Local data structures

J.F. Jardine*

jardine@uwo.ca

March 4, 2023

Abstract

Local data structures are systems of neighbourhoods within data sets. Specifications of neighbourhoods can arise in multiple ways, for example, from global geometric structure (stellar charts), combinatorial structure (weighted graphs), desired computational outcomes (natural language processing), or sampling. These examples are discussed, in the context of a theory of neighbourhoods.

This theory is a step towards a mathematical understanding of clustering for large data sets. These clusters can only be approximated in practice, but approximations can be constructed from neighbourhoods via patching arguments that are derived from the Healy-McInnes UMAP construction. The patching arguments are enabled by changing the theoretical basis for data set structure, from metric spaces to extended pseudo metric spaces.

Introduction

This paper is a preliminary discussion of the mathematics of local structures for large data sets.

Potential objects of study include subsets $\mathcal{U} \subset \mathbb{R}^N$, where the data set $\mathcal{U}$ (or "universe") is essentially infinite, meaning that $\mathcal{U}$ is too large to analyze with available computational devices.

Alternatively, there may not be a metric space structure on the data set $\mathcal{U}$. Such objects $\mathcal{U}$ can arise as vertices of large weighted graphs Γ, which could describe data transfers that occur during a time interval. Other examples arise in the "bag of words" model natural language processing, which model has a combinatorial structure that is not graph theoretic.

There could, finally, be no apparent geometric or combinatorial structure for $\mathcal{U}$, and its structure near a point may have to be approximated (or learned) by iterated sampling.

*Supported by NSERC.

In general, one wants to break up a data set $\mathcal{U}$ into smaller computable pieces N that cover $\mathcal{U}$ in the sense that every $x \in \mathcal{U}$ is in some neighbourhood N, in the hope/expectation that analyses of the neighbourhoods N can be assembled to a full or at least useful partial analysis of the universal data set $\mathcal{U}$. This is essentially the approach taken by the mapper algorithm [**?**] (see Remark **??** below), and it can make perfect sense for clustering at relatively small distance scales.

The elements of a neighbourhood N should be close to x in some sense, but one has to address the question of how to find such neighbourhoods in a sea of data $\mathcal{U}$. If there is no prior information about the structure or genesis of $\mathcal{U}$, the phrase "close to x" may not have much meaning. In good cases, there is information about local geometric or combinatorial structures that allows one to get started.

Most generally, a neighbourhood N of a point x in a data set $\mathcal{U}$ is a suitably sized subset of $\mathcal{U}$ which contains x. If $\mathcal{U}$ is a metric space (or an extended pseudo metric space) then N has a diameter $s(N)$, which is the maximum distance $d(x, y)$ for $y \in N$.

The inclusion $N \subset \mathcal{U}$ determines an inclusion of Vietoris-Rips complexes $V(N) \subset V(\mathcal{U})$.

If every $x \in \mathcal{U}$ has a specific choice of neighbourhood N_x, as in Section 3, then the collection of all such neighbourhoods determines an inclusion of filtered complexes

$$N(\mathcal{U}) := \cup_{x \in \mathcal{U}}\ V(N_x) \subset V(\mathcal{U}),$$

which complexes are filtered by distance in the usual way. I say that $N(\mathcal{U})$ is the *neighbourhood complex* that is defined by the family of neighbourhoods $N = \{N_x\}$.

The neighbourhood complex $V(N)$ is the mapper complex for the covering $V(N_x) \subset V(\mathcal{U})$ of the global Vietoris-Rips complex $V(\mathcal{U})$, as in [**?**].

Every element $y \neq x$ in a neighbourhood N_x determines a ray

$$\{x, y\} \subset N_x \subset \mathcal{U},$$

and the collection of such rays determines a filtered subcomplex

$$R(N_x) = \vee_{y \neq x}\ V(\{x, y\}) \subset V(N_x).$$

Taking the union

$$R(\mathcal{U}) = \cup_{x \in \mathcal{U}}\ R(N_x) \subset V(\mathcal{U})$$

defines the *ray subcomplex* $R(\mathcal{U})$, which is a subcomplex of both $V(\mathcal{U}$ and $N(\mathcal{U})$.

The ray subcomplex $R(\mathcal{U})$ is a filtered (or weighted) graph.

If the neighbourhoods N_x consist of sets of k-nearest neighbours for the points of $\mathcal{U}$, then the ray subcomplex $R(\mathcal{U})$ is the k-nearest neighbours graph, which is a well-studied object. The k-nearest neighbours graph is used to construct the UMAP graph [**?**], [**?**], [**?**].

The inclusions

$$R(\mathcal{U}) \subset N(\mathcal{U}) \subset V(\mathcal{U})$$

of filtered complexes induce surjections

$$\pi_0 R_s(\mathcal{U}) \to \pi_0 N_s(\mathcal{U}) \to \pi_0 V_s(\mathcal{U})$$

for distance parameters s, which are analyzed in special cases in Sections 3 and 4. There are good comparison results for finite s for bounded neighbourhoods, which is the subject of Section 4. See Lemma **??**, Lemma **??**, Lemma **??** and Lemma **??**.

In that setting, the neighbourhood complex $N_s(\mathcal{U})$ for bounded neighbourhoods has the same 1-skeleton as the global Vietoris-Rips complex $V_s(\mathcal{U})$ at small distance scales s, which makes the neighbourhood complex $N_s(\mathcal{U})$ a good approximation of $V_s(\mathcal{U})$ for clustering for such s.

At higher distance scales, the clusters of the ray complex $R_s(\mathcal{U})$ coincide with those of the neighbourhood complex $N_s(\mathcal{U})$. The outcome is that, for clustering, the neighbourhood complex $N(\mathcal{U})$ is a bridge between the ray complex $R(\mathcal{U})$ (a UMAP-like object) and the full Vietoris-Rips complex $V(\mathcal{U})$.

The basic ideas and constructions of this paper appear in the Sections 2 and 3, along with a discussion of the relationship between neighbourhoods and sequences of nearest neighbours. With a view to potential applications (as in Section 5), we generally assume that $\mathcal{U}$ is an extended pseudo metric space, or an ep-metric space. The basic ideas around ep-metric spaces are summarized in Section 1.

Subsequent results and calculations are determined by choices of neighbourhoods, which choices vary with the geometric or combinatorial structures of specific examples.

The definitions and results of Sections 4, 6 and 7 are based on naive examples (or thought experiments) that motivate and illustrate these ideas:

1) The Gaia Archive $\mathcal{U}$ is a database of roughly a billion stars in the Milky Way. The raw data for the Archive is a set of scans that has been collected by the Gaia Space Observatory spacecraft, starting in 2014. The scans return high resolution photometric and spectral data for stars within small apertures, and so the archive is constructed from an assembly of local data. The positions of the stars in the archive relative to the Sun are determined, after repeated observations and much computation.

These positions can be expressed as a function $p : \mathcal{U} \to \mathbb{R}^3$ that determines the members of the Archive $\mathcal{U}$ uniquely. The position function p is a type of dimension reduction. In the language of the mapper construction, it is a filter function [**?**].

