SIAM J. APPL. ALGEBRA GEOMETRY
Vol. 5, No. 4, pp. 675–700

Robust Hierarchical Clustering for Directed Networks: An Axiomatic Approach[*]

Gunnar Carlsson[†], Facundo Mémoli[‡], and Santiago Segarra[§]

Abstract. We provide a complete taxonomic characterization of robust hierarchical clustering methods for directed networks following an axiomatic approach. We begin by introducing three practical properties associated with the notion of robustness in hierarchical clustering: linear scale preservation, stability, and excisiveness. Linear scale preservation enforces imperviousness to change in units of measure, whereas stability ensures that a bounded perturbation in the input network entails a bounded perturbation in the clustering output. Excisiveness refers to the local consistency of the clustering outcome. Algorithmically, excisiveness implies that we can reduce computational complexity by only clustering a subset of our data while theoretically guaranteeing that the same hierarchical outcome would be observed when clustering the whole dataset. In parallel to these three properties, we introduce the concept of representability, a generative model for describing clustering methods through the specification of their action on a collection of networks. Our main result is to leverage this generative model to give a precise characterization of all robust—i.e., excisive, linear scale preserving, and stable—hierarchical clustering methods for directed networks. We also address the implementation of our methods and describe an application to real data.

Key words. hierarchical clustering, networks, robustness, consistency, stability, excisiveness

AMS subject classifications. 68R10, 68R12, 62H30, 91C20

DOI. 10.1137/20M1359201

1. Introduction.

The concept of clustering, i.e., partitioning a dataset into groups such that objects in one group are more similar to each other than they are to objects outside the group, is a fundamental tool for the advancement of knowledge in a wide range of disciplines from, e.g., medicine [54] to marketing [45]. Motivated by its relevance, literally hundreds of clustering algorithms have been developed in the past decades [29, 33, 41, 42, 43, 48, 52], mainly for the application to finite metric spaces but also for the increasingly relevant case of directed networks [46], in which the dissimilarity from node x to node x' may differ from the one from x' to x [5, 28, 38, 40, 44, 49, 50, 51, 58]. Directionality naturally arises in multiple contexts [35, 36]. Apart from the canonical example of a food web, information networks such as scientific citations or the World Wide Web are typically directed. Gene-regulatory networks are highly nonreciprocal and this lack of reciprocity needs to be accounted for when, e.g., grouping (clustering) genes that might have similar functional properties [56]. Moreover, in social networks, pairwise relations are rarely purely symmetric and this asymmetry is key

[*] Received by the editors August 11, 2020; accepted for publication (in revised form) August 4, 2021; published electronically October 19, 2021. Authors are listed in alphabetical order.
https://doi.org/10.1137/20M1359201
Funding: The second author was funded by NSF RI 1422400, NSF DMS 1723003, and NSF CCF 1740761.

[†] Department of Mathematics, Stanford University, Stanford, CA 94305 USA (gunnar@math.stanford.edu).

[‡] Departments of Mathematics and of Computer Science and Engineering, Ohio State University, Columbus, OH 43210 USA (memoli@math.osu.edu).

[§] Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA (segarra@rice.edu).

to accurately separate leaders from followers [47]. In these settings, effectively extracting knowledge from real and noisy data requires a theoretical understanding of robust techniques to analyze directed networks.

Although the theoretical underpinnings of clustering are not in general as well developed as its practical applications [4, 26, 53], the foundations of clustering in metric spaces have been developed over the past two decades [1, 7, 8, 31, 32, 37, 57]. For the specific case of hierarchical clustering [29, 33, 58], where instead of a single partition we look for a family of partitions indexed by a resolution parameter, some theoretical understanding has been achieved for the case of finite metric spaces [7, 8] and for the more general case of directed networks [9, 10, 12, 13, 39]. Of special interest to our work is [13], where two axioms encoding desirable features of hierarchical clustering methods were proposed, and an infinite but bounded family of methods satisfying these axioms (denominated the family of admissible methods) was identified. However, the disadvantage of this approach is that these two axioms are not sufficient to ensure robustness of the clustering methods abiding by them.

In the current paper, we build upon [8], [10], and [13] and deepen the characterization of hierarchical clustering methods on directed networks to identify those with robustness properties, which we view as encoding practical relevance. We define and analyze three properties of practical relevance, namely *excisiveness, linear scale invariance, and stability*, and say that a *hierarchical clustering method is robust if it possesses these three properties.*

Contributions. The contributions of this paper are threefold:

(i) We introduce a formal definition of robustness for hierarchical clustering methods based on the properties of excisiveness, linear scale preservation, and stability. Furthermore, we determine a subset of robust methods among a set of established admissible methods.

(ii) We introduce the notion of representability as a generative model for hierarchical clustering methods, where a method is defined through the specification of its local behavior. We also show that every representable method can be factorized into two well-defined operations: a symmetrizing operation followed by a well-established hierarchical clustering method.

(iii) We relate the two aforementioned notions by showing that, within the so-called admissible methods, representability is equivalent to robustness. This novel characterization result implies that any admissible and robust clustering method can be generated using a collection of representers.

Related work. The study of robust clustering methods has been an active area of research for a few decades now [2, 20, 23]. If we focus on the specific case of hierarchical clustering, it has long been known that certain linkage functions—such as Ward's linkage [30]—are more tolerant to noise in the input data than others. Moreover, relatively ad hoc methods—like Wishart's method [55]—have been developed to avoid shortcomings of other linkage functions by discarding low-density regions of the data. More recently, improvements over these classical methods [14], novel schemes that specifically focus on the imperviousness to outliers [24, 34], and robust algorithms that can accommodate categorical attributes [25] have been developed. Active methods have also been proposed where the similarities are selectively sampled before linkage, thus gaining robustness to a limited fraction of anomalous similarities [21]. From a more theoretical viewpoint, [3] proposes a robust linkage function to provably cluster data that satisfies a "good neighborhood" property, which is a relaxation of the strict separation property (where all points are more similar to points in their own cluster than to points in any other cluster). The current paper has two main differences with the aforementioned body of work:

- The definition of robustness here considered is novel and precisely defined based on the three properties of excisiveness, linear scale preservation, and stability.
- The result of our theoretical study is not just a single hierarchical clustering method that can be shown to be robust but rather *a generative model to construct all possible robust hierarchical clustering methods under the considered framework.*

Paper outline. After introducing basic concepts about clustering and networks (section 2), in section 3 we present a formal definition of robustness by introducing the properties of excisiveness (section 3.1), linear scale preservation (section 3.2), and stability (section 3.3). Representability, a notion introduced in section 4, provides a generative model for clustering where a method is defined through the specification of its behavior in a collection of special networks. In section 4.1, we show that every representable clustering method can be decomposed into a symmetrizing operation followed by the application of single linkage clustering. Although seemingly unrelated with the robustness properties previously mentioned, representability is a key notion to characterize clustering methods. Indeed, in section 5 we present our main result stating that an admissible clustering method is representable if and only if it is robust. Finally, in section 6, we illustrate the main result by implementing a representable clustering method, testing it on a real-world economic network, and confirming its robustness.

2. Preliminaries.

A network N is defined as a pair (X, A_X) where X is a finite set of n points or nodes and $A_X : X \times X \to \mathbb{R}_+$ is a dissimilarity function. Dissimilarities $A_X(x, x')$ from x to x' are nonnegative, and 0 if and only if $x = x'$, but may not satisfy the triangle inequality and may be directed or asymmetric, i.e., $A_X(x, x') \neq A_X(x', x)$ for some $x, x' \in X$. Given a positive real α, define the multiple of a network $\alpha * N := (X, \alpha\, A_X)$. Let $\mathcal{N}$ denote the collection of all networks. Networks $N \in \mathcal{N}$ can have different node sets X and different dissimilarities A_X. We focus our study on directed networks since these general structures include, as particular cases, undirected networks and finite metric spaces.

The output of hierarchically clustering the network $N = (X, A_X)$ is a dendrogram D_X, which is a collection of partitions $D_X(\delta)$ indexed by the resolution parameter $\delta \geq 0$ satisfying the following conditions:

- The partitions in $D_X(\delta)$ are nested, i.e., if x and x' are in the same partition at resolution δ, then they stay co-clustered for all larger resolutions $\delta' > \delta$.
- $D_X(0) = \big\{\{x\}, x \in X\big\}$, i.e., for the resolution parameter $\delta = 0$ each point $x \in X$ must form its own cluster.
- $D_X(\delta_0) = \big\{X\big\}$, i.e., for some sufficiently large resolution δ_0 all nodes must belong to the same cluster.

From these fundamental requirements and a technical condition of continuity (for all δ there exists $\epsilon > 0$ such that $D_X(\delta) = D_X(\gamma)$ for $\gamma \in [\delta, \delta + \epsilon]$) it follows that dendrograms can be represented as trees [7, sect. 3.1]. The interpretation of a dendrogram is that of a structure which yields different clusterings at different resolutions. When x and x' are co-clustered at resolution δ in D_X we say that they are equivalent at that resolution and write $x \sim_{D_X(\delta)} x'$.

Given a network (X, A_X) and $x, x' \in X$, a chain $C(x, x')$ is an ordered sequence of nodes in X, $C(x, x') = [x = x_0, x_1, \ldots, x_{l-1}, x_l = x']$, which starts at x and finishes at x'. The *links* of a chain are the edges connecting consecutive nodes of the chain in the direction given by it. We define the *cost* of chain $C(x, x')$ as the maximum dissimilarity $\max_{i | x_i \in C(x,x')} A_X(x_i, x_{i+1})$

encountered when traversing its links in order. The directed minimum chain cost $\tilde{u}^*_X(x,x')$ between x and x' is defined as the minimum cost among all the chains connecting x to x',

$$\tilde{u}^*_X(x,x') := \min_{C(x,x')} \max_{i | x_i \in C(x,x')} A_X(x_i, x_{i+1}). \tag{2.1}$$

An ultrametric u_X on the set X is a function $u_X : X \times X \to \mathbb{R}_+$ that satisfies symmetry $u_X(x,x') = u_X(x',x)$, identity $u_X(x,x') = 0 \iff x = x'$, and the strong triangle inequality

$$u_X(x,x') \leq \max\big(u_X(x,x''), u_X(x'',x')\big) \tag{2.2}$$

for all $x, x', x'' \in X$. For a given dendrogram D_X, consider the minimum resolution at which x and x' are co-clustered and define

$$u_X(x,x') := \min\big\{\delta \geq 0 \,|\, x \sim_{D_X(\delta)} x'\big\}. \tag{2.3}$$

It can be shown that the function u_X as defined in (2.3) is an ultrametric on the set X, from where it follows that dendrograms and finite ultrametric spaces are equivalent [7]. However, ultrametrics are more convenient than dendrograms for the results developed in this paper.

A hierarchical clustering method is defined as a map $\mathcal{H} : \mathcal{N} \to \mathcal{D}$ from the collection of networks $\mathcal{N}$ to the collection of all dendrograms $\mathcal{D}$ or, equivalently, as a map $\mathcal{H} : \mathcal{N} \to \mathcal{U}$ mapping every (possibly directed) network into the collection $\mathcal{U}$ of networks with ultrametrics as dissimilarity functions.

This loose definition of a hierarchical clustering method allows the existence of a wide variety of methods, most of them of little practical utility. Thus, in section 2.1 we recall an axiomatic construction built to select a subfamily of admissible clustering methods.

For future reference, we say that two methods $\mathcal{H}$ and $\mathcal{H}'$ are *equivalent*, denoted $\mathcal{H} \equiv \mathcal{H}'$, if $\mathcal{H}(N) = \mathcal{H}'(N)$ for all networks $N \in \mathcal{N}$. We also recall the definition of single linkage hierarchical clustering $\mathcal{H}^{\text{SL}}$ of *symmetric or undirected* networks with output ultrametrics $u^{\text{SL}}_X(x,x') := \min_{C(x,x')} \max_i A_X(x_i, x_{i+1})$.

