

MIT Open Access Articles

Hierarchical clustering of asymmetric networks

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Carlsson, Gunnar et al. "Hierarchical Clustering of Asymmetric Networks." *Advances in Data Analysis and Classification* 12, 1 (November 2017): 65–105 © Springer-Verlag

As Published: <http://dx.doi.org/10.1007/s11634-017-0299-5>

Publisher: Springer-Verlag

Persistent URL: <http://hdl.handle.net/1721.1/115058>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



Hierarchical Clustering of Asymmetric Networks

Gunnar Carlsson · Facundo Mémoli ·
Alejandro Ribeiro · Santiago Segarra

Received: date / Accepted: date

Abstract This paper considers networks where relationships between nodes are represented by directed dissimilarities. The goal is to study methods that, based on the dissimilarity structure, output hierarchical clusters, i.e., a family of nested partitions indexed by a connectivity parameter. Our construction of hierarchical clustering methods is built around the concept of admissible methods, which are those that abide by the axioms of value – nodes in a network with two nodes are clustered together at the maximum of the two dissimilarities between them – and transformation – when dissimilarities are reduced, the network may become more clustered but not less. Two particular methods, termed reciprocal and nonreciprocal clustering, are shown to provide upper and lower bounds in the space of admissible methods. Furthermore, alternative clustering methodologies and axioms are considered. In particular, modifying the axiom of value such that clustering in two-node networks occurs at the minimum of the two dissimilarities entails the existence of a unique admissible clustering method. Finally, the developed clustering methods are implemented to analyze the internal migration in the United States.

Keywords Hierarchical clustering · Asymmetric network · Directed graph · Axiomatic construction · Reciprocal clustering · Nonreciprocal clustering

Work in this paper is supported by NSF CCF-1217963, NSF CAREER CCF-0952867, NSF IIS-1422400, NSF CCF-1526513, AFOSR FA9550-09-0-1-0531, AFOSR FA9550-09-1-0643, NSF DMS-0905823, and NSF DMS-0406992.

G. Carlsson
Department of Mathematics, Stanford University
E-mail: gunnar@math.stanford.edu

F. Mémoli
Department of Mathematics and Department of Computer Science and Engineering, Ohio State University
E-mail: memoli@math.osu.edu

A. Ribeiro
Department of Electrical and Systems Engineering, University of Pennsylvania
E-mail: aribeiro@seas.upenn.edu

S. Segarra
Institute for Data, Systems, and Society, Massachusetts Institute of Technology
E-mail: segarra@mit.edu

1 Introduction

The problem of determining clusters in a data set admits different interpretations depending on whether the underlying data is metric, symmetric but not necessarily metric, or asymmetric. Of these three classes of problems, clustering of metric data is the most studied one in terms of both, practice and theoretical foundations. In terms of practice there are literally hundreds of methods, techniques, and heuristics that can be applied to the determination of hierarchical and nonhierarchical clusters in finite metric spaces – see, e.g., Xu and Wunsch (2005). Theoretical foundations of clustering methods, while not as well developed as their practical applications (Ben-David et al, 2006; Guyon et al, 2009; Von Luxburg and Ben-David, 2005), have been evolving over the past decade (Ackerman and Ben-David, 2008; Carlsson and Mémoli, 2010a,b, 2013; Kleinberg, 2002; Zadeh and Ben-David, 2009). Of particular relevance to our work is the case of hierarchical clustering where, instead of a single partition, we look for a family of partitions indexed by a resolution parameter; see e.g., Jain and Dubes (1988); Lance and Williams (1967); Zhao and Karypis (2005). In this context, it has been shown by Carlsson and Mémoli (2010a) that single linkage (Jain and Dubes, 1988, Ch. 4) is the unique hierarchical clustering method that satisfies three reasonable axioms. These axioms require that the hierarchical clustering of a metric space with two points merges the two points at a resolution given by the distance between them, that there be no non-singleton clusters at resolutions smaller than the smallest distance in the space, and that when distances shrink, the metric space may become more clustered but not less. It should be noted that in the previous discussion and throughout the paper we consider the traditional definition of a metric space (Burago et al, 2001) where the dissimilarities are positive-definite, symmetric, and satisfy the triangle inequality. As opposed to Sato (1988), Chino and Shiraiwa (1993), Okada and Iwamoto (1996), and Chino (2012) where a variety of metrics – including asymmetric metrics on general Minkowski spaces and symmetric metrics on a (complex) Hilbert space – are considered, we adopt the convention that asymmetric dissimilarities as well as those that do not satisfy the triangle inequality fall in the non-metric category.

When we remove the condition that the data be metric, we move into the realm of clustering in weighted networks, i.e. a set of nodes with pairwise and possibly *directed* dissimilarities represented by edge weights. For the undirected case, the knowledge of theoretical underpinnings is incipient but practice is well developed. Determining clusters in this undirected context is often termed community detection and is formulated in terms of finding cuts such that the edges between different groups have high dissimilarities – meaning points in different groups are dissimilar from each other – and the edges within a group have small dissimilarities – which means that points within the same cluster are similar to each other (Newman and Girvan, 2002, 2004; Shi and Malik, 2000). An alternative approach for clustering nodes in graphs is the idea of spectral clustering (Bach and Jordan, 2004; Chung, 1997; Ng et al, 2002; Von Luxburg, 2007). When a graph contains several connected components its Laplacian matrix has multiple eigenvectors associated with the null eigenvalue and the nonzero elements of the corresponding eigenvectors identify the different connected components. The underlying idea of spectral clustering is that

different communities should be identified by examining the eigenvectors associated with eigenvalues close to zero.

Further relaxing symmetry so that we can allow for asymmetric relationships between nodes (Saito and Yadohisa, 2004) reduces the number of available methods that can deal with such data (Boyd, 1980; Hubert, 1973; Meila and Pentney, 2007; Murtagh, 1985; Pentney and Meila, 2005; Slater, 1976, 1984; Tarjan, 1983; Zhou et al, 2005). Examples of these methods are the adaptation of spectral clustering to asymmetric graphs by using a random walk perspective (Pentney and Meila, 2005), the use of weighted cuts of minimum aggregate cost (Meila and Pentney, 2007), and methods based on the initial decomposition of asymmetric dissimilarity matrices into symmetric plus skew-symmetric parts and ulterior clustering of these two components (Vicari, 2014, 2015).

In spite of these contributions, the rarity of clustering methods for asymmetric networks is expected because the interpretation of clusters as groups of nodes that are closer to each other than to the rest is difficult to generalize when nodes are close in one direction but far apart in the other.

Although it is difficult to articulate a general intuition for clustering of asymmetric networks, there are nevertheless some behaviors that we should demand from any reasonable clustering method. Following Kleinberg (2002) and Carlsson and Mémoli (2010a), the perspective taken in this paper is to impose these desired behaviors as axioms and proceed to characterize the space of methods that are admissible with respect to them. While different axiomatic constructions are discussed here, the general message is that strong structure can be induced by seemingly weak axioms.

In Section 2 we introduce notions related to network theory and clustering needed for the development of the results presented in this paper. In particular, we revisit the known equivalence between dendrograms and ultrametrics (Section 2.1), which is instrumental to our proofs. The axioms of value and transformation are stated formally in Section 3 but they correspond to the following intuitions:

(A1) *Axiom of Value.* For a network with two nodes, the nodes are clustered together at a resolution equal to the maximum of the two intervening dissimilarities.

(A2) *Axiom of Transformation.* If we consider a domain network and map it into a target network in a manner such that no pairwise dissimilarity is increased by the mapping, then the resolution level at which two nodes in the target network become part of the same cluster is not larger than the level at which they were clustered together in the original domain network.

A hierarchical clustering method satisfying axioms (A1) and (A2) is said to be *admissible*. Our first theoretical study is the relationship between clustering and mutual influence in networks of arbitrary size (Section 4). In particular, we show that the outcome of any admissible hierarchical clustering method is such that a necessary condition for two nodes to cluster together is the existence of chains that allow for direct or indirect influence between the nodes. Two hierarchical clustering methods that abide by axioms (A1) and (A2) are derived in Section 5. The first method, *reciprocal clustering*, requires clusters to form through edges exhibiting low dissimilarity in both directions whereas the second method, *nonreciprocal clustering*, allows clusters to form through cycles of small dissimilarity. A fundamental result regarding admissible methods is the proof that any clustering method that satisfies axioms (A1)

and (A2) lies between reciprocal and nonreciprocal clustering in a well-defined sense (Section 6). Specifically, any clustering method that satisfies axioms (A1) and (A2) forms clusters at resolutions larger than the resolutions at which they are formed with nonreciprocal clustering, and smaller than the resolutions at which they are formed with reciprocal clustering. When restricted to symmetric networks, reciprocal and nonreciprocal clustering yield equivalent outputs, which coincide with the output of single linkage (Section 6.1). This observation is consistent with the existence and uniqueness result by Carlsson and Mémoli (2010a) since axioms (A1) and (A2) are reduced to two of the axioms considered there when we restrict attention to metric data. The derivations in our paper show that the existence and uniqueness result by Carlsson and Mémoli (2010a) is true for all symmetric, not necessarily metric, datasets and that a third axiom considered there is redundant because it is implied by the other two.

In some applications the requirement for bidirectional influence in the Axiom of Value is not justified as unidirectional influence suffices to establish proximity. Only requiring unidirectional influence leads to the study of alternative axiomatic constructions and their corresponding admissible hierarchical clustering methods (Section 7). We first propose an Alternative Axiom of Value in which clusters in two-node networks are formed at the minimum of the two dissimilarities. Under this axiomatic framework we define unilateral clustering as a method in which influence propagates through chains of nodes that are close in at least one direction (Section 7.1). Contrary to the case of admissibility with respect to (A1)-(A2) in which a range of methods exist, unilateral clustering is the unique method that is admissible with respect to the Alternative Axiom of Value. Moreover, an agnostic position where nodes in two-node networks are allowed to cluster at any resolution between the minimum and the maximum dissimilarity between them is also studied (Section 7.2).

Lastly, in Section 8 the developed clustering methods are exemplified through their application to the network of internal migration between states of the U.S. for the year 2011. The purpose of this example is to understand which information can be extracted by performing hierarchical clustering analyses based on the different methods proposed. The migration network example illustrates the different clustering outputs obtained when we consider the Axiom of Value (A1) or the Alternative Axiom of Value (A1'') as conditions for admissibility. Unilateral clustering, the unique method compatible with (A1''), forms clusters around influential states like California and Texas by merging each of these states with other smaller ones around them (Section 8.3). On the other hand, methods compatible with (A1) like reciprocal clustering, tend to first merge states with balanced bidirectional influence such as two different populous states or states sharing urban areas. In this way, reciprocal clustering sees California first merging with Texas for being two very influential states and Washington merging with Oregon for sharing the urban area of Portland (Section 8.1). Moreover, the similarity between the reciprocal and nonreciprocal outcomes (Section 8.2) indicates that no other clustering method satisfying axiom (A1) would reveal new information for this specific dataset. Concluding remarks are presented in Section 9. All proofs not included in the main body of the text can be found in the Appendix (Section 10).

Preliminary versions of some of the results here presented appeared in Carlsson et al (2013b) and Carlsson et al (2013a), two short conference papers that served the

purpose of announcing our results. More precisely, Carlsson et al (2013b) discusses some of the ideas presented in Sections 3, 5, and 6, whereas the alternative axiomatic construction developed here in Section 7 as well as the portion of Section 8 devoted to unilateral clustering were previewed in Carlsson et al (2013a).

2 Preliminaries

We define a network N_X to be a pair (X, A_X) where X is a finite set of points or nodes and $A_X : X \times X \rightarrow \mathbb{R}_+$ is a dissimilarity function. The dissimilarity $A_X(x, x')$ between nodes $x \in X$ and $x' \in X$ is assumed to be non-negative for all pairs (x, x') and 0 if and only if $x = x'$. We do not, however, require A_X to be a metric on the finite set X . Specifically, dissimilarity functions A_X need not satisfy the triangle inequality and, more consequential for the problem considered here, they may be asymmetric in that it is possible to have $A_X(x, x') \neq A_X(x', x)$ for some $x \neq x'$. We further define \mathcal{N} as the set of all networks N_X . Networks in \mathcal{N} can have different node sets X as well as different dissimilarity functions A_X .

The smallest non-trivial networks contain two nodes p and q and two dissimilarities α and β as depicted in Fig. 2. The following special networks appear often throughout our paper: consider the dissimilarity function $A_{p,q}$ with $A_{p,q}(p, q) = \alpha$ and $A_{p,q}(q, p) = \beta$ for some $\alpha, \beta > 0$ and define the *two-node network* $\Delta_2(\alpha, \beta)$ with parameters α and β as $\Delta_2(\alpha, \beta) := (\{p, q\}, A_{p,q})$.

By a clustering of the set X we mean a partition P_X of X ; i.e., a collection of sets $P_X = \{B_1, \dots, B_J\}$ which are pairwise disjoint, $B_i \cap B_j = \emptyset$ for $i \neq j$, and are required to cover X , $\cup_{i=1}^J B_i = X$. The sets B_1, B_2, \dots, B_J are called the *blocks* or *clusters* of P_X . We define the *power set* $\mathcal{P}(X)$ of X as the set containing every subset of X , thus $B_i \in \mathcal{P}(X)$ for all i . An equivalence relation \sim on X is a binary relation such that for all $x, x', x'' \in X$ we have that (1) $x \sim x$, (2) $x \sim x'$ if and only if $x' \sim x$, and (3) $x \sim x'$ and $x' \sim x''$ imply $x \sim x''$.

A partition $P_X = \{B_1, \dots, B_J\}$ of X induces and is induced by an equivalence relation \sim_{P_X} on X where, for all $x, x' \in X$, we have that $x \sim_{P_X} x'$ if and only if x and x' belong to the same block. In this paper we focus on hierarchical clustering methods. The output of hierarchical clustering methods is not a single partition P_X but a nested collection D_X of partitions $D_X(\delta)$ indexed by a resolution parameter $\delta \geq 0$. In consistency with our previous notation, for a given D_X , we say that two nodes x and x' are equivalent at resolution $\delta \geq 0$ and write $x \sim_{D_X(\delta)} x'$ if and only if nodes x and x' are in the same block of $D_X(\delta)$. The nested collection D_X is termed a *dendrogram* and is required to satisfy the following two properties plus a technical condition (Carlsson and Mémoli, 2010a):

(D1) *Boundary conditions.* For $\delta = 0$ the partition $D_X(0)$ clusters each $x \in X$ into a separate singleton and for some δ_0 sufficiently large $D_X(\delta_0)$ clusters all elements of X into a single set, $D_X(0) = \{\{x\}, x \in X\}$, $D_X(\delta_0) = \{X\}$ for some $\delta_0 > 0$.

(D2) *Hierarchy.* As δ increases clusters can be aggregated but not split. I.e., for any $\delta_1 < \delta_2$ and any pair of points x, x' for which $x \sim_{D_X(\delta_1)} x'$, the relation $x \sim_{D_X(\delta_2)} x'$ must hold as well.

The interpretation of a dendrogram is that of a structure which yields different clusterings at different resolutions. At resolution $\delta = 0$ each point is in a cluster of its own. As the resolution parameter δ increases, nodes start forming clusters. According to condition (D2), nodes become ever more clustered since once they join together in a cluster, they stay together in the same cluster for all larger resolutions. Eventually, the resolutions become coarse enough so that all nodes become members of the same cluster and stay that way as δ keeps increasing. A dendrogram can be represented as a rooted tree; see e.g. Fig. 1.

Denoting by \mathcal{D} the space of all dendrograms we define a hierarchical clustering method as a function

$$\mathcal{H} : \mathcal{N} \rightarrow \mathcal{D}, \quad (1)$$

from the space of networks \mathcal{N} to the space of dendrograms \mathcal{D} such that the underlying node set X is preserved. For the network $N_X = (X, A_X)$ we denote by $D_X = \mathcal{H}(X, A_X)$ the output of clustering method \mathcal{H} .

