34--[* --a4-- 0]
```

```
n1      n2      n3      n4      n5      n6      n7      n8
[*]--a1--[0 --p12-- * --a2-- 0]--p23--[*]--a3--[0 --p34-- * --a4-- 0]
```

```
n1      n2      n3      n4      n5      n6      n7      n8
[*]--a1--[0 --p12-- * --a2-- 0]--p23--[* --a3-- 0]--p34--[* --a4-- 0]
```

Clusters 3 and 4 will form out of at least nodes $n_5, \dots, n_8$ The forth cluster keeps n_7, n_8 forever. Therefore the third cluster center will be either in the middle of $[n_5, n_6]$ or to the left of it and it will be so as long as the second cluster does not get control over n_5.

So after centroid update, the second cluster center lies to the left of the middle of $[n_3, n_4]$. Note that the distance of the center of the third cluster to n_5 is less than $a_3/2$, and that of the second cluster more than $a_2/2 + p_{23}$. Therefore in the next step the second cluster will not get n_5 and so its distance will remain above $a_2/2 + p_{23}$ and it will not change as long as it does not get control over n_5, but it cannot and so this will stay forever so.

The first cluster can capture n_2 in the first step only if $a_1 < (2p_{12} + a_2)/3$. But we assumed the contrary, that is that $a_1 > (2p_{12} + a_2)/3$. So it will never capture it. The thesis of the Lemma holds in this case.

Therefore, combined with the previous case, thesis of the Lemma holds for $a_1 > p_{12} < a_2$ altogether. By symmetry, it holds for $a_4 > p_{34} < a_3$ too.

3.4. Case $a_1 < p_{12} > a_2$

Let us investigate the case when $k = 4$ AND $a_1 < p_{12} > a_2$. By symmetry, also $a_4 < p_{34} > a_3$, because all the other relations of these distances were already discussed and the Lemma held for them.

3.4.1. Case: $a_2 < p_{23} < a_3$

Let us investigate the case when $k = 4\text{AND}a_1 < p_{12} > a_2$ AND $a_4 < p_{34} > a_3$ AND $a_2 < p_{23} < a_3$. Let us look at the seeding $s_1 = n_2, s_2 = n_4, s_3 = n_6, s_4 = n_7$. One of the following clusterings will emerge in step 2.

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- *]--p12--[0 --a2-- * --p23-- 0]--a3--[*]--p34--[* --a4-- 0]
```

Cluster 1 gets $\{n_1, n_2\}$, and cluster 2 gets $\{n_3, n_4, n_5\}$ or Cluster 1 gets $\{n_1, n_2, n_3\}$, and cluster 2 gets $\{n_4, n_5\}$ In subsequent steps Cluster 1 keeps $\{n_1, n_2\}$, but cluster 2 may loose n_3 to cluster 1 . Hence the distance of the second cluster center to n_5 will be equal or smaller than $(2p_{23} + a_2)/3 < p_{23}$ hence the third cluster will never gain control over n_5 as its distance is at least a_3. The thesis of the Lemma holds in this case.

3.4.2. Case: $a_2 > p_{23} > a_3$

Let us investigate the case when $k = 4\text{AND}a_1 < p_{12} > a_2\text{AND}a_2 > p_{23} > a_3$. By a symmetric argument, The thesis of the Lemma holds in the case $a_2 > p_{23} > a_3$

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- *]--p12--[*]--a2--[0 --p23-- * --a3-- 0]--p34--[* --a4-- 0]
```

3.4.3. Case: $a_2 > p_{23} < a_3$

Let us investigate the case when $k = 4\text{AND}a_1 < p_{12} > a_2\text{AND}a_4 < p_{34} > a_3$ AND $a_2 > p_{23} < a_3$.

3.4.3.1. Case: $a_2 > (2p_{23} + a_3)/3$

Let us investigate the case when $k = 4\text{AND}a_1 < p_{12} > a_2\text{AND}a_4 < p_{34} > a_3$ AND $a_2 > p_{23} < a_3$ AND $a_2 > (2p_{23} + a_3)/3$. Under the seeding $s_1 = n_2, s_2 = n_3, s_3 = n_5, s_4 = n_7$

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- *]--p12--[*]--a2--[0 --p23-- * --a3-- 0]--p34--[* --a4-- 0]
```

The first cluster will form of n_1, n_2, the forth of n_7, n_8. They will never loose control over these nodes. The second cluster will not capture n_4, because its distance to it amounts to a_2, and the distance of the third cluster center to it amounts to at most $(2p_{23} + a_3)/3$ which is smaller than a_2 by the assumption.