2.1. Admissible hierarchical clustering methods.

In [10, 13], the authors impose the following two requirements on clustering methods:

(A1) *Axiom of value.* Given a two-node network $N = (\{p,q\}, A_{p,q})$ with $A_{p,q}(p,q) = \alpha$, and $A_{p,q}(q,p) = \beta$, the ultrametric $(X, u_{p,q}) = \mathcal{H}(N)$ output by $\mathcal{H}$ satisfies

$$u_{p,q}(p,q) = \max(\alpha, \beta). \tag{2.4}$$

(A2) *Axiom of transformation.* Given networks $N_X = (X, A_X)$ and $N_Y = (Y, A_Y)$ and a dissimilarity reducing map $\phi : X \to Y$, i.e., a map ϕ such that for all $x, x' \in X$ it holds that $A_X(x,x') \geq A_Y(\phi(x), \phi(x'))$, the outputs $(X, u_X) = \mathcal{H}(N_X)$ and $(Y, u_Y) = \mathcal{H}(N_Y)$ satisfy

$$u_X(x,x') \geq u_Y(\phi(x), \phi(x')). \tag{2.5}$$

We say that node x is able to influence node x' at resolution δ if the dissimilarity from x to x' is not greater than δ. In two-node networks, our intuition dictates that a cluster is formed if nodes p and q are able to influence each other. Thus, axiom (A1) states that in a network with two nodes, the dendrogram D_X has them merging at the maximum value of the two dissimilarities between them. Axiom (A2) captures the intuition that if a network is transformed such that some nodes become more similar but no pair of nodes increases its dissimilarity, then the transformed network should cluster at lower resolutions than the original one. Formally, (A2) states that a contraction of the dissimilarity function A_X entails a contraction of the associated ultrametric u_X.

A hierarchical clustering method $\mathcal{H}$ is *admissible* if it satisfies axioms (A1) and (A2). Two admissible methods of interest are reciprocal and nonreciprocal clustering, as defined next.

Reciprocal and nonreciprocal clustering. The *reciprocal* clustering method $\mathcal{H}^{\mathrm{R}}$ outputs the ultrametric $(X, u_X^{\mathrm{R}}) = \mathcal{H}^{\mathrm{R}}(X, A_X)$ defined as

$$u_X^{\mathrm{R}}(x,x') := \min_{C(x,x')} \max_{i \mid x_i \in C(x,x')} \bar{A}_X(x_i, x_{i+1}), \tag{2.6}$$

where $\bar{A}_X(x,x') := \max(A_X(x,x'), A_X(x',x))$ for all $x, x' \in X$. The *nonreciprocal* clustering method $\mathcal{H}^{\mathrm{NR}}$ outputs the ultrametric $(X, u_X^{\mathrm{NR}}) = \mathcal{H}^{\mathrm{NR}}(X, A_X)$ given by

$$u_X^{\mathrm{NR}}(x,x') := \max\Big(\tilde{u}_X^*(x,x'),\ \tilde{u}_X^*(x',x)\Big). \tag{2.7}$$

Intuitively, in (2.6) we search for chains $C(x,x')$ linking nodes x and x'. Then, for a given chain, we walk from x to x' and determine the maximum dissimilarity, in either the forward or the backward direction, across all links in the chain. The reciprocal ultrametric $u_X^{\mathrm{R}}(x,x')$ is the minimum of this value across all possible chains; see Figure 1. Putting it differently, reciprocal clustering joins x and x' at resolution δ if it is possible to go back and forth at maximum cost δ through the same chain. By contrast, nonreciprocal clustering permits different chains. In (2.7), we implicitly consider forward chains $C(x,x')$ going from x to x' and backward chains $C(x',x)$ from x' to x. We then determine the respective maximum dissimilarities and search independently for the forward and backward chains that minimize the respective maximum dissimilarities. The nonreciprocal ultrametric $u_X^{\mathrm{NR}}(x,x')$ is the maximum of these two minimum values; see Figure 2.

These two methods exemplify extremal behaviors. Indeed, reciprocal and nonreciprocal clustering bound the ultrametrics generated by all admissible methods, as stated next.

Theorem 2.1 ([13, Theorem 4]). *Consider an arbitrary network $N = (X, A_X)$ and let u_X^{R} and u_X^{NR} be the associated reciprocal and nonreciprocal ultrametrics as defined in (2.6) and (2.7). Then, for any admissible method $\mathcal{H}$ the output ultrametric $(X, u_X) = \mathcal{H}(X, A_X)$ is such that for all pairs x, x',*

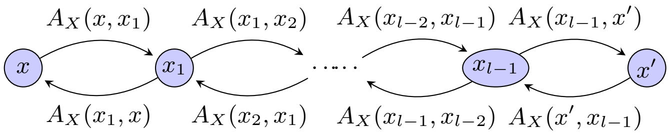


Figure 1. *Reciprocal clustering. Nodes x, x' cluster at resolution δ if they can be joined with a bidirectional chain of maximum dissimilarity δ (cf. (2.6)).*

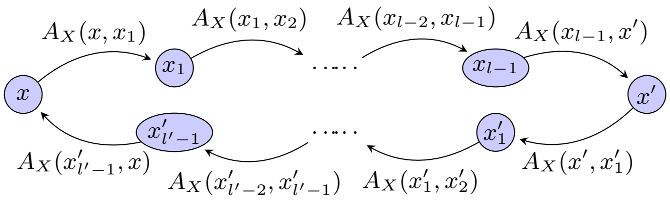


Figure 2. *Nonreciprocal clustering. Nodes x, x' cluster at resolution δ if they can be joined in both directions with possibly different chains of maximum dissimilarity δ (cf. (2.7)).*

$$u_X^{\mathrm{NR}}(x, x') \le u_X(x, x') \le u_X^{\mathrm{R}}(x, x'). \tag{2.8}$$

In particular, $u_X^{\mathrm{NR}} = u_X^{\mathrm{R}}$ whenever (X, A_X) is undirected.

According to Theorem 2.1, nonreciprocal clustering yields uniformly minimal ultrametrics while reciprocal clustering yields uniformly maximal ultrametrics among all methods satisfying (A1)–(A2). Moreover, the existence of admissible methods strictly different from $\mathcal{H}^{\mathrm{NR}}$ and $\mathcal{H}^{\mathrm{R}}$ has been shown [11, Prop. 1]. For symmetric networks, reciprocal and nonreciprocal clustering coincide, implying that there is a unique admissible method, which is equivalent to the well-known single linkage hierarchical clustering method [29, Chap. 4]. In section 3, we present practical properties—excisiveness, linear scale preservation, and stability—which are not shared by every admissible method, and we use them to further winnow the set of clustering methods.

3. Robust hierarchical clustering methods.

We formalize the notion of a robust hierarchical clustering method as one that satisfies three properties: excisiveness (section 3.1), linear scale preservation (section 3.2), and stability (section 3.3). In this section, we define, analyze, and give examples of methods satisfying these three properties.

3.1. Excisiveness.

Consider a clustering method $\mathcal{H}$ and a given network $N = (X, A_X)$. Denote by $(X, u_X) = \mathcal{H}(N)$ the ultrametric output, as D_X the output dendrogram, and, for a given resolution δ, denote the dendrogram's partition by $D_X(\delta) = \{B_1(\delta), \dots, B_{J(\delta)}(\delta)\}$, where each block $B_i(\delta)$ represents a cluster at resolution δ. Consider then the induced subnetworks N_i^δ associated with each block $B_i(\delta)$ of $D_X(\delta)$ defined as

$$N_i^\delta := \Big(B_i(\delta),\ A_X\big|_{B_i(\delta)\times B_i(\delta)}\Big), \tag{3.1}$$

where $A_X|_{B_i(\delta)\times B_i(\delta)}$ denotes the restriction of A_X to the nodes in $B_i(\delta)$. In terms of ultrametrics, networks N_i^δ are such that their node set $B_i(\delta)$ satisfies

$$\begin{aligned} u_X(x, x') &\le \delta \quad \text{for all } x, x' \in B_i(\delta), \\ u_X(x, x'') &> \delta \quad \text{for all } x \in B_i(\delta),\ x'' \notin B_i(\delta). \end{aligned} \tag{3.2}$$

Two related ultrametrics can be defined on the node set represented by any block $B_i(\delta)$: first, the result of restricting the output clustering ultrametric u_X to $B_i(\delta)$, and second, the ultrametric obtained when applying the clustering method $\mathcal{H}$ to the subnetwork N_i^δ. If the two intervening ultrametrics are the same for every network N, all i, and all $\delta > 0$, then we say that the method $\mathcal{H}$ is excisive, as we formally define next.

(P1) *Excisiveness.* We say that $\mathcal{H}$ is *excisive* if, for any arbitrary network N, for all subnetworks N_i^δ (cf. (3.1)) at all resolutions $\delta > 0$ it holds that

$$\mathcal{H}\left(N_i^\delta\right) = \left(B_i(\delta),\ u_X\big|_{B_i(\delta)\times B_i(\delta)}\right). \tag{3.3}$$

The appeal of excisive methods is that they exhibit local consistency in the following sense. For a given resolution δ, when we cluster the subnetworks as defined in (3.1), we obtain a dendrogram on the node set $B_i(\delta)$ for every i. Excisiveness ensures that when clustering the whole network and cutting the output dendrogram at resolution δ, the branches obtained coincide with the previously computed dendrograms for every subnetwork; see Figure 3. Our notion of excisiveness is inspired by [8], where a related concept was analyzed for nonhierarchical clustering of finite metric spaces.

Excisiveness entails a tangible practical advantage when hierarchically clustering big data. In applications, one often begins by performing a coarse clustering at an exploratory phase. Notice that the computational cost of obtaining this coarse partition, which corresponds to *one* particular resolution, is smaller than that of computing the whole dendrogram. After having done this, one focuses on relevant clusters—via the subsequent application of the clustering method—in order to reveal the whole hierarchical structure of this subset of the data. An excisive method guarantees that the result obtained from this two-step procedure coincides with the more computationally intensive clustering of the whole dataset. A specific example of this computational gain is presented next.

Example 3.1 (single linkage computation). Focus on the application of single linkage hierarchical clustering to a finite metric space of n points. Single linkage is an excisive clustering method, as can be concluded by combining Proposition 3.2 below with the fact that, for finite metric spaces, reciprocal and nonreciprocal clustering coincide with single linkage (cf. Theorem 2.1). Consider two different ways of computing the output dendrogram for a

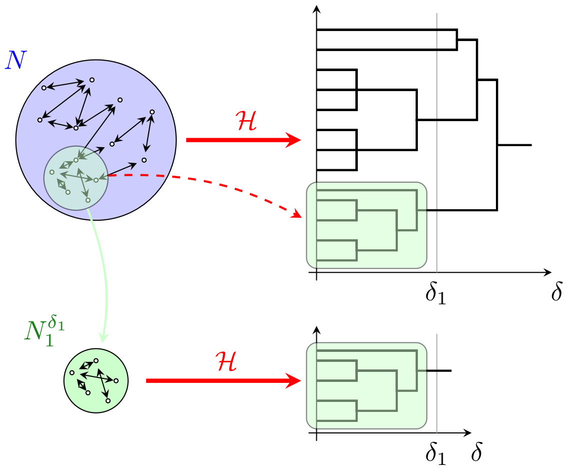


Figure 3. *The clustering method $\mathcal{H}$ is excisive. Given an arbitrary network N (blue) the method $\mathcal{H}$ outputs the dendrogram on the top right, where the green branch corresponds to the subnetwork $N_1^{\delta_1}$. If we consider the isolated subnetwork $N_1^{\delta_1}$ and apply $\mathcal{H}$, excisiveness guarantees that the obtained dendrogram is equivalent to the green branch in the original one.*

subspace of the aforementioned finite metric space. The first approach is to hierarchically cluster the whole finite metric space and then extract the relevant branch. The computational cost of single linkage is equivalent to that of finding a minimum spanning tree in an undirected graph which, for a complete graph, is of cost $\mathcal{O}(n^2)$ [15, Theorem 1.1].[1] The second approach consists of first obtaining the partition given by a single linkage corresponding to *one* coarse resolution. This is equivalent to finding the connected components in a graph where only the edges of weight smaller than the resolution are present. Assuming that the average degree of each node in this graph is α, the computational cost of finding the connected components is $\mathcal{O}(\max(n, n\,\alpha/2)) = \mathcal{O}(n\,\alpha/2)$ as long as $\alpha \geq 2$ [27]. After this, we pick the subspace of interest and find its minimum spanning tree. Assuming that the subspace contains $\beta\, n$ nodes, the cost of finding the minimum spanning tree is $\mathcal{O}(\beta^2 n^2)$. Consequently, the cost of the first approach is $\mathcal{O}(n^2)$, whereas the cost of the second one is $\mathcal{O}(n\,\alpha/2) + \mathcal{O}(\beta^2 n^2)$. This entails an asymptotic reduction of order β^{-2}. In the extreme case where $\beta = \beta_0/n$ so that the subspace of interest is independent of the size of the whole network, there is a reduction in computational complexity from quadratic to linear in n. Still, excisiveness ensures that the outputs of both approaches coincide, allowing us to follow the second—more efficient—approach.