In the description of hierarchical clustering methods the concepts of *chain*, *chain cost*, and *minimum chain cost* are important. Given a network (X, A_X) and $x, x' \in X$, a chain from x to x' is any *ordered* sequence of nodes $[x = x_0, x_1, \dots, x_{l-1}, x_l = x']$, which starts at x and finishes at x' . We will frequently use the notation $C(x, x')$ to denote one such chain. We say that $C(x, x')$ links or connects x to x' . Given two chains $C(x, x') = [x = x_0, x_1, \dots, x_l = x']$ and $C(x', x'') = [x' = x'_0, x'_1, \dots, x'_{l'} = x'']$ such that the end point of the first one coincides with the starting point of the second one, we define the *concatenated chain* $C(x, x') \uplus C(x', x'')$ as

$$C(x, x') \uplus C(x', x'') := [x = x_0, \dots, x_l = x' = x'_0, \dots, x'_{l'} = x'']. \quad (2)$$

Observe that the chain $C(x, x') = [x = x_0, x_1, \dots, x_{l-1}, x_l = x']$ and its reverse $[x' = x_l, x_{l-1}, \dots, x_1, x_0 = x]$ are different entities even if the intermediate hops are the same. The *links* of a chain are the edges connecting its consecutive nodes in the direction imposed by the chain. We define the *cost* of a given chain $C(x, x') = [x = x_0, \dots, x_l = x']$ as $\max_{i | x_i \in C(x, x')} A_X(x_i, x_{i+1})$, i.e., the maximum dissimilarity encountered when traversing its links in order. The directed minimum chain cost $\tilde{u}_X^*(x, x')$ between x and x' is then defined as the minimum cost among all the chains connecting x to x' ,

$$\tilde{u}_X^*(x, x') := \min_{C(x, x')} \max_{i | x_i \in C(x, x')} A_X(x_i, x_{i+1}). \quad (3)$$

In asymmetric networks the minimum chain costs $\tilde{u}_X^*(x, x')$ and $\tilde{u}_X^*(x', x)$ are different in general but they are equal on symmetric networks. In this latter case, the costs $\tilde{u}_X^*(x, x') = \tilde{u}_X^*(x', x)$ are instrumental in the definition of single linkage clustering (Carlsson and Mémoli, 2010a). Indeed, for resolution δ , single linkage makes x and x' part of the same cluster if and only if they can be linked through a chain of cost not exceeding δ . Formally, the equivalence classes at resolution δ in the single linkage dendrogram SL_X over a symmetric network (X, A_X) are defined by

$$x \sim_{SL_X(\delta)} x' \iff \tilde{u}_X^*(x, x') = \tilde{u}_X^*(x', x) \leq \delta. \quad (4)$$

We further define a *loop* as a chain of the form $C(x, x)$ for some $x \in X$ such that $C(x, x)$ contains at least one node other than x . Since a loop is a particular case of a

chain, the cost of a loop is given by its largest dissimilarity. Furthermore, consistently with (3), we define the *minimum loop cost* $\text{mlc}(X, A_X)$ of a network (X, A_X) as the minimum across all possible loops of each individual loop cost,

$$\text{mlc}(X, A_X) := \min_x \min_{C(x,x)} \max_{i|x_i \in C(x,x)} A_X(x_i, x_{i+1}), \quad (5)$$

where, we recall, $C(x, x)$ contains at least one node different from x . Another relevant property of a network (X, A_X) is the *separation* of the network $\text{sep}(X, A_X)$ which we define as its minimum positive dissimilarity,

$$\text{sep}(X, A_X) := \min_{x \neq x'} A_X(x, x'). \quad (6)$$

Notice that from (5) and (6) we must have $\text{sep}(X, A_X) \leq \text{mlc}(X, A_X)$. Further observe that in the particular case of networks with symmetric dissimilarities the two quantities coincide, i.e., $\text{sep}(X, A_X) = \text{mlc}(X, A_X)$.

When one restricts attention to networks (X, A_X) having dissimilarities A_X that conform to the definition of a finite metric space – i.e., dissimilarities A_X are symmetric and satisfy the triangle inequality – it has been shown by Carlsson and Mémoli (2010a) that single linkage is the unique hierarchical clustering method satisfying axioms (A1)-(A2) in Section 3 plus a third axiom stating that clusters cannot form at resolutions smaller than the minimum distance between different points of the space. In the case of asymmetric networks the space of admissible methods is richer, as we demonstrate throughout this paper.

2.1 Dendrograms as ultrametrics

Dendrograms are convenient graphical representations but otherwise cumbersome to handle. A mathematically more convenient representation is obtained when one identifies dendrograms with finite *ultrametric* spaces. An ultrametric defined on the set X is a metric function $u_X : X \times X \rightarrow \mathbb{R}_+$ that satisfies a stronger triangle inequality as we formally define next.

Definition 1 Given a node set X , an ultrametric u_X is a non-negative function $u_X : X \times X \rightarrow \mathbb{R}_+$ satisfying the following properties:

- (i) *Identity.* The ultrametric $u_X(x, x') = 0$ if and only if $x = x'$ for all $x, x' \in X$.
- (ii) *Symmetry.* For all pairs of points $x, x' \in X$ it holds that $u_X(x, x') = u_X(x', x)$.
- (iii) *Strong triangle inequality.* Given $x, x', x'' \in X$, the ultrametrics $u_X(x, x'')$, $u_X(x, x')$, and $u_X(x', x'')$ satisfy the strong triangle inequality

$$u_X(x, x'') \leq \max(u_X(x, x'), u_X(x', x'')). \quad (7)$$

Since (7) implies the usual triangle inequality $u_X(x, x'') \leq u_X(x, x') + u_X(x', x'')$ for all $x, x', x'' \in X$, ultrametric spaces are particular cases of metric spaces.

Our interest in ultrametrics stems from the fact that it is possible to establish a structure preserving bijective mapping between dendrograms and ultrametrics as proved by the following construction and theorem; see also Fig. 1.

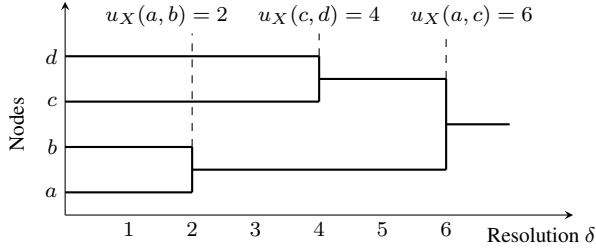


Fig. 1 Equivalence of dendrograms and ultrametrics for $X = \{a, b, c, d\}$. Given a dendrogram D_X define the function $u_X(x, x') := \min \{\delta \geq 0 \mid x \sim_{D_X(\delta)} x'\}$. This function is an ultrametric because it satisfies the identity property, the strong triangle inequality (7) and is symmetric.

Consider the map $\Psi : \mathcal{D} \rightarrow \mathcal{U}$ from the space of dendrograms to the space of networks endowed with ultrametrics, defined as follows: for a given dendrogram D_X over the finite set X write $\Psi(D_X) = (X, u_X)$, where we define $u_X(x, x')$ for all $x, x' \in X$ as the smallest resolution at which x and x' are clustered together $u_X(x, x') := \min\{\delta \geq 0 \mid x \sim_{D_X(\delta)} x'\}$. We also consider the map $\Upsilon : \mathcal{U} \rightarrow \mathcal{D}$ constructed as follows: for a given ultrametric u_X on the finite set X and each $\delta \geq 0$ define the relation $\sim_{u_X(\delta)}$ on X as $x \sim_{u_X(\delta)} x' \iff u_X(x, x') \leq \delta$. Further define $D_X(\delta) := \{X \text{ mod } \sim_{u_X(\delta)}\}$ and $\Upsilon(X, u_X) := D_X$.

Theorem 1 (Carlsson and Mémoli 2010a) *The maps Ψ and Υ are both well defined. Furthermore, $\Psi \circ \Upsilon$ is the identity on \mathcal{U} and $\Upsilon \circ \Psi$ is the identity on \mathcal{D} .*

Given the equivalence between dendrograms and ultrametrics established by Theorem 1 we can regard hierarchical clustering methods \mathcal{H} as inducing ultrametrics in node sets X based on dissimilarity functions A_X . However, ultrametrics are particular cases of dissimilarity functions. Thus, we can reinterpret the method \mathcal{H} as a map [cf. (1)]

$$\mathcal{H} : \mathcal{N} \rightarrow \mathcal{U} \quad (8)$$

mapping the space of networks \mathcal{N} to the space $\mathcal{U} \subset \mathcal{N}$ of networks endowed with ultrametrics. For all $x, x' \in X$, the ultrametric value $u_X(x, x')$ induced by \mathcal{H} is the minimum resolution at which x and x' are co-clustered by \mathcal{H} . Observe that the outcome of a hierarchical clustering method defines an ultrametric in the set X even when the original data does not correspond to a metric, as is the case of asymmetric networks. We say that two methods \mathcal{H}_1 and \mathcal{H}_2 are *equivalent*, and we write $\mathcal{H}_1 \equiv \mathcal{H}_2$, if and only if $\mathcal{H}_1(N) = \mathcal{H}_2(N)$ for all $N \in \mathcal{N}$.

3 Axioms of value and transformation

To study hierarchical clustering methods on asymmetric networks we start from intuitive notions that we translate into the axioms of value and transformation discussed in this section.

The Axiom of Value is obtained from considering the two-node network $\Delta_2(\alpha, \beta)$ defined in Section 2 and depicted in Fig. 2. 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

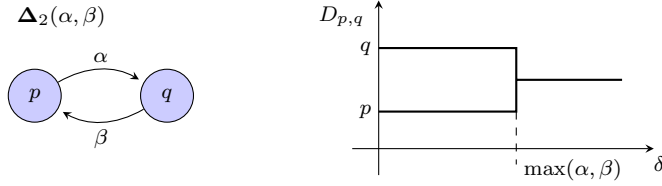


Fig. 2 Axiom of Value. Nodes in a two-node network cluster at the minimum resolution at which both can influence each other.

able to influence each other. This implies that the output dendrogram should be such that p and q are part of the same cluster at resolutions $\delta \geq \max(\alpha, \beta)$ that allow direct mutual influence. Conversely, we expect nodes p and q to be in separate clusters at resolutions $0 \leq \delta < \max(\alpha, \beta)$ that do *not* allow for mutual influence. At resolutions $\delta < \min(\alpha, \beta)$ there is no influence between the nodes and at resolutions $\min(\alpha, \beta) \leq \delta < \max(\alpha, \beta)$ there is unilateral influence from one node over the other. In either of the latter two cases the nodes are different in nature. If we think of dissimilarities as, e.g., trust, it means one node is trustworthy whereas the other is not. If we think of the network as a Markov chain, at resolutions $0 \leq \delta < \max(\alpha, \beta)$ the states are different singleton equivalence classes – one of the states would be transient and the other one absorbent. Given that, according to (8), a hierarchical clustering method is a map \mathcal{H} from networks to ultrametrics, we formalize this intuition as the following requirement on the set of admissible maps:

(A1) *Axiom of Value.* The ultrametric $(\{p, q\}, u_{p,q}) = \mathcal{H}(\Delta_2(\alpha, \beta))$ produced by \mathcal{H} applied to the two-node network $\Delta_2(\alpha, \beta)$ satisfies $u_{p,q}(p, q) = \max(\alpha, \beta)$.

Clustering nodes p and q together at resolution $\delta = \max(\alpha, \beta)$ is somewhat arbitrary, as any monotone increasing function of $\max(\alpha, \beta)$ would be admissible. As a value claim, however, it means that the clustering resolution parameter δ is expressed in the same units as the elements of the dissimilarity function.

The second restriction on the space of allowable methods \mathcal{H} formalizes our expectations for the behavior of \mathcal{H} when confronted with a transformation of the underlying set X and the dissimilarity function A_X ; see Fig. 3. Consider networks $N_X = (X, A_X)$ and $N_Y = (Y, A_Y)$ and denote by $D_X = \mathcal{H}(X, A_X)$ and $D_Y = \mathcal{H}(Y, A_Y)$ the corresponding dendrogram outputs. If we map all the nodes of the network $N_X = (X, A_X)$ into nodes of the network $N_Y = (Y, A_Y)$ in such a way that no pairwise dissimilarity is increased we expect the latter network to be more clustered than the former at any given resolution. Intuitively, nodes in N_Y are more capable of influencing each other, thus, clusters should be formed more easily. In terms of the respective dendrograms we expect that nodes co-clustered at resolution δ in D_X are mapped to nodes that are also co-clustered at this resolution in D_Y . In order to formalize this notion, we introduce the concept of a *dissimilarity-reducing map*. Given two networks $N_X = (X, A_X)$ and $N_Y = (Y, A_Y)$, map $\phi : X \rightarrow Y$ is dissimilarity reducing if it holds that $A_X(x, x') \geq A_Y(\phi(x), \phi(x'))$ for all $x, x' \in X$.

The Axiom of Transformation that we introduce next is a formal statement of the intuition described above:

(A2) *Axiom of Transformation.* Consider two networks $N_X = (X, A_X)$ and $N_Y = (Y, A_Y)$ and a dissimilarity-reducing map $\phi : X \rightarrow Y$, i.e. a map ϕ such that for

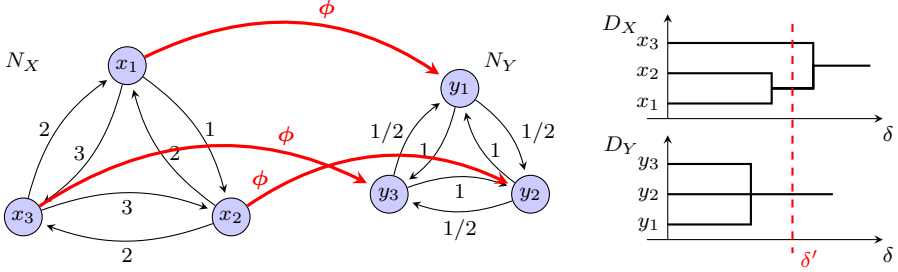


Fig. 3 Axiom of Transformation. If the network N_X can be mapped to the network N_Y using a dissimilarity-reducing map ϕ , then for every resolution δ nodes clustered together in $D_X(\delta)$ must also be clustered in $D_Y(\delta)$. E.g., since points x_1 and x_2 are clustered together at resolution δ' , their image through ϕ , i.e. $y_1 = \phi(x_1)$ and $y_2 = \phi(x_2)$, must also be clustered together at this resolution.

all $x, x' \in X$ it holds that $A_X(x, x') \geq A_Y(\phi(x), \phi(x'))$. Then, for all $x, x' \in X$, the output ultrametrics $(X, u_X) = \mathcal{H}(X, A_X)$ and $(Y, u_Y) = \mathcal{H}(Y, A_Y)$ satisfy

$$u_X(x, x') \geq u_Y(\phi(x), \phi(x')). \quad (9)$$

We say that a hierarchical clustering method \mathcal{H} is admissible with respect to (A1) and (A2), or *admissible* for short, if it satisfies axioms (A1) and (A2).

For the particular case of symmetric networks (X, A_X) we defined the single linkage dendrogram SL_X through the equivalence relations in (4). According to Theorem 1 this dendrogram is equivalent to an ultrametric space that we denote by (X, u_X^{SL}) . More specifically, as is well known (Carlsson and Mémoli, 2010a), the single linkage ultrametric u_X^{SL} in symmetric networks is given by

$$u_X^{SL}(x, x') = \tilde{u}_X^*(x, x') = \tilde{u}_X^*(x', x) = \min_{C(x, x')} \max_{i | x_i \in C(x, x')} A_X(x_i, x_{i+1}), \quad (10)$$

where we also used (3) to write the last equality.

3.1 Comments about the axiomatic framework

The axiomatic treatment of clustering goes back at least to Kleinberg (2002) who cogently argued about his proposed axioms. These axioms were scale invariance, richness (surjectivity or non-blindness to partitions), and a notion of consistency which is akin to our axiom of transformation. Kleinberg (2002) concluded that no flat (non-hierarchical) clustering method could exist that satisfies his three axioms simultaneously. These axioms, and rich classes of variants have received a very thorough treatment in the last decade, e.g., (Ackerman and Ben-David, 2008; Carlsson and Mémoli, 2010a,b, 2013; Zadeh and Ben-David, 2009).

In the work of Carlsson and Mémoli (2010a) carried out in the context of hierarchical clustering of finite metric spaces, the additional freedom inherent in not having to assign just a single partition to a dataset but a whole family of them (i.e. a dendrogram), permits the existence of a unique method that satisfies the axioms. This method turns out to be the very well known single linkage hierarchical clustering method (Jain and Dubes, 1988, Ch. 4). In the setting of the present paper,

we explore the consequences of imposing analogues of the axioms of Carlsson and Mémoli (2010a) to the setting of hierarchical clustering, but where now one wishes to process data in the form of *directed* networks. The intuition is that the methods that arise in this way will be related to single linkage and will therefore inherit some of its parsimony and nice theoretical properties.

We now analyze two features related to (A2). The reader should be aware that Sections 4 and 7 will offer more details about variants of (A1), and their interaction with (A2).

The axiom of transformation and relabelings of data points. The axiom of transformation in particular implies the following:

Proposition 1 *Assume (X, A_X) and (Y, A_Y) are networks such that there exists a bijection $\phi : X \rightarrow Y$ with the property that $A_X(x, x') = A_Y(\phi(x), \phi(x'))$ for all $x, x' \in X$. Then, any method \mathcal{H} abiding by the axiom of transformation will produce ultrametric spaces $(X, u_X) = \mathcal{H}(X, A_X)$ and $(Y, u_Y) = \mathcal{H}(Y, A_Y)$ such that*

$$u_X(x, x') = u_Y(\phi(x), \phi(x')) \text{ for all } x, x' \in X.$$

Before proving this relation, we remark what this property means: re-sorting or relabeling the points – modeled via a bijection – in a given dataset alters the result of applying an admissible method *only* in relabeling the outputs in a coherent manner. The consequence is that if a practitioner is after a method that behaves in a predictable manner when the order in which data points are listed is changed, then it is natural to require that the method satisfies the axiom of transformation.

Proof of Proposition 1: From the facts that \mathcal{H} satisfies the axiom of transformation and $A_X(x, x') \geq A_Y(\phi(x), \phi(x'))$ for all $x, x' \in X$, we have that

$$u_X(x, x') \geq u_Y(\phi(x), \phi(x')) \text{ for all } x, x' \in X. \quad (11)$$

On the other hand, since $\phi : X \rightarrow Y$ is a bijection, we can define the (inverse) function $\phi^{-1} : Y \rightarrow X$ such that $A_Y(y, y') = A_X(\phi^{-1}(y), \phi^{-1}(y'))$ for all $y, y' \in Y$. In particular, one has that $A_Y(y, y') \geq A_X(\phi^{-1}(y), \phi^{-1}(y'))$. Invoking again the satisfaction of the axiom of transformation for the method \mathcal{H} , it follows that

$$u_Y(y, y') \geq u_X(\phi^{-1}(y), \phi^{-1}(y')) \text{ for all } y, y' \in Y. \quad (12)$$

Again, because ϕ is a bijection, the above implies that

$$u_Y(\phi(x), \phi(x')) \geq u_X(x, x') \text{ for all } x, x' \in X. \quad (13)$$

Finally, equations (11) and (13) imply the claim. ■

The axiom of transformation and idempotency. One particular property implied by our axioms is that any method satisfying them will be automatically idempotent. Namely, if one applies the method once, then ulterior applications of the same method do not change the output, providing a notion of self-consistency.

