

Morse Theory and an Impossibility Theorem for Graph Clustering

Fabio Strazzeri*

F.STRAZZERI@SOTON.AC.UK

Rubén J. Sánchez-García

R.SANCHEZ-GARCIA@SOTON.AC.UK

Editor:

Abstract

Kleinberg introduced three natural clustering properties, or axioms, and showed they cannot be simultaneously satisfied by any clustering algorithm. We present a new clustering property, Monotonic Consistency, which avoids the well-known problematic behaviour of Kleinberg's Consistency axiom, and the impossibility result. Namely, we describe a clustering algorithm, Morse Clustering, inspired by Morse Theory in Differential Topology, which satisfies Kleinberg's original axioms with Consistency replaced by Monotonic Consistency. Morse clustering uncovers the underlying flow structure on a set or graph and returns a partition into trees representing basins of attraction of critical vertices. We also generalise Kleinberg's axiomatic approach to sparse graphs, showing an impossibility result for Consistency, and a possibility result for Monotonic Consistency and Morse clustering.

Keywords: data clustering, graph clustering, axiomatic framework, Morse theory, Morse flow

Introduction

Given a set of objects and a pairwise similarity function, a *clustering algorithm* is a formal procedure that groups together objects which are similar and separate the ones which are not (Jain and Dubes, 1988), mimicking the human ability to categorize and group together objects by similarity. Methods and approaches to clustering algorithms have been growing for decades (Jain and Dubes, 1988; Jain et al., 1999; Aggarwal and Reddy, 2013), with clustering becoming a standard data analytic technique (Jain, 2010). This has been complemented by an interest in underlying principles and general desirable properties (sometimes called *axioms*) of clustering algorithms (Fisher and Ness, 1971), especially as clustering is an infamously ill-defined problem in the abstract (Jain, 2010; von Luxburg et al., 2012).

A more recent interest in the axiomatic approach was sparked by Kleinberg's impossibility theorem (Kleinberg, 2003). In the spirit of Arrow's impossibility theorem in social science (Arrow, 1950), Kleinberg gives three natural properties a clustering algorithm should have, namely Scale Invariance, Richness, and Consistency, then proves that they cannot be simultaneously satisfied.

Several authors have since criticised Kleinberg's approach, particularly the Consistency axiom (Ben-David and Ackerman, 2009; Ackerman et al., 2010; Correa-Morris, 2013), and proposed alternative frameworks that circumvent the impossibility result. For instance, by restricting clustering functions to k -partitions, for a fixed k , the axioms can coexist (Zadeh and Ben-David, 2012); if we allow arbitrary parameters, Kleinberg's axioms are compatible when applied to a parametric family of a clustering algorithm, as discussed

in (Correa-Morris, 2013); and, by replacing partitions by dendograms as the output of a clustering function, the authors in (Carlsson and Memoli, 2010) show a possibility and uniqueness result satisfied by single-linkage hierarchical clustering. In all these cases, Kleinberg’s impossibility is avoided by either restricting or extending the definition of clustering function. Other authors shift the axiomatic focus to clustering quality measures (Ben-David and Ackerman, 2009; Laarhoven and Marchiori, 2014; Yu and Xu, 2014), or cost functions (Karayiannis, 1999; Puzicha et al., 2000).

In this article, we remain close to Kleinberg’s original setting and directly address the problematic behaviour of the Consistency axiom instead, which we replace by a weaker condition that we call *Monotonic Consistency*, where the rate of expansion, respectively contraction, of inter-, respectively intra-, cluster distances is not arbitrary, but globally controlled by an expansive function η (Section 1.2). In essence, η controls the inter-cluster expansion, while its inverse η^{-1} controls the intra-cluster contraction. As η is a function on distances, not pairs of points, the control is global, with points at similar distances experiencing the same expansion or contraction. Without this global condition, we recover Outer or Inner Consistency, each incompatible with Scale Invariance and Richness (Ackerman et al., 2010).

Monotonic Consistency avoids the problematic behaviour of Consistency (see Section 1.5), and, moreover, we show a possibility result: Monotonic Consistency, Scale Invariance and Richness are mutually compatible clustering axioms (Corollary 16). As far as we know, this is the only alternative in the literature to the Consistency axiom that is compatible with Richness and Scale Invariance without modifying the definition of clustering function.

Our possibility result depends on a new clustering method called *Morse Clustering*, inspired by Morse Theory in Differential Topology. Although naturally a vertex-weighted clustering algorithm (in the sense of (Ackerman and Ben-David, 2016)), an unweighted version (which we call *agnostic* Morse Clustering) satisfies Kleinberg’s original axioms, with Consistency replaced by Monotonic Consistency.

After briefly introducing Morse theory (Section 2.1), we describe the general Morse flow algorithm (Section 2.2) and explain how it induces a partition on a graph that we call Morse partition (Section 2.3). The Morse flow (and hence the partition) depends on a way of comparing vertices and edges locally (formally, a choice of vertex and edge preorders). Then, at every vertex, the maximal edge (if it exists) is chosen for the flow if it is also *admissible*, that is, if it represents an ‘uphill’ direction. The remaining edges are removed, and what remains is a partition of the graph into trees, each representing a ‘basin of attraction’ of a critical vertex (a sink of the flow).

In contrast to other affinity or hill climbing algorithms, Morse clustering first extracts the maximal edge at a vertex v (if it exists and is unique), and then uses it in the flow only if it is also ascending; otherwise the flow stops at v . This design choice works well in practice (Appendix B), produces a rich family of clustering algorithms (Section 2.4 and Appendix A) and, crucially, allows us to distinguish local maxima (Section 2.2) without introducing additional scaling or threshold parameters.

We present three instances of Morse Clustering, corresponding to three choices of vertex and edge preorders, then show that each of them satisfy a pair of Kleinberg’s original axioms, and that all of them satisfy Monotonic Consistency (Section 2.4). In particular, one of them

satisfy Monotonic Consistency, Scale Invariance and Richness, which are therefore mutually compatible clustering axioms (Corollary 16).

Our last contribution is a generalisation of Kleinberg's axiomatic approach to graph clustering (Section 3). A distance function d on a set X can be represented by a complete graph G with vertex set X and edges weighted by $d(u, v) > 0$. In fact, many clustering algorithms (including Morse Clustering) work on this graph representation. A classical example is Single Linkage, which, in fact, only depends on a minimum spanning tree of G (Gower and Ross, 1969). A natural generalisation of Kleinberg's setting is, therefore, the case when G is an arbitrary, rather than complete, graph. That is, we fix a graph G and consider distances supported on the edge set (this is the natural setting of graph clustering Schaeffer (2007)). In Section 3, we justify this approach, consider Kleinberg's axioms in this graph clustering setting, show that the impossibility result still holds, even when Richness is relaxed naturally to Connected-Richness (partitions where every cluster is a connected subgraph), and give a possibility result for Monotonic Consistency and the same instance of Morse Clustering. As the sparse case (G arbitrary) contains the complete case (G complete), Kleinberg's impossibility theorem (Kleinberg, 2003) is now a particular case of our graph clustering impossibility result (Theorem 20).

In this paper we consider Morse Clustering on edge-weighted graphs with an fixed indexing on the vertices, but the extent of the algorithm is broader. We can apply Morse Clustering to any vertex-annotated graph, where the annotation may come from the graph structure itself (e.g. vertex centrality) or from additional metadata provided by the user. We may call these *unsupervised* respectively *supervised* Morse Clustering, as opposed to the *agnostic* Morse Clustering presented in this article for theoretical and axiomatic purposes only. Morse Clustering, as a practical clustering algorithm of its own right, is studied in (Strazzeri and Sanchez-Garcia) (see also (Strazzeri, 2019), (Schofield et al., 2019) and Appendix B in this article) in the supervised and unsupervised form described above. Note that the use of vertex annotations to improve clustering has been explored elsewhere in the literature such as in (Newman and Clauset, 2016).

1. Monotonic Consistency

In this section, we introduce a weakening of the Consistency axiom that we call *Monotonic Consistency*. We start with a review of Kleinberg's original axioms and the problematic behaviour of Consistency.

1.1 A critique of Kleinberg's axioms

Given a set X of n objects that we want to compare, a *dissimilarity* on X is a pairwise function

$$d : X \times X \rightarrow \mathbb{R}$$

such that $d(i, j) = d(j, i) \geq 0$, and $d(i, j) = 0$ if and only if $i = j$, for all $i, j \in X$. We will adhere to the convention in the literature and refer to d from now on as a *distance*, although it may not satisfy the triangle inequality. Following (Kleinberg, 2003), we define a *clustering algorithm* on X as a map

$$F : \{d \text{ distance on } X\} \rightarrow \{\mathcal{P} \text{ partition of } X\}. \quad (1)$$

A *partition* of X is a disjoint union $X = X_1 \cup \dots \cup X_k$, and we call each X_i a *cluster* of the partition. If $\mathcal{P} = \{X_1, \dots, X_n\}$ is a partition of X and $x, y \in X$, we use the notation $x \sim_{\mathcal{P}} y$ if x and y belong to the same cluster of \mathcal{P} , and $x \not\sim_{\mathcal{P}} y$ if not.

Kleinberg (Kleinberg, 2003) introduced three natural properties for a clustering algorithm, then proved that they cannot be simultaneously satisfied by any clustering algorithm F . These properties are:

- **Scale Invariance:** Given a distance d on X and $\alpha > 0$, we have $F(d) = F(\alpha \cdot d)$;
- **Richness:** Given a partition \mathcal{P} of X , there exists a distance d on X such that $F(d) = \mathcal{P}$;
- **Consistency:** Given two distances d and d' on X with $\mathcal{P} = F(d)$, if d' is a \mathcal{P} -*transformation* of d , that is,

$$\begin{cases} d'(v, w) \leq d(v, w) & \text{if } v \sim_{\mathcal{P}} w, \text{ and} \\ d'(v, w) \geq d(v, w) & \text{if } v \not\sim_{\mathcal{P}} w, \end{cases} \quad (2)$$

then $F(d) = F(d')$.

Kleinberg also showed that each pair of these properties can be simultaneously satisfied, in fact by three different versions of Single Linkage.

Our first contribution is a weakening of the Consistency property which is both very natural, and can coexist with Richness and Scale-Invariance. To motivate our definition, we first discuss the problematic behaviour of Kleinberg's Consistency in the presence of Richness and Scale Invariance (see also (Correa-Morris, 2013; Ackerman et al., 2010; Zadeh and Ben-David, 2012)). Given F a consistent and scale-invariant clustering algorithm, and two different partitions $F(d_1) \neq F(d_2)$, it can be shown (Kleinberg, 2003, Theorem 3.1) that each partition is not the refinement of the other (a partition \mathcal{P} is a *refinement* of \mathcal{Q} if each cluster of \mathcal{P} is contained in a cluster of \mathcal{Q}). In particular, given a distance d and associated partition $\mathcal{P} = F(d)$, we can never obtain a partition identical to \mathcal{P} but with one, or more, of its clusters further subdivided (Fig. 1). On the other hand, consider any distance d' satisfying

$$\begin{cases} d'(v, w) < d(v, w) & \text{if } v, w \in C_1, \\ d'(v, w) < d(v, w) & \text{if } v, w \in C_2, \\ d'(v, w) = d(v, w) & \text{otherwise,} \end{cases}$$

where C is a cluster of \mathcal{P} and $C = C_1 \cup C_2$ is an arbitrary partition of C . Note that any such d' is a \mathcal{P} -transformation of d . This means that we can arbitrarily emphasize the subcluster structure, to the point that it could be more natural to consider C_1 and C_2 as separate clusters (Fig. 1), while Consistency implies $F(d) = F(d')$ regardless.

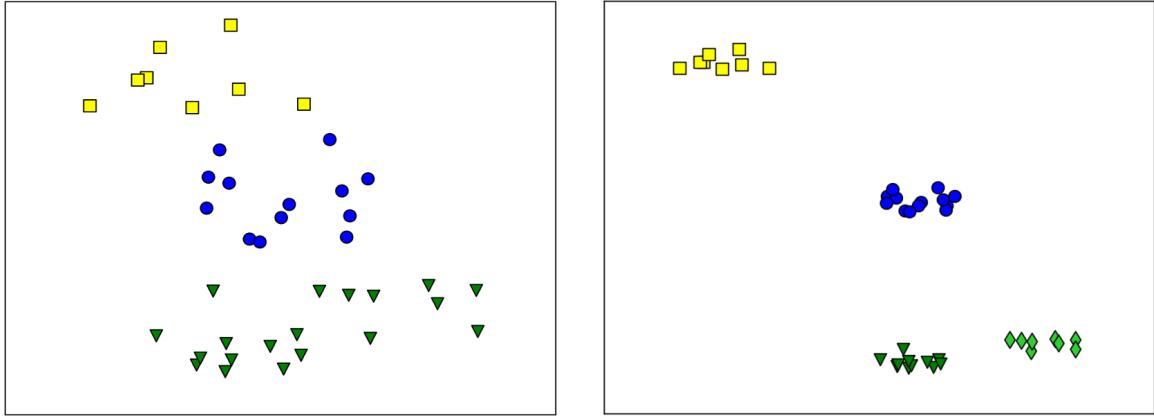


Figure 1: Problematic behaviour of the Consistency axiom. We can arbitrarily emphasize any subcluster structure without affecting the output of the clustering algorithm. This behaviour is explicitly avoided by Monotonic Consistency (Section 1.5).