The thesis of the Lemma holds.

3.4.3.2. Case: $a_2 < (2p_{23} + a_3)/3$

Let us investigate the case when $k = 4 \text{AND} a_1 < p_{12} > a_2 \text{AND} a_4 < p_{34} > a_3$ AND$a_2 > p_{23} < a_3$ AND $a_2 < (2p_{23} + a_3)/3$.

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- *]--p12--[0 --a2-- *]--p23--[0 --a3-- *]--p34--[* --a4-- 0]
```

Under the seeding $s_1 = n_2, s_2 = n_4, s_3 = n_6, s_4 = n_7$, the first cluster will form of n_1, n_2, the forth of n_7, n_8. They will never loose control over these nodes. The third cluster will capture n_5 only if $a_3 < (2p_{23} + a_2)/3 < a_2$. But we assumed $a_2 < (2p_{23} + a_3)/3$ which implies that $a_2 < (2p_{23} + a_3)/3 < a_3$. These two requirements are contradictory. The thesis of the Lemma holds. Combinwed with the former case, thesis of the Lemma holds already when $a_2 > p_{23} < a_3$.

3.4.4. Case: $a_2 < p_{23} > a_3$

Let us investigate the case when $k = 4 \text{AND} a_1 < p_{12} > a_2$ AND $a_4 < p_{34} > a_3$ AND $a_2 < p_{23} > a_3$. This case means that, informally speaking, the gap between clusters has to be bigger than each the distance within the cluster neighbouring with the gap.

Consider the seeding S1: $s_1 = n_2, s_2 = n_5, s_3 = n_7, s_4 = n_8$. One of the following clustering may emerge in Step 2.

```
n1       n2       n3       n4       n5       n6       n7       n8
[0 --a1-- * --p12-- 0 --a2-- 0]--p23--[* --a3-- 0]--p34--[*]--a4--[*]
```

```
n1       n2       n3       n4       n5       n6       n7       n8
[0 --a1-- * --p12-- 0]--a2--[0 --p23-- * --a3-- 0]--p34--[*]--a4--[*]
```

```
n1       n2       n3       n4       n5       n6       n7       n8
[0 --a1-- *]--p12--[0 --a2-- 0 --p23-- * --a3-- 0]--p34--[*]--a4--[*]
```

The first cluster captures for sure n_1, n_2 and never looses them. If the first cluster would capture $n_1, \dots, n_4$, it would not lose it in further iteration. Thesis holds. So we are left with the clusterings $\Gamma_1 = \{\{n_1, n_2, n_3\}, \{n_4, n_5, n_6\}, \{n_7\}, \{n_8\}\}$,
$\Gamma_3 = \{\{n_1, n_2\}, \{n_3, n_4, n_5, n_6\}, \{n_7\}, \{n_8\}\}$.

By symmetric seeding S2 $s_1 = n_1, s_2 = n_2, s_3 = n_4, s_4 = n_7$. we need to consider only $\Gamma_2 = \{\{n_1\}, \{n_2\}, \{n_3, n_4, n_5\}, \{n_6, n_7, n_8\}\}$. $\Gamma_4 = \{\{n_1\}, \{n_2\}, \{n_3, n_4, n_5, n_6\}, \{n_7, n_8\}\}$ after the initialization step.

```
n1       n2       n3       n4       n5       n6       n7       n8
[*]--a1--[*]--p12--[0 --a2-- * --p23-- 0]--a3--[0 --p34-- * --a4-- 0]
```

```
n1       n2       n3       n4       n5       n6       n7       n8
[*]--a1--[*]--p12--[0 --a2-- * --p23-- 0 --a3-- 0]--p34--[* --a4-- 0]
```

If after initialization with S2 we would obtain clustering Γ_4, then this means that $p_{34} > p_{23} + a_3$. On the other hand, it is obvious that if under seeding S1 we obtain any of the clusterings Γ_1 or Γ_3, the second cluster will never get node n_2, therefore its center will reside to the right of n_4, therefore the third cluster would never capture n_6. So the thesis of Lemma holds in this case. So we do not need to consider Γ_4 any more.

By symmetry, also if Γ_3 occurs under the seeding S1, the Lemma holds.

So we need to consider S1 leading ton Γ_1 and S2 leading to Γ_2.

In case of Γ_1 after S1, in the step 3 can relocate in such a way that step 2 in the next iteration cluster 1 can take over n_4.

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- 0 --p12-- 0 --a2-- 0]--p23--[0 --a3-- 0]--p34--[*]--a4--[*]
```

In this case cluster 2 will never regain n_4. The Thesis holds.