Proposition 2 *Suppose $\mathcal{H} : \mathcal{N} \rightarrow \mathcal{U}$ satisfies (A1) and (A2). Then, it is idempotent, i.e. for all $(X, A_X) \in \mathcal{N}$ one has $\mathcal{H}(\mathcal{H}(X, A_X)) = \mathcal{H}(X, A_X)$.*

The proof of Proposition 2 (which is deferred to the appendix), utilizes concepts and results that will be introduced in Sections 5 and 6. From the proposition, it follows that if a practitioner is after a method enjoying idempotency, then requiring the axioms of transformation and value will guarantee this property.

4 Influence modalities

The Axiom of Value states that, in order for two nodes to belong to the same cluster, they have to be able to exercise mutual influence on each other. When we consider a network with more than two nodes the concept of mutual influence is more difficult because it is possible to have direct influence as well as indirect chains of influence through other nodes. In this section we introduce two intuitive notions of mutual influence in networks of arbitrary size and show that they can be derived from the axioms of value and transformation. Besides their intrinsic value, these influence modalities are important for later developments in this paper; see, e.g. the proof of Theorem 4.

Consider first the intuitive notion that for two nodes to be part of a cluster there has to be a way for each of them to exercise influence on the other, either directly or indirectly. To formalize this idea, recall the concept of minimum loop cost (5); see Fig. 4. For this network, the loops $[a, b, a]$ and $[b, a, b]$ have maximum cost 2 corresponding to the link (b, a) in both cases. All other two-node loops have cost 3. All of the counterclockwise loops, e.g., $[a, c, b, a]$, have cost 3 and any of the clockwise loops have cost 1. Thus, the minimum loop cost of this network is $\text{mlc}(X, A_X) = 1$.

For resolutions $0 \leq \delta < \text{mlc}(X, A_X)$ it is impossible to find chains of mutual influence with maximum cost smaller than δ between any pair of points. Indeed, suppose we can link x to x' with a chain of maximum cost smaller than δ , and also link x' to x with a chain having the same property. Then, we can form a loop with cost smaller than δ by concatenating these two chains. Thus, the intuitive notion that clusters cannot form at resolutions for which it is impossible to observe mutual influence can be translated into the requirement that no clusters can be formed at resolutions $\delta < \text{mlc}(X, A_X)$. In terms of ultrametrics, this implies that it must be $u_X(x, x') \geq \text{mlc}(X, A_X)$ for any $x \neq x' \in X$ as we formally state next:

(PI) Property of Influence. For any network $N_X = (X, A_X)$ the ultrametric $(X, u_X) = \mathcal{H}(X, A_X)$ is such that $u_X(x, x')$ for distinct nodes cannot be smaller than the minimum loop cost $\text{mlc}(X, A_X)$ [cf. (5)] of the network, i.e. $u_X(x, x') \geq \text{mlc}(X, A_X)$ for all $x \neq x'$.

Since for the network in Fig. 4 the minimum loop cost is $\text{mlc}(X, A_X) = 1$, then the Property of Influence implies that $u_X(x, x') \geq \text{mlc}(X, A_X) = 1$ for any pair of nodes $x \neq x'$. Equivalently, the output dendrogram is such that for resolutions

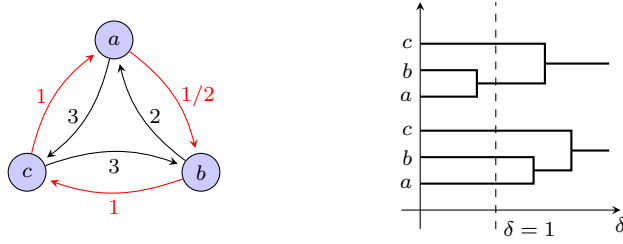


Fig. 4 Property of Influence. No clusters can be generated at resolutions for which it is impossible to form influence loops. Here, the loop of minimum cost is formed by circling the network clockwise where the maximum cost encountered is $A_X(b, c) = A_X(c, a) = 1$. The top dendrogram is invalid because a and b cluster at resolution $\delta < 1$ whereas the bottom dendrogram satisfies the Property of Influence (P1).

$\delta < \text{mlc}(X, A_X) = 1$ each node is in its own block. Observe that (P1) does not imply that a cluster with more than one node is formed at resolution $\delta = \text{mlc}(X, A_X)$ but states that achieving this minimum resolution is a necessary condition for the formation of clusters.

A second intuitive statement about influence in networks of arbitrary size comes in the form of the *Extended Axiom of Value*. To introduce this concept define a family of canonical asymmetric networks $\Delta_n(\alpha, \beta) := (\{1, \dots, n\}, A_{n, \alpha, \beta})$, with $n \in \mathbb{N}$ and $\alpha, \beta > 0$, where the underlying node set $\{1, \dots, n\}$ consists of the first n natural numbers and the dissimilarity value $A_{n, \alpha, \beta}(i, j)$ between points i and j depends on whether $i > j$ or not. For points $i > j$ we let $A_{n, \alpha, \beta}(i, j) = \alpha$ whereas for points $i < j$ we have $A_{n, \alpha, \beta}(i, j) = \beta$. Recall that, by definition, $A_{n, \alpha, \beta}(i, i) = 0$. In the network $\Delta_n(\alpha, \beta)$ all pairs of nodes have dissimilarities α in one direction and β in the other direction. This symmetry entails that all nodes should cluster together at the same resolution, and the requirement of mutual influence along with consistency with the Axiom of Value entails that this resolution should be $\max(\alpha, \beta)$. Before formalizing this definition notice that having clustering outcomes that depend on the ordering of the nodes in the space $\{1, \dots, n\}$ is not desirable. Thus, we consider a permutation $\Pi = \{\pi_1, \pi_2, \dots, \pi_n\}$ of $\{1, 2, \dots, n\}$ and the action $\Pi(A)$ on a dissimilarity function A , which we define by $\Pi(A)(i, j) = A(\pi_i, \pi_j)$ for all i and j . Define now the network $\Delta_n(\alpha, \beta, \Pi) := (\{1, \dots, n\}, \Pi(A_{n, \alpha, \beta}))$. With this definition we can now formally introduce the Extended Axiom of Value as follows:

(A1') *Extended Axiom of Value*. Consider the network $\Delta_n(\alpha, \beta, \Pi) = (\{1, \dots, n\}, \Pi(A_{n, \alpha, \beta}))$. Then, for all indices $n \in \mathbb{N}$, constants $\alpha, \beta > 0$, and permutations Π of $\{1, \dots, n\}$, the outcome $(\{1, \dots, n\}, u) = \mathcal{H}(\Delta_n(\alpha, \beta, \Pi))$ satisfies $u(i, j) = \max(\alpha, \beta)$, for all pairs of nodes $i \neq j$.

Observe that the Axiom of Value (A1) is subsumed into the Extended Axiom of Value for $n = 2$. Further note that the minimum loop cost of $\Delta_n(\alpha, \beta, \Pi)$ is $\max(\alpha, \beta)$. Combining this with the Property of Influence (P1), it follows that for the network $\Delta_n(\alpha, \beta, \Pi)$ we must have $u(i, j) \geq \text{mlc}(\Delta_n(\alpha, \beta)) = \max(\alpha, \beta)$ for $i \neq j$. By the Extended Axiom of Value (A1') we have $u(i, j) = \max(\alpha, \beta)$ for $i \neq j$, which means that (A1') and (P1) are compatible requirements. We can then conceive of alternative axiomatic formulations where admissible methods are required to abide by the Axiom of Transformation (A2), the Property of Influence (P1), and either the (regular) Axiom of Value (A1) or the Extended Axiom of Value (A1') – Axiom (A1)

and (P1) are compatible because (A1) is a particular case of (A1') which we already argued is compatible with (P1). We will see in the following section that these two alternative axiomatic formulations are equivalent to each other in the sense that a clustering method satisfies one set of axioms if and only if it satisfies the other. We further show that (P1) and (A1') are implied by (A1) and (A2). As a consequence, it follows that both alternative axiomatic formulations are equivalent to simply requiring fulfillment of axioms (A1) and (A2).

4.1 Equivalent axiomatic formulations

We begin by stating the equivalence between admissibility with respect to (A1)-(A2) and (A1')-(A2). Furthermore, a theorem stating that methods admissible with respect to (A1) and (A2) satisfy the Property of Influence (P1) is presented next.

Theorem 2 *Assume the hierarchical clustering method \mathcal{H} satisfies the Axiom of Transformation (A2). Then, \mathcal{H} satisfies the Axiom of Value (A1) if and only if it satisfies the Extended Axiom of Value (A1').*

The Extended Axiom of Value (A1') is stronger than the (regular) Axiom of Value (A1). However, Theorem 2 shows that when considered together with the Axiom of Transformation (A2), both axioms of value are equivalent in the restrictions they impose in the set of admissible clustering methods \mathcal{H} . In the following theorem we show that the Property of Influence (P1) can be derived from axioms (A1) and (A2).

Theorem 3 *If a clustering method \mathcal{H} satisfies the Axiom of Value (A1) and the Axiom of Transformation (A2), then it satisfies the Property of Influence (P1).*

The fact that (P1) is implied by (A1) and (A2) as claimed by Theorem 3 implies that adding (P1) as a third axiom on top of these two is moot. In the discussion leading to the introduction of the Axiom of Value (A1) in Section 3 we argued that the intuitive notion of a cluster dictates that it must be possible for co-clustered nodes to influence each other. In the discussion leading to the definition of the Property of Influence (P1) at the beginning of this section we argued that in networks with more than two nodes the natural extension is that co-clustered nodes must be able to influence each other either directly or through their indirect influence on other intermediate nodes. The Property of Influence is a codification of this intuition because it states the impossibility of cluster formation at resolutions where influence loops cannot be formed. While (P1) and (A1) seem quite different and seemingly independent, we have shown in this section that if a method satisfies axioms (A1) and (A2) it must satisfy (P1). Therefore, requiring direct influence on a two-node network as in (A1) restricts the mechanisms for indirect influence propagation so that clusters cannot be formed at resolutions that do not allow for mutual, possibly indirect, influence as stated in (P1). In that sense the restriction of indirect influence propagation in (P1) is not just *intuitively* reasonable but *formally* implied by the more straightforward restrictions on direct influence in (A1) and dissimilarity-reducing maps in (A2).

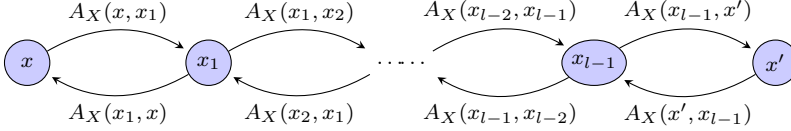


Fig. 5 Reciprocal clustering. Nodes x and x' are clustered at resolution δ if they can be joined with a (reciprocal) chain whose maximum dissimilarity is smaller than or equal to δ in both directions [cf. (15)].

5 Reciprocal and nonreciprocal clustering

Pick any network $N_X = (X, A_X) \in \mathcal{N}$. One particular clustering method satisfying axioms (A1)-(A2) can be constructed by considering the *symmetric* dissimilarity

$$\bar{A}_X(x, x') := \max(A_X(x, x'), A_X(x', x)), \quad (14)$$

for all $x, x' \in X$. This effectively reduces the problem to clustering of symmetric data, a scenario in which single linkage clustering in (4) is known to satisfy axioms analogous to (A1)-(A2) (Carlsson and Mémoli, 2010a). Drawing upon this connection we define the *reciprocal* clustering method \mathcal{H}^R with output $(X, u_X^R) = \mathcal{H}^R(X, A_X)$ as the one for which the ultrametric $u_X^R(x, x')$ between points x and x' is given by

$$u_X^R(x, x') := \min_{C(x, x')} \max_{i | x_i \in C(x, x')} \bar{A}_X(x_i, x_{i+1}). \quad (15)$$

An illustration of the definition in (15) is shown in Fig. 5. We search for chains $C(x, x')$ linking nodes x and x' . For a given chain we walk from x to x' and for every link, connecting say x_i with x_{i+1} , we determine the maximum dissimilarity in both directions, i.e. the value of $\bar{A}_X(x_i, x_{i+1})$. We then determine the maximum across all the links in the chain. The reciprocal ultrametric $u_X^R(x, x')$ between points x and x' is the minimum of this value across all possible chains. Recalling the equivalence of dendrograms and ultrametrics provided by Theorem 1, we know that R_X , the dendrogram produced by reciprocal clustering, clusters x and x' together for resolutions $\delta \geq u_X^R(x, x')$. Combining the latter observation with (15), we can write the reciprocal clustering equivalence classes as

$$x \sim_{R_X(\delta)} x' \iff \min_{C(x, x')} \max_{i | x_i \in C(x, x')} \bar{A}_X(x_i, x_{i+1}) \leq \delta. \quad (16)$$

Comparing (16) with the definition of single linkage in (4) with $\tilde{u}_X^*(x, x')$ as defined in (3), we see that reciprocal clustering is equivalent to single linkage for the symmetrized network $N = (X, \bar{A}_X)$ where dissimilarities between nodes are symmetrized to the maximum value of each directed dissimilarity.

For the method \mathcal{H}^R specified in (15) to be a properly defined hierarchical clustering method, we need to establish that u_X^R is a valid ultrametric. It is clear that $u_X^R(x, x') = 0$ only if $x = x'$ and that $u_X^R(x, x') = u_X^R(x', x)$ because the definition is symmetric on x and x' . To verify that the strong triangle inequality in (7) holds, let $C^*(x, x')$ and $C^*(x', x'')$ be chains that achieve the minimum in (15) for $u_X^R(x, x')$ and $u_X^R(x', x'')$, respectively. The maximum cost in the concatenated chain $C(x, x'') = C^*(x, x') \cup C^*(x', x'')$ does not exceed the maximum cost in each individual chain. Thus, while the cost may be smaller on a different chain, the chain

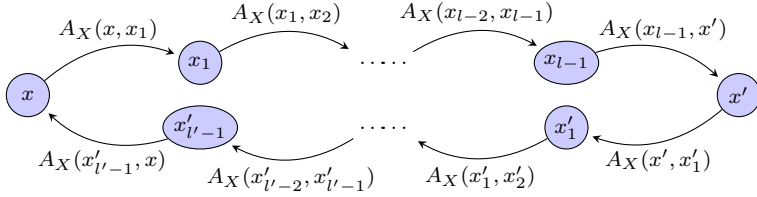


Fig. 6 Nonreciprocal clustering. Nodes x and x' are clustered at resolution δ if they can be joined in both directions with possibly different chains of maximum dissimilarity not greater than δ [cf. (19)].

$C(x, x'')$ suffices to bound $u_X^R(x, x'') \leq \max(u_X^R(x, x'), u_X^R(x', x''))$ as in (7). It is also possible to prove that \mathcal{H}^R satisfies axioms (A1)-(A2), as we do next.

Proposition 3 *The reciprocal clustering method \mathcal{H}^R is valid and admissible. I.e., u_X^R in (15) is an ultrametric for all networks and \mathcal{H}^R satisfies axioms (A1)-(A2).*

Proof: That u_X^R conforms to the definition of an ultrametric was proved in the paragraph preceding this proposition. To see that the Axiom of Value (A1) is satisfied, pick an arbitrary two-node network $\Delta_2(\alpha, \beta)$ as defined in Section 2 and denote by $(\{p, q\}, u_{p,q}^R) = \mathcal{H}^R(\Delta_2(\alpha, \beta))$ the output of applying the reciprocal clustering method to $\Delta_2(\alpha, \beta)$. Since every possible chain from p to q must contain p and q as consecutive nodes, applying the definition in (15) yields $u_{p,q}^R(p, q) = \max(A_{p,q}(p, q), A_{p,q}(q, p)) = \max(\alpha, \beta)$. Axiom (A1) is thereby satisfied.