We propose a more restrictive definition of Consistency which avoids this type of behaviour. The idea is to globally fix the rate at which we can increase (decrease) the intra-cluster (inter-cluster) distances. We do this restricting to \mathcal{P} -transformations obtained through a particular class of functions, which we describe next.

1.2 Expansive and contractive maps

Definition 1 Let X and Y be subsets of \mathbb{R} . We call a continuous map $\eta: X \rightarrow Y$ expansive if

$$|\eta(x) - \eta(y)| \geq |x - y| \quad \text{for all } x, y \in X. \quad (3)$$

By reversing the inequality, we define a contractive map.

Expansive maps can be defined more generally for maps between metric spaces (Gottschalk and Hedlund, 1955) as maps that do not decrease distances between pairs of points, and we have added the continuity hypothesis for convenience (see Remark 3). We will use expansive maps to expand and contract distances with respect to a partition, namely, $d'(u, v) = \eta(d(u, v))$ if u and v belong to different clusters, and $d'(u, v) = \eta^{-1}(d(u, v))$ if they belong to the same cluster. In particular, we take $X = Y = [0, \infty)$ in the definition above, and assume $\eta(0) = 0$. The following lemma summarises some useful properties.

Lemma 2 Let $\eta: [0, \infty) \rightarrow [0, \infty)$ be a continuous expansive map with $\eta(0) = 0$. Then:

- (i) η is strictly increasing, a bijection, and satisfies $\eta(x) \geq x$ for all x ;
- (ii) η^{-1} is strictly increasing, a contractive map, and satisfies $\eta^{-1}(x) \leq x$ for all x .

Proof (i) By contradiction, if η is not strictly increasing, we can find $x > y$ with $\eta(x) \leq \eta(y)$, so that $\eta(0) = 0 \leq \eta(x) \leq \eta(y)$ and, by the Intermediate Value Theorem, we can find $z \in [0, y]$ such that $\eta(z) = \eta(x)$, a contradiction. The growth condition is immediate from the expansion property (3) for $y = 0$,

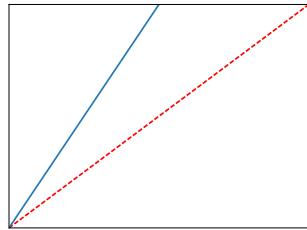
$$|\eta(x)| = \eta(x) \geq |x| = x,$$

for all $x \in [0, \infty)$. Since η is strictly increasing, it is injective. It is also surjective: The growth condition above gives $\eta(x) \rightarrow \infty$ as $x \rightarrow \infty$ and, together with $\eta(0) = 0$ and continuity, we have that η takes all values in $[0, \infty)$.

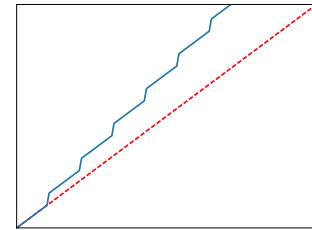
(ii) Since η is bijective, it has an inverse η^{-1} . The inverse of a (strictly) increasing function is also (strictly) increasing. To show this, and the two remaining properties, one can simply use the corresponding properties of η in (i) on $x' = \eta(x)$ and $y' = \eta(y)$. \blacksquare

Example 1 The following are examples of continuous expansive functions $\eta: [0, \infty) \rightarrow [0, \infty)$ with $\eta(0) = 0$.

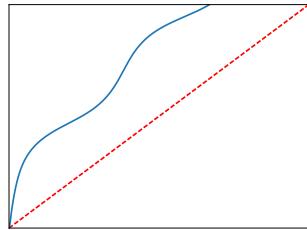
1. (Linear) $\eta(x) = \alpha x$ for $\alpha \geq 1$ (Fig. 2(a)).
2. (Piecewise linear) $\eta = \eta(d_i) + \alpha_i(x - d_i)$ for $x \in [d_i, d_{i+1}]$, where $0 = d_1 < d_2 < \dots < d_n$, $\eta(0) = 0$, and $\alpha_i \geq 1$, for all i (Fig. 2(b)).
3. (Differentiable) A differentiable function $\eta: [0, \infty) \rightarrow [0, \infty)$ with $\eta(0) = 0$ is expansive if and only if $\eta'(x) \geq 1$ for all x (Fig. 2(c)).
4. (Graphical criterion) A continuous function $\eta: [0, \infty) \rightarrow [0, \infty)$ is expansive if and only if the function $\eta(x) - x$ is increasing (this follows from Remark 3).



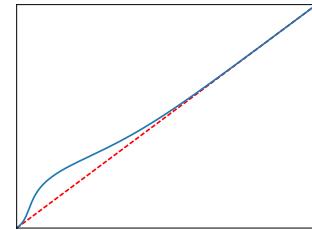
(a) Linear



(b) Piecewise linear



(c) Differentiable



(d) Counterexample

Figure 2: Examples of expansive functions and one counterexample (solid blue lines). At each point, the function grows at least as fast as the line $y = x$ (dashed red line).

Remark 3 If η is increasing, Eq. (3) is equivalent to

$$\eta(x) - \eta(y) \geq x - y \quad \text{for all } x \geq y. \quad (4)$$

In fact, this equation alone implies η increasing and thus Eq. (3). We could drop the continuity hypothesis in Definition 1, and define an expansive function simply by Eq. (4). In practice, however, a monotonic transformation (Definition 4) can always be realised by a continuous, piecewise linear function η (Lemma 6).

1.3 Monotonic transformations

In Kleinberg's original Consistency axiom, arbitrary transformations that increase inter-cluster distances and decrease intra-cluster distances are allowed. To avoid an impossibility result, we restrict to transformations obtained via an expansive function η , as follows. Recall that we write $x \sim_{\mathcal{P}} y$ if x and y are in the same cluster with respect to a partition \mathcal{P} , and $x \not\sim_{\mathcal{P}} y$ if not.

Definition 4 Let d be a distance on a set X , and \mathcal{P} a partition of X . A \mathcal{P} -monotonic transformation of d is any distance d' on X such that

$$\begin{cases} d(x, y) = \eta(d'(x, y)) & \text{if } x \sim_{\mathcal{P}} y, \text{ and} \\ d(x, y) = \eta^{-1}(d'(x, y)) & \text{if } x \not\sim_{\mathcal{P}} y, \end{cases} \quad (5)$$

for some expansive map $\eta: [0, \infty) \rightarrow [0, \infty)$, and all $x, y \in X$. (Note that such η necessarily satisfies $\eta(0) = 0$.)

Definition 5 A clustering algorithm F is Monotonic Consistent if $F(d') = F(d)$ whenever d' is a $F(d)$ -monotonic transformation of d .

Note that, given d and \mathcal{P} , d' is uniquely determined by η . Since $\eta(x) \geq x$ and $\eta^{-1}(x) \leq x$ for all x (Lemma 2), the distance function d' increases inter-cluster distances and decreases intra-cluster distances (hence Consistency implies Monotonic Consistency). However, our allowed transformations do so globally (d' depends on distances between points, not the actual points) and monotonically (the rates at which we expand or contract distances are the inverse of one another). Finally, note that \mathcal{P} -monotonic transformations can be composed and this corresponds to the composition $\eta_2 \circ \eta_1$ of expansive maps.

Example 2 The following are examples of \mathcal{P} -monotonic transformations.

1. (Linear) Let $\eta(x) = \alpha x$, $\alpha \geq 1$. The corresponding \mathcal{P} -monotonic transformation multiplies inter-cluster distances by α , and intra-cluster distances by $1/\alpha$. This is similar to Inner and Outer Consistency, introduced in (Ackerman et al., 2010), except that the expansion and contraction rates are not arbitrary, but the reciprocal of one another.
2. (Linear step function) This is the function

$$\eta(x) = \begin{cases} x & 0 \leq x \leq d_1, \\ \alpha(x - d_1) + d_1 & d_1 \leq x \leq d_2, \\ (x - d_2) + \alpha d_2 & d_2 \leq x, \end{cases} \quad (6)$$

for some $0 \leq d_1 < d_2$ and $\alpha > 1$. The associated \mathcal{P} -monotonic transformation preserves (inter- or intra-cluster) distances below d_1 , scales distances between d_1 and d_2 as in Figure 2b, and (necessarily) translates distances above d_2 , adding $\eta(d_2) = \alpha d_2$ to inter-cluster distances, and subtracting $\eta(d_2)$ to intra-cluster distances. Note that d_2 can be equal to $+\infty$ and so the third line in Equation (6) becomes obsolete.

3. (Piecewise linear) This is generalises both (1) and (2): For the piecewise linear η as in Figure 2c, we have a rate of expansion/contraction α_i , and a translation by $\eta(d_i)$, for distances in the interval $[d_i, d_{i+1}]$ where η is linear. It can be shown that each piecewise linear function is a composition of linear step functions.

Below, we show that every \mathcal{P} -monotonic transformation is induced by a piecewise linear η , or, equivalently, by a finite composition of linear step functions.

1.4 Characterisation of monotonic transformations

Although d' is uniquely determined by η , this η is not unique, that is, different choices of η may result in the same \mathcal{P} -monotonic transformation d' . Indeed, any expansive η interpolating the points $(d'(x, y), d(x, y))$ with $x \sim_{\mathcal{P}} y$ and $(d(x, y), d'(x, y))$ with $x \not\sim_{\mathcal{P}} y$ necessarily gives the same \mathcal{P} -monotonic transformation d' , by Eq. (5). In particular, we can always assume η to be piecewise linear in Definition 4, and, in fact, we can determine whether such function exists directly from d' , as the next result shows.

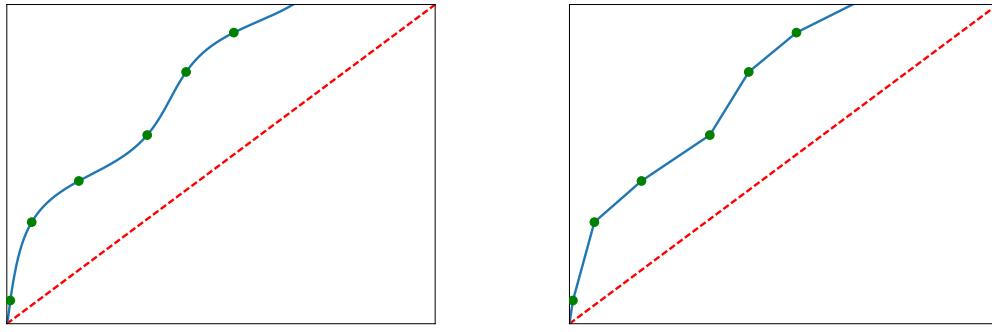


Figure 3: Expansive map (left) and linear interpolation (right) through the points in the subset S (as in Lemma 6). Both maps determine the same \mathcal{P} -monotonic transformation d' of a distance d . In the linear interpolation (right), the slope of each successive segment must be at least 1.

Lemma 6 *Let d and d' be distances on a finite set X and \mathcal{P} a partition of X . Then d' is a \mathcal{P} -monotonic transformation of d if and only if a linear interpolation of the points*

$$S = \{(d(x, y), d'(x, y)) \mid x \sim_{\mathcal{P}} y\} \cup \{(d'(x, y), d(x, y)) \mid x \not\sim_{\mathcal{P}} y\} \subseteq \mathbb{R}^2$$

is a well-defined expansive map $\eta: [0, \infty) \rightarrow [0, \infty)$.

Proof Clearly, if there exists a linear interpolation η of the points in S such that it is a well-defined expansive map, then d' is a \mathcal{P} -monotonic transformation of d , by definition.

Now assume d' is a \mathcal{P} -monotonic transformation of d . Then we can write

$$S = \{(d(x, y), \eta(d(x, y))) \mid x \sim_{\mathcal{P}} y\} \cup \{(d'(x, y), \eta(d'(x, y))) \mid x \not\sim_{\mathcal{P}} y\},$$

where $\eta: [0, \infty) \rightarrow [0, \infty)$ is an expansive map. To define a linear interpolation of S we will assume that S is ordered lexicographically

$$S = \{(x_0, y_0), (x_1, y_1), \dots, (x_N, y_N)\},$$

where $y_i = \eta(x_i)$ for $0 \leq i \leq N$ and $x_i < x_{i+1}$. We can assume the latter since η is injective: if $x_i = x_{i+1}$ then $y_i = y_{i+1}$. Consider now the linear interpolation of S consisting of segments between consecutive pairs of points (x_i, y_i) and (x_{i+1}, y_{i+1}) . As every point in S is of the form $(x, \eta(x))$, we have that the slope of each segment is

$$\frac{\eta(x_{i+1}) - x_{i+1}}{\eta(x_i) - x_i} \geq 1,$$

as η is expansive, Eq. (4). From this we have that the linear interpolation above, effectively a discretization of η , is in fact a well-defined expansive map. \blacksquare

1.5 Avoidance of problematic behaviour

Recall that Kleinberg's Consistency axiom allows us to arbitrarily emphasize any subcluster structure within a cluster without changing the output of the clustering algorithm (Fig. 1). We now show how this problematic behaviour is explicitly avoided by Monotonic Consistency. Suppose that we have a set X and a partition $\mathcal{P} = F(d)$ with respect to a clustering algorithm F and a distance d on X . Choose a cluster C and a partition $C = C_1 \cup C_2$ that we wish to emphasize on a new distance d' which (necessarily) decreases the intra-cluster distances, but in a way that distances within each C_1 and C_2 decrease much faster than distances between C_1 and C_2 , in order to achieve the behaviour depicted in Fig. 1.