The other possibility is that instead in the step 2 of the next iteration either cluster 2 takes over n_3 or n_3 remains in cluster 1 .

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- 0 --p12-- 0]--a2--[0 --p23-- 0 --a3-- 0]--p34--[*]--a4--[*]
```

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- 0]--p12--[0 --a2-- 0 --p23-- 0 --a3-- 0]--p34--[*]--a4--[*]
```

By analogy in case of Γ_2 after S2 in the next iteration in step 2 we need to consider only either cluster 3 takes over n_6 or n_6 remains in cluster 4.

```
n1       n2        n3       n4        n5       n6        n7       n8
[*]--a1--[*]--p12--[0 --a2-- 0 --p23-- 0]--a3--[0 --p34-- 0 --a4-- 0]
```

```
n1       n2        n3       n4        n5       n6        n7       n8
[*]--a1--[*]--p12--[0 --a2-- 0 --p23-- 0 --a3-- 0]--p34--[0 --a4-- 0]
```

3.4.4.1. Case: under Γ_2 cluster 3 takes over node n_6 and at the same time under Γ_1 cluster 2 takes over node n_3

Let us investigate the case when $k = 4 \text{AND} a_1 < p_{12} > a_2 \text{AND} a_4 < p_{34} > a_3$ AND $a_2 < p_{23} > a_3$ AND under Γ_2 cluster 3 takes over node n_6 and at the same time under Γ_1 cluster 2 takes over node n_3. So consider the situation that under Γ_2 cluster 3 takes over node n_6

```
n1       n2        n3       n4        n5       n6        n7       n8
[*]--a1--[*]--p12--[0 --a2-- 0 --p23-- 0]--a3--[0 --p34-- 0 --a4-- 0]
```

```
n1       n2        n3       n4        n5       n6        n7       n8
[*]--a1--[*]--p12--[0 --a2-- 0 --p23-- 0 --a3-- 0]--p34--[0 --a4-- 0]
```

This implies

$$(2p_{23} + a_2)/3 + a_3 < (2p_{34} + a_4)/3$$
$$2p_{23} + a_2 + 3a_3 < (2p_{34} + a_4)$$
$$(2p_{23} + a_2 + 3a_3)/4 < (2p_{34} + a_4)/4 < 3p_{34}/4 < p_{34}$$

At the same time under Γ_1 let cluster 2 take over node n_3.

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- 0 --p12-- 0]--a2--[0 --p23-- 0 --a3-- 0]--p34--[*]--a4--[*]
```

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- 0]--p12--[0 --a2-- 0 --p23-- 0 --a3-- 0]--p34--[*]--a4--[*]
```

This implies:

$$(2p_{23} + a_3)/3 + a_2 < (2p_{12} + a_1)/3$$
$$(2p_{23} + a_3 + 3a_2)/4 < p_{12}$$

The conditions $(2p_{23} + a_3 + 3a_2)/4 < p_{12}$ and $(2p_{23} + a_2 + 3a_3)/4 < p_{34}$ mean that the cluster consisting of nodes n_3, n_4, n_5, n_6 will never lose any of its members to the clusters on either of its sides. The thesis holds.

3.4.4.2. Case: that under Γ_2 cluster 3 takes over node n_6 and at the same time under Γ_1 clusters 1 and 2 are not taking over mutually their elements

Let us investigate the case when $k = 4 \text{AND} a_1 < p_{12} > a_2 \text{AND} a_4 < p_{34} > a_3$ AND $a_2 < p_{23} > a_3$ AND that under Γ_2 cluster 3 takes over node n_6 and at the same time under Γ_1 clusters 1 and 2 are not taking over mutually their elements.

So consider the situation that under Γ_2 cluster 3 takes over node n_6

```
n1       n2        n3       n4        n5       n6        n7       n8
[*]--a1--[*]--p12--[0 --a2-- 0 --p23-- 0]--a3--[0 --p34-- 0 --a4-- 0]
```

```
n1       n2        n3       n4        n5       n6        n7       n8
[*]--a1--[*]--p12--[0 --a2-- 0 --p23-- 0 --a3-- 0]--p34--[0 --a4-- 0]
```

This implies

$$(2p_{23} + a_2)/3 + a_3 < (2p_{34} + a_4)/3$$
$$(2p_{23} + a_2 + 3a_3)/4 < (2p_{34} + a_4)/4$$

At the same time assume that under Γ_1 clusters 1 and 2 are not taking over mutually their elements.