To show fulfillment of Axiom (A2), consider two networks (X, A_X) and (Y, A_Y) , a dissimilarity-reducing map $\phi : X \rightarrow Y$ and define $(X, u_X^R) := \mathcal{H}^R(X, A_X)$ and $(Y, u_Y^R) := \mathcal{H}^R(Y, A_Y)$. For an arbitrary pair of nodes $x, x' \in X$, denote by $C_X^*(x, x') = [x = x_0, \dots, x_l = x']$ a chain that achieves the minimum reciprocal cost in (15) so as to write $u_X^R(x, x') = \max_{i|x_i \in C_X^*(x, x')} \bar{A}_X(x_i, x_{i+1})$. Consider the transformed chain $C_Y(\phi(x), \phi(x')) = [\phi(x) = \phi(x_0), \dots, \phi(x_l) = \phi(x')]$ in the set Y . Since the transformation ϕ does not increase dissimilarities we have that for all links in this chain $A_Y(\phi(x_i), \phi(x_{i+1})) \leq A_X(x_i, x_{i+1})$ and $A_Y(\phi(x_{i+1}), \phi(x_i)) \leq A_X(x_{i+1}, x_i)$. This implies that

$$\max_{i|\phi(x_i) \in C_Y(\phi(x), \phi(x'))} \bar{A}_Y(\phi(x_i), \phi(x_{i+1})) \leq u_X^R(x, x'). \quad (17)$$

Further note that $C_Y(\phi(x), \phi(x'))$ is a particular chain joining $\phi(x)$ and $\phi(x')$ whereas the reciprocal ultrametric is the minimum across all such chains. Therefore,

$$u_Y^R(\phi(x), \phi(x')) \leq \max_{i|\phi(x_i) \in C_Y(\phi(x), \phi(x'))} \bar{A}_Y(\phi(x_i), \phi(x_{i+1})). \quad (18)$$

Substituting (17) in (18), the fulfillment of Axiom (A2) follows. ■

In reciprocal clustering, nodes x and x' belong to the same cluster at a resolution δ whenever we can go back and forth from x to x' at a maximum cost δ through the same chain. By contrast, in *nonreciprocal* clustering we relax the restriction about the chain being the same in both directions and cluster nodes x and x' together if there are chains, possibly different, linking x to x' and x' to x . To state this definition in terms of ultrametrics consider a given network $N = (X, A_X)$ and recall the definition

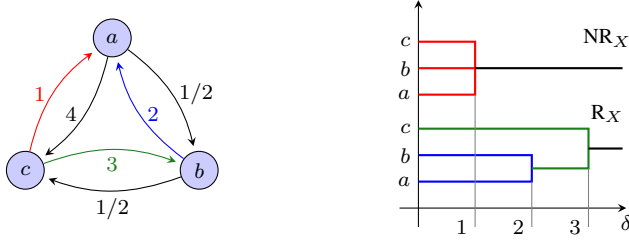


Fig. 7 Reciprocal and nonreciprocal dendrograms. An example network with its corresponding reciprocal (bottom) and nonreciprocal (top) dendrograms.

of the unidirectional minimum chain cost \tilde{u}_X^* in (3). We define the nonreciprocal clustering method \mathcal{H}^{NR} with output $(X, u_X^{\text{NR}}) = \mathcal{H}^{\text{NR}}(X, A_X)$ as the one for which the ultrametric $u_X^{\text{NR}}(x, x')$ between points x and x' is given by the maximum of the unidirectional minimum chain costs $\tilde{u}_X^*(x, x')$ and $\tilde{u}_X^*(x', x)$ in each direction,

$$u_X^{\text{NR}}(x, x') := \max \left(\tilde{u}_X^*(x, x'), \tilde{u}_X^*(x', x) \right). \quad (19)$$

An illustration of the definition in (19) is shown in Fig. 6. We consider forward chains $C(x, x')$ going from x to x' and backward chains $C(x', x)$ going from x' to x . For each of these chains we determine the maximum dissimilarity across all the links in the chain. We then search independently for the best forward chain $C(x, x')$ and the best backward chain $C(x', x)$ that minimize the respective maximum dissimilarities across all possible chains. The nonreciprocal ultrametric $u_X^{\text{NR}}(x, x')$ between points x and x' is the maximum of these two minimum values.

As it is the case with reciprocal clustering we can verify that u_X^{NR} is a properly defined ultrametric and that, as a consequence, the nonreciprocal clustering method \mathcal{H}^{NR} is properly defined. Identity and symmetry are immediate. For the strong triangle inequality consider chains $C^*(x, x')$ and $C^*(x', x'')$ that achieve the minimum costs in $\tilde{u}_X^*(x, x')$ and $\tilde{u}_X^*(x', x'')$ as well as the chains $C^*(x'', x')$ and $C^*(x', x)$ that achieve the minimum costs in $\tilde{u}_X^*(x'', x')$ and $\tilde{u}_X^*(x', x)$. The concatenation of these chains permits concluding that $u_X^{\text{NR}}(x, x'') \leq \max(u_X^{\text{NR}}(x, x'), u_X^{\text{NR}}(x', x''))$, which is the strong triangle inequality in (7). The method \mathcal{H}^{NR} also satisfies axioms (A1)-(A2) as the following proposition states.

Proposition 4 *The nonreciprocal clustering method \mathcal{H}^{NR} is valid and admissible. I.e., u_X^{NR} in (19) is an ultrametric for all networks and \mathcal{H}^{NR} satisfies axioms (A1)-(A2).*

Proof: That \mathcal{H}^{NR} outputs valid ultrametrics was already argued prior to the statement of Proposition 4. The proof for admissibility of \mathcal{H}^{NR} is omitted since it is analogous to that of admissibility of \mathcal{H}^{R} (cf. Theorem 3). ■

The reciprocal and nonreciprocal dendrograms for an example network are shown in Fig. 7. Notice that these dendrograms are *different*. In the reciprocal dendrogram nodes a and b cluster together at resolution $\delta = 2$ due to their direct connections $A_X(a, b) = 1/2$ and $A_X(b, a) = 2$. Node c joins this cluster at resolution $\delta = 3$ because it links bidirectionally with b through the chain $[b, c]$ whose maximum cost is $A_X(c, b) = 3$. The optimal reciprocal chain linking a and c is $[a, b, c]$ whose maximum cost is also $A_X(c, b) = 3$. In the nonreciprocal dendrogram we can link nodes

with different chains in each direction. As a consequence, a and b cluster together at resolution $\delta = 1$ because the directed cost of the chain $[a, b]$ is $A_X(a, b) = 1/2$ and the directed cost of the chain $[b, c, a]$ is $A_X(c, a) = 1$. Similar chains demonstrate that a and c as well as b and c also cluster together at resolution $\delta = 1$.

In more generality, we now discuss an application scenario and illustrate that \mathcal{H}^R and \mathcal{H}^{NR} are suited to unveil different aspects of the underlying data. Consider a social network (X, A_X) consisting of n agents $X = \{x_1, \dots, x_n\}$ moving inside a certain fixed environment from say time 0 to time T (Smith et al, 2016). Assume that when agents come within close proximity of each other they establish a hand-shaking protocol and exchange identity information. E.g., if our agents are actually people meeting at a party, then they may exchange phone numbers. Nevertheless, the exchange is done in a probabilistic fashion: they always exchange information but one of the two agents involved might forget it right away. These possible lapses of memory indicate that the process is inherently asymmetric. The possibly asymmetric weights $A_X(x_i, x_j)$ that we associate to this network record the first time that agent x_i interacts with and successfully records the contact details of agent x_j . We assume that $A_X(x_i, x_i) = 0$ for all i . Some questions that arise in this scenario are related to the ability of the network to propagate information. Specifically, two questions can be naturally posed in this situation: (1) What is the first time when it is possible to relay messages between any pair of agents in the network? (2) In a communication context, what is the first time that a two way full-duplex communication channel (Choi et al, 2010) can be established between any pair of agents? It follows from the above discussion that applying \mathcal{H}^{NR} to (X, A_X) will answer question (1), whereas, in order to answer question (2) one must apply \mathcal{H}^R .

Remark 1 Let $(X, A_X) \in \mathcal{N}$ where X contains n nodes. As pointed out after (16), in order to compute u_X^R one needs to apply standard single linkage hierarchical clustering to the symmetrized network (X, \bar{A}_X) . This means that for the case of reciprocal clustering, complexity of order $O(n^2)$ can be achieved by leveraging the well-known equivalence between single linkage clustering and the search of a minimum spanning tree (Hu, 1961; Müllner, 2011). For the case of nonreciprocal clustering, Tarjan’s method (Tarjan, 1983) can be directly applied with complexity $O(n^2 \log n)$.¹

6 Extremal ultrametrics

Given that we have constructed two admissible methods satisfying axioms (A1)-(A2), the question whether these two constructions are the only possible ones arises and, if not, whether they are special in some sense. We prove in this section that reciprocal and nonreciprocal clustering are a peculiar pair in that all possible admissible clustering methods are contained between them in a well-defined sense. To explain this sense properly, observe that since reciprocal chains [cf. Fig. 5] are particular cases of nonreciprocal chains [cf. Fig. 6] we must have that $u_X^{NR}(x, x') \leq u_X^R(x, x')$ for all pairs of nodes x, x' . I.e., nonreciprocal ultrametrics do not exceed reciprocal ultrametrics. An important characterization is that any method \mathcal{H} satisfying axioms (A1)-(A2) yields ultrametrics that lie between u_X^{NR} and u_X^R as we formally state next.

¹ MATLAB implementations of the clustering methods here discussed can be downloaded at http://www.mit.edu/~segarra/clustering_methods_ADAC.zip

Theorem 4 Consider an admissible clustering method \mathcal{H} satisfying axioms (A1)-(A2). For an arbitrary given network $N = (X, A_X)$ denote by $(X, u_X) = \mathcal{H}(N)$ the output of \mathcal{H} applied to N . Then, for all pairs of nodes $x, x' \in X$

$$u_X^{\text{NR}}(x, x') \leq u_X(x, x') \leq u_X^{\text{R}}(x, x'), \quad (20)$$

where u_X^{NR} and u_X^{R} denote the nonreciprocal and reciprocal ultrametrics as defined by (19) and (15), respectively.

According to Theorem 4, nonreciprocal clustering applied to a given network $N = (X, A_X)$ yields a uniformly minimal ultrametric among those output by all clustering methods satisfying axioms (A1)-(A2). Reciprocal clustering yields a uniformly maximal ultrametric. Any other clustering method abiding by (A1)-(A2) yields an ultrametric such that the value $u_X(x, x')$ for any two points $x, x' \in X$ lies between the values $u_X^{\text{NR}}(x, x')$ and $u_X^{\text{R}}(x, x')$ assigned by nonreciprocal and reciprocal clustering. In terms of dendrograms, (20) implies that among all possible clustering methods, the smallest possible resolution at which nodes are clustered together is the one corresponding to nonreciprocal clustering. The highest possible resolution is the one that corresponds to reciprocal clustering.

6.1 Hierarchical clustering on symmetric networks

Restrict attention to the subspace $\mathcal{M} \subset \mathcal{N}$ of symmetric networks, that is $N = (X, A_X) \in \mathcal{M}$ if and only if $A_X(x, x') = A_X(x', x)$ for all $x, x' \in X$. When restricted to the space \mathcal{M} reciprocal and nonreciprocal clustering are equivalent methods because, for any pair of points, minimizing nonreciprocal chains are always reciprocal – more precisely there may be multiple minimizing nonreciprocal chains but at least one of them is reciprocal. To see this formally, first fix $x, x' \in X$ and observe that in symmetric networks the symmetrization in (14) is unnecessary because $\bar{A}_X(x_i, x_{i+1}) = A_X(x_i, x_{i+1}) = A_X(x_{i+1}, x_i)$ and the definition of reciprocal clustering in (15) reduces to

$$u_X^{\text{R}}(x, x') = \min_{C(x, x')} \max_{i | x_i \in C(x, x')} A_X(x_i, x_{i+1}) = \min_{C(x', x)} \max_{i | x_i \in C(x', x)} A_X(x_i, x_{i+1}). \quad (21)$$

Further note that the costs of any given chain $C(x, x') = [x = x_0, x_1, \dots, x_{l-1}, x_l = x']$ and its reciprocal $C(x', x) = [x' = x_l, x_{l-1}, \dots, x_1, x_0 = x]$ are the same. It follows that directed minimum chain costs $\tilde{u}_X^*(x, x') = \tilde{u}_X^*(x', x)$ are equal and according to (19) equal to the nonreciprocal ultrametric

$$u_X^{\text{NR}}(x, x') = \tilde{u}_X^*(x, x') = \tilde{u}_X^*(x', x) = u_X^{\text{R}}(x, x'). \quad (22)$$

To write the last equality in (22) we used the definitions of $\tilde{u}_X^*(x, x')$ and $\tilde{u}_X^*(x', x)$ in (3) which are correspondingly equivalent to the first and second equality in (21).

By further comparison of the ultrametric definition of single linkage in (10) with (22) the equivalence of reciprocal, nonreciprocal, and single linkage clustering in symmetric networks follows

$$u_X^{\text{NR}}(x, x') = u_X^{\text{SL}}(x, x') = u_X^{\text{R}}(x, x'). \quad (23)$$

The equivalence in (22) along with Theorem 4 demonstrates that when considering the application of hierarchical clustering methods $\mathcal{H} : \mathcal{M} \rightarrow \mathcal{U}$ to symmetric networks, there exist a unique method satisfying (A1)-(A2). The equivalence in (23) shows that this method is single linkage. Before stating this result formally let us define the symmetric version of the Axiom of Value:

(B1) *Symmetric Axiom of Value.* Consider a symmetric two-node network $\Delta_2(\alpha, \alpha)$. The ultrametric $(\{p, q\}, u_{p,q}) = \mathcal{H}(\Delta_2(\alpha, \alpha))$ satisfies $u_{p,q}(p, q) = \alpha$.

Since there is only one dissimilarity in a symmetric network with two nodes, (B1) states that they cluster together at the resolution that connects them to each other. We can now prove that single linkage is the unique hierarchical clustering method in symmetric networks that is admissible with respect to (B1) and (A2).

Corollary 1 *Let $\mathcal{H} : \mathcal{M} \rightarrow \mathcal{U}$ be a hierarchical clustering method for symmetric networks and \mathcal{H}^{SL} be the single linkage method with output ultrametrics as defined in (10). If \mathcal{H} satisfies axioms (B1) and (A2) then $\mathcal{H} \equiv \mathcal{H}^{\text{SL}}$.*

Proof: When restricted to symmetric networks, (B1) and (A1) are equivalent statements. Thus, \mathcal{H} satisfies the hypotheses of Theorem 4 and, as a consequence, (20) is true for any pair of points x, x' of any network $N \in \mathcal{M}$. But by (23) nonreciprocal, single linkage, and reciprocal ultrametrics coincide. Thus, we can reduce (20) to $u_X^{\text{SL}}(x, x') \leq u_X(x, x') \leq u_X^{\text{SL}}(x, x')$, implying that $\mathcal{H} \equiv \mathcal{H}^{\text{SL}}$. ■

The uniqueness result claimed by Corollary 1 strengthens the uniqueness result by Carlsson and Mémoli (2010a, Theorem 18). To explain the differences consider the symmetric version of the Property of Influence. In a symmetric network there is always a loop of minimum cost of the form $[x, x', x]$ for some pair of points x, x' . Indeed, say that $C^*(x^*, x^*)$ is one of the loops achieving the minimum cost in (5) and let $A_X(x, x') = \text{mlc}(X, A_X)$ be the maximum dissimilarity in this loop. Then, the cost of the loop $[x, x', x]$ is $A_X(x, x') = A_X(x', x) = \text{mlc}(X, A_X)$ which means that either the loop $C^*(x^*, x^*)$ was already of the form $[x, x', x]$ or that the cost of the loop $[x, x', x]$ is the same as $C^*(x^*, x^*)$. In any event, there is a loop of minimum cost of the form $[x, x', x]$ which implies that in symmetric networks we must have

$$\text{mlc}(X, A_X) = \min_{x \neq x'} A_X(x, x') = \text{sep}(X, A_X), \quad (24)$$

We can introduce the symmetric version of the Property of Influence:

(Q1) *Symmetric Property of Influence.* For any symmetric network $N_X = (X, A_X)$ the output $(X, u_X) = \mathcal{H}(N_X)$ is such that $u_X(x, x')$ for distinct points cannot be smaller than the network separation, i.e. $u_X(x, x') \geq \text{sep}(N_X)$ for all $x \neq x'$.

Carlsson and Mémoli (2010a) define admissibility with respect to (B1), (A2), and (Q1), which corresponds to conditions (I), (II), and (III) of their Theorem 18. Corollary 1 shows that Property (Q1) is redundant when given axioms (B1) and (A2) – respectively, Condition (III) by Carlsson and Mémoli (2010a, Theorem 18) is redundant when given conditions (I) and (II). Corollary 1 also shows that single linkage is the unique admissible method for all symmetric, not necessarily metric, networks.

7 Alternative axiomatic constructions

The axiomatic framework that we adopted allows alternative constructions by modifying the underlying set of axioms. Among the axioms in Section 3, the Axiom of Value (A1) is perhaps the most open to interpretation. Although we required the two-node network in Fig. 2 to first cluster into one single block at resolution $\max(\alpha, \beta)$ corresponding to the largest dissimilarity and argued that this was reasonable in most situations, it is also reasonable to accept that in some situations the two nodes should be clustered together as long as one of them is able to influence the other. To account for this possibility we replace the Axiom of Value by the following alternative.

(A1'') *Alternative Axiom of Value.* The ultrametric $(\{p, q\}, u_{p,q}) := \mathcal{H}(\Delta_2(\alpha, \beta))$ output by \mathcal{H} from the two-node network $\Delta_2(\alpha, \beta)$ satisfies $u_{p,q}(p, q) = \min(\alpha, \beta)$.

Axiom (A1'') replaces the requirement of bidirectional influence in Axiom (A1) to unidirectional influence; see Fig. 8. We say that a clustering method \mathcal{H} is admissible with respect to the alternative axioms if it satisfies axioms (A1'') and (A2).

The property of influence (P1), which is a keystone in the proof of Theorem 4, is not compatible with the Alternative Axiom of Value (A1''). Indeed, just observe that the minimum loop cost of the two-node network in Fig. 8 is $\text{mlc}(\Delta_2(\alpha, \beta)) = \max(\alpha, \beta)$ whereas in (A1'') we are requiring the output ultrametric to be $u_{p,q}(p, q) = \min(\alpha, \beta)$. We therefore have that Axiom (A1'') itself implies $u_{p,q}(p, q) = \min(\alpha, \beta) < \max(\alpha, \beta) = \text{mlc}(\Delta_2(\alpha, \beta))$ for the cases when $\alpha \neq \beta$. Thus, we reformulate (P1) into the Alternative Property of Influence (P1') that we define next.

(P1') *Alternative Property of Influence.* For any network $N_X = (X, A_X)$ the output ultrametric $(X, u_X) = \mathcal{H}(N_X)$ is such that $u_X(x, x')$ for distinct points cannot be smaller than the separation of the network, $u_X(x, x') \geq \text{sep}(N_X)$ for all $x \neq x'$.

