# Characterization, Stability and Convergence of Hierarchical Clustering Methods

**Gunnar Carlsson**                                   GUNNAR@MATH.STANFORD.EDU
*Department of Mathematics,*
*Stanford University,*
*Stanford, CA 94305*

**Facundo Mémoli**∗                                   MEMOLI@MATH.STANFORD.EDU
*Department of Mathematics,*
*Stanford University,*
*Stanford, CA 94305*

**Editor:** Ulrike von Luxburg

## Abstract

We study hierarchical clustering schemes under an axiomatic view. We show that within this framework, one can prove a theorem analogous to one of J. Kleinberg (Kleinberg, 2002), in which one obtains an existence and uniqueness theorem instead of a non-existence result. We explore further properties of this unique scheme: stability and convergence are established. We represent dendrograms as ultrametric spaces and use tools from metric geometry, namely the Gromov-Hausdorff distance, to quantify the degree to which perturbations in the input metric space affect the result of hierarchical methods.

**Keywords:** Clustering, hierarchical clustering, stability of clustering, Gromov-Hausdorff distance.

---

∗. Corresponding author.

# Contents

# 1. Notation

| Symbol | Meaning |
| --- | --- |
| $\mathbb{R}$ | Real numbers. |
| $\mathbb{R}^d$ | $d$-dimensional Euclidean space. |
| $\mathbb{N}$ | Natural numbers. |
| $((a))$ | A square symmetric matrix with elements $a_{ij}$ which are usually distances. |
| $(X, d)$ | Metric space $X$ with metric $d$, page 9. |
| $(X, u)$ | Ultrametric space $X$ with ultrametric $u$, page 9. |
| $\mathcal{X}, \mathcal{X}_n$ | Collection of all finite (resp. $n$ point) metric spaces, page 9. |
| $\mathcal{U}, \mathcal{U}_n$ | Collection of all finite (resp. $n$ point) ultrametric spaces, page 9. |
| $\mathcal{C}(X)$ | Collection of all non-empty subsets of the set $X$, page 9. |
| $\mathcal{U}(X)$ | Collection of all ultrametrics over the finite set $X$, page 9. |
| $\mathcal{P}(X)$ | Collection of all partitions of the finite set $X$, page 9. |
| $\Pi, \mathcal{B}, \mathcal{A}$ | A partition of a finite set and blocks of that partition, respectively, page 9. |
| $\sim, [a], A\backslash \sim$ | An equivalence relation, the equivalence class of a point and the quotient space, page 9. |
| $\sim_r$ | An equivalence relation with a parameter $r \geqslant 0$, page 10. |
| $S^{k-1}(r)$ | Sphere of radius $r$ and dimension $k-1$ embedded in $\mathbb{R}^k$, page 9. |
| $\mathcal{L}(W)$ | Maximal metric $\leqslant W$, page 10. |
| $\theta : [0, \infty) \to \mathcal{P}(X)$ | A dendrogram over the finite set $X$, 11. |
| $\mathcal{D}(X)$ | Collection of all dendrograms over the finite set $X$, page 11. |
| $\theta^*$ | Dendrogram over the finite set $X$ arising from $\sim_r$, 13. |
| $\ell^{\mathrm{SL}}, \ell^{\mathrm{CL}}, \ell^{\mathrm{AL}}$ | Linkage functions, page 15. |
| $\theta^{\mathrm{SL}}, \theta^{\mathrm{AL}}, \theta^{\mathrm{CL}}$ | Dendrograms arising from linkage functions, 15. |
| $\mathfrak{T}$ | A hierarchical clustering method seen as a map $\mathfrak{T} : \mathcal{X} \to \mathcal{U}$, page 23. |
| $\mathfrak{T}^*$ | A HC method arising from the maximal sub-dominant ultrametric, page 24. |
| $u_\theta$ | An ultrametric obtained from the dendrogram $\theta$, page 22. |
| $\theta^u$ | A dendrogram obtained from the ultrametric $u$, page 23. |
| $\Psi$ | A bijective map between $\mathcal{D}(X)$ and $\mathcal{U}(X)$, page 21. |
| $\Delta_n$ | Metric space isometric to an $n$ point unit simplex, page 29. |

$L_n$            Metric space isometric to $n$ points on a line, page 29.

$d_{\mathcal{H}}^Z$            Hausdorff distance between subsets of the metric space $Z$, page 37.

$\mathfrak{T}^{\mathrm{SL}}, \mathfrak{T}^{\mathrm{CL}}, \mathfrak{T}^{\mathrm{CL}}$            Standard linkage based HC methods seen as maps from $\mathcal{X}$ to $\mathcal{U}$, page 24.

$\mathrm{dis}(f), \mathrm{dis}(f, g)$            Distortion of a map $f$ and joint distortion of a pair of maps $f$ and $g$, page 27.

$d_{\mathcal{GH}}$            Gromov-Hausdorff distance between metric spaces, pages 27, 40.

$\mathrm{sep}(X)$            Separation of the metric space $X$, page 10.

$\mathrm{diam}\,(X)$            Diameter of the metric space $X$, page 9.

$P_n$            All the $n!$ permutations of elements of the set $\{1, \ldots, n\}$.

$\Gamma_{X,Y}$            A function used to measure metric distortion, page 40.

$(X, d, \mu)$            An mm-space, $(X, d)$ a compact metric space, $\mu$ a Borel probability measure, page 43.

$\mathrm{supp}\,[\mu]$            Support of the probability measure $\mu$, page 43.

$\mathbf{P}_\mu$            Probability with respect to the law $\mu$.

## 2. Introduction

Clustering techniques play a very central role in various parts of data analysis. They can give important clues to the structure of data sets, and therefore suggest results and hypotheses in the underlying science. Many of the interesting methods of clustering available have been applied to good effect in dealing with various datasets of interest. However, despite being one of the most commonly used tools for unsupervised exploratory data analysis, and despite its extensive literature, very little is known about the theoretical foundations of clustering methods. These points have been recently made prominent by Ben-David and von Luxburg in (von Luxburg and Ben-David, 2005; Ben-David et al., 2006).

The general question of which methods are "best", or most appropriate for a particular problem, or how significant a particular clustering is has not been addressed too frequently. This lack of theoretical guarantees can be attributed to the fact that many methods involve particular choices to be made at the outset, for example how many clusters there should be, or the value of a particular thresholding parameter. In addition, some methods depend on artifacts in the data, such as the particular order in which the observations are listed.

In (Kleinberg, 2002), J. Kleinberg proves a very interesting impossibility result for the problem of even defining a clustering scheme with some rather mild invariance properties. He also points out that his results shed light on the trade-offs one has to make in choosing clustering algorithms.

**Standard clustering methods** take as input a finite metric space $(X, d)$ and output a partition of $X$. Let $\mathcal{P}(X)$ denote the set of all possible partitions of the set $X$. Kleinberg (Kleinberg, 2002) discussed this situation in an axiomatic way and identified a set of reasonable properties of standard clustering schemes, namely, scale invariance, richness and consistency. Fix a standard clustering method $f$ and a metric space $(X, d)$ and let $f(X, d) = \Pi \in \mathcal{P}(X)$. Kleinberg identified the following desirable properties of a clustering scheme:

- **Scale Invariance**: For all $\alpha > 0$, $f(X, \alpha \cdot d) = \Pi$.

- **Richness**: Fix any finite set $X$. Then for all $\Pi \in \mathcal{P}(X)$, *there exists* $d_\Pi$, a metric on $X$ s.t. $f(X, d_\Pi) = \Pi$.

- **Consistency**: Let $\Pi = \{B_1, \ldots, B_\ell\}$. Let $\hat{d}$ be any metric on $X$ s.t.

    1. for all $x, x' \in B_\alpha$, $\hat{d}(x, x') \leqslant d(x, x')$ and
    2. for all $x \in B_\alpha$, $x' \in B_{\alpha'}$, $\alpha \neq \alpha'$, $\hat{d}(x, x') \geqslant d(x, x')$.

    Then, $f(X, \hat{d}) = \Pi$.

He then proved, in the same spirit of Arrow's impossibility theorem, that no clustering scheme satisfying these conditions simultaneously can exist.

**Theorem 1 ((Kleinberg, 2002))** *There exists no clustering algorithm that satisfies scale invariance, richness and consistency.*

Then, in particular, Kleinberg's axioms rule out single, average and complete linkage (standard) clustering. Clusters in any of these three methods can be obtained by first constructing a hierachical decomposition of space (such as those provided by hierarchical clustering methods) and then selecting the partition that arises at a given, fixed, threshold.

A natural question is whether Kleinberg's impossibility results still holds when one admits clustering schemes that do not try to return a fixed partition of a space, but are allowed to return a *hierarchical decomposition*.

Furthermore, datasets can exhibit multiscale structure and this can render standard clustering algorithms inapplicable in certain situations, see Figure 1. This further motivates the use of **Hierarchical clustering methods.** Hierarchical methods take as input a finite metric space $(X, d)$ and output a hierarchical family of partitions of $X$.
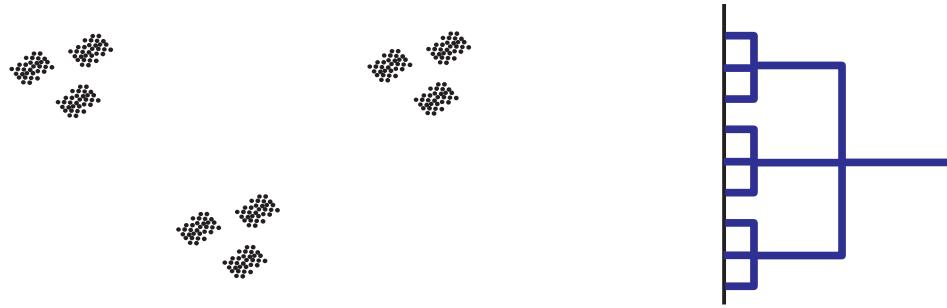


Figure 1: Dataset with multiscale structure. Any standard clustering algorithm will fail to capture the structure of the data.

These hierarchical families of partitions that constitute the output of hierarchical methods receive the name of *dendrograms*. Dendrograms come in two versions: *proximity* and *threshold* dendrograms. These two types of dendrograms differ in whether they retain some proximity information about the underlying clusters that they represent or not: proximity dendrograms do retain such information whereas threshold dendrograms do not. Practicioners of statistical data analysis seem to work almost exclusively with proximity dendrograms. For this reason we opt to carry out our analysis under the model that hierarchical methods take as input a finite metric space $X$ and output a proximity dendrogram over $X$, see Remark 3.

6

We remind the reader that we are using the term standard clustering methods to refer to procedures that take a finite metric space as input and output a fixed single partition of the metric space.

In a similar spirit to Kleinberg's theorem, we prove in Theorem 18 that in the context of hierarchical methods, one obtains *uniqueness* instead of non-existence. We emphasize that our result can be interpreted as a *relaxation* of the theorem proved by Kleinberg, in the sense that allowing clustering schemes that output a nested family of partitions in the form of a proximity dendrogram, instead of a fixed partition, removes the obstruction to existence. The unique HC method characterized by our theorem turns out to be single linkage hierarchical clustering.

We stress the fact that our result assumes that outputs of hierarchical methods are proximity dendrograms, whereas Kleinberg's Theorem applies to flat/standard clustering, a situation in which the output contains no proximity information between clusters.

In order to state and prove our results we make use of the well known **equivalent representation** of dendrograms, the output of HC methods, using *ultrametrics*. This already appears in the book Jardine and Sibson (Jardine and Sibson, 1971) and in the classical work of Hartigan and others, see (Hartigan, 1985), (Jain and Dubes, 1988, §3.2.3) and references therein.

In recent years, the theme of studying the properties of metrics with prescribed generalized curvature properties has been studied intensively. In particular, the work of Gromov (Gromov, 1987) has been seminal, and many interesting results have been proved concerning objects other than metric spaces, such as finitely generated groups, depending on these methods. The curvature conditions can be formulated in terms of properties of triangles within the metric spaces, and the most extreme of these properties is that embodied in ultrametric spaces. A second idea of Gromov's is to make the collection of all metric spaces into its own metric space, and the resulting metric gives a very useful and natural way to distinguish between metric spaces (Gromov, 2007). This metric is known as the Gromov-Hausdorff distance and its restriction to the subclass of ultrametric spaces is therefore a very natural object to study.

**Stability**   of some kind is clearly a desirable property of clustering methods and, therefore, a point of interest is studying whether results obtained by a given clustering algorithm are *stable* to perturbations in the input data. Since input data are modelled as finite metric spaces, and the output of hierarchical methods can be regarded as finite ultrametric spaces, the Gromov-Hausdorff distance provides a natural tool for studying *variability* or *perturbation* of the inputs and outputs of hierarchical clustering methods.

After observing in §4.6 that average and complete linkage clustering are not stable in the metric sense alluded to above, we prove in Proposition 26 that single linkage does enjoy a kind of stability:

**Proposition 2** *Let $(X, d_X)$ and $(Y, d_Y)$ be two finite metric spaces and let $(X, u_X)$ and $(Y, u_X)$ be the two (finite metric ultrametric spaces) corresponding outputs yielded by single linkage HC. Then,*

$$d_{\mathcal{GH}}\big((X, u_X), (Y, u_Y)\big) \leqslant d_{\mathcal{GH}}\big((X, d_X), (Y, d_Y)\big).$$

*Here, $d_{\mathcal{GH}}$ stands for the Gromov-Hausdorff distance.*
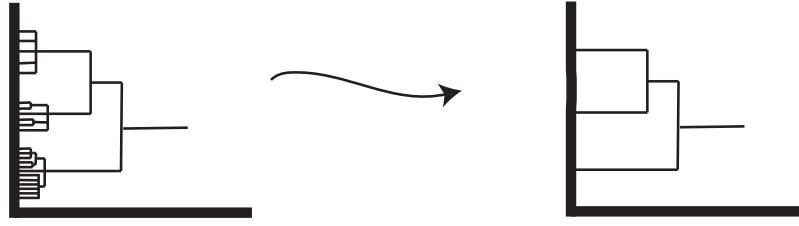


Figure 2: Convergence of dendrograms. We formalize this concept by equivalently representing dendrogram as ultrametrics and then computing the Gromov-Hausdorff distance between the resulting metrics. We prove in Theorem 30 that by taking increasingly many i.i.d. samples from a given probability distribution $\mu$ on a metric space,then with probability 1 one recovers a multiscale representation of the supprt of $\mu$.

This result is very important for the convergence theorems which we prove in the later parts of the paper. These results describe in a very precise way the fact that for compact metric spaces $X$, the results of clustering the finite subsets of $X$ yields a collection of dendrograms which ultimately converge to the dendrogram for $X$. In order for this to happen, one needs the metric on the ultrametric spaces as well as the behavior of the clustering construction on the Gromov-Hausdorff distance, which is what Proposition 2 does. The issue of stability is further explored in §6.

**Probabilistic convergence.** Finally, in Theorem 30 we also prove that for random i.i.d. observations $\mathbb{X}_n = \{x_1, \ldots, x_n\}$ with probability distribution $\mu$ compactly supported in a metric space $(X, d)$, the result $(\mathbb{X}_n, u_{\mathbb{X}_n})$ of applying single linkage clustering to $(\mathbb{X}_n, d)$ **converges almost surely** in the Gromov-Hausdorff sense to an ultrametric space that recovers the multiscale structure of the *support* of $\mu$, see Figure 20. This can be interpreted as a refinement of a previous observation (Hartigan, 1985) that SLHC is insensitive to the distribution of mass of $\mu$ in its support.

**Organization of the paper.** This paper is organized as follows: §1 provides a list of all the notation defined and used throughout the paper; §3 introduces the terminology and basic concepts that we use in our paper; §4.2 reviews hierarchical clustering

methods in general; §4.3 discusses the representation of dendrograms as ultrametric spaces and establishes the equivalence of both repersentations; and §4.5 delves into the issue of constructing a notion of distance between dendrograms which is based in the equivalence of dendrograms and ultrametrics; §4.6 comments on issues pertaining to the theoretical properties of HC methods. In §5 we present our characterization result, Theorem 18, for SL in a spirit similar to the axiomatic treatment of Kleinberg. We delve into the stability and convergence questions of SL in §6, where we introduce all the necessary concepts from Metric Geometry. Proposition 26 and Theorem 28 contain our results for the deterministic case. In §6.3 we prove a probabilistic convergence result Theorem 30 that hinges on a general sampling theorem for measure metric spaces, Theorem 34. Finally, we conclude the paper with a discussion on future directions.

For clarity of exposition, we have chosen to move most of the proofs in this paper to and appendix. The ones which remain in the main text are intended to provide intuition which would not otherwise be there.

## 3. Background and notation

A **metric space** is a pair $(X, d)$ where $X$ is a set and $d : X \times X \to \mathbb{R}^+$ satisfies

1. For all $x, x' \in X$, $d(x', x) = d(x, x') \geqslant 0$ and $d(x, x') = 0$ if and only if $x = x'$.

2. For all $x, x', x'' \in X$, $d(x, x'') \leqslant d(x, x') + d(x', x'')$.

A metric space $(X, u)$ is an **ultrametric space** if and only if for all $x, x', x'' \in X$,

$$\max(u(x, x'), u(x', x'')) \geqslant u(x, x''). \tag{3.1}$$

Ultrametric spaces are therefore metric spaces which satisfy a stronger type of triangle inequality. It is interesting to observe that this ultrametric triangle inequality (3.1) implies that all triangles are *isosceles*.[1]

Notice that by iterating the ultrametric property one obtains that if $x_1, x_2, \ldots, x_k$ is any set of $k$ points in $X$, then

$$\max \big( u(x_1, x_2), u(x_2, x_3), \ldots, u(x_{k-1}, x_k) \big) \geqslant u(x_1, x_k).$$

For a fixed finite set $X$, we let $\mathcal{U}(X)$ denote the collection of all ultrametrics on $X$. For $n \in \mathbb{N}$ let $\mathcal{X}_n$ (resp. $\mathcal{U}_n$) denote the collection of all metric spaces (resp. ultra-metric spaces) with $n$ points. Let $\mathcal{X} = \bigsqcup_{n \geqslant 1} \mathcal{X}_n$ denote the collection of all finite metric spaces and $\mathcal{U} = \bigsqcup_{n \geqslant 1} \mathcal{U}_n$ all finite ultrametric spaces. For $(X, d) \in \mathcal{X}$ let

$$\operatorname{sep}(X, d) := \min_{x \neq x'} d(x, x') \quad \text{and} \quad \operatorname{diam}(X, d) := \max_{x, x'} d(x, x')$$

---

1. Indeed, assume that all sides $a, b, c$ of a triangle in a given ultrametric space are different. Then, without loss of generality $a > b > c$. But then, $a > \max(a, b)$ which violates (3.1). Hence, there must be at least two equal sides in every triangle in an ultrametric space.

be the *separation* and the *diameter* of $X$, respectively.

We now recall the definition of an **equivalence relation**. Given a set $A$, a *binary relation* is a subset $S \subset A \times A$. One says that $a$ and $a'$ are *related* and writes $a \sim a'$ whenever $(a, a') \in S$. $S$ is called an *equivalence relation* if and only if for all $a, b, c \in A$, all the following hold true:

- Reflexivity: $a \sim a$.

- Symmetry: if $a \sim b$ then $b \sim a$.

- Transitivity: if $a \sim b$ and $b \sim c$ then $a \sim c$.

The *equivalence class* of $a$ under $\sim$, denoted $[a]$, is defined as all those $a'$ which are related to $a$: $[a] = \{a' \in A, \text{s.t. } a' \sim a\}$. Finally, the *quotient space* $A \backslash \sim$ is the collection of all equivalence classes: $A \backslash \sim := \{[a], a \in A\}$.

We now construct our first example which will be crucial in our presentation.

**Example 1 ($r$-equivalence)** *Given a finite metric space $(X, d)$ and $r \geqslant 0$ we say that points $x, x' \in X$ are* **r-equivalent** *(denoted $x \sim_r x'$) if and only if there exists points $x_0, x_1, \ldots, x_t \in X$ with $x_0 = x$, $x_t = x'$ and $d(x_i, x_{i+1}) \leqslant r$ for $i = 0, \ldots, t - 1$. It is easy to see that $\sim_r$ is indeed an equivalence relation on $X$.*

*This definition embodies the simple idea of partitioning a finite metric space into path connected components, where the granularity of this partitioning is specified by the parameter $r \geqslant 0$, see Figure 1.*

For a finite set $X$, and a symmetric function $W : X \times X \to \mathbb{R}^+$ let $\mathcal{L}(W)$ denote the *maximal metric* on $X$ less than of equal to $W$ (Bridson and Haefliger, 1999), i.e.

$$\mathcal{L}(W)(x, x') = \min \left\{ \sum_{i=0}^{m-1} W(x_i, x_{i+1}) \, \big| \, x = x_0, \ldots, x_m = x \right\}$$

for $x, x' \in X$.

