Journal of Classification 33:118-140 (2016)
DOI: 10.1007/s00357-016-9198-2

On the Properties of α-Unchaining Single Linkage Hierarchical Clustering

Alvaro Martínez-Pérez

University of Castilla-La Mancha, Spain

Abstract: In the election of a hierarchical clustering method, theoretic properties may give some insight to determine which method is the most suitable to treat a clustering problem. Herein, we study some basic properties of two hierarchical clustering methods: α-unchaining single linkage or $SL(\alpha)$ and a modified version of this one, $SL^*(\alpha)$. We compare the results with the properties satisfied by the classical linkage-based hierarchical clustering methods.

Keywords: Hierarchical clustering; Single linkage; Chaining effect; Weakly unchaining; α-bridge-unchaining.

1. Introduction

Kleinberg (2002) studied the problem of clustering in an axiomatic way. He proposed a few basic properties that any clustering function F should satisfy. Let X be any finite set and d any distance function. Let $\mathcal{P}(X)$ denote the set of all possible partitions of X. Fix a clustering method F so that $F(X, d) = \Pi \in \mathcal{P}(X)$. The properties proposed by Kleinberg were:

The author was partially supported by MTM 2012-30719. The author would like to express his gratitude to the anonymous referees whose helpful comments greatly improved the presentation and readability of this paper.

Corresponding Author's Address: A. Martínez-Pérez, Departamento de Análisis Económico y Finanzas, Universidad de Castilla-La Mancha, Talavera de la Reina 45600 Spain, email: alvaro.martinezperez@uclm.es.

Published online: 4 April 2016

- Scale invariance: For all $\alpha > 0$, $F(X, \alpha \cdot d) = \Pi$
- Richness: Given a finite set X, for every $\Pi \in \mathcal{P}(X)$ there exists a metric d_Π on X such that $F(X, d_\Pi) = \Pi$.
- Consistency: Let $\Pi = \{B_1, ..., B_n\}$. Let d' be any distance on X such that

  1) for all $x, x' \in B_i$, $d'(x, x') \leq d(x, x')$ and
  2) for all $x \in B_i$, $x' \in B_j$, $i \neq j$, $d'(x, x') \geq d(x, x')$.

  Then, $F(X, d') = \Pi$.

Then, he proved that no standard clustering function can satisfy these three conditions simultaneously. This does not mean that defining a clustering function is impossible. The impossibility only holds when the unique input in the algorithm is the space and the set of distances. It can be avoided including, for example, the number of clusters to be obtained as part of the input. See Ackerman, Ben-David and Locker (2010a) and Zadeht and Ben-David (2009). Another option is to consider hierarchical clustering where the output can be presented as a nested family of partitions, a dendrogram or as an ultrametric space over the same underlying set as the input. The main advantage of this approach is that hierarchical clustering methods output multiscale information of the data set. If the multiscale structure is a significant information of the problem, the output of a standard clustering method, if not irrelevant, would be at least incomplete. For more details see Carlson and Mémoli (2008; 2010).

Carlsson and Mémoli (2010) also studied a property-based approach to hierarchical clustering methods taking as input a finite metric space. They prove, see Theorem 3.1 below, that single linkage hierarchical clustering ($SL\ HC$) is the only hierarchical clustering algorithm which satisfies simultaneously three properties:

- (I) If the input consists of two points at a distance p, the output is the same space.
- (II) A non-increasing map between two metric spaces induces a non-increasing map between the outputs (when considered as ultrametric spaces).
- (III) The distance between any pair of points in the output is greater or equal than the minimal distance between two points in the input.

It is worth mentioning that there is another characterization of SL by Zadeh and Ben-David in the setting of partitional clustering. See Zadeh and Ben-David (2009).

Also Carlsson and Mémoli (2010) prove that $SL\ HC$ exhibits some interesting properties. In particular, it is stable in the Gromov-Hausdorff sense, this is, if two metric spaces are close in the Gromov-Hausdorff metric, then applying the algorithm, the ultrametric spaces obtained are also close in this metric. However, there is a basic weakness in $SL\ HC$ which is the *chaining effect*, see Murtagh (1983) and Wishart (1969). The chaining effect can be seen as the tendency of the algorithm to merge two blocks when the minimal distance between them is small. This might be a problem in many practical situations. Consider, for example, two clusters in $\mathbb{R}^n$ following a multivariate normal distribution. If both clusters are close enough these blocks are merged together very soon by $SL\ HC$, independently of the distribution of the points in the data set.

In Martínez-Pérez (2013), we tried to offer some solution to this effect. We proposed a modified version of $SL\ HC$ called α-unchaining single linkage (or $SL(\alpha)$). This method shows some sensitivity to the density distribution of the data set, so it is capable of detecting blocks when the minimal distance between them is small. We also defined a second version of this method, $SL^*(\alpha)$, to detect blocks when they are connected by a chain of points.

Thus, we were able to offer some solution to these chaining effects but, in exchange, we lost some of the good properties of SL. In particular, $SL(\alpha)$ is no longer stable in the Gromov-Hausdorff sense. In fact, as we proved in Martínez-Pérez (2015), there is no stable solution to this chaining effect in the range of almost-standard linkage-based HC methods using ℓ^{SL}.

Now, the question is when should we use $SL(\alpha)$ and $SL^*(\alpha)$. Among the large variety of clustering methods the best option usually depends on the particular clustering problem. But how do we choose the most suitable algorithm for the task? Ackerman, Ben David and Loker propose to study significant properties of some popular clustering functions. See Ackerman, Ben-David and Locker (2010a; 2010b). The idea is finding abstract significant properties concerning the output of the algorithms which illustrate the difference between applying one clustering method or another. Then, the practitioner should decide which properties are important for the problem under study and choose the algorithm which satisfies them.

In Martínez-Pérez (2013; 2015), we tried to give concrete definitions to express the chaining and stability properties of $SL(\alpha)$ and $SL^*(\alpha)$. Herein, we complete the work by analyzing some basic properties on these methods.

We prove that $SL(\alpha)$ and $SL^*(\alpha)$ are *permutation invariant*, this is, the order in which the points are introduced in the algorithm does not affect the output of the algorithm. They are also *rich*, meaning that for every possible output of the data set, there is a metric in the input such that the algorithm produces that output. In the weighted setting, these methods are not *weight-robust*, this is, assigning different weights to the points may offer different outputs of the algorithm. Notice that this property can be considered to choose between complete linkage (CL), which is weight-robust, and average linkage (AL) which is not. See Ackerman, Ben-David, Branzei and Loker (2012).

$SL(\alpha)$ and $SL^*(\alpha)$ are obtained by adding some unchaining conditions to $SL\ HC$. Thus, we pay special attention to the characteristic properties of SL following the characterization from Carlsson and Mémoli (2010). In order to illustrate the difference between SL and the new methods, we define three natural properties for HC methods which can be summarized as follows:

- A HC method is *faithful* if the algorithm leaves ultrametric spaces invariant, meaning that if the input is an ultrametric space, then the output is the same ultrametric space.
- A HC method is *lower-bounded by* SL if whenever two points are at distance ε in the output, then there exists a ε-chain between them in the input.
- A HC method is *non-expanding in the inclusion* if adding points to the input will never increase the distance between the points in the output.

Being faithful can be a desirable property, especially when the data set is close to an ultrametric space. For example, suppose that the data set is an approximation of a set of points which is contained in an ultrametric space (this can be the case of any tree-like data set, for example, a phylogenetic tree). Then if the measures of the distances between the points were perfect, applying a faithful algorithm we obtain a result which is exactly the real dendrogram. If the method is also semi-stable, see Martínez-Pérez (2015), we can assure that if the measures are precise, then the output will be close the real data.