Observe that the Alternative Property of Influence (P1') coincides with the Symmetric Property of Influence (Q1) defined in Section 6.1. This is not surprising because for symmetric networks the Axiom of Value (A1) and the Alternative Axiom of Value (A1'') impose identical restrictions. Moreover, since the separation of a network cannot be larger than its minimum loop cost, the Alternative Property of Influence (P1') is implied by the (regular) Property of Influence (P1), but not vice versa.

The Alternative Property of Influence (P1') states that no clusters are formed at resolutions at which there are no unidirectional influences between any pair of nodes and is consistent with the Alternative Axiom of Value (A1''). Moreover, in studying methods admissible with respect to (A1'') and (A2), (P1') plays a role akin to the one played by (P1) when studying methods that are admissible with respect to (A1) and (A2). In particular, as (P1) is implied by (A1) and (A2), (P1') is true if (A1'') and (A2) hold as we assert in the following theorem.

Theorem 5 *If a clustering method \mathcal{H} satisfies the Alternative Axiom of Value (A1'') and the Axiom of Transformation (A2) then it also satisfies the Alternative Property of Influence (P1').*

Theorem 5 admits the following interpretation. In (A1'') we require two-node networks to cluster at the resolution where unidirectional influence occurs. When we consider (A1'') in conjunction with (A2) we can translate this requirement into a

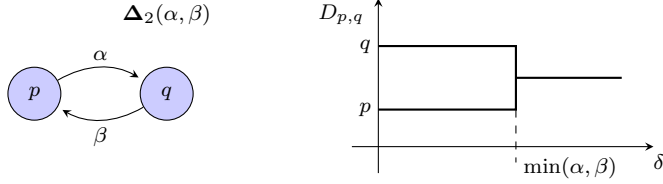


Fig. 8 Alternative Axiom of Value. For a two-node network, nodes are clustered together at the minimum resolution at which one of them can influence the other.

statement about clustering in arbitrary networks. Such requirement is the Alternative Property of Influence (PI') which prevents nodes to cluster at resolutions at which no influence exists between *any* two nodes.

7.1 Unilateral clustering

Mimicking the developments in Sections 3-6, we move on to identify and define methods that satisfy axioms (A1'')-(A2) and then bound the range of admissible methods with respect to these axioms. To do so, let $N = (X, A_X)$ be a given network and consider the dissimilarity function $\hat{A}_X(x, x') := \min(A_X(x, x'), A_X(x', x))$, for all $x, x' \in X$. Notice that, as opposed to the definition of \bar{A}_X , where the symmetrization is done by means of a max operation, \hat{A} is defined by using a min operation. We define the *unilateral* clustering method \mathcal{H}^U with output ultrametric $(X, u_X^U) = \mathcal{H}^U(N)$, where u_X^U is defined as

$$u_X^U(x, x') := \min_{C(x, x')} \max_{i | x_i \in C(x, x')} \hat{A}_X(x_i, x_{i+1}), \quad (25)$$

for all $x, x' \in X$. To show that \mathcal{H}^U is a properly defined clustering method, we need to establish that u_X^U as defined in (25) is a valid ultrametric. However, comparing (25) and (10) we see that $\mathcal{H}^U(X, A_X) \equiv \mathcal{H}^{SL}(X, \hat{A}_X)$, i.e. applying the unilateral clustering method to an asymmetric network (X, A_X) is equivalent to applying single linkage clustering method to the symmetrized network (X, \hat{A}_X) . Since we know that single linkage produces a valid ultrametric when applied to any symmetric network such as (X, \hat{A}_X) , (25) is a properly defined ultrametric. Furthermore, it can be shown that \mathcal{H}^U satisfies axioms (A1'') and (A2).

Proposition 5 *The unilateral clustering method \mathcal{H}^U with output ultrametrics defined in (25) satisfies axioms (A1'') and (A2).*

In the case of admissibility with respect to (A1) and (A2), nonreciprocal and reciprocal clustering are two different admissible methods which bound every other possible clustering method satisfying (A1)-(A2) (cf. Theorem 4). In contrast, in the case of admissibility with respect to (A1'') and (A2), unilateral clustering is the *unique* admissible method as stated in the following theorem.

Theorem 6 *Let \mathcal{H} be a hierarchical clustering method satisfying axioms (A1'') and (A2). Then, $\mathcal{H} \equiv \mathcal{H}^U$ where \mathcal{H}^U is the unilateral clustering.*

By Theorem 6, the space of methods that satisfy the Alternative Axiom of Value (A1'') and the Axiom of Transformation (A2) is inherently simpler than the space of methods that satisfy the (regular) Axiom of value (A1) and the Axiom of Transformation (A2). Further note that in the case of symmetric networks, for all $x, x' \in X$ we have $\hat{A}_X(x, x') = A_X(x, x') = A_X(x', x)$ and as a consequence unilateral clustering is equivalent to single linkage as it follows from comparison of (10) and (25). Thus, the result in Theorem 6 reduces to the statement in Corollary 1, which was derived upon observing that in symmetric networks reciprocal and nonreciprocal clustering yield identical outcomes. The fact that reciprocal, nonreciprocal, and unilateral clustering all coalesce into single linkage when restricted to symmetric networks is consistent with the fact that the Axiom of Value (A1) and the Alternative Axiom of Value (A1'') are both equivalent to the Symmetric Axiom of Value (B1) when restricted to symmetric dissimilarities.

7.2 Agnostic Axiom of Value

Axiom (A1) stipulates that every two-node network $\Delta_2(\alpha, \beta)$ is clustered into a single block at resolution $\max(\alpha, \beta)$, whereas Axiom (A1'') stipulates that they should be clustered at $\min(\alpha, \beta)$. One can also be agnostic with respect to this issue and say that both of these situations are admissible. An agnostic version of axioms (A1) and (A1'') is given next.

(A1''') *Agnostic Axiom of Value.* The ultrametric $(X, u_{p,q}) = \mathcal{H}(\Delta_2(\alpha, \beta))$ produced by \mathcal{H} applied to the two-node network $\Delta_2(\alpha, \beta)$ satisfies $\min(\alpha, \beta) \leq u_X(p, q) \leq \max(\alpha, \beta)$.

Since fulfillment of (A1) or (A1'') implies fulfillment of (A1'''), any admissible clustering method with respect to the original axioms (A1)-(A2) or with respect to the alternative axioms (A1''')-(A2) must be admissible with respect to the agnostic axioms (A1''')-(A2). In this sense, (A1''')-(A2) is the most general combination of axioms described in this paper. For methods that are admissible with respect to (A1''') and (A2) we can bound the range of outcome ultrametrics as stated next.

Theorem 7 *Consider a clustering method \mathcal{H} satisfying axioms (A1''') and (A2). For an arbitrary given network $N = (X, A_X)$ denote by $(X, u_X) = \mathcal{H}(X, A_X)$ the outcome of \mathcal{H} applied to N . Then, for all pairs of nodes $x, x' \in X$*

$$u_X^U(x, x') \leq u_X(x, x') \leq u_X^R(x, x'), \quad (26)$$

where $u_X^U(x, x')$ and $u_X^R(x, x')$ denote the unilateral and reciprocal ultrametrics as defined by (25) and (15), respectively.

By Theorem 7, given an asymmetric network (X, A_X) , any hierarchical clustering method abiding by axioms (A1''') and (A2) produces outputs contained between those corresponding to two methods. The first method, unilateral clustering, symmetrizes A_X by calculating $\hat{A}_X(x, x') = \min(A_X(x, x'), A_X(x', x))$ for all $x, x' \in X$ and computes single linkage on (X, \hat{A}_X) . The other method, reciprocal clustering, symmetrizes A_X by calculating $\bar{A}_X(x, x') = \max(A_X(x, x'), A_X(x', x))$ for all $x, x' \in X$ and computes single linkage on (X, \bar{A}_X) .

Remark 2 Consider a network (X, A_X) with n nodes. From the discussion after (25) it follows that unilateral clustering arises from applying standard single linkage hierarchical clustering to the symmetrized network (X, \hat{A}_X) . As explained in Remark 1, it then follows that unilateral clustering can be computed with complexity $O(n^2)$ by leveraging the equivalence between single linkage and the minimum spanning tree problem (Hu, 1961; Müllner, 2011).²

8 An illustrative example

We apply the hierarchical clustering methods developed throughout the paper to analyze the internal migration network between states of the United States (U.S.) for year 2011. The number of migrants from state to state, including the District of Columbia (DC) as a separate entity, is published yearly by the geographical mobility section of the U.S. census bureau³. We denote by S , with cardinality $|S| = 51$, the set containing every state plus DC and by $M : S \times S \rightarrow \mathbb{R}_+ \cup \{+\infty\}$ the migration flow similarity function given by the U.S. census bureau in which $M(s, s')$ is the number of individuals that migrated from state s to state s' . We then construct the asymmetric network $N_S = (S, A_S)$ with dissimilarities A_S such that $A_S(s, s) = 0$ for all $s \in S$ and $A_S(s, s') := 1 - M(s, s') / \sum_{t \neq s'} M(t, s')$ for all $s \neq s' \in S$. The normalization $M(s, s') / \sum_{t \neq s'} M(t, s')$ can be interpreted as the probability that an immigrant to state s' comes from state s . Dissimilarities $A_S(s, s')$ focus attention on the composition of migration flows rather than on their magnitude. A small dissimilarity from state s to state s' implies that from all the immigrants into s' a high percentage comes from s . E.g., if 85% of the immigration into s' comes from s , then $A_S(s, s') = 1 - 0.85 = 0.15$. The application of hierarchical clustering to migration data has been extensively investigated by Slater; see (Slater, 1976, 1984).

8.1 Reciprocal clustering \mathcal{H}^R

The dendrogram obtained from applying the reciprocal clustering method \mathcal{H}^R defined in (15) to the migration network N_S is shown in Fig. 9-(a); see Remark 1. Figures 9-(b) through 9-(e) illustrate the partitions that are obtained at four representative resolutions $\delta_1^R = 0.895$, $\delta_2^R = 0.921$, $\delta_3^R = 0.933$, and $\delta_4^R = 0.947$. States marked with the same color other than white are co-clustered at the given resolution whereas states in white are singleton clusters. For a given δ , states that are clustered together in partitions produced by \mathcal{H}^R are those connected by a chain of intense bidirectional migration flows in the sense dictated by the resolution under consideration.

Reciprocal clusters form between states that have strong migration flows in both directions. This results in a clustering pattern dominated by geographical proximity, with the tightest clusters formed between states that share metropolitan areas.

² See footnote 1.

³ Available at <https://www.census.gov/data/tables/time-series/demo/geographic-mobility/state-to-state-migration.html>.

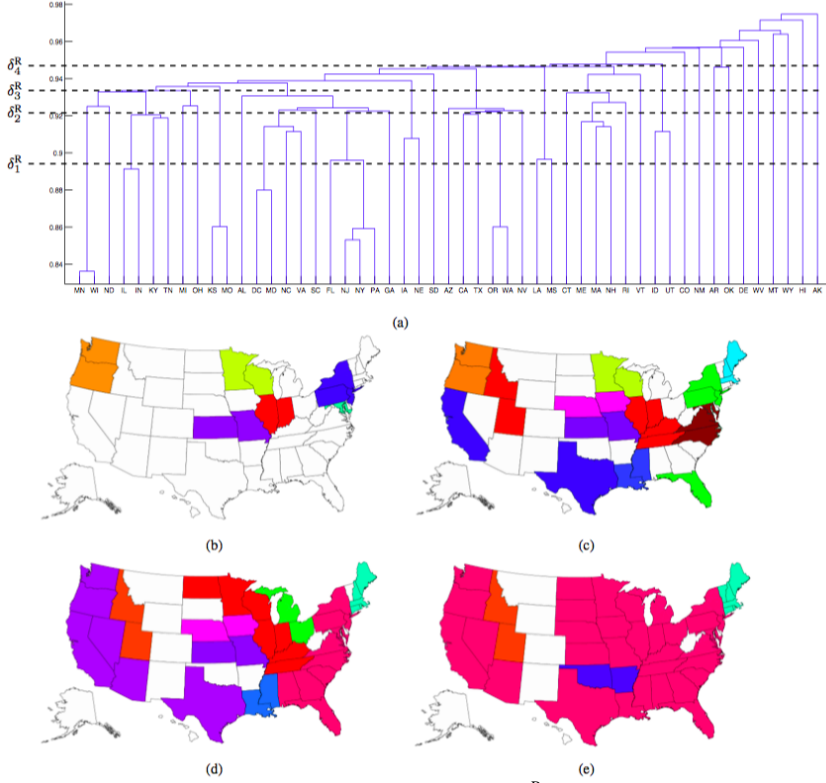


Fig. 9 (a) Reciprocal dendrogram. Output of clustering method \mathcal{H}^R when applied to the migration network N_S . (b) Clusters at resolution δ_1^R . States that share urban metropolitan areas merge together first. States in white form singleton clusters at this resolution. (c) Clusters at resolution δ_2^R . Clusters are highly determined by geographical proximity except for Texas and Florida. (d) Clusters at resolution δ_3^R . The two coasts form separate clusters. (e) Clusters at resolution δ_4^R . Most of the nation forms a single cluster. Observe New England's relative isolation.

With the exceptions of California, Florida, and Texas that we discuss below, all states merge into clusters with other neighboring states. In particular, the first non-singleton clusters to form are pairs of neighboring states that join together at resolutions smaller than δ_1^R as shown in Fig. 9-(b). The formation of these clusters can be explained by the fact that these states share respective metropolitan areas. For example, Minneapolis and Duluth are shared by Minnesota and Wisconsin, and Portland is shared by Oregon and Washington. Even while crossing state lines, migration within shared metropolitan areas corresponds to people moving to different neighborhoods or suburbs and occurs frequently enough to suggest it is the reason behind the clusters formed at low resolutions in the reciprocal dendrogram.

As we continue to increase the resolution, the only two clustering exceptions to geographic proximity appear at δ_2^R , shown in Fig. 9-(c). These exceptions are the merging of Florida into the northeastern cluster formed by New Jersey, Pennsylvania, and New York as well as the formation of a cluster made of California and Texas. This anomaly occurs among the four states with the most intense outgoing and incoming migration in the country during 2011. The data analyzed shows that people move

from all over the United States to New York, California, Texas, and Florida. Hence, the proportion of incoming migration from neighboring states is not as significant as for other states. Thus, these four states have a strong influence on the immigration into their neighboring states but, given the mechanics of \mathcal{H}^R , the lack of influence in the opposite direction is the reason why Texas joins California and Florida joins New York before forming a cluster with their neighbors. If we require only unidirectional influence as in Section 8.3, then these four states first join their neighboring states as observed in Fig. 11.

Higher resolutions see the appearance of three regional clusters in the Atlantic Coast, Midwest, and New England, as well as a cluster composed of the West Coast states plus Texas; see Fig. 9-(d). The New England cluster illustrated in Fig. 9-(e) shows a remarkable degree of migrational isolation with respect to the rest of the country. This indicates that people living in New England tend to move within the region, that people outside New England rarely move into the area, or both.

8.2 Nonreciprocal clustering \mathcal{H}^{NR}

The outcome of applying the nonreciprocal clustering method \mathcal{H}^{NR} defined in (19) to N_S is shown in Fig. 10; see Remark 1. Comparing the reciprocal and nonreciprocal dendrograms in figs. 9-(a) and 10 shows that the nonreciprocal clustering method merges any pair of states into a common cluster at a resolution not higher than the resolution at which they are co-clustered by reciprocal clustering. This is as it should be because the uniform dominance of nonreciprocal ultrametrics by reciprocal ultrametrics holds for all networks [cf. (20)]. E.g., for the reciprocal method, Colorado and Florida become part of the same cluster at resolution $\delta = 0.954$ whereas for the nonreciprocal case they become part of the same cluster at resolution $\delta = 0.939$.

Nonreciprocal clustering detects migration cycles. However, the overall similarity between the outputs of reciprocal and nonreciprocal clustering shows that migration cycles are rare. This similarity also limits the possible output of any other method satisfying the axioms of value and transformation (cf. Theorem 4).

There are many striking similarities between the reciprocal and nonreciprocal dendrograms in figs. 9-(a) and 10. In both dendrograms, the first three clusters to emerge are the pair Minnesota and Wisconsin, followed by the pair New York and New Jersey which are in turn co-clustered with Pennsylvania. Indeed, the first seven groupings coincide with those in the reciprocal dendrogram that we attributed to shared metropolitan areas [cf. Fig. 9-(b)].

For the migration network N_S , nonreciprocal clustering may be able to detect migration cycles of arbitrary length that are overlooked by reciprocal clustering. E.g., if people in state A tend to move predominantly to B , people in B move predominantly to C , and people in C move predominantly to A , nonreciprocal clustering merges these three states according to this migration cycle but reciprocal clustering does not. The overall similarity of the reciprocal and nonreciprocal dendrograms in figs. 9-(a) and 10 suggests that migration cycles are rare in the United States. Notice that highly symmetric data would also correspond to similar reciprocal and nonreciprocal dendrograms. Nevertheless, another consequence of highly symmetric data

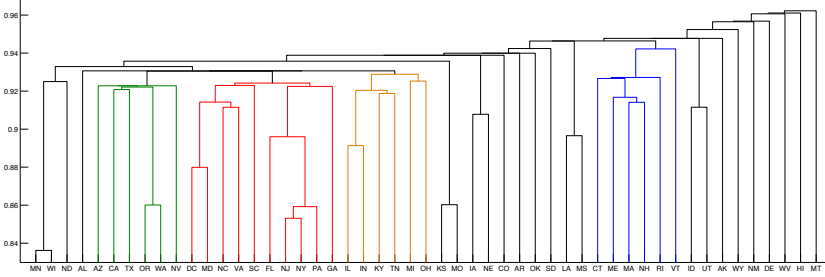


Fig. 10 Nonreciprocal dendrogram. Dendrogram obtained when applying the nonreciprocal method \mathcal{H}^{NR} to the state-to-state migration network N_S . The resemblance with the dendrogram in Fig. 9-(a) indicates that migration cycles are not ubiquitous.

would be to obtain a similar unilateral dendrogram. This is not the case, as can be seen in Section 8.3.