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- 0 --p12-- 0]--a2--[0 --p23-- 0 --a3-- 0]--p34--[*]--a4--[*]
```

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- 0 --p12-- 0]--a2--[0 --p23-- 0 --a3-- 0]--p34--[*]--a4--[*]
```

But this means that

$$(2p_{23} + a_2 + 3a_3)/4 > (p_{23} + 2a_3)/3$$

Taking into account that $(2p_{34} + a_4)/4 < p_{34}$, we obtain from the above equations that therefore $(p_{23} + 2a_3)/3 < p_{34}$. This means that under Γ_1 the cluster 3 cannot take over n_6 so that the clustering Γ_1 remain stable. The thesis holds.

By symmetry the situation that under Γ_1 cluster 2 takes over node n_3 and at the same time under Γ_2 clusters 3 and 4 are not taking over mutually their elements supports the thesis also.

3.4.4.3. Case: under Γ_1 clusters 1 and 2 are not taking over mutually their elements and at the same time under Γ_2 clusters 3 and 4 are not taking over mutually their elements

Let us investigate the case when $k = 4 \text{AND} a_1 < p_{12} > a_2 \text{AND} a_4 < p_{34} > a_3$ AND $a_2 < p_{23} > a_3$ AND under Γ_1 clusters 1 and 2 are not taking over mutually their elements and at the same time under Γ_2 clusters 3 and 4 are not taking over mutually their elements. Consider the following seeding S7: $s_1 = n_4, s_2 = n_6, s_3 = n_7, s_4 = n_8$

```
n1       n2        n3       n4        n5       n6        n7       n8
[0 --a1-- 0 --p12-- 0 --a2-- *]--p23--[0 --a3-- *]--p34--[*]--a4--[*]
```

We get the clustering $\Gamma_7 = \{\{n_1, n_2, n_3, n_4\}, \{n_5, n_6\}, \{n_7\}, \{n_8\}\}$. We need to prevent the first cluster to keep these initial elements, therefore the following must hold:

$$(a_1 + 2p_{12} + 3a_2)/4 > p_{23} + a_3/2$$

hence

$$(3p_{12} + 3a_2)/4 > p_{23} + a_3/2$$
$$3/4p_{12} + 1/4a_2 > p_{23} + a_3/2 - a_2/2$$

and by analogy under the seeding: $s_1 = n_1, s_2 = n_2, s_3 = n_3, s_4 = n_5$

```
n1      n2      n3      n4      n5      n6      n7      n8
[*]--a1--[*]--p12--[* --a2-- 0]--p23--[* --a3-- 0 --p34-- 0 --a4-- 0]
```

implying the clustering $\Gamma_4 = \{\{n_1\}, \{n_2\}, \{n_3, n_4\}, \{n_5, n_6, n_7, n_8\}\}$.

$$3/4p_{34} + 1/4a_3 > p_{23} + a_2/2 - a_3/2$$

Either $a_2/2 - a_3/2 \geq 0$ or $a_3/2 - a_2/2 \geq 0$. Assume the latter without restraining the generality Hence

$$3/4p_{12} + 1/4a_2 > p_{23}$$
$$p_{12} > p_{23}$$

Let us consider the clustering Γ_2. In order for the cluster 2 to capture node n_3 the following needs to hold:

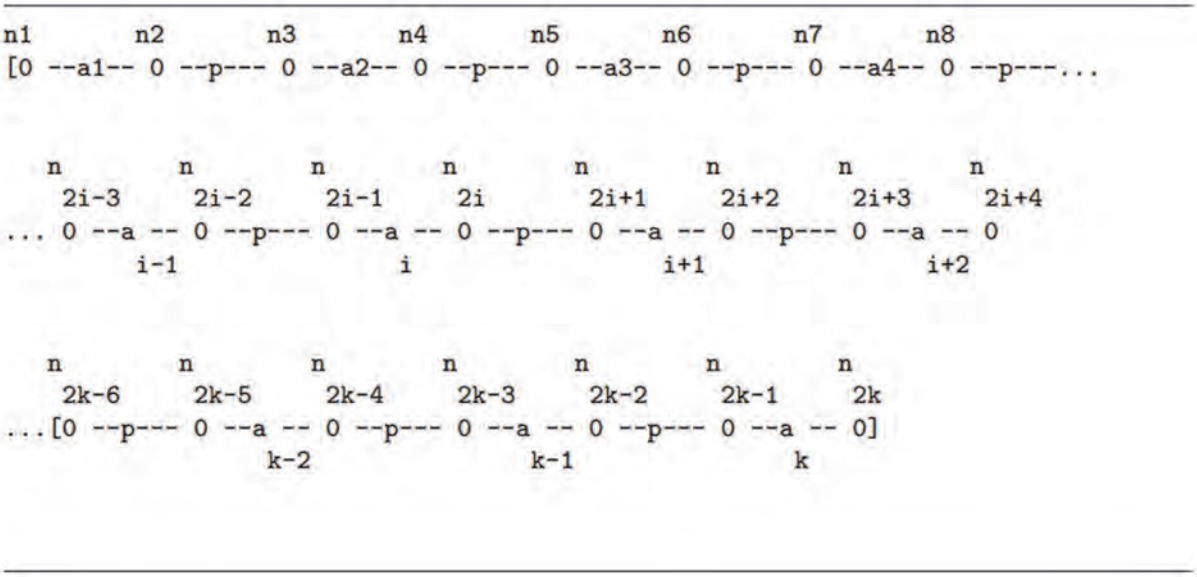
$$p_{12} < (2a_2 + p_{23})/3 < 3p_{23}/3 = p_{23}$$

which contradicts the previously derived condition $p_{12} > p_{23}$. This case supports the thesis either.

Hence the violation of k-richness in the probabilistic sense for $k = 4$ is proven.

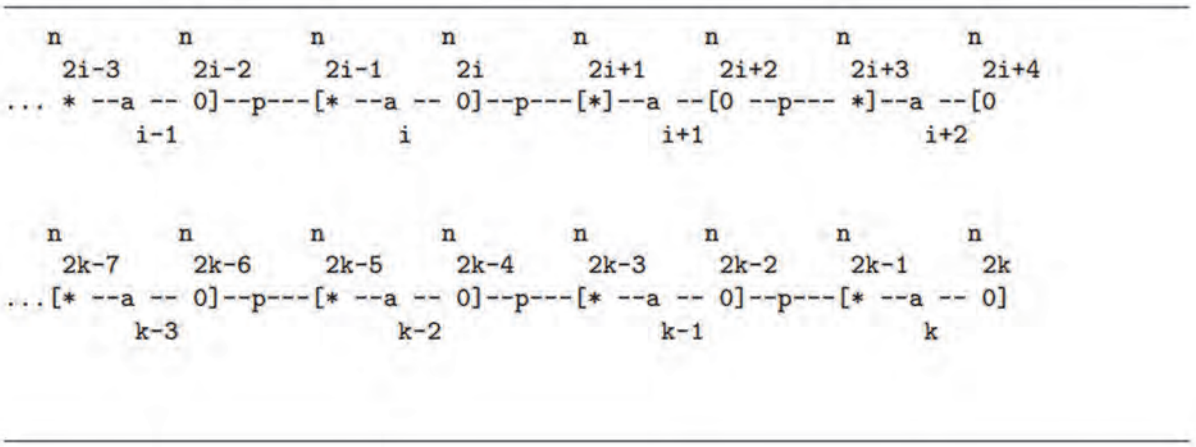
4. Proof of Lemma 2

Let us investigate the case when $k > 4$. For k greater than 4 , consider clustering into k two element clusters. The nodes shall be denoted as above $n_1, \ldots, n_{2k}$, the distance between elements of cluster j shall be a_j and the distance between cluster j and $j + 1$ shall be denoted by $p_{j,j+1}$. If the clustering should exist at all, the following must hold: $|a_j - a_{j+1}| < 2p_{j,j+1}$.



we will refrain from showing indexes of p in the figures as they are selfevident.

Let us look at the situation when $p_{i-1,i} > a_i$. Consider a seeding such that for $j >= is_j = n_{2j-1}$ for some i.



This ensures that under no step of k-means the j^{th} cluster ($j \geq i$ will contain node n_{2j+1}. This can be shown by induction. Directly after seeding, in step 2 of k-means, the cluster k will contain at least nodes n_{2k-1} and n_{2k} and maybe n_{2k-2}, but it cannot contain the node n_{2k+1} as there is no such node. The cluster $j, i <= j < k$ contains at least the node n_{2j-1}, and maybe n_{2j} if not contained in the next cluster, maybe n_{2j-2} if not contained in the previous cluster. It does not contain n_{2j+1} because there is the next seed.