Being lower-bounded by SL gives a lower bound to the distance between points in the output. It is a very common property. For example, consider any linkage function such that the distance between two blocks is always greater or equal than the minimal distance between them. Then, the corresponding linkage-based hierarchical clustering algorithm will be lower-bounded by SL.

Clearly, SL is faithful, lower-bounded by SL and non-expanding in the inclusion. In fact, we prove that these three properties offer an alternative characterization of SL. See Theorem 3.11.

Many algorithms are faithful and lower-bounded by SL. In particular, $CL\ HC$, $AL\ HC$, $SL(\alpha)$ and $SL^*(\alpha)$ satisfy these properties. Thus, being non-expanding in the inclusion is a key property to illustrate the difference between SL and other methods as those mentioned above. Also, considering the original characterization from Carlsson and Mémoli (2010), it is trivial to check that AL, CL, $SL(\alpha)$ and $SL^*(\alpha)$ satisfy I and III but not II and therefore, property II can be used to distinguish those algorithms from SL. However, since II implies being non-expanding in the inclusion, we believe that the last one is a better option for the task.

The results obtained in Martínez-Pérez (2013;2015) and herein are summarized in Table 1.

2. Background and Notation

A dendrogram over a finite set is a nested family of partitions. This is usually represented as a rooted tree.

Let $\mathcal{P}(X)$ denote the collection of all partitions of a finite set $X = \{x_1, ..., x_n\}$. Then, a dendrogram can also be described, see Carlsson and Mémoli (2010), as a map $\theta\colon [0, \infty) \to \mathcal{P}(X)$ such that:

1. $\theta(0) = \{\{x_1\}, \{x_2\}, ..., \{x_n\}\}$,
2. there exists T such that $\theta(t) = X$ for every $t \geq T$,
3. if $r \leq s$ then $\theta(r)$ refines $\theta(s)$,
4. for all r there exists $\varepsilon > 0$ such that $\theta(r) = \theta(t)$ for $t \in [r, r + \varepsilon]$.

Notice that conditions 2 and 4 imply that there exist $t_0 < t_1 < ... < t_m$ such that $\theta(r) = \theta(t_{i-1})$ for every $r \in [t_{i-1}, t_i)$, $i = 0, 1, ..., m$ and $\theta(r) = \theta(t_m) = \{X\}$ for every $r \in [t_m, \infty)$.

For any partition $\{B_1, ..., B_k\} \in \mathcal{P}(X)$, the subsets B_i are called *blocks*.

Let $\mathcal{D}(X)$ denote the collection of all possible dendrograms over a finite set X. Given some $\theta \in \mathcal{D}(X)$, let us denote $\theta(t) = \{B_1^t, ..., B_{k(t)}^t\}$. Therefore, the nested family of partitions is given by the corresponding partitions at $t_0, ..., t_m$, this is, $\{B_1^{t_i}, ..., B_{k(t_i)}^{t_i}\}$, $i = 0, ..., m$.

An *ε-chain* is a finite sequence of points $x_0, ..., x_N$ that are separated by distances less or equal than ε: $d(x_i, x_{i+1}) \leq \varepsilon$. Two points are *$\varepsilon$-connected* if there is an ε-chain joining them. Any two points in an *ε-connected set* can be linked by an ε-chain. An *ε-component* is a maximal ε-connected subset.

An *ultrametric space* is a metric space (X, d) such that for all $x, y, z \in X$, $d(x, y) \leq \max\{d(x, z), d(z, y)\}$. Given a finite metric space X let $\mathcal{U}(X)$ denote the set of all ultrametrics over X.

There is a well known equivalence between trees and ultrametrics. See Hughes (2004) and Martínez-Pérez and Morón (2009) for a complete

Table 1. Overview of the properties satisfied by the hierarchical clustering methods discussed in this work.

	SL	CL	AL	$SL(\alpha)$	$SL^*(\alpha)$
Permutation invariant	✓	✓	✓	✓	✓
Rich	✓	✓	✓	✓	✓
Weight-robust	✓	✓	✗	✗	✗
Faithful	✓	✓	✓	✓	✓
Non-expanding in the inclusion	✓	✗	✗	✗	✗
Lower bounded by SL	✓	✓	✓	✓	✓
Semi-stable	✓	✓	✓	✓	✗
Stable	✓	✗	✗	✗	✗
Strongly chaining	✓	✗	✗	✗	✗
Completely chaining	✓	✗	✗	✗	✗
Weakly unchaining	✗	✗	✗	✓	✓
α-bridge-unchaining	✗	✗	✗	✗	✓

exposition of how to build categorical equivalences between them. In particular, this may be translated into an equivalence between dendrograms and ultrametrics:

Thus, a hierarchical clustering method $\mathfrak{T}$ can be presented as an algorithm whose output is a dendrogram or an ultrametric space.

Notation: Let $\mathfrak{T}_{\mathcal{D}}(X, d)$ denote that we apply the algorithm $\mathfrak{T}$ to the metric space (X, d) and we present the output as a dendrogram, θ_X (or just θ if there is no ambiguity about the space). Let $\mathfrak{T}_{\mathcal{U}}(X, d)$ denote that we apply the algorithm $\mathfrak{T}$ to the metric space (X, d) an we present the output as an ultrametric space, (X, u_X) (or just (X, u) if there is no ambiguity). The dendrogram θ_X and the ultrametric space (X, u_X) are two presentations of the same object.

In Carlsson and Mémoli (2010), the authors use a recursive procedure to redefine $SL\ HC$, $AL\ HC$ and $CL\ HC$. The main advantage of this procedure is that it allows to merge more than two clusters at the same time. Therefore, AL and $CL\ HC$ can be made permutation invariant. In Martínez-Pérez (2013), we gave an alternative presentation of this recursive procedure as a first step to define $SL(\alpha)$ and $SL^*(\alpha)$. Let us recall here, for completeness, this presentation.

For $x, y \in X$ and any (standard) clustering C of X, $x \sim_C y$ if x and y belong to the same cluster in C and $x \not\sim_C y$, otherwise.

Two (standard) clusterings $C = (C_1, ..., C_k)$ of (X, d) and $C' = (C'_1, ...C'_k)$ of (X', d') are isomorphic clusterings, denoted $(C, d) \cong (C', d')$, if there exists a bijection $\phi : X \to X'$ such that for all $x, y \in X$, $d(x, y) = d'(\phi(x), \phi(y))$ and $x \sim_C y$ if and only if $\phi(x) \sim_{C'} \phi(y)$.

A linkage function can be though of as a way of measuring the distance between two blocks. We use the definition of linkage function from Ackerman, Ben-David and Loker (2010b):

Definition 2.1. A *linkage function* is a function

$$\ell : \{(X_1, X_2, d) \,|\, d \text{ is a distance function over } X_1 \cup X_2\} \to R^+$$

such that,

1. ℓ is *representation independent*: For all (X_1, X_2) and (X_1', X_2'), if they are clustering-isomorphic, $((X_1, X_2), d) \cong ((X_1', X_2'), d')$, then $\ell(X_1, X_2, d) = \ell(X_1', X_2', d')$.
2. ℓ is *monotonic*: For all (X_1, X_2) if d' is a distance function over $X_1 \cup X_2$ such that for all $x \sim_{\{X_1,X_2\}} y$, $d(x,y) = d'(x,y)$ and for all $x \not\sim_{\{X_1,X_2\}} y$, $d(x,y) \leq d'(x,y)$ then $\ell(X_1, X_2, d) \leq \ell(X_1, X_2, d')$.
3. Any pair of clusters can be made arbitrarily distant: For any pair of data sets (X_1, d_1), (X_2, d_2), and any r in the range of ℓ, there exists a distance function d that extends d_1 and d_2 such that $\ell(X_1, X_2, d) > r$.