However similar, the reciprocal and nonreciprocal dendrograms in figs. 9-(a) and 10 are not identical. E.g., the last state to merge with the rest of the country in the reciprocal dendrogram is Alaska at resolution $\delta = 0.975$ whereas the last state to merge in the nonreciprocal dendrogram is Montana at resolution $\delta = 0.962$ with Alaska joining the rest of the country at resolution $\delta = 0.948$. Given the mechanics of \mathcal{H}^{NR} , this must occur due to the existence of a cycle of migration involving Alaska which is stronger than the bidirectional exchange between Alaska and any other state.

One can leverage the results presented in this paper to make universal claims about the clustering outputs for *any* admissible hierarchical clustering method. Indeed, from Theorem 4 we know that any clustering method satisfying the axioms of value and transformation applied to the migration network N_S yields a dendrogram such that the resolution at which any pair of states merges in a common cluster is bounded by the resolutions at which the same pair of states is co-clustered in the reciprocal and nonreciprocal dendrograms. Given the similarity between the dendrograms in figs. 9-(a) and 10, we can assert that any other hierarchical clustering method satisfying the axioms of value and transformation would lead to similar conclusions about the migrational patterns in the United States.

8.3 Unilateral clustering \mathcal{H}^{U}

The dendrogram obtained from applying the unilateral clustering method \mathcal{H}^{U} defined in (25) to the network N_S is shown in Fig. 11-(a); see Remark 2. The colors in the dendrogram correspond to the clusters formed at resolution $\delta_1^{\text{U}} = 0.872$ which are also shown in the map in Fig. 11-(b). States that are clustered together in unilateral partitions are those connected by a chain of intense unidirectional migration flows in the sense dictated by the resolution parameter.

Unilateral clustering permits interactions in either direction. Hence, clusters are centered around populous states whose migration outflows are naturally dominant.

Unilateral clustering may detect one-way migration flows that are overlooked by reciprocal and nonreciprocal clustering. E.g., if people in state A tend to move to B but people in B rarely move to A either directly or through intermediate states, uni-

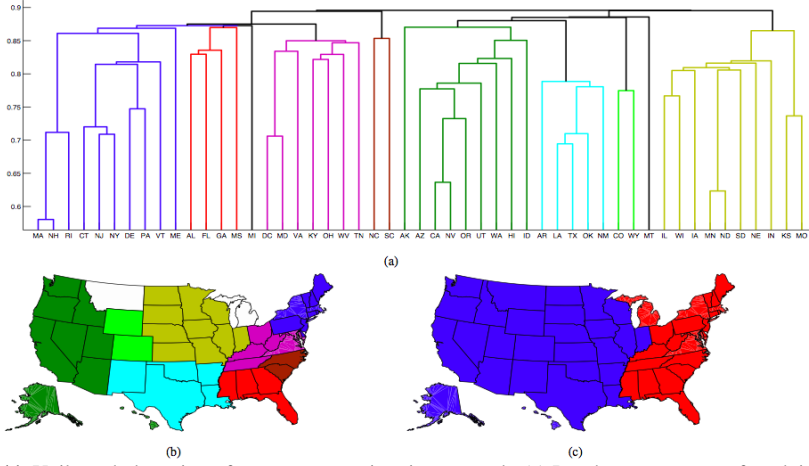


Fig. 11 Unilateral clustering of state-to-state migration network. (a) Dendrogram output of applying the unilateral clustering method \mathcal{H}^U to the network of state-to-state migration N_S . Clusters at resolution $\delta_1^U = 0.872$ are highlighted in color. (b) Highlighted clusters are identified in a map. Clusters tend to form around high populated states. (c) Map colored according to the partition at resolution $\delta_2^U = 0.896$. Two clear clusters, east and west, arise.

lateral clustering merges these two states according to the one-way intense flow from A to B but reciprocal and nonreciprocal clustering do not. The differences between the unilateral dendrogram in Fig. 11-(a) with the reciprocal and nonreciprocal dendrograms in figs. 9-(a) and 10 indicate that migration flows which are intense in one way but not in the other are common. E.g., the first two states to merge in the unilateral dendrogram in Fig. 11-(a) are Massachusetts and New Hampshire at resolution $\delta = 0.580$ because from all the people that moved into New Hampshire, 42% came from Massachusetts, this being the highest value in all the country. The flow in the direction from New Hampshire to Massachusetts is lower, only 9% of the immigrants entering the latter come from the former. This is the reason why these two states are not the first to merge in the reciprocal and nonreciprocal dendrograms.

The relation between geographical proximity and tendency to form clusters is even more determinant here than in reciprocal and nonreciprocal clustering since the exceptions of Texas, California, and Florida do not occur in this case. Indeed, California first merges with Nevada at resolution $\delta = 0.637$, Texas with Louisiana at $\delta = 0.694$, and Florida with Alabama at $\delta = 0.830$, the three pairs of states being neighbors. Moreover, from Fig. 11-(b) it is immediate that at resolution δ_1^U every non-singleton cluster is formed by a set of neighboring states.

Clusters tend to form around populous states. In Fig. 11-(b), the six clusters with more than two states contain the seven states with largest population. The data suggests that the reason for this is that populous states have a strong influence on the immigration into neighboring states. Indeed, if we focus on the cyan cluster formed around Texas, the proportional immigration into Louisiana, New Mexico, Oklahoma, and Arkansas coming from Texas is 31%, 22%, 29%, and 21% respectively. The opposite is not true, since the immigration in the opposite direction is of 5%, 3%, 4%, and 3%, respectively. However, this flow in the opposite sense is not required for unilateral clustering to merge the states into one cluster.

Unilateral clustering detects an east-west division of migration flows in the United States. Fig. 11-(c) shows the two clusters formed at resolution δ_2^U , corresponding to a migrational flow of 10.45%. This implies that for any two different states within the same cluster we can find a unilateral chain where every flow is at least 10.45%. More interestingly, there is no pair of states, one from the east and one from the west, with a flow of 10.45% or more in any direction.

9 Conclusions and discussion

We presented an axiomatic construction of hierarchical clustering for asymmetric networks. Even though the notion of proximity between nodes – hence, the concept of clustering – is unclear when we are given directed dissimilarities, we determined desirable properties that clustering methods should satisfy. These properties were translated into the axioms of value and transformation. We then presented two clustering methods – reciprocal and nonreciprocal – that abide by these axioms. In reciprocal clustering, node clusters are formed based on path of bidirectional influence whereas in nonreciprocal clustering the influence in both directions can be propagated via different paths. More interestingly, we showed that any other method satisfying both axioms must be contained between reciprocal and nonreciprocal clustering in a well-defined sense. We also analyzed alternative axiomatic constructions. The construction based on the Extended Axiom of Value, though seemingly stronger, was shown to be equivalent to the original axiomatic framework. A different construction, based on an Alternative Axiom of Value, gave rise to a unique admissible clustering method, named unilateral clustering where unidirectional influence is sufficient for the formation of clusters. Finally, when applied to symmetric networks, all hierarchical clustering methods considered here boil down to single linkage, in which case the characterization results presented generalize and expand existing results for clustering of finite metric spaces.

Although the results in this paper suggest that many different admissible methods may exist, Theorem 4 forces all of them to coincide with single linkage when the input network is symmetric. This implies that natural extensions of other clustering methods – such as complete or Ward’s linkage (Jain and Dubes, 1988) – to asymmetric networks will not be admissible. Notice that this is not a declaration of the practical validity of these methods but rather a clear-cut classification that can be useful for the practitioner. More precisely, if a practitioner regards the proposed axioms as reasonable properties, then an admissible method must be chosen, whereas if a non-admissible method is used then at least one of the axioms will be violated.

Clustering of asymmetric data is an inherently mismatched problem because we are seeking to study a symmetric relationship (grouping) on asymmetric data. This mismatch implies that there may be more than one way of creating groups and our results indicate that there are three essentially different ways in which groups (clusters) can be created in asymmetric data: through bidirectional relationships (reciprocal clustering), cycles (nonreciprocal clustering), and unidirectional relationships (unilateral clustering). Each of these three ways of creating clusters is fundamental in its own way because they all follow from some basic axiomatic construction. In Section 8 we illustrated these differences and explained how the methods help to uncover

different properties of the dataset of interest. E.g., strong reciprocal clusters formed around shared metropolitan areas, Alaska is involved in a migration cycle, and unilateral clusters form around populous states. Alternatively, one can avoid the mentioned symmetry mismatch by accepting an *asymmetric ultrametric* as the output of the clustering process. We have followed this approach in (Carlsson et al, 2014), where we refer to the process of constructing asymmetric ultrametrics as *quasi-clustering*. It is worth noting that asymmetric ultrametrics cannot be represented using a dendrogram. Indeed, given that asymmetric ultrametrics keep directionality information, we must not only represent groupings but also influences between groups, as well as between elements of groups. Overall, a quasi-clustering analysis is complementary of the clustering analyses that we propose here.

Several applications produce network data exhibiting weights with both positive and negative signs – a scenario not contemplated by the theoretical results developed in this paper. Therefore, an important path of further development is to investigate the degree to which our results may be extended to such more general situation.

10 Appendix: Proofs

Proof of Proposition 2: We prove that any method \mathcal{H} satisfying axioms (A1)-(A2) is idempotent by showing that it satisfies that $\mathcal{H}(X, u_X) = (X, u_X)$, for all $(X, u_X) \in \mathcal{U}$. Consider the application of admissible methods \mathcal{H} to the ultrametric network $U_X = (X, u_X)$. Since U_X is symmetric, from (22) we have that $u_X^{\text{NR}} = u_X^{\text{R}}$. Thus, if we show that \mathcal{H}^{R} is idempotent, we know that \mathcal{H}^{NR} is as well. Moreover, from (20) it would follow that every admissible method is idempotent. Consequently, we need to show that $\mathcal{H}^{\text{R}}(X, u_X) = (X, u_X)$, for all $(X, u_X) \in \mathcal{U}$.

Denoting by $(X, u_X^{\text{R}}) = \mathcal{H}^{\text{R}}(X, u_X)$ the outcome of applying \mathcal{H}^{R} to U_X , we can write for all $x, x' \in X$ [cf. (15)]

$$u_X^{\text{R}}(x, x') = \min_{C(x, x')} \max_{i | x_i \in C(x, x')} u_X(x_i, x_{i+1}), \quad (27)$$

where there is no need to take the maximum between $u_X(x_i, x_{i+1})$ and $u_X(x_{i+1}, x_i)$ since U_X is symmetric. Given a chain $C(x, x')$ and using the fact that u_X is an ultrametric it follows from the strong triangle inequality in (7) that $u_X(x, x') \leq \max_{i | x_i \in C(x, x')} u_X(x_i, x_{i+1})$. Since the previous inequality is valid for *all* chains and the value of $u_X^{\text{R}}(x, x')$ in (27) comes from the cost of some chain, we have that $u_X^{\text{R}}(x, x') \geq u_X(x, x')$, for all $x, x' \in X$. Also, by considering the particular chain $C(x, x') = [x, x']$ with cost $u_X(x, x')$, it follows from (27) that $u_X^{\text{R}}(x, x') \leq u_X(x, x')$, for all $x, x' \in X$. Combining these inequalities, we have that $u_X^{\text{R}}(x, x') = u_X(x, x')$ for all $x, x' \in X$, as wanted \blacksquare

Proof of Theorem 2: In proving Theorem 2, we make use of the following lemma.

Lemma 1 *Let $N = (X, A_X)$ be any network and δ any positive constant. Suppose that $x, x' \in X$ are such that their associated minimum chain cost [cf. (3)] satisfies $\tilde{u}_X^*(x, x') \geq \delta$. Then, there exists a partition $P_\delta(x, x') = \{B_\delta(x), B_\delta(x')\}$ of the node set X into blocks $B_\delta(x)$ and $B_\delta(x')$ with $x \in B_\delta(x)$ and $x' \in B_\delta(x')$ such that $A_X(b, b') \geq \delta$, for all points $b \in B_\delta(x)$ and $b' \in B_\delta(x')$.*

Proof: We prove this by contradiction. If a partition $P_\delta(x, x') = \{B_\delta(x), B_\delta(x')\}$ with $x \in B_\delta(x)$ and $x' \in B_\delta(x')$ satisfying Lemma 1 does not exist for all pairs of points $x, x' \in X$ satisfying $\tilde{u}_X^*(x, x') \geq \delta$, then there is at least one pair of nodes $x, x' \in X$ satisfying $\tilde{u}_X^*(x, x') \geq \delta$ such that for *all* partitions of X into two blocks $P = \{B, B'\}$ with $x \in B$ and $x' \in B'$ we can find at least a pair of elements $b_P \in B$ and $b'_P \in B'$ for which

$$A_X(b_P, b'_P) < \delta. \quad (28)$$

Begin by considering the partition $P_1 = \{B_1, B'_1\}$ where $B_1 = \{x\}$ and $B'_1 = X \setminus \{x\}$. Since (28) is true for all partitions having $x \in B$ and $x' \in B'$ and x is the unique element of B_1 , there must exist a node $b'_{P_1} \in B'_1$ such that

$$A_X(x, b'_{P_1}) < \delta. \quad (29)$$

Hence, the chain $C(x, b'_{P_1}) = [x, b'_{P_1}]$ composed of these two nodes has cost smaller than δ . Moreover, since $\tilde{u}_X^*(x, b'_{P_1})$ represents the minimum cost among all chains $C(x, b'_{P_1})$ linking x to b'_{P_1} , we can assert that $\tilde{u}_X^*(x, b'_{P_1}) \leq A_X(x, b'_{P_1}) < \delta$. Consider now the partition $P_2 = \{B_2, B'_2\}$ where $B_2 = \{x, b'_{P_1}\}$ and $B'_2 = X \setminus B_2$. From (28), there must exist a node $b'_{P_2} \in B'_2$ that satisfies at least one of the two following conditions: i) $A_X(x, b'_{P_2}) < \delta$, or ii) $A_X(b'_{P_1}, b'_{P_2}) < \delta$. If i) is true, the chain $C(x, b'_{P_2}) = [x, b'_{P_2}]$ has cost smaller than δ . If ii) is true, we combine the dissimilarity bound with the one in (29) to conclude that the chain $C(x, b'_{P_2}) = [x, b'_{P_1}, b'_{P_2}]$ has cost smaller than δ . In either case we conclude that there exists a chain $C(x, b'_{P_2})$ linking x to b'_{P_2} whose cost is smaller than δ . Therefore, the minimum chain cost must satisfy $\tilde{u}_X^*(x, b'_{P_2}) < \delta$. We can repeat this process iteratively where, e.g., partition P_3 is composed by $B_3 = \{x, b'_{P_1}, b'_{P_2}\}$ and $B'_3 = X \setminus B_3$, to obtain partitions P_1, P_2, \dots, P_{n-1} and corresponding nodes $b'_{P_1}, b'_{P_2}, \dots, b'_{P_{n-1}}$ such that the associated minimum chain cost satisfies $\tilde{u}_X^*(x, b'_{P_i}) < \delta$, for all i . Observe that nodes b'_{P_i} are distinct by construction and distinct from x . Since there are n nodes in the network it must be that $x' = b'_{P_k}$ for some $i \in \{1, \dots, n-1\}$, entailing that $\tilde{u}_X^*(x, x') < \delta$, and reaching a contradiction. ■

Continuing with the proof of Theorem 2, to show that (A1)-(A2) imply (A1')-(A2) let \mathcal{H} be a method that satisfies (A1) and (A2) and denote by $(\{1, 2, \dots, n\}, u_{n,\alpha,\beta}) = \mathcal{H}(\Delta_n(\alpha, \beta, \Pi))$. We want to prove that (A1') is satisfied which means that we have to show that for all indices $n \in \mathbb{N}$, constants $\alpha, \beta > 0$, permutations Π of $\{1, \dots, n\}$, and points $i \neq j$, we have $u_{n,\alpha,\beta}(i, j) = \max(\alpha, \beta)$. We will do so by showing both i) $u_{n,\alpha,\beta}(i, j) \leq \max(\alpha, \beta)$, and ii) $u_{n,\alpha,\beta}(i, j) \geq \max(\alpha, \beta)$, for all $n \in \mathbb{N}$, $\alpha, \beta > 0$, Π , and $i \neq j$.