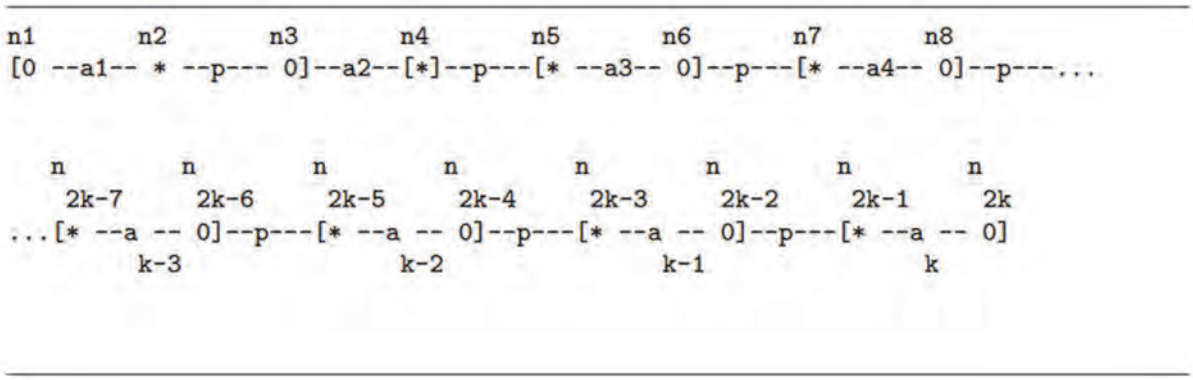
Now consider what happens when cluster centres of clusters emerged this way are computed (step 3 of k-means). Define the vector $v_{j,1}$ as one from the centre of cluster j to n_{2j+1}. It

amounts to at least $\lceil a_j/2 + p_{j,j+1} \rceil$ at this moment, for $j > 4$. Define the vector $v_{j+1,2}$ as one from n_{2j+1} to the centre of cluster $j+1$. It amounts after the initial step to at most $\lceil a_{j+1}/2 \rceil (4 < j < k)$. Therefore cluster j cannot expand in the next step to capture n_{2j+1}, because $|a_j - a_{j+1}| < 2p_{j,j+1}$ implies $-a_j + a_{j+1} < 2p_{j,j+1}$, that is $a_{j+1}/2 < p_{j,j+1} + a_j/2$. Therefore the vector $v_{j,1}$ will not decrease and $v_{j,2}$ will not increase because $v_{j,1} + v_{j,2}$ is constant for j (distance between n_{2j+1} and n_{2j-1}) - this is shown by induction on $j = k-1, k-2, \ldots, i$ under the condition that cluster $i-1$ would not capture n_{2i-1}. Cluster $i-1$ will not capture n_{2i-1}, because the cluster i cannot capture n_{2i+1}, therefore its center will lie to the left of n_{2i}, therefore its distance to n_{2i-1} will amount to a_i at most, while the distance of cluster $i-1$ center will be at distance of at least $p_{i-1,i}$ from n_{2i-1}, and by assumption $p_{i-1,i} > a_i$. We will exploit this partial seeding below.

4.1. Case $a_1 < p_{12} < a_2$

Let us investigate the case when $k > 4$ AND $a_1 < p_{12} < a_2$.

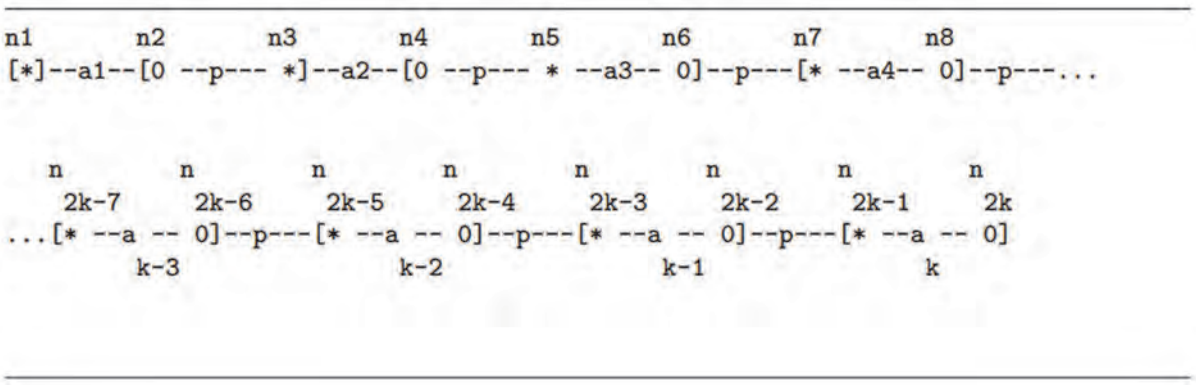
Let us choose the seeds $s_1 = n_2, s_2 = n_4$ and $s_3, s_4, \ldots, s_k$ at nodes n_{2j-1} (that is to the right of s_2.)