Having established that excisiveness is a property of practical relevance, we seek to study its relation with the axiomatic approach reviewed in section 2.1. One can show that there exist clustering methods that, while satisfying axioms (A1)–(A2), are also excisive. Indeed, the reciprocal and nonreciprocal clustering methods introduced in section 2.1 are excisive, as we state next.

Proposition 3.2. *The reciprocal $\mathcal{H}^{\mathrm{R}}$ and nonreciprocal $\mathcal{H}^{\mathrm{NR}}$ methods with output ultrametrics defined in* (2.6) *and* (2.7)*, respectively, are excisive as defined in* (P1).

Proof. Given an arbitrary network $N = (X, A_X)$, denote by $(X, u^{\mathrm{R}}_X) = \mathcal{H}^{\mathrm{R}}(N)$ the output ultrametric when applying $\mathcal{H}^{\mathrm{R}}$ to N. Pick an arbitrary resolution δ and focus on a subnetwork $N_i^\delta = (X_i^\delta, A_{X_i^\delta})$ as defined in (3.1). Denote by $(X_i^\delta, u^{\mathrm{R}}_{X_i^\delta}) = \mathcal{H}^{\mathrm{R}}(N_i^\delta)$ the clustering output when applying $\mathcal{H}^{\mathrm{R}}$ to N_i^δ. We want to show that

$$u^{\mathrm{R}}_{X_i^\delta} \equiv u^{\mathrm{R}}_X\big|_{X_i^\delta \times X_i^\delta}. \tag{3.4}$$

Since the network N, the resolution δ, and the subnetwork index i were chosen arbitrarily, (3.4) would imply that the reciprocal clustering method $\mathcal{H}^{\mathrm{R}}$ is excisive (cf. (P1)), as wanted. We first show that

$$u^{\mathrm{R}}_{X_i^\delta}(x, x') \geq u^{\mathrm{R}}_X(x, x') \tag{3.5}$$

for all nodes $x, x' \in X_i^\delta$. Notice that the inclusion map $\phi : X_i^\delta \to X$ from network N_i^δ to N such that $\phi(x) = x$ is a dissimilarity reducing map as defined in (A2). Hence, since $\mathcal{H}^{\mathrm{R}}$ satisfies the axiom of transformation (A2), inequality (3.5) must hold. In order to show the opposite inequality, pick arbitrary nodes $x, x' \in X_i^\delta$ and assume that

$$u^{\mathrm{R}}_X(x, x') = \alpha. \tag{3.6}$$

[1] Notice that the mentioned complexity omits an inverse Ackermann term which is small in practice. Also, randomized algorithms with expected complexity $\mathcal{O}(n^2)$ have been derived; see the discussion in [15].

From (3.2), we know that $\alpha \leq \delta$. From the definition of $\mathcal{H}^{\mathrm{R}}$ in (2.6), equality (3.6) implies that there exists a chain $C(x,x') = [x = x_0, x_1, \ldots, x_l = x']$ where the maximum dissimilarity in both directions between consecutive nodes is α. However, notice that part of this chain can be used to join any two nodes x_j and x_k where $j,k \in \{0,1,\ldots,l\}$ with dissimilarities not larger than α. This implies that $u_X^{\mathrm{R}}(x_j,x_k) = \alpha$ for $j,k \in \{0,1,\ldots,l\}$ and from the definition of subnetwork (cf. (3.2)) we must have that $x_j \in X_i^\delta$ for all $j \in \{0,1,\ldots,l\}$. Consequently, when applying the reciprocal clustering method $\mathcal{H}^{\mathrm{R}}$ to N_i^δ, the nodes in the chain $C(x,x')$ are contained in its node set X_i^δ, allowing us to write (cf. (2.6))

$$u^{\mathrm{R}}_{X_i^\delta}(x,x') \leq \max_{j | x_j \in C(x,x')} \bar{A}_{X_i^\delta}(x_j,x_{j+1}) = \alpha = u_X^{\mathrm{R}}(x,x'), \tag{3.7}$$

where the inequality comes from the fact that we picked one particular chain $C(x,x')$ instead of minimizing across all possible chains. Since $x,x' \in X_i^\delta$ were picked arbitrarily, (3.7) implies that $u^{\mathrm{R}}_{X_i^{\mathrm{R}}}(x,x') \leq u_X^{\mathrm{R}}(x,x')$ for all $x,x' \in X_i^\delta$. Combining this inequality with (3.5), equivalence (3.4) follows and we show excisiveness of $\mathcal{H}^{\mathrm{R}}$. A similar argument can be used to show excisiveness of nonreciprocal clustering $\mathcal{H}^{\mathrm{NR}}$. ■

Despite Proposition 3.2, *excisiveness is not implied by admissibility* with respect to (A1) and (A2). To see this, consider the admissible semireciprocal clustering method $\mathcal{H}^{\mathrm{SR}(t)}$ introduced in [11] and briefly explained next.

Semireciprocal clustering presents an intermediate behavior between reciprocal and nonreciprocal clustering. In reciprocal clustering, we minimize the cost of a chain in both directions simultaneously, whereas in nonreciprocal clustering we minimize the cost in both directions separately. However, semireciprocal clustering adopts an intermediate position. In order to formalize this, we denote by $C_t(x,x')$ a chain starting at x and finishing at x' with at most t nodes while we reserve the notation $C(x,x')$ to denote a chain linking x with x' with no maximum imposed on the number of nodes in the chain. Given a network $N = (X, A_X)$, define as $A_X^{\mathrm{SR}(t)}(x,x')$ the minimum cost incurred when traveling from node x to node x' using a chain of at most t nodes. That is,

$$A_X^{\mathrm{SR}(t)}(x,x') := \min_{C_t(x,x')} \max_{i | x_i \in C_t(x,x')} A_X(x_i,x_{i+1}). \tag{3.8}$$

The family of semireciprocal clustering methods $\mathcal{H}^{\mathrm{SR}(t)}$ with output $(X, u_X^{\mathrm{SR}(t)}) = \mathcal{H}^{\mathrm{SR}(t)}(N)$ is defined as

$$u_X^{\mathrm{SR}(t)}(x,x') := \min_{C(x,x')} \max_{i | x_i \in C(x,x')} \bar{A}_X^{\mathrm{SR}(t)}(x_i,x_{i+1}), \tag{3.9}$$

where the function $\bar{A}_X^{\mathrm{SR}(t)}(x_i,x_{i+1})$ is computed as follows:

$$\bar{A}_X^{\mathrm{SR}(t)}(x_i,x_{i+1}) := \max\left(A_X^{\mathrm{SR}(t)}(x_i,x_{i+1}), A_X^{\mathrm{SR}(t)}(x_{i+1},x_i)\right). \tag{3.10}$$

We can interpret (3.9) as the application of reciprocal clustering (cf. (2.6)) to a network with dissimilarities given by $A_X^{\mathrm{SR}(t)}$ in (3.8), i.e., a network with dissimilarities given by the

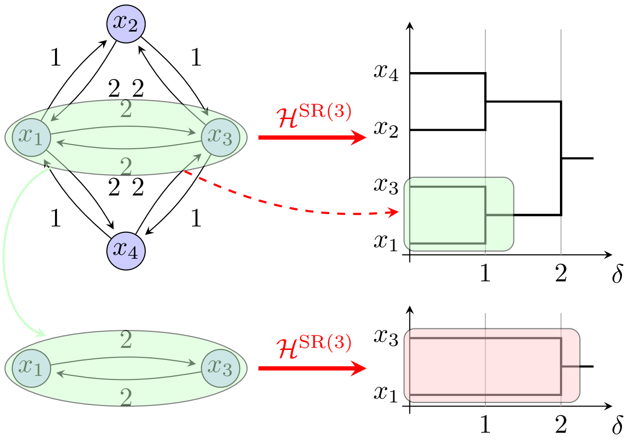


Figure 4. *Admissibility does not imply excisiveness. The admissible method $\mathcal{H}^{\mathrm{SR}(3)}$ does not satisfy the excisiveness condition since the green branch in the top dendrogram differs from the red branch in the lower one (cf. Figure 3).*

optimal choice of chains of constrained length t. Semireciprocal clustering methods satisfy axioms (A1)–(A2); see [11, Prop. 4].

To see that admissibility does not imply excisiveness, consider the network in Figure 4 and its dendrogram corresponding to the semireciprocal clustering method $\mathcal{H}^{\mathrm{SR}(3)}$. For a resolution $\delta = 1.5$, focus on the subnetwork $N_1^{1.5} = (\{x_1, x_3\}, A_{\{1,3\}})$ with $A_{\{1,3\}}(x_1, x_3) = A_{\{1,3\}}(x_3, x_1) = 2$. When the clustering method $\mathcal{H}^{\mathrm{SR}(3)}$ is applied to this subnetwork, the output dendrogram (red) differs from the corresponding branch in the original dendrogram (green). This counterexample shows that excisiveness cannot be derived from axioms (A1) and (A2).

3.2. Linear scale preservation. Consider a network $N_X = (X, A_X)$ and a linear function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ where $\psi(z) = \alpha\, z$ for some $\alpha > 0$. Define the network $N_X^\psi := (X, \psi \circ A_X)$ with the same set of nodes and linearly scaled dissimilarities. With this notation in place, we formally define our second robustness property.

(P2) *Linear scale preservation.* We say that $\mathcal{H}$ is *linear scale preserving* if for any arbitrary network N_X and function ψ satisfying the above requirements, the outputs $(X, u_X) := \mathcal{H}(N_X)$ and $(X, u_X^\psi) := \mathcal{H}(N_X^\psi)$ satisfy

$$u_X^\psi = \psi \circ u_X. \tag{3.11}$$

For linear scale preserving methods, the ultrametric outcomes vary according to the same linear function that transforms the dissimilarity function. Hence, the hierarchical structure output by these methods is invariant with respect to units. In terms of dendrograms, linear scale preservation entails that a transformation of dissimilarities with an appropriate linear function ψ results in a dendrogram where the order in which nodes are clustered together is the same while the resolution at which mergings occur changes linearly according to ψ.

In practice, linear scale preservation is a desirable property. For example, if we want to hierarchically cluster finite metric spaces—which are particular cases of asymmetric networks where every dissimilarity is symmetric and the triangle inequality is satisfied—the hierarchy of the output should not depend on the unit used to measure distances. Equivalently, the choice of units does not alter the nature of a given metric space; thus, if we measure distances in, e.g., meters or centimeters, we should obtain the same structure when clustering both. Linear scale preserving methods guarantee this behavior for arbitrary asymmetric networks.

The reciprocal and nonreciprocal clustering methods introduced in section 2.1 are linear scale preserving.