Let $u, v \in C_1$ distinct and $w \in C_2$, and call $x = d(u, v)$, $x' = d'(u, v)$, $y = d(u, w)$ and $y' = d'(u, w)$. We impose $x' \leq x$ and $y' \leq y$, and, in addition, we want to make $x - x'$ large while keeping $y - y'$ small (Fig. 4). This is not possible if d' is a \mathcal{P} -monotonic transformation of d , as follows. Let η be an expansive map realising d' . Then $x = \eta(x')$ and $y = \eta(y')$. Assume first $x \leq y$. Then Eq. (4) gives

$$\eta(y') - y' \geq \eta(x') - x' \iff y - y' \geq x - x'. \quad (7)$$

This implies that if we want to reduce the distances inside of a subcluster ($x - x'$ large), we need to reduce the distances between the clusters ($y - y'$) by at least the same amount. The remaining case, $x \geq y$, follows from η^{-1} being a decreasing function (Lemma 2),

$$x \geq y \implies x' = \eta^{-1}(x) \geq \eta^{-1}(y) = y', \quad (8)$$

so that we cannot decrease the intra-cluster distance x without also decreasing the inter-cluster distance y .

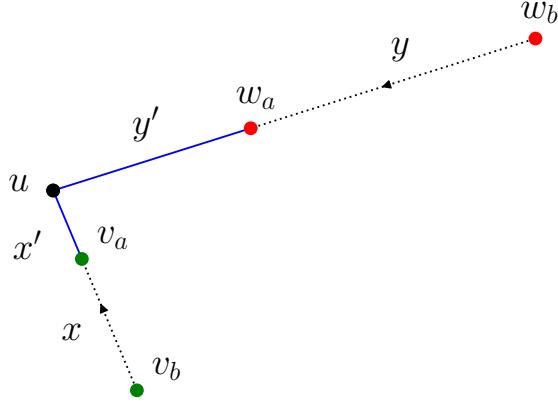


Figure 4: Avoidance of the problematic behaviour by Monotonic Consistency. A \mathcal{P} -monotonic transformation of d reduces the distance from u to v by $x - x'$, and the distance from u to w by $y - y'$ (here shown with subscripts ‘b’ and ‘a’ indicating ‘before’ and ‘after’ the transformation). Then either $x' \geq y'$ (Eq. (8)), or $y - y' \geq x - x'$ (Eq. (7)). In either case, we cannot separate u and v from w within the same cluster.

We finish Section 1 by exploring Monotonic Consistency for Single Linkage, and for metrics.

1.6 Single Linkage does not satisfy Monotonic Consistency

We will show that Monotonic-Consistency, a weakening of Consistency, can be satisfied together with Richness and Scale-Invariance by a particular instance of Morse clustering (Corollary 16). This is in contrast with Single Linkage, which, with different stopping conditions, satisfies each pair of Kleinberg’s axioms (Kleinberg, 2003). The instance of Single-Linkage satisfying Richness and Scale Invariance, namely Scale- α Single Linkage with $0 < \alpha < 1$, does not satisfy Monotonic Consistency, as we show next. Recall that Scale- α Single Linkage applied to (X, d) returns the connected components of the graph with vertex X and edges (i, j) such that $d(i, j) < \alpha \max_{s, t \in X} d(s, t)$.

Lemma 7 *Let $\alpha \in (0, 1)$. Then Scale- α Single-Linkage does not satisfy Monotonic Consistency.*

Proof Let X be any set with at least three points, \mathcal{P} any partition of X with at least two clusters, and $x, y \in X$ such that $x \not\sim_{\mathcal{P}} y$. Define d on X as follows

$$d(u, v) = \begin{cases} \frac{\alpha}{2}, & \text{if } u \sim_{\mathcal{P}} v, \\ 1, & \text{if } u = x, v = y, \\ \alpha, & \text{otherwise.} \end{cases}$$

Let $d_{\max} = \max_{s, t \in X} d(s, t) = 1$. If we represent (X, d) by a complete graph with vertex set X and edges (i, j) , $i \neq j$, weighted by $d(i, j) > 0$, Scale- α Single-Linkage returns the connected component of the graph obtained after removing all edges (i, j) with value

$d(i, j) \geq \alpha d_{\max} = \alpha$, in this case. Consequently, Scale- α Single-Linkage applied to d returns the original partition \mathcal{P} .

Let d' be the \mathcal{P} -monotonic transformation of d given by

$$\eta(x) = \frac{x^2 + x}{\alpha}.$$

(Note that $\eta(0) = 0$ and $\eta'(x) = \frac{2x+1}{\alpha} > 1$ for all x , so η is indeed expansive.) Then

$$d'(u, v) = \begin{cases} \eta^{-1}\left(\frac{\alpha}{2}\right) = \frac{-1 + \sqrt{1 + 2\alpha^2}}{2}, & \text{if } u \sim_{\mathcal{P}} v, \\ \eta(1) = \frac{2}{\alpha}, & \text{if } u = x, v = y, \\ \eta(\alpha) = 1 + \alpha, & \text{otherwise.} \end{cases}$$

We now have $d'_{\max} = \eta(1) = \frac{2}{\alpha}$ and thus scale- α Single-Linkage removes the edges (i, j) with $d(i, j) \geq \alpha d'_{\max} = 2$. Since $\alpha < 1$, the only removed edge is $d(x, y)$ and, since X has at least three points, the algorithm returns the trivial partition $\{X\}$, clearly not \mathcal{P} . \blacksquare

1.7 Monotonic Consistency for metrics

A *metric* is a distance (in the sense of this article) which also satisfies the triangle inequality, $d(u, w) \leq d(u, v) + d(v, w)$ for all u, v, w . Metrics arise naturally when X is embedded in a metric space such as \mathbb{R}^m , and, in fact, for many clustering algorithms (for example k -means clustering), the distance function is always a metric. It is therefore natural to ask whether Monotonic Consistency is a useful property in this context, namely, whether a non-trivial (that is, η not the identity) \mathcal{P} -monotonic transformation of a metric can be a metric. (If not, Monotonic Consistency would become an empty axiom for metrics.) Of course, not every \mathcal{P} -monotonic transformation of a metric will be a metric, but we show below that, given a metric d and an arbitrary partition \mathcal{P} , we can always find \mathcal{P} -monotonic transformations of d which are metrics.

Given a distance d on a set X , we call a triple of points $i, j, k \in X$ *aligned* if they are distinct and $d(i, k) = d(i, j) + d(j, k)$.

Theorem 8 *Let X be a set, \mathcal{P} a partition of X , and d a distance on X such that no triple of nodes is aligned. Then there exists a constant $c(d, \mathcal{P}) > 1$ such that, for all $s \in [1, c(d, \mathcal{P}))$, the \mathcal{P} -monotonic transformation of d given by $\eta(x) = sx$ is a metric. Moreover, there is a universal constant $c(d)$ independent of the partition, that is, $1 < c(d) \leq c(d, \mathcal{P})$ for all partitions \mathcal{P} of X .*

Proof Let d' be the \mathcal{P} -monotonic transformation of d given by $\eta(x) = sx$ for some $s \geq 1$. We will find conditions on s to guarantee that d' satisfies the triangle inequality. Let $i, j, k \in X$ distinct (if not, the triangle inequality is automatically satisfied). We want to show that

$$d(i, k) \leq d(i, j) + d(j, k) \implies d'(i, k) \leq d'(i, j) + d'(j, k).$$

Recall that

$$d'(i, j) = \begin{cases} s d(i, j) & \text{if } i \sim_{\mathcal{P}} j, \\ \frac{d(i, j)}{s} & \text{otherwise.} \end{cases}$$

If i, j and k are in the same cluster then clearly

$$\frac{d(i, k)}{s} \leq \frac{d(i, j)}{s} + \frac{d(j, k)}{s}.$$

If they are all in pairwise different clusters, then

$$s d(i, k) \leq s d(i, j) + s d(j, k).$$

If i and k are in the same cluster but j is not, then (recall $s \geq 1$)

$$\frac{d(i, k)}{s} \leq d(i, k) \leq s d(i, j) + s d(j, k).$$

Since i and k are interchangeable in the triangle inequality above, the only remaining case is when i and j are in the same cluster, but k is not. In this case, we want to show that

$$s d(i, k) \leq \frac{d(i, j)}{s} + s d(j, k). \quad (9)$$

If $d(i, k) \leq d(j, k)$ then $s d(i, k) \leq s d(j, k)$ and Eq. (9) is automatically satisfied. If $d(i, k) > d(j, k)$, Eq. (9) is satisfied if and only if

$$s^2 (d(i, k) - d(j, k)) \leq d(i, j) \iff s \leq \sqrt{\frac{d(i, j)}{d(i, k) - d(j, k)}}.$$

Define

$$\begin{aligned} c(d, \mathcal{P}) &= \min_{\substack{i \sim_{\mathcal{P}} j, i \not\sim_{\mathcal{P}} k \\ d(i, k) > d(j, k)}} \sqrt{\frac{d(i, j)}{d(i, k) - d(j, k)}} \quad \text{and} \\ c(d) &= \min_{d(i, k) > d(j, k)} \sqrt{\frac{d(i, j)}{d(i, k) - d(j, k)}}. \end{aligned}$$

Clearly, $c(d) \leq c(d, \mathcal{P})$ for all partitions \mathcal{P} . To finish the proof, note that the triangle inequality for d guarantees $c(d) \geq 1$, and $c(d) = 1$ if and only if there is an aligned triple of points. \blacksquare

Defining the minimum of an empty set as infinity, we might have $c(d, \mathcal{P}) = \infty$ (or $c(d) = \infty$), meaning that the \mathcal{P} -monotonic transformation of d given by $\eta(x) = sx$ is a metric for any $s \geq 1$, and Theorem 8 still holds. Of course, this would only occur if for all i, j, k with $i \sim_{\mathcal{P}} j$ and $i \not\sim_{\mathcal{P}} k$, we have $d(i, k) = d(j, k)$.

2. Morse Clustering

In this section we consider the clustering algorithm **Morse** in the form of three variants: **SiR-Morse**, **k -Morse** and **δ -Morse** (the last two are described in full detail in Appendix A). Each of them satisfy one pair of the original Kleinberg axioms, and all of them satisfy Monotonic Consistency. In particular, one of them (**SiR-Morse**) satisfies Scale Invariance and Richness, showing that our three axioms can be simultaneously satisfied (Corollary 16). Morse clustering is inspired by Topology and Differential Geometry, namely Morse theory (Milnor, 1963) and its discretisation due to Forman (Forman, 1998). We start with a brief introduction to both continuous and discrete Morse theory and explain how they motivate our clustering algorithm.

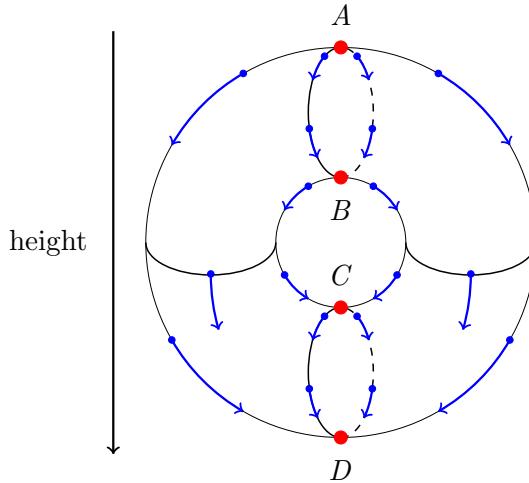


Figure 5: Morse function (vertical height) on a torus, critical points (red), and associated flow (blue). The flow represents a unique maximal descent (or ascend, if we reverse time) path that a particle, such as a drop of water, would follow on the surface. It is defined everywhere except at four critical points, which can be thought of as a flow source (A), sink (D) or a combination of both (B, C). The number and type of critical points, for any Morse function, is a topological invariant of the torus.

2.1 Morse theory

Topology is the mathematical study of ‘shape’ (Prasolov, 1995). It considers properties of a *space* (such as a 2D surface, or 3D object) which are invariant under *continuous deformations* such as stretching, bending or collapsing. A *topological invariant* is a property, for example whether the space is disconnected, which is invariant under such deformations. A standard approach in Topology is to study a space via functions defined on the space. *Morse theory* (Milnor, 1963) considers potential-like functions called *Morse functions* and their associated *flow* on the space, defined by the unique direction of maximal descent at every point, except at a few *critical points* (see Fig. 5).