To prove i), define the two-node network $N_{\max} := \Delta_2(\max(\alpha, \beta), \max(\alpha, \beta))$ and define $(\{p, q\}, u_{p,q}) := \mathcal{H}(N_{\max})$. Since \mathcal{H} abides by (A1),

$$u_{p,q}(p, q) = \max(\max(\alpha, \beta), \max(\alpha, \beta)) = \max(\alpha, \beta). \quad (30)$$

Consider now the map $\phi_{i,j} : \{p, q\} \rightarrow \{1, \dots, n\}$ from N_{\max} to the permuted canonical network $\Delta_n(\alpha, \beta, \Pi)$ where $\phi_{i,j}(p) = i$ and $\phi_{i,j}(q) = j$. Since dissimilarities in $\Delta_n(\alpha, \beta, \Pi)$ are either α or β and the dissimilarities in the two-node network are $\max(\alpha, \beta)$ it follows that the map $\phi_{i,j}$ is dissimilarity reducing regardless of the particular values of i and j . Since the method \mathcal{H} was assumed to satisfy (A2) as well, we

must have $u_{p,q}(p, q) \geq u_{n,\alpha,\beta}(\phi_{i,j}(p), \phi_{i,j}(q)) = u_{n,\alpha,\beta}(i, j)$. Inequality i) follows from substituting (30) into this last expression.

In order to show inequality ii), pick two arbitrary distinct nodes $i, j \in \{1, \dots, n\}$ in the node set of $\Delta_n(\alpha, \beta, \Pi)$. Denote by $C(i, j)$ and $C(j, i)$ two minimizing chains in the definition (3) of the directed minimum chain costs $\tilde{u}_{n,\alpha,\beta}^*(i, j)$ and $\tilde{u}_{n,\alpha,\beta}^*(j, i)$ respectively. Observe that at least one of the following two inequalities must be true $\tilde{u}_{n,\alpha,\beta}^*(i, j) \geq \max(\alpha, \beta)$ or $\tilde{u}_{n,\alpha,\beta}^*(j, i) \geq \max(\alpha, \beta)$. Indeed, if both inequalities were false, the concatenation of $C(i, j)$ and $C(j, i)$ would form a loop $C(i, i) = C(i, j) \uplus C(j, i)$ of cost strictly less than $\max(\alpha, \beta)$. This cannot be true because $\max(\alpha, \beta)$ is the minimum loop cost of the network $\Delta_n(\alpha, \beta, \Pi)$.

Without loss of generality assume $\tilde{u}_{n,\alpha,\beta}^*(i, j) \geq \max(\alpha, \beta)$ is true and consider $\delta = \max(\alpha, \beta)$. By Lemma 1 we are therefore guaranteed to find a partition of the node set $\{1, \dots, n\}$ into two blocks $B_\delta(i)$ and $B_\delta(j)$ with $i \in B_\delta(i)$ and $j \in B_\delta(j)$ such that for all $b \in B_\delta(i)$ and $b' \in B_\delta(j)$ it holds that

$$\Pi(A_{n,\alpha,\beta})(b, b') \geq \delta = \max(\alpha, \beta). \quad (31)$$

Define a two-node network $N_{\min} := \Delta_2(\max(\alpha, \beta), \min(\alpha, \beta)) = (\{r, s\}, A_{r,s})$ where $A_{r,s}(r, s) = \max(\alpha, \beta)$ and $A_{r,s}(s, r) = \min(\alpha, \beta)$ and define $(\{r, s\}, u_{r,s}) := \mathcal{H}(N_{\min})$. Since the method \mathcal{H} satisfies (A1) we must have

$$u_{r,s}(r, s) = \max(\max(\alpha, \beta), \min(\alpha, \beta)) = \max(\alpha, \beta). \quad (32)$$

Consider the map $\phi'_{i,j} : \{1, \dots, n\} \rightarrow \{r, s\}$ such that $\phi'_{i,j}(b) = r$ for all $b \in B_\delta(i)$ and $\phi'_{i,j}(b') = s$ for all $b' \in B_\delta(j)$. The map $\phi'_{i,j}$ is dissimilarity reducing because

$$\Pi(A_{n,\alpha,\beta})(k, l) \geq A_{r,s}(\phi'_{i,j}(k), \phi'_{i,j}(l)), \quad (33)$$

for all $k, l \in \{1, \dots, n\}$. To see the validity of (33) consider three different possible cases. If k and l belong both to the same block, i.e., either $k, l \in B_\delta(i)$ or $k, l \in B_\delta(j)$, then $\phi'_{i,j}(k) = \phi'_{i,j}(l)$ and $A_{r,s}(\phi'_{i,j}(k), \phi'_{i,j}(l)) = 0$, immediately satisfying (33). If $k \in B_\delta(j)$ and $l \in B_\delta(i)$ it holds that $A_{r,s}(\phi'_{i,j}(k), \phi'_{i,j}(l)) = A_{r,s}(s, r) = \min(\alpha, \beta)$ which cannot exceed $\Pi(A_{n,\alpha,\beta})(k, l)$ which is either equal to α or β . If $k \in B_\delta(i)$ and $l \in B_\delta(j)$, then we have $A_{r,s}(\phi'_{i,j}(k), \phi'_{i,j}(l)) = A_{r,s}(r, s) = \max(\alpha, \beta)$ but we also have $\Pi(A_{n,\alpha,\beta})(k, l) = \max(\alpha, \beta)$ as it follows by taking $b = k$ and $b' = l$ in (31), thus, again satisfying (33).

Since \mathcal{H} fulfills the Axiom of Transformation (A2) we must have

$$u_{n,\alpha,\beta}(i, j) \geq u_{r,s}(\phi'_{i,j}(i), \phi'_{i,j}(j)) = u_{r,s}(r, s). \quad (34)$$

Substituting (32) in (34) we obtain the inequality ii). Combining both inequalities i) and ii), it follows that $u_{n,\alpha,\beta}(i, j) = \max(\alpha, \beta)$. Thus, admissibility with respect to (A1)-(A2) implies admissibility with respect to (A1')-(A2). The opposite implication is immediate since (A1) is a particular case of (A1'), concluding the proof. ■

Proof of Theorem 3: We show that if a clustering method satisfies axioms (A1') and (A2) then it satisfies the Property of Influence (P1). Notice that this result, combined with Theorem 2, implies the statement of Theorem 3. The following lemma is instrumental in the ensuing proof.

Lemma 2 Let $N = (X, A_X)$ be an arbitrary network with n nodes and $\Delta_n(\alpha, \beta) = (\{1, \dots, n\}, A_{n,\alpha,\beta})$ be the canonical network with $0 < \alpha \leq \text{sep}(X, A_X)$ and $\beta = \text{mlc}(X, A_X)$. Then, there exists a bijective dissimilarity-reducing map $\phi : X \rightarrow \{1, \dots, n\}$, i.e. $A_X(x, x') \geq A_{n,\alpha,\beta}(\phi(x), \phi(x'))$, for all $x, x' \in X$.

Proof: To construct the map ϕ consider the function $P : X \rightarrow \mathcal{P}(X)$ from the node set X to its power set $\mathcal{P}(X)$ such that $P(x) := \{x' \in X \mid x' \neq x, A_X(x', x) < \beta\}$, for all $x \in X$. Having $r \in P(s)$ for some $r, s \in X$ implies that $A_X(r, s) < \beta = \text{mlc}(X, A_X)$. An important observation is that we must have a node $x \in X$ whose P -image is empty. Otherwise, pick a node $x_n \in X$ and construct the chain $[x_0, x_1, \dots, x_n]$ where the i th element x_{i-1} of the chain is in the P -image of x_i . From the definition of P it follows that all dissimilarities along this chain satisfy $A_X(x_{i-1}, x_i) < \beta = \text{mlc}(X, A_X)$. But since the chain $[x_0, x_1, \dots, x_n]$ contains $n + 1$ elements, at least one node must be repeated. Hence, we have found a loop for which all dissimilarities are bounded above by $\beta = \text{mlc}(X, A_X)$, which is impossible because it contradicts the definition of the minimum loop cost in (5). We can then find a node x_{i_1} for which $P(x_{i_1}) = \emptyset$. Fix $\phi(x_{i_1}) = 1$.

Select now a node $x_{i_2} \neq x_{i_1}$ whose P -image is either $\{x_{i_1}\}$ or \emptyset , which we write jointly as $P(x_{i_2}) \subseteq \{x_{i_1}\}$. Following a similar reasoning to the previous one, such a node must exist and fix $\phi(x_{i_2}) = 2$. Repeat this process k times so that at step k we have $\phi(x_{i_k}) = k$ for a node $x_{i_k} \notin \{x_{i_1}, x_{i_2}, \dots, x_{i_{k-1}}\}$ whose P -image is a subset of the nodes already picked, i.e., $P(x_{i_k}) \subseteq \{x_{i_1}, \dots, x_{i_{k-1}}\}$. Since all the nodes x_{i_k} are different, the map ϕ with $\phi(x_{i_k}) = k$ is bijective. By construction, ϕ is such that for all $l > k$, $x_{i_l} \notin P(x_{i_k})$. From the definition of P , this implies that the dissimilarity from x_{i_l} to x_{i_k} must satisfy $A_X(x_{i_l}, x_{i_k}) \geq \beta$, for all $l > k$. Moreover, from the definition of the canonical matrix $A_{n,\alpha,\beta}$ we have that $A_{n,\alpha,\beta}(\phi(x_{i_l}), \phi(x_{i_k})) = A_{n,\alpha,\beta}(l, k) = \beta$ for all $l > k$. By combining these two expressions, we conclude that $A_X(x, x') \geq A_{n,\alpha,\beta}(\phi(x), \phi(x'))$ is true for all points with $\phi(x) > \phi(x')$. When $\phi(x) < \phi(x')$, we have $A_{n,\alpha,\beta}(\phi(x), \phi(x')) = \alpha$ which was assumed to be bounded above by the separation of the network (X, A_X) , thus, $A_{n,\alpha,\beta}(\phi(x), \phi(x'))$ is not greater than any positive dissimilarity in the range of A_X . ■

Continuing the main proof of Theorem 3, consider a given arbitrary network $N = (X, A_X)$ with $X = \{x_1, x_2, \dots, x_n\}$ and define $(X, u_X) := \mathcal{H}(X, A_X)$. The method \mathcal{H} is known to satisfy (A1') and (A2) and we want to show that it satisfies (P1) for which we need to show that $u_X(x, x') \geq \text{mlc}(X, A_X)$ for all $x \neq x'$.

Consider the canonical network $\Delta_n(\alpha, \beta) = (\{1, \dots, n\}, A_{n,\alpha,\beta})$ with $\beta = \text{mlc}(X, A_X)$ being the minimum loop cost of the network N and $\alpha > 0$ a constant not exceeding the separation of the network. Thus, we have $\alpha \leq \text{sep}(X, A_X) \leq \text{mlc}(X, A_X) = \beta$. Note that networks N and $\Delta_n(\alpha, \beta)$ have equal number of nodes.

Defining $(\{1, \dots, n\}, u_{\alpha,\beta}) := \mathcal{H}(\Delta_n(\alpha, \beta))$, since \mathcal{H} satisfies the Extended Axiom of Value (A1'), then for all indices $i, j \in \{1, \dots, n\}$ with $i \neq j$ we have

$$u_{\alpha,\beta}(i, j) = \max(\alpha, \beta) = \beta = \text{mlc}(X, A_X). \quad (35)$$

Further, focus on the bijective dissimilarity-reducing map considered in Lemma 2 and notice that since \mathcal{H} satisfies (A2) it follows that for all $x, x' \in X$

$$u_X(x, x') \geq u_{\alpha,\beta}(\phi(x), \phi(x')). \quad (36)$$

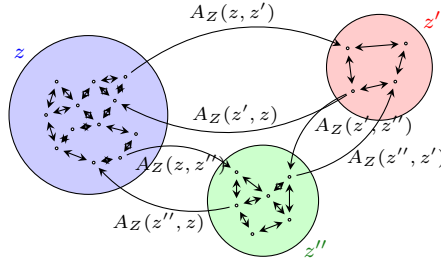


Fig. 12 Network of equivalence classes for a given resolution. The Axiom of Transformation permits relating the clustering in the original network and the clustering in the network of equivalence classes.

Since the equality in (35) is true for all $i \neq j$ and since all points $x \neq x'$ are mapped to points $\phi(x) \neq \phi(x')$ because ϕ is bijective, (36) implies $u_X(x, x') \geq \beta = \text{mlc}(X, A_X)$, for all distinct $x, x' \in X$. ■

Proof of Theorem 4: We prove the theorem by showing both inequalities in (20). **Proof of $u_X^{\text{NR}}(x, x') \leq u_X(x, x')$:** Recall that validity of (A1)-(A2) implies validity of (P1) by Theorem 3. Consider the nonreciprocal clustering equivalence relation $\sim_{\text{NR}_X(\delta)}$ at resolution δ according to which $x \sim_{\text{NR}_X(\delta)} x'$ if and only if x and x' belong to the same nonreciprocal cluster at resolution δ . Notice that this is true if and only if $u_X^{\text{NR}}(x, x') \leq \delta$. Further consider the set $Z := X \bmod \sim_{\text{NR}_X(\delta)}$ of corresponding equivalence classes and the map $\phi_\delta : X \rightarrow Z$ that maps each point of X to its equivalence class. Notice that x and x' are mapped to the same point z if they belong to the same cluster at resolution δ .

We define the network $N_Z := (Z, A_Z)$ by endowing Z with the dissimilarity A_Z derived from the dissimilarity A_X as

$$A_Z(z, z') := \min_{x \in \phi_\delta^{-1}(z), x' \in \phi_\delta^{-1}(z')} A_X(x, x'). \quad (37)$$

The dissimilarity $A_Z(z, z')$ compares all the dissimilarities $A_X(x, x')$ between a member of the equivalence class z and a member of the equivalence class z' and sets $A_Z(z, z')$ to the value corresponding to the least dissimilar pair; see Fig. 12. Notice that according to construction, the map ϕ_δ is dissimilarity reducing $A_X(x, x') \geq A_Z(\phi_\delta(x), \phi_\delta(x'))$, because we either have $A_Z(\phi_\delta(x), \phi_\delta(x')) = 0$ if x and x' are co-clustered at resolution δ , or $A_X(x, x') \geq \min_{x \in \phi_\delta^{-1}(z), x' \in \phi_\delta^{-1}(z')} A_X(x, x') = A_Z(\phi_\delta(x), \phi_\delta(x'))$ if they are mapped to different equivalent classes.

Consider now an arbitrary method \mathcal{H} satisfying axioms (A1)-(A2) and denote by $(Z, u_Z) = \mathcal{H}(N_Z)$ the outcome of \mathcal{H} when applied to N_Z . To apply Property (P1) to this outcome we determine the minimum loop cost of N_Z in the following claim.

Claim 1 *The minimum loop cost of the network N_Z is $\text{mlc}(N_Z) > \delta$.*

Proof: Assume that Claim 1 is not true, denote by $C(z, z) = [z, z', \dots, z^{(l)}, z]$ a loop of cost smaller than δ and consider arbitrary nodes $x \in \phi_\delta^{-1}(z)$ and $x' \in \phi_\delta^{-1}(z')$. By definition, given two nodes in the same equivalence class, we can always find a chain from one to the other of cost not larger than δ . Moreover, since we are assuming that $A_Z(z, z') \leq \delta$, this implies that there exists at least one node x_1 belonging to class z

and another node x_2 belonging to z' such that $A_X(x_1, x_2) \leq \delta$. Combining these two facts, we can guarantee the existence of a chain from x to x' of cost not larger than δ , since we can go first from x to x_1 then from x_1 to x_2 and finally from x_2 to x' without encountering dissimilarities greater than δ . In a similar way, we can go from x' to x by constructing a chain that goes through all the equivalence classes in $C(z, z)$, i.e., from z' to z'' then to $z^{(3)}$ and so on until we reach z . Since we can go from x to x' and back with chains of cost not exceeding δ , it follows that $u_X^{\text{NR}}(x, x') \leq \delta$ contradicting the assumption that x and x' belong to different equivalent classes. Therefore, the assumption that Claim 1 is false cannot hold. ■

Continuing with the main proof, since the minimum loop cost of N_Z satisfies Claim 1 it follows from Property (P1) that $u_Z(z, z') > \delta$ for all pairs of distinct equivalent classes z, z' . Further note that, since ϕ_δ is dissimilarity reducing, Axiom (A2) implies that $u_X(x, x') \geq u_Z(z, z')$. Combining these facts, we can conclude that when x and x' map to different equivalence classes it holds that $u_X(x, x') \geq u_Z(z, z') > \delta$. Recall that x and x' mapping to different equivalence classes is equivalent to $u_X^{\text{NR}}(x, x') > \delta$. Consequently, we can claim that $u_X^{\text{NR}}(x, x') > \delta$ implies $u_X(x, x') > \delta$, or, in set notation that $\{(x, x') : u_X^{\text{NR}}(x, x') > \delta\} \subseteq \{(x, x') : u_X(x, x') > \delta\}$. Since the previous expression is true for arbitrary $\delta > 0$ it implies that $u_X^{\text{NR}}(x, x') \leq u_X(x, x')$ for all $x, x' \in X$ as in the first inequality in (20). ■

Proof of $u_X(\mathbf{x}, \mathbf{x}') \leq u_X^{\text{R}}(\mathbf{x}, \mathbf{x}')$: To prove the second inequality in (20) consider points x and x' with reciprocal ultrametric $u_X^{\text{R}}(x, x') = \delta$. Let $C^*(x, x') = [x = x_0, \dots, x_l = x']$ be a chain achieving the minimum in (15) so that we can write

$$\delta = u_X^{\text{R}}(x, x') = \max_{i|x_i \in C^*(x, x')} \max(A_X(x_i, x_{i+1}), A_X(x_{i+1}, x_i)). \quad (38)$$

Turn attention to the symmetric two-node network $\Delta_2(\delta, \delta) = (\{p, q\}, A_{p,q})$ with $A_{p,q}(p, q) = A_{p,q}(q, p) = \delta$ and define $(\{p, q\}, u_{p,q}) := \mathcal{H}(\Delta_2(\delta, \delta))$. Notice that according to Axiom (A1) we have $u_{p,q}(p, q) = \max(\delta, \delta) = \delta$.