For technical reasons, it is usually assumed that a linkage function has a countable range. Say, the set of nonnegative algebraic real numbers.

Some standard choices for ℓ are:

- Single linkage: $\ell^{SL}(B, B') = \min_{(x,x') \in B \times B'} d(x, x')$
- Complete linkage: $\ell^{CL}(B, B') = \max_{(x,x') \in B \times B'} d(x, x')$
- Average linkage: $\ell^{AL}(B, B') = \frac{\sum_{(x,x') \in B \times B'} d(x,x')}{\#(B) \cdot \#(B')}$ where $\#(X)$ denotes the cardinality of the set X.

A linkage-based hierarchical clustering method $\mathfrak{T}$ is determined by a linkage function. The algorithm begins with a partition where the blocks are the single points. Then, in each step, every pair of blocks at minimal distance is merged. Notice that this allows multiple blocks to be merged at the same time. As we showed in Martínez-Pérez (2015), this definition could cause technical complications constructing the dendrogram if the linkage function is not *increasing*, this is, if the minimal distance between blocks does not necessarily increase at every step. One solution is to include being increasing as a basic property on the linkage-function. Otherwise, linkage-based methods can be formally presented as follows.

Let (X, d) be a finite metric space where $X = \{x_1, ..., x_n\}$. Let ℓ be some linkage function. Then, the linkage-based algorithm $\mathfrak{T}^\ell(X, d)$ outputs a dendrogram θ^ℓ defined as follows:

1. Let $\Theta_0 := \{\{x_1\}, ..., \{x_n\}\}$ and $R_0 = 0$.
2. For every $i \geq 1$, while $\Theta_{i-1} \neq \{X\}$, let $R_i := \min\{\ell(B, B') \mid B, B' \in \Theta_{i-1},\ B \neq B'\}$.
3. Let Θ_i be the result of merging every pair of blocks $B, B' \in \Theta_{i-1}$ such that $\ell(B, B') = R_i$.
4. Finally, let $\theta^\ell \colon [0, \infty) \to \mathcal{P}(X)$ be such that $\theta^\ell(r) := \Theta_{i(r)}$ with $i(r) := \max\{i \mid R_i \leq r\}$.

In Martínez-Pérez (2015), the methods defined by applying this algorithm for some linkage function ℓ are called *standard linkage-based HC* methods.

Let us now recall the definition of $SL(\alpha)$ and $SL^*(\alpha)$. Further explanations, figures and easy examples can be found in Martínez-Pérez (2013).

2.1 Definition of $SL(\alpha)$

This method is defined to treat the chaining effect produced by blocks which are close using the minimal distance, ℓ^{SL}. See Figure 1. These type of blocks are merged together very soon by SL HC so they are usually undetected by this algorithm.

Given a finite metric space (X, d), let $F_t(X, d)$ be the Rips (or Vietoris-Rips) complex of (X, d). Let us recall that the Rips complex of a metric space (X, d) is a simplicial complex whose vertices are the points of X and $[v_0, ..., v_k]$ is a simplex of $F_t(X, d)$ if and only if $d(v_i, v_j) \leq t$ for every i, j. Given any subset $Y \subset X$, by $F_t(Y)$ we refer to the subcomplex of $F_t(X)$ defined by the vertices in Y. A simplex $[v_0, ..., v_k]$ has dimension k. The dimension of a simplicial complex is the maximal dimension of its simplices.

Notice that densely packed points produce high dimensional simplices in the Rips complex. We will consider this as a sign that the cluster is significant. Low dimensional simplices mean that there are few points close, so they might be suspicious of being noise or poor measurements.

$SL(\alpha)$ is a HC method that depends on a parameter $\alpha \in \mathbb{N}$. We use single linkage, ℓ^{SL}, to measure the distance between blocks and define the method as a linkage-based algorithm.

The algorithm starts with a partition $\Theta_0 = \{\{x_1\}, ..., \{x_n\}\}$. For technical reasons we define $D = \{0 = t_0 < t_1 < \cdots < t_m\}$ to be the ordered set of distances between points in X and iterate the algorithm from $i = 1$ to $i = m$.

In the step i, two blocks, B, B' from the partition Θ_{i-1} are merged if the following two conditions are satisfied simultaneously:

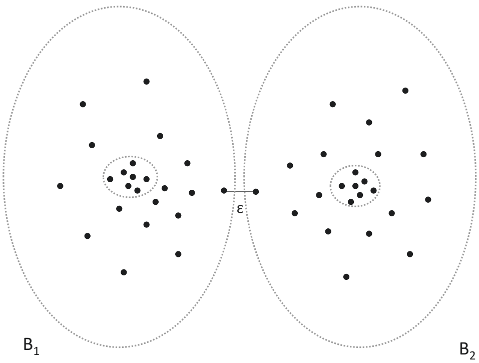

Figure 1. The minimal distance between the blocks B_1 and B_2 is ε. The clustering $\{B_1, B_2\}$ would not be detected by $SL\ HC$.

- $\ell^{SL}(B, B') \leq t_i$
- there is a simplex $\Delta \in F_{t_i}(B \cup B')$ such that $\Delta \cap B \neq \emptyset$, $\Delta \cap B' \neq \emptyset$ and $\alpha \cdot dim(\Delta) \geq \min\{dim(F_{t_i}(B)), dim(F_{t_i}(B'))\}$.

Therefore, two blocks are merged if the minimal distance is t_i (as in $SL\ HC$) unless both blocks contain simplices whose dimension is more than α times the dimension of the simplices that are common to both blocks. In this case, the connection is dismissed.

Finally, the resulting dendrogram is defined as $\theta_\alpha(t) := \Theta_i$ for every $t \in [t_i, t_{i+1})$.

This construction is generalized in Martínez-Pérez (2015) to define the class of *almost-standard linkage-based* HC methods. This class allows to incorporate to a linkage-based algorithm some extra condition on two clusters to be merged.

Let us recall that given some space X with a distance function d, a sequence of points $x_0, x_1, ..., x_n$ is a *t-chain* if $d(x_{i-1}, x_i) \leq t$ for every $1 \leq i \leq n$.

Remark 2.2 It is immediate, by construction, that if two points x, x' belong to the same block of $\theta_\alpha(t_i)$ then, necessarily, there exists a t_i-chain, $x = x_0, x_1, ..., x_n = x'$ joining them

2.2 Definition of $SL^*(\alpha)$

This method is defined on the basis of $SL(\alpha)$ to avoid two blocks to be merged when there is a chain of single points or small blocks joining them. This method introduces a distinction between big blocks, which are considered as the relevant information, and small blocks which are considered as secondary in the formation of new clusters.

The algorithm starts again with the single points $\Theta_0^* := \{\{x_1\}, ..., \{x_n\}\}$.

Then, given the ordered set of distances between points in X, $D = \{0 = t_0 < t_1 < \cdots < t_m\}$, the algorithm iterates from $i = 1$ to $i = m$.

For every i, two blocks, B, B', from the partition Θ_{i-1}^* are related, $B \sim^* B'$, if the following conditions (the same as in $SL(\alpha)$) are satisfied simultaneously:

- $\ell^{SL}(B, B') \leq t_i$
- there is a simplex $\Delta \in F_{t_i}(B \cup B')$ such that $\Delta \cap B \neq \emptyset$, $\Delta \cap B' \neq \emptyset$ and $\alpha \cdot dim(\Delta) \geq \min\{dim(F_{t_i}(B)), dim(F_{t_i}(B'))\}$.

Now let us define a graph $G_\alpha^{t_i}$ whose vertices are the blocks in Θ_{i-1}^* and there is an edge joining the blocks B, B' if and only if $B \sim^* B'$.