For a finite set $X$, we let $\mathcal{C}(X)$ denote the collection of all non-empty subsets of $X$. By $\mathcal{P}(X)$ we denote the set of all partitions of $X$. For a given partition $\Pi \in \mathcal{P}(X)$ we refer to each $\mathcal{B} \in \Pi$ as a *block* of $\Pi$. For partitions $\Pi, \Pi' \in \mathcal{P}(X)$, we say that $\Pi$ is *coarser* than $\Pi'$, or equivalently that $\Pi'$ is a *refinement* of $\Pi$, if for every block $\mathcal{B}' \in \Pi'$ there exists a block $\mathcal{B} \in \Pi$ s.t. $\mathcal{B}' \subset \mathcal{B}$.

For $k \in \mathbb{N}$ and $r > 0$ let $S^{k-1}(r) \subset \mathbb{R}^k$ denote the $(k-1)$ dimensional sphere with radius $r$. By $((a))$ we will denote a matrix of elements $a_{ij}$.
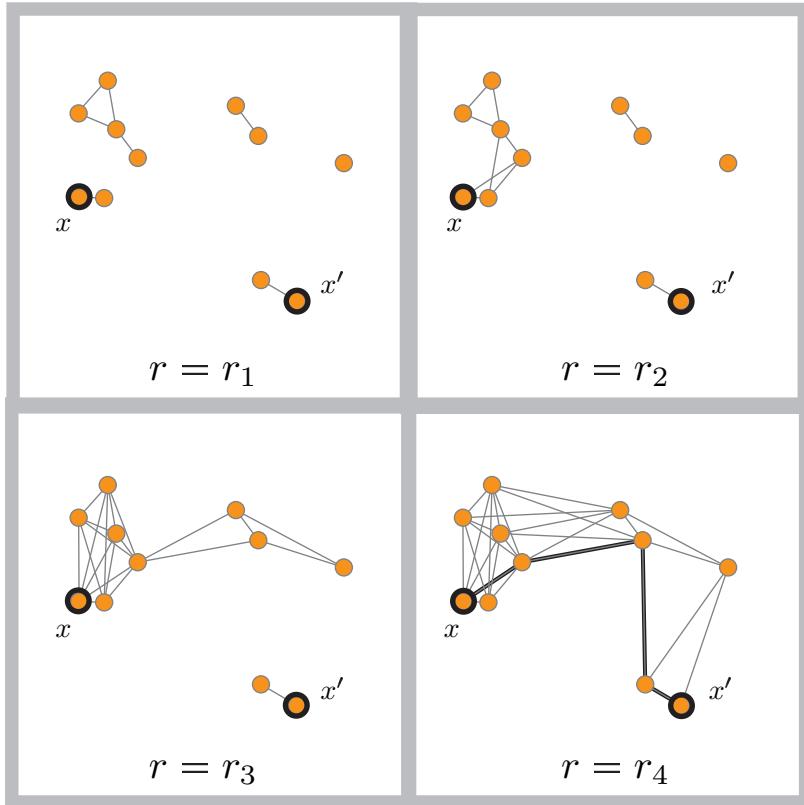
Figure 3: Illustration of the equivalence relation $\sim_r$. A finite metric space $X$ is specified by the points in orange which are endowed with the Euclidean distance. This construction can be understood as allowing the creation of edges joining two points whenever the distance between them does not exceed $r$. Then, two points $x$ and $x'$ in black are deemed $r$-equivalent if one can find a sequence of edges on the resulting graph connecting $x$ to $x'$. From left to right and top to bottom we show the resulting graph one obtains for 4 increasing values of $r$. The points $x$ and $x'$ are not $r$-equivalent when $r = r_1, r_2$ or $r_3$, but they are $r_4$-equivalent.

## 4. Hierarchical clustering: formulation

In this section we formaly define hierarchical clustering methods as maps that assign a dendrogram to a finite metric space. First, in §4.1 formalize the standard concept of dendrogram; then, in §4.2 we present a formal treatment of HC methods which emphasizes the need for a formulation that is insensitive to arbitrary choices such as the labels given to the points in the dataset. Finally, in §4.3 we prove that the collection of all dendrograms over a finite set is in a one to one correspondence with the collection of all ultrametrics on this set. We then redefine HC methods as maps

from the collection of finite metric spaces to the collection all finite ultrametric spaces. This change in perspective permits a natural formulation and study of the *stability* and *convergence* issues in later sections of the paper. In particular, in §4.5, we discuss the construction of notions of *distance between dendrograms* by appealing to the ultrametric representation. These notions are instrumental for the arguments in §6.

Finally, in §4.6, we disgress on some critiques to the classical HC methods. The situation with HC methods is seemingly paradoxical in that SL is the one that seems to enjoys the best theoretical properties while CL and AL, despite exhibiting some undesirable behaviour, are the usual choices of practicioners.

## 4.1 Dendrograms

A **dendrogram** over a finite set $X$ is defined to be nested family of partitions, usually represented graphically as a rooted tree. Dendrograms are meant to represent a hierarchical decompositions of the underlying set $X$, such as those that are produced by hierarchical clustering algorithms, and therefore the nested family of partitions provided must satisfy certain conditions. We formally describe dendrograms as pairs $(X, \theta)$, where $X$ is a finite set and $\theta : [0, \infty) \to \mathcal{P}(X)$. The parameter of $\theta$ usually represents a certain notion of *scale* and it is reflected in the height of the different levels, see Figure 4.1. We require that $\theta$ satisfies:

1. $\theta(0) = \{\{x_1\}, \ldots, \{x_n\}\}$. This condition means that the initial decomposition of space is the finest possible: the space itself.

2. There exists $t_0$ s.t. $\theta(t)$ is the *single block partition* for all $t \geqslant t_0$. This condition encondes the fact that for large enough $t$, the partition of the space becomes trivial.

3. If $r \leqslant s$ then $\theta(r)$ *refines* $\theta(s)$. This condition ensures that the family of partitions provided by the dendrogram is indeed nested.

4. For all $r$ there exists $\varepsilon > 0$ s.t. $\theta(r) = \theta(t)$ for $t \in [r, r+\varepsilon]$. (technical condition)

Let $\mathcal{D}(X)$ denote the collection of all possible dendrograms over a given finite set $X$. When understood from context, we will omit the first component of a dendrogram $(X, \theta) \in \mathcal{D}(X)$ and refer to $\theta$ as a dendrogram over $X$.

**Remark 3 (About our definition of dendrogram)** *Our definition coincides with what Jain and Dubes call* **proximity dendrograms** *in (Jain and Dubes, 1988, §3.2). We stress that we view the parameter t in our definition as part of the information about the hierarchical clustering. Jain and Dubes also discuss a simpler version of dendrograms, which they call* threshold dendrograms, *which retain merely the order in which succesive partitions are created. These of course can be viewed as functions*
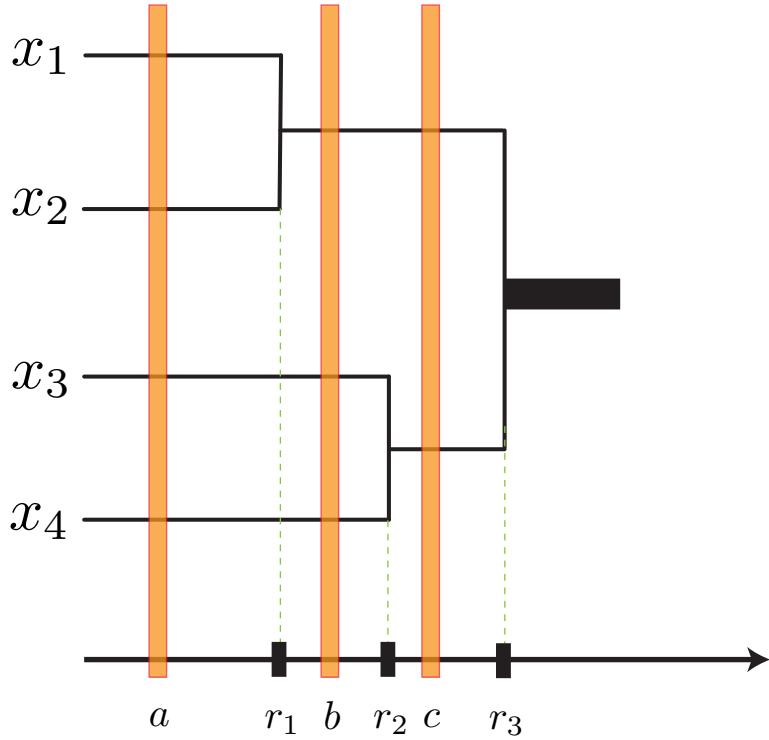
Figure 4: A graphical representation of a dendrogram over the set $X = \{x_1, x_2, x_3, x_4\}$. Let $\theta$ denote the dendrogram. Notice for example that $\theta(a) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}\}$; $\theta(b) = \{\{x_1, x_2\}, \{x_3\}, \{x_4\}\}$; $\theta(c) = \{\{x_1, x_2\}, \{x_3, x_4\}\}$; and $\theta(t) = \{x_1, x_2, x_3, x_4\}$ for any $t \geqslant r_3$.

from $\mathbb{N}$ into $\mathcal{P}(X)$ satisfying the constraints (1), (2) and (3) above, instead of having the domain $[0, \infty)$.

It seems that proximity dendrograms are the type of dendrograms that are most often employed by practicioners and statisticians, see for example the dendrograms provided by the statistical software R (`http://www.r-project.org/`) and by matlab's statistics toolbox (`http://www.mathworks.com/products/statistics/`, 2009), whereas threshold dendrograms are more popular in the Machine Learning and Computer Science communities.

Usually, Hierarchical Clustering methods are defined as those maps that to each finite metric space $(X, d)$ assign a dendrogram over $X$.

Using the definitions above we now construct our first example.

**Example 2** *For each finite metric space $(X, d)$ let $(X, \theta^*) \in \mathcal{D}(X)$ be given by $\theta^*(r) = X \backslash \sim_r$. In other words, for each $r \geqslant 0$, $\theta^*(r)$ returns the partition of $X$ into $\sim_r$-equivalence classes. Recall (Example 1) that two points $x$ amd $x'$ are $\sim_r$ equivalent if and only if one can find a sequence of points $x_0, x_1, \ldots, x_k$ s.t. the first of them is $x$ and the last one is $x'$ and all the hops are smaller than $r$: $\max_i d_X(x_i, x_{i+1}) \leqslant r$. We will see below that this definition coincides with single linkage hierarchical clustering. See Figure 2 for an illustration of this concept.*
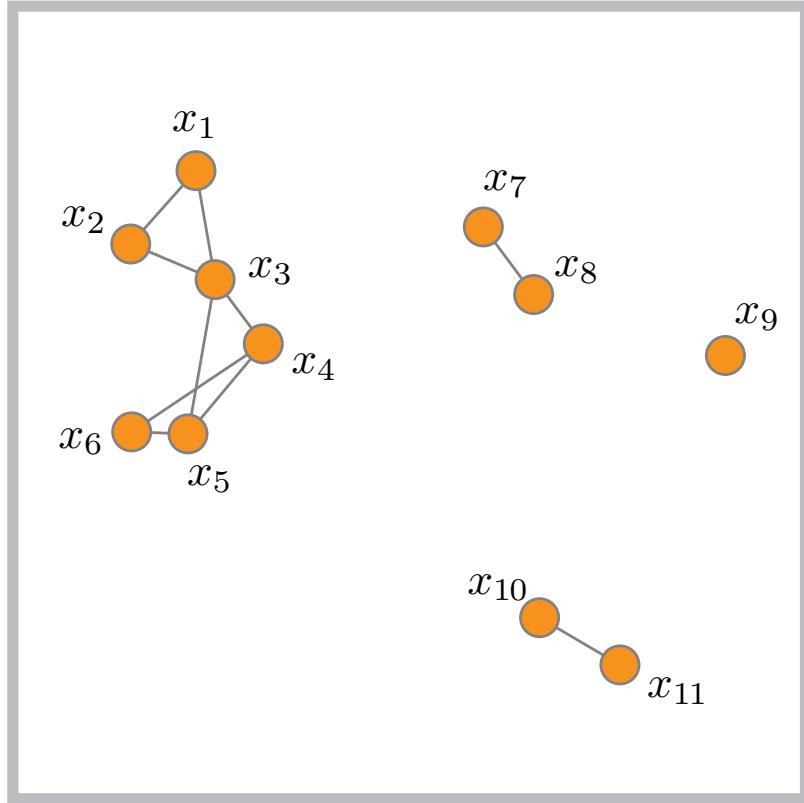


Figure 5: For the same finite metric space $X$ of Example 1 and the value $r = r_2$, $X \backslash \sim_{r_2} = \big\{ \{x_1, x_2, x_3, x_4, x_5, x_6\}, \{x_7, x_8\}, \{x_9\}, \{x_{10}, x_{11}\} \big\}$, that is, $\sim_{r_2}$ splits $X$ into four path connected components.

*In order to build up intuition about our definitions, we prove that $(X, \theta^*)$ is indeed a dendrogram. Since $X$ is a metric space, $x \sim_0 x'$ if and only if $x = x'$. Thus condition (1) above is satisfied. Clearly, for $t \geqslant diam(X, d)$, $x \sim_t x'$ for all $x, x'$, and thus condition (2) holds. Fix $0 \leqslant r \leqslant s$ and let $\mathcal{B}$ be a maximal connected component of $\theta^*(r)$ and let $x, x' \in \mathcal{B}$. Then, by definition of $\theta^*(r)$, $x \sim_r x'$. But it follows from the definition of $\sim_r$ that if $x \sim_r x'$, then $x \sim_s x'$ for all $s \geqslant r$. Hence, $x, x'$ are in the same block of $\theta^*(s)$ and condition (3) holds. Condition (4) holds since clearly $\theta^*$ is right continuous, has finitely many discontinuity points, and is piecewise constant.*

14

We now need to discuss a formal description of agglomerative HC methods.

## 4.2 A general description of Agglomerative Hierarchical Clustering methods

In this section we give a description of agglomerative HC methods that is suitable for our theoretical analyses. Standard algorithmic descriptions of HC methods typically make the assumption that in the merging process there are only two points at minimal linkage value of eachother. For example, the formulation of Lance and Williams (1967) does not specifically explain how to deal with the case when more than two points are candidates for merging. In practice one could argue that if at a certain stage, say, three points are at minimal linkage value of eachother, then one could proceed to merge them two at a time, according to some predefined rule that depends on the indices of the points.

Whereas this *tie breaking* strategy seems reasonable from a computational point of view, it invariably leads to dendrograms that depend on the ordering of the points. This is no doubt an undesirable feature that can be translated into, for example, that the results of the clustering methods depend on the order in which the data samples were obtained. Single linkage HC is exempted from this problem however, because of the fact that at each stage only *minimal distances* are taken into account. In contrast, complete and average linkage will produce results that do not behave well under reordering of the points.

The problems arising from ad hoc tie breaking are often not even mentioned in books on clustering. A notable exception is the book (Jain and Dubes, 1988), especially Section §3.2.6, where the reader can find a careful exposition of these issues.

Below, we formulate HC methods in a way that is independent of these extraneous features. In order to do so , we need to have some kind of **invariance** in the formulation. More precisely, let $(X, d_X)$ be the input metric space, where we assume that $X = \{1, \ldots, n\}$ consists of exactly $n$ points. Write $(X, \theta_X)$ is the output dendrogram of a given HC method applied to $(X, d_X)$. Let $\pi$ be a permutation of the indices $\{1, 2, \ldots, n\}$, and $(Y, d_Y)$ be the metric space with points $\{1, \ldots, n\}$ and permuted metric: $d_Y(i, j) := d_X(\pi_i, \pi_j)$ for all $i, j \in \{1, \ldots, n\}$; further, denote by $(Y, \theta_Y)$ the output dendrogram of the same HC method applied on $(Y, d_Y)$. Then, we require that for all permutations $\pi$, the result of computing the dendrogram first and then permuting the result **is the same** as the result of first permuting the input distance matrix and then computing the output dendrogram:

$$\pi \circ \theta_X(t) = \theta_Y(t), \text{ for all } t \geqslant 0. \tag{4.2}$$

Formally, the action of a permutation $\pi$ over a partition (such as $\theta_X(t)$) above must be understood in the following sense: if $P = \{\mathcal{B}_1, \ldots, \mathcal{B}_r\}$ is a partition of $\{1, 2, \ldots, n\}$, then $\pi \circ P$ is the partition with blocks $\{\pi \circ \mathcal{B}_i, \ 1 \leqslant i \leqslant r\}$, where in turn $\pi \circ \mathcal{B}_i$ consists of all those indices $\pi_j$ for $j \in \mathcal{B}_i$.

We elaborate on this in the next example. We first recall the usual definition of CLHC, and then construct a simple metric space consisting of five points where this usual formulation of CL fails to exhibit invariance to permutations.

### 4.2.1 THE STANDARD FORMULATION OF COMPLETE LINKAGE HC

We assume $(X, ((d)))$ is a given finite metric space. In this example, we use the formulas for CL but the structure of the iterative procedure in this example is common to all HC methods (Jain and Dubes, 1988, Chapter 3). Let $\theta$ be the dendrogram to be constructed in this example.

1. Set $X_0 = X$ and $D_0 = ((d))$ and set $\theta(0)$ to be the partition of $X$ into singletons.

2. Search the matrix $D_0$ for the smallest non-zero value, i.e. find $\delta_0 = \text{sep}(X_0)$, and find all pairs of points $\{(x_{i_1}, x_{j_1}), (x_{i_2}, x_{j_2}) \ldots, (x_{i_k}, x_{j_k})\}$ at distance $\delta_0$ from eachother, i.e. $d(x_{i_\alpha}, x_{j_\alpha}) = \delta_0$ for all $\alpha = 1, 2, \ldots, k$, where one orders the indices s.t. $i_1 < i_2 < \ldots < i_k$.

3. Merge the <u>first pair of elements</u> in that list, $(x_{i_1}, x_{j_1})$, into a single group. The procedure now removes $\overline{(x_{i_1}, x_{j_1})}$ from the initial set of points and adds a point $c$ to represent the cluster formed by both: define $X_1 = (X_0 \backslash \{x_{i_1}, x_{j_1}\}) \cup \{c\}$. Define the dissimilarity matrix $D_1$ on $X_1 \times X_1$ by $D_1(a, b) = D_0(a, b)$ for all $a, b \neq c$ and $D_1(a, c) = D_1(c, a) = \max(D_0(x_{i_1}, a), D_0(x_{j_1}, a))$ (this step is the only one that depends on the choice corresponding to CL). Finally, set

$$\theta(\delta) = \{x_{i_1}, x_{j_1}\} \cup \bigcup_{i \neq i_1, j_1} \{x_i\}.$$

4. The construction of the dendrogram $\theta$ is completed by repeating the previous steps until all points have been merged into a single cluster.

**Example 3 (about the standard formulation of complete linkage)** *The crux of the problem lies in step 3 of the procedure outlined above. The choice to merge just the first pair of points in the list causes the procedure to not behave well under relabeling of the points in the sense of (4.2).*

*An explicit example is the following: consider the metric space $(\{1, 2, 3, 4, 5\}, ((d)))$ with five points and distance matrix*

$$((d)) = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ \left( \begin{array}{ccccc} 0 & 1 & 2 & 5 & 5 \\ 1 & 0 & 3 & 6 & 6 \\ 2 & 3 & 0 & 3 & 7 \\ 5 & 6 & 3 & 0 & 4 \\ 6 & 7 & 4 & 6 & 0. \end{array} \right) \end{array}$$

*This metric space arises from considering the graph metric on the graph depicted in Figure 6. Under CLHC (as defined in §4.2.1), and under the action of all possible permutations of the labels of its 5 points, this metric space produces 3 different **non-equivalent** dendrograms, see Figure 7. This is an undesirable feature, as discussed at length in (Jain and Dubes, 1988, Chapter 3).*
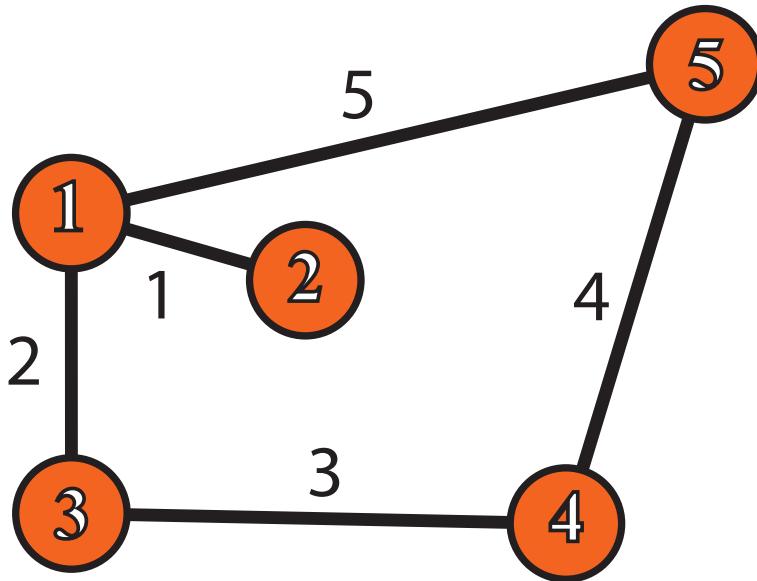


Figure 6:  A finite metric space that under permutations leads to different outputs of the usual CL HC algorithm, see text for details. The metric is defined by the graph distance on the weighted graph shown.