Forman (1998) introduced a discrete version of Morse Theory which applies to discretisations of continuous spaces, such as a polygonal mesh of a continuous surface. Such discretisation decomposes the space into vertices, edges, triangles, etc. called *simplices*. A *discrete Morse function* assigns a real number to each simplex under certain combinatorial restrictions, and we have associated notions of *critical simplex*, and *discrete Morse flow* (Fig. 6).

Discrete Morse theory can be applied to clustering by representing a set X with distance d as an undirected weighted graph G with vertex set X , and an edge between i and j if $d(i, j) > 0$, and no such edge otherwise. (This is an all-to-all, or complete, graph.) A graph is a discretisation of a curve and hence discrete Morse theory applies. To obtain a partition of X using Morse theory, first we extend the edge weights given by the distances $d(i, j) > 0$ to a Morse function on the graph by assigning weights to the vertices as well. This Morse function determines a unique flow on the vertices which, in turn, gives a natural partition of

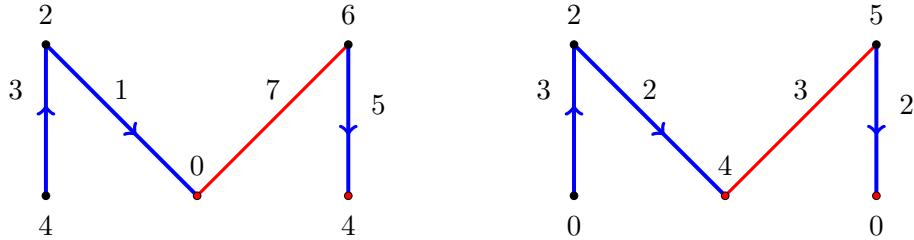


Figure 6: Morse flows on the same discrete space (here a small graph) associated to a discrete Morse function (left) and to arbitrary edge and vertex weight (right). Our algorithm (Algorithm 1) generalises the standard construction (Forman, 1998) of the Morse flow associated to a discrete Morse function (left) to the Morse flow of an arbitrary edge and vertex weighted on a graph (right). In both cases, we show the associated discrete Morse flow as blue directed edges and critical simplices (vertices and edges) in red. A particle on a vertex has a unique direction of descent following the blue arrow, except at the two critical vertices shown in red, both sinks of the flow. After removing the critical edge, we have two connected components, each a tree rooted at a critical vertex (Algorithm 2).

the vertex set. The clusters are the connected components of the graph after removing the critical edges (edges not participating in the flow), and each cluster becomes a tree rooted at a critical vertex (a sink of the flow), see Fig. 6. We describe this in detail next.

2.2 Morse flow

Let X be a finite set and d a distance (dissimilarity) on X . The Morse clustering of (X, d) is obtained from the Morse flow on the graph representation of (X, d) , by removing the edges not participating in the flow. In turn, the Morse flow is determined by the direction of maximal descent at every vertex together with the initial and final vertex weights (to guarantee a descending path). In its more general form, rather than weights, we only need a way of comparing vertices and edges locally. Formally, this consists on a choice of vertex and edge preorders.

A *preorder* on a set is a binary relation \preceq that is reflexive ($a \preceq a$ for all a) and transitive ($a \preceq b$ and $b \preceq c$ implies $a \preceq c$ for all a, b, c). We write $a \prec b$ if $a \preceq b$ and $b \not\preceq a$ (that is, $b \preceq a$ does not hold). A preorder is *total* if $a \preceq b$ or $b \preceq a$ for all a, b . Our main examples are the total preorders induced by an edge or vertex weight function on a graph (Example 3). By a *graph* $\mathcal{G} = (V, E)$ we mean a non-empty vertex set V and an edge set $E \subseteq V \times V$ so that $(u, v) \in E$ represents a directed edge from u to v . A graph is *undirected* if $(v, w) \in E$ whenever $(w, v) \in E$, for all $v, w \in V$, *loopless* if $(v, v) \notin E$ for all $v \in V$, and *finite* if V (and therefore E) is a finite set.

Example 3 Let $\mathcal{G} = (V, E)$ be a graph.

- (1) (*Edge weights*) For any function $w: E \rightarrow \mathbb{R}$, the relation $e \preceq f$ if $w(e) \leq w(f)$ is a total preorder on E .
- (2) (*Vertex weights*) For any function $w: V \rightarrow \mathbb{R}$, the relation $u \preceq v$ if $w(u) \leq w(v)$ is a total preorder on V .

A distance d on a set X is an edge weight function for the complete graph with vertex set X , and hence induces a total edge preorder on the graph representation of X . Similarly, a labelling $X = \{x_1, x_2, \dots, x_n\}$ induces a vertex weight $w(x_i) = i$ and hence a total preorder on the vertices $V = X$ of such graph representation.

Remark 9 *A preorder is an order if it is also anti-symmetric ($a \preceq b$ and $b \preceq a$ implies $a = b$). Our examples above are not necessarily orders, as we may have $w(a) = w(b)$ with $a \neq b$. If \preceq is a total order, $a \prec b$ is equivalent to $a \preceq b$ and $a \neq b$. Note that any total preorder on a set is induced by a weight function $w: X \rightarrow \mathbb{N}$.*

Morse clustering applies to an arbitrary finite graph \mathcal{G} with a choice of edge and vertex preorders \preceq_E and \preceq_V . First, it finds the (ascending) Morse flow associated to $(\mathcal{G}, \preceq_E, \preceq_V)$ (Algorithm 1), then the vertex partition associated to the Morse flow, that is, the connected components of the graph after removing the critical edges (Algorithm 2); see also Fig. 6. First, we need to introduce some notation and terminology.

Given a node v , we define the set of *edges at v* as

$$E_v = \{(v, w) \in E\}.$$

A *maximal edge at v* is a maximum for E_v with respect to the edge preorder, that is, an edge $e \in E_v$ such that $f \preceq e$ for all $f \in E_v$. Note that a maximal edge at v may not exist (e.g. if the preorder is not total), or it may not be unique (e.g. if $w(e) = w(f)$ for some edge weights). An edge (v, w) is *ascending*, or *admissible*, if $v \prec_V w$. If an edge is the unique maximal edge at v , and it is ascending, we call it a *non-critical* or *flow edge at v* .

We define the *Morse flow* of a graph \mathcal{G} with a choice of preorders \preceq_E and \preceq_V as the map $\Phi: V \rightarrow V$ given by

$$\Phi(v) = \begin{cases} w & \text{if } (v, w) \text{ is the unique maximal edge at } v, \text{ and it is ascending } v \prec_V w, \\ v & \text{otherwise.} \end{cases}.$$

Remark 10 *There is a similar notion of descending edges and Morse flow. For simplicity, we define Morse flow as ascending, and achieve descending flows simply by reversing the vertex preorder \preceq_V .*

Input: graph $\mathcal{G} = (V, E)$, edge preorder \preceq_E , vertex preorder \preceq_V
Output: Morse flow $\Phi: V \rightarrow V$

```

foreach  $v \in V$  do
    if maximal edge  $e_v = (v, w)$  of  $E_v$  exists and it is unique and  $v \prec_V w$  then
         $\Phi(v) = w$ 
    else
         $\Phi(v) = v$ 
    end
end

```

Algorithm 1: Morse flow algorithm.

Graphically, we can represent a Morse flow as directed edges (v, w) whenever $\Phi(v) = w$ (blue directed edges in Fig. 6). Edges not participating in the flow (red edges in Fig. 6) are

called *critical edges*, and fixed points of the flow, $\Phi(v) = v$, are called *critical vertices* (red vertices in Fig. 6), and correspond to ‘sinks’ of the flow.

The Morse flow of a graph can be computed using Algorithm 1. This algorithm can be easily implemented so that each edge is visited only once, and thus has linear time complexity $\mathcal{O}(m)$ where m is the number of edges.

An important observation is that we *first* use the edge preorder to extract the maximal edge at a vertex (if it exists and is unique), and *then* use it in the flow only if it is also ascending. In particular, if the maximal edge does not exist, or it is not unique, or, crucially, it is not ascending, we define $\Phi(v) = v$, that is, the flow stops at v (Fig. 7). This design choice works well in practice (Appendix B), produces a rich family of clustering algorithms (Section 2.4 and Appendix A) and, crucially, allows us to distinguish local maxima (Fig. 8) without introducing additional scaling/threshold parameters.

Note that the reverse procedure (evaluating the vertex preorder first) gives a ‘hill climbing’ or ‘affinity’ algorithm: among all ascending edges, we choose the maximal (‘closest’) one. The difference, although subtle, is important: The ‘rock climbing’ algorithm, in contrast with the Morse flow, fails to detect local maxima if they are close (Fig. 8).

Note that the Morse flow algorithm work for both directed and undirected networks. For undirected networks (the case we are concern with in this paper), each undirected edge $\{(v, w), (w, v)\}$ is evaluated twice, once at v and once at w . Since at most one of (v, w) or (w, v) is admissible (ascending), at most one of them belongs to the flow. The fact that the flow is (strictly) ascending, also means that there cannot be any cycles in the flow.

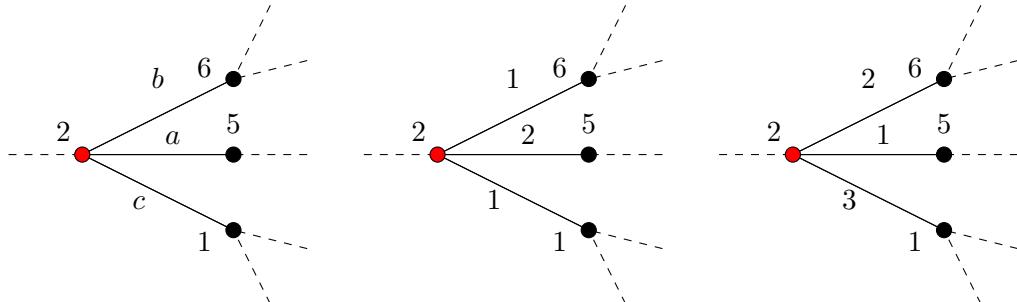


Figure 7: Three ways in which critical vertices arise. Left: We suppose the preorder $a \preceq_V b$ and $a \preceq_V c$ with b and c not related, and hence there is no maximal edge (partial preorder). Middle: The maximal edge is not unique (the order is total but there is a tie). Right: the maximal edge (labelled 3) is not admissible. In all cases, the vertex labelled 2, shown in red, is critical. The solid edges might still be part of the flow (directed towards the critical vertex) depending on the weights in the dashed edges.

Recall that an edge is critical if it does not participate in the flow, and a vertex is critical if it is a fixed point, $\Phi(v) = v$, that is, a ‘sink’ of the flow. Formally, we define

$$V_{\text{crit}} = \{v \in V \mid \Phi(v) = v\} \quad \text{and} \quad E_{\text{crit}} = \{(v, w) \in E \mid \Phi(v) \neq w\}.$$

We will see that, after removing the critical edges, what remains is a partition of the graph into a disjoint union of directed trees rooted at critical vertices (edge directions given

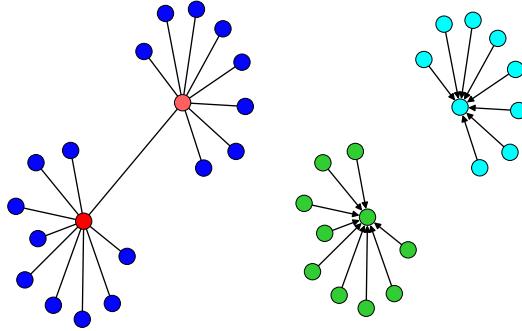


Figure 8: Morse clustering can separate nearby local maxima. Left: Toy graph with node weights shown by colour from low (dark blue) to high (red, critical vertex). Right: Output of the Morse flow and clustering algorithm (two clusters, shown by colour). The closest (highest weight) edge at each red vertex is a ‘downhill’ edge to a blue vertex. Hence both vertices, and the edge between them, are critical.

by the flow). The cluster associated to a critical vertex v is

$$T_v = \{w \in V \mid \Phi^N(w) = v \text{ for some } N \geq 0\}. \quad (10)$$

Here Φ^N is the composition of Φ with itself N times (and Φ^0 is the identity map), so that $\Phi^N(w)$ is the vertex at which we arrive from w after following the flow N steps (across N edges). In the dynamical system terminology, we can describe each T_v as the ‘basin of attraction’ of v .

Let us write \mathcal{T}_v for the subgraph with vertex set T_v and edge set all non-critical edges between vertices in T_v . Recall that the *depth* of a rooted tree is the maximal distance to its root.