Notice that $SL(\alpha)$ just merges together the blocks that form the connected components of this graph to define Θ_i. In $SL^*(\alpha)$ we distinguish between big blocks and small blocks in each connected component using the parameter α.

Let A be any connected component of $G_\alpha^{t_i}$, denoted $A \in cc(G_\alpha^{t_i})$, with blocks $B_1, ..., B_r$. We call *big blocks* of A those blocks $B_i \in A$ such that

$$\alpha \cdot \#(B_i) \geq \max_{1 \leq l \leq r}\{\#(B_l)\}. \tag{1}$$

The rest of the blocks of A are called *small blocks*.

Let $H_\alpha(A)$ be the subgraph of A whose vertices are the big blocks and $S_\alpha(A)$ be the subgraph of A whose vertices are the small blocks.

Then, $B, B' \in \Theta_{i-1}^*$ are merged in Θ_i^* if B, B' belong to the same connected component, let us say A, and one of the following conditions holds:

iii) $\exists\, C \in cc(H_\alpha(A))$ such that $B, B \in C$.

iv) $B \in C \in cc(H_\alpha(A))$, $B' \in C' \in cc(S_\alpha(A))$ and there is no big block in $A \backslash C$ adjacent to any block in C'.

Finally, the resulting dendrogram is defined as $\theta_\alpha^*(t) := \Theta_i^*$ for every $t \in [t_i, t_{i+1})$.

Remark 2.3 At step iii, if $H_\alpha(A)$ is connected, then $B_{j_1} \cup \cdots \cup B_{j_r}$ defines a block of $\theta_\alpha(t_i)$.

Remark 2.4 Notice that if two points x, x' belong to the same block of $\theta^*_\alpha(t_i)$ then, necessarily, there exists a t_i-chain, $x = x_0, x_1, ..., x_n = x'$ joining them.

3. Single Linkage Hierarchical Clustering

In this section we recall some basic properties and the characterization of $SL\ HC$ from Carlsson and Mémoli (2010). We also propose some alternatives. Our first intention is to find significant properties to compare SL and $SL(\alpha)$.

3.1 Characterization of SL

Carlsson and Mémoli provided the following axiomatic characterization of $SL\ HC$:

Let us recall that given a finite metric space (X, d), $sep(X, d) := min_{x \neq x'} d(x, x')$.

Theorem 3.1 (Carlsson and Mémoli 2010, Theorem 18) *Let $\mathfrak{T}$ be a hierarchical clustering method such that:*

(I) $\mathfrak{T}_\mathcal{U}\Big(\{p,q\}, \begin{pmatrix} \delta & 0 \\ 0 & \delta \end{pmatrix}\Big) = \Big(\{p,q\}, \begin{pmatrix} \delta & 0 \\ 0 & \delta \end{pmatrix}\Big)$ *for all* $\delta > 0$.

(II) Given two finite metric spaces X, Y and $\phi\colon X \to Y$ such that $d_X(x,x') \geq d_Y(\phi(x), \phi(x'))$ for all $x, x' \in X$, then

$$u_X(x,x') \geq u_Y(\phi(x), \phi(x'))$$

also holds for all $x, x' \in X$, where $\mathfrak{T}_\mathcal{U}(X, d_X) = (X, u_X)$ and $\mathfrak{T}_\mathcal{U}(Y, d_Y) = (Y, u_Y)$.

(III) For any metric space (X, d),

$$u(x,x') \geq sep(X,d) \text{ for all } x \neq x' \in X$$

where $\mathfrak{T}_\mathcal{U}(X,d) = (X,u)$.

Then, $\mathfrak{T}$ is exactly single linkage hierarchical clustering.

Notation: For the particular case of $SL\ HC$, if there is no need to distinguish the metric space, let us denote $\mathfrak{T}^{SL}_\mathcal{D}(X,d) = \theta_{SL}$ and $\mathfrak{T}^{SL}_\mathcal{U}(X,d) = (X, u_{SL})$.

Notation: Given two metrics d, d' defined on a set X, let us denote $d \leq d'$ if $d(x,x') \leq d'(x,x')\ \forall\, x, x' \in X$.

Definition 3.2 A hierarchical clustering method $\mathfrak{T}$ is *lower-bounded by SL* if whenever two points are at distance ε in the output, then there exists a ε-chain between them in the input (i.e. if for every metric space (X, d), $u \geq u_{SL}$, where $\mathfrak{T}_{\mathcal{U}}(X, d) = (X, u)$).

The following propositions follow immediately from the proof of Theorem 18 in Carlsson and Mémoli (2010).

Proposition 3.3 *For any metric space (X, d), if $\mathfrak{T}$ satisfies conditions II and III, then $\mathfrak{T}$ is lower-bounded by SL (i.e. $u \geq u_{SL}$).*

Also, it is readily seen that if a HC algorithm is lower-bounded by SL, then $\mathfrak{T}$ satifies III.

Proposition 3.4 *If $\mathfrak{T}$ satisfies conditions I and II, then $u_{SL} \geq u$.*

In fact, Proposition 3.4 can be improved introducing the following definition.

Definition 3.5 A hierarchical method $\mathfrak{T}$ is *non-expanding in the inclusion* if for any metric space (Y, d) and any subset $X \subset Y$, if $i \colon X \to Y$ is the inclusion map, then $u_X(x, x') \geq u_Y(i(x), i(x'))$.

This means that, by adding points to the input, the ultrametric distance in the output may turn smaller but never bigger. Clearly, property II implies being non-expanding in the inclusion.

Remark 3.6 Carlsson and Mémoli (2008) study hierarchical clustering (or persistent clustering) and introduce some notion of functoriality. In order to compare the application of the algorithm on different data sets, it would be very useful to consider the algorithm as a functor from the data sets (metric spaces or sets with a distance function) to dendrograms. The authors consider three different categories whose objects are finite metric spaces: $\underline{\mathcal{M}}^{iso}$ whose morphisms are isometries, $\underline{\mathcal{M}}^{mon}$ whose morphisms are distance non-increasing maps (i.e. non-expanding maps) which are inclusions as set maps, and $\underline{\mathcal{M}}^{gen}$ whose morphisms are distance non-increasing maps.

In this sense, it can be seen that being non-expanding in the inclusion is equivalent to being functorial in the category $\underline{\mathcal{M}}^{sub}$ whose objects are finite metric spaces and whose morphisms are inclusions (this is, isometries which are inclusions as set maps). Since $\underline{\mathcal{M}}^{sub} \subset \underline{\mathcal{M}}^{mon} \subset \underline{\mathcal{M}}^{gen}$, if the method is not non-expanding in the inclusion, then it is not functorial neither in $\underline{\mathcal{M}}^{mon}$ nor $\underline{\mathcal{M}}^{gen}$.

If we accept that HC algorithms must be faithful and lower-bounded by SL, then the uniqueness Theorem 3.11 below, implies that SL is the unique method which defines a functor from data sets to dendrograms in the categories $\underline{\mathcal{M}}^{sub}$, $\underline{\mathcal{M}}^{mon}$ and $\underline{\mathcal{M}}^{gen}$. This is closely related with the uniqueness theorem on the functoriality on $\underline{\mathcal{M}}^{gen}$, see Theorem 4.1 in Carlsson and Mémoli (2008).

The proof of Proposition 3.4 in Carlsson and Mémoli (2010) can be trivially adapted to obtain the following.