Proposition 3.3. *The reciprocal $\mathcal{H}^{\mathrm{R}}$ and nonreciprocal $\mathcal{H}^{\mathrm{NR}}$ methods with output ultrametrics defined in* (2.6) *and* (2.7)*, respectively, are linear scale preserving as defined in* (P2)*.*

Proof. This proposition follows as a particular case of our main result (Theorem 5.1, to be introduced in section 5) after demonstrating that $\mathcal{H}^{\mathrm{R}}$ and $\mathcal{H}^{\mathrm{NR}}$ are representable methods; cf. Figure 8 and the associated discussion. Notice that Proposition 3.2 can also be shown as a particular case of the more general Theorem 5.1, but we decided to include its proof to demonstrate the technique to show one of these simpler results independently. Nonetheless, the current proof is omitted to avoid redundancy. ■

Proposition 3.3 notwithstanding, linear scale preservation is a condition independent of axioms (A1) and (A2). This can be seen by analyzing the behavior of the admissible method $\mathcal{H}^{\mathrm{R/NR}}(\beta)$ introduced in [11] and briefly explained next.

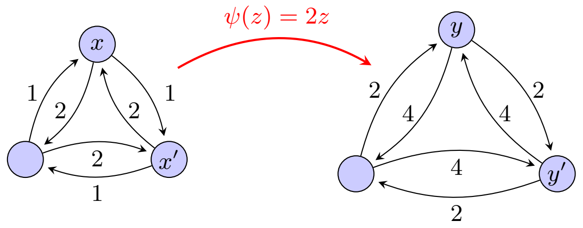
The grafting clustering method $\mathcal{H}^{\mathrm{R/NR}}(\beta)$ is constructed by pasting branches of the nonreciprocal dendrogram into corresponding branches of the reciprocal dendrogram. To define this precisely, one computes the reciprocal and nonreciprocal dendrograms and cuts all branches of the reciprocal dendrogram at resolution $\beta > 0$. Then, replace the cut branches of the reciprocal tree by the corresponding branches—i.e., those with the same leaves—of the nonreciprocal tree. This hybrid dendrogram is the output of method $\mathcal{H}^{\mathrm{R/NR}}(\beta)$. In terms of ultrametrics, we can define this pasting formally as follows:

$$u_X^{\mathrm{R/NR}}(x,x';\beta) := \begin{cases} u_X^{\mathrm{NR}}(x,x') & \text{if } u_X^{\mathrm{R}}(x,x') \leq \beta, \\ u_X^{\mathrm{R}}(x,x') & \text{if } u_X^{\mathrm{R}}(x,x') > \beta. \end{cases} \tag{3.12}$$

The ultrametric in (3.12) is valid and $\mathcal{H}^{\mathrm{R/NR}}(\beta)$ satisfies axioms (A1) and (A2); see [11, Prop. 1]. However, the method $\mathcal{H}^{\mathrm{R/NR}}(\beta)$ is not linear scale preserving, as can be seen from a simple counterexample. Consider the three-node network in Figure 5 as well as its transformation after applying the linear function $\psi(z) = 2\,z$. The figure illustrates the fact that the reciprocal and nonreciprocal ultrametrics are transformed by ψ, as it should be given Proposition 3.3. However, we see that the ultrametric output by $\mathcal{H}^{\mathrm{R/NR}}(\beta)$ (for $\beta = 3$) is multiplied by 4 instead of by 2, thus violating (P2).

3.3. Stability. As a third important robustness property, we introduce the notion of stability. This concept requires the definition of a metric $d_{\mathcal{N}}$ between networks. This metric is a generalization of the Gromov–Hausdorff distance [6, Chapter 7.3], originally conceived as a metric between compact metric spaces, to the more general collection of networks $\mathcal{N}$.

Whenever two networks N_X and N_Y are related by a simple redefinition of the node labels, we say that they are isomorphic and we write $N_X \cong N_Y$. The collection of networks where



$$u_X^{\mathrm{NR}}(x,x') = 1 \xrightarrow{= 2 \times u_X^{\mathrm{NR}}(x,x')} u_Y^{\mathrm{NR}}(y,y') = 2$$

$$u_X^{\mathrm{R}}(x,x') = 2 \xrightarrow{= 2 \times u_X^{\mathrm{R}}(x,x')} u_Y^{\mathrm{R}}(y,y') = 4$$

$$u_X^{\mathrm{R/NR}}(x,x';\beta) = 1 \xrightarrow[\beta = 3]{\neq 2 \times u_X^{\mathrm{R/NR}}(x,x';\beta)} u_Y^{\mathrm{R/NR}}(y,y';\beta) = 4$$

Figure 5. *Admissibility does not imply linear scale preservation. Reciprocal and nonreciprocal clustering are linear scale preserving while* $\mathcal{H}^{\mathrm{R/NR}}(\beta)$ *is not.*

all isomorphic networks are represented by a single point is called the collection of networks modulo isomorphism and denoted as $\mathcal{N} \mod \cong$. For node sets X and Y consider subsets $R \subseteq X \times Y$ of the Cartesian product set $X \times Y$ with elements $(x, y) \in R$. The set R is a *correspondence* between X and Y if for all $x_0 \in X$ we have at least one element $(x_0, y) \in R$ and for all $y_0 \in Y$ we have at least one element $(x, y_0) \in R$. The metric $d_{\mathcal{N}}$ between networks N_X and N_Y takes the value

$$d_{\mathcal{N}}(N_X, N_Y) := \frac{1}{2} \min_{R} \max_{(x,y),(x',y') \in R} \big| A_X(x, x') - A_Y(y, y') \big|. \tag{3.13}$$

Definition (3.13) is a verbatim generalization of the Gromov–Hausdorff distance in [6, Theorem 7.3.25] except that the dissimilarity functions A_X and A_Y are not restricted to be metrics. For this more general case, $d_{\mathcal{N}}$ is still a legitimate metric in the space $\mathcal{N} \mod \cong$; see [12, sect. A.4] for a proof of this fact. See [19] for applications to the stability of persistent homology over networks. The case of possibly *infinite* networks was studied in [17, 18, 16].[2] With this definition in place, we formally introduce the property of stability.

(P3) *Stability.* We say that $\mathcal{H}$ is *stable* if there exists a *finite* constant $L = L(\mathcal{H}) \geq 0$ such that, for any two networks N_X and N_Y, it holds that

$$d_{\mathcal{N}}\big(\mathcal{H}(N_X), \mathcal{H}(N_Y)\big) \leq L \cdot d_{\mathcal{N}}(N_X, N_Y). \tag{3.14}$$

Stability ensures that small perturbations on a network result in small perturbations in the associated ultrametric. More precisely, perturbations of size at most ε on a given network—as measured by $d_{\mathcal{N}}$—result in perturbations in the clustering results which are bounded by $L\,\varepsilon$. In other words, every stable hierarchical clustering method $\mathcal{H}$ is *Lipschitz* as a map from $(\mathcal{N}, d_{\mathcal{N}})$ into itself, making them suitable for practical applications. See [7, Remark 17] for

[2] These papers consider a notion of network (X, A_X) a bit more general than the one we considered here: the authors do not require A_X to satisfy any conditions except being real valued.

a discussion of the fact that average and complete linkage hierarchical methods are unstable when applied to finite metric spaces.

Mimicking the developments in previous subsections, one can show that the reciprocal and nonreciprocal clustering methods are stable.

Proposition 3.4. *The reciprocal $\mathcal{H}^{\mathrm{R}}$ and nonreciprocal $\mathcal{H}^{\mathrm{NR}}$ methods with output ultrametrics defined in* (2.6) *and* (2.7)*, respectively, are stable as defined in* (P3).

Proof. This proposition follows as a particular case of our main result (Theorem 5.1), after demonstrating that $\mathcal{H}^{\mathrm{R}}$ and $\mathcal{H}^{\mathrm{NR}}$ are representable methods; cf. Figure 8 and the associated discussion. ■

We say that a clustering method is *robust if it satisfies the properties of excisiveness* (P1), *linear scale preservation* (P2), *and stability* (P3). Given that robustness is an important practical feature, we want to characterize the family of robust admissible methods. From Propositions 3.2, 3.3, and 3.4 we know that reciprocal and nonreciprocal clustering belong to this family. Our objective is to find if other methods are contained within this family and, more importantly, to provide a comprehensive description of these. To this end, we introduce the concept of representability next.

4. Representability.

We define a representable hierarchical clustering method as one where the clustering of arbitrary networks is specified through the clustering of particular examples that we call *representers*. Representers are possibly asymmetric networks $\omega = (X_\omega, A_\omega)$ with the distinction that the dissimilarity function A_ω need not be defined for all pairs of nodes, i.e., $\mathrm{dom}(A_\omega) \neq X_\omega \times X_\omega$. In this sense, representers are more general objects than networks as introduced in section 2.

Given an arbitrary network $N = (X, A_X)$, and a representer $\omega = (X_\omega, A_\omega)$, we define the *expansion constant* of a map $\phi : X_\omega \to X$ from ω to N as

$$L(\phi; \omega, N) := \max_{\substack{(z,z')\in\mathrm{dom}(A_\omega)\\ z\neq z'}} \frac{A_X(\phi(z), \phi(z'))}{A_\omega(z, z')}. \tag{4.1}$$

Notice that $L(\phi; \omega, N)$ is the minimum multiple of ω such that the map ϕ is dissimilarity reducing as defined in (A2) from $L(\phi; \omega, N) * \omega$ to N. Notice as well that the maximum in (4.1) is computed for pairs (z, z') in the domain of A_ω. Pairs not belonging to the domain can be mapped to any dissimilarity without modifying the value of the expansion constant. We define the optimal multiple $\lambda^\omega_X(x, x')$ between x and x' in X with respect to ω as

$$\lambda^\omega_X(x, x') := \min\left\{ L(\phi; \omega, N) \mid \phi : X_\omega \to X,\ x, x' \in \mathrm{Im}(\phi)\right\}. \tag{4.2}$$

Equivalently, $\lambda^\omega_X(x, x')$ is the minimum expansion constant among those maps that have x and x' in their image. That is, it is the minimum multiple needed for the existence of a dissimilarity reducing map from a multiple of ω to N that has x and x' in its image.

We can now define the representable method $\mathcal{H}^\omega$ associated with a given representer ω by defining the cost of a chain $C(x, x') = [x = x_0, \ldots, x_l = x']$ linking x to x' as the maximum optimal multiple $\lambda_X^\omega(x_i, x_{i+1})$ between consecutive nodes in the chain. The ultrametric u_X^ω associated with output $(X, u_X^\omega) = \mathcal{H}^\omega(X, A_X)$ is given by the minimum chain cost

$$u_X^\omega(x,x') := \min_{C(x,x')} \max_{i | x_i \in C(x,x')} \lambda_X^\omega(x_i, x_{i+1}), \tag{4.3}$$

for all $x, x' \in X$. Representable methods are generalized to cases in which we are given a nonempty collection Ω of representers ω. In such case, we define the function λ_X^Ω as

$$\lambda_X^\Omega(x,x') \; := \; \inf_{\omega\in\Omega} \; \lambda_X^\omega(x,x') \tag{4.4}$$

for all $x, x' \in X$. The value $\lambda_X^\Omega(x,x')$ is the infimum across all optimal multiples given by the different representers $\omega \in \Omega$. For a given network $N = (X, A_X)$, the representable clustering method $\mathcal{H}^\Omega$ associated with the collection of representers Ω is the one with outputs $(X, u_X^\Omega) = \mathcal{H}^\Omega(X, A_X)$ such that the ultrametric u_X^Ω is given by

$$u_X^\Omega(x,x') := \min_{C(x,x')} \max_{i | x_i \in C(x,x')} \lambda_X^\Omega(x_i, x_{i+1}) \tag{4.5}$$

for all $x, x' \in X$. See Figure 6 for an illustrative example.