Theorem 11 *Let $\mathcal{G} = (V, E)$ be a finite graph with edge and vertex preorders \preceq_E and \preceq_V , and associated Morse flow $\Phi: V \rightarrow V$. Then:*

- (i) *The Morse flow stabilises, that is, there is $N \geq 0$ such that $\Phi^N = \Phi^{N+1}$;*
- (ii) *$\{T_v \mid v \in V_{\text{crit}}\}$ is a partition of V ;*
- (iii) *\mathcal{T}_v is a directed (edge directions given by the flow) rooted tree with root v ;*
- (iv) *Within T_v , the vertex v is the only critical vertex, and it is maximal with respect to the vertex preorder;*
- (v) $\max\{\text{depth}(T_v) \mid v \in V_{\text{crit}}\} = \min\{N \geq 0 \mid \Phi^N = \Phi^{N+1}\};$
- (vi) *The graph $(V, E \setminus E_{\text{crit}})$ equals the disjoint union of the graphs \mathcal{T}_v for $v \in V_{\text{crit}}$.*

Proof (i) Let $v \in V$. By the definition of the Morse flow, either $\Phi(v) = v$ (a critical vertex), or $\Phi(v) = w$ and $v \prec_V w$ (which implies $v \neq w$ by reflexivity). Therefore, the

sequence $v = v_0, v_1, v_2, \dots$ where $v_i = \Phi^i(v)$, must contain a critical vertex before the first repetition: otherwise, we would have $v_i \prec_V v_{i+1} \prec_V \dots \prec_V v_k = v_i$ and thus $v_i \prec_V v_i$ by transitivity, a contradiction. Since the graph is finite, say $|V| = n$, there will be repetition in any subset of $n + 1$ vertices. Consequently, there is a critical vertex $v_k = \Phi^k(v)$ in the sequence above and, in fact, $k \leq n$. All in all, the flow stabilises after at most $n = |V|$ steps. (The case $N = 0$ can occur if all vertices are critical.)

(ii) Let $v \in V$. By the argument in (i), the sequence $v_i = \Phi^i(v)$ ($i \geq 0$) stabilises, that is, there is $k \geq 0$ such that $\{v_i \mid 0 \leq i \leq k\}$ are distinct, non-critical, and $v_j = v_k$ critical for all $j \geq k$. In particular, $v \in T_{v_k}$, by Eq. (10). This shows that every vertex belong to a set T_v for $v \in V_{\text{crit}}$, and that these sets must be disjoint.

(iii) Since all edges in \mathcal{T}_v are non-critical, we have $v \prec_V w$ across each edge and thus a cycle would imply $u \prec_V u$ for some vertex u , a contradiction. All edges are directed and point towards the root v , by the discussion above.

(iv) For each critical vertex w , we have $\Phi(w) = w$ hence $w \in T_w$. Since, by (ii), they form a partition of the vertex set, v is the only critical vertex in T_v . Every (directed) edge (u, w) in \mathcal{T}_v is not critical, hence admissible, so that $u \prec_V w$. As v is the root of the tree \mathcal{T}_v , it must then be maximal with respect to \preceq_V .

(v) It suffices to show that, for any $v \in V_{\text{crit}}$, and any $N \geq 0$ such that $\Phi^N = \Phi^{N+1}$, we have $\text{depth}(\mathcal{T}_v) \leq N$. Let $k = \text{depth}(\mathcal{T}_v)$. Then there is $w \in T_v$ such that $w_i = \Phi^i(w)$, $i \geq 0$, stabilises after exactly k steps. In particular, w_0, \dots, w_{k-1} are all distinct and hence $k \leq N$.

(vi) Let $\mathcal{G}_1 = (V, E \setminus E_{\text{crit}})$ and \mathcal{G}_2 the disjoint union of the graphs \mathcal{T}_v for $v \in V_{\text{crit}}$. Since $\{T_v \mid v \in V_{\text{crit}}\}$ is a partition of V , both \mathcal{G}_1 and \mathcal{G}_2 have the same vertex set. We show they also have the same edge set and hence they are equal. The edges in \mathcal{G}_2 are non-critical thus a subset of $E \setminus E_{\text{crit}}$. Conversely, given a non-critical edge (v, w) in \mathcal{G}_1 , we have $w = \Phi(v)$ and the sequence $v, \Phi(v) = w, \Phi^2(v), \dots$ shows that v and w belong to the same tree critical tree, and thus this tree contains the edge (v, w) . ■

2.3 Morse clustering algorithm

The *Morse partition* of a graph \mathcal{G} with a choice of vertex and edge preorders \preceq_V and \preceq_E is the partition of the vertex set given by the connected components of the graph $\mathcal{G}_{\text{Morse}} = (V, E \setminus E_{\text{crit}})$. By Theorem 11, there is a cluster for each critical vertex, and, in fact, $\mathcal{G}_{\text{Morse}}$ is a disjoint union of directed rooted trees with roots at the critical vertices.

A complete algorithm that returns the Morse clustering of $(\mathcal{G}, \preceq_V, \preceq_E)$ is given below (Algorithm 2). Its time complexity is clearly linear on the number of vertices and edges. Alternatively, the Morse flow and clustering can be computed simultaneously one edge at a time, by keeping a list of critical edges and of the maximal edge at each vertex. Therefore,

the time complexity of (any instance of) Morse clustering is $\mathcal{O}(m)$ where m is the number of edges of the graph.

Input: graph $\mathcal{G} = (V, E)$, edge preorder \preceq_E , vertex preorder \preceq_V

Output: partition \mathcal{P} of V

$n \leftarrow |V|$

initialise G_{Morse} to an empty graph on n vertices

$\Phi \leftarrow \text{MorseFlow}(\mathcal{G}, \preceq_E, \preceq_V)$

foreach $v \in V$ **do**

if $\Phi(v) \neq v$ **then**

 | add edge $(v, \Phi(v))$ to G_{Morse}

end

end

$\mathcal{P} \leftarrow$ connected components of G_{Morse}

Algorithm 2: Morse clustering algorithm.

We finish with a useful result to determine when two Morse partitions are equal.

Lemma 12 *Let Φ and Φ' be Morse flows on X with associated Morse partitions \mathcal{P} and \mathcal{P}' . If $x \sim_{\mathcal{P}} \Phi'(x)$ for all $x \in X$, then \mathcal{P}' is a refinement of \mathcal{P} .*

Proof Write $\mathcal{P} = \{X_1, \dots, X_n\}$ and $\mathcal{P}' = \{X'_1, \dots, X'_{n'}\}$. Write x_i , respectively x'_j , for the critical vertex in X_i , respectively X'_j , for all i, j . Choose $N \geq 1$ such that both Φ and Φ' stabilise, that is, $\Phi^N = \Phi^{N+1}$ and $(\Phi')^N = (\Phi')^{N+1}$. We need to show that, for each j there is i such that $X'_j \subseteq X_i$.

Let $x \in X'_j$ and consider the flow paths

$$\begin{aligned} p(x) &= \{x, \Phi(x), \dots, \Phi^N(x) = x_i\} \text{ and} \\ p'(x) &= \{x, \Phi'(x), \dots, (\Phi')^N(x) = x'_j\} \end{aligned}$$

By definition of Morse partition, all points in $p(x)$ are in the same cluster of \mathcal{P} , namely X_i , and all points in $p'(x)$ in the same cluster of \mathcal{P}' , namely X'_j . By hypothesis, $(\Phi')^n(x) \sim_{\mathcal{P}} (\Phi')^{n+1}(x)$ for all $n \geq 0$, so $p'(x) \subseteq X_i$. In particular $x'_j \sim_{\mathcal{P}} x_i$.

Given any other $y \in X'_j$,

$$\begin{aligned} p(y) &= \{y, \Phi(y), \dots, \Phi^N(y) = x_k\} \subseteq X_k \text{ and} \\ p'(y) &= \{y, \Phi'(y), \dots, (\Phi')^N(y) = x'_j\} \subseteq X'_j, \end{aligned}$$

for a possibly different cluster X_k . Again, by hypothesis, we have $p'(y) \subseteq X_k$ and, in particular, $x'_j \sim_{\mathcal{P}} x_k$. Then $x'_j \in X_i \cap X_k \neq \emptyset$ and hence $i = k$, as distinct clusters are disjoint. Since y was arbitrary, we conclude that $X'_j \subseteq X_i$. \blacksquare

Note that two partitions are equal if and only if each is the refinement of the other, or if they have the same size (number of clusters) and one is the refinement of the other.

2.4 A possibility theorem for Monotonic Consistency

Different choices of edge and vertex preorders result in different instances of Morse Clustering. We now show an instance of Morse Clustering that satisfies Scale-Invariance and Richness (Theorem 14) as well as Monotonic Consistency (Theorem 15), proving that these three axioms are mutually compatible.

Remark 13 *One can in fact define two further instances of Morse Clustering, that we call k -Morse and δ -Morse, which satisfy the other two pairs of Kleinberg’s axioms, namely Consistency and Scale-Invariance (k -Morse) and Consistency and Richness (δ -Morse). Furthermore, they both satisfy Monotonic Consistency. Full details are given in Appendix A.*

Let (X, d) be a set with a distance function, and consider the complete graph with vertex set X . Let us fix, once and for all, a labelling $X = \{x_1, x_2, \dots, x_n\}$, which we will use to create the vertex preorders (see the remarks at the end of this section on labelling). We also assume that X has at least three points.

Our instance of Morse Clustering is called **SiR-Morse** (Scale-invariant and Rich), and corresponds to the choices of vertex and edge preorders given below.

SiR-Morse

- $v_i \preceq_V v_j$ if $i \leq j$
- $(v, w) \preceq_E (v, t)$ if $d(v, w) \geq d(v, t)$

Note that the vertex preorder is a total order, and the edge preorder is also locally total (at each vertex). The corresponding Morse flow chooses, at each vertex v , the edge with smallest distance, if it is unique and admissible. On the other hand, if more than one edge at v achieves the smallest distance, or if such edge is not admissible, then v is critical, that is, the Morse flow fixes v , $\Phi(v) = v$.

Theorem 14 *SiR-Morse is Scale-Invariant and Rich.*

Proof (Scale-invariance) Scale-Invariance does not affect the vertex or edge preorders, since \preceq_V is independent of d and, for \preceq_E , $d(v, w) \leq d(v, t)$ if and only if $\alpha d(v, w) \leq \alpha d(v, t)$ for all $\alpha > 0$. Hence the output of **SiR-Morse** for (X, d) and for $(X, \alpha d)$ are the same.

(Richness) Consider $V = V_1 \cup \dots \cup V_k$ an arbitrary partition of V . Let v_i be the maximal vertex in V_i (\preceq_V is a total order) and define a distance d as follows

$$d(v, w) = \begin{cases} 1, & \text{if } v, w \in V_i \text{ for some } i, \text{ and either } v = v_i \text{ or } w = v_i, \\ 2, & \text{otherwise,} \end{cases}$$

for all $v \neq w$. If $v \in V_i$, the edge to v_i is always admissible and the largest with respect to \preceq_E , so $\Phi(v) = v_i$ for the Morse flow, and we recover the partition $V_1 \cup \dots \cup V_k$. \blacksquare

Theorem 15 *SiR-Morse satisfies Monotonic Consistency.*

Proof Let d be a distance on X , \mathcal{P} the output partition of **SiR-Morse** on (X, d) , and d' a \mathcal{P} -monotonic transformation of d . We want to show that **SiR-Morse** produces the same partition on (X, d') . We will prove that, in fact, the associated Morse flows Φ and Φ' are identical.

Let η be a monotonic transformation realising d' , that is,

$$\begin{aligned} d(u, v) &= \eta(d'(u, v)) && \text{if } u \sim_{\mathcal{P}} v, \text{ and} \\ d(u, v) &= \eta^{-1}(d'(u, v)) && \text{if } u \not\sim_{\mathcal{P}} v. \end{aligned}$$

Let $v \in X$ and consider first the case $w = \Phi(v) \neq v$. Then, by the definition of **SiR-Morse** preorders,

$$d(v, w) < d(v, s) \text{ for all } s \neq v, w.$$

To prove that $\Phi'(v) = w$, we need to show that $d'(v, w) < d'(v, s)$ for all $s \neq v, w$. We have two subcases.

1. If $s \sim_{\mathcal{P}} v$, we have $d'(v, w) = \eta^{-1}(d(v, w))$ and $d'(v, s) = \eta^{-1}(d(v, s))$, so

$$d(v, w) < d(v, s) \text{ implies } d'(v, w) < d'(v, s),$$

as η^{-1} is increasing (Lemma 2).

2. If $s \not\sim_{\mathcal{P}} v$, we have $d'(v, w) = \eta^{-1}(d(v, w))$ and $d'(v, s) = \eta(d(v, s))$, so

$$d(v, w) < d(v, s) \text{ implies } d'(v, w) \leq d(v, w) < d(v, s) \leq d'(v, s),$$

as $\eta^{-1}(x) \leq x \leq \eta(x)$ for all x (Lemma 2).

In conclusion, we have $d'(v, w) < d'(v, s)$ for all $s \neq v, w$ so $\Phi'(v) = w$.

The remaining case is $\Phi(v) = v$. Suppose, by contradiction, that $w = \Phi'(v) \neq v$. This implies $v \prec_V w$ and $d'(v, w) < d'(v, s)$ for all $s \neq v, w$. Note that, since v is critical and therefore maximal within its cluster, we have $v \not\sim_{\mathcal{P}} w$. On the other hand, $\Phi(v) = v$ means that either the unique maximal edge is not admissible, or it is admissible but the maximum is not unique.