A cluster $\{n_1, n_2, n_3\}$ will form around s_1 and the center of this cluster will eventually lie to the right of n_2. Hence the next cluster to the right of it will have no possibility to gain control over n_3 because it is closer to n_2 than to n_4. Hence the relation $a_1 < p_{12} < a_2$ supports the thesis.

4.2. Case $a_1 > p_{12} > a_2$

Let us investigate the case when $k > 4 \text{AND} a_1 > p_{12} > a_2$. Assume the following seeding: $s_1 = n_1, , s_j = n_{2j-1}$ for $j = 2, \ldots, k$.

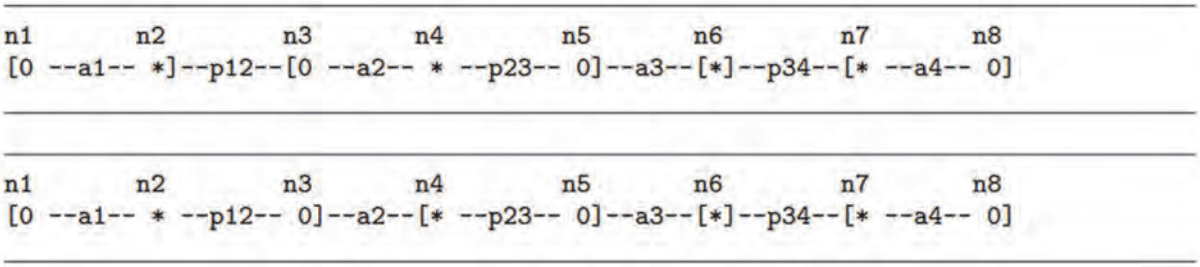


This is the case discussed above in the introduction where $i = 2$. Obviously, cluster 1 cannot take over n_2, even if second cluster gets n_4 because the second cluster would not get n_5. Therefore a cluster $\{n_1, n_2\} \in \Gamma$ cannot form.

Therefore the relation $a_1 > p_{12} > a_2$ also supports the thesis.

4.3. Case $a_m < p_{m,m+1} < a_{m+1}$ for some m

Let us investigate the case when $k > 4$ AND $a_m < p_{m,m+1} < a_{m+1}$ for some m. Let us now discuss the case $a_m < p_{m,m+1} < a_{m+1}$. Let us look at the seeding $s_j = n_{2j}$ for $j = 1, \dots, m-1, s_m = n_{2m}, s_{m+1} = n_{2m+2}, s_j = n_{2(j)-1}$ for $j = m+2, \dots, k$.



Clusters $1, \dots, m$ resemble clusters $k, \dots, i$ from the above case BB (but in reverse order) what ensures that the cluster $m+1$ will never get the node n_{2m+1}. Therefore this case has to be rejected. So either $a_m > p_{m,m+1} < a_{m+1}$ for each m or $a_m < p_{m,m+1} > a_{m+1}$ for each m or

4.4. Case $a_m > p_{m,m+1} < a_{m+1}$ for each m

Let us investigate the case when $k > 4$ AND $a_m > p_{m,m+1} < a_{m+1}$ for each m. Let a_i be the longest. Make a seeding $s_j = n_{2j-1}$ for $j = 1, \dots, i$. $s_j = n_{2j-2}$ for $j = i+1, k$. Initially clusters will form: $\{n_1\}, \{n_{2j-2}, n_{2j-1}\}$ for $j = 2, \dots, k-1$. and $\{n_{2k-2}, n_{2k-1}, n_{2k}\}$. No cluster

j will ever take over node n_{2j+1}, as previously stated because $a_1 > p_{12}$. The question is if it can take over n_{2j}. Considers clusters i and $i+1$. with nodes $\{n_{2i-2}, n_{2i-1}\}$ and $\{n_{2i}, n_{2i+1}\}$ resp. initially. Clusters $1,..i$ are stable until cluster $i+1$ changes. In the extreme case the cluster $i+1$ can take over n_{2i+2}. In this case the distance from the cluster center to n_{2i} amounts to $\left(2p_{i,i+1} + a_{i+1}\right)/3)a_{i+1}$ which means that cluster i will not get the node n_{2i} so that the required clustering cannot be formed. So this case needs to be rejected.