As we mentioned, not all dissimilarities are necessarily defined in representers. However, the issue of whether a representer is connected or not plays a prominent role in the validity and admissibility of representable methods. We say that a representer $\omega = (X_\omega, A_\omega)$ is *weakly connected* if for every pair of nodes $z, z' \in X_\omega$ we can find a chain $C(z,z') = [z = z_0, \ldots, z_l = z']$ such that either $(z_i, z_{i+1}) \in \mathrm{dom}(A_\omega)$ or $(z_{i+1}, z_i) \in \mathrm{dom}(A_\omega)$ or both for all $i = 0, \ldots, l-1$. Moreover, we say that Ω is *uniformly bounded* if and only if there exists a finite $M > 0$ such that

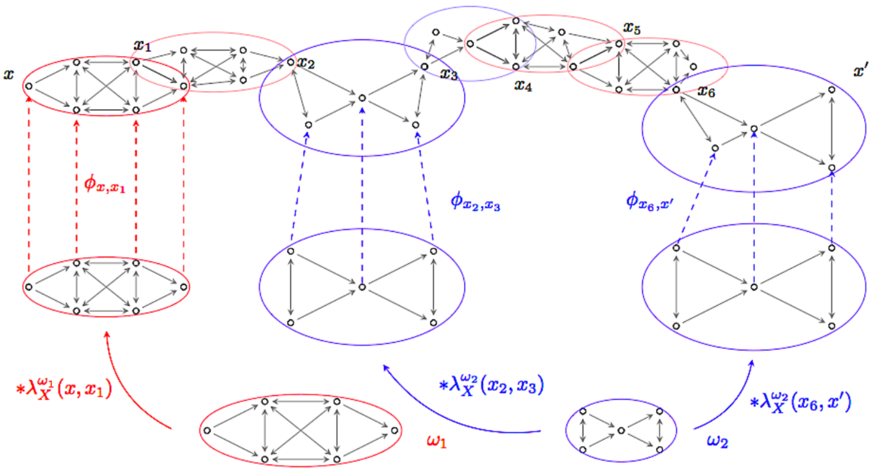


Figure 6. *Representable method $\mathcal{H}^\Omega$ with ultrametric output as in (4.5). The collection of representers $\Omega = \{\omega_1, \omega_2\}$ is shown at the bottom. In order to compute $u_X^\Omega(x,x')$ we link x and x' through a chain, e.g., $[x, x_1, \ldots, x_6, x']$ in the figure, and link pairs of consecutive nodes with multiples of the representers. The ultrametric value $u_X^\Omega(x,x')$ is given by minimizing over all paths joining x and x' the maximum multiple of a representer used to link consecutive nodes in the path (cf. (4.5)).*

$$\max_{(z,z')\in\mathrm{dom}(A_\omega)} A_\omega(z,z') \leq M \tag{4.6}$$

for all $\omega \in \Omega$. For any representer ω, let $\mathrm{sep}(\omega) := \min_{(z,z')\in\mathrm{dom}(A_\omega)} A_\omega(z,z')$, and, for a family Ω of representers, we define $\mathrm{sep}(\Omega) := \inf_{\omega\in\Omega} \mathrm{sep}(\omega)$. We can now formally define the notion of representability.

(P4) *Representability.* We say that a clustering method $\mathcal{H}$ is *representable* if there exists a uniformly bounded collection Ω of weakly connected representers each with a finite number of nodes and $\mathrm{sep}(\Omega) > 0$ such that $\mathcal{H} \equiv \mathcal{H}^\Omega$ where $\mathcal{H}^\Omega$ has output ultrametrics as in (4.5).

It can be shown that indeed under the conditions in (P4), (4.5) defines a valid ultrametric, as stated next.[3]

Proposition 4.1. *For every collection of representers Ω satisfying the conditions in* (P4), (4.5) *defines a valid ultrametric.*

Representability allows the definition of universal hierarchical clustering methods from given representative examples. Every representer $\omega \in \Omega$ can be understood as defining a specific structure that can be considered as a cluster unit. The scaling of this cluster unit (cf. (4.2)) and its replication throughout the network (cf. (4.3)) signal the resolution at which nodes become part of the same cluster. For nodes x and x' to cluster together at resolution δ, we need to construct a path from x to x' with overlapping versions of representers scaled by parameters not larger than δ. When we have multiple representers, we can use any of them to build these chains (cf. (4.4) and (4.5)).

Although seemingly unrelated, the property of representability (P4) is tightly related to the more practical requirements of excisiveness (P1), linear scale preservation (P2), and stability (P3), as will be formally shown in section 5.

4.1. Factorization of representable methods. The following factorization property for representable methods has practical value in itself and will be instrumental to showing our main result in Theorem 5.1. Every representable clustering method factors into the composition of two maps: a symmetrizing map that depends on Ω, followed by single linkage hierarchical clustering. This is formally stated next.

Proposition 4.2. *Every representable clustering method $\mathcal{H}^\Omega$ admits a decomposition of the form $\mathcal{H}^\Omega \equiv \mathcal{H}^{\mathrm{SL}} \circ \Lambda^\Omega$, where $\Lambda^\Omega : \mathcal{N} \to \mathcal{N}^{\mathrm{sym}}$ is a map from the collection of asymmetric networks $\mathcal{N}$ to that of symmetric networks $\mathcal{N}^{\mathrm{sym}}$ and $\mathcal{H}^{\mathrm{SL}} : \mathcal{N}^{\mathrm{sym}} \to \mathcal{U}$ is the single linkage clustering method for symmetric networks.*

Proof. The proof is just a matter of identifying elements in (4.5). Define the function Λ^Ω as the one that maps the network $N = (X, A_X)$ into $\Lambda^\Omega(X, A_X) = (X, \lambda_X^\Omega)$, where the dissimilarity function λ_X^Ω has values given by (4.4). That (X, λ_X^Ω) is a symmetric network—i.e., that λ_X^Ω satisfies symmetry and identity—is shown in the proof of Proposition 4.1. Comparing the definitions of the output ultrametrics of the representable method $\mathcal{H}^\Omega$ in (4.5) and of the single linkage method in section 2, we conclude that

$$\mathcal{H}^\Omega(X, A_X) = \mathcal{H}^{\mathrm{SL}}(X, \lambda_X^\Omega) = \mathcal{H}^{\mathrm{SL}}\big(\Lambda^\Omega(X, A_X)\big), \tag{4.7}$$

as wanted. ■

[3] Longer proofs such as the one associated with this result have been deferred to the appendix.

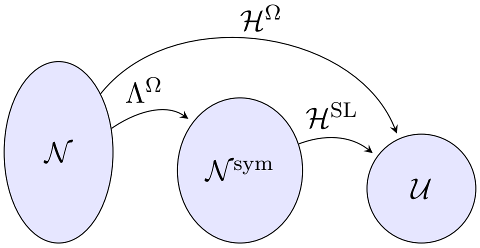


Figure 7. *Decomposition of representable methods. A representable method can be decomposed into a map from the collection of asymmetric networks to the collection of symmetric networks composed with the single linkage map into the collection of ultrametrics. See Proposition 4.2.*

Representable clustering methods, as all other hierarchical clustering methods, are maps from the collection of asymmetric networks $\mathcal{N}$ to the collection of ultrametrics $\mathcal{U}$; see Figure 7. Proposition 4.2 allows the decomposition of these maps into two components with definite separate roles. The first element of the composition is the function Λ^{Ω} whose objective is to symmetrize the original, possibly asymmetric, dissimilarity function. This transformation is followed by an application of single linkage $\mathcal{H}^{\mathrm{SL}}$ with the goal of inducing an ultrametric structure on this symmetric, but not necessarily ultrametric, intermediate network. Proposition 4.2 attests that there may be many different ways of inducing a symmetric structure depending on the selection of the representers in Ω but that there is a unique method to induce ultrametric structure. This unique method is single linkage hierarchical clustering.

From an algorithmic perspective, Proposition 4.2 implies that the computation of ultrametrics arising from representable methods requires a symmetrizing operation that depends on Ω followed by application of a single linkage algorithm; see, e.g., [22]. A related decomposition result is derived in [8, Theorem 6.3] for clustering in metric spaces. Proposition 4.2 is a significant extension of this result which applies not only to finite metric spaces but also to asymmetric networks in general. For the case of metric spaces, when sensitivity to density might be a desirable property, suitable choices of representers Ω are known to induce this behavior; see [8, section 6.7 and Definition 7.2].

The converse of Proposition 4.2 is not true, i.e., the composition of *any* symmetrizing map followed by single linkage need not correspond to a representable method for some family of representers Ω. To see this, consider the grafting method $\mathcal{H}^{\mathrm{R/NR}}(\beta)$ introduced in section 3.2. We can think of the application of $\mathcal{H}^{\mathrm{R/NR}}(\beta)$ as a symmetrizing map ($\mathcal{H}^{\mathrm{R/NR}}(\beta)$ itself) followed by single linkage. In this case, the application of single linkage would be moot, since the image of $\mathcal{H}^{\mathrm{R/NR}}(\beta)$ is already an ultrametric (this follows from the fact that single linkage attains the maximal subdominant ultrametric [7, Corollary 14]). Thus, we have argued that $\mathcal{H}^{\mathrm{R/NR}}(\beta)$ can be decomposed as in Figure 7; however, $\mathcal{H}^{\mathrm{R/NR}}(\beta)$ is not representable since we have shown that it is not linear scale preserving and that would violate our main result in Theorem 5.1.

5. A generative model for robust hierarchical clustering methods.

Our main theorem establishes the equivalence between the classes of representable and robust hierarchical clustering methods.

Theorem 5.1. *Given an admissible hierarchical clustering method $\mathcal{H}$, it is robust,* (P1), (P2), (P3)*, if and only if it is representable,* (P4).

Intuitively, the relationship between representability and robustness stated in Theorem 5.1 originates from the fact that both concepts address the locality of clustering methods. Representability (P4) implies that the method can be interpreted as an extension of particular cases or representers. In a related fashion, excisiveness (P1) requires the clustering of local subnetworks to be consistent with the clustering of the entire network.

The importance of Theorem 5.1 resides in relating implicit properties of a clustering method with practical relevance such as linear scale preservation with a generative model of clustering methods such as representability. Thus, when designing a clustering method for a particular application, if robustness is a desirable property, then Theorem 5.1 asserts that representability must be considered as a generative model. Conversely, it is unclear how to establish directly whether a given clustering method is representable. However, Theorem 5.1 provides an indirect way to prove representability via the analysis of excisiveness, linear scale preservation, and stability.

In section 3, we described an admissible method (grafting) which is not linear scale preserving and another one (semireciprocal clustering) which is not excisive. Hence, Theorem 5.1 states that neither of these methods is representable. Conversely, by combining Theorem 5.1 with Propositions 3.2, 3.3, and 3.4, we can ensure that the reciprocal $\mathcal{H}^{\mathrm{R}}$ and nonreciprocal $\mathcal{H}^{\mathrm{NR}}$ methods are both representable. Indeed, in Figure 8 we exhibit the collections of representers associated with each of the two methods, i.e., $\mathcal{H}^{\mathrm{R}} \equiv \mathcal{H}^{\omega_{\mathrm{R}}}$ and $\mathcal{H}^{\mathrm{NR}} \equiv \mathcal{H}^{\Omega_{\mathrm{NR}}}$.

To see why the equivalence stated in Theorem 5.1 is true for the case of reciprocal clustering, pick an arbitrary network $N = (X, A_X)$ and notice that the expansion constant (cf. (4.1)) of any map ϕ from ω_{R} to N is equal to

$$L(\phi; \omega_{\mathrm{R}}, N) = \max\big(A_X(\phi(z), \phi(z')), A_X(\phi(z'), \phi(z))\big), \tag{5.1}$$

where z and z' denote the two nodes of the representer ω_{R}. Moreover, from the definition of optimal multiple between nodes $x, x' \in X$, we know that nodes x and x' must be the images of z and z' under ϕ which implies that

$$\lambda_X^{\omega_{\mathrm{R}}}(x, x') = \max(A_X(x, x'), A_X(x', x)). \tag{5.2}$$

By tering
(2.6

ıair of
arbi (x', x)
thaı l)), to

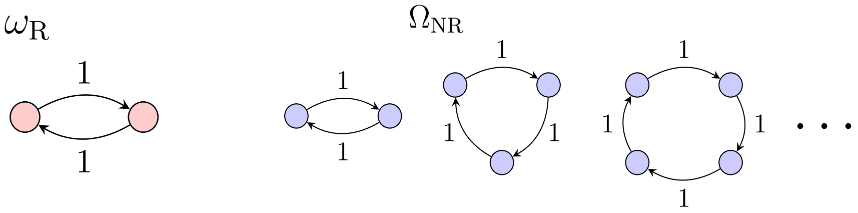


Figure 8. *$\mathcal{H}^{\mathrm{R}}$ can be represented by one representer ω_{R} while $\mathcal{H}^{\mathrm{NR}}$ requires a countably infinite collection Ω_{NR} of representers.*

obtain a loop. The maximum dissimilarity in this loop is equal to $\max(\tilde{u}^*_X(x,x'), \tilde{u}^*_X(x',x))$ which is exactly $u^{\mathrm{NR}}_X(x,x')$ (cf. (2.7)). Furthermore, if this loop is composed of k nodes, then we may pick the representer in Ω_{NR} with exactly k nodes and map it injectively to the loop. Since by construction x and x' belong to the image of the map and its expansion constant is equal to the maximum dissimilarity in the loop $u^{\mathrm{NR}}_X(x,x')$, we obtain that $\lambda^{\Omega_{\mathrm{NR}}}_X(x,x') = u^{\mathrm{NR}}_X(x,x')$, from which the result follows.