We now re-define general HC methods in a way that they satisfy (4.2).

### 4.2.2 THE PERMUTATION INVARIANT FORMULATION.

Here we consider the family of Agglomerative Hierarchical clustering techniques, (Jain and Dubes, 1988, Chapter 3). We define these by the recursive procedure described next. The main difference with §4.2.1 lies that in Step 3 we will allow for more than just two points into the same cluster and also, it could happen, for example, that four points $A, B, C, D$ merge into two different clusters $\{A, B\}$ and $\{C, D\}$ at the same time.

Let the finite metric space $(X, d)$ be given where $X = \{x_1, \ldots, x_n\}$ and let $L$ denote a family of *linkage functions* on $X$:

$$L := \{\ell : \mathcal{C}(X) \times \mathcal{C}(X) \to \mathbb{R}^+\}$$

17
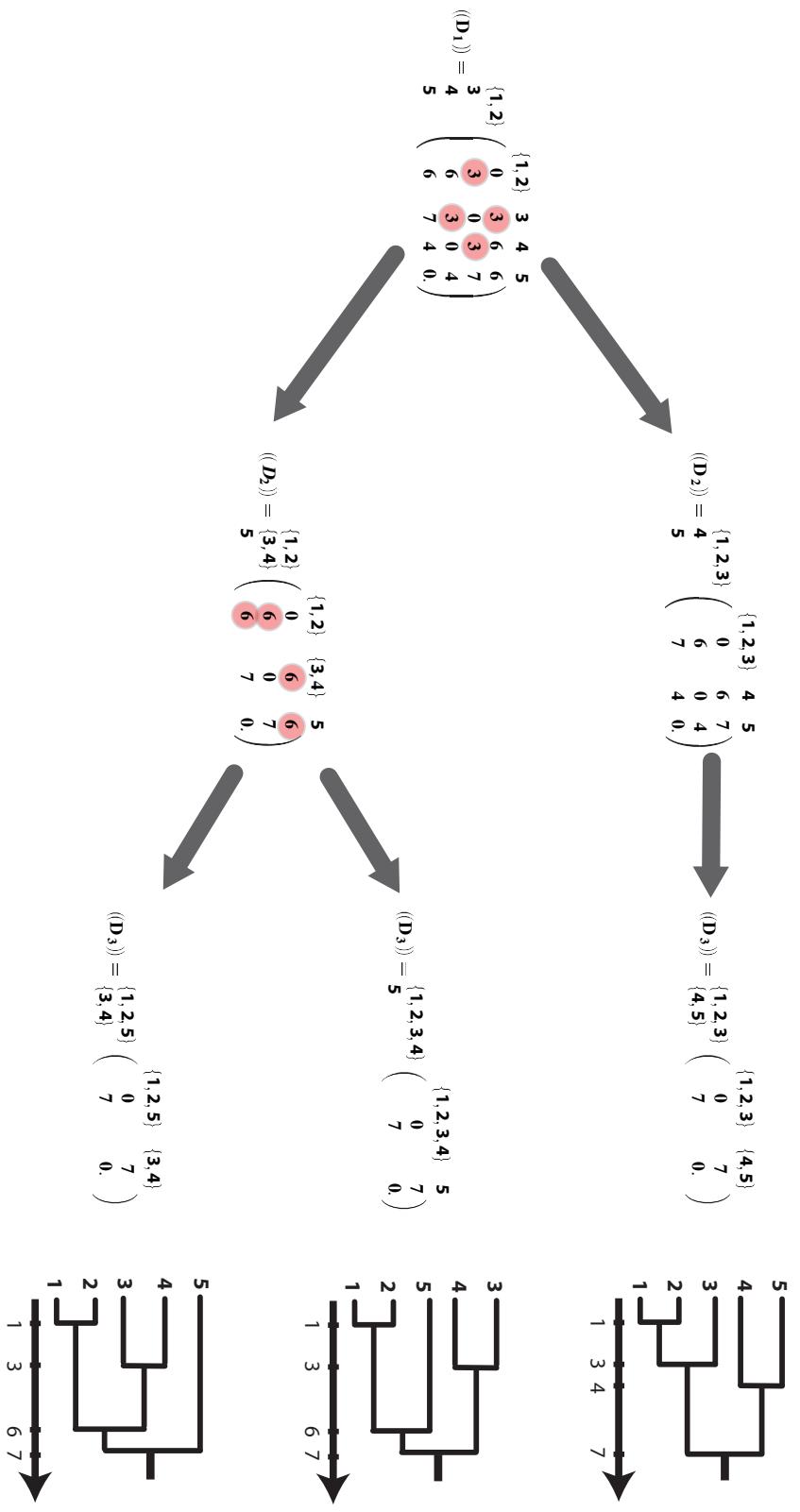
Figure 7: The five point metric space of Example 3 leads to 3 non-equivalent CL-dendrograms. Here we show the agglomerative procedure corresponding to the usual order dependent definition of CL HC, see text in Example 3 for details. Notice that in $D_1$, there are two pairs at minimal linkage value (this value is 3, in red circles): $(\{1,2\},3)$ on one hand, and $(3,4)$ on the other. The top row shows the agglomeration process that follows after merging $\{1,2\}$ with 3. The bottom row shows that when one instead chooses to merge 3 and 4 instead, then at the next step there is again ambiguity (represented again by the red circles). The three dendrograms shown on the right are the three possible outputs one finds when choosing different orders for the agglomeration processes. Clearly, these three dendrograms are not equivalent under permutations of the labels of their base points.

with the property all that $\ell \in L$ are bounded non-negative functions. These functions assign a non-negative value to each pair of non-empty subsets of $X$, and provide a certain measure of *distance* between two clusters. Let $\mathcal{B}, \mathcal{B}' \in \mathcal{C}(X)$, then, some possible standard choices for $\ell$ are:

- *Single linkage*: $\ell^{\mathrm{SL}}(\mathcal{B}, \mathcal{B}') = \min_{x \in \mathcal{B}} \min_{x' \in \mathcal{B}'} d(x, x')$;

- *Complete linkage*: $\ell^{\mathrm{CL}}(\mathcal{B}, \mathcal{B}') = \max_{x \in \mathcal{B}} \max_{x' \in \mathcal{B}'} d(x, x')$; and

- *Average linkage*: $\ell^{\mathrm{AL}}(\mathcal{B}, \mathcal{B}') = \frac{\sum_{x \in \mathcal{B}} \sum_{x' \in \mathcal{B}'} d(x, x')}{\#\mathcal{B} \cdot \#\mathcal{B}'}$.

- *Hausdorff linkage*: $\ell^{\mathrm{HL}}(\mathcal{B}, \mathcal{B}') = d_{\mathcal{H}}(\mathcal{B}, \mathcal{B}').$[2]

The permutation invariant formulation is as follows:

1. Fix $\ell \in L$. For each $R > 0$ consider the equivalence relation $\sim_{\ell,R}$ on blocks of a partition $\Pi \in \mathcal{P}(X)$, given by $\mathcal{B} \sim_{\ell,R} \mathcal{B}'$ if and only if there is a sequence of blocks $\mathcal{B} = \mathcal{B}_1, \ldots, \mathcal{B}_s = \mathcal{B}'$ in $\Pi$ with $\ell(\mathcal{B}_k, \mathcal{B}_{k+1}) \leqslant R$ for $k = 1, \ldots, s-1$.

2. Consider the sequences $R_1, R_2, \ldots \in [0, \infty)$ and $\Theta_1, \Theta_2, \ldots \in \mathcal{P}(X)$ given by $\Theta_1 := \{x_1, \ldots, x_n\}$, and recursively for $i \geqslant 1$ by $\Theta_{i+1} = \Theta_i / \sim_{\ell, R_i}$ where

$$R_i := \min\{\ell(\mathcal{B}, \mathcal{B}'); \mathcal{B}, \mathcal{B}' \in \Theta_i, \mathcal{B} \neq \mathcal{B}'\}.$$
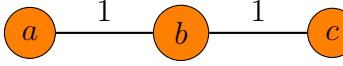
   Note that this process necessarily ends in finitely many steps. This construction reflects the fact that at step $i$ one agglomerates those clusters at distance $\leqslant R_i$ from eachother (as measured by the linkage function $\ell$). More than two clusters could be merged at any given step.

3. Finally, we define $\theta^\ell : [0, \infty) \to \mathcal{P}(X)$ by $r \mapsto \theta^\ell(r) := \Theta_{i(r)}$ where $i(r) := \max\{i | R_i \leqslant r\}$.

**Remark 4 (About our definition of HC methods)** *Note that, unlike the usual definition of agglomerative hierarchical clustering §4.2.1 ((Jain and Dubes, 1988, §3.2)), at each step of the inductive definition we allow for more than two clusters to be merged. Of course, the standard formulation can be recovered if one assumes that at each step $i$ of the algorithm, there exist only two blocks $\mathcal{B}$ and $\mathcal{B}'$ in $\Theta_i$ s.t. $R_i = \ell(\mathcal{B}, \mathcal{B}')$. Then, at each step, only two blocks will be merged.*

---

2. The Hausdorff distance is defined in Definition 21.

**Example 4** *Note for example that for the five point metric space in Example 3, the result of applying CL (according to the permutation invariant formulation) is the dendrogram in Figure 8 (a). It also follows, for example, that when applied to the metric space* $L_3 := \left( \{a, b, c\}, \left( \begin{smallmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{smallmatrix} \right) \right)$, *which can be represented by three points on a line:*  *, SL, AL and CL all yield the same dendrogram, which is shown in Figure 8 (b).*
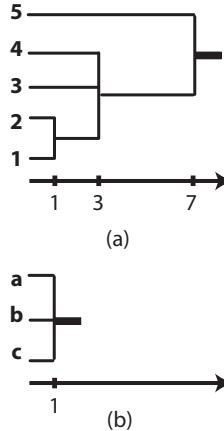


Figure 8: (a) shows the result of applying the permutation invariant formulation of CL to the five point metric space of Example 3 (see also Figure 6). (b) shows the dendrogram that one obtains as output of (the permutation invariant formulation of) SL, AL and CL applied to the metric space $L_3$.

**Proposition 5** *We have the following properties of the construction above:*

- *For $i = 1, 2, \ldots$, $\Theta_{i+1}$ is coarser than $\Theta_i$ and*

- *$R_{i+1} \geqslant R_i$.*

- *$\theta^\ell$ is a dendrogram over $X$.*

**Proof** The only non trivial claim is that $R_{i+1} \geqslant R_i$, which can be proved by induction on $i$. ■

**Remark 6** *From this point forward, all references to SL, AL, and CL clustering will be to the permutation invariant formulation, in which more than two clusters can be merged at a given step.*

The following result is clear, and we omit its proof.

**Proposition 7** *The above construction of hierarchical clustering algorithms (including SL, AL, and CL) yields algorithms which are permutation invariant.*

**A simplification for SL HC.** In the particular case of SL, there is an alternative formulation that uses the equivalence relation introduced in Example 1 and its associated dendrogram (Example 2). The proof of the following Proposition is deferred to the appendix.

**Proposition 8** *Let $(X, d)$ be a finite metric space and $\theta^{SL}$ be the dendrogram over $X$ obtained by the single linkage agglomerative procedure described above, and let $\theta^*$ be the dendrogram over $X$ constructed in Example 2. Then, $\theta^{SL}(r) = \theta^*(r)$ for all $r \geqslant 0$.*

### 4.3 Dendrograms as Ultrametric spaces

The representation of dendrograms as ultrametrics is well known and it appears in the book by Jardine and Sibson (Jardine and Sibson, 1971), it has already been used in the work of Hartigan (Hartigan, 1985), and is touched upon in the classical reference of Jain and Dubes (Jain and Dubes, 1988, §3.2.3).

We now present the main ideas regarding this change in perspective which we will adopt for all subsequent considerations. The formulation of the output of hierarchical clustering algorithms as ultrametric spaces is powerful when one is proving stability results, as well as results about the approximation of the dendrograms of metric spaces by their finite subspaces. This is so because of the fact that once a dendrogram is regarded as a metric space, the Gromov-Hausdorff metric provides a very natural notion of distance on the output, in which the right kind of stability results are easily formulated. We state these theorems in §6.

The main result in this section is that dendrograms and ultrametrics are **equivalent**.

**Theorem 9** *Given a finite set $X$, there is a bijection $\Psi : \mathcal{D}(X) \to \mathcal{U}(X)$ between the collection $\mathcal{D}(X)$ of all dendrograms over $X$ and the collection $\mathcal{U}(X)$ of all ultrametrics over $X$ such that for any dendrogram $\theta \in \mathcal{D}(X)$ the ultrametric $\Psi(\theta)$ over $X$ generates the same hierarchical decomposition as $\theta$, i.e.*

$$(*) \quad \text{for each } r \geqslant 0, \ x, x' \in \mathcal{B} \in \theta(r) \iff \Psi(\theta)(x, x') \leqslant r.$$

*Furthermore, this bijection is given by*

$$\Psi(\theta)(x, x') = \min\{r \geqslant 0 \,|\, x, x' \text{ belong to the same block of } \theta(r)\}.$$

In order to establish the above theorem, we first construct certain natural mappings from $\mathcal{D}(X)$ to $\mathcal{U}(X)$ and from $\mathcal{U}(X)$ to $\mathcal{D}(X)$, and we then prove they are inverses of eachother and satisfy $(*)$.

**From dendrograms to ultrametrics.** Let $X$ be a finite set and $\theta : [0, \infty) \to \mathcal{P}(X)$ a dendrogram over $X$. Consider the symmetric map $u_\theta : X \times X \to \mathbb{R}^+$ given by

$$(x, x') \mapsto \min\{r \geq 0 \,|\, x, x' \text{ belong to the same block of } \theta(r)\}. \qquad (4.3)$$

See Figure 9 for an illustration of this definition. Note that condition (4) in the definition of dendrograms guarantees that $u_\theta$ is well defined. It is easy to see that $u_\theta$ defines an ultrametric on $X$:

**Lemma 10** *Let $X$ be a finite set and $(X, \theta) \in \mathcal{D}(X)$. Then $u_\theta : X \times X \to \mathbb{R}^+$ defined in (4.3) is an ultrametric.*



$$((u_\theta)) = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{pmatrix} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ 0 & r_1 & r_3 & r_3 \\ r_1 & 0 & r_3 & r_3 \\ r_3 & r_3 & 0 & r_2 \\ r_3 & r_3 & r_2 & 0 \end{pmatrix}$$
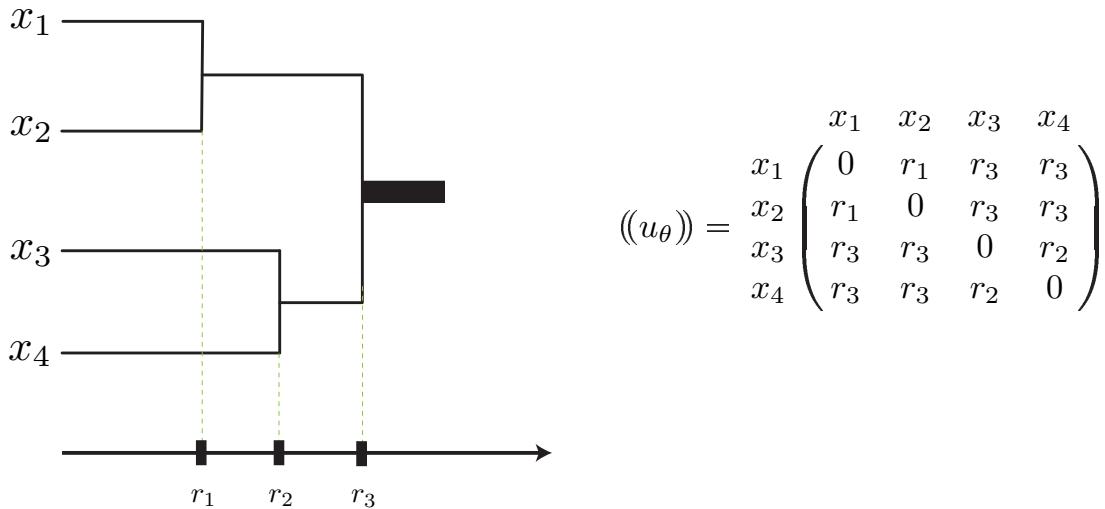
Figure 9: A graphical representation of a dendrogram $\theta$ over $X = \{x_1, x_2, x_3, x_3\}$ and the ultrametric $u_\theta$. Notice for example, that according to (4.3), $u_\theta(x_1, x_2) = r_1$ since $r_1$ is the first value of the (scale) parameter for which $x_1$ and $x_2$ are merged into the same cluster. Similarly, since $x_1$ and $x_3$ are merged into the same cluster for the first time when the parameter equals $r_3$, then $u_\theta(x_1, x_3) = r_3$.

**From ultrametrics to dendrograms.** Conversely, given an ultrametric $u : X \times X \to \mathbb{R}^+$, its associated dendrogram

$$\theta^u : [0, \infty) \to \mathcal{P}(X)$$

can be obtained as follows: for each $r \geq 0$ let $\theta^u(r)$ be the collection of equivalence classes of $X$ under the relation $x \sim x'$ if and only if $u(x, x') \leq r$. That this defines

22

an equivalence relation follows immediately from the fact that $u$ is an ultrametric. Indeed, assume that $x \sim x'$ and $x' \sim x''$ for some $r \geqslant 0$. Then, $u(x, x') \leqslant r$ and $u(x', x'') \leqslant r$. Now, by the ultrametric property, $\max\big(u(x, x'), u(x', x'')\big) \geqslant u(x, x'')$ and hence $u(x, x'') \leqslant r$ as well. We conclude that $x \sim x''$ thus establishing the transitivity of $\sim$.

**Example 5** *Consider the ultrametric $u$ on $X = \{x_1, x_2, \ldots, x_6\}$ given by*

$$
((u)) = 
\begin{array}{c c}
 & \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\
\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} & 
\left(\begin{array}{cccccc}
0 & 2 & 2 & 5 & 6 & 6 \\
2 & 0 & 2 & 5 & 6 & 6 \\
2 & 2 & 0 & 5 & 6 & 6 \\
5 & 5 & 5 & 0 & 6 & 6 \\
6 & 6 & 6 & 6 & 0 & 4 \\
6 & 6 & 6 & 6 & 4 & 0.
\end{array}\right)
\end{array}
$$

*Then, for example $\theta^u(0) = \big\{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\big\}$, $\theta^u(3) = \big\{\{x_1, x_2, x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\big\}$, $\theta^u(4.5) = \big\{\{x_1, x_2, x_3\}, \{x_4\}, \{x_5, x_6\}\big\}$, $\theta^u(5.5) = \big\{\{x_1, x_2, x_3, x_4\}, \{x_5, x_6\}\big\}$ and $\theta^u(7) = \big\{x_1, x_2, x_3, x_4, x_5, x_6\big\}$. A graphical representation of the dendrogram $\theta^u$ is given in Figure 10.*

**The conclusion of the proof of Theorem 9.** It is easy to check that (1) given any dendrogram $\theta$ on X, $\theta^{u_\theta} = \theta$ and (2) given any ultrametric $u$ on X, $u_{\theta^u} = u$. Now, let $\Psi : \mathcal{D}(X) \to \mathcal{U}(X)$ be defined by $\theta \mapsto \Psi(\theta) := u_\theta$. By construction we see that $\Psi : \mathcal{D}(X) \to \mathcal{U}(X)$ is a bijection and that $\Psi^{-1}$ is given by $u \mapsto \theta^u$. From (4.3) we see that $\Psi$ satisfies ($*$). Hence, we obtain Theorem 9.