4.5. Case $a_m < p_{m,m+1} > a_{m+1}$ for each m

Let us investigate the case when $k > 4$ AND $a_m < p_{m,m+1} > a_{m+1}$ for each m. Consider an $i = 5$. As $p_{45} > a_5 > a_5/2$, the 4 th cluster will never acquire n_9. So it is only possible for cluster 5 to acquire n_8 or nodes with lower indexes. If this happens, the probabilistic k-richness definition is violated. If it does not acquire it at any point in time, then k-richness definition is violated due to conditions described in the case $k = 4$ above. If cluster 5 acquires n_8, then the argument there can repeated under the condition that a_4 is close to zero.

This completes the proof.

5. Concluding Remarks

We have demonstrated in this paper, that contrary to claims of Ackerman et al. [1], the k-means-random is not probabilistically k-rich.

Missing probabilistic k-richness of k-means-random means that no matter how the distances between clusters are, there exists an upper limit on probability that the true cluster structure in the data will be detected.

In order to characterize the k-richness properties of the k-means-random algorithm, the concept of weak probabilistic k-richness needs to be introduced, as done in [9]. The interested reader is advised to study that paper to see that there exists also a lower limit on probability that the true cluster structure in the data will be detected.

References

1. Ackerman M., Ben-David S., and Loker D. Towards property-based classification of clustering paradigms. In J.D. Lafferty, C.K.I. Williams, J. ShaweTaylor, R.S. Zemel, and

A. Culotta, editors, Advances in Neural Information Processing Systems 23, pages 10-18. Curran Associates, Inc., 2010.

2. Ackerman M., Ben-David S., Loker D., and Sabato S. Clustering oligarchies. In Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013, pages 66-74, 2013.

3. Bandyopadhyay S. and Murty M. N. Axioms to characterize efficient incremental clustering. In Proceedings of the 23rd International Conference on Pattern Recognition, pages 450-455. IEEE, 2016.

4. Ben-David S. Attempts to axiomatize clustering, 2005. NIPS Workshop, December 2005.

5. Ben-David S. and Ackerman M. Measures of clustering quality: A working set of axioms for clustering. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, Advances in Neural Information Processing Systems 21, pages 121-128. Curran Associates, Inc., 2009.

6. Correa-Morris J. An indication of unification for different clustering approaches. Pattern Recognition, 46(9):2548-2561, 2013.

7. Hopcroft J. and Kannan R. Computer science theory for the information age, 2012. chapter 8.13.2. A Satisfiable Set of Axioms. page 272ff.

8. Kleinberg J. An impossibility theorem for clustering. In Proc. NIPS 2002, pages 446-453, 2002. http://books.nips.cc/papers/files/nips15/LT17.pdf.

9. Kłopotek R. A. and Kłopotek M. A. On probabilistic krichness of the k-means algorithms. In Giuseppe Nicosia, Panos M. Pardalos, Renato Umeton, Giovanni Giuffrida, and Vincenzo Sciacca, editors, Machine Learning, Optimization, and Data Science - 5th International Conference, LOD 2019, Siena, Italy, September 10-13, 2019, Proceedings, volume 11943 of Lecture Notes in Computer Science, pages 259-271. Springer, 2019.

10. Meilă M. Comparing clusterings: An axiomatic view. In Proceedings of the 22Nd International Conference on Machine Learning, ICML '05, pages 577-584, New York, NY, USA, 2005. ACM.

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

12. Laarhoven T. and Marchiori E. Axioms for graph clustering quality functions. Journal of Machine Learning Research, 15:193-215, 2014.

13. Wierzchoń S.T. and Kłopotek M. A. Modern Clustering Algorithms. Springer Verlag Series: Studies in Big Data 34. Springer, 2018.

14. Zadeh R. B. and Ben-David S. A uniqueness theorem for clustering. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, pages 639-646, Arlington, Virginia, United States, 2009. AUAI Press.

15. Mohammad Z. A. A PAC-Theory of Clustering with Advice. PhD thesis, University of Waterloo, 2018.