From observation, if x is a star in the archive $\mathcal{U}$, then there is a neighbourhood $N_x \subset \mathcal{U}$ of stars close to x such that N_x has a computable number of elements. We could insist that N_x is a bounded neighbourhood, in that it has a bounded radius $s(N_x)$ and it contains at most k elements for some choice of integer k.

This is an explicitly geometric example, which is closely aligned with methods that are presented in Section 4.

2) For some data sets, there is a graph structure Γ with no apparent ambient metric space.

For example, a collection of data transfers between computer accounts within a (short enough) time interval can be given the structure of a sparse directed weighted graph, as in Example **??** below. The number of bytes transmitted by a transfer is its weight.

The vertices of these graphs have low valence. One knows all of the transfers $e : x \leftrightarrow y$ for each account x, and from this one builds a computable neighbourhood $N_k(x)$ of accounts which are separated from x by at most k transfer steps (or hops).

One needs a way of assigning weights $d(x, y)$ to the various $y \in N_k(x)$. Starting with an account x, one could expect that the accounts y with which it does the most "business" are the closest to x. The elements y of $N_k(x)$ which are closest to x are defined "inversely" by the sum $\Sigma(x, y)$ of all weights of directed edge paths between x and y. Then the distance $d(x, y)$ can be defined by

$$d(x, y) = e^{-\Sigma(x,y)}$$

between x and y for each $y \in N_k(x)$.

From the data of neighbourhoods and weights, the Healy-McInnes UMAP machine generates a global ep-metric D on the set Z of vertices of the graph Γ, with clusters given by the directed set $\pi_0 V(Z, D)$, or equivalently (Theorem **??**) by the directed set $\pi_0 R(N)$ arising from the rays of the various neighbourhoods $N_k(x)$.

The point, ultimately, is that one uses the graph structure to find computable weighted neighbourhoods $N_k(x)$ for all vertices x of a sparse weighted directed graph Γ. These local structures then patch together to define a global ep-metric on the full set of vertices of Γ, along with cluster constructions.

These ideas appear in Section 6. In broad outline, they apply equally well to all sparse weighted graphs.

There is a fundamental idea in play here: the UMAP construction creates global space-level structure and cluster computations from local information given by weighted neighbourhoods, with or without the existence of an ambient metric.

This observation is applied repeatedly in examples that are displayed here. We specify neighbourhoods with weights, and then feed these neighbourhoods to general machinery.

The relevant theoretical features of the UMAP construction are summarized in Section 5. That section contains an alternate presentation of the UMAP graph, which is constructed by patching together rays without invoking most of the standard methods of UMAP — see Theorem **??**.

3) Section 7 is a discussion of neighbourhoods of words in the "continuous bag of words" model from natural language processing (NLP). With such neigh-

bourhoods in hand (and with appropriate definitions of weights), one again uses UMAP methods to construct an ep-metric space structure on the set of words $\mathcal{L}$ that of a corpus.

The methods of Section 7 extend to any finite set of strings of data elements, in which a local metric can be defined by proximity within strings.

In the examples displayed so far, the local nature of a data set varies within a given geometric or combinatorial structure. These structures are in part determined by desired computational outcomes, and they are the starting points for calculations.

One could, finally, be presented with a very large cloud of points $\mathcal{U}$ with an ep-metric space structure, but with no other information, from which one wants to approximate (or discover) a neighbourhood N_x for a given point $x \in \mathcal{U}$.

There seems to be no choice in such a case but to apply brute force methods that are based on repeated random sampling, with the goal of learning a description of a neighbourhood, or "k-complete" neighbourhood N_x for x. A potential method for doing so is described in Section 8.

The k-complete neighbourhoods of this paper (see Sections 2 and 4) are strongly related to sets of k-nearest neighbours for a point x, but have the benefit of being uniquely defined, and are therefore easier to manipulate theoretically. Of course, the positive integer k must be specified up front.

Contents

1 Extended pseudo metric spaces

An *extended pseudo-metric space* (X, d), here called an *ep-metric space*, is a set X together with a function $d : X \times X \to [0, \infty]$ such that the following conditions hold:

1) $d(x, x) = 0$,

2) $d(x, y) = d(y, x)$,

3) $d(x, z) \leq d(x, y) + d(y, z)$.

There is no condition that $d(x, y) = 0$ implies x and y coincide — this is where the adjective "pseudo" comes from, and the gadget is "extended" because we allow infinite distance.

A metric space (X, d) is an ep-metric space for which $d(x, y) = 0$ implies $x = y$, and all distances $d(x, y)$ are finite.

There is a category $\mathbf{ep-met}$ of ep-metric spaces, with morphisms $f : (X, d) \to (Y, d')$ given by functions $f : X \to Y$ which are non-expanding in the sense that $d'(f(x), f(y)) \leq d(x, y)$ for all $x, y \in X$.

The category $\mathbf{ep-met}$ is a cocomplete in the sense that it has all small colimits.

In effect, the coproduct $\sqcup_i\ (X_i, di)$ is the disjoint union set $\sqcup_i\ X_i$, equipped with the ep-metric d defined by

$$d(x,y) = \begin{cases} d_i(x,y) & \text{if } x,y \in X_i \text{ for some } i, \text{ and} \\ \infty & \text{otherwise.} \end{cases}$$

Coequalizers are constructed from a quotient function. Suppose that (X,d) is an ep-metric space and that $p : X \to Y$ is a surjective function. Then Y has an ep-metric D such that for any pair $z, w \in Y$,

$$D(z,w) = \inf_P \ \sum d(x_i, y_i),$$

where each "path" P consists of pairs of points $\{x_i, y_i\},\ i \leq n$ in X such that $z = p(x_0)$, $w = p(y_n)$ and $p(y_i) = p(x_{i+1})$ for $i \leq n-1$. The function p defines a map $p : (X,d) \to (Y,D)$ of ep-metric spaces that has the universal property of quotients.

Example 1. Suppose that (X,d) and (X,d') are ep-metric spaces having the same set of oelements X. Then the amalgamation (wedge) $(X,d) \vee (X,d')$ in the ep-metric space category is an ep-metric space structure on X with

$$D(z,w) = \inf_P \ \sum D(x_i, x_{i+1}),$$

where each path P is a string of elements $z = x_0, x_i, \ldots, x_n = w$ of X and

$$D(x_i, y_i) = \min\ \{d(x_i, x_{i+1}), d'(x_i, x_{i+1})\}.$$

Each finite ep-metric space $\mathcal{U}$ has a family of Vietoris-Rips complexes $V_s(\mathcal{U})$, which are parameterized by distance s. Explicitly, $V_s(\mathcal{U})$ is the abstract simplicial complex (or poset) whose simplices are the finite subsets $\sigma = \{x_0, \ldots, x_k\}$ of $\mathcal{U}$ such that $d(x_i, x_j) \leq s$. The simplex σ is a k-simplex, and it has cardinality $k+1$.

As in the standard case, there is an ascending family of complexes

$$V_s(\mathcal{U}) \subset V_t(\mathcal{U}),\ s \leq t,$$

with $\mathcal{U} = V_0(\mathcal{U})$ (discrete complex on the set $\mathcal{U}$).