Proposition 3.7 *If $\mathfrak{T}$ satisfies I and is non-expanding in the inclusion, then $u_{SL} \geq u$.*

Proof. Let $x, x' \in (X, d)$ such that $u_{SL}(x, x') = \delta$. Then, there exists a δ-chain $x = x_0, x_1, ..., x_n = x'$ such that $\max_i d(x_{i-1}, x_i) = \delta$. By I, if $X_i = \{x_{i-1}, x_i\}$, $\mathfrak{T}(X_i, d|_{X_i}) = (X_i, d|_{X_i})$ and $u_{X_i}(x_{i-1}, x_i) \leq \delta$. Then, since $\mathfrak{T}$ is non-expanding in the inclusion, $u(x_{i-1}, x_i) \leq u_{X_i}(x_{i-1}, x_i) \leq \delta$ and, by the properties of the ultrametric, $u(x, x') \leq \delta$.
■

As we mentioned above, a hierarchical clustering method can be seen as an algorithm that takes as input some space with a distance function and gives as output an ultrametric space. Thus, another natural condition to ask on a hierarchical clustering method is that applying it to an ultrametric space we obtain the same ultrametric space.

Definition 3.8 We say that $\mathfrak{T}$ is *faithful* if given an ultrametric space as input, then the output is the same ultrametric space (i.e. for any ultrametric space (X, d), $u(x, y) = d(x, y)$, where $\mathfrak{T}_{\mathcal{U}}(X, d) = (X, u)$).

It can be readily seen that $SL\ HC$ is faithful:

Proposition 3.9 *If (X, d) is an ultrametric space, then $u_{SL}(x, y) = d(x, y)$ for every $x, y \in X$.*

Proof. By definition, it is clear that $u_{SL}(x, y) \leq d(x, y)$ for every $x, y \in X$.

Let us see that, if (X, d) is an ultrametric space, then $u_{SL}(x, y) \geq d(x, y)$. $u_{SL}(x, y) = \inf\{t \mid \text{there exists a } t\text{-chain joining } x \text{ to } y\}$. Suppose $u_{SL}(x, y) = t$ and let $x = x_0, x_1, ..., x_n = y$ be a t-chain joining x to y. By the properties of the ultrametric, $d(x_{i-1}, x_{i+1}) \leq \max\{d(x_{i-1}, x_i), d(x_i, x_{i+1})\} \leq t$ for every $1 \leq i \leq n - 1$. Therefore, $d(x, y) \leq t$ and $u_{SL}(x, y) \geq d(x, y)$.
■

Richness property for HC methods can be defined in the same way Kleinberg did for standard clustering. Thus, a HC method $\mathfrak{T}$ is *rich* if given a finite set X, for every $\theta \in \mathcal{D}(X)$ there exists a metric d_θ on X such that $\mathfrak{T}_{\mathcal{D}}(X, d_\theta) = \theta$. Therefore, being faithful immediately implies being rich.

Corollary 3.10 *$\mathfrak{T}^{SL}$ is rich.*

By Proposition 3.9, $SL\ HC$ is faithful. It is trivial to see that it is lower-bounded by SL. Also, since property (II) implies being non-expanding in the inclusion, it follows from Theorem 3.1 that SL is non-expanding in the inclusion. In fact, these properties give an alternative characterization of $SL\ HC$.

Theorem 3.11 *$SL\ HC$ is the unique hierarchical clustering method which is faithful, non-expanding in the inclusion and lower-bounded by SL.*

Proof. Trivially, being faithful implies property (I). By Proposition 3.7, if a method $\mathfrak{T}$ satisfies property (I) and is non-expanding in the inclusion, then $u_{SL} \geq u$. If $\mathfrak{T}$ is also lower-bounded by SL, then $u = u_{SL}$. Therefore, $\mathfrak{T}$ is equivalent to SL.
■

3.2 Stability of SL

Let us recall the definition of Gromov-Hausdorff distance from Burago, Burago and Ivanov (2001). See also Gromov (2007).

Let (X, d_X) and (Y, d_Y) be two metric spaces. A correspondence (between A and B) is a subset $R \in A \times B$ such that

- $\forall\, a \in A$, there exists $b \in B$ s.t. $(a, b) \in R$
- $\forall\, b \in B$, there exists $a \in A$ s.t. $(a, b) \in R$

Let $\mathcal{R}(A, B)$ denote the set of all possible correspondences between A and B.

Let $\Gamma_{X,Y} : X \times Y \times X \times Y \to \mathbb{R}^+$ given by

$$(x, y, x', y') \mapsto |d_X(x, x') - d_Y(y, y')|.$$

Then, the *Gromov-Hausdorff distance* between X and Y is:

$$d_{\mathcal{GH}}(X, Y) := \frac{1}{2} \inf_{R \in \mathcal{R}(X,Y)} \sup_{(x,y)(x',y') \in R} \Gamma_{X,Y}(x, y, x', y').$$

The *Gromov-Hausdorff metric* gives a notion of distance between metric spaces. One of the advantages of this metric is that it is well defined for metric spaces of different cardinality. In Carlsson and Mémoli (2010) this metric is used to prove that $\mathfrak{T}^{SL}$ holds some stability under small perturbations on the metric. The authors prove that if two metric spaces are close (in the Gromov-Hausdorff metric) then the corresponding ultrametric spaces obtained as output of the algorithm are also close. In Martínez-Pérez

(2015), we studied Gromov-Hausdorff stability of linkage-based HC methods defining the following conditions.

Notation: Let $(\mathcal{M}, d_{GH})$ denote the set of finite metric spaces with the Gromov-Hausdorff metric and $(\mathcal{U}, d_{GH})$ denote the set of finite ultrametric spaces with the Gromov-Hausdorff metric.

Definition 3.12 A HC method $\mathfrak{T}$ is *semi-stable in the Gromov-Hausdorff sense* if for any sequence of finite metric spaces $((X_k, d_k))_{k\in\mathbb{N}}$ in $(\mathcal{M}, d_{GH})$ such that $\lim_{k\to\infty}(X_k, d_k) = (U, d) \in \mathcal{U}$ then $\lim_{k\to\infty} \mathfrak{T}_{\mathcal{U}}(X_k, d_k) = \mathfrak{T}_{\mathcal{U}}(U, d)$.

Definition 3.13 A HC method $\mathfrak{T}$ is *stable in the Gromov-Hausdorff sense* if

$$\mathfrak{T}_{\mathcal{U}} \colon (\mathcal{M}, d_{GH}) \to (\mathcal{U}, d_{GH})$$

is continuous.

A hierarchical clustering method is said to be *permutation invariant* if it yields the same dendrogram under permutation of the points in the sample, this is, if the output of the algorithm does not depend on the order by which the data is introduced. Although this is not the easiest way to check this property, it may be noticed that being stable in the Gromov-Hausdorff sense implies being permutation invariant.

The following result is a consequence of Proposition 26 in Carlsson and Mémoli (2010).

Proposition 3.14 *SL HC is stable in the Gromov-Hausdorff sense. In particular, it is semi-stable and permutation invariant.*

4. Basic Properties of $SL(\alpha)$ and $SL^*(\alpha)$

In this section, we study some basic properties on $SL(\alpha)$ and $SL^*(\alpha)$. In particular, we check those seen at Section 3.

The following result is clear from the definition.

Proposition 4.1 *$SL(\alpha)$ and $SL^*(\alpha)$ are permutation invariant algorithms.*

Notation: Let (X, d) be a finite metric space. Let us recall that if there is no ambiguity on the metric space we denote $\mathfrak{T}_{\mathcal{D}}^{SL}(X, d) = \theta_{SL}$ and $\mathfrak{T}_{\mathcal{U}}^{SL}(X, d) = u_{SL}$. Let us denote $\mathfrak{T}_{\mathcal{D}}^{SL(\alpha)}(X, d) = \theta_\alpha$ and $\mathfrak{T}_{\mathcal{U}}^{SL(\alpha)}(X, d) = u_\alpha$. Similarly, let $\mathfrak{T}_{\mathcal{D}}^{SL^*(\alpha)}(X, d) = \theta_\alpha^*$ and $\mathfrak{T}_{\mathcal{U}}^{SL^*(\alpha)}(X, d) = u_\alpha^*$.