First we show that $d(v, w)$ is also a minimal distance at v (possibly not unique). Suppose, by contradiction, $d(v, s) < d(v, w)$ for some $s \neq v, w$. There are two subcases.

1. If $s \sim_{\mathcal{P}} v$, then we have $d'(v, s) = \eta^{-1}(d(v, s))$ and $d'(v, w) = \eta(d(v, w))$, so

$$d(v, s) < d(v, w) \text{ implies } d'(v, s) \leq d(v, s) < d(v, w) \leq d'(v, w),$$

as $\eta^{-1}(x) \leq x \leq \eta(x)$ (Lemma 2).

2. If $s \not\sim_{\mathcal{P}} v$, then we have $d'(v, s) = \eta(d(v, s))$ and $d'(v, w) = \eta(d(v, w))$, so

$$d(v, s) < d(v, w) \text{ implies } d'(v, s) < d'(v, w),$$

as η is increasing (Lemma 2).

	Scale-Invariance	Richness	Consistency	Monotonic-Consistency
SiR-Morse	✓	✓	✗	✓
k-Morse	✓	✗	✓	✓
δ-Morse	✗	✓	✓	✓

Table 1: Clustering axioms and three instances of Morse clustering.

In either case, we have $d'(v, s) < d'(v, w)$, a contradiction to the minimality of $d'(v, w)$.

Since $d(v, w)$ is a minimal distance and $v \prec_V w$, but $\Phi(v) = v \neq w$, the minimal distance (maximal edge) cannot be unique. Let $d(v, s) = d(v, w)$ for some $s \neq v, w$. We have, again, two subcases.

1. If $s \sim_P v$, then we have $d'(v, s) = \eta^{-1}(d(v, s))$ and $d'(v, w) = \eta(d(v, w))$, so

$$d(v, s) = d(v, w) \text{ implies } d'(v, s) \leq d(v, s) = d(v, w) \leq d'(v, w),$$

as $\eta^{-1}(x) \leq x \leq \eta(x)$ (Lemma 2).

2. If $s \not\sim_P v$, then we have $d'(v, s) = \eta(d(v, s))$ and $d'(v, w) = \eta(d(v, w))$, so

$$d(v, s) = d(v, w) \text{ implies } d'(v, s) = d'(v, w),$$

as η is injective (Lemma 2).

This implies that $d'(v, s) \leq d'(v, w)$, so $d'(v, w)$ cannot be the unique minimal distance for d' at v , a contradiction. ■

Corollary 16 *Scale Invariance, Richness and Monotonic Consistency are mutually compatible clustering axioms.*

We have summarised the clustering axioms satisfied by our three instances of Morse Clustering in Table 1.

We finish this section with a few remarks on vertex labelling and tie-breaking. Note that our choices of vertex preorders depend on an arbitrary but fixed choice of vertex labelling $X = \{x_1, \dots, x_n\}$. Such a choice is implicit in (Kleinberg, 2003), where it is used as a tie-breaking procedure for Single Linkage clustering. For Morse Clustering, on the other hand, this vertex labelling represents a choice of a vertex potential function and is fundamental to the algorithm, as only ‘uphill’ or ascending edges are admissible. Nevertheless, the results in this section apply to an arbitrary, but fixed, labelling or ordering of the elements in X , and this suffices in our axiomatic setting. We call this *agnostic* Morse Clustering. In a practical setting, however, Morse Clustering is a clustering algorithm for edge- and vertex-weighted networks, with vertex weights obtained from either the network structure alone (e.g. vertex degrees), or provided otherwise by the practitioner. We call these *unsupervised*, respectively *supervised*, Morse Clustering. Examples of both are given in Appendix B.

3. Axiomatic Approach to Graph Clustering

In this section, we consider the axiomatic approach in the context of graph clustering, that is, of distances supported on a given graph G . Mathematically, we allow the distance function to take the value 0 (Definition 17). Conceptually, there are two different approaches depending on whether 0 is considered a numerical value (minimum distance) or indicating that the distance is ‘not defined’. Each approach results in a different definition of clustering algorithm, namely Eqs. (11) and (12).

In Section 3.1, we argue that the first approach is essentially equivalent to that of (Laarhoven and Marchiori, 2014), where a possibility theorem for Kleinberg’s axioms is shown. The second approach, on the other hand, is closer to the usual interpretation of graph clustering, or partitioning, in network and computer science (Newman, 2018; Schaeffer, 2007). In this case, we prove an impossibility result for Consistency (3.2), and a possibility result for Monotonic Consistency, even when Richness is replaced by the more natural Connected-Richness axiom (Section 3.3).

3.1 Two approaches

If we allow a distance function to take the value 0 between pairs of distinct points, we obtain the following.

Definition 17 *A pseudo-distance on a set X is a function $d : X \times X \rightarrow \mathbb{R}$ such that $d(v, w) = d(w, v) \geq 0$ and $d(v, v) = 0$ for all $v, w \in X$ (that is, we allow $d(u, v) = 0$ for $u \neq v$).*

Note that we can represent a pseudo-distance on X as a graph with vertex set X in the usual way: an edge between i and j if $d(i, j) > 0$, and no such edge if $d(i, j) = 0$.

A direct generalisation of Kleinberg’s definition of clustering algorithm, Eq. (1), is

$$F : \{d \text{ pseudo-distance on } X\} \rightarrow \{\mathcal{P} \text{ partition of } X\}. \quad (11)$$

Kleinberg’s original axioms make sense in this setting, however a possibility result now holds: the function that returns the connected components of the graph representation of d (as above), is clearly scale-invariant, rich and consistent (cf. (Laarhoven and Marchiori, 2014)).

The issue is that 0 is a special numerical value: it is the absolute minimum for any pairwise distance, and does not scale ($c \cdot 0 = 0$ for all c). In particular, although Scale-Invariance does not change the underlying graph, the Consistency axiom can create and eliminate edges. Hence this approach focuses on the set X rather than on a fixed graph G .

Remark 18 *If we also allow an absolute maximum distance, ∞ , then $1/d$ is a similarity function in the sense of (Laarhoven and Marchiori, 2014) and Consistency for d is equivalent to Consistent Improvement (Laarhoven and Marchiori, 2014) for $1/d$.*

Instead, we suggest a more natural approach when the focus is on the graph $G = (V, E)$: we allow arbitrary positive distances on edges while keeping $d(u, v) = 0$ whenever $(u, v) \notin E$. In clustering problems, we are normally interested in minimising the edge cut (Schaeffer, 2007), and hence the absence of an edge is significant. In fact, the underlying hypothesis in

network analysis, where clustering is a very well established technique, is that the structure of the network (mathematically a graph, typically very sparse) carries information. For this reason, we fix a graph G and restrict to distances supported on (the edges of) G , and define distances, and clustering algorithms, accordingly.

Definition 19 *A pseudo-distance on a graph $G = (V, E)$ is a pseudo-distance d on the vertex set V that is supported on the edge set, that is, $d(v, w) \neq 0$ if and only if $(v, w) \in E$. (Equivalently, a positive weight function on undirected edges.)*

Note that, for this definition to make sense, G must be loopless and undirected (we will assume this from now on). Given a graph $G = (V, E)$, we define a *graph clustering algorithm* as any function

$$F : \{d \text{ pseudo-distance on } G\} \rightarrow \{\mathcal{P} \text{ partition of } V\}. \quad (12)$$

Clearly, a distance on a set X is the same as a pseudo-distance on the complete graph with vertex set $V = X$. Hence this so-called *sparse* setting generalises Kleinberg's setting from a complete to an arbitrary (but fixed) graph on X .

3.2 An impossibility theorem for graph clustering

Kleinberg's axioms can be stated in the graph clustering setting above (Eq. (12)), as follows.

- **Scale-invariance:** For any pseudo-distance d on G and $\alpha > 0$, we have $F(d) = F(\alpha \cdot d)$;
- **Richness:** Given a partition \mathcal{P} , there exists a pseudo-distance d on G such that $F(d) = \mathcal{P}$;
- **Consistency:** Given pseudo-distances d and d' on G with $\mathcal{P} = F(d)$, if d' is a \mathcal{P} -transformation of d , that is,

$$\begin{cases} d'(v, w) \leq d(v, w) & \text{if } v \sim_{\mathcal{P}} w, \text{ and} \\ d'(v, w) \geq d(v, w) & \text{if } v \not\sim_{\mathcal{P}} w, \end{cases} \quad (13)$$

then $F(d') = F(d)$.

(If G is a complete graph these axioms coincide with Kleinberg's for the set $X = V$.)

In the sparse setting it seems natural to restrict to *connected partitions*, that is, partitions where each cluster is a connected subgraph of G , as otherwise we would be grouping together objects which are unknown to be similar or not, in apparent contradiction with the very principle of clustering. Therefore, we define a weaker Richness axiom:

- **Connected-Richness:** Given a connected partition \mathcal{P} , there exists a pseudo-distance d on G such that $F(d) = \mathcal{P}$.

Similarly, we will only consider connected graphs from now on (it seems sensible to assume $F(G) = F(G_1) \cup F(G_2)$ whenever G is the disjoint union of G_1 and G_2).

Connected-Richness is clearly equivalent to Richness in the complete case. In the sparse case, however, many graph clustering algorithms, such as Single Linkage, or Morse Clustering (Algorithms 1 and 2), always produce a connected partition (which seems very sensible in any case). Since clustering algorithms cannot create new edges, such algorithms cannot satisfy Richness in its general form. Since Richness implies Connected-Richness, our impossibility result also holds for Scale-Invariance, Consistency and Richness.

Theorem 20 (An Impossibility Theorem for Graph Clustering) *Let G be a connected graph with at least three vertices, and F a graph clustering algorithm on G . Then F cannot satisfy Scale-Invariance, Consistency and Connected-Richness.*

Before proving this theorem, we introduce some notation. Given a pseudo-distance d on $G = (V, E)$ and a partition \mathcal{P} of V , let $g(\mathcal{P}, d) = (x, y)$ and $h(\mathcal{P}, d) = (p, q)$ where

$$\begin{aligned} x &= \max \{d(u, v) \mid (u, v) \in E, u \sim_{\mathcal{P}} v\}, \\ p &= \min \{d(u, v) \mid (u, v) \in E, u \sim_{\mathcal{P}} v\}, \\ y &= \min \{d(u, v) \mid (u, v) \in E, u \not\sim_{\mathcal{P}} v\}, \\ q &= \max \{d(u, v) \mid (u, v) \in E, u \not\sim_{\mathcal{P}} v\}, \end{aligned}$$

the maximal (minimal) intra (inter) cluster distances, and, if \mathcal{P} is the trivial partition, we set $y = q = 0$.

We observe that, if d and d' are pseudo-distances on G and \mathcal{P} is a partition of V , the condition $h(\mathcal{P}, d) = g(\mathcal{P}, d')$ guarantees that d' is a \mathcal{P} -transformation of d .

Proof Note that, in any connected graph, we can always remove a vertex so that the remaining graph is connected. For example, if T is a spanning tree of G , v any vertex, and s the vertex realising the maximal (shortest path) distance from v in T , then the graph induced by $V \setminus \{s\}$ must still be connected. Since $|V| \geq 3$, we can repeat the argument on $V \setminus \{s\}$ and find $t \neq s$ such that $\mathcal{P} = \{\{s\}, X \setminus \{s\}\}$ and $\mathcal{P}' = \{\{s\}, \{t\}, X \setminus \{s, t\}\}$ are connected partitions.

Since F satisfies Connected-Richness, there exist pseudo-distances d and d' on G such that $F(d) = \mathcal{P}$ and $F(d') = \mathcal{P}'$. Let $h(\mathcal{P}, d) = (p, q)$ and $h(\mathcal{P}', d') = (p', q')$. Since F satisfies Consistency, we can assume $p < q$ and $p' < q'$. Also, note that p, q and q' cannot be zero.

Let d^* be the pseudo-distance on G defined by $d^*(s, v) = q$ if $v \neq s$, $d^*(t, v) = p$ if $v \neq s, t$, and $d^*(u, v) = (pp')/q'$ if $u, v \neq s, t$. Then $g(\mathcal{P}, d^*) = (p, q)$, since the only inter-cluster distance value is q , and the only intra-cluster distance values are p and $p(p'/q') < p$. Therefore, $g(\mathcal{P}, d^*) = h(\mathcal{P}, d)$, hence d^* is a \mathcal{P} -transformation of d , by the observation before the proof, and, consequently, $F(d^*) = F(d)$, by Consistency.