The limiting object $V_\infty(\mathcal{U})$ is a simplex $\Delta^{\mathcal{U}}$ with vertices $\mathcal{U}$, but it is not the case that $V_\infty(\mathcal{U})$ is a union of the subobjects $V_s(\mathcal{U})$ with s finite. Write

$$V_{<\infty}(\mathcal{U}) = \cup_{s<\infty}\ V_s(\mathcal{U}).$$

The simplicial set $V_{<\infty}(\mathcal{U})$ is a finite disjoint union of contractible components.

2 Neighbourhoods

Suppose that (Z, d) is a finite ep-metric space and that $x \in Z$.

In all of the following,

$$Z(x, s) = \{y \in Z \mid d(x, y) \leq s\}$$

is the closed ball of radius s in Z that is centred at x.

A **neighbourhood** N of x is a subset N of Z with $x \in N$ and $d(x, y) < \infty$.

A neighbourhood N aquires an ep-metric space structure from Z, and defines a filtered subcomplex $V(N) \subset V(Z)$ of the Vietoris-Rips complex $V(Z)$.

The **radius** $s(N)$ of the neighbourhood N is defined by

$$s(N) = \max_{y \in N} \ d(x, y).$$

Then $s(N) < \infty$ by assumption.

The neighbourhood N is said to be **complete** if $N = Z(x, s_N)$.

A neighbourhood N of x is a set of **nearest neighbours** if $d(x, z) \geq s(N)$ for all $z \in Z - N$. If $N = \{x, x_1, \ldots, x_k\}$ is a set of nearest neighbours (i.e. with cardinality $k + 1$), then N is a set of **k-nearest neighbours**.

A nearest neighbour y for x with $d(x, y) < \infty$ can be identified with a neighbourhood $N = \{x, y\}$ of nearest neighbours. This means that $d(x, y) \leq d(x, z)$ for all $z \in Z - \{x\}$. The distance $d(x, y)$ could be 0 in general.

Every complete neighbourhood $N = Z(x, s_N)$ is a set of nearest neighbours for x, and is a set of n-nearest neighbours, where $n = |N| - 1$.

Lemma 2. *Suppose that $N = \{x, x_1, \ldots, x_k\}$ is a set of nearest neighbours for x, and that the x_i are ordered such that*

$$d(x, x_1) \leq d(x, x_2) \leq \cdots \leq d(x, x_k).$$

Then x_i is a nearest neighbour of x in the subset $Z - \{x_1, \ldots, x_{i-1}\}$.

Proof. We have

$$d(x, x_i) \leq d(x, x_{i+1}) \leq \cdots \leq d(x, x_k) \leq d(x, z)$$

for all z outside of N. It follows that $d(x, x_i)) \leq d(x, w)$ for all $w \in Z - \{x_1, \ldots x_{i-1}\}$. □

Lemma 3. *Suppose that the neighbourhood N is a set of nearest neighbours for x and $z \in X - N$ is chosen such that $d(x, z) < \infty$ and $d(x, z) \leq d(x, v)$ for all $v \in Z - N$. Then the set $N \cup \{z\}$ is a set of nearest neighbours for x.*

Proof. The radius s_z of $N \cup \{z\}$ is $d(x, z)$. Choose $v \in Z - (N \cup \{z\})$. Then $s(N) \leq d(x, v)$, and $d(x, z) \leq d(x, v)$ by the minimality of $d(x, z)$. It follows that $s(N \cup \{z\}) \leq d(x, v)$. □

Remark 4. Applying Lemma **??** inductively gives nearest neighbourhoods N of x of all possible finite cardinalities $|N|$ with $|N| \leq |Z|$.

There is a function $d_x : Z \to [0, \infty]$ with $d_x(y) = d(x, y)$. A nearest neighbour for x is an element $z \in Z - \{x\}$ such that $d_x(z) < \infty$ and $d_x(z)$ is minimal.

For such an element z, write $s = d_x(z)$. Then s is the minimum finite value of the image $d_x(Z)$, and $z \in Z_x(s)$, where

$$Z_x(s) = d_x^{-1}(s)$$

is the fibre (pre-image) of d_x over s.

Lemma 5. *Suppose that N is a set of nearest neighbours for x, and suppose that $\{s_1, \ldots, s_p\}$ is the set of elements of the image $d_x(N)$, with $s_1 < \ldots s_p$. Then $\{s_1, \ldots, s_p\}$ is a set of smallest finite elements of $d_x(Z)$, and*

$$N = Z_x(s_1) \cup \cdots \cup Z_x(s_{p-1}) \sqcup F$$

where $F \subset Z_x(s_p)$.

Proof. This is proved by induction on $|N|$, using Lemma **??** and Lemma **??**. □

If the neighbourhood $N = \{x, x_1, \ldots, x_k\}$ is a set of nearest neighbours of x with

$$d(x, x_1) \leq \cdots \leq d(x, x_k),$$

one says that $(x_1, x_2, \ldots, x_k)$ is a **sequence of k-nearest neighbours** for x.

Lemma 6. *Suppose that $\{y_1, \ldots, y_k\}$ is a set of distinct elements of $X - \{x\}$ with*

$$d(x, y_1) \leq d(x, y_2) \leq \cdots \leq d(x, y_k) < \infty.$$

If $(x_1, \ldots, x_k)$ is a sequence of k-nearest neighbours for x, then $d(x, x_i) \leq d(x, y_i)$ for $1 \leq i \leq k$.

Proof. $d(x, x_1) \leq d(x, y_1)$, since x_1 is a nearest neighbour.

Suppose that $d(x, x_i) \leq d(x, y_i)$ for $i \leq r$. Then

1) If $d(x, x_r) < d(x, y_{r+1})$ then $d(x, x_{r+1}) \leq d(x, y_{r+1})$ by minimality.

2) If $d(x, x_r) = d(x, y_{r+1})$ then y_{r+1} is a nearest neighbour of x in $Z - \{x_1, \ldots, x_r\}$, and so $d(x, x_{r+1}) = d(x, y_{r+1})$. □

Corollary 7. *Suppose that $(x_1, \ldots, x_k)$ and $(y_1, \ldots, y_k)$ are sequences of k-nearest neighbours for x. Then $d(x, x_i) = d(x, y_i)$ for all i.*

Corollary 8. *Suppose that W is a finite ep-metric space, and the inclusion $Z \subset W$ induces an ep-metric structure on the subset Z. Supppose that $x \in Z$. Suppose that $(w_1, \ldots, w_k)$ and $(z_1, \ldots, z_k)$ are sequences of k-nearest neighbours for x in W and Z, respectively. Then $d(x, w_i) \leq d(x, z_i)$ for $1 \leq i \leq k$.*

Lemma 9. *Suppose that $(x_1, \dots, x_k)$ is a sequence of nearest neighbours for x in Z, and that $\{y_1, \dots, y_k\}$ is a sequence of distinct elements of Z with $d(x, y_1) \leq \dots \leq d(x, y_k) < \infty$.*

If $d(x, x_i) = d(y, y_i)$ for all i, then $(y_1, \dots, y_k)$ is a sequence of nearest neighbours for x.

Proof. $d(x, y_1) = d(x, x_1) \leq d(x, z)$ for all z, so that y_1 is a nearest neighbour for x in Z.

Inductively, suppose that $\{y_1, \dots, y_i\}$ is a set of nearest neighbours for x.