---

From now, whenever given a dendrogram $\theta_X$ over a set $X$, we will be using the notation $\Psi(\theta_X)$ for the ultrametric associated to $X$ given by Theorem 9. In a similar manner, given an ultrametric $u$ on $X$, $\Psi^{-1}(u)$ will denote the dendrogram over $X$ given by Theorem 9.

---

## 4.4 Reformulation of Hierarchical Clustering using Ultrametrics

In the sequel, appealing to Theorem 9 which states the **equivalence between ultrametrics and dendrograms**, we represent dendrograms as *ultrametric* spaces. Then, any hierarchical clustering method can be regarded as a map from finite metric spaces into finite ultrametric spaces. This motivates the following definition:

**Definition 11** *A hierarchical clustering method is defined to be a map*

$$
\mathfrak{T} : \mathcal{X} \to \mathcal{U} \quad s.t. \quad \mathcal{X}_n \ni (X, d) \mapsto (X, u) \in \mathcal{U}_n, \ n \in \mathbb{N}.
$$
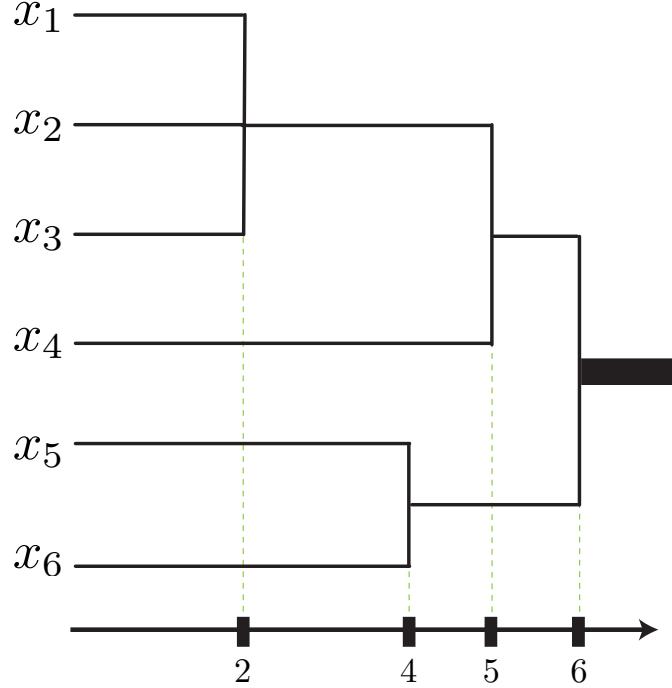
Figure 10: A graphical representation of the dendrogram $\theta^u$ of Example 5, see the text for details.

**Example 6** *For a given finite metric space $(X, d)$ consider the HC method $\mathfrak{T}^{\mathrm{SL}}$ given by $\mathfrak{T}^{\mathrm{SL}}(X, d) = (X, \Psi(\theta^{\mathrm{SL}}))$, where $\theta^{\mathrm{SL}}$ is the single linkage dendrogram over $X$ defined in §4.2. Similarly, we define $\mathfrak{T}^{\mathrm{CL}}$ and $\mathfrak{T}^{\mathrm{AL}}$.*

**Example 7 (maximal sub-dominant ultrametric)** *There is a canonical construction: Let $\mathfrak{T}^* : \mathcal{X} \to \mathcal{U}$ be given by $(X, d) \mapsto (X, u^*)$ where*

$$u^*(x, x') := \min \left\{ \max_{i=0,\ldots,k-1} d(x_i, x_{i+1}), \; s.t. \; x = x_0, \ldots, x_k = x' \right\}. \qquad (4.4)$$

*We remark that the minimum above is taken over $k \in \mathbb{N}$ and all $k + 1$-tuples of points $x_0, x_1, \ldots, x_k$ in $X$ s.t. $x_0 = x$ and $x_k = x'$. Notice that for all $x, x' \in X$, $u^*(x, x') \leqslant d(x, x')$.*

*This construction is sometimes known as the **maximal sub-dominant ultrametric** and it has the property that if $u \leqslant d$ is any other ultrametric on $X$, then $u \leqslant u^*$. The Lemma below proves that this canonical construction is equivalent to the ultrametric induced by the equivalence relation in Example 1.*

**Lemma 12** *For $(X, d) \in \mathcal{X}$ write $\mathfrak{T}^*(X, d) = (X, u^*)$ and let $(X, \theta^*) \in \mathcal{D}(X)$ be the dendrogram arising from the construction in Example 2. Then, $u^* = \Psi(\theta^*)$.*

24

**Remark 13** *Notice that another way of stating the Lemma above is that $x \sim_r x'$ if and only if $u^*(x, x') \leqslant r$.*

It turns out that $\mathrm{T}^*$ yields **exactly** *single linkage clustering* as defined in §4.2.

**Corollary 14** *One has that $\mathfrak{T}^{SL} = \mathfrak{T}^*$.*

Equivalently, for any finite metric space $X$, the single linkage dendrogram $\theta^{\mathrm{SL}}$ on $X$ agrees with $\Psi^{-1}(u^*)$.
**Proof** The proof follows easily from Proposition 8 and Lemma 12. ∎

> We emphasize that, as it follows from Corollary 14, $\mathfrak{T}^*$ produces ultrametric outputs which are exactly those corresponding to SLHC. We will use this fact strongly in the sequel.

**Interpretation of the ultrametric.** For a HC method $\mathfrak{T}$ and $(X, d) \in \mathcal{X}$, let $\mathfrak{T}(X, d) = (X, u)$. The intuition that arises from (4.3) is that for two points $x, x' \in X$, $u(x, x')$ measures the minimal **effort** method $\mathfrak{T}$ makes in order to join $x$ to $x'$ into the same cluster.

We note in particular that a desirable property of a HC algorithm should be that upon shrinking some of the distances in the input metric space, the corresponding "efforts" also decrease. This property is exactly verified by $\mathfrak{T}^*$. Indeed, let $X$ be a finite set and $d_1$ and $d_2$ two metrics on $X$ s.t. $d_1 \geqslant d_2$. Write $\mathfrak{T}^*(X, d_1) = (X, u_1^*)$ and $\mathfrak{T}^*(X, d_2) = (X, u_2^*)$. Then, it follows immediately from equation (4.4) that $u_1^* \geqslant u_2^*$ (compare with Kleinberg's *consistency* property, pp 5).

Observe that CL and AL HC *fail to satisfy this property*. An example is provided in Figure 19.

We see in Theorem 18 that a condition of this type, together with two more natural normalizing conditions, *completely characterizes SLHC*.

## 4.5 Comparing results of Hierarchical Clustering methods

One of the goals of this paper is to study the stability of clustering methods to perturbations in the input metric space. In order to do so one needs to define certain suitable notions of distance between dendrograms. We choose to do this by appealing to the ultrametric representation of dendrograms, which provides a natural way of defining a distance between hierarchical clusterings. We now delve into the construction.

Consider first the simple case of two different dendrograms $(X, \alpha)$ and $(X, \beta)$ over the same fixed finite set $X$. In this case, as a tentative measure of dissimilarity between the dendrograms we look at the maximal difference between the associated ultrametrics given by Theorem 9: $u_\alpha = \Psi(\alpha)$ and $u_\beta = \Psi(\beta)$: $\max_{x,x' \in X} |u_\alpha(x, x') - u_\beta(x, x')|$. There is a natural interpretation of the condition that $\max_{x,x' \in X} |u_\alpha(x, x') -$

$u_\beta(x,x')| \leqslant \varepsilon$: if we look at the graphical representation of the dendrograms $\alpha$ and $\beta$, then the transition horizontal lines in Figure 11 have to occur within $\varepsilon$ of eachother.[3] This is easy to see by recalling that by (4.3),

$$u_\alpha(x,x') = \min\{r \geqslant 0| \, x, x' \text{ belong to the same block of } \alpha(r)\}$$

and

$$u_\beta(x,x') = \min\{r \geqslant 0| \, x, x' \text{ belong to the same block of } \beta(r)\}.$$

For the example in Figure 11, we then obtain that $\max_i |r_i - r'_i| \leqslant \varepsilon$, which is not surprising since $r_2 = u_\alpha(x_1, x_2)$, $r'_2 = u_\beta(x_1, x_2)$, etc.
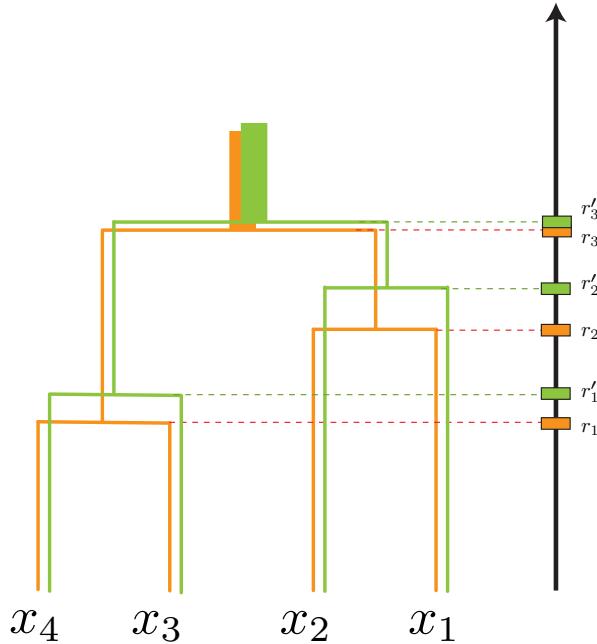


Figure 11: Two different dendrograms $(X, \alpha)$ and $(X, \beta)$ over the same underlying set $X = \{x_1, x_2, x_3, x_4\}$. Let $\alpha$ be the dendrogram represented in orange and $\beta$ be the one in green. The condition that $\|u_\alpha - u_\beta\|_{L^\infty(X \times X)} \leqslant \varepsilon$ is equivalent to the horizontal dotted lines corresponding to $r_i$ and $r'_i$ ($i = 1, 2, 3$) being within $\varepsilon$ of eachother.

Now, in a slightly more general situation we may be faced with the task of comparing two different dendrograms $\alpha$ and $\beta$ without knowing (or caring about) the

---

3. These lines represent values of the scale parameter for which there is a merging of blocks of the partitions encoded by the dendrograms.

exact labels of the points. In this case, a natural solution is to look at the minimum of the maximum difference of the corresponding ultrametrics under all possible permutations, namely:

$$\min_{\pi \in P_n} \max_{x,x' \in X} |u_\alpha(x, x') - u_\beta(\pi(x), \pi(x'))|, \tag{4.5}$$

where $n$ is the cardinality of $X$ and $P_n$ is the collection of all permutations of $n$ elements.

The most general case arises when we do not know whether the dendrograms come from the same underlying set or not. This situation may arise, for example, when comparing the results of clustering two different samples, of possibly different sizes, coming from the same dataset. One may want to be able to compare two such clusterings as a way to ascertain whether the sample size is sufficient for capturing the structure of the underlying dataset.

Assume then that we are given $(X_1, \alpha)$ and $(X_2, \beta)$, two different dendrograms, defined possibly over two *different sets* $X_1$ and $X_2$ of different cardinality. This potential difference in cardinality in the two sets forces us to consider transformations other than mere permutations. A natural solution, which can be interpreted as a *relaxation* of the permutation based distance (4.5) discussed above, is to consider maps $f : X_1 \to X_2$ and $g : X_2 \to X_1$ and look at their *distortions*:

$$\mathrm{dis}(f) := \max_{x,x' \in X_1} |u_\alpha(x, x') - u_\beta(f(x), f(x'))|,$$

$$\mathrm{dis}(g) := \max_{x,x' \in X_2} |u_\alpha(g(x), g(x')) - u_\beta(x, x)|.$$

The next natural step would be to optimize over the choice of $f$ and $g$, for example by minimizing the maximum of the two distortions:

$$\min_{f,g} \max \big(\mathrm{dis}(f), \mathrm{dis}(g)\big).$$

This construction is depicted in Figure 12. Roughly speaking, this idea leads to the Gromov-Hausdorff distance. The difference lies in the fact that in standard definition of the Gromov-Hausdorff distance, one also considers a term that measures the degree to which $f$ and $g$ are inverses of eachother. Being more precise, given the maps $f$ and $g$, this term, called the *joint distortion* of $f$ and $g$ is given by

$$\mathrm{dis}(f, g) := \max_{x \in X, y \in Y} |u_\alpha(x, g(y)) - u_\beta(y, f(x))|.$$

One defines the Gromov-Hausdorff distance between $(X_1, u_\alpha)$ and $(X_2, u_\beta)$ by

$$d_{\mathcal{GH}}\big(X_1, X_2\big) := \frac{1}{2} \min_{f,g} \max \big(\mathrm{dis}(f), \mathrm{dis}(g), \mathrm{dis}(f, g)\big).^4 \tag{4.6}$$

---

4. The factor $\frac{1}{2}$ is of course inmaterial but kept here for coherence with the standard definition.

We now see exactly how the inclusion of the new term enforces $f$ and $g$ to be approximate inverses of eachother. Assume that for some $\varepsilon > 0$ $d_{\mathcal{GH}}(X_1, X_2) < \varepsilon$, then, in particular, there exist maps $f$ and $g$ such that $|u_\alpha(x, g(y)) - u_\beta(y, f(x))| \leqslant 2\varepsilon$ for all $x \in X_1$ and $y \in X_2$. Choosing $y = f(x)$, in particular, we obtain that $u_\alpha(x, g(f(x))) \leqslant 2\varepsilon$ for all $x \in X_1$. Similarly one obtains that $u_\beta(y, f(g(y))) \leqslant 2\varepsilon$ for all $y \in X_2$. These two inequalities measure the degree to which $f \circ g$ and $g \circ f$ differ from the identities, and thus, measure the degree to which $f$ and $g$ fail to be inverses of eachother. This is a useful feature when one considers *convergence issues* such as we do in §6.

**Interpretation of the Gromov-Hausdorff distance in terms of dendrograms.**
Assume that $d_{\mathcal{GH}}\big((X_1, u_\alpha), (X_2, u_\beta)\big) \leqslant \frac{\eta}{2}$ for some $\eta \geqslant 0$. Then there exist maps $f : X \to Y$ and $g : Y \to X$ such that the following conditions hold (see Figure 13):

- If $x, x'$ fall in the same block of $\alpha(t)$ then $f(x), f(x')$ belong to the same block of $\beta(t')$ for all $t' \geqslant t + \eta$.

- If $y, y'$ fall in the same block of $\beta(t)$ then $g(y), g(y')$ belong to the same block of $\alpha(t')$ for all $t' \geqslant t + \eta$.

For the next section we do not need to make use of the full generality in these considerations: there we only compare dendrograms defined over the same underlying set. A more detailed use and additional material about the Gromov-Hausdorff ideas is given in §6.

We finish this section with a precise result regarding the stability of dendrograms arising from SLHC.

The following Lemma deals with the situation when we have a fixed finite set $P$ and two different metrics on $P$ and then we compute the result of applying $\mathfrak{T}^*$ each of these metrics. This lemma is a particular case of our main stability result, Proposition 26 in §6. In the interest of clarity, we prove it here to provide some intuition about the techniques.

**Lemma 15** *Let $P$ be a fixed finite set and let $d_1, d_2$ be two metrics on $P$. Write $\mathfrak{T}^*(P, d_i) = (P, u_i)$, $i = 1, 2$. Then,*

$$\max_{p,q \in P} |u_1(p, q) - u_2(p, q)| \leqslant \max_{p,q \in P} |d_1(p, q) - d_2(p, q)|.$$

**Proof** Let $\eta = \max_{p,q \in P} |d_1(p, q) - d_2(p, q)|$. Let $p_0, \ldots, p_k \in P$ be s.t. $p_0 = p$, $p_k = q$ and $\max_i d_1(p_i, p_{i+1}) = u_1(p, q)$. Then, by definition of $u_2$ (which is the minimum over all chains of the maximal hop measured with metric $d_2$) and the fact that $d_2 \leqslant d_1 + \eta$:

$$u_2(p, q) \leqslant \max_i d_2(p_i, p_{i+1}) \leqslant \max_i \big(\eta + d_1(p_i, p_{i+1})\big) = \eta + u_1(p, q).$$

Similarly, $u_1(p, q) \leqslant \eta + u_2(p, q)$, and hence $|u_1(p, q) - u_2(p, q)| \leqslant \eta$. The claim follows since $p, q \in P$ are arbitrary. ∎
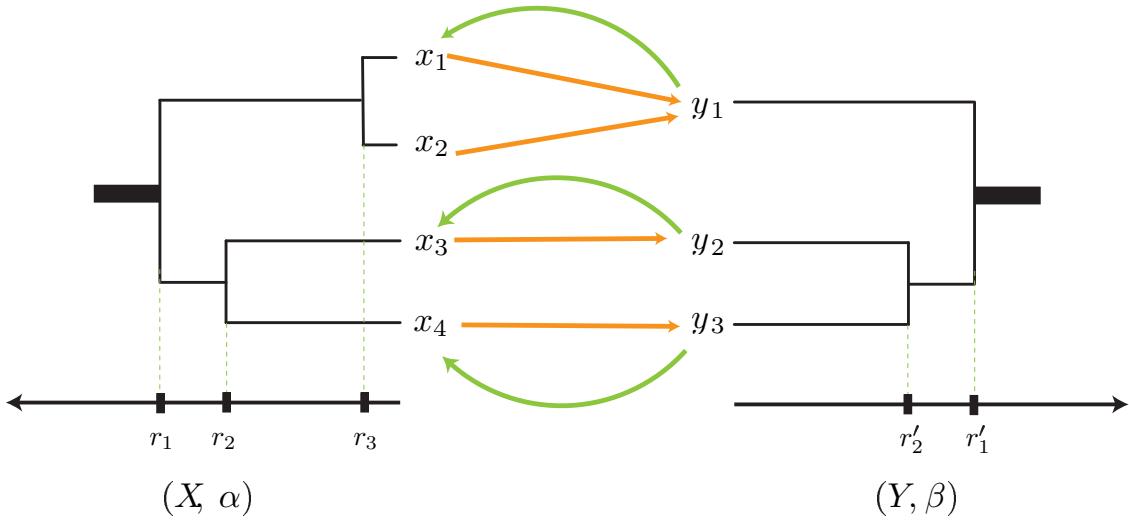
Figure 12: In this example, two different dendrograms, $(X, \alpha)$ and $(Y, \beta)$, are given. In orange arrows we show a map $f : X \to Y$ and in green arrows the map $g : Y \to X$. With simple explicit computations one sees that these choices of maps $f$ and $g$ incur distortions $\mathrm{dis}(f) = \mathrm{dis}(f, g) = \max\left(r_3, |r_1 - r_1'|, |r_2 - r_2'|\right)$ and $\mathrm{dis}(g) = \max\left(|r_1 - r_1'|, |r_2 - r_2'|\right)$, respectively. Hence, we see that $d_{\mathcal{GH}}\left((X, \Psi(\alpha), (Y, \Psi(\beta))\right) \leqslant \frac{1}{2} \max\left(r_3, |r_1 - r_1'|, |r_2 - r_2'|\right)$.

## 4.6 Some remarks about Hierarchical Clustering methods

Practitioners of clustering often prefer AL and CL to SL because it is perceived that the former two methods tend to produce clusters which are more coherent conceptually, and which are in a non-technical sense viewed as more compact. In fact, SL exhibits the so called **chaining effect** which makes it more likely to produce clusterings which separate items which conceptually should be together. We view these observations as evidence for the idea that good clustering schemes need to take some notion of density into account, rather than straightforward geometric information alone. One can loosely argue that given the actual definition of the linkage functions used by AL and CL, these two methods do enjoy some sort of sensitivity to density. Unfortunately, AL and CL are **unstable**, and in particular, **discontinuous** in a very precise sense (see Remark 16 below), whereas SL enjoys all the nice theoretical properties that the other two methods lack.

In this section we review this seemingly paradoxical situation.