On the other hand, $g(\mathcal{P}', \alpha d^*) = \alpha g(\mathcal{P}', d^*)$ for any α positive constant. If we choose $\alpha = q'/p$ then we have $g(\mathcal{P}', \alpha d^*) = \alpha((pp')/q', p) = (p', q') = h(\mathcal{P}', d')$ so, by the same argument as above, αd^* is a \mathcal{P}' -transformation of d' and thus $F(\alpha d^*) = F(d') = \mathcal{P}'$, by Consistency. Since F satisfies Scale-Invariance, this implies $F(\alpha d^*) = F(d^*) = F(d) = \mathcal{P}$ and, therefore, $\mathcal{P} = \mathcal{P}'$, clearly a contradiction. \blacksquare

3.3 Monotonic Consistency for graph clustering

Next we consider Monotonic Consistency and **Morse** Clustering in the sparse setting. We can extend Monotonic-Consistency to connected graphs by considering monotonic transformations (Definition 4) of pseudo-distances on a given graph.

- **Monotonic-Consistency:** Given pseudo-distances d and d' on G with $\mathcal{P} = F(d)$, if d' is a \mathcal{P} -monotonic transformation of d , then $F(d') = F(d)$.

The input of the Morse Clustering algorithm (Algorithm 2) is an arbitrary graph, and the output flow always induces a connected partition (Theorem 11). Therefore, we can consider Morse Clustering, and hence any of its instances, as graph clustering algorithms.

The three instances of Morse Clustering discussed in Section 2.4 satisfy the analogous axioms as in the complete case except that we need to allow the vertex labelling (arbitrary but prefixed in the complete case) to be part of the algorithm to satisfy Connected-Richness. This is a necessary condition: once a vertex labelling (or preorder) is fixed, only ‘uphill’ edges are admissible, preventing certain configurations to occur (for example, u and v cannot be in the same cluster if all paths from u to v contain a vertex lower than both). This is not an intrinsic limitation of Morse Clustering but reflects the fact that it is fundamentally a vertex-weighted clustering algorithm, that is, both distance and vertex preorder are part of the input data.

We can either allow the (so far arbitrary and prefixed) vertex labelling to be part of the algorithm, or to restrict to partitions compatible with such a committed choice. Formally, given a vertex preorder \preceq_V on V , we say that a partition $\mathcal{P} = \{V_1, \dots, V_k\}$ of V is *compatible with* \preceq_V if there is a rooted spanning tree T_i of (the subgraph induced by) V_i rooted at a vertex v_i such that every directed edge in T_i (edges directed towards the root) is admissible with respect to \preceq_V . Note that v_i is necessarily the maximal vertex in T_i with respect to the preorder, and that \mathcal{P} is necessarily a connected partition.

Remark 21 *One can show that \mathcal{P} is compatible with \preceq_V if and only if for every $u \sim_{\mathcal{P}} v$ there exists a path from u to v such that no vertex in the path is strictly less than both u and v .*

Clearly, for every partition there is a choice of compatible preorder \preceq_V . This is also true for the **SiR** and δ -**Morse** vertex preorders: given a partition, there is a choice of labelling $V = \{v_1, \dots, v_n\}$ such that the preorder is compatible with the partition.

Formally, we define Morse-Richness for a Morse clustering algorithm F on a graph $G = (V, E)$ with a choice of vertex preorder \preceq_V as follows.

- **Morse-Richness:** Given a partition \mathcal{P} of V compatible with \preceq_V , there exists a pseudo-distance d on G and a vertex preorder such that $F(d) = \mathcal{P}$.

Morse-Richness is equivalent to Connected-Richness if we accept the vertex labelling as an input of the algorithm.

Now we can show that the three instances of Morse Clustering satisfy the analogous axioms as in Section 2 (see Table 1), including a possibility theorem for Monotonic-Consistency and **SiR-Morse**.

Theorem 22 Let $G = (V, E)$ be a graph, and consider **SiR-Morse**, **k -Morse** and **δ -Morse** as graph clustering algorithms on G , for some fixed labelling $V = \{v_1, \dots, v_n\}$. Then:

- (i) **SiR-Morse** satisfies Scale-Invariance, Morse-Richness and Monotonic Consistency.
- (ii) **k -Morse** satisfies Scale-Invariance and Consistency.
- (iii) **δ -Morse** satisfies Morse-Richness and Consistency.

Proof i The proofs of Scale Invariance and Monotonic Consistency are identical (they do not use the fact that G is a complete graph) as those in Theorem 14. For Morse-Richness, consider $V = V_1 \cup \dots \cup V_k$ an arbitrary connected partition of V . For each V_i , choose a spanning tree T_i and a root v_i such that each edge in T_i is admissible.

Define a pseudo-distance d on G as follows. If (s, t) is an edge on T_i , then $d(s, t)$ is the maximum of the distance from s to v_i in T_i and the distance from t to v_i in T_i (by distance in a tree we simply mean the ‘hop’ distance). If (s, t) is an edge not in any spanning tree, then $d(s, t) = |V|$.

With this choice, v_i is critical and, if $v \in V_i$, then the maximal edge at v is the one connecting it to a vertex in T_i closer to v_i , and it is admissible. All in all, the associated tree $T_{v_i} = T_i$ and the Morse flow recovers the original partition.

ii The proof of Scale Invariance is identical to that in Theorem 23. For Consistency, let d be a pseudo-distance on G , \mathcal{P} the partition given by **k -Morse**, and d' a \mathcal{P} -transformation of d , that is,

$$\begin{cases} d(v, w) \geq d'(v, w), & \text{if } v \sim_{\mathcal{P}} w, \\ d(v, w) \leq d'(v, w), & \text{otherwise.} \end{cases}$$

Let Φ respectively Φ' be the Morse flow corresponding to d respectively d' . As in the proof of Theorem 23, for all $i > n - k$ we have that $\Phi(v_i) = v_i = \Phi'(v_i)$, critical.

Suppose now $\Phi(v_i) = v_i$ for some $i \leq n - k$. Let $J = \{v_j \mid (v_i, v_j) \in E, v_i \prec_V v_j\}$, the admissible edges from v_i . By the definition of the edge preorder, if there are admissible edges ($J \neq \emptyset$) then the maximal admissible edge exists and it is unique. Since v_i is critical, we must have $J = \emptyset$. Since there are no admissible edges at v_i , we also have $\Phi'(v_i) = v_i$. All in all, Φ and Φ' have the same number of critical points and therefore \mathcal{P} and \mathcal{P}' have the same number of clusters (possibly more than k). The rest of the proof goes as in the proof of Theorem 24.

iii The proof of Consistency is identical to that in Theorem 25. For Morse-Richness, consider $V = V_1 \cup \dots \cup V_k$ an arbitrary connected partition of V , and choose a spanning tree T_i and a root v_i such that each edge in T_i is admissible.

Define a pseudo-distance d on G as follows. If (s, t) is an edge in some T_i , then $d(s, t) = \delta/2$, and if (s, t) is not an edge in any T_i then $d(s, t) = \delta$. By the definition of edge preorder, v_i is critical and the maximal edge at $v \in V_i \setminus \{v_i\}$ is the only edge in T_i connecting v to a vertex closer to v_i in T_i . All in all, the tree associated to v_i by the Morse flow is T_i and hence we recover the original partition. ■

Acknowledgments

We thank Francisco Belchí Guillamón for valuable discussions about Monotonic Consistency and metrics. Fabio Strazzeri was supported by a PhD studentship by Mathematical Sciences and the Institute for Life Sciences at the University of Southampton.

References

M. Ackerman and S. Ben-David. A characterization of linkage-based hierarchical clustering. *Journal of Machine Learning Research*, 17(232):1–17, 2016.

M. Ackerman, S. Ben-David, and D. Loker. Towards property-based classification of clustering paradigms. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 10–18. Curran Associates, Inc., 2010.

C. C. Aggarwal and C. K. Reddy. *Data Clustering: Algorithms and Applications*. CRC press, 2013.

K. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346, 1950.

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

J. Bigler et al. A severe asthma disease signature from gene expression profiling of peripheral blood from U-BIOPRED cohorts. *American Journal of Respiratory and Critical Care Medicine*, 195(10):1311–1320, 2017.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

G. Carlsson and F. Memoli. Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research*, 11:1425–1470, 2010.

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

L. Fisher and J. W. Van Ness. Admissible clustering procedures. *Biometrika*, 58(1):91–104, 1971.

R. Forman. Morse theory for cell complexes. *Advances in Mathematics*, 134:90–145, 1998.

W. H. Gottschalk and G. A. Hedlund. *Topological dynamics*, volume 36. American Mathematical Soc., 1955.

J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Applied statistics*, pages 54–64, 1969.

A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., 1988.

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

N. B. Karayiannis. An axiomatic approach to soft learning vector quantization and clustering. *IEEE Transactions on Neural Networks*, 10(5):1153–1165, 1999.

J. Kleinberg. *An Impossibility Theorem for Clustering*. MIT Press, 2003.

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

A. Lancichinetti and S. Fortunato. Community detection algorithms: a comparative analysis. *Physical Review E*, 80(5):056117, 2009.

J. Milnor. *Morse Theory*. Princeton University Press, 1963.

M. E. J. Newman. *Networks*. Oxford university press, 2018.

M. E. J. Newman and A. Clauset. Structure and inference in annotated networks. *Nature Communications*, 7, 2016.

V. Prasolov. *Intuitive topology*, volume 4. American Mathematical Soc., 1995.

J. Puzicha, T. Hofmann, and J. M. Buhmann. A theory of proximity based clustering: Structure detection by optimization. *Pattern Recognition*, 33(4):617–634, 2000.

M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

S. E. Schaeffer. Graph clustering. *Computer science review*, 1(1):27–64, 2007.

J. P. R. Schofield et al. A Topological data analysis network model of asthma based on blood gene expression profiles. *bioRxiv preprint* <http://dx.doi.org/10.1101/516328>, 2019.

F. Strazzeri. *A Morse-theoretical clustering algorithm for annotated networks and spectral bounds for fuzzy clustering*. PhD thesis, University of Southampton, 2019.

F. Strazzeri and R. J. Sanchez-Garcia. A Morse-theoretic clustering algorithm (in preparation).

U. von Luxburg, R. C. Williamson, and I. Guyon. Clustering: Science or art? In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pages 65–79. PMLR, 2012.

J. Yu and Z. Xu. Categorization axioms for clustering results. *arXiv preprint arXiv:1403.2065*, 2014.

R. Zadeh and S. Ben-David. A uniqueness theorem for clustering. *CoRR*, abs/1205.2600, 2012.

Appendix A. Two further instances of Morse Clustering

In this appendix, we define and study the two further instances of Morse clustering mentioned in the Main Text (Section 2.4), namely **k -Morse** and **δ -Morse**. They illustrate the versatility of Morse clustering and show that, for suitable choices of vertex and edge pre-orders, Morse clustering can satisfy each pair of Kleinberg’s axioms, in analogy to the three instances of Single-Linkage clustering with the same property in (Kleinberg, 2003). We keep the notation and terminology from Section 2.4.

Let $k \geq 1$ be an integer. Next, we present a Morse algorithm that guarantees a partition with k clusters (Theorem 23), and thus it cannot be rich. However it satisfies Consistency and Scale Invariance (Theorem 24).

k -Morse

- $v_i \preceq_V v_j$ if $i = j$ or $i + k < j$
- $(v, w) \preceq_E (v, t)$ if
 - $w \preceq_V v \preceq_V t$, or
 - $d(v, w) > d(v, t)$ and $v \preceq_V t$, or
 - $d(v, w) = d(v, t)$ and $w \preceq_V t$.

For this choice of vertex preorder, there are exactly k critical vertices, $v_n, v_{n-1}, \dots, v_{n-k+1}$, and hence k clusters (see Theorem 23 below). The edge preorder is defined such that admissible edges are always greater than non-admissible ones, and admissible ones are compared using distances, with the vertex preorder used as tie-breaking procedure. In particular, if there are admissible edges at v , the maximal admissible edge at v exists and it is unique.

Theorem 23 *k -Morse always produces a partition with k clusters.*

Proof If $v_i \in X$ with $i > n - k$ then there are no vertices greater than v_i with respect to \preceq_V hence no admissible edges at v and thus $\Phi(v_i) = v_i$ critical. On the other hand, v_i with $i \leq n - k$ cannot be critical, as there are admissible edges $(v_i, v_j) \in E_{v_i}$ for all $j > i + k$, so the maximum exists and it is unique. All in all, there are exactly k critical vertices $v_n, v_{n-1}, \dots, v_{n-k+1}$ and therefore exactly k clusters. ■

Theorem 24 *k -Morse is Consistent and Scale-Invariant.*

Proof (Scale-invariance) A distance transformation $d' = \alpha \cdot d$ for $\alpha > 0$ does not affect the k -Morse vertex or edge preorder, hence we obtain the same partition.

(Consistency) Let d be a distance in X , \mathcal{P} the partition given by k -Morse on (X, d) , and d' a \mathcal{P} -transformation of d , that is,

$$d(v, w) \geq d'(v, w), \text{ if } v \sim_{\mathcal{P}} w, \quad (14)$$

$$d(v, w) \leq d'(v, w), \text{ otherwise.} \quad (15)$$

Let Φ respectively Φ' be the Morse flow corresponding to d respectively d' . The critical points depend on the vertex preorder alone, hence, as in the proof of Theorem 23, we have $\Phi(v_i) = v_i = \Phi'(v_i)$ for all $i > n - k$ and thus \mathcal{P} and \mathcal{P}' have the same number of clusters. Therefore, it suffices to show that $x \sim_{\mathcal{P}} \Phi'(x)$ for all $x \in X$, by Lemma 12.