Suppose that $d(x, x_i) = d(x, x_{i+1})$. Then $d(x, y_i) = d(x, y_{i+1})$, and so $\{y_1, \dots, y_{i+1}\}$ is a set of nearest neighbours.

If $d(x, x_i) < d(x, x_{i+1})$, then $\{y_1, \dots, y_i\} = \{x_1, \dots, x_i\}$ by comparing fibres $Z_x(s)$, so that y_{i+1} is the nearest neighbour of x in $Z - \{y_1, \dots, y_i\}$. □

We close this section with a discussion of k-complete neighbourhoods.

The image of the distance function $d_x : Z \to [0, \infty]$ has the form

$$\mathrm{Im}(d_x) = \{s_1, s_2, \dots\},$$

where there are strict inequalities $s_i < s_{i+1}$ for all i. The data set Z is a disjoint union of non-empty fibres of d_x:

$$Z = p_x^{-1}(s_1) \sqcup p_x^{-1}(s_2) \sqcup \dots = Z_x(s_1) \sqcup Z_x(s_2) \sqcup \dots.$$

For each s_i, there is a unique complete neighbourhood $Z(x, s_i)$ of x, with

$$Z(x, s_i) = d_x^{-1}(s_1) \sqcup \dots \sqcup d_x^{-1}(s_i).$$

The complete neighbourhoods of x form a finite ascending tower

$$Z(x, s_1) \subset Z(x, s_2) \subset Z(x, s_3) \subset \dots$$

Any complete neighbourhood N with $Z(x, s_i) \subsetneqq N$ must have strictly greater radius $s_N > s_i$.

Suppose that k is a positive integer and that $|Z| \geq k$. Then there is a smallest number i such that $|Z(x, s_i)| \geq k$. In this case, the neighbourhood $Z(x, s_i)$ is k-complete.

Alternatively, the k-complete neighbourhood N of x is the smallest complete neighbourhood such that $|N| \geq k$.

The element $x \in Z$ has a unique k-complete neighbourhood N in Z, provided that $|Z| \geq k$. The k-complete neighbourhood N is a well defined object, while there may be multiple sets of k-nearest neighbours of x

Lemma 10. *Suppose that $Z_1, Z_2 \subset \mathcal{U}$, and that $x \in Z_i$. Suppose that $N_i \subset Z_i$ is the k-complete neighbourhood of x in Z_i, and suppose that N is the k-complete neighbourhood of x in $Z = Z_1 \cup Z_2$. Then $N \subset N_1 \cup N_2$.*

Proof. The set N is a set of nearest neighbours for x in $Z_1 \cup Z_2$, and $Z_i \cap N$ is a set of nearest neighbours for x in Z_i.

In effect, if $z \in Z_i$ is not in $Z_i \cap N$ then z is not in N, so that $d(x,z) \geq s_N$, while $s_N \geq s_{Z_i \cap N}$.

If there is an $s < s_N$ such that $|Z_i(x,s)| \geq k$, then $|Z(x,s)| \geq k$ for $s < s_N$, and so $N = Z(x, s_N)$ is not k-complete. It follows that $|Z_i(x,s)| < k$ for $s < s_N$, and so $Z_i(x, s_N) \subset N_i$.

Thus, $N \subset N_1 \cup N_2$, as claimed □

Lemma **??** leads to a method of approximating k-complete neighbourhoods for a point x in a very large data set $\mathcal{U}$.

In effect, if $Z_i \subset \mathcal{U}, 1 \leq i \leq p$ is a collection of subsets of $\mathcal{U}$ with $x \in Z_i$, and if $N_i \subset Z_i$ is a k-complete neighbourhood of x in Z_i, then the k-complete neighboourhood N of x in $Z_1 \cup \cdots \cup Z_p$ is the k-complete neighbourhood of x in the much smaller object $N_1 \cup \cdots \cup N_p$.

3 Topological constructions

Suppose that (Z,d) is a finite ep-metric spac.

Suppose given a set of neighbourhoods N_x for each $x \in Z$. Recall that the neighbourhood N_x has a diameter $s(N_x) < \infty$.

Each neighbourhood N_x determines a filtered subcomplex $V(N_x) \subset V(Z)$ of the Vietoris-Rips complex $V(Z)$.

The inclusions $\{x,y\} \subset V(N_x)$, $y \in N_x - \{x\}$, induce filtered simplicial complex maps

$$R(N_x) := \vee_y \, \Delta^1_{\geq s} \subset V(N_x) \subset V(Z). \tag{1}$$

The copies of Δ^1 are defined by rays $\{x,y\}$ of weights s.

Remark 11. More properly, if the ray $\{x,y\}$ has weight $t = d(x,y)$, then the corresponding 1-simplex of $R(N_x)$ is the filtered simplex $\Delta^1_{\geq t}$ such that

$$(\Delta^1_{\geq t})_s = \begin{cases} \emptyset & \text{if } s < t, \text{ and} \\ \Delta^1 & \text{if } s \geq t. \end{cases}$$

It is better, sometimes, to say that $R(N_x)$ is **covered** by simplices $\Delta^1_{\geq s}$ corresponding to rays $\{x,y\}$ of weight s. This simply reflects the fact that the obvious map

$$\sqcup_y \, \Delta^1_{\geq s} \to \vee_y \, \Delta^1_{\geq s} = R(N_x)$$

is an epimorphism of filtered complexes.

The full union

$$R(N) = \cup_x \, R(N_x) \subset V(Z)$$

is the **ray subcomplex** of $V(Z)$, for the collection of neighbourhoods $N = \{N_x\}$.

The ray subcomplex $R(N)$ is a filtered (or weighted) graph. If the neighbourhoods N_x consist of k-nearest neighbours, then $R(N)$ is the k-nearest neighbours (kNN) graph.

The neighbourhoods N_x generate an abstract simplicial complex $V(N) \subset V(Z)$ whose simplices are the subsets $\sigma \subset N_x$ of the various neighbourhoods N_x. The resulting filtered simplicial complex can be written

$$V(N) = \cup_x V(N_x) \subset V(Z).$$

The subcomplex $V(N)$ of $V(Z)$ is called the **neighbourhood complex**.

The inclusions $R(N_x) \subset V(N_x)$ induce an inclusion $R(N) \subset V(N)$, so we have inclusions

$$R(N) \subset V(N) \subset V(Z) \tag{2}$$

of filtered complexes, with corresponding inclusions

$$R_s(N) \subset V_s(N) \subset V_s(Z) \tag{3}$$

of the various filtration stages.

The induced functions

$$\pi_0 R_s(N) \to \pi_0 V_s(N) \to \pi_0 V_s(Z)$$

in path components (or clusters) are surjective for all parameters t, since all complexes have the same vertex set, namely Z.

Remark 12. The neighbourhood complex $V(N) = \cup_x V(N_x)$ is covered by the subcomplexes $V(N_x)$, in the sense that there is a surjection

$$\bigsqcup_x V(N_x) \to V(N).$$

This covering has an associated Čech resolution, and there is a natural coequalizer

$$\bigsqcup_{x,y} V_s(N_x) \cap V_s(N_y) \rightrightarrows \bigsqcup_x V_s(N_x) \to V_s(N)$$

in simplicial sets, where s is the distance parameter. The path component functor preserves colimits, so there is a coequalizer

$$\bigsqcup_{x,y} \pi_0(V_s(N_x) \cap V_s(N_y)) \rightrightarrows \bigsqcup_x \pi_0 V_s(N_x) \to \pi_0 V_s(N)$$

in diagrams of sets, or clusters.