Focus now on transformations $\phi_i : \{p, q\} \rightarrow X$ given by $\phi_i(p) = x_i$, $\phi_i(q) = x_{i+1}$ so as to map p and q to subsequent points in the chain $C^*(x, x')$ used in (38). Since it follows from (38) that $A_X(x_i, x_{i+1}) \leq \delta$ and $A_X(x_{i+1}, x_i) \leq \delta$ for all i , it is just a simple matter of notation to observe that

$$A_X(\phi_i(p), \phi_i(q)) \leq A_{p,q}(p, q) = \delta, \quad A_X(\phi_i(q), \phi_i(p)) \leq A_{p,q}(q, p) = \delta. \quad (39)$$

Since according to (39) transformations ϕ_i are dissimilarity reducing, it follows from Axiom (A2) that $u_X(x_i, x_{i+1}) = u_X(\phi_i(p), \phi_i(q)) \leq u_{p,q}(p, q) = \delta$, for all i . To complete the proof we use the fact that since u_X is an ultrametric and $C^*(x, x') = [x = x_0, \dots, x_l = x']$ is a chain joining x and x' the strong triangle inequality dictates [cf. (7)] that $u_X(x, x') \leq \max_i u_X(x_i, x_{i+1}) \leq \delta$. The proof of the second inequality in (20) follows by substituting $\delta = u_X^{\text{R}}(x, x')$ [cf. (38)]. ■

Having showed both inequalities in (20), the global proof concludes. ■

Proof of Theorem 5: Suppose there exists a clustering method \mathcal{H} that satisfies axioms (A1'') and (A2) but does not satisfy Property (P1'). This means that there exists a network $N = (X, A_X)$ with output ultrametrics $(X, u_X) = \mathcal{H}(N)$ for which

$u_X(x_1, x_2) < \text{sep}(X, A_X)$ for at least one pair of nodes $x_1 \neq x_2 \in X$. Focus on a symmetric two-node network $\Delta_2(s, s) = (\{p, q\}, A_{p,q})$ with $A_{p,q}(p, q) = A_{p,q}(q, p) = s = \text{sep}(X, A_X)$ and define $(X, u_{p,q}) = \mathcal{H}(\Delta_2(s, s))$. From Axiom (A1''), we must have that

$$u_{p,q}(p, q) = \min(\text{sep}(X, A_X), \text{sep}(X, A_X)) = \text{sep}(X, A_X). \quad (40)$$

Construct the map $\phi : X \rightarrow \{p, q\}$ from the network N to $\Delta_2(s, s)$ that takes node x_1 to $\phi(x_1) = p$ and every other node $x \neq x_1$ to $\phi(x) = q$. No dissimilarity can be increased when applying ϕ since every dissimilarity is mapped either to zero or to $\text{sep}(X, A_X)$ which is by definition the minimum dissimilarity in the original network [cf. (6)]. Hence, ϕ is dissimilarity reducing and from Axiom (A2) it follows that $u_X(x_1, x_2) \geq u_{p,q}(\phi(x_1), \phi(x_2)) = u_{p,q}(p, q)$. By substituting (40) into the previous expression, we contradict $u_X(x_1, x_2) < \text{sep}(X, A_X)$ proving that such method \mathcal{H} cannot exist. ■

Proof of Proposition 5: To show fulfillment of (A1''), consider the network $\Delta_2(\alpha, \beta)$ and define $(\{p, q\}, u_{p,q}^U) := \mathcal{H}^U(\Delta_2(\alpha, \beta))$. Since every chain connecting p and q must contain these two nodes as consecutive nodes, applying the definition in (25) yields $u_{p,q}^U(p, q) = \min(A_{p,q}(p, q), A_{p,q}(q, p)) = \min(\alpha, \beta)$, and Axiom (A1'') is thereby satisfied. In order to show fulfillment of Axiom (A2), the proof is analogous to the one developed in Proposition 3. The proof only differs in the appearance of minimizations instead of maximizations to account for the difference in the definitions of unilateral and reciprocal ultrametrics [cf. (25) and (15)]. ■

Proof of Theorem 6: Given an arbitrary network (X, A_X) , denote by \mathcal{H} a clustering method that fulfills axioms (A1'') and (A2) and define $(X, u_X) := \mathcal{H}(X, A_X)$. Then, we show the theorem by proving the following inequalities for all nodes $x, x' \in X$,

$$u_X^U(x, x') \leq u_X(x, x') \leq u_X^U(x, x'). \quad (41)$$

Proof of leftmost inequality in (41): Consider the unilateral clustering equivalence relation $\sim_{U_X(\delta)}$ at resolution δ according to which $x \sim_{U_X(\delta)} x'$ if and only if x and x' belong to the same unilateral cluster at resolution δ . That is, $x \sim_{U_X(\delta)} x' \iff u_X^U(x, x') \leq \delta$. Further, as in the proof of Theorem 4, consider the set Z of equivalence classes at resolution δ . That is, $Z := X \text{ mod } \sim_{U_X(\delta)}$. Also, consider the map $\phi_\delta : X \rightarrow Z$ that maps each point of X to its equivalence class. Notice that x and x' are mapped to the same point z if and only if they belong to the same block at resolution δ , consequently $\phi_\delta(x) = \phi_\delta(x') \iff u_X^U(x, x') \leq \delta$. We define the network $N_Z = (Z, A_Z)$ by endowing Z with the dissimilarity function A_Z derived from A_X as explained in (37). For further details on this construction, review the corresponding proof in Theorem 4 and see Fig. 12. We stress the fact that the map ϕ_δ is dissimilarity reducing for all δ .

Claim 2 *The separation of the equivalence class network N_Z is $\text{sep}(N_Z) > \delta$.*

Proof: First, observe that by definition of unilateral clustering (25), we know that,

$$u_X^U(x, x') \leq \min(A_X(x, x'), A_X(x', x)), \quad (42)$$

since a two-node chain between nodes x and x' is a particular chain joining the two nodes whereas the ultrametric is calculated as the minimum over all chains. Now, assume that $\text{sep}(N_Z) \leq \delta$. Therefore, by (37) there exists a pair of nodes x and x' that belong to different equivalence classes and have $A_X(x, x') \leq \delta$. However, if x and x' belong to different equivalence classes, they cannot be clustered at resolution δ , hence, $u_X^U(x, x') > \delta$. Inequalities $A_X(x, x') \leq \delta$ and $u_X^U(x, x') > \delta$ cannot hold simultaneously since they contradict (42). Thus, it must be that $\text{sep}(N_Z) > \delta$. ■

Define $(Z, u_Z) := \mathcal{H}(Z, A_Z)$ and, since $\text{sep}(N_Z) > \delta$ (cf. Claim 2), it follows from Property (P1') that for all $z \neq z'$ it holds $u_Z(z, z') > \delta$. Further, recalling that ϕ_δ is a dissimilarity-reducing map, from Axiom (A2) we must have $u_X(x, x') \geq u_Z(\phi_\delta(x), \phi_\delta(x')) = u_Z(z, z')$ for some $z, z' \in Z$. This fact, combined with $u_Z(z, z') > \delta$, entails that when $\phi_\delta(x)$ and $\phi_\delta(x')$ belong to different equivalence classes $u_X(x, x') \geq u_Z(\phi_\delta(x), \phi_\delta(x')) > \delta$. Notice now that $\phi_\delta(x)$ and $\phi_\delta(x')$ belonging to different equivalence classes is equivalent to $u_X^U(x, x') > \delta$. Hence, we can state that $u_X^U(x, x') > \delta$ implies $u_X(x, x') > \delta$ for any arbitrary $\delta > 0$. In set notation, $\{(x, x') : u_X^U(x, x') > \delta\} \subseteq \{(x, x') : u_X(x, x') > \delta\}$. Since the previous expression is true for arbitrary $\delta > 0$, this implies that $u_X^U(x, x') \leq u_X(x, x')$, proving the left inequality in (41). ■

Proof of rightmost inequality in (41): Consider two nodes x and x' with unilateral ultrametric value $u_X^U(x, x') = \delta$. Let $C^*(x, x') = [x = x_0, \dots, x_l = x']$ be a minimizing chain in the definition (25) so that we can write

$$\delta = u_X^U(x, x') = \max_{i | x_i \in C^*(x, x')} \min \left(A_X(x_i, x_{i+1}), A_X(x_{i+1}, x_i) \right). \quad (43)$$

Consider the two-node network $\Delta_2(\delta, M) = (\{p, q\}, A_{p,q})$ where $M := \max_{x, x'} A_X(x, x')$ and define $(\{p, q\}, u_{p,q}) := \mathcal{H}(\{p, q\}, A_{p,q})$. Notice that according to Axiom (A1'') we have $u_{p,q}(p, q) = u_{p,q}(q, p) = \min(\delta, M) = \delta$, where the last equality is enforced by the definition of M .

Focus now on each link of the minimizing chain in (43). For every successive pair of nodes x_i and x_{i+1} , we must have

$$\max \left(A_X(x_i, x_{i+1}), A_X(x_{i+1}, x_i) \right) \leq M, \quad (44)$$

$$\min \left(A_X(x_i, x_{i+1}), A_X(x_{i+1}, x_i) \right) \leq \delta. \quad (45)$$

Expression (44) is true since M is defined as the maximum dissimilarity in A_X . Inequality (45) is justified by (43), since δ is defined as the maximum among links of the minimum distance in both directions of the link. This observation allows the construction of dissimilarity-reducing maps $\phi_i : \{p, q\} \rightarrow X$,

$$\phi_i := \begin{cases} \phi_i(p) = x_i, \phi_i(q) = x_{i+1}, & \text{if } \hat{A}_X(x_i, x_{i+1}) = A_X(x_i, x_{i+1}) \\ \phi_i(q) = x_i, \phi_i(p) = x_{i+1}, & \text{otherwise.} \end{cases} \quad (46)$$

In this way, we can map p and q to subsequent nodes in the chain $C(x, x')$ used in (43). Inequalities (44) and (45) combined with the map definition in (46) guarantee

that ϕ_i is a dissimilarity-reducing map for every i . Since clustering method \mathcal{H} satisfies Axiom (A2), it follows that

$$u_X(\phi_i(p), \phi_i(q)) \leq u_{p,q}(p, q) = \delta, \quad \text{for all } i. \quad (47)$$

Substituting $\phi_i(p)$ and $\phi_i(q)$ in (47) by the corresponding nodes given by the definition (46), we can write $u_X(x_i, x_{i+1}) = u_X(x_{i+1}, x_i) \leq \delta$, for all i , where the symmetry property of ultrametrics was used. To complete the proof we invoke the strong triangle inequality (7) and apply it to $C(x, x') = [x = x_0, \dots, x_l = x']$, the minimizing chain in (43). As a consequence, $u_X(x, x') \leq \max_i u_X(x_i, x_{i+1}) \leq \delta$. The proof of the right inequality in (41) is completed by substituting $\delta = u_X^U(x, x')$ [cf. (43)] into the last previous expression. ■

Having proved both inequalities in (41), unilateral clustering is the only method that satisfies axioms (A1'') and (A2), completing the global proof. ■

Proof of Theorem 7: The leftmost inequality in (26) can be proved using the same method of proof used for the leftmost inequality in (41) within the proof of Theorem 6. The proof of the rightmost inequality in (26) is equivalent to the proof of the rightmost inequality in Theorem 4. ■

References

- Ackerman M, Ben-David S (2008) Measures of clustering quality: A working set of axioms for clustering. In: Neural Info. Process. Syst. (NIPS), pp 121–128
- Bach F, Jordan M (2004) Learning spectral clustering. In: Neural Info. Process. Syst. (NIPS), pp 305–312
- Ben-David S, Von Luxburg U, Pál D (2006) A sober look at clustering stability. In: Conf. Learning Theory (COLT), pp 5–19
- Boyd JP (1980) Asymmetric clusters of internal migration regions of France. IEEE Trans Syst Man Cybern 2:101–104
- Burago D, Burago Y, Ivanov S (2001) A Course in Metric Geometry, AMS Graduate Studies in Math., vol 33. American Mathematical Society
- Carlsson G, Mémoli F (2010a) Characterization, stability and convergence of hierarchical clustering methods. J Mach Learn Res 11:1425–1470
- Carlsson G, Mémoli F (2010b) Multiparameter hierarchical clustering methods. In: Conf. Intl. Fed. Classif. Soc. (IFCS), Springer-Verlag, pp 63–70
- Carlsson G, Mémoli F (2013) Classifying clustering schemes. Found Comp Math 13(2):221–252
- Carlsson G, Mémoli F, Ribeiro A, Segarra S (2013a) Alternative axiomatic constructions for hierarchical clustering of asymmetric networks. In: Global Conf. on Signal and Info. Process. (GlobalSIP), pp 791–794
- Carlsson G, Mémoli F, Ribeiro A, Segarra S (2013b) Axiomatic construction of hierarchical clustering in asymmetric networks. In: Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP), pp 5219–5223
- Carlsson G, Mémoli F, Ribeiro A, Segarra S (2014) Hierarchical quasi-clustering methods for asymmetric networks. JMLR W&CP: Intl Conf Mach Learn 32(1):352–360
- Chino N (2012) A brief survey of asymmetric MDS and some open problems. Behaviormetrika 39(1):127–165
- Chino N, Shiraiwa K (1993) Geometrical structures of some non-distance models for asymmetric MDS. Behaviormetrika 20(1):35–47
- Choi JI, Jain M, Srinivasan K, Levis P, Katti S (2010) Achieving single channel, full duplex wireless communication. In: Proc. Intl. Conf. Mobile Comp. and Netw., ACM, pp 1–12
- Chung FR (1997) Spectral Graph Theory, vol 92. American Mathematical Soc.
- Guyon I, Von Luxburg U, Williamson RC (2009) Clustering: Science or art. In: NIPS 2009 wksp. on Clustering Theory

- Hu TC (1961) The maximum capacity route problem. *Operations Research* 9(6):898–900
- Hubert L (1973) Min and max hierarchical clustering using asymmetric similarity measures. *Psychometrika* 38(1):63–72
- Jain A, Dubes RC (1988) *Algorithms for Clustering Data*. Prentice Hall Advanced Reference Series, Prentice Hall Inc.
- Kleinberg JM (2002) An impossibility theorem for clustering. In: *Neural Info. Process. Syst. (NIPS)*, pp 446–453
- Lance GN, Williams WT (1967) A general theory of classificatory sorting strategies 1: Hierarchical systems. *Computer J* 9(4):373–380
- Meila M, Pentney W (2007) Clustering by weighted cuts in directed graphs. *SIAM Intl Conf Data Mining* pp 135–144
- Müllner D (2011) Modern hierarchical, agglomerative clustering algorithms. *ArXiv e-prints* 1109.2378
- Murtagh F (1985) *Multidimensional Clustering Algorithms*. Compstat Lectures, Vienna: Physica Verlag
- Newman M, Girvan M (2002) Community structure in social and biological networks. *Proc Ntnl Acad Sci* 99(12):7821–7826
- Newman M, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E* 69, 026113
- Ng A, Jordan M, Weiss Y (2002) On spectral clustering: Analysis and an algorithm. In: *Neural Info. Process. Syst. (NIPS)*, pp 849–856
- Okada A, Iwamoto T (1996) University enrollment flow among the Japanese prefectures: A comparison before and after the joint first stage achievement test by asymmetric cluster analysis. *Behaviormetrika* 23(2):169–185
- Pentney W, Meila M (2005) Spectral clustering of biological sequence data. In: *Ntnl. Conf. Artificial Intel.*, pp 845–850
- Saito T, Yadohisa H (2004) *Data Analysis of Asymmetric Structures: Advanced Approaches in Computational Statistics*. CRC Press
- Sato Y (1988) An analysis of sociometric data by MDS in Minkowski space. *Statistical Theory and Data Analysis II* pp 385–396
- Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell* 22(8):888–905
- Slater P (1976) Hierarchical internal migration regions of France. *IEEE Trans Syst Man Cybern* 4:321–324
- Slater P (1984) A partial hierarchical regionalization of 3140 US counties on the basis of 1965–1970 intercounty migration. *Env Plan A* 16(4):545–550
- Smith Z, Chowdhury S, Mévoli (2016) Hierarchical representations of network data with optimal distortion bounds. In: *Asilomar Conf. Signals, Systems and Computers*, pp 1773–1777
- Tarjan RE (1983) An improved algorithm for hierarchical clustering using strong components. *Inf Process Lett* 17(1):37–41
- Vicari D (2014) Classification of asymmetric proximity data. *Journal of Classification* 31(3):386–420
- Vicari D (2015) CLUSKEXT: Clustering model for skew-symmetric data including external information. *Advances in Data Analysis and Classification* pp 1–22
- Von Luxburg U (2007) A tutorial on spectral clustering. *Stat Comp* 17(4):395–416
- Von Luxburg U, Ben-David S (2005) Towards a statistical theory of clustering. In: *PASCAL wksp. on Statistics and Optimization of Clustering*
- Xu R, Wunsch D (2005) Survey of clustering algorithms. *IEEE Trans Neural Netw* 16(3):645–678
- Zadeh RB, Ben-David S (2009) A uniqueness theorem for clustering. In: *Conf. Uncert. Artif. Intell. (UAI)*, pp 639–646
- Zhao Y, Karypis G (2005) Hierarchical clustering algorithms for document datasets. *Data Min Knowl Discov* 10:141–168
- Zhou D, Schölkopf B, Hofmann T (2005) Semi-supervised learning on directed graphs. In: *Neural Info. Process. Syst. (NIPS)*, pp 1633–1640