For each $n \in \mathbb{N}$ let $L_n$ be a metric space with $n$ points $P = \{p_1, \ldots, p_n\}$ and metric $d_{L_n}(p_i, p_j) = |i - j|$, $i, j \in \{1, \ldots, n\}$. Similarly, let $\Delta_n$ be the metric space with the same underlying set and metric $d_{\Delta_n}(p_i, p_j) = 1$, $i, j \in \{1, \ldots, n\}$, $i \neq j$. Clearly, the
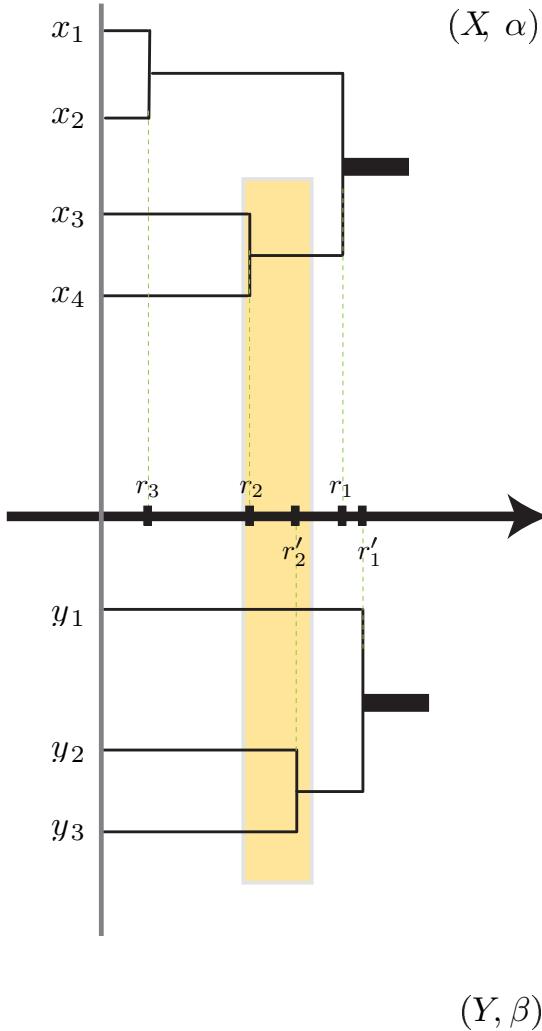
Figure 13: These are the same dendrograms as in Figure 12. Let $r_1 = \frac{5}{3}$, $r_2 = 1$, $r_3 = \frac{1}{3}$, $r_1' = \frac{11}{6}$ and $r_2' = \frac{4}{3}$. For the maps $f$ and $g$ s.t. $f(x_1) = f(x_2) = y_1$, $f(x_3) = y_2$, $f(x_4) = y_3$, $g(y_1) = x_1$, $g(y_2) = x_3$ and $g(y_3) = x_4$, using the formulas computed in Figure 12 we see that $\mathrm{dis}(f) = \mathrm{dis}(g) = \mathrm{dis}(f, g) = \frac{1}{3}$ and hence $d_{\mathcal{GH}}\big((X, \Psi(\alpha)), (Y, \Psi(\beta))\big) \leqslant \frac{1}{6}$. Now notice for instance that $x_3$ and $x_4$ fall in the same block of $\alpha(r_2) = \alpha(1)$ and that $y_2 = f(x_3)$ and $y_3 = f(x_4)$ fall in the same block of $\beta(t')$ for all $t' \geqslant r_2 + 2 \cdot \frac{1}{6} = 1 + \frac{1}{3} = \frac{4}{3} = r_2'$.

metric space $L_n$ is isometric to points equally spaced on a line in Euclidean space whereas (s.t. two adjacent points are at distance 1 from eachother) $\Delta_n$ is isometric to the $(n-1)$-unit-simplex as a subset of $\mathbb{R}^{n-1}$.

Clearly, the outputs of Single Linkage HC applied to both $L_n$ and $\Delta_n$ coincide for all $n \in \mathbb{N}$:

$$\mathfrak{T}^*(P, d_{L_n}) = \mathfrak{T}^*(P, d_{\Delta_n}) = \big(P, ((\gamma))\big) \tag{4.7}$$

where $\gamma_{ij} = 0$ if $i = j$ and $\gamma_{ij} = 1$ if $i \neq j$, for all $n \in \mathbb{N}$, see Figure 14.
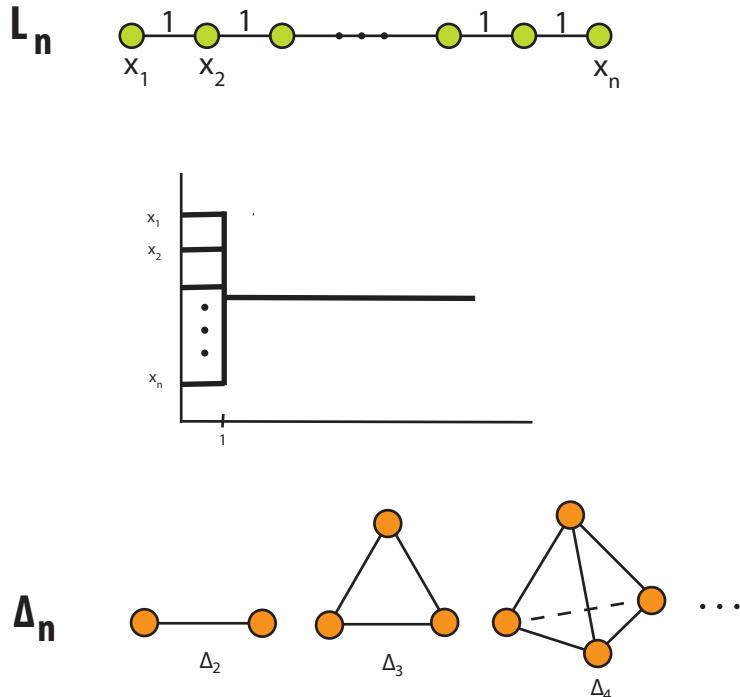


Figure 14: The metric spaces $L_n$ and $\Delta_n$ both have $n$ points. Single linkage HC applied to either of them yields the dendrogram in the center.

By appealing to the Euclidean realizations of $L_n$ and $\Delta_n$, one can define *perturbed* versions of these two metric spaces. Indeed, fix $\varepsilon > 0$ and let $\{a_1, \ldots, a_n\} \subset [0, \varepsilon/2]$ and $\{b_1, \ldots, b_n\} \subset S^{n-1}(\varepsilon/2)$. Define $L_n^\varepsilon$ to be the metric space with underlying set $P$ and metric $d_{L_n^\varepsilon}(p_i, p_j) = |i - j + a_i - a_j|$. Similarly, define $\Delta_n^\varepsilon$ to be the metric space with underlying set $P$ and metric $d_{\Delta_n^\varepsilon}(p_i, p_j) = \|s_i - s_j + b_i - b_j\|$.

Notice that by construction,

$$\max_{i,j} |d_{L_n}(p_i, p_j) - d_{L_n^\varepsilon}(p_i, p_j)| \leqslant \varepsilon \tag{4.8}$$

and

$$\max_{i,j} |d_{\Delta_n}(p_i, p_j) - d_{\Delta_N^\varepsilon}(p_i, p_j)| \leqslant \varepsilon. \tag{4.9}$$

31

We thus say that the spaces $(P, d_{L_n^\varepsilon})$ and $(P, d_{\Delta_n^\varepsilon})$ are perturbed versions of $(P, d_{L_n})$ and $(P, d_{\Delta_n})$, respectively.

**Remark 16 (About a critique to SL)** *Single linkage is generally regarded as a poor choice in practical applications. The reason for this is the so called* chaining *effect observed experimentally, which is central to the criticism to SL made in (Lance and Williams, 1967) (see also the discussion in (Wishart, 1969, pp. 296)). The following two observations are important:*

**(O1)** *It is generally argued that since $(P, d_{L_n^\varepsilon})$ corresponds to points on the vicinity of a line, whereas $(P, d_{\Delta_n^\varepsilon})$ corresponds to points in the close vicinity of a $(n-1)$-simplex, then the cluster formed by points on the latter metric space is more* compact *or* denser *than the one formed by the former, and thus more meaningful.*

**(O2)** *The outputs of SL to the spaces $(P, d_{L_n^\varepsilon})$ and $(P, d_{\Delta_n^\varepsilon})$ are very similar and this similarity is of order $\varepsilon$.*

*Indeed, if we write $\mathfrak{T}^*(P, d_{L_n^\varepsilon}) = (P, u_{L_n^\varepsilon})$ and $\mathfrak{T}^*(P, d_{\Delta_n^\varepsilon}) = (P, u_{\Delta_n^\varepsilon})$, then, by the triangle inequality for the $L^\infty$ norm,*

$$
\begin{aligned}
\|u_{L_n^\varepsilon} - u_{\Delta_n^\varepsilon}\|_{L^\infty(P\times P)} \;\leqslant\;\; & \|u_{L_n^\varepsilon} - u_{L_n^0}\|_{L^\infty(P\times P)} \\
+ \;\; & \|u_{L_n^0} - u_{\Delta_n^0}\|_{L^\infty(P\times P)} \\
+ \;\; & \|u_{\Delta_n^0} - u_{\Delta_n^\varepsilon}\|_{L^\infty(P\times P)}.
\end{aligned} \tag{4.10}
$$

*As we pointed out in (4.7) at the beginning of Section §4.6,*

$$
u_{L_n^0} = u_{L_n} = ((\gamma)) = u_{\Delta_n} = u_{\Delta_n^0},
$$

*thus, (4.10) simplifies into:*

$$
\begin{aligned}
\|u_{L_n^\varepsilon} - u_{\Delta_n^\varepsilon}\|_{L^\infty(P\times P)} \;\leqslant\;\; & \|u_{L_n^\varepsilon} - u_{L_n^0}\|_{L^\infty(P\times P)} \\
+ \;\; & \|u_{\Delta_n^\varepsilon} - u_{\Delta_n^0}\|_{L^\infty(P\times P)} \\
& \textit{(and by Lemma 15:)} \\
\leqslant\;\; & \|d_{L_n^\varepsilon} - d_{L_n^0}\|_{L^\infty(P\times P)} \\
+ \;\; & \|d_{\Delta_n^\varepsilon} - d_{\Delta_n^0}\|_{L^\infty(P\times P)}.
\end{aligned} \tag{4.11}
$$

*Hence, by (4.11) and the construction of $d_{L_n^\varepsilon}$ and $d_{\Delta_n^\varepsilon}$ (equations (4.8) and (4.9)), we conclude that*

$$
\|u_{L_n^\varepsilon} - u_{\Delta_n^\varepsilon}\|_{L^\infty(P\times P)} \leqslant 2\varepsilon.
$$

*This means that for any small perturbations of $L_n$ and $\Delta_n$, the output of SL to these perturbations are at a small distance from eachother, as we claimed.*

*When put together, observations **(O1)** and **(O2)** suggest that SL is unable to pick denser associations of data, such as cliques, over sparser ones, such as linear structures. This feature is undesirable in practical applications where often times one would like to regard clusters as modes of an underlying distribution (Wishart, 1969; Hartigan, 1981).*

*It is then the case that in practical applications, CL and especially AL are preferred over SL. These two methods have the property that they indeed somehow favor the association of compact subsets of points. For CL this can be explained easily using the concept of* maximal clique *(maximally connected sub-graphs of a given graph) (Jain and Dubes, 1988, Section 3.2.1). Let $d_k$ be the diameter of the cluster created in step $k$ of CL clustering and define a graph $G(k)$ as the graph that links all data points with a distance of at most $d_k$. Then the clusters after step $k$ are the maximal cliques of $G(k)$. This observation reinforces the perception that CL yields clusters that are dense as measured by the presence of cliques. The sensitivity of AL to density has been discussed by Hartigan in (Hartigan, 1985, Section 3) and is basically due to the averaging performed in the definition of its linkage function.*

*A more principled way of taking density into account, that does not depend on ad hoc constructions which destroy the stability property, would be to explicitly build the density into the method. In (Carlsson and Mémoli, 2009) we study **multiparameter clustering methods**, which are similar to HC methods but we* track *connected components in a multiparameter landscape. We also study the classification and stability properties of multiparameter clustering methods.*

**Remark 17 (Instability of CL and AL)** *It turns out that CL and AL, despite not exhibiting the undesirable feature of the chaining effect, and despite being regarded as more sensitive to density, are **unstable** in a precise sense. Consider for example CL and let $n = 3$. In the construction of $(P, d_L^\varepsilon)$ above let $a_1 = a_2 = 0$ and $a_3 = \varepsilon$, then*

$$((d_L)) = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \end{array} \begin{pmatrix} p_1 & p_2 & p_3 \\ 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix} \text{ and } ((d_L^\varepsilon)) = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \end{array} \begin{pmatrix} p_1 & p_2 & p_3 \\ 0 & 1 & 2+\varepsilon \\ 1 & 0 & 1+\varepsilon \\ 2+\varepsilon & 1+\varepsilon & 0 \end{pmatrix}.$$

*Write $\mathfrak{T}^{CL}(P, d_L) = (P, u_L)$ and $\mathfrak{T}^{CL}(P, d_L^\varepsilon) = (P, u_L^\varepsilon)$. Clearly,*

$$((u_L)) = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \end{array} \begin{pmatrix} p_1 & p_2 & p_3 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \text{ and } ((u_L^\varepsilon)) = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \end{array} \begin{pmatrix} p_1 & p_2 & p_3 \\ 0 & 1 & 2+\varepsilon \\ 1 & 0 & 2+\varepsilon \\ 2+\varepsilon & 2+\varepsilon & 0 \end{pmatrix}.$$

*Notice that despite $\max_{i,j} |d_L(p_i, p_j) - d_L^\varepsilon(p_i, p_j)| = \varepsilon$, $\max_{i,j} |u_L(p_i, p_j) - u_L^\varepsilon(p_i, p_j)| = 1 + \varepsilon > 1$ for all $\varepsilon > 0$. We thus conclude that CL is not stable under small perturbations of the metric. Note that in particular, it follows that CL is **not continuous**. The same construction can be adapted for AL. See Figure 15.*
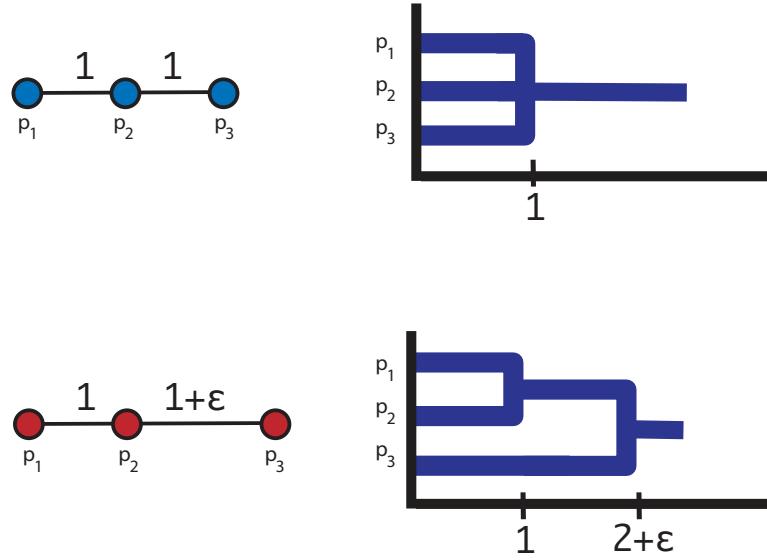
Figure 15: Complete Linkage is not stable to small perturbations in the metric. On the left we show two metric spaces that are metrically very similar. To the right of each of them we show their CL dendrogram outputs. Regardless of $\varepsilon > 0$, the two outputs are always very dissimilar. We make the notion of *similarity* between dendrogram precise in §6 by interpreting dendrograms as ultrametric spaces and by computing the Gromov-Hausdorff distance between these ultrametric spaces.

## 5. A characterization theorem for SL hierarchical clustering

In this section we obtain a **characterization** of SL hierarchical clustering in terms of some simple axioms. The main axiom, (II) below, says that the clustering scheme has a prescribed behavior under distance non-increasing maps of metric space. The behavior is that the map of metric spaces should induce a map of clusters, i.e. that if two points in the domain space belong to the same cluster, then so do their images in the clustering of the image metric space. This notion, referred to as **functoriality** in the mathematics literature, appears to us to be a very natural one, and it is closely related to Kleinberg's consistency property (cf. pp. 5) for ordinary clustering methods; see Remark 19 for an interpretation of our axioms.

**Theorem 18** *Let $\mathfrak{T}$ be a hierarchical clustering method s.t.*

**(I)** $\mathfrak{T}\left(\{p,q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right)\right) = \left(\{p,q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right)\right)$ *for all $\delta > 0$.*

**(II)** *Whenever $X, Y \in \mathcal{X}$ and $\phi : X \to Y$ are 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 $\mathfrak{T}(X, d_X) = (X, u_X)$ and $\mathfrak{T}(Y, d_Y) = (Y, u_Y)$. prop*

**(III)** *For all $(X, d) \in \mathcal{X}$,*

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

*where $\mathfrak{T}(X, d) = (X, u)$.*

*Then $\mathfrak{T} = \mathfrak{T}^*$, that is, $\mathfrak{T}$ is exactly single linkage hierarchical clustering.*

**Remark 19 (Interpretation of the conditions)** *Let $(X, d) \in \mathcal{X}$ and write $\mathfrak{T}(X, d) = (X, u)$. The intuition is that $u(x, x')$ measures the* effort *method $\mathfrak{T}$ makes in order to join $x$ to $x'$ into the same cluster.*

*Condition (I) is clear, the two-point metric space contains only one degree of freedom which has to determine unambiguously the behavior of any clustering method $\mathfrak{T}$. In terms of dendrograms, this means that the two point metric space $\left( \{A, B\}, \left( \begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix} \right) \right)$ must be mapped to the dendrogram where $A$ and $B$ are merged at parameter value $\delta$, see Figure 16.*

*Condition (II) is crucial and roughly says that whenever one shrinks some distances (even to zero) to obtain a new (pseudo) metric space, then the corresponding efforts in this new space have to be smaller than the efforts in the original metric space. This is consistent with the notion that reducing the distance between two points (without increasing all other distances) makes them more likely to belong to the same cluster.*

*Let $\theta_X = \Psi^{-1}(u_X)$ and $\theta_Y = \Psi^{-1}(u_Y)$ be the dendrograms associated to $u_X$ and $u_Y$. In terms of dendrograms, this means that if two points $x, x' \in X$ are in the same block of $\theta_X(t)$ for some $t > 0$, then $\phi(x)$ and $\phi(x')$ must be in the same block of $\theta_Y(t)$. see Figure 17.*

*Condition (III) expresses the fact that in order to join two points $x, x' \in X$, any clustering method $\mathfrak{T}$ has to make an effort of at least the separation $sep(X, d)$ of the metric space. In terms of dendrograms, this means that $\theta_X(t)$ has to equal the partition of $X$ into singletons for all $0 \leq t < sep(X, d)$. See Figure 18.*

**Remark 20** *It is interesting to point out why complete linkage and average linkage hierarchical clustering, as defined in §4.2.2, fail to satisfy the conditions in Theorem 18. It is easy to see that conditions (I) and (III) are always satisfied by CL and AL.*
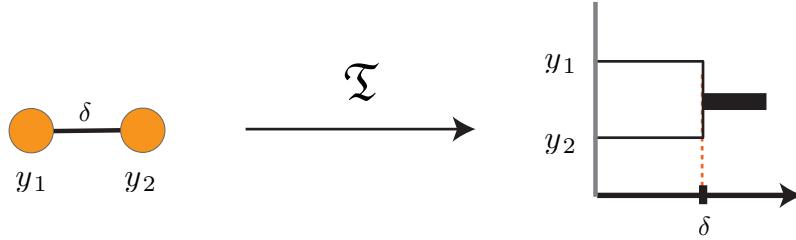
Figure 16: **Interpretation of Condition I**: For all $\delta > 0$ the two point metric space on the left must be mapped by $\mathfrak{T}$ into the dendrogram on the right.

*Consider the metric spaces $X = \{A, B, C\}$ with metric given by the edge lengths $\{4, 3, 5\}$ and $Y = (A', B', C')$ with metric given by the edge lengths $\{4, 3, 2\}$, as given in Figure 19. Obviously, the map $\phi$ from $X$ to $Y$ with $\phi(A) = A'$, $\phi(B) = B'$ and $\phi(C) = C'$ is s.t.*

$$d_Y(\phi(x), \phi(x')) \leqslant d_X(x, x') \ \ \textit{for all } x, x' \in \{A, B, C\}.$$

*It is easy to check that*

$$(\!(u_X)\!) = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{ccc} A & B & C \\ \begin{pmatrix} 0 & 5 & 3 \\ 5 & 0 & 5 \\ 3 & 5 & 0 \end{pmatrix} \end{array} \ and \ (\!(u_Y)\!) = \begin{array}{c} \\ A' \\ B' \\ C' \end{array} \begin{array}{ccc} A' & B' & C' \\ \begin{pmatrix} 0 & 2 & 4 \\ 2 & 0 & 4 \\ 4 & 4 & 0 \end{pmatrix} \end{array}.$$

*Note that for example $3 = u_X(A, C) < u_Y(\phi(A), \phi(C)) = u_Y(A', C') = 4$ thus violating property (II). The same construction yields a counter-example for average linkage.*

## 6. Metric stability and convergence of $\mathfrak{T}^*$

The Proposition and Theorem below assert the metric stability and consistency/convergence of the method $\mathfrak{T}^*$ (i.e. of SLHC, by virtue of Proposition 14. We use the notion of Gromov-Hausdorff distance between metric spaces (Burago et al., 2001). This notion of distance permits regarding the collection of all compact metric spaces as a metric space in itself.