Let $x \in X$. If x is critical, $\Phi(x) = \Phi'(x)$ as they have the same critical points, so clearly $x \sim_{\mathcal{P}} \Phi(x) = \Phi'(x)$. If x is not critical, let $w = \Phi(x)$ and $t = \Phi'(x)$. The maximality and the definition of \preceq_E implies

$$d(x, w) \leq d(x, t) \text{ and } d'(x, t) \leq d'(x, w).$$

Since $\Phi(x) = w$, they are in the same cluster, $x \sim_{\mathcal{P}} w$, and thus $d'(x, w) \leq d(x, w)$, by Eq. (14) above. All in all,

$$d'(x, t) \leq d'(x, w) \leq d(x, w) \leq d(x, t). \quad (16)$$

Now, if $d'(x, t) < d(x, t)$, they are necessarily in the same cluster, $x \sim_{\mathcal{P}} t$, by Eqs. (14) and (15) above. The remaining case $d'(x, t) = d(x, t)$ implies equalities in Eq. (16), and, by the definition of the edge preorders and the maximality of (x, w) with respect to d , we have $w = t$. In both cases, $x \sim_{\mathcal{P}} t = \Phi'(x)$. \blacksquare

Let $\delta > 0$. The final instance of Morse clustering satisfies Consistency and Richness, and is given by the following choices of preorders.

δ -Morse

- $v_i \preceq_V v_j$ if $i \leq j$
- $(v, w) \preceq_E (v, t)$ if
 - $w = t$, or
 - $d(v, t) < \min\{d(v, w), \delta\}$ and $v \preceq_V t$, or
 - $d(v, w) = d(v, t) < \delta$ and $v \preceq_V w \preceq_V t$.

With this preorder, only admissible edges with distance less than the threshold parameter δ are considered for the flow. Among those edges, we choose the one with minimal distance, using the vertex preorder to resolve ties. Note that, if there are admissible edges at distance less than δ , the maximum admissible edge exists and it is unique.

Theorem 25 δ -Morse satisfies Consistency and Richness.

Proof (Richness) Consider an arbitrary partition $X = X_1 \cup \dots \cup X_k$ and define the distance function

$$d(v, w) = \begin{cases} \frac{\delta}{2}, & \text{if } v, w \text{ are in the same cluster, and} \\ \delta, & \text{otherwise,} \end{cases}$$

for $v \neq w$. Let x_i be the largest vertex in X_i with respect to \preceq_V and $v \in X_i$ arbitrary. By the definition of d and the edge preorder, we have that (v, x_i) is the maximum admissible edge at v . Also, x_i is critical: the maximum edge at x_i is of the form (x_i, w) for $w \in X_i$, hence not admissible or, if $|X_i| = 1$, any edge in E_{x_i} is maximal, hence unique (since $|X| \geq 3$). Therefore, δ -Morse reproduces the partition $X_1 \cup \dots \cup X_k$ (in fact, each cluster is a directed star with root x_i).

(Consistency) Let d be a distance in X , \mathcal{P} the partition given by δ -Morse on (X, d) , and d' a \mathcal{P} -transformation of d , that is,

$$d(v, w) \geq d'(v, w), \text{ if } v \sim_{\mathcal{P}} w, \quad (17)$$

$$d(v, w) \leq d'(v, w), \text{ otherwise.} \quad (18)$$

Let Φ respectively Φ' be the Morse flow corresponding to d respectively d' . Let $s \in X$ arbitrary, $v = \Phi(s)$ and $w = \Phi'(s)$ with $v, w \neq s$. As in the proof of Theorem 24, we have

$$d'(s, w) \leq d'(s, v) \leq d(s, v) \leq d(s, w).$$

Then either $d'(s, w) < d(s, w)$, and so $s \sim_{\mathcal{P}} w$ by Eq. (17), or $d'(s, w) = d(s, w)$, which implies, by the definition of edge preorder, $v = w$, and thus $s \sim_{\mathcal{P}} w = \Phi'(s)$ too. As s was arbitrary, we conclude that \mathcal{P}' is a refinement of \mathcal{P} , by Lemma 12. To prove that they are equal, it suffices to show that they have the same critical points (i.e. the same number of clusters), that is, $\Phi(v) = v$ if and only if $\Phi'(v) = v$.

Suppose that $\Phi(v_i) = v_i$ and $\Phi'(v_i) = v_j$, $i \neq j$. Since the vertex preorder is strictly increasing along the flow, $v_i \prec_V v_j$, that is, $i < j$. By the definition of Morse clustering, $v_i \sim_{\mathcal{P}'} v_j$ hence $v_i \sim_{\mathcal{P}} v_j$, since \mathcal{P}' is a refinement. However, this contradicts v_i being maximal in its \mathcal{P} cluster as $i < j$.

Now suppose $\Phi'(v_i) = v_i$ and $\Phi(v_i) = v_j$, $i \neq j$. The edge from v_i to v_j is in the flow Φ , so $d(v_i, v_j) < \delta$, however it is not in the flow Φ' so $d'(v_i, v_j) \geq \delta$. However, as $d'(v_i, v_j) \leq d(v_i, v_j) < \delta$, we have that v_i has at least one admissible edge. By the definition of \preceq_E , v_i cannot be critical for Φ , that is, a unique maximal edge that is admissible must exist.

Finally, since $v_i \sim_{\mathcal{P}} v_j$ and d' is a \mathcal{P} -transformation, we have $d'(v_i, v_j) \leq d(v_i, v_j)$, and we arrive to a contradiction. \blacksquare

Appendix B. Morse Clustering in Practice

The algorithm **Morse** presented here is intrinsically a clustering algorithm for graphs with vertex annotations, such as vertex weights. Node annotation, sometimes called ‘metadata’ in the literature, is often readily available but ignored in network analysis, particularly in clustering. However, node annotation can help improve community detection algorithms

(Newman and Clauset, 2016). The node annotation exploited by **Morse** can be application-dependent, in the form of external metadata provided by the user, but also graph-dependent measures, such as vertex degree, local clustering coefficient, or other network structural measures. We refer to the former as *supervised*, and the latter as *unsupervised*, Morse clustering.

In this appendix, we show (unsupervised) Morse Clustering (**SiR-Morse**) in practice: first using the LFR benchmark (Section B.1), then in a bioinformatics study (Section B.2). We summarise the main results with full details available in (Strazzeri, 2019), respectively (Schofield et al., 2019).

B.1 LFR benchmark

In (Lancichinetti and Fortunato, 2009) the authors present the Lancichinetti–Fortunato–Radicchi (LFR) algorithm which generates benchmark networks with a priori known communities with heterogeneous node degree distribution and community sizes, given by power laws of predefined exponents. We chose this benchmark test because it reproduces typical vertex degree and community (cluster) sizes in real-world networks (Newman, 2018). We studied the performance of **Morse** against the LFR benchmark using NMI (Normalised Mutual Information) as evaluation measure. We used the number of common neighbours as edge weights, $w(i, j) = |N(i) \cap N(j)|$, and two centrality measures as node weights, that is the annotation is $f : V \rightarrow \mathbb{R}^2$, where the entries are the maximal and the sum of the incident edge weights, $w_1(i) = \max_{(i,j) \in E} w(i, j)$ and $w_2(i) = \sum_{(i,j) \in E} w(i, j)$ (we compare nodes using w_1 and resolve ties using w_2). The results are shown in Figure 9 (left) and compared to two state-of-the-art clustering algorithms, namely Blondel et al. (Blondel et al., 2008) and Infomap (Rosvall and Bergstrom, 2008). We can see that **Morse** detects the ground-truth communities for μ_t (mixing parameter) up to 0.5, showing a similar performance to Blondel et al., although not as good as the best performing algorithm, Infomap.

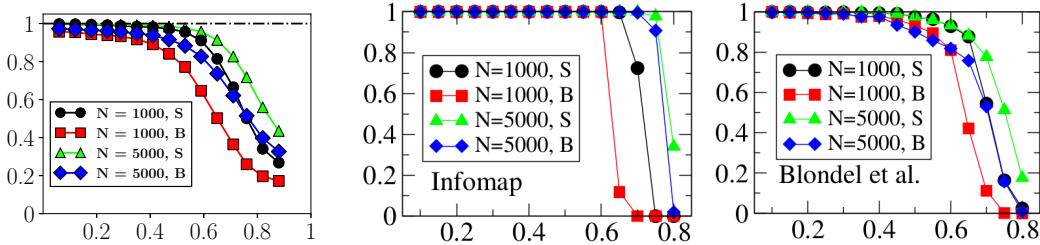


Figure 9: LFR benchmark tests for the **Morse** algorithm. Following Lancichinetti and Fortunato (2009), we show four curves, corresponding to two different network sizes (1000 and 5000 nodes) and two ranges of community sizes: **S** (Small, 10–50 nodes per community) and **B** (Big, 20–1000 nodes per community). We plot a mixing parameter (x -axis, ‘noise’) versus NMI (y -axis, clustering ‘correctness’). Each data point represents average performance over 100 realisations.

B.2 Asthma study

We applied **Morse** in a real-world scenario and showed how it helps to identify phenotypes of asthma based on blood gene expression profiles, using data from the U-BIOPRED¹ study.

Transcriptomics data from 606 participants in an asthma study were clustered into overlapping groups of patients by Topological Data Analysis (TDA), following Bigler et al. (2017), using the Ayasdi Platform². This TDA procedure generated a network with groups of patients represented by nodes and connections denoting overlap between groups. In Figure 10, we show this network with two node measures (cluster size, and compactness), and in Figure 11, the output of our **Morse** clustering algorithm using a combination of these node measures. The subsequent validation study of the 9 largest clusters confirmed a very good to excellent prediction of cluster classification in the validation set, based on a logistic regression model identifying predictors of the clusters in the training set.

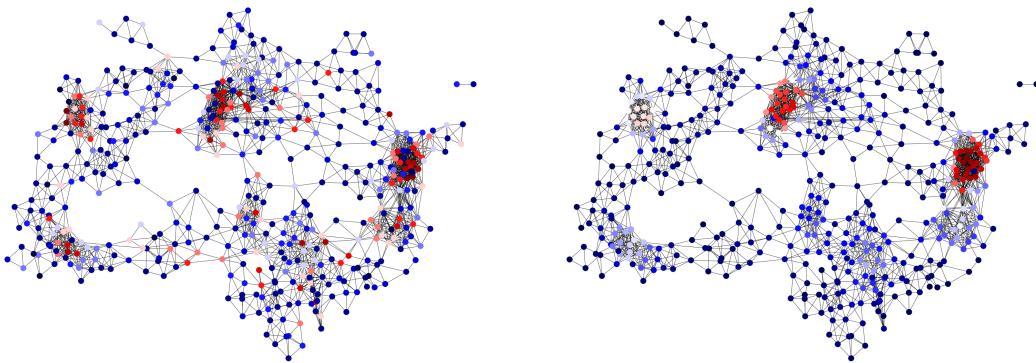


Figure 10: Network representation of transcriptomics data from an asthma study. Each node represents a group of similar patients, with overlaps (the same patient may belong to more than one group) denoted by edges. The network is coloured by group size (left) and cluster compactness (right)—a measured based on the average distance between in-group patients—from blue (low) to red (high). These two node measures were used by the Morse clustering shown in Figure 11.

1. <http://www.europeanlung.org/en/projects-and-research/projects/u-biopred/home>
 2. <https://www.ayasdi.com/platform/>

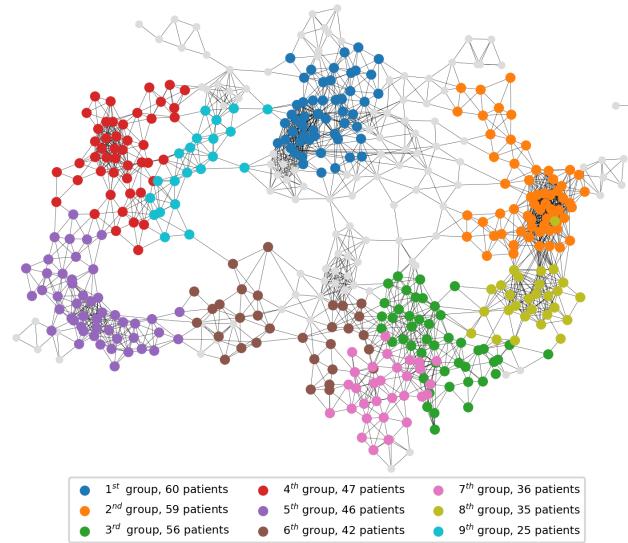


Figure 11: Morse clustering of the TDA-based network of Figure 10 with respect to patient overlap (edge weight) and a combination of patient group size and compactness (Fig. 10). We only show the 9 largest clusters (with respect to overall patient number). Of these, 7 clusters were reproducibly identified in an independent validation set using ROC analysis of logistic regression classification models, see Schofield et al. (2019).