The directed set $\pi_0 V_s(N)$ is the cluster object given by the mapper construction for the covering of Z by the family of neighbourhoods $\{N_x\}$ [**?**].

Lemma 13. *Suppose that $x, y \in Z$. There is a path from x to y in $R(N)$ if and only if there is a sequence of elements*

$$x = x_0, x_1, \ldots, x_r = y$$

and neighbourhoods N_{x_i} of x_i, such that $N_{x_i} \cap N_{x_{i+1}} \neq \emptyset$ for all i.

Proof. Suppose that

$$x = z_0, \ldots, z_p = y$$

is a sequence of points such that $x_{i+1} \in N_{x_i}$ or $x_i \in N_{x_{i+1}}$ for neighbourhoods N_{x_i} and $N_{x_{i+1}}$ of x_i and x_{i+1}, respectively. If $x_{i+1} \in N_{x_i}$ then $N_{x_i} \cap N_{x_{i+1}} \neq \emptyset$. Similarly, if $x_i \in N_{x_{i+1}}$ then $N_{x_i} \cap N_{x_{i+1}} \neq \emptyset$.

Suppose, conversely, that $v \in N_{x_i} \cap N_{x_{i+1}}$. Then there is an edge $x_i \to v$ in N_{x_i} and an edge $x_{i+1} \to v$ in $N_{x_{i+1}}$, so that there is a path

$$x_i \to v \leftarrow x_{i+1}$$

through neighbourhoods. □

By definition, the ray complex $R(N)$ is a filtered subcomplex of $V(Z)$. The subcomplex $R_s(N) \subset V_s(Z)$ is generated by rays $\{x, y\}$ with $d(x, y) \leq s$.

We have the following analog of Lemma **??**:

Lemma 14. *Suppose that $x, y \in Z$. For each parameter value s, there is a path from x to y in $R_s(N)$ if and only if there is a sequence of elements*

$$x = x_0, x_1, \ldots, x_r = y$$

and neighbourhoods N_{x_i} of x_i, such that $(N_{x_i})_s \cap (N_{x_{i+1}})_s \neq \emptyset$ for all i.

4 Bounded neighbourhoods

4.1 k-bounded neighbourhoods

In some examples (such as stellar charts), it is natural that neighbourhoods N of x have bounded cardinality and radius: $|N| \leq k+1$ for some k and $s(N) \leq S$, with both k and S fixed.

From this point of view, for a fixed x, the **k-bounded neighbourhoods** N of x are the subsets of $Z(x, S)$ which contain x and have at most $k+1$ elements. Again, $Z(x, S)$ is the ball of radius S in Z, which is centred on x.

We assume that $k \geq 1$ henceforth.

A point x can have more than one k-bounded neighbourhood. The k-bounded neighbourhoods of x are ordered by inclusion, and the family has maximal elements. We have the following:

1) The maximal k-bounded neighbourhboods $N \subset Z(x, S)$ either have cardinality $k + 1$ or satisfy $N = Z(x, S)$.

2) All sets N of k-nearest neighbours with $s(N) \leq S$ are maximal.

3) If $N = \{x\}$ is maximal, then x is an isolated point for the parameter S.

The corresponding neighbourhood complex $V(k-N)$ is the filtered subcomplex of $V(Z)$ that is generated by the subobjects $V(N)$ for all k-bounded neighbourhoods N of all x, and $R(k-N)$ is the associated ray subcomplex. As in (**??**), we have a sequence of inclusions

$$R(k-N) \subset V(k-N) \subset V(Z).$$

If $t \leq S$ and $\sigma = \{x_0, \ldots, x_n\}$ is an n-simplex of $V(Z)_t$ with $n \leq k$, then σ is a k-bounded neighbourhood of x_0. In effect, σ has at most $k+1$ elements, of maximal distance $t \leq S$ from x_0. It follows that $V_t(k-N)_n = V_t(Z)_n$ for $n \leq k$, or that $\mathrm{sk}_k V_t(k-N) = \mathrm{sk}_k V_t(Z)$. In particular, $\mathrm{sk}_1 V_t(k-N) = \mathrm{sk}_1 V_t(Z)$ since $k \geq 1$, and so the simplicial sets $V_t(k-N)$ and $V_t(Z)$ have the same path components.

We have shown the following:

Lemma 15. *Suppose that $t \leq S$, and construct the neighbourhood complex $V(k-N)$ from k-bounded neighbourhoods as above. Then the function*

$$\pi_0 V_t(k-N) \to \pi_0 V_t(Z)$$

is a bijection.

Suppose that $t \geq S$, and that $\{x, y\}$ is a 1-simplex of $V_t(k-N)$. Then $\{x, y\} \subset N$ for a k-bounded neighbourhood N of some z. Further, $d(z, x) \leq s(N)$ and $d(z, y) \leq s(N)$, so that $d(z, x), d(z, y) \leq s(N) \leq S \leq t$. It follows that there is a path

$$x \leftarrow z \rightarrow y$$

in $R_t(k-N)$, and so the function

$$\pi_0 R_t(k-N) \to \pi_0 V_t(k-N)$$

is a bijection.

We have proved

Lemma 16. *Suppose that $t \geq S$. Then the induced function*

$$\pi_0 R_t(k-N) \to \pi_0 V_t(k-N)$$

is a bijection.

Write

$$R(k-N) = \cup_t \; R_t(k-N).$$

Then the map

$$\pi_0 R_t(k-N) \to \pi_0 R(k-N)$$

is a bijection for $t \geq S$, because $R_t(k-N) = R(k-N)$ in that range.

We therefore have the following:

Corollary 17. *The functions*

$$\pi_0 R(k-N) \leftarrow \pi_0 R_t(k-N) \rightarrow \pi_0 V_t(k-N)$$

are bijections for all $t \geq S$.

Write

$$N(x) = \cup_N \; V(N)$$

in $V(Z)$, where the union is indexed over all k-bounded neighbourhoods N of x. Let $R(x) \subset N(x)$ be the associated ray subcomplex. We have the inclusions

$$VR(x) \subset VN(x) \subset V(Z(x,S)).$$

Suppose that $t \leq S$ and $n+1 \leq k$. Suppose that $\sigma = \{x_0, \ldots, x_n\}$ is a non-degenerate n-simplex of $V(Z(x,S))_t$. Write

$$\sigma_x = \{x, x_0, \ldots, x_n\}.$$

Then $|\sigma_x| \leq k+1$, so that σ_x is a k-bounded neighbourhood of x, and so $\sigma = d_0\sigma_x$ is in the image of the composite

$$V(\sigma_x)_t \rightarrow VN(x)_t \rightarrow V(Z(x,S))_t.$$

It follows that $\mathrm{sk}_{k-1} V_t N(x) = \mathrm{sk}_{k-1} V_t(Z(x,S))$ for $t \leq S$. In particular, the map

$$\pi_0 V_t N(x) = \pi_0 V_t(Z(x,S))$$

is a bijection if $k \geq 2$ and $t \leq S$.