This seemingly abstract construction is in fact very useful. Finite metric spaces are by now ubiquitous in virtually all areas of data analysis, and the idea of assigning a metric to the collection of all of them is in fact quite an old one. For Euclidean metric spaces, for example, the idea of constructing a metric was used by Kendall (Kendall et al., 1999) and Bookstein (Bookstein et al., 1985) in constructing a *statistical shape theory*, motivated by the ideas about form of biological organisms developed by D'Arcy Thompson.
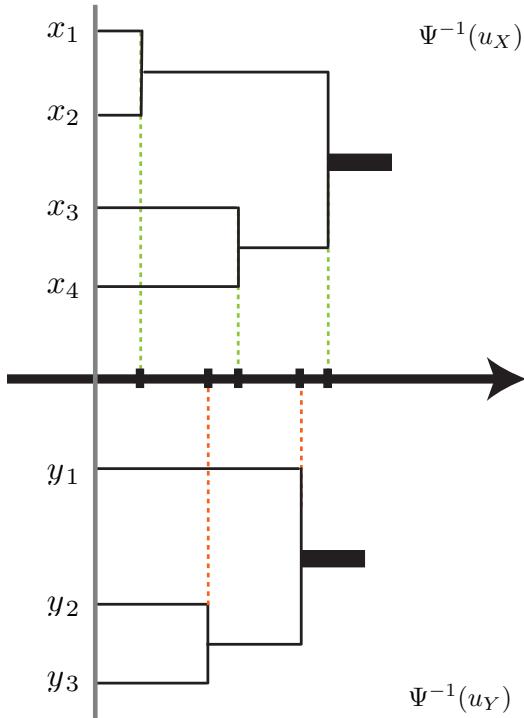
Figure 17: **Interpretation of Condition II**: Assume that $\phi : X \to Y$ is a distance non-increasing map such that $\phi(x_1) = \phi(x_2) = y_1$, $\phi(x_3) = y_2$ and $\phi(x_4) = y_3$. Then, Condition (II) requires that if $x, x' \in X$ are merged into the same cluster of $\Psi^{-1}(u_X)$ at parameter value $t$, then $\phi(x)$ and $\phi(x')$ must merge into the same cluster of $\Psi^{-1}(u_Y)$ for some parameter value $\leqslant t$. In the Figure, this translates into the condition that vertical dotted lines corresponding to mergings of pairs of points in $X$ should happen at parameter values greater than or equal than the parameter values for which correponding points in $Y$ (via $\phi$) are merged into the same cluster. For example, $\phi(x_1), \phi(x_2)$ merge into the same cluster at parameter value 0. The condition is clearly verified for this pair since by definition of $\phi$, $\phi(x_1) = \phi(x_2) = y_1$. Take now $x_3$ and $x_4$: clearly the vertical line that shows the parameter value for which they merge is to the right of the vertical line showing the parameter value for which $y_2 = \phi(x_3)$ and $y_3 = \phi(x_4)$ merge.

### 6.1 The Gromov-Hausdorff distance and examples

**Definition 21** *Let $(Z, d_Z)$ be a compact metric space. The Hausdorff distance between any two compact subsets $A, B$ of $Z$ is defined by*

$$d_{\mathcal{H}}^Z(A, B) := \max \left( \max_{a \in A} \min_{b \in B} d_Z(a, b), \max_{b \in B} \min_{a \in A} d_Z(a, b) \right).$$

37
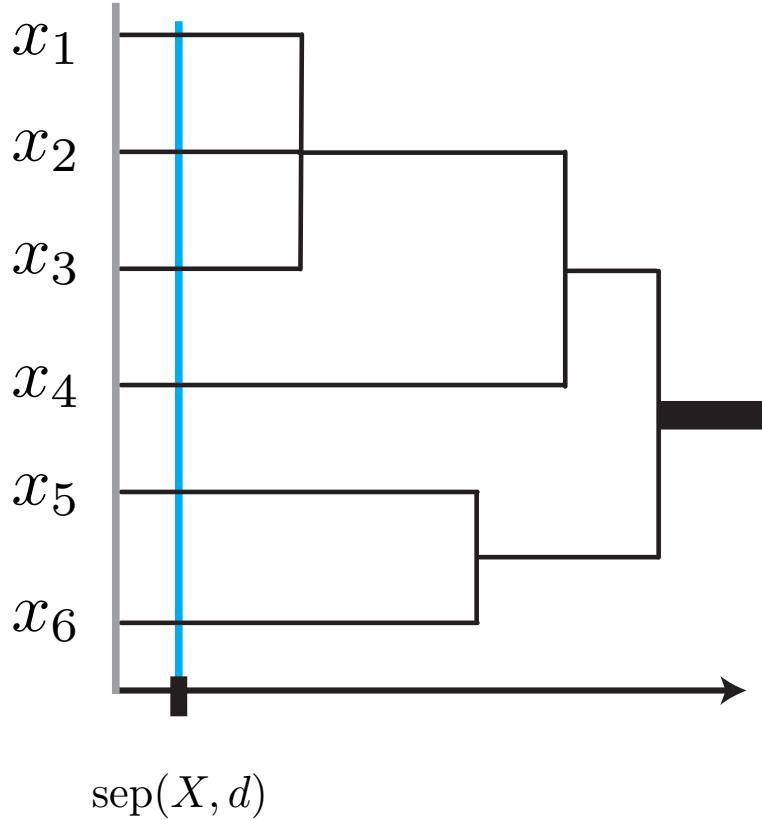
Figure 18: **Interpretation of Condition III**: The vertical line at parameter value $t = \text{sep}(X, d)$ must intersect the horizontal lines of the dendrogram before any two points are merged.

**Remark 22** *Let $\mathbb{Z} = \{z_1, \ldots, z_n\} \subset Z$. Then, $d_{\mathcal{H}}^Z(\mathbb{Z}, Z) \leqslant \delta$ for some $\delta \geqslant 0$ if and only if $Z \subset \bigcup_{i=1}^n \overline{B(z_i, \delta)}$. In other words, $d_{\mathcal{H}}^Z(\mathbb{Z}, Z)$ describes the minimal $\delta$ s.t. $\mathbb{Z}$ is a $\delta$-net for $Z$ and therefore measures how well $\mathbb{Z}$ covers $Z$.*

The Gromov-Hausdorff distance $d_{\mathcal{GH}}(X, Y)$ between compact metric spaces $(X, d_X)$ and $(Y, d_Y)$ was orignally defined to be the infimal $\varepsilon > 0$ s.t. there exists a metric $d$ on $X \bigsqcup Y$ with $d_{|X \times X} = d_X$ and $d_{|Y \times Y} = d_Y$ for which the Hausdorff distance between $X$ and $Y$ (as subsets of $(X \bigsqcup Y, d)$) is less than $\varepsilon$ (Gromov, 1987). There is, however, an alternative expression for the GH distance that is better suited for our purposes which we now recall.

**Definition 23 (Correspondence)** *For sets $A$ and $B$, a subset $R \subset A \times B$ is a correspondence (between $A$ and $B$) if and and only if*
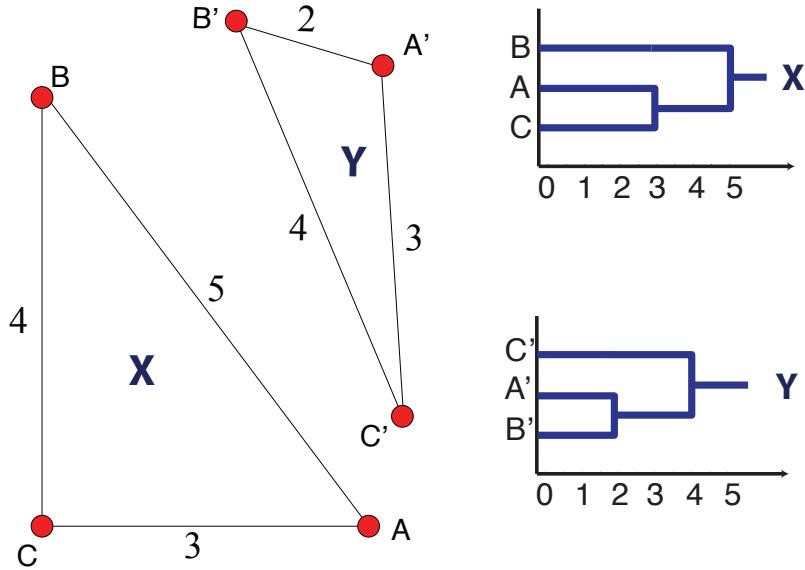
Figure 19:  An example that shows why complete linkage fails to satisfy condition (2) of Theorem 18.

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

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

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

We now give several examples to illustrate this definition.

**Example 8** Let $A = \{a_1, a_2\}$ and $B = \{b_1, b_2, b_3\}$. In this case, $R_1 = \{(a_1, b_1), (a_2, b_2), (a_1, b_3)\}$ is a correspondence but $R_2 = \{(a_1, b_1), (a_2, b_2)\}$ is not.

**Example 9** Let $A$ and $B$ be finite s.t. $\#A = \#B = n$. In this case, if $\pi$ is any permutation matrix of size $n$, then $\{(a_i, b_{\pi_i}), i = 1, \ldots, n\} \in \mathcal{R}(A, B)$.

**Example 10** Let $\phi : X \to Y$ and $\psi : Y \to X$ be given maps. Then, one can construct a correspondence out of these maps, call it $R(\phi, \psi)$ given by

$$\{(x, \phi(x)),\ x \in X\} \bigcup \{(\psi(y), y),\ y \in Y\}.$$

For metric spaces $(X, d_X)$ and $(Y, d_Y)$. Let $\Gamma_{X,Y} : X \times Y \times X \times Y \to \mathbb{R}^+$ be given by

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

Then, by (Burago et al., 2001, Theorem 7.3.25) the **Gromov-Hausdorff distance** between $X$ and $Y$ is equal to

$$d_{\mathcal{GH}}(X, Y) := \frac{1}{2} \inf_{R \in \mathcal{R}(X,Y)} \sup_{(x,y),(x',y') \in R} \Gamma_{X,Y}(x, y, x', y'). \qquad (6.12)$$

It can be seen (it is an easy computation) that in (6.12) one can restrict the infimum to those correspondences that arise from maps $\phi$ and $\psi$ such as those constructed in Example 10. Then, one recovers expression (4.6) which we gave in §4.5, namely, that actually

$$d_{\mathcal{GH}}(X, Y) := \frac{1}{2} \inf_{\phi, \psi} \max \big(\operatorname{dis}(\phi), \operatorname{dis}(\psi), \operatorname{dis}(\phi, \psi)\big). \qquad (6.13)$$

**Remark 24** *Expression (6.13) defines a **distance** on the set of (isometry classes of) finite metric spaces, (Burago et al., 2001, Theorem 7.3.30). From now on let $\mathcal{G}$ denote the collection of all (isometry classes of) compact metric spaces. We say that $\{(X_n, d_{X_n})\}_{n \in \mathbb{N}} \subset \mathcal{G}$ Gromov-Hausdorff converges to $X \in \mathcal{G}$ if and only if $d_{\mathcal{GH}}(X_n, X) \to 0$ as $n \uparrow \infty$.*

**Example 11** *Fix $(X, d_X) \in \mathcal{G}$. Consider the sequence $\{(X, \frac{1}{n} \cdot d_X)\}_{n \in \mathbb{N}} \subset \mathcal{G}$. Then, $X_n$ Gromov-Hausdorff converges to the metric space consisting of a single point.*

**Remark 25 (Gromov-Hausdorff distance and Hausdorff distance)** *Let $(X, d_X)$ be a compact metric space. Then, if $X' \subset X$ is compact and we endow $X'$ with the metric $d_{X'}$ equal to the restriction of $d_X$, then*

$$d_{\mathcal{GH}}\big((X, d_X), (X', d_{X'})\big) \leqslant d_{\mathcal{H}}^X\big(X', X\big).$$

*This is easy to see by defining the correspondence $R$ between $X$ and $X'$ given by*

$$R = \{(x', x'),\ x' \in X'\} \cup \{(x, x'),\ x \in V(x'),\ x' \in X'\},$$

*where $V(x') := \{x \in X,\ d_X(x, x') \leqslant d_X(x, z),\ z \in X' \backslash \{x'\}\}$. Indeed, since then, for all $(x_1, x_1'), (x_2, x_2') \in R$,*

$$\frac{1}{2}|d_X(x_1, x_2) - d_X(x_1', x_2')| \leqslant \frac{1}{2}\big(d_X(x_1, x_1') + d_X(x_2, x_2')\big) \leqslant \max_{x \in X} \min_{x' \in X'} d_X(x, x') = d_{\mathcal{H}}^X\big(X, X'\big).$$

**Example 12** *Consider a finite set $M$ and $d, d' : M \times M \to \mathbb{R}^+$ two metrics on $M$. Then, the GH distance between $(M, d)$ and $(M, d')$ is bounded above by the $L^\infty$ norm of the difference between $d$ and $d'$:*

$$d_{\mathcal{GH}}\big((M, d), (M, d')\big) \leqslant \frac{1}{2}\|d - d'\|_{L^\infty(M \times M)}.$$

*To prove this it is enough to consider the correspondence $R \in \mathcal{R}(M, M)$ given by $R = \{(m, m),\ m \in M\}$.*

Notice that as an application, for the metric spaces $(P, d_{L_n^\varepsilon})$ and $(P, d_{\Delta_n^\varepsilon})$ discussed in §4.6, one has that

$$d_{\mathcal{GH}}\big((P, d_{L_n}), (P, d_{L_n^\varepsilon})\big) \leq \frac{\varepsilon}{2} \ and$$

$$d_{\mathcal{GH}}\big((P, d_{\Delta_n}), (P, d_{\Delta_n^\varepsilon})\big) \leq \frac{\varepsilon}{2}.$$

## 6.2 Stability and Convergence results

Our first result states that SL HC is stable in the Gromov-Hausdorff sense and it is a generalization of Lemma 15.

**Proposition 26** *For any two finite metric spaces $(X, d_X)$ and $(Y, d_Y)$*

$$d_{\mathcal{GH}}\big((X, d_X), (Y, d_Y)\big) \geq d_{\mathcal{GH}}\big(\mathfrak{T}^*(X, d_X), \mathfrak{T}^*(Y, d_Y)\big).$$

**Remark 27** *This Proposition generalizes Lemma 15. Notice for example that in case $X$ and $Y$ are finite, they need not have the same number of points. This feature is important in order to be able to make sense of situations such as the one depicted in Figure 2 in pp. 8, where one is trying to capture the connectivity (i.e. clustering) properties of an underlying 'continuous' space by taking finitely (but increasingly) many samples from this space and applying some form of HC to this finite set. Theorem 28 below deals with exactly this situation. See Figure 20.*

Let $(Z, d_Z)$ be a compact metric space. Given a finite index set $A$ and a (finite) collection of disjoint compact subsets of $Z$, $\{U^{(\alpha)}\}_{\alpha \in A}$, let $W_A : A \times A \to \mathbb{R}^+$ be given by

$$(\alpha, \alpha') \mapsto \min_{\substack{z \in U^{(\alpha)} \\ z' \in U^{(\alpha')}}} d_Z(z, z').$$

A metric space $(A, d_A)$ arises from this construction, where $d_A = \mathcal{L}(W_A)$. We say that $(A, d_A)$ is the metric space with underlying set $A$ **arising** from $\{U^{(\alpha)}\}_{\alpha \in A}$. Notice that $\operatorname{sep}(A, d_A)$ equals the minimal separation between any two sets $U^{(\alpha)}$ and $U^{(\alpha')}$ ($\alpha \neq \alpha'$). More precisely,

$$\operatorname{sep}(A, d_A) = \min_{\substack{\alpha, \alpha' \in A, \\ \alpha \neq \alpha'}} \min_{\substack{z \in U^{(\alpha)} \\ z' \in U^{(\alpha')}}} d_Z(z, z').$$

We now state a metric stability and convergence result, see Figure 20. The proof of this result is deferred to §A.

**Theorem 28** *Assume $(Z, d_Z)$ is a compact metric space. Let $X$ and $X'$ be any two finite subsets of $Z$ and let $d_X = d_{Z|_{X \times X}}$ and $d_{X'} = d_{Z|_{X' \times X'}}$. Write $\mathfrak{T}^*(X, d_X) = (X, u_X)$ and $\mathfrak{T}^*(X', d_{X'}) = (X', u_{X'})$. Then,*

41
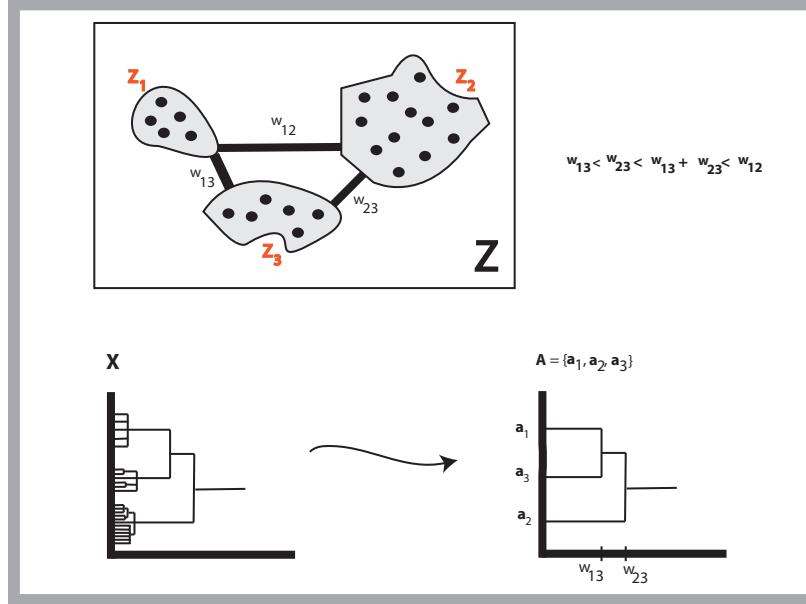
Figure 20: Illustration of Theorem 28. *Top*: A space $Z$ composed of 3 disjoint path connected parts, $Z^{(1)}, Z^{(2)}$ and $Z^{(3)}$. The black dots are the points in the finite sample $X$. In the figure, $w_{ij} = W(i,j)$, $1 \leqslant i \neq j \leqslant 3$. *Bottom Left*: The dendrogram representation of $(X, u_X)$. *Bottom Right* The dendrogram representation of $(Z, u_Z)$. Note that $d_Z(z_1, z_2) = w_{13} + w_{23}$, $d_Z(z_1, z_3) = w_{13}$ and $d_Z(z_2, z_3) = w_{23}$. As $r \to 0$, $(X, u_X) \to (Z, u_Z)$ in the Gromov-Hausdorff sense, see text for details.

1. *(Finite Stability)* $d_{\mathcal{GH}}\big((X, u_X), (X', u_{X'})\big) \leqslant d_{\mathcal{H}}^Z(X, Z) + d_{\mathcal{H}}^Z(X', Z)$.