In general, one can design representers Ω different from those in Figure 8 to capture diverse structures in the directed network under study, thus leading to representable methods that go beyond reciprocal and nonreciprocal clustering. An example of this is given in section 6. Moreover, regardless of the particular choice of Ω, Theorem 5.1 guarantees that the resulting hierarchical clustering method will be robust. This is also illustrated through a numerical experiment in the next section.

6. Experimental illustration.

The U.S. Department of Commerce publishes a yearly table of inputs and outputs organized by economic sectors.[4] We focus on a specific section of this table, called *uses*, that corresponds to the inputs to production for different industrial sectors. More precisely, we are given a set I of 61 industrial sectors as defined by the North American Industry Classification System and a similarity function $U : I \times I \to \mathbb{R}_+$, where $U(i,i')$ represents how much of the production of sector i (in dollars) is used as an input of sector i'. Based on this, we define the network $N_I = (I, A_I)$ where the dissimilarity function A_I satisfies $A_I(i,i) = 0$ for all $i \in I$ and, for $i \neq i' \in I$, is given by

$$A_I(i,i') := \left(\frac{U(i,i')}{\sum_k U(k,i')} \right)^{-1}. \tag{6.1}$$

The normalization in (6.1) can be interpreted as the proportion of the input to productive sector i' that comes from sector i. Consequently, we focus on the relative combination of inputs of a sector rather than the size of the economic sector itself. Moreover, we compute the inverse of this normalized quantity to obtain a measure A_I that represents dissimilarities. That is, if most of the productive input of i' comes from i, then the normalization would output a number close to 1 and the dissimilarity measure $A_I(i,i')$ would be small.

We hierarchically cluster the network N_I of economic sectors using the representable method $\mathcal{H}^\omega$ associated with the representer ω in Figure 9 (left); see section 6.1 for details. From the structure of ω, the method $\mathcal{H}^\omega$ clusters two nodes if they can be joined via cycles of at most three nodes with strong connection in one direction—represented by the dissimilarities equal to 1—while simultaneously having not too weak connections in the opposite direction—represented by the dissimilarities equal to 3.

In Figure 9 (left), we present the output dendrogram when the method $\mathcal{H}^\omega$ is applied to N_I. Implementation details of this particular clustering method can be found in section 6.1. Theorem 5.1 guarantees that if we take a branch of the dendrogram in Figure 9 (left), e.g., the one highlighted in red, and focus on a subnetwork of the economic network spanned by the corresponding industrial sectors and cluster this subnetwork, we obtain a dendrogram equivalent to the red branch. Indeed, this is the case as can be seen in Figure 9 (right).

[4] Available at http://www.bea.gov/industry/io_annual.htm.

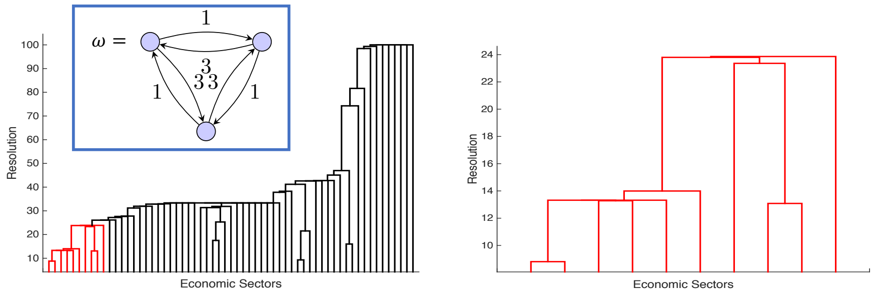


Figure 9. *(Left) Dendrogram obtained when clustering the economic network N_I using the representable clustering method $\mathcal{H}^\omega$, for the representer ω shown. (Right) Illustration of excisiveness. When clustering the subnetwork spanned by the economic sectors corresponding to the red branch in the left, the output dendrogram matches the branch.*

Similarly, we can multiply the economic network by a scalar and cluster the resulting multiple network and we are guaranteed to obtain a multiple of the original dendrogram (cf. (P2)), and small perturbations to the network result in small perturbations to the output dendrogram (cf. (P3)).

6.1. Implementation of the representable method $\mathcal{H}^\omega$.

First notice that for an arbitrary network $N_X = (X, A_X)$, the disimilarity function A_X can be represented as a matrix which, as it does not lead to confusion, we also denote as $A_X \in \mathbb{R}^{n\times n}$. Define the matrix B_X where each element is given by

$$[B_X]_{ij} = \min_k \max\Big([A_X]_{ij}, [A_X]_{jk}, [A_X]_{ki}, [A_X]_{ji}/3, [A_X]_{kj}/3, [A_X]_{ik}/3\Big). \tag{6.2}$$

By comparing (6.2) with (4.1), it follows that the element i, j of matrix B_X stores the minimum of the expansion constant of a map ϕ from the representer ω to the network N_X with nodes i and j in its image *and* mapping a unit dissimilarity in ω to the directed dissimilarity from i to j. From this interpretation of B_X it follows immediately that the symmetric matrix $\Lambda_X := \min(B_X, B_X^T)$ contains as elements the optimal multiples, i.e., $[\Lambda_X]_{i,j} = \lambda_X^\omega(i,j)$. To see this, notice that the optimal map from ω to N_X attaining the minimum expansion constant in (4.2) must contain nodes i and j in its image and must map a unit dissimilarity in ω either to the directed dissimilarity from i to j or from j to i, thus $\lambda_X^\omega(i,j) = \min([B_X]_{ij}, [B_X]_{ji})$.

Finally, we compute the output ultrametric as in (4.3), which is equivalent to applying single linkage clustering to the symmetric network (X, Λ_X); thus any known single linkage algorithm [22] can be used for this last step.

7. Conclusion.

We defined *robustness* of hierarchical clustering methods via the fulfillment of three properties: excisiveness (the clustering output of a subnetwork does not depend on the information beyond the subnetwork), linear scale preservation (the clustering output is not modified by a change of units), and stability (a small perturbation in the network entails a small perturbation in the clustering output). As a generative model for hierarchical clustering methods we introduced the concept of representability. The behavior of representable methods is determined by specifying their output on a collection of representers. Moreover, we showed that every representable method can be decomposed into two phases: a symmetrizing map Λ^Ω followed by single linkage clustering. This decomposition result enables the decoupled implementation of hierarchical clustering methods of practical relevance. Our main result was the proof that, within the set of admissible hierarchical clustering methods, the subset of representable methods coincides with the class of robust methods as determined by the three aforementioned properties.

For future work, it seems interesting to understand how the complexity of computing Λ^Ω depends on the structure of the collection of associated representers Ω. It also seems of particular interest to expand the list of desirable practical properties in order to get a more stringent characterization of the methods that are relevant in practice. We can consider, e.g., the notion of scale preservation for general dissimilarity transformations *not* restricted to linear transformations as done in this paper. One overarching aim is to identify conditions that need to be imposed on the representers so that the associated representable method complies with the stricter notion of practicality.

Appendix A. Relegated proofs.

A.1. Proof of Proposition 4.1. Given a collection Ω of representers $\omega = (X_\omega, A_\omega)$, we want to see that for an arbitrary network $N = (X, A_X)$ the output $(X, u_X^\Omega) = \mathcal{H}^\Omega(X, A_X)$ satisfies the identity, symmetry, and strong triangle inequality properties of an ultrametric. To show that the strong triangle inequality in (2.2) is satisfied let $C^*(x, x')$ and $C^*(x', x'')$ be minimizing chains for $u_X^\Omega(x, x')$ and $u_X^\Omega(x', x'')$, respectively. Consider then the chain $C(x, x'')$ obtained by concatenating $C^*(x, x')$ and $C^*(x', x'')$, in that order. Notice that the maximum over i of the optimal multiples $\lambda_X^\Omega(x_i, x_{i+1})$ in $C(x, x'')$ does not exceed the maximum multiples in each individual chain. Thus, the maximum multiple in the concatenated chain $C(x, x'')$ suffices to bound $u_X^\Omega(x, x'') \leq \max(u_X^\Omega(x, x'), u_X^\Omega(x', x''))$ by (4.5) as in (2.2).

To show the symmetry property, $u_X^\Omega(x, x') = u_X^\Omega(x', x)$ for all $x, x' \in X$, first notice that a direct implication of the definition of optimal multiples in (4.2) is that $\lambda_X^\omega(x, x') = \lambda_X^\omega(x', x)$ for every representer ω. From (4.4) we then obtain that λ_X^Ω is symmetric, from where symmetry of u_X^Ω immediately follows.

For the identity property, i.e., $u_X^\Omega(x, x') = 0$ if and only if $x = x'$, we first show that if $x = x'$ we must have $u_X^\Omega(x, x') = 0$. Pick any $x \in X$, let $x' = x$, and pick the chain $C(x, x) = [x, x]$ starting and ending at x with no intermediate nodes as a candidate minimizing chain in (4.5). While this particular chain need not be optimal in (4.5) it nonetheless holds that

$$0 \leq u_X^\Omega(x, x) \leq \lambda_X^\Omega(x, x), \tag{A.1}$$

where the first inequality holds because all costs $\lambda_X^\Omega(x_i, x_{i+1})$ in (4.5) are nonnegative since they correspond to the expansion constant of some map, which is nonnegative by definition (4.1). Notice that for the cost $\lambda_X^\omega(x, x)$ in (4.2), we minimize the expansion constant among maps $\phi_{x,x}$ that are only required to have node x in its image. Thus, consider the map that takes all the nodes in any representer $\omega \in \Omega$ into node $x \in X$. From (4.1), the expansion constant of this map is zero, which implies by (4.2) that $\lambda_X^\omega(x, x) = 0$ for all $\omega \in \Omega$. Combining this result with (4.4) we then get that $\lambda_X^\Omega(x, x) = 0$ and from (A.1) we conclude that $u_X^\Omega(x, x) = 0$.

In order to show that the condition $u_X^\Omega(x, x') = 0$ implies that $x = x'$ we prove that if $x \neq x'$ we must have $u_X^\Omega(x, x') > \alpha > 0$ for some strictly positive constant α. In proving this, we make use of the following claim.