Suppose that $t \geq S$, and that y and z are vertices of $N(x)$. Then $d(x,y), d(x,z) \leq S \leq t$, and it follows that the map

$$* = \pi_0 R(x)_t \rightarrow \pi_0 N(x)_t$$

is a bijection.

Every $y \in Z(x,S)$ is a member of a k-bounded neighbourhood $\{x,y\}$ since $k \geq 1$. It follows that the maps

$$\pi_0 R(x)_t \rightarrow \pi_0 N(x)_t \rightarrow \pi_0 V_t Z(x,S)$$

are surjective.

We have proved:

Lemma 18. *Suppose that the complexes* $N(x)$ *and* $R(x)$ *are defined as above. Suppose that* $k \geq 2$. *Then we have the following:*

1) *If* $t \leq S$ *then the map* $\pi_0 N(x)_t \rightarrow \pi_0 V_t(Z(x,s))$ *is a bijection.*

2) *If* $t \geq S$ *then the maps*

$$* = \pi_0 R(x)_t \rightarrow \pi_0 N(x)_t \rightarrow \pi_0 V(Z(x,S))_t$$

are bijections.

4.2 Complete neighbourhoods

Suppose that Z is a finite ep-metric space, and that each $x \in Z$ has a fixed complete neighbourhood $N_x = Z(x, r_x)$. Form the associated filtered complexes

$$R(N) \subset V(N) \subset V(Z),$$

for Z and the system of neighbourhoods $N = \{N_x\}$.

Example 19. Suppose that $S > 0$ is a fixed distance parameter and $k > 1$ is a fixed integer.

Say that a neighbourhood N_x of $x \in Z$ is **complete k-bounded** if N_x has the form

$$N_x = Z(x, s_x) \cap Z(x, S),$$

where $Z(x, s_x)$ is the unique k-complete neighbourhood of x (see Section 2).

There are two possibilities: $N_x = Z(x, s_x)$, in which case N_x is k-complete, or $N_x = Z(x, S)$ and $|Z(x, S)| < k$. In either case, the neighbourhood N_x is uniquely determined and is complete.

The use of *complete* k-bounded neighbourhoods gives a different perspective for the stellar chart example. For a fixed (and appropriate) distance S and positive integer k, the complete k-bounded neighbourhoods N_x of stars x in a globular cluster would be k-complete neighbourhoods of small radius, while stars in an outer spiral arm are more likely to have neighbourhoods N_x of smaller cardinality.

Lemma 20. *Suppose that Z is a finite ep-metric space, and that each $x \in Z$ has a fixed complete neighbourhood $N_x = Z(x, r_x)$.*

1) Suppose that $t \leq r_x$ for all x. Then the functions

$$\pi_0 R_t(N) \to \pi_0 V_t(N) \to \pi_0 V_t(Z)$$

are bijections.

2) Suppose that $t \geq r_x$ for all x. Then the map

$$\pi_0 R_t(N) \to \pi_0 V_t(N)$$

is a bijection.

3) Suppose that $t \geq S \geq r_x$ for all x. Then the map

$$\pi_0 R_S(N) \to \pi_0 R_t(N)$$

is a bijection.

Proof. For 1), suppose that $\{x, y\}$ is a 1-simplex of length t. Then $y \in N_x$ since $t \leq r_x$, and $\{x, y\}$ is a ray of N_x. It follows that there are equalities of 1-skeleta

$$\mathrm{sk}_1 R_t(N) = \mathrm{sk}_1 V_t(N) = \mathrm{sk}_1 V_t(Z),$$

and the statement follows.

For statement 2), suppose that $\{x, y\}$ is a 1-simplex of $V_t(N_z) \subset V_t(N)$. Then there are 1-simplices $x \leftarrow z \rightarrow y$ in $V_t(N_z)$ since $t \geq r_z$. This is true for all z, and it follows that the function

$$\pi_0 R_t(N) \rightarrow \pi_0 V_t(N)$$

is a bijection.

To prove statement 3), every ray (1-simplex) $\{x, y\}$ of $R(N)$ has length $\leq S$, so that $R_S(N) = R_t(N)$. □

Corollary 21. *Suppose that $t \geq S \geq r_x$ for all $x \in Z$. Then the inclusion $V_S(N) \subset V_t(N)$ of neighbourhood complexes induces a bijection*

$$\pi_0 V_S(N) \rightarrow \pi_0 V_t(N).$$

Proof. The Corollary follows from statements 2) and 3) of the Lemma **??** □

Remark 22. Suppose that Z' is the subset of elements $x \in Z$ such that $Z(x, r_x) = \{x\}$, and let $Z" = Z - Z'$. Then

1) $V_t(Z) = Z' \sqcup V_t(Z'')$,

2) $V(N)_t = Z' \sqcup V(N'')_t$,

3) $R(N)_t = Z' \sqcup R(N'')_t$,

for $t \leq r_z$, all z, where Z' is a discrete set. Here,

$$Z'' = \cup_{y \in Z''} \, Z(y, r_y),$$

and N'' is the system of neighbourhoods $Z(y, n_y)$ for $y \in Z''$.

5 The UMAP construction

One starts with a neighbourhood N_x for each vertex x of a data set Z, with positive weights $d(x, y)$ for each $y \in N_x - \{x\}$. The subset $\{x, y\}$ for such a y is said to be a ray.

The weight $d(x, y)$ defines an ep-metric space structure on the set $\{x, y\}$. Form the ep-metric space

$$Z_x = \vee_{y \in N(x) - \{x\}} \, \{x, y\},$$

from the rays $\{x, y\}$, for each $x \in Z$. This structure is extended to an ep-metric space structure (Z, D_x) on the full set of vertices Z of Γ, by setting

$$(Z, D_x) = (\sqcup_{z \in Z - Z_x} \, \{z\}) \sqcup Z_x$$

in ep-metric spaces.

The ep-metric space

$$(Z, D) = \vee_{x \in Z} \ (Z, D_x)$$

and the UMAP complex

$$V(Z, N) = \vee_{x \in Z} \ V(Z, D_x)$$

are formed by amalgamating along vertices (elements of Z), in ep-metric spaces and filtered complexes, respectively.

It is crucial, for these ep-metric space constructions, to know that the category of ep-metric spaces is cocomplete — see Section 1.

The following excision statement for path components is Lemma 2 of [**?**]:

Theorem 23. *The canonical map $V(Z, N) \to V(Z, D)$ induces isomorphisms*

$$\pi_0 V(Z, N)_s \xrightarrow{\cong} \pi_0 V(Z, D)_s$$

for s finite.

Theorem **??** is proved by observing that distances in (Z, D) are computed from paths through neighbourhoods N_x.

We shall need the following local computation:

Lemma 24. *Suppose that $y \in N(x) - \{x\}$ defines the ray $\{x, y\}$. Then*

$$d(x, y) = D_x(x, y).$$

in Z_x.