The following two results state that if α is big enough then the algorithms $SL(\alpha)$ and $SL^*(\alpha)$ respectively, are equivalent to SL.

Proposition 4.2 *Let (X, d) be a finite metric space with $X = \{x_1, ..., x_n\}$. If $\alpha \geq \frac{n-2}{2}$, then $\theta_{SL} = \theta_\alpha$.*

Proof. Let $\mathfrak{T}_{\mathcal{D}}^{SL}(X,d) = \theta_{SL}$, $\mathfrak{T}_{\mathcal{D}}^{SL(\alpha)}(X,d) = \theta_\alpha$.

We know that $\theta_{SL}(t_0) = \theta_\alpha(t_0)$.

Suppose $\theta_{SL}(t_{i-1}) = \theta_\alpha(t_{i-1})$. Let us see that if $\alpha \geq \frac{n-2}{2}$, condition i already implies ii and the edges of the graph $G_\alpha^{t_i}$ are those defined by condition i. Let B_1, B_2 be two blocks in $\theta_\alpha(t_{i-1})$ such that $\min\{d(x,y) \,|\, x \in B_1,\, y \in B_2\} \leq t_i$. For any simplex Δ, $\alpha{\cdot}dim(\Delta) \geq \alpha$ and $\min\{\#(B_1), \#(B_2)\} \leq \frac{n}{2}$. Since $\alpha \geq \frac{n-2}{2}$, $\alpha \cdot dim(\Delta) \geq \alpha \geq \min\{\#(B_1) - 1, \#(B_2) - 1\} \geq \min\{dim(F_{t_i}(B_1)), dim(F_{t_i}(B_2))\}$. Therefore, $\theta_{SL}(t_i) = \theta_\alpha(t_i)$.

Thus, $\theta_{SL}(t) = \theta_\alpha(t)$ for every $t \geq 0$.

■

Proposition 4.3 *Let (X,d) be a finite metric space with $X = \{x_1, ..., x_n\}$. If $\alpha \geq n-1$, then $\theta_{SL} = \theta_\alpha^*$.*

Proof. Let $\mathfrak{T}_{\mathcal{D}}^{SL}(X,d) = \theta_{SL}$ and $\mathfrak{T}_{\mathcal{D}}^{SL^*(\alpha)}(X,d) = \theta_\alpha^*$.

We know that $\theta_{SL}(t_0) = \theta_\alpha^*(t_0)$.

Suppose $\theta_{SL}(t_{i-1}) = \theta_\alpha^*(t_{i-1})$. As we saw in the proof of Proposition 4.2, since $\alpha \geq n-1 > \frac{n-2}{2}$, condition i already implies ii and the edges of the graph $G_\alpha^{t_i}$ are those defined by condition i.

Now, let $A = \{B_1, ..., B_r\}$ be any connected component of $G_\alpha^{t_i}$.

If the subgraph $H_\alpha(A)$ is not connected, then there are at least three blocks B_{i_1}, B_{i_2}, B_{i_3} in A, such that $1 \leq \#(B_{i_1}) < \frac{1}{\alpha} \max_{1 \leq l \leq r}\{\#(B_l)\}$ and $\#(B_{i_2}), \#(B_{i_3}) \geq \frac{1}{\alpha} \max_{1 \leq l \leq r}\{\#(B_l)\}$. Trivially, $\max_{1 \leq l \leq r}\{\#(B_l)\} \leq n-2$. Hence, there is a contradiction since $1 \leq \frac{n-2}{\alpha} \leq \frac{n-2}{n-1} < 1$.

Hence, $H_\alpha(A)$ is connected and, as we saw in Remark 2.3, all the blocks in A are identified. Therefore, $\theta_\alpha^*(t_i) = \theta_{SL}(t_i)$.

Thus, $\theta_{SL}(t) = \theta_\alpha^*(t)$ for every $t \geq 0$.

■

Proposition 4.4 *$SL(\alpha)$ and $SL^*(\alpha)$ are lower-bounded by SL (i.e. $u_{SL} \leq u_\alpha$ and $u_{SL} \leq u_\alpha^*$) for every $\alpha \in \mathbb{N}$.*

Proof. As we saw at Remarks 2.2 and 2.4, if two points $x, x' \in X$ belong to the same block of $\theta_\alpha(t)$ (resp. $\theta_\alpha^*(t)$), they belong, in particular, to the same t-component of X and, therefore, to the same block of $\theta_{SL}(t)$. Thus, $u_{SL}(x,x') \leq u_\alpha(x,x')$ (resp. $u_{SL}(x,x') \leq u_\alpha^*(x,x')$).

■

The following result proves that if the input is an ultrametric space, then $SL(\alpha)$ and $SL^*(\alpha)$ are equivalent to SL.

Proposition 4.5 *If (X,d) is an ultrametric space, then $\theta_\alpha = \theta_{SL} = \theta_\alpha^*$ for every α.*

Proof. By definition, $\theta_\alpha(t_0) = \theta_{SL}(t_0) = \theta^*_\alpha(t_0)$. Suppose $\theta_\alpha(t_{i-1}) = \theta_{SL}(t_{i-1}) = \theta^*_\alpha(t_{i-1}) = \{B_1, ..., B_n\}$. Let us see that $\theta_\alpha(t_i) = \theta_{SL}(t_i) = \theta^*_\alpha(t_i)$.

Let B_i, B_j be such that $\min\{d(x,y) \,|\, x \in B_i,\ y \in B_j\} \leq t_i$. Since B_i, B_j are (t_{i-1})-components, by the properties of the ultrametric, $d(x,y) \leq t_i$ for every $(x,y) \in B_1 \times B_2$.

Therefore, every pair of points in $B_1 \cup B_2$ define a simplex in $F_{t_i}(B_1 \cup B_2)$ and condition ii holds for every α. Thus, there is an edge defined between B_i and B_j. This proves that $\theta_\alpha = \theta_{SL}$.

Now, let B_i, B_j be two blocks in the same connected component of $G^{t_i}_\alpha$. Then, by the properties of the ultrametric, $\{B_i, B_j\}$ is an edge of $G^{t_i}_\alpha$. Hence, for any connected component A of $G^{t_i}_\alpha$, $H_\alpha(A)$ is connected and, by Remark 2.3, $\theta^*_\alpha(t_i)$ is defined by the connected components of $G^{t_i}_\alpha$. This proves that $\theta^*_\alpha = \theta_{SL}$.
■

Since SL is faithful, it follows from Proposition 4.5 that:

Corollary 4.6 *$SL(\alpha)$ and $SL^*(\alpha)$ are faithful.*

Corollary 4.7 *$SL(\alpha)$ and $SL^*(\alpha)$ are rich.*

Notice that $SL(\alpha)$ and $SL^*(\alpha)$ are faithful and lower-bounded by SL. Since they are different from SL, by Theorem 3.11, it follows that they are not non-expansive in the inclusion. In the following example from Martínez-Pérez (2013), we show how including some space in a larger data set can expand the distance between points in the output applying $SL(\alpha)$.

Example 4.8 *Let (X,d) be the graph from Figure 2.*

Suppose the edges in N_1, N_2 have length 1 and the rest have length 3. The distances between vertices are measured as the minimal length of a path joining them.