2. *(Approximation bound) Assume in addition that $Z = \bigsqcup_{\alpha \in A} Z^{(\alpha)}$ where $A$ is a finite index set and $Z^{(\alpha)}$ are compact, disjoint and path-connected sets. Let $(A, d_A)$ be the finite metric space with underlying set $A$ arising from $\{Z^{(\alpha)}\}_{\alpha \in A}$. Let $\mathfrak{T}^*(A, d_A) = (A, u_A)$. Then, if $d_{\mathcal{H}}^Z(X, Z) < sep(A, d_A)/2$,*

$$d_{\mathcal{GH}}\big((X, u_X), (A, u_A)\big) \leqslant d_{\mathcal{H}}^Z(X, Z).$$

3. *(Convergence) Under the hypotheses of (2), let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of finite subsets of $Z$ s.t. $d_{\mathcal{H}}^Z(X_n, Z) \to 0$ as $n \to \infty$, and $d_{X_n}$ be the metric on $X_n$ given by the restriction of $d_Z$ to $X_n \times X_n$. Then, one has that*

$$d_{\mathcal{GH}}\big(\mathfrak{T}^*(X_n, d_{X_n}), (A, u_A)\big) \to 0 \text{ as } n \to \infty.$$

**Remark 29 (Interpretation of the statement)** *Assertion (1) guarantees that if $X, X'$ are both dense samples of $Z$, then the result of applying $\mathfrak{T}^*$ to both sets are very close in the Gromov-Hausdorff sense.*

*Assertions (2) and (3) identify the limiting behavior of the construction $\mathfrak{T}^*(X_n, d_{X_n})$ as $X_n$ becomes denser and denser in $X$, see Figure 20.*

### 6.3 A probabilistic convergence result

In this section, we prove a precise result which describes how the dendrograms attached to compact metric spaces by single linkage clustering can be obtained as the limits of the dendrograms attached to finite subsets of the metric space. The result is by necessity probabilistic in nature. This kind of result is of great importance, since we are often interested in infinite metric spaces but typically do not have access to more than finitely many random samples from the metric space.

Theorem 30 and Corollary 32 below proves that for random i.i.d. observations $\mathbb{X}_n = \{x_1, \dots, x_n\}$ with probability distribution $\mu$ compactly supported in a metric space $(X, d)$, the result $(\mathbb{X}_n, u_{\mathbb{X}_n})$ of applying single linkage clustering to $(\mathbb{X}_n, d)$ converges almost surely in the Gromov-Hausdorff sense to an ultrametric space that recovers the multiscale structure of the *support* of $\mu$, see Figure 20. This is a refinement of a previous observation of Hartigan (Hartigan, 1985) that SLHC is insensitive to the distribution of mass of $\mu$ in its support.

The proof of this theorem relies on Theorem 34, a probabilistic covering theorem of independent interest. In order to state and prove our theorems we make use of the formalism of **metric measure spaces**.

A triple $(X, d_X, \mu_X)$, where $(X, d_X)$ is a metric space and $\mu_X$ is a Borel probability measure on $X$ with compact support will be called an **mm-space** (short for measure metric space). The support supp $[\mu_X]$ of a measure $\mu_X$ on $X$ is the minimal closed set $A$ (w.r.t. inclusion) s.t. $\mu_X(X \backslash A) = 0$. Measure metric spaces are considered in the work of Gromov and are useful in different contexts, see (Gromov, 2007, Chapter $3\frac{1}{2}$). For a mm-space $X$ let $f_X : \mathbb{R}^+ \to \mathbb{R}^+$ be defined by

$$r \mapsto \min_{x \in \text{supp}[X]} \mu_X(B_X(x, r)).$$

Note also that by construction $f_X(\cdot)$ in non-decreasing and $f_X(r) > 0$ for all $r > 0$. Let also $F_X : \mathbb{N} \times \mathbb{R}^+ \to \mathbb{R}^+$ be defined by $(n, \delta) \mapsto \frac{e^{-nf_X(\delta/4)}}{f_X(\delta/4)}$. Note that for fixed $\delta_0 > 0$, (1) $F_X(\cdot, \delta_0)$ is decreasing in its argument, and (2) $\sum_{n \in \mathbb{N}} F_X(n, \delta_0) < \infty$.

**Theorem 30** *Let $(Z, d_Z, \mu_Z)$ be a mm-space and write supp $[\mu_Z] = \bigcup_{\alpha \in A} U^{(\alpha)}$ for a finite index set $A$ and $\mathbf{U} = \{U^{(\alpha)}\}_{\alpha \in A}$ a collection of disjoint, compact, path-connected subsets of $Z$. Let $(A, d_A)$ be the metric space arising from $\mathbf{U}$ and let $\delta_A := sep(A, d_A)/2$.*

*For each $n \in \mathbb{N}$, let $\mathbb{Z}_n = \{z_1, z_2, \dots, z_n\}$ be a collection of $n$ independent random variables (defined on some probability space $\Omega$ with values in $Z$) with distribution $\mu_Z$, and let $d_{\mathbb{Z}_n}$ be the restriction of $d_Z$ to $\mathbb{Z}_n \times \mathbb{Z}_n$. Then, for $\zeta \geqslant 0$ and $n \in \mathbb{N}$,*

$$\mathbf{P}_{\mu_Z}\Big(d_{\mathcal{GH}}\big(\mathfrak{T}^*(\mathbb{Z}_n, d_{\mathbb{Z}_n}), \mathfrak{T}^*(A, d_A)\big) > \zeta\Big) \leqslant F_Z\big(n, \min(\zeta, \delta_A/2)\big). \qquad (6.14)$$

**Corollary 31** *Under the hypotheses of Theorem 30, for any pre-specified probability level* $p \in (0, 1)$ *and tolerance* $\zeta \geqslant 0$, *if*

$$n \geqslant \frac{\ln \frac{1}{1-p} - \ln f_X(\delta/4)}{f_X(\delta/4)},$$

*then* $\mathbf{P}_{\mu_Z}\Big(d_{\mathcal{GH}}\big(\mathfrak{T}^*(\mathbb{Z}_n, d_{\mathbb{Z}_n}), \mathfrak{T}^*(A, d_A)\big) \leqslant \zeta\Big) \geqslant p$, *where* $\delta := \min(\zeta, \delta_A/2)$.

**Corollary 32** *Under the hypotheses of Theorem 30,* $\mathfrak{T}^*(\mathbb{Z}_n, d_{\mathbb{Z}_n}) \xrightarrow{n} \mathfrak{T}^*(A, d_A)$ *in the Gromov-Hausdorff sense* $\mu_Z$-*almost surely.*

**Proof** [Proof of Corollary 32] The proof follows immediately from the expression for $F_X$ and the Borel-Cantelli Lemma. ∎

**Remark 33** *Note that the convergence theorem above implies that in the limit,* $\mathfrak{T}^*(X_n, d_{X_n})$ *only retains information about the support of the probability measure but not about the way the mass is distributed inside the support, cf. (Hartigan, 1985).*

**Example 13** ($Z \subset \mathbb{R}^d$) *Let* $\rho : \mathbb{R}^d \to \mathbb{R}^+$ *be a density function with compact, support* $Z$ *and* $\mu$ *be its associated probability measure. Then* $(\mathbb{R}^d, \|\cdot\|, \mu)$ *satisfies the assumptions in the theorem. If one makes additional smoothness assumptions on* $\rho$, *in this particular case one can relate* $F_Z(n, \zeta)$ *to geometrical properties of the boundary of* $supp\,[\rho]$.

**Example 14** ($Z$ **is a Riemannian manifold**) *In more generality,* $Z$ *could be a Riemannian manifold and* $\mu$ *a probability measure absolutely continuous w.r.t. to the Riemannian area measure on* $Z$.

## 7. Discussion

We have obtained novel characterization, stability and convergence theorems for SL HC. Our theorems contemplate both the deterministic and the stochastic case. Our characterization theorem can be interpreted as a relaxation of Kleinberg's impossibility result for standard clustering methods in that by allowing the output of clustering methods to be hierarchical, one obtains existence and uniqueness.

Our stability results seem to be novel and complement classical observations that CL and AL are discontinuous as maps from finite metric spaces into dendrograms.

Our convergence results also seem to be novel and they refine a previous observation by Hartigan about the information retained about an underlying density by SL clustering of an i.i.d. collection of samples from that density. Our setting for the stochastic convergence results is quite general in that we do not assume the underlying space to be a smooth manifold and we do not assume the underlying probability measure to have a density with respect to any reference measure.

We understand that SL HC is not sensitive to variations in the density (see also (Hartigan, 1981)). In our future work we will be looking at ways of further relaxing the notions of clustering that can cope with the problem of detecting "dense" clusters, in the same spirit as (Wishart, 1969; Stuetzle, 2003). A follow up paper (Carlsson and Mémoli, 2009) presents a systematic treatment of this with a more general framework.

Some recent works have also addressed the characterization of clustering schemes in the hierarchical case. The authors of the present paper reported a characterization for proximity dendrograms (Carlsson and Mémoli, 2008) using the language of category theory. Zadeh and Ben-David (Zadeh and Ben-David, 2009) gave a characterization for threshold dendrograms.[5] More classical is the work of Jardine and Sibson (Jardine and Sibson, 1971) who also ultimately view HC methods as maps form finite metric spaces to finite ultrametric spaces.

It is interesting to consider the situation when one requires the map $\phi$ in our characterization theorem (Theorem 18) to be 1 to 1 on points. In this case, a much wider class of hierarchical schmemes becomes possible including for example a certain version of *clique clustering*. The restriction on the nature of $\phi$ would be called restriction of *functoriality* by a mathematician. The classification question of clustering methods that arises becomes mathematically interesting and we are currently exploring it (Carlsson and Mémoli, 2009.; Carlsson and Mémoli, 2008).

## Acknowledgments

---

5. Recall that the difference between these two types of dendrograms is that proximity dendrograms retain the linkage value at which mergings take place whereas threshold dendrograms only record the order, see Remark 3.

# Appendix A. Proofs

**Proof** [Proof of Proposition 8] The claim follows from the following claim, which we prove by induction on $i$:

*Claim*: For all $i \geqslant 2$, $x, x' \in X$ are s.t. there exists $\mathcal{B} \in \Theta_i$ with $x, x' \in \mathcal{B}$ if and only if $x \sim_{R_{i-1}} x'$.

**Proof** [Proof of the Claim] For $i = 2$ the claim is clearly true. Fix $i > 2$.

Assume that $x, x' \in X$ and $\mathcal{B} \in \Theta_{i+1}$ are such that $x, x' \in \mathcal{B}$. If $x, x'$ belong to the same block of $\Theta_i$ there is nothing to prove. So, assume that $x \in \mathcal{A}$ and $x' \in \mathcal{A}'$ with $\mathcal{A} \neq \mathcal{A}'$ and $\mathcal{A}, \mathcal{A}' \in \Theta_i$. Then, it must be that there exist blocks $\mathcal{A} = \mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_s = \mathcal{A}'$ of $\Theta_i$ s.t. $\ell^{\mathrm{SL}}(\mathcal{A}_t, \mathcal{A}_{t+1}) \leqslant R_i$ for $t = 1, \ldots, s - 1$. Pick $x_1, y_1 \in \mathcal{A}_1$, $x_2, y_2 \in \mathcal{A}_2$, $\ldots$, $x_s, y_s \in \mathcal{A}_s$ s.t. $x_1 = x$ and $y_s = x'$ and $d(y_t, x_{t+1}) = \ell^{\mathrm{SL}}(\mathcal{A}_t, \mathcal{A}_{t+1}) \leqslant R_i$ for $t = 1, \ldots, s - 1$, see the Figure 21.
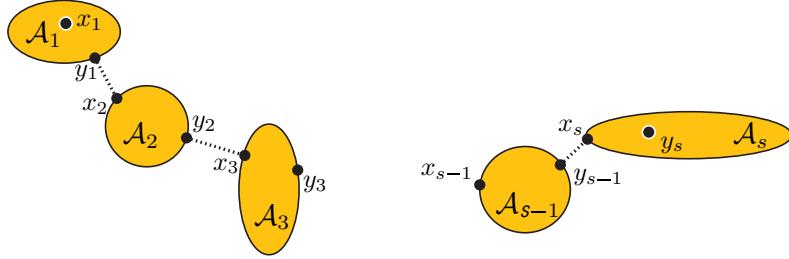


Figure 21: Construction used in the proof of Proposition 8.

Notice that by the inductive hypothesis we have $x_t \sim_{R_{i-1}} y_t$ for $t = 1, \ldots, s$. It follows that $x \sim_r x'$ for $r = \max(R_i, R_{i-1})$. By Proposition 5, $r = R_i$ and hence $x \sim_{R_i} x'$.

Assume now that $x \sim_{R_i} x'$. If $x, x'$ belong to the same block of $\Theta_i$ there's nothing to prove since $\Theta_{i+1}$ is coarser than $\Theta_i$ and hence $x, x'$ will also belong to the same block of $\Theta_{i+1}$. Assume then that $x \in \mathcal{B}$ and $x' \in \mathcal{B}'$ for $\mathcal{B}, \mathcal{B}' \in \Theta_i$ with $\mathcal{B} \neq \mathcal{B}'$. Let $x = x_1, x_2, \ldots, x_s = x'$ be points in $X$ with $d(x_t, x_{t+1}) \leqslant R_i$ for $t = 1, \ldots, s - 1$. Also, for $t = 1, \ldots, s - 1$ let $\mathcal{B}_t$ be the block of $\Theta_i$ to which $x_t$ belongs. But then, by construction

$$R_i \geqslant d(x_t, x_{t+1}) \geqslant \min_{z \in \mathcal{B}_t, z' \in \mathcal{B}_{t+1}} d(z, z') = \ell^{\mathrm{SL}}(\mathcal{B}_t, \mathcal{B}_{t+1}) \text{ for } t = 1, \ldots, s - 1,$$

and hence $\mathcal{B}_1 \sim_{\ell^{\mathrm{SL}}, R_i} \mathcal{B}_s$. In particular, $\mathcal{B}_1 \cup \mathcal{B}_s \subset \mathcal{A}$ for some $\mathcal{A} \in \Theta_{i+1}$ and thus $x, x'$ belong to the same block in $\Theta_{i+1}$. ∎

∎

**Proof** [Proof of Lemma 10] Obviously $u_\theta$ is non-negative. Pick $x, x', x'' \in X$ and let $r_1, r_2 \geqslant 0$ be s.t. $x, x'$ belong to the same block of $\theta(r_1)$ and $x', x''$ belong to the same block of $\theta(r_2)$. These numbers clearly exist by condition (2) in the definition of dendrograms. Then, there exist a block $\mathcal{B}$ of $\theta\big(\max(r_1, r_2)\big)$ s.t. $x, x'' \in \mathcal{B}$ and hence $u_\theta(x, x'') \leqslant max(r_1, r_2)$. The conclusion follows since $r_1 \geqslant u_\theta(x, x')$ and $r_2 \geqslant u_\theta(x', x'')$ are arbitrary.

Now, let $x, x' \in X$ be such that $u_\theta(x, x') = 0$. Then $x, x'$ are in the same block of $\theta(0)$. Condition (1) in the definition of dendrograms implies that $x = x'$. ■

**Proof** [Proof of Lemma 12] Pick $x, x' \in X$ and let $r := u_{\theta*}(x, x')$. Then, according to (4.3), there exist $x_0, x_1, \ldots, x_t \in X$ with $x_0 = x$, $x_t = x'$ and $\max_i d(x_i, x_{i+1}) \leqslant r$. From (4.4) we conclude that then $u^*(x, x') \leqslant r$ as well. Assume now that $u^*(x, x') \leqslant r$ and let $x_0, x_1, \ldots, x_t \in X$ be s.t. $x_0 = x$, $x_t = x'$ and $\max_i d(x_i, x_{i+1}) \leqslant r$. Then, $x \sim_r x'$ and hence again by recalling (4.3), $u_{\theta*}(x, x') \leqslant r$. This finishes the proof. ■

**Proof** [Proof of Theorem 18] Pick $(X, d) \in \mathcal{X}$. Write $\mathfrak{T}(X, d) = (X, u)$ and $\mathfrak{T}^*(X, d) = (X, u^*)$.

**(A)** We prove that $u^*(x, x') \geqslant u(x, x')$ for all $x, x' \in X$. Pick $x, x' \in X$ and let $\delta := u^*(x, x')$. Let $x = x_0, \ldots, x_n = x'$ be s.t.

$$\max_i d(x_i, x_{i+1}) = u^*(x, x') = \delta.$$

Consider the two point metric space $(Z, e) := (\{p, q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right))$. Fix $i \in \{0, \ldots, n-1\}$. Consider $\phi : \{p, q\} \to X$ given by $p \mapsto x_i$ and $q \mapsto x_{i+1}$. By condition (I) we have $\mathfrak{T}(Z_\delta) = Z_\delta$. Note that $\delta = e(p, q) \geqslant d(\phi(p), \phi(q)) = d(x_i, x_{i+1})$ and hence by condition (II),

$$\delta \geqslant u(x_i, x_{i+1}).$$

Then, since $i$ was arbitrary, we obtain $\delta \geqslant \max_i u(x_i, x_{i+1})$. Now, since $u$ is an ultrametric on $X$, we know that $\max_i u(x_i, u_{i+1}) \geqslant u(x, x')$ and hence $\delta \geqslant u(x, x')$.

**(B)** We prove that $u^*(x, x') \leqslant u(x, x')$ for all $x, x' \in X$. Fix $r > 0$. Let $(X_r, d_r)$ be the metric space with underlying set $X_r$ given by the equivalence classes of $X$ under the relation $x \sim_r x'$. Let $\phi_r : X \to X_r$ be given by $x \mapsto [x]_r$ where $[x]_r$ denotes the equivalence class of $x$ under $\sim_r$. Let $\tilde{d}_r : X_r \times X_r \to \mathbb{R}^+$ be given by

$$\tilde{d}_r(z, z') = \min_{\substack{x \in \phi_r^{-1}(z) \\ x' \in \phi_r^{-1}(z')}} d(x, x')$$

and let $d_r = \mathcal{L}(\tilde{d}_r)$. Note that, by our construction, $\phi_r$ is such that for all $x, x' \in X$,

$$d(x, x') \geqslant d_r(\phi_r(x), \phi_r(x')).$$

Indeed, assume the contrary. Then for some $x, x' \in X$ one has that $d(x, x') < d_r(\phi_r(x), \phi_r(x'))$. But, from the definition of $d_r$ it follows that $d(x, x') < d_r(\phi_r(x), \phi_r(x')) \leqslant \tilde{d}_r(\phi_r(x), \phi_r(x')) = \min\{d(\overline{x}, \overline{x}'), \text{ s.t. } \overline{x} \sim_r x; \overline{x}' \sim_r x'\}$. This is a contradiction since $x \sim_r x$ and $x' \sim_r x'$.

Write $\mathfrak{T}(X_r, d_r) = (X_r, u_r)$. Then, by condition (III),

$$u(x, x') \geqslant u_r(\phi_r(x), \phi_r(x')) \tag{A.15}$$

for all $x, x' \in X$. Note that

$$\text{sep}(X_r, d_r) > r. \tag{A.16}$$

Indeed, for otherwise, there would be two points $x, x' \in X$ with $[x]_r \neq [x']_r$ and $r \geqslant d(x, x') \geqslant u^*(x, x')$. But this gives a contradiction by Remark 13.

*Claim*: $u^*(x, x') > r$ implies that $u_r(\phi_r(x), \phi_r(x')) > r$.

Assuming the claim, let $x, x' \in X$ be s.t. $u^*(x, x') > r$, then by equation (A.15),

$$u(x, x') \geqslant u_r(\phi_r(x), \phi_r(x')) > r.$$

That is, we have obtained that for any $r > 0$,

$$\{(x, x') \text{ s.t. } u^*(x, x') > r\} \subseteq \{(x, x') \text{ s.t. } u(x, x') > r\},$$

which implies that $u^*(x, x') \leqslant u(x, x')$ for all $x, x' \in X$.

*Proof of the claim.* Let $x, x' \in X$ be s.t. $u^*(x, x') > r$. Then, $[x]_r \neq [x']_r$. By definition of $\phi_r$, also, $\phi_r(x) \neq \phi_r(x')$ and hence, by condition (III) and equation (A.16):

$$u_r(\phi_r(x), \phi_r(x')) \geqslant \text{sep}(X_r, d_r) > r.$$

∎

**Proof** [Proof of Proposition 26] Write $\mathfrak{T}^*(X, d_X) = (X, u_X)$ and $\mathfrak{T}^*(Y, d_Y) = (Y, u_Y)$. Let $\eta = d_{\mathcal{GH}}\big((X, d_X), (Y, d_Y)\big)$ and $R \in \mathcal{R}(X, Y)$ s.t. $|d_X(x, x') - d_Y(y, y')| \leqslant 2\eta$ for all $(x, y), (x', y') \in R$. Fix $(x, y)$ and $(x', y') \in R$. Let $x_0, \ldots, x_m \in X$ be s.t. $x_0 = x$, $x_m = x'$ and $d_X(x_i, x_{i+1}) \leqslant u_X(x, x')$ for all $i = 0, \ldots, m-1$. Let $y = y_0, y_1, \ldots, y_{m-1}, y_m = y' \in Y$ be s.t. $(x_i, y_i) \in R$ for all $i = 0, \ldots, m$ (this is possible by definition of $R$). Then, $d_Y(y_i, y_{i+1}) \leqslant d_X(x_i, x_{i+1}) + \eta \leqslant u_X(x, x') + \eta$ for all $i = 0, \ldots, m-1$ and hence $u_Y(y, y') \leqslant u_X(x, x') + 2\eta$. By exchanging the roles of $X$ and $Y$ one obtains the inequality $u_X(x, x') \leqslant u_Y(y, y') + 2\eta$. This means $|u_X(x, x') - u_Y(y, y')| \leqslant 2\eta$. Since $(x, y), (x', y') \in R$ are arbitrary, and upon recalling the expression of the Gromov-Hausdorff distance given by (6.12) we obtain the desired conclusion. ∎

**Proof** [Proof of Theorem 28] By Proposition 26 and the triangle inequality for the Gromov-Hausdorff distance,

$$d_{\mathcal{GH}}(X, Z) + d_{\mathcal{GH}}(X', Z) \geqslant d_{\mathcal{GH}}((X, u_X)), (X', u_{X'}))).$$

Now, (1) follows from Remark 25.