Proof. The number $D_x(x, y)$ is the minimum of all sums $\sum_j \ d(x_j, x_{j+1})$, for paths

$$P : \ x = x_0, x_1, \ldots, x_p = y$$

through rays in Z_x. The ray $\{x_{p-1}, y\}$ must be the ray $\{x, y\}$, so that

$$D_x(x, y) \geq \sum_j \ d(x_j, x_{j+1}) \geq d(x, y).$$

The subobject $\{x, y\}$ is a ray, so that $\{x, y\}$ is a path, and $D_x(x, y) \leq d(x, y)$. $\square$

Each $y \in N_x - \{x\}$ determines an inclusion of filtered complexes

$$V(\{x, y\}) \subset V(Z_x, D_x) \subset V(Z, D).$$

The simplicial set $V_t(\{x, y\}$ consists of vertices $\{x, y\}$ for $t < d_x(x, y)$, and has 1-simplices

$$\{x\} \subset \{x, y\} \supset \{y\}$$

for $t \geq d_x(x, y)$.

Remark 25. Recall that $V(\{x, y\})$ is the barycentric subdivison of a filtered 1-simplex that would be defined by imposing a total order on the set $\{x, y\}$.

Suppose that $\{x, u\}$ is a ray in N_x and that $\{y, v\}$ is a ray of N_y, and consider the composite monomorphisms

$$V(\{x, u\}) \subset V(Z_x) \subset V(Z), \quad V(\{y, v\}) \subset V(Z_y) \subset V(Z).$$

Suppose that $d_x(x, u) \leq d_y(y, v)$.

Generally, $V(X) = BP(X)$, where $P(X)$ is a poset of generating simplices. In the case at hand, therere is a pullback diagram

$$\begin{array}{ccc} BP(\{x,u\} \cap \{y,v\}) & \longrightarrow & BP(\{y,v\}) \\ \downarrow & & \downarrow \\ BP(\{x,u\}) & \longrightarrow & BP(Z) \end{array}$$

The intersection $\{x, u\} \cap \{y, v\}$ is at most a 2-element set. If $\{x, u\} \cap \{y, v\} = \emptyset$ the pullback is empty, and if $\{x, u\} \cap \{y, v\}$ is a point the pullback is a point.

If $\{x, u\} \cap \{y, v\}$ is a 2-element set, then $\{x, u\} = \{y, v\}$, and there is a commutative diagram

$$\begin{array}{ccc} BP(\{x,u\}) & \xleftarrow{\theta} & BP(\{y,v\} \\ & \searrow \quad \swarrow & \\ & BP(Z) & \end{array}$$

where θ "reduces weight". It follows, in this case, that there is a pullback

$$\begin{array}{ccc} BP(\{y,v\}) & \xrightarrow{1} & BP(\{y,v\}) \\ {\scriptstyle\theta}\downarrow & & \downarrow \\ BP(\{x,u\}) & \longrightarrow & BP(Z) \end{array} \tag{4}$$

The ray complex $R(N_x) \subset V(Z_x)$ is the wedge of rays

$$R(N_x) = \vee_{y \in N_x - \{x\}} \, V(\{x, y\}).$$

The filtered complex monomorphisms

$$\phi_x : R(N_x) \to V(Z_x, D_x) \to V(Z, D),$$

together define a monomorphism

$$\phi : R(N) = \cup_{x \in Z} \, R(N_x) \subset V(Z),$$

and we say that the union $R(N)$ is the ray subcomplex of $V(Z)$. The ray complex $R(N)$ is a weighted graph.

The ray complex $R(N)$ is a union of (or is covered by) filtered subcomplexes $V(\{x, y\})$, which are defined by rays $\{x, y\}$ and their weights $d(x, y)$. The intersections (pullbacks)

$$\begin{array}{ccc} BP(\{y,v\} \cap \{y,v\}) & \longrightarrow & BP(\{y,v\}) \\ \downarrow & & \downarrow \\ BP(\{x,u\}) & \longrightarrow & R(N) \end{array}$$

are constructed in $V(Z)$ as above, since $R(N) \subset V(Z)$ is a monomorphism. It follows that the ray complex $R(N)$ is a union of rays, with possible adjustments of weights in intersections, as in the pullback diagram (**??**).

Remark 26. The present description of the ray complex $R(N)$ is independent of distances in the space (Z, D). It generalizes the description of the ray complex that appears in Section 3, which uses a fixed ambient ep-metric.

There is, finally, an excision result that makes $R(N)$ a candidate for the UMAP graph, as follows:

Theorem 27. *The filtered complex map $\phi : R(N) \subset V(Z, D)$ induces isomporphisms*

$$\phi_* : \pi_0 R_s(N) \xrightarrow{\cong} \pi_0 V_s(Z, D)$$

for all $s \geq 0$.

Proof. The proof is similar to that of Theorem **??**.

The map ϕ is the identity on vertices, so the functions ϕ_* are surjective.

Suppose that there is a 1-simplex $\{z, w\}$ of $V(Z, D)_s$. Then there is a path

$$z = x_0, x_1, \ldots, x_p = w$$

through rays $\{x_i, x_{i+1}\}$ such that

$$\sum_{i=0}^{p} d(x_i, x_{i+1}) \leq s.$$

by Lemma **??**. But then $d(x_i, x_{i+1}) \leq s$ for all i, so that z and w are in the same path component of the simplicial set $R(N)_s$.

It follows that the functions ϕ_* are injective. □

6 Weighted directed graphs

A weighted directed graph Γ consists of a set of edges $e : x \to y$, such that each edge e has a weight $w(e) > 0$. For the present discussion, the vertices of Γ are faces of the edges. I write Z for the set of vertices of Γ.

Trivial examples are given by 1-skeleta $\mathrm{sk}_1 K$ of oriented simplicial complexes K, with weights $w(e) = 1$ for each 1-simplex $e : x \to y$, and such that every vertex is a face of some non-degenerate 1-simplex e.

A weighted directed graph Γ is said to be **sparse** if all vertices x of Γ have low valence. This means that each vertex of Γ is in the boundary of a small (i.e. computable) number of edges.

Suppose that x is a vertex of a transfer graph Γ. A **path** $P : x \dashrightarrow y$ from x to another vertex y in Γ is a string of edges

$$P : \ x = y_0 \xrightarrow{e_1} y_1 \xrightarrow{e_2} \dots \xrightarrow{e_p} y_p = y. \tag{5}$$

Say that the integer p is the length $\ell(P)$ of the path P.

Remark 28. The collection of all paths $P : x \dashrightarrow y$ in the graph Γ form a weighted graph $P(\Gamma)$ having the same vertices as the graph Γ. The paths $P : x \dashrightarrow y$ and $Q : y \dashrightarrow z$ are composeable: the concatenation of P with Q defines a path $P \circ Q : x \dashrightarrow z$. Thus, $P(\Gamma)$ has more structure: $P(\Gamma)$ is the free category on the graph Γ.

The path graph $P(\Gamma)$ is not sparse in general.

The weight $w(P)$ of the path P can be defined by

$$w(P) = \min_i \{w(e_i)\}. \tag{6}$$

Remark 29. The definition of the weight of a path is somewhat arbitrary, and depends on applications. The assignment of (**??**) is motivated by graphs of data transfers, which are discussed below. One could, alternatively, set

$$w(P) = \sum_i \ w(e_i).$$

Fix a positive integer k.

The **neighbourhood** $N_k(x)$ is the collection of all vertices y, which appear in paths

$$Q : \ z_0 \to z_1 \to \cdots \to z_p,$$

having length $p = \ell(Q) \leq k$, such that $x = z_i$ for some i

We assign a weight (or distance) $d(x, y)$ for all y in the neighbourhood $N_k(x)$.