Let $Z := \{x_0, y_0\}$ and $d'(x_0, y_0) = 3$. Let $i\colon (Z, d') \to (X, d)$ be the inclusion map. It is immediate to check that applying either $SL(\alpha)$ or $SL^*(\alpha)$ with $\alpha < 3$ we obtain ultrametric spaces (Z, u_Z), (X, u_X) such that $u_Z(x_0, y_0) < u_X(x_0, y_0)$.

Ackerman, Ben-David, Branzei and Loker (2012) study clustering in the weighted setting. A weight function on the data set X is a map $w : X \to \mathbb{R}^+$. The authors analyze the influence of weight functions on different clustering algorithms. If the weight map w is evaluated in $\mathbb{N}$, it can be interpreted as the data set having $w(x)$ copies of the point x in the same location. In this particular case, their work is related to Fisher and Van Ness (1971) where the authors studied how the duplication of points affected the output of different clustering algorithms.

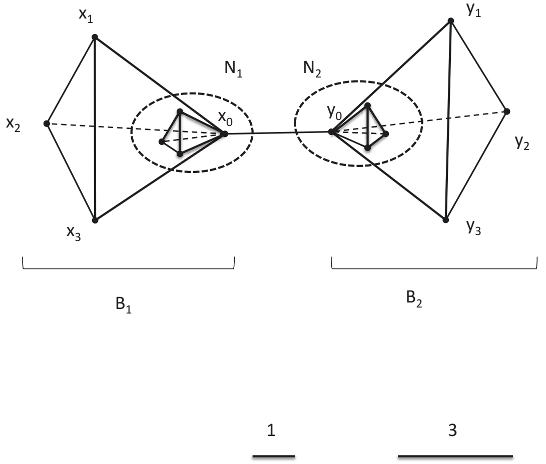

Figure 2. $SL(\alpha)$ is not non-expanding in the inclusion.

Linkage-based algorithms can be defined in the weighted setting, as usual, based on a linkage function. SL and CL work exactly the same way in the weighted setting since ℓ^{SL} and ℓ^{CL} are not affected by the weight function. For the case of AL, let $w(X) = \sum_{x \in X} w(x)$. Then, the average-linkage linkage function is

$$\ell^{AL}(X_1, X_2, d, w) = \sum_{x \in X_1, y \in X_2} \frac{d(x,y) \cdot w(x) \cdot w(y)}{w(X_1) \cdot w(X_2)}.$$

In Ackerman, Ben-David, Branzei and Loker (2012) the authors consider the property of being weight-robust proving that this condition distinguishes between AL and CL hierarchical k-clustering. We adapt the definition to offer an easier and more intuitive exposition of what is being weight-robust. Our definition implies the one from Ackerman, Ben-David, Branzei and Loker (2012) without being exactly equivalent. Nevertheless, the results herein work for both definitions.

Definition 4.9 *A hierarchical clustering algorithm $\mathfrak{T}$ is* weight-robust *if for all (X, d) and all weight functions w, w', $\mathfrak{T}(w[X], d) = \mathfrak{T}(w'[X], d)$.*

Since ℓ^{SL} and ℓ^{CL} are independent from the weight given to the points, it is readily seen that SL and CL are weight-robust. See Fisher and Van Ness (1971). However, as it is proved in Ackerman, Ben-David, Branzei and Loker (2012), AL is not weight-robust.

Let us see how the definition of $SL(\alpha)$ and $SL^*(\alpha)$ can be extended to the weighted setting. The distance between blocks is measured using ℓ^{SL}

which is not affected by the weight. Therefore, we only need to define how to compute the dimension of the simplices in the Rips Complex. Given a weight function $w : X \to \mathbb{R}^+$, let a simplex $[x_0, ..., x_n]$ have dimension $\sum_{i=0}^n w(x_i) - 1$. This definition is consistent with considering for each point x, $w(x)$ copies in the same location.

Proposition 4.10 *$SL(\alpha)$ and $SL^*(\alpha)$ are not weight-robust.*

Proof. Consider the space (X, d) shown in Figure 2 where every vertex has weight 1.

Applying $SL(\alpha)$ with $\alpha = 1$ we obtain a dendrogram θ_α such that

$\theta_\alpha(t) = \{\{x_0\}, ..., \{x_6\}, \{y_0\}, ..., \{y_6\}\}$ for every $t < 1$,

$\theta_\alpha(t) = \{\{x_1\}, \{x_2\}, \{x_3\}, N_1, N_2, \{y_1\}, \{y_2\}, \{y_3\}\}$ for every $1 \leq t < 3$, where $N_1 = \{x_0, x_4, x_5, x_6\}$ and $N_2 = \{y_0, y_4, y_5, y_6\}$.

$\theta_\alpha(t) = \{B_1, B_2\}$ for every $3 \leq t < 5$

$\theta_\alpha(t) = X$ for every $5 \leq t$.

Applying $SL^*(\alpha)$ the same dendrogram is obtained.

Now, suppose we triplicate in (X, d) the points x_0, y_0, this is, we consider $w(x_0) = w(y_0) = 3$ and $w(x_i) = 1 = w(y_i)$ for every $1 \leq i \leq 6$. Then, let us apply $SL(\alpha)$ on the new space (X, d').

$\theta_\alpha(t) = \{\{x_0\}, , ..., \{x_6\}, \{y_0\}, ..., \{y_6\}\}$ for every $t < 1$,

$\theta_\alpha(t) = \{\{x_1\}, \{x_2\}, \{x_3\}, N_1, N_2, \{y_1\}, \{y_2\}, \{y_3\}\}$ for every $1 \leq t < 3$.

Now, for $t = 3$, notice that the Rips complexes $F_3(N_1)$, $F_3(N_2)$ have dimension 5. Also, there is a complex $\{x_0, y_0\}$ in $F_3(X, d)$ which intersects N_1 and N_2 and has dimension 5. Since $\alpha = 1$, N_1 and N_2 are merged. Thus, $\theta_\alpha(t) = X$ for every $3 \leq t$.

Applying $SL^*(\alpha)$ the same dendrogram is obtained. Therefore, both methods are not weight-robust.

■

5. Chaining Effect and Stability

In this section we recall, for completeness, the results about chaining and stability properties of $SL(\alpha)$ and $SL^*(\alpha)$ obtained in Martínez-Pérez (2013; 2015).

Chaining effect is treated in a somehow imprecise way in the literature. There are several types of effects which can be included under this name. $SL(\alpha)$ is defined to treat some particular type of chaining effect which is the tendency of the algorithm to merge two blocks when the minimal distance between them is small. This is clearly one of the main problems of SL. In Martínez-Pérez (2013), we defined concrete properties to bound the kind of chaining effects we are considering. These properties were being *strongly chaining* and being *completely chaining* which might be undesired chaining effects in many practical problems.

A HC method is *strongly chaining* if given two blocks B_1, B_2 such that the minimal distance between them is smaller that the minimal ε such that B_1 is ε-connected, then the clustering $\{B_1, B_2\}$ does not appear at any level of the resulting dendrogram.

Roughly speaking, a HC method is *completely chaining* if given two blocks B_1, B_2 such that they are connected by an ε-chain $x_0, x_1, ..., x_k$ with $x_0 \in B_1$ and $x_k \in B_2$ where B_1 is not ε-connected, then B_1 is never a block in the resulting dendrogram.

Then, we prove that SL is strongly chaining and completely chaining while AL, CL, $SL(\alpha)$ and $SL^*(\alpha)$ are neither strongly chaining nor completely chaining.

We also define two properties to illustrate the type of chaining effects that can be avoided using $SL(\alpha)$ and $SL(\alpha)$ respectively. These are being α-weakly unchaining and α-bridge unchaining.