We now prove the second claim. Let $\delta > 0$ be s.t. $\min_{\alpha \neq \beta} W_A(\alpha, \beta) \geqslant \delta$. For each $z \in Z$ let $\alpha(z)$ denote the index of the path connected component of $Z$ s.t. $z \in Z^{(\alpha(z))}$. Since $r := d_{\mathcal{H}}^Z(X, Z) < \frac{\delta}{2}$, it is clear that $\#(Z^{(\alpha)} \cap X) \geqslant 1$ for all $\alpha \in A$. It follows that $R = \{(x, \alpha(x)) | x \in X\}$ belongs to $\mathcal{R}(X, A)$. We prove below that for all $x, x' \in X$,

$$u_A(\alpha(x), \alpha(x')) \overset{(I)}{\leqslant} u_X(x, x') \overset{(II)}{\leqslant} u_A(\alpha(x), \alpha(x')) + 2r.$$

By putting (I) and (II) together we will have $d_{\mathcal{GH}}((X, u_X), (A, u_A)) \leqslant r$.

Let's prove (I). It follows immediately from the definition of $d_A$ and $W_A$ that for all $y, y' \in X$,

$$W_A(\alpha(y), \alpha(y')) \leqslant d_X(y, y').$$

From the definition of $d_A$ it also follows that $W_A(\alpha, \alpha') \geqslant d_A(\alpha, \alpha')$ for all $\alpha, \alpha' \in A$. Then, in order to prove (I) pick $x_0, \ldots, x_m$ in $X$ with $x_0 = x$, $x_m = x'$ and $\max_i d_X(x_i, x_{i+1}) \leqslant u_X(x, x')$. Consider the points in $A$ given by

$$\alpha(x) = \alpha(x_0), \alpha(x_1), \ldots, \alpha(x_m) = \alpha(x').$$

Then,

$$d_A(\alpha(x_i), \alpha(x_{i+1})) \leqslant W_A(\alpha(x_i), \alpha(x_{i+1})) \leqslant d_X(x_i, x_{i+1}) \leqslant u_X(x, x')$$

for $i = 0, \ldots, m - 1$ by the observations above. Then, $\max_i d_A(\alpha(x_i), \alpha(x_{i+1})) \leqslant d_X(x, x')$ and by recalling the definition of $u_A(\alpha(x), \alpha(x'))$ we obtain (I).

We now prove (II). Assume first that $\alpha(x) = \alpha(x') = \alpha$. Fix $\epsilon_0 > 0$ small. Let $\gamma : [0, 1] \to Z^{(\alpha)}$ be a continuous path s.t. $\gamma(0) = x$ and $\gamma(1) = x'$. Let $z_1, \ldots, z_m$ be points on image($\gamma$) s.t. $z_0 = x$, $z_m = x'$ and $d_X(z_i, z_{i+1}) \leqslant \epsilon_0$, $i = 0, \ldots, m - 1$. By hypothesis, one can find $x = x_0, x_1, \ldots, x_{m-1}, x_m = x'$ s.t. $d_Z(x_i, z_i) \leqslant r$. Thus,

$$\max_i d_X(x_i, x_{i+1}) \leqslant \epsilon_0 + 2r$$

and hence $u_X(x, x') \leqslant \epsilon_0 + 2r$. Let $\epsilon_0 \to 0$ to obtain the desired result.

Now if $\alpha = \alpha(x) \neq \alpha(x') = \beta$, let $\alpha_0, \alpha_1, \ldots, \alpha_l \in A$ be s.t. $\alpha_0 = \alpha(x)$, $\alpha_l = \alpha(x')$ and $d_A(\alpha_j, \alpha_{j+1}) \leqslant u_A(\alpha, \beta)$ for $j = 0, \ldots, l - 1$.

By definition of $d_A$, for each $j = 0, \ldots, l - 1$ one can find a *chain*

$$C_j = \left\{ \alpha_j^{(0)}, \ldots, \alpha_j^{(r_j)} \right\} \quad \text{s.t.} \quad \alpha_j^{(0)} = \alpha_j, \; \alpha_j^{(r_j)} = \alpha_{j+1}$$
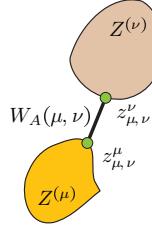
and

$$\sum_{i=0}^{r_j-1} W_A\left(\alpha_j^{(i)}, \alpha_j^{(i+1)}\right) = d_A(\alpha_j, \alpha_{j+1}) \leqslant u_A(\alpha, \beta).$$

Since $W_A$ takes non-negative values, then, for fixed $j \in \{0, \ldots, l-1\}$, it follows that

$$W_A(\alpha_j^{(i)}, \alpha_j^{(i+1)}) \leqslant u_A(\alpha, \beta) \text{ for all } i = 0, \ldots, r_j - 1.$$

Consider the chain $C = \{\hat{\alpha}_0, \ldots, \hat{\alpha}_s\}$ in $A$ joining $\alpha$ to $\beta$ given by the concatenation of all the $C_j$. By eliminating repeated consecutive elements in $C$, if necessary, one can assume that $\hat{\alpha}_i \neq \hat{\alpha}_{i+1}$. By construction $W_A(\hat{\alpha}_i, \hat{\alpha}_{i+1}) \leqslant u_A(\alpha, \beta)$ for $i \in \{0, \ldots, s-1\}$, and $\hat{\alpha}_0 = \alpha$, $\hat{\alpha}_s = \beta$. We will now lift $C$ into a chain in $Z$ joining $x$ to $x'$. Note that by compactness, for all $\nu, \mu \in A$, $\nu \neq \mu$ there exist $z_{\nu,\mu}^\nu \in Z^{(\nu)}$ and $z_{\nu,\mu}^\mu \in Z^{(\mu)}$ s.t. $W_A(\nu, \mu) = d_Z(z_{\nu,\mu}^\nu, z_{\nu,\mu}^\mu)$.



Consider the chain $G$ in $Z$ given by

$$G = \left\{ x, z_{\hat{\alpha}_0, \hat{\alpha}_1}^{\hat{\alpha}_0}, z_{\hat{\alpha}_0, \hat{\alpha}_1}^{\hat{\alpha}_1}, \ldots, z_{\hat{\alpha}_{s-1}, \hat{\alpha}_s}^{\hat{\alpha}_s}, x' \right\}.$$

For each point $g \in G \subset Z$ pick a point $x(g) \in X$ s.t. $d_Z(g, x(g)) \leqslant r$. Note that this is possible by definition of $r$ and also, that $x(g) \in Z^{(\alpha(g))}$ since $r < \delta/2$.

Let $G' = \{x_0, x_1, \ldots, x_t\}$ be the resulting path in $X$. Notice that if $\alpha(x_k) \neq \alpha(x_{k+1})$ then

$$d_X(x_k, x_{k+1}) \leqslant 2r + W_A(\alpha(x_k), \alpha(x_{k+1})) \tag{A.17}$$

by the triangle inequality. Also, by construction, for $k \in \{0, \ldots, t-1\}$,

$$W_A(\alpha(x_k), \alpha(x_{k+1})) \leqslant u_A(\alpha, \beta). \tag{A.18}$$

Now, we claim that

$$u_X(x, x') \leqslant \max_k W_A(\alpha(x_k), \alpha(x_{k+1})) + 2r. \tag{A.19}$$

This claim will follow from (A.17) and the simple observation that

$$u_X(x, x') \leqslant \max_k u_X(x_k, x_{k+1}) \leqslant \max_k d_X(x_k, x_{k+1})$$

which in turn follows from the fact that $u_X$ is the ultrametric on $X$ defined by (4.4), see remarks in Example 7. If $\alpha(x_k) = \alpha(x_{k+1})$ we already proved that $u_X(x_k, x_{k+1}) \leqslant 2r$.

50

If on the other hand $\alpha(x_k) \neq \alpha(x_{k+1})$ then (A.19) holds. Hence, we have that without restriction, for all $x, x' \in X$,

$$u_X(x, x') \leqslant \max_k W_A(\alpha(x_k), \alpha(x_{k+1})) + 2r.$$

and hence the claim. Combine this fact with (A.18) to conclude the proof of (II). Claim (3) follows immediately from (2).

∎

## A.1  The proof of Theorem 30

We will make use of the following general covering theorem in the proof of Theorem 30.

**Theorem 34** *Let $(X, d, \mu)$ be an mm-space and $\mathbb{X}_n = \{x_1, x_2, \ldots, x_n\}$ a collection of $n$ independent random variables (defined on some probability space $\Omega$, and with values in $X$) and identically distributed with distribution $\mu$. Then, for any $\delta > 0$,*

$$\mathbf{P}_\mu\left(d_{\mathcal{H}}^X\left(\mathbb{X}_n, supp\left[\mu_X\right]\right) > \delta\right) \leqslant F_X(n, \delta).$$

**Proof**  Consider first a fixed point $x \in \text{supp}\left[\mu_X\right]$ and $h > 0$. Then, since $x_1, \ldots, x_n$ are i.i.d., for all $i$, $\mathbf{P}_\mu(x_i \in B_X(x, h)) = \mu(B_X(x, h))$. We then have:

$$
\begin{aligned}
\mathbf{P}_\mu\left(\left\{x \notin \bigcup_{i=1}^n B_X(x_i, h)\right\}\right) &= \mathbf{P}_\mu\left(\bigcap_{i=1}^n \{x \notin B_X(x_i, h)\}\right) \\
&= \mathbf{P}_\mu\left(\bigcap_{i=1}^n \{x_i \notin B_X(x, h)\}\right) \\
&= \prod_{i=1}^n \mathbf{P}_\mu\left(\{x_i \notin B_X(x, h)\}\right) \quad \text{(by independence)} \\
&= \left(1 - \mu_X(B_X(x, h))\right)^n \\
&\leqslant \left(1 - f_X(h)\right)^n.
\end{aligned}
$$
(A.20)

We now obtain a similar bound for the probability that a ball of radius $\delta/2$ around $x$ is within $\delta$ of a point in $\mathbb{X}_n$. Notice that the following inclusion of events holds:

$$\left\{B_X(x, \delta/2) \subset \bigcup_{i=1}^n B_X(x_i, \delta)\right\} \supseteq \left\{x \in \bigcup_{i=1}^n B_X(x_i, \delta/2)\right\}.$$
(A.21)

Indeed, assume that the event $\{x \in \bigcup_{i=1}^n B_X(x_i, \delta/2)\}$ holds. Then, $x \in B_X(x_i, \delta/2)$ for some $i \in \{1, \ldots, n\}$. Pick any $x' \in B_X(x, \delta/2)$, then by the triangle inequality, $d_X(x', x_i) \leqslant d_X(x', x) + d_X(x, x_i) < \delta/2 + \delta/2 = \delta$, thus $x' \in B_X(x_i, \delta)$. Since $x'$ is an

arbitrary point in $B_X(x, \delta/2)$ we are done. Now, from (A.21) and (A.20) (for $h = \delta/2$) above, we find

$$\mathbf{P}_\mu \left( \left\{ B_X(x, \delta/2) \nsubseteq \bigcup_{i=1}^{n} B_X(x_i, \delta) \right\} \right) \leqslant (1 - f_X(\delta/2))^n. \qquad \text{(A.22)}$$

Now, consider a maximal $\delta/4$-packing of supp $[\mu_X]$ by balls with centers $\{p_1, \ldots, p_N\}$. Then, clearly, supp $[\mu_X] = \bigcup_{j=1}^{N} B_X(p_j, \delta/2)$. Such a packing always exists since supp $[\mu_X]$ is assumed to be compact (Burago et al., 2001). Notice that $N$, the cardinality of the packing, can be bounded by $1/f_X(\delta/4)$. Indeed, since $B_X(p_\alpha, \delta/4) \cap B_X(p_\beta, \delta/4) = \varnothing$ for $\alpha \neq \beta$, we have

$$
\begin{aligned}
1 = \mu_X(\text{supp} [\mu_X]) &= \mu_X \left( \bigcup_{j=1}^{N} B_X(p_j, \delta/2) \right) \\
&\geqslant \mu_X \left( \bigcup_{j=1}^{N} B_X(p_j, \delta/4) \right) \\
&= \sum_{j=1}^{N} \mu_X(B_X(p_j, \delta/4)) \\
&\geqslant N \cdot f_X(\delta/4)
\end{aligned}
$$

and the claim follows. Now, we finish the proof by first noting that since $\mathbb{X}_n \subset$ supp $[\mu_X]$, the following inclusion of events holds:

$$\{d_{\mathcal{H}}^X(\mathbb{X}_n, \text{supp} [\mu_X]) > \delta\} \subseteq \left\{ X \nsubseteq \bigcup_{i=1}^{n} B_X(x_i, \delta) \right\}$$

and hence, using the union bound, then (A.22) and the bound on $N$, we find:

$$
\begin{aligned}
\mathbf{P}_\mu\left(d_{\mathcal{H}}^X\left(\mathbb{X}_n, \operatorname{supp}\left[\mu_X\right]\right) > \delta\right) &\leqslant \mathbf{P}_\mu\left(X \nsubseteq \bigcup_{i=1}^n B_X(x_i, \delta)\right) \\
&= \mathbf{P}_\mu\left(\bigcup_{j=1}^N \left\{B_X(p_j, \delta/2) \nsubseteq \bigcup_{i=1}^n B_X(x_i, \delta)\right\}\right) \\
&\leqslant N \cdot \max_{j=1,\ldots,N} \mathbf{P}_\mu\left(B_X(p_j, \delta/2) \nsubseteq \bigcup_{i=1}^n B_X(x_i, \delta)\right) \\
&\leqslant \frac{1}{f_X(\delta/4)} \cdot (1 - f_X(\delta/2))^n \\
&\leqslant \frac{1}{f_X(\delta/4)} \cdot (1 - f_X(\delta/4))^n \quad (\text{since } f_X(\cdot) \text{ is non-decreasing}) \\
&\leqslant \frac{1}{f_X(\delta/4)} e^{-n\, f_X(\delta/4)} \quad (\text{by the inequality } (1-t) \leqslant e^{-t},\ \forall t \in \mathbb{R}) \\
&= F_X(n, \delta) \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\text{(A.23)}
\end{aligned}
$$

thus concluding the proof. ∎

**Proof** [Proof of Theorem 30] For each $n \in \mathbb{N}$, introduce the random variables $r_n := d_{\mathcal{H}}^Z\left(\mathbb{Z}_n, \operatorname{supp}\left[\mu_Z\right]\right)$ and $g_n := d_{\mathcal{GH}}\left(\mathfrak{T}^*(\mathbb{Z}_n, d_{\mathbb{Z}_n}), \mathfrak{T}^*(A, d_A)\right)$. Fix $\zeta' = \delta_A/2$. Note that by Theorem 28 (2) once $r_n \leqslant \zeta$ for some $\zeta \leqslant \zeta'$ we know that $g_n \leqslant r_n$ a.s. Hence, we have

$$
\mathbf{P}\left(g_n > \zeta\right) \leqslant \mathbf{P}\left(r_n > \zeta\right) \leqslant F_X(n, \zeta), \quad\quad\quad\quad\text{(A.24)}
$$

where the last inequality follows from Lemma 34.

Meanwhile, if $\zeta \geqslant \zeta'$ is arbitrary, then $\mathbf{P}\left(g_n > \zeta\right) \leqslant \mathbf{P}\left(g_n > \zeta'\right)$. By (A.24) (for $\zeta = \zeta'$) we find $\mathbf{P}\left(g_n > \zeta\right) \leqslant \mathbf{P}\left(r_n > \zeta'\right) \leqslant F_X(n, \zeta')$ for all $\zeta \geqslant \zeta'$. Thus, we have found that

$$
\mathbf{P}\left(g_n > \zeta\right) \leqslant \begin{cases} F_X(n, \zeta') & \text{for } \zeta \geqslant \zeta'. \\ F_X(n, \zeta) & \text{for } \zeta \leqslant \zeta'. \end{cases}
$$

The conclusion now follows. ∎

# References

Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In Gábor Lugosi and Hans-Ulrich Simon, editors, *COLT*, volume 4005 of *Lecture Notes in Computer Science*, pages 5–19. Springer, 2006. ISBN 3-540-35294-5.

Fred L. Bookstein, Barry Chernoff, Ruth L. Elder, Julian M. Humphries Jr., Gerald R. Smith, and Richard E. Strauss. *Morphometrics in evolutionary biology : the geometry of size and shape change, with examples from fishes.* Academy of Natural Sciences of Philadelphia., 1985.

Martin R. Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences].* Springer-Verlag, Berlin, 1999. ISBN 3-540-64324-9.

D. Burago, Y. Burago, and S. Ivanov. *A Course in Metric Geometry*, volume 33 of *AMS Graduate Studies in Math.* American Mathematical Society, 2001.

G. Carlsson and F. Mémoli. Persistent Clustering and a Theorem of J. Kleinberg. *ArXiv e-prints*, August 2008.

G. Carlsson and F Mémoli. Multiparameter clustering methods. In Claus Weihs Hermann Locarek-Junge, editor, *'Classification as a Tool for Research'. Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft fr Klassifikation e.V.*, Berlin-Heidelberg-New York, to appear 2009. Springer.

Gunnar Carlsson and Facundo Mémoli. Classifying clustering schemes. Technical report, 2009.

M. Gromov. Hyperbolic groups. In *Essays in group theory*, volume 8 of *Math. Sci. Res. Inst. Publ.*, pages 75–263. Springer, New York, 1987.

Misha Gromov. *Metric structures for Riemannian and non-Riemannian spaces.* Modern Birkhäuser Classics. Birkhäuser Boston Inc., Boston, MA, english edition, 2007. ISBN 978-0-8176-4582-3; 0-8176-4582-9. Based on the 1981 French original, With appendices by M. Katz, P. Pansu and S. Semmes, Translated from the French by Sean Michael Bates.

J. A. Hartigan. Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.*, 76(374):388–394, 1981. ISSN 0162-1459.

J. A. Hartigan. Statistical theory in clustering. *J. Classification*, 2(1):63–76, 1985. ISSN 0176-4268.

Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data.* Prentice Hall Advanced Reference Series. Prentice Hall Inc., Englewood Cliffs, NJ, 1988. ISBN 0-13-022278-X.

Nicholas Jardine and Robin Sibson. *Mathematical taxonomy.* John Wiley & Sons Ltd., London, 1971. Wiley Series in Probability and Mathematical Statistics.

D. G. Kendall, D. Barden, T. K. Carne, and H. Le. *Shape and shape theory.* Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1999. ISBN 0-471-96823-4.

Jon M. Kleinberg. An impossibility theorem for clustering. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 446–453. MIT Press, 2002. ISBN 0-262-02550-7.

G. N. Lance and W. T. Williams. A general theory of classificatory sorting strategies 1. Hierarchical systems. *Computer Journal*, 9(4):373–380, February 1967. ISSN 0010-4620.

Werner Stuetzle. Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample. *J. Classification*, 20(1):25–47, 2003. ISSN 0176-4268.

`http://www.mathworks.com/products/statistics/`. Matlab statistics toolbox, 2009.

`http://www.r-project.org/`. R statistical software.

U. von Luxburg and S. Ben-David. Towards a statistical theory of clustering. Technical report, PASCAL workshop on clustering, London, 2005.

D. Wishart. Mode analysis: a generalization of nearest neighbor which reduces chaining effects. In *Numerical Taxonomy*, pages 282–311. Academic Press, 1969.

Reza Zadeh and Shai Ben-David. A uniqueness theorem for clustering. In *Proceedings of the Proceedings of the Twenty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-09)*, page 8, Corvallis, Oregon, 2009. AUAI Press.