For $y \in N_k(x)$, define the **weight sum** $\Sigma(x, y)$ by

$$\Sigma(x, y) = \Big(\sum_{P : x \dashrightarrow y, \ell(P) \leq k} w(P) \Big) + \Big(\sum_{Q : y \dashrightarrow x, \ell(P) \leq k} w(Q) \Big).$$

In a graph of transactions, the weight sum $\Sigma(x, y)$ represents the total value of all transactions between x and y. If $\Sigma(x, y)$ has a large value, then there is more business between x and y, and these objects should be closer in some

sense. To express this relationship, use the Shannon information function to define a distance

$$d(x, y) = e^{-\Sigma(x,y)} \tag{7}$$

for $y \in N_k(x)$.

Other approaches to defining a distance $d(x, y)$ for the vertices y of $N_k(x)$ are certainly possible.

We end up with a computable neighbourhood $N_k(x)$ of vertices in a sparse directed graph Γ for each of its vertices x, with distances (weights) $d(x, y)$ for $y \in N_k(x) - \{x\}$.

These are the inputs for the UMAP construction, which is described in Section 4.

Example 30 (Data transfers). A data transfer $e : x \to y$ from a computer account x to a different account y has a weight $w(e)$, which is the number of bytes transferred. The transfer e also has source and target time stamps, $s(e)$ and $t(e)$, respectively, with $s(e) < t(e)$. Thus (provisionally), a graph Γ of data transfers has edges $e : (x, s(e)) \to (y, t(e))$ with $x \neq y$, and its vertices consist of pairs (x, t), where x is a computer account and t is either a source or a target timestamp for some edge.

There may be multiple vertices (x, t) for a fixed account x. Suppose that $t_0 < t_2 < \cdots < t_p$ are the timestamps for a fixed account x. Say that the list $\{t_0, \ldots, t_p\}$ is the simplex of timestamps for the account x.

For $x \neq y$, an edge $E : (x, s) \to (y, t)$ of the transfer graph Γ consists of a transfer $e : (x, s(e)) \to (y, t(e))$, together with relations $s \leq s(e)$ and $t(e) \leq t$ in the simplices of timestamps for the accounts x and y, respectively. The weight $w(E)$ is the weight $w(e)$ of the transfer e. The set vertices Z of Γ consists of all pairs (x, s) of accounts x and timestamps s of transfers.

If all timestamps lie within a small enough interval, then the transfer graph Γ is sparse.

This example motivates the definitions of weights of paths and distances within neighbourhoods that are seen above.

Explicitly, a path

$$P : \ (x, s) = x_0 \xrightarrow{E_1} x_1 \xrightarrow{E_2} \ldots \xrightarrow{E_p} x_p = (y, t)$$

in Γ consists of edges

$$E_i : x_i = (x_i, s_i) \to (x_{i+1}, s_{i+1}) = x_{i+1}$$

with $x_i \neq x_{i+1}$, and each such edge has weight $w(E_i)$.

The weight $w(P)$ of the path P is defined by

$$w(P) = \min_i \{w(E_i)\},$$

as in (**??**)). The weight $w(P)$ represents the maximum amount of data that could be transferred from (x, s) to (y, t) along the path P.

Fix a positive integer k and an element $x = (x, s)$ in the transfer graph Γ, and define the neighbourhood $N_k(x)$ as vertices of paths crossing x of length at most k.

The weight sum $\Sigma(x, y)$ for $y \in N_k(x)$ is defined by

$$\Sigma(x, y) = (\sum_{P: x \dashrightarrow y, \ell(P) \leq k} w(P)) + (\sum_{Q: y \dashrightarrow x, \ell(P) \leq k} w(Q)),$$

and the weight $d(x, y)$ of the ray $\{x, y\}$ has the form

$$d(x, y) = e^{-\Sigma(x,y)}.$$

Remark 31 (**Undirected graphs).** The directed structure for the graph Γ is a central feature of the examples discussed above. Analogous local to global methods apply equally well to construct ep-metric spaces and UMAP complexes for undirected graphs.

Suppose that Ω is a sparse weighted graph, with weights $w(e)$ for the edges e of Ω. One assumes that the vertices of Ω are faces of its edges.

Suppose that x is a vertex of Ω. Say that $y \in N_k[x]$ if there is a path (path), or string of edges

$$P : \ x = x_0 \overset{e_1}{\leftrightarrow} x_1 \overset{e_2}{\leftrightarrow} \dots \overset{e_p}{\leftrightarrow} x_p = y$$

with $p \leq k$.

Again there are choices, but define the weight $w(P)$ of the path $P : x \leftrightarrow y$ by

$$w(P) = \min_i \{w(e_i)\}.$$

Fix a vertex x and a positive integer k. Define $N_k(x)$ to be the set of all vertices of Ω which lie on paths of length $\ell(P)$ at most k that pass through x.

Write

$$\Sigma(x, y) = \sum_{P: x \dashrightarrow y, \ell(P) \leq k} w(P),$$

and set

$$d(x, y) = e^{-\Sigma(x,y)}$$

for $y \in N_k(x)$.

One uses the weights $d(x, y)$ to construct an ep-metric on the neighbourhood $N_k(x)$. These ep-metrics patch together, to give an ep-metric on the full set Z of vertices of Ω.

7 Bags of words

In the "bag of words" model for natural language processing (see, for example [**?**]), one starts with a collection $C = \{C_1, \dots, C_N\}$ of **documents** C_i, where each $C_i = (t_{i,1}, \dots, t_{i,M_i})$ is a sequence of **tokens** (ie. words, phrases, etc.), with possible repetitions. The sequence C is the **corpus**.

The sequence C_i is a function

$$C_i : \underline{M}_i \to \mathcal{T},$$

where $\mathcal{T}$ is the set of distinct tokens in all C_i, and $\underline{M}_i = \{1, 2, \ldots, M_i\}$. The sequence C_i may have repeats, so the function C_i is not injective in general.

The usual thing is to amalgamate some tokens (by root words, or whatever), to form a surjective map $\ell : \mathcal{T} \to \mathcal{L}$. The set $\mathcal{L}$ is the **vocabulary** and its elements are called **words**.

Write p for the composite function

$$p : \sqcup_i \, \underline{M}_i \xrightarrow{C} \mathcal{T} \xrightarrow{\ell} \mathcal{L},$$

and let $p_i : \underline{M}_i \to \mathcal{L}$ be the restriction of p to the summand $\underline{M}_i$.

We assume that there are no common tokens ("stop words") or rare tokens in the set $\mathcal{T}$, however these are determined. This means that the fibres $p^{-1}(w)$ of the function p are neither too large nor too small, and in particular are computationally manageable. The function p and its fibres are the objects of interest for this discussion.

The fibres $p^{-1}(w)$ are the instances of the word $w \in \mathcal{L}$ in the corpus C.

Remark 32. In more generality, we could have functions $p_i : \underline{M}_i \to Z$ which cover a set Z, in the sense that the amalgamated function

$$p : \sqcup_i \, \underline{M}_i \to Z$$

is surjective. Here, the restriction of p to the summand $\underline{