Claim A.1. *Given a network $N = (X, A_X)$, a weakly connected representer $\omega = (X_\omega, A_\omega)$, and a dissimilarity reducing map $\phi : X_\omega \to X$ whose image satisfies $|Im(\phi)| \geq 2$, there exists a pair of points $(z, z') \in dom(A_\omega)$ for which $\phi(z) \neq \phi(z')$.*

Proof. Suppose that $\phi(z^1) = x^1$ and $\phi(z^2) = x^2$, with $x^1 \neq x^2 \in X$. These nodes can always be found since $|\mathrm{Im}(\phi)| \geq 2$. By our hypothesis, the network is weakly connected. Hence, there must exist a chain $C(z^1, z^2) = [z^1 = z_0, z_1, \ldots, z_l = z^2]$ linking z^1 and z^2 for which either $(z_i, z_{i+1}) \in \mathrm{dom}(A_\omega)$ or $(z_{i+1}, z_i) \in \mathrm{dom}(A_\omega)$ for all $i = 0, \ldots, l-1$. Focus on the image of this chain under the map ϕ, $C(x^1, x^2) = [x^1 = \phi(z_0), \phi(z_1), \ldots, \phi(z_l) = x^2]$. Notice that not all the nodes are necessarily distinct. However, since the extreme nodes are different by construction, at least one pair of consecutive nodes must differ, say, $\phi(z_p) \neq \phi(z_{p+1})$. Due to ω being weakly connected, in the original chain we must have either (z_p, z_{p+1}) or $(z_{p+1}, z_p) \in \mathrm{dom}(A_\omega)$. Hence, either $z = z_p$ and $z' = z_{p+1}$ or vice versa must fulfill the statement of the claim. ■

Returning to the main argument, observe that since pairwise dissimilarities in all networks $\omega \in \Omega$ are uniformly bounded, the maximum dissimilarity across all links of all representers

$$d_{\max} = \sup_{\omega \in \Omega} \max_{(z,z') \in \mathrm{dom}(A_\omega)} A_\omega(z, z') \tag{A.2}$$

is guaranteed to be finite. Define the separation of the network as its minimum positive dissimilarity, i.e., $\mathrm{sep}(X, A_X) := \min_{x \neq x'} A_X(x, x')$, and pick any real α such that $0 < \alpha < \mathrm{sep}(X, A_X)/d_{\max}$. Then for all $(z, z') \in \mathrm{dom}(A_\omega)$ and all $\omega \in \Omega$ we have

$$\alpha\ A_\omega(z, z') < \mathrm{sep}(X, A_X). \tag{A.3}$$

Claim A.1 implies that regardless of the map ϕ chosen, this map transforms some defined dissimilarity in ω, i.e., $A_\omega(z, z')$ for some $(z, z') \in \mathrm{dom}(A_\omega)$, into a dissimilarity in N. Moreover, every positive dissimilarity in N is greater than or equal to the network separation $\mathrm{sep}(X, A_X)$. Hence, (A.3) implies that there cannot be any dissimilarity reducing map ϕ with $|\mathrm{Im}(\phi)| \geq 2$ from $\alpha * \omega$ to N for any $\omega \in \Omega$. From (4.2), this implies that for all $x \neq x' \in X$ and for all ω we have that $\lambda_X^\omega(x, x') > \alpha > 0$. Hence, from (4.4) we conclude that $\lambda_X^\Omega(x, x') > \alpha > 0$, which in turn implies that the ultrametric value between two different nodes $u_X^\Omega(x, x')$ must be strictly positive.

A.2. Proof of Theorem 5.1.

We first prove that (P4) implies (P1)–(P3). Notice that the expansion constants of arbitrary maps (4.1) satisfy

$$L(\phi; \omega, \alpha * N) = \alpha\, L(\phi; \omega, N) \tag{A.4}$$

for any positive constant $\alpha > 0$. That (P4) implies (P2) follows by combining the linear relation in (A.4) with the definition of a representable method in (4.5).

To show that representability implies excisiveness (P1), we must prove that (3.3) is true for a general representable clustering method $\mathcal{H}^\Omega$. Hence, consider a network $N = (X, A_X)$, a resolution $\delta > 0$, and a subnetwork $N_i^\delta = (B_i(\delta),\ A_X|_{B_i(\delta) \times B_i(\delta)})$ as defined in (3.1), and define the output ultrametrics $(X, u_X^\Omega) = \mathcal{H}^\Omega(N)$ and $(X, u_{N_i^\delta}^\Omega) = \mathcal{H}^\Omega(N_i^\delta)$. Since the identity map from N_i^δ to N is dissimilarity reducing, admissibility of $\mathcal{H}^\Omega$ implies (cf. the axiom of transformation, (A2))

$$u^\Omega_{N_i^\delta}(x,x') \geq u^\Omega_X(x,x') \tag{A.5}$$

for all $x, x' \in B_i(\delta)$. In order to show the reverse inequality, pick arbitrary nodes $x, x' \in B_i(\delta)$. From the definition of subnetwork (3.2), it must be that

$$u^\Omega_X(x,x') \leq \delta, \qquad u^\Omega_X(x,x'') > \delta \tag{A.6}$$

for all $x'' \notin B_i(\delta)$. The leftmost inequality in (A.6) implies that there exists a minimizing chain $C(x,x') = [x = x_0, x_1, \ldots, x_l = x']$ in definition (4.5) and a series of maps $\phi_{x_j,x_{j+1}}$ for all j determining the optimal multiples $\lambda^\Omega_X(x_j, x_{j+1}) \leq \delta$. Notice that the ultrametric value between any two nodes in the images of the maps $\phi_{x_j,x_{j+1}}$ is smaller than or equal to δ. Hence, from (A.6) we have that the minimizing chain $C(x,x')$ and the image of every optimal dissimilarity reducing map are contained in $B_i(\delta)$ so that the same chain can be used to compute $u^\Omega_{N_i^\delta}(x,x')$. This implies that

$$u^\Omega_{N_i^\delta}(x,x') \leq u^\Omega_X(x,x') \tag{A.7}$$

for all $x, x' \in B_i(\delta)$. Combining (A.5) with (A.7) we obtain (3.3), showing that (P4) implies (P1).

To prove that (P4) implies (P3), we resort to Proposition 4.2, where we have that $\mathcal{H}^\Omega \equiv \mathcal{H}^{\text{SL}} \circ \Lambda^\Omega$. In [7] it was shown that $d_\mathcal{N}(\mathcal{H}^{\text{SL}}(N_X), \mathcal{H}^{\text{SL}}(N_Y)) \leq d_\mathcal{N}(N_X, N_Y)$ for any N_X and N_Y in $\mathcal{N}$. Thus, in order to establish our claim it is enough to prove that there exists a finite constant $L = L(\Omega) \geq 0$ such that

$$d_\mathcal{N}(\Lambda^\Omega(N_X), \Lambda^\Omega(N_Y)) \leq L\, d_\mathcal{N}(N_X, N_Y). \tag{A.8}$$

We claim this to be true for $L(\Omega) := \big(\text{sep}(\Omega)\big)^{-1}$.

In order to verify this, assume that $\eta = d_\mathcal{N}(N_X, N_Y)$ and pick any correspondence R between X and Y such that $|A_X(x,x') - A_Y(y,y')| \leq 2\eta$ for all (x,y) and (x',y') in R (cf. (3.13)). Fix any two pairs (x,y) and (x',y') in R. For any representer $\omega \in \Omega$, let $\phi : \omega \to X$ be any map such that $x, x' \in \text{Im}(\phi)$. Moreover, consider any function $\varphi : X \to Y$ such that $\varphi(x) = y$ and $\varphi(x') = y'$ and $(x'', \varphi(x'')) \in R$ for all $x'' \in X$. Notice that the definition of correspondence ensures that at least one such function φ exists. Then, we have

$$L(\varphi\circ\phi;\omega,N_Y) \leq \max_{\substack{(z,z')\in\text{dom}(A_\omega)\\ z\neq z'}} \frac{A_X(\phi(z),\phi(z'))}{A_\omega(z,z')} + 2\eta\ \text{sep}(\omega)^{-1} = L(\phi;\omega,N_X) + 2\eta\ \text{sep}(\omega)^{-1}. \tag{A.9}$$

By construction, $y, y' \in \text{Im}(\varphi\circ\phi)$. Thus, $L(\varphi\circ\phi;\omega,N_Y)$ is an upper bound for the optimal multiple $\lambda^\omega_Y(y,y')$ so from (A.9) it follows that

$$\lambda^\omega_Y(y,y') \leq L(\phi;\omega,N_X) + 2\eta\ \text{sep}(\omega)^{-1}. \tag{A.10}$$

This inequality is valid for all functions $\phi : \omega \to X$ s.t. $x, x' \in \mathrm{Im}(\phi)$. Thus, for the particular map ϕ minimizing $L(\phi; \omega, N_X)$, (A.10) becomes $\lambda_Y^\omega(y, y') \leq \lambda_X^\omega(x, x') + 2\eta \ \mathrm{sep}(\omega)^{-1}$. By symmetry, we obtain $|\lambda_X^\omega(x, x') - \lambda_Y^\omega(y, y')| \leq 2\eta \ \mathrm{sep}(\omega)^{-1}$ for all $(x, y), (x', y') \in R$. It then follows that

$$|\lambda_X^\Omega(x, x') - \lambda_Y^\Omega(y, y')| \leq 2\eta \ \mathrm{sep}(\Omega)^{-1}, \tag{A.11}$$

as claimed, where the fact that we require $\mathrm{sep}(\Omega) > 0$ guarantees that (A.11) is well-defined. This completes the proof that (P4) implies (P1)–(P3).

To prove the converse statement, consider an arbitrary admissible clustering method $\mathcal{H}$ which is excisive, linear scale preserving, and stable. We will construct a representable method $\mathcal{H}^\Omega$ such that $\mathcal{H} \equiv \mathcal{H}^\Omega$.

Denote by $(X, u_X) = \mathcal{H}(X, A_X)$ an arbitrary output ultrametric and define the collection of representers Ω as follows:

$$\Omega = \left\{ \omega \ \middle| \ \omega = \frac{1}{\max\limits_{x,x' \in B_i(\delta)} u_X(x, x')} * N_i^\delta, |B_i(\delta)| > 1, \delta > 0 \right\} \tag{A.12}$$

for all resolutions $\delta > 0$ and $N_i^\delta := (B_i(\delta), A_X|_{B_i(\delta) \times B_i(\delta)})$ being a subnetwork of all possible networks $N = (X, A_X)$ given the method $\mathcal{H}$. In other words, we pick as representers the collection of all possible subnetworks generated by the method $\mathcal{H}$, each of them scaled by the inverse of the maximum ultrametric obtained in such subnetwork. Notice that from the definition of subnetwork (3.2) we have that

$$\max_{x,x' \in B_i(\delta)} u_X(x, x') \leq \delta, \tag{A.13}$$

which appears in the denominator of the definition (A.12) for every representer $\omega \in \Omega$.

We show equivalence of methods $\mathcal{H}$ and $\mathcal{H}^\Omega$ by showing that the ultrametric outputs coincide for every network. Pick an arbitrary network $N = (X, A_X)$ and two different nodes $x, x' \in X$ and define $\alpha := u_X(x, x')$. Since Ω was built considering all possible networks, including N, there is a representer $\omega \in \Omega$ that corresponds to the subnetwork N_i^α at resolution α that contains x and x'. From (A.13), the inclusion map ϕ from $\alpha * \omega$ to N such that $\phi(x) = x$ is dissimilarity reducing and $x, x' \in \mathrm{Im}(\phi)$. From definition (4.2) this implies that $\lambda_X^\omega(x, x') \leq \alpha$. By substituting in (4.4) and further substitution in (4.5) we obtain that $u_X^\Omega(x, x') \leq \alpha$. Recalling that $\alpha = u_X(x, x')$ and that we chose the network N and the pair of nodes x, x' arbitrarily, we may conclude that $u_X^\Omega \leq u_X$ for every network N.

In order to show the other direction of the inequality, we must first observe that for every representer, the ultrametric value given by $\mathcal{H}$ between any pair of nodes in the representer is upper bounded by 1. To see this, given a representer $\omega = (X_\omega, A_{X_\omega})$ associated with the subnetwork N_i^δ in (A.12) we have that

$$u_{X_\omega}(\tilde{x}, \tilde{x}') = \frac{1}{\max\limits_{x,x' \in B_i(\delta)} u_X(x, x')} \, u_{B_i(\delta)}(\tilde{x}, \tilde{x}') = \frac{1}{\max\limits_{x,x' \in B_i(\delta)} u_X(x, x')} \, u_X|_{B_i(\delta) \times B_i(\delta)}(\tilde{x}, \tilde{x}') \leq 1$$

for all $\tilde{x}, \tilde{x}' \in X_\omega$. The first equality in (A.2) is implied by the definition of ω in (A.12) and linear scale preservation of $\mathcal{H}$. The second equality is derived from excisiveness of $\mathcal{H}$.