A method is *α-weakly unchaining* if it is able to detect the clustering $\{B_1, B_2\}$ where

- B_1, B_2 are ε-connected,
- there is a single pair of points $x_0 \in B_1, y_0 \in B_2$ with $d(x, y) = t < \varepsilon$,
- $d(x, y) > \varepsilon$ for every $(x_0, y_0) \neq (x, y) \in B_1 \times B_2$,
- $F_t(B_1) > \alpha$, $F_t(B_2) > \alpha$.

Roughly speaking, a method is *α-bridge unchaining* if it is able to detect in $X = B_1 \cup \{x_1, ..., x_{k-1}\} \cup B_2$ the clusters B_1, B_2 where

- ε is the minimal number such that B_1, B_2 are ε-connected,
- $dim(F_\varepsilon(B_1)), dim(F_\varepsilon(B_2)) > \alpha$
- there is an ε-chain $x_0, ..., x_k$ with $x_0 \in B_1, x_k \in B_k$ and $x_2, ..., x_{k-1} \notin B_1 \cup B_2$

In particular, being strongly chaining implies that the method is not weakly unchaining for any α and being completely chaining implies that the method is not bridge unchaining for any α.

In Martínez-Pérez (2013), we proved that $SL(\alpha)$ and $SL^*(\alpha)$ are α-weakly unchaining (and SL, CL, AL are not). We also proved that $SL^*(\alpha)$ is α-bridge unchaining (and SL, CL, AL, $SL(\alpha)$ are not).

In Martínez-Pérez (2015), we studied the stability of linkage-based hierarchical clustering algorithms. We used Gromov-Hausdorff metric to define stability as in Carlsson and Mémoli (2010). In particular, we defined a method to be *stable* if for every pair of finite data sets which are close in the Gromov-Hausdorff distance, then the ultrametric spaces obtained in the output are still close in the same metric (this is, the algorithm is continuous as a functor from data sets to ultrametric spaces). We defined a method to be *semi-stable* if given a sequence of finite data sets which is convergent to an ultrametric space in the Gromov-Hausdorff metric then the ultrametric spaces obtained in the output also converge to the output obtained from the ultrametric space.

We proved that $SL(\alpha)$ is semi-stable in the Gromov-Hausdorff sense. Unfortunately, most of the good stability properties of SL do not hold. $SL(\alpha)$ and $SL^*(\alpha)$ are not stable in the Gromov-Hausdorff sense. Also, it is not difficult to check that $SL^*(\alpha)$ is not semi-stable in the Gromov-Hausdorff sense.

To complete the results of Table 1 let us recall that SL, CL and AL are permutation invariant, see Carlsson and Mémoli (2010); SL, CL are weight-robust and AL is not weight-robust, see Ackerman, Ben-David, Branzei and Loker (2012); SL, AL and CL are faithful, see Martínez-Pérez (2015). Richness can be obtained as an immediate consequence of being faithful.

6. Conclusions

We prove that $SL(\alpha)$ and $SL^*(\alpha)$ are faithful (i.e. the algorithm leaves ultrametric spaces invariant), lower-bounded by SL (i.e. if the distance between two points in the output is ε, then there is a ε-chain between them in the input), permutation invariant (i.e. the output of the algorithm does not depend on the order by which the data is introduced) and rich (i.e. given a data set, for every possible output, there is a metric in the input such that the application of the algorithm yields that output).

These properties are satisfied by many algorithms, SL, CL and AL among them. In the spirit of Kleinberg impossibility result we may consider being faithful, lower-bounded by SL, permutation invariant and rich, as basic desirable conditions for any hierarchical clustering algorithm.

We prove that SL is the only hierarchical clustering algorithm which is simultaneously faithful, lower-bounded by SL and non-expanding in the inclusion. In particular, the last property is not satisfied by the algorithms defined to treat the chaining effects: $SL(\alpha)$ and $SL^*(\alpha)$. This means that if we are analyzing only a subset of the data set, which may be practical if the whole data set is too big to compute, using SL, the distance between the points in the output is greater or equal than the distance that would result from applying the algorithm to the whole data set. Applying any other algorithm, this can not be assured. Typically, the distance between the points in the output when only a subspace is considered may result either smaller or greater applying any other algorithm.This is the case with CL, AL, $SL(\alpha)$ or $SL^*(\alpha)$.

We also proved that $SL(\alpha)$ and $SL^*(\alpha)$ are not weight-robust. Therefore, in both cases assigning weights to the points may have a deep influence in the output of the algorithm. This is also the case applying AL but not if we are using SL or CL.

The chaining and unchaining properties of these methods were studied in Martínez-Pérez (2013). The stability properties of linkage-based methods were analyzed in Martínez-Pérez (2015). The main results from those papers together with the results obtained herein are summarized in Table 1.

References

ACKERMAN, M., BEN-DAVID, S., and LOKER, D. (2010a), "Towards Property-Based Algorithms Among Hierarchical Clustering Methods", *Advances in Neural Information Processing Systems, 23,* 10–18.

ACKERMAN, M., BEN-DAVID, S., and LOKER, D. (2010b), "Characterization of Linkage-Based Clustering", *Proceedings of the 23rd International Conference on Learning Theory (COLT).*

ACKERMAN, M., BEN-DAVID, S., BRANZEI, S., and LOKER, D. (2012), "Weighted Clustering," *Proceedings of the 26th AAAI Conference on Artificial Intelligence.*

BURAGO, D., BURAGO, Y., and IVANOV, S. (2001), "A Course in Metric Geometry," in *Graduate Studies in Mathematics 33*, Providence RI: AMS.

CARLSSON, G., and MÉMOLI, F. (2008), "Persistent Clustering and a Theorem of J. Kleinberg," ArXiv:0808.2241.

CARLSSON, G., and MÉMOLI, F. (2010), "Characterization, Stability and Convergence of Hierarchical Clustering Methods," *Journal of Machine Learning Research, 11*, 1425–1470.

FISHER, L., and VAN NESS, J. (1971), "Admissible Clustering Procedures," *Biometrika, 58*, 91–104.

GROMOV, M. (2007), *Metric Structures for Riemannian and Non-Riemannian Spaces*, Modern Birkhäuser Classics, Boston MA: Birkhäuser Boston Inc.

HUGHES, B. (2004), "Trees and Ultrametric Spaces: A Categorical Equivalence," *Advances in Mathematics, 189*, 148–191.

KLEINBERG, J.M. (2002), "An Impossibility Theorem for Clustering," in *Advances in Neural Information Processing Systems 15* (NIPS 2002), eds. S. Becker, S. Thrun, and K. Obermayer, Cambridge MA: MIT Press, pp. 446–453.

MARTÍNEZ-PÉREZ, A., and MORÓN M.A. (2009), "Uniformly Continuous Maps Between Ends of $\mathbb{R}$-trees", *Mathematische Zeitschrift, 263(3)*, 583–606.

MARTÍNEZ-PÉREZ, A. (2013), "A Density-Sensitive Hierarchical Clustering Method," arXiv:1210.6292v2 [cs.LG].

MARTÍNEZ-PÉREZ, A. (2015), "Gromov-Hausdorff Stability of Linkage-Based Hierarchical Clustering Methods," *Advances in Mathematics*, to appear.

MURTAGH, F. (1983), "A Survey of Recent Advances in Hierarchical Clustering Algorithms", *The Computer Journal, 26(4)*, 354–359.

WISHART, D. (1969), “Mode Analysis: A Generalization of Nearest Neighbor Which Reduces Chaining Effects”, in *Numerical Taxonomy*, ed. A.J. Cole, New York: Academic Press, pp. 282–311.

ZADEH, R.B., and BEN-DAVID, S. (2009), “A Uniqueness Theorem for Clustering”, *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, arXiv:1205.2600 [cs.LG].