Pick an arbitrary network $N = (X, A_X)$ and a pair of nodes $x, x' \in X$ and define $\beta := u_X^\Omega(x, x')$. This means that there exists a minimizing chain $C(x, x') = [x' = x_0, x_1, \dots, x_l = x']$ such that for every consecutive pair of nodes we can find a dissimilarity reducing map $\phi_{x_j, x_{j+1}}$ from $\beta * \omega_j$ to N for some representer $\omega_j \in \Omega$ such that $x_j, x_{j+1} \in \mathrm{Im}(\phi_{x_j, x_{j+1}})$. Focus on a particular pair of consecutive nodes x_j, x_{j+1} and denote by p_j, p_{j+1} two respective preimages on $\omega_j = (X_{\omega_j}, A_{X_{\omega_j}})$ under the map $\phi_{x_j, x_{j+1}}$. Without loss of generality, we can assume that $x_j \neq x_{j+1}$ for all j. The preimages need not be unique. Denote by $\beta * \omega_j = (X_{\omega_j}^\beta, \beta\, A_{X_{\omega_j}})$ the β multiple of the representer ω_j. Since $\phi_{x_j, x_{j+1}}$ is a dissimilarity reducing map from $\beta * \omega_j$ to N, the axiom of transformation, (A2), implies that

$$u_{X_{\omega_j}^\beta}(p_j, p_{j+1}) \geq u_X(x_j, x_{j+1}). \tag{A.14}$$

Moreover, we can assert that

$$u_{X_{\omega_j}^\beta}(p_j, p_{j+1}) = \beta\, u_{X_{\omega_j}}(p_j, p_{j+1}) \leq \beta, \tag{A.15}$$

where the equality is due to linear scale preservation and the inequality is justified by (A.2). From the combination of (A.14) and (A.15) we obtain that $u_X(x_j, x_{j+1}) \leq \beta$. Since this is true for an arbitrary pair of consecutive nodes in $C(x, x')$, from the strong triangle inequality we have that $u_X(x, x') \leq \max_j u_X(x_j, x_{j+1}) \leq \beta$. Recalling that $\beta = u_X^\Omega(x, x')$ and that the network N was arbitrary, we can conclude that $u_X^\Omega \geq u_X$ for every network $N = (X, A_X)$. Combining this with $u_X^\Omega \leq u_X$, we conclude that $u_X^\Omega = u_X$, completing the proof.

Acknowledgment. The authors would like to thank Alejandro Ribeiro for his valuable technical input and his thoughtful comments on an earlier version of this manuscript.

REFERENCES

[1] M. Ackerman and S. Ben-David, *Measures of clustering quality: A working set of axioms for clustering*, in Neural Information Processing Systems, 2008, pp. 121–128.

[2] L. N. F. Ana and A. K. Jain, *Robust data clustering*, in Proceedings of the Conference on Computer Vision and Pattern Recognition, Vol. 2, 2003, pp. 128–133.

[3] M.-F. Balcan, Y. Liang, and P. Gupta, *Robust hierarchical clustering*, J. Mach. Learn. Res., 15 (2014), pp. 3831–3871.

[4] S. Ben-David, U. Von Luxburg, and D. Pál, *A sober look at clustering stability*, in Proceedings of the Conference on Learning Theory, 2006, pp. 5–19.

[5] J. P. Boyd, *Asymmetric clusters of internal migration regions of France*, IEEE Trans. Syst. Man Cybern., 2 (1980), pp. 101–104.

[6] D. Burago, Y. Burago, and S. Ivanov, *A Course in Metric Geometry*, AMS, Providence, RI, 2001.

[7] G. Carlsson and F. Memoli, *Characterization, stability and convergence of hierarchical clustering methods*, J. Mach. Learn. Res., 11 (2010), pp. 1425–1470.

[8] G. Carlsson and F. Memoli, *Classifying clustering schemes*, Found. Comput. Math., 13 (2013), pp. 221–252.

[9] G. Carlsson, F. Memoli, A. Ribeiro, and S. Segarra, *Alternative axiomatic constructions for hierarchical clustering of asymmetric networks*, in Proceedings of GlobalSIP, 2013, pp. 791–794.

[10] G. Carlsson, F. Memoli, A. Ribeiro, and S. Segarra, *Axiomatic construction of hierarchical clustering in asymmetric networks*, in Proceedings of ICASSP, 2013, pp. 5219–5223.

[11] G. Carlsson, F. Memoli, A. Ribeiro, and S. Segarra, *Hierarchical clustering methods and algorithms for asymmetric networks*, in Proceedings of the Asilomar Conference on Signals, Systems, and Computers, 2013, pp. 1773–1777.

[12] G. Carlsson, F. Memoli, A. Ribeiro, and S. Segarra, *Hierarchical quasi-clustering methods for asymmetric networks*, in Proceedings of ICML, 2014, pp. 352–360.

[13] G. Carlsson, F. Mémoli, A. Ribeiro, and S. Segarra, *Hierarchical clustering of asymmetric networks*, Adv. Data. Anal. Classif., 12 (2018), pp. 65–105.

[14] K. Chaudhuri and S. Dasgupta, *Rates of convergence for the cluster tree*, in Advances in Neural Information Processing Systems, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, eds., Curran Associates, 2010, pp. 343–351.

[15] B. Chazelle, *A minimum spanning tree algorithm with inverse-Ackermann type complexity*, J. ACM, 47 (2000), pp. 1028–1047.

[16] S. Chowdhury and F. Mémoli, *Convergence of Hierarchical Clustering and Persistent Homology Methods on Directed Networks*, preprint, arXiv:1711.04211, 2017.

[17] S. Chowdhury and F. Mémoli, *Distances and Isomorphism Between Networks and the Stability of Network Invariants*, preprint, arXiv:1708.04727, 2017.

[18] S. Chowdhury and F. Mémoli, *The Metric Space of Networks*, preprint, arXiv:1804.02820, 2018.

[19] S. Chowdhury and F. Mémoli, *Persistent path homology of directed networks*, in Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, 2018, pp. 1152–1169.

[20] R. N. Dave and R. Krishnapuram, *Robust clustering methods: A unified view*, IEEE Trans. Fuzzy Systems, 5 (1997), pp. 270–293.

[21] B. Eriksson, G. Dasarathy, A. Singh, and R. Nowak, *Active clustering: Robust and efficient hierarchical clustering using adaptively selected similarities*, in Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, G. Gordon, D. Dunson, and M. Dudik, eds., PMLR 15, 2011, pp. 260–268.

[22] H. Gabow, Z. Galil, T. Spencer, and R. Tarjan, *Efficient algorithms for finding minimum spanning trees in undirected and directed graphs*, Combinatorica, 6 (1986), pp. 109–122.

[23] L. A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar, *A review of robust clustering methods*, Adv. Data Anal. Classif., 4 (2010), pp. 89–109.

[24] S. Guha, R. Rastogi, and K. Shim, *CURE: An efficient clustering algorithm for large databases*, SIGMOD Rec., 27 (1998), pp. 73–84.

[25] S. Guha, R. Rastogi, and K. Shim, *Rock: A robust clustering algorithm for categorical attributes*, Information Systems, 25 (2000), pp. 345–366.

[26] I. Guyon, U. V. Luxburg, and R. C. Williamson, *Clustering: Science or art*, in NIPS Workshop on Clustering Theory, 2009.

[27] J. Hopcroft and R. Tarjan, *Algorithm* 447*: Efficient algorithms for graph manipulation*, Commun. ACM, 16 (1973), pp. 372–378.

[28] L. Hubert, *Min and max hierarchical clustering using asymmetric similarity measures*, Psychometrika, 38 (1973), pp. 63–72.

[29] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall Advanced Reference Series, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[30] J. H. Ward, Jr., *Hierarchical grouping to optimize an objective function*, J. Amer. Statist. Assoc., 58 (1963), pp. 236–244.

[31] J. M. Kleinberg, *An impossibility theorem for clustering*, in Neural Information Processing Systems, 2002, pp. 446–453.

[32] T. V. Laarhoven and E. Marchiori, *Axioms for graph clustering quality functions*, J. Mach. Learn. Res., 15 (2014), pp. 193–215.

[33] G. N. Lance and W. T. Williams, *A general theory of classificatory sorting strategies* 1*: Hierarchical systems*, Computer J., 9 (1967), pp. 373–380.

[34] S. Lattanzi, S. Leonardi, V. Mirrokni, and I. Razenshteyn, *Robust hierarchical k-center clustering*, in Proceedings of the Conference on Innovations in Theoretical Computer Science, 2015, pp. 211–218.

[35] R. S. MacKay, S. Johnson, and B. Sansom, *How directed is a directed network?*, R. Soc. Open Sci., 7 (2020), 201138.

[36] A. G. Marques, S. Segarra, and G. Mateos, *Signal processing on directed graphs: The role of edge directionality when processing and learning from network data*, IEEE Signal Processing Magazine, 37 (2020), pp. 99–116.

[37] M. Meila, *Comparing clusterings: An axiomatic view*, in Proceedings of ICML, ACM, 2005, pp. 577–584.

[38] M. Meila and W. Pentney, *Clustering by weighted cuts in directed graphs*, Proceedings of the International Conference on Data Mining, SIAM, Philadelphia, 2007, pp. 135–144.

[39] F. Mémoli and G. V. F. Pinto, *Motivic Clustering Schemes for Directed Graphs*, preprint, arXiv:2001.00278, 2020.

[40] F. Murtagh, *Multidimensional Clustering Algorithms*, Compstat Lectures, Physica Verlag, Vienna, 1985.

[41] M. Newman and M. Girvan, *Community structure in social and biological networks*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 7821–7826.

[42] M. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Phys. Rev. E, 69 (2004), 026113.

[43] A. Ng, M. Jordan, and Y. Weiss, *On spectral clustering: Analysis and an algorithm*, in Neural Information Processing Systems, 2002, pp. 849–856.

[44] W. Pentney and M. Meila, *Spectral clustering of biological sequence data*, in National Conference on Artificial Intelligence, 2005, pp. 845–850.

[45] G. Punj and D. W. Stewart, *Cluster analysis in marketing research: Review and suggestions for application*, J. Marketing Res., 20 (1983), pp. 134–148.

[46] T. Saito and H. Yadohisa, *Data Analysis of Asymmetric Structures: Advanced Approaches in Computational Statistics*, CRC Press, Boca Raton, FL, 2004.

[47] S. Segarra, G. Mateos, A. G. Marques, and A. Ribeiro, *Blind identification of graph filters*, IEEE Trans. Signal Process., 65 (2017), pp. 1146–1159.

[48] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Mach. Intell., 22 (2000), pp. 888–905.

[49] P. B. Slater, *Hierarchical internal migration regions of France*, IEEE Trans. Syst. Man Cybern., 4 (1976), pp. 321–324.

[50] P. B. Slater, *A partial hierarchical regionalization of* 3140 *US counties on the basis of* 1965-1970 *intercounty migration*, Env. Plan. A, 16 (1984), pp. 545–550.

[51] R. E. Tarjan, *An improved algorithm for hierarchical clustering using strong components*, Inform. Process. Lett., 17 (1983), pp. 37–41.

[52] U. Von Luxburg, *A tutorial on spectral clustering*, Stat. Comput., 17 (2007), pp. 395–416.

[53] U. Von Luxburg and S. Ben-David, *Towards a statistical theory of clustering*, in PASCAL Workshop on Statistics and Optimization of Clustering, 2005.

[54] D. Walsh and L. Rybicki, *Symptom clustering in advanced cancer*, Supp. Care Cancer, 14 (2006), pp. 831–836.

[55] D. Wishart, *Mode analysis: A generalization of nearest neighbor which reduces chaining effects*, in Numerical Taxonomy, Academic Press, New York, 1969, pp. 282–311.

[56] H. Yu and M. Gerstein, *Genomic analysis of the hierarchical structure of regulatory networks*, Proc. Natl. Acad. Sci. USA, 103 (2006), pp. 14724–14731.

[57] R. B. Zadeh and S. Ben-David, *A uniqueness theorem for clustering*, in Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, 2009, pp. 639–646.

[58] Y. Zhao and G. Karypis, *Hierarchical clustering algorithms for document datasets*, Data Min. Knowl. Discov., 10 (2005), pp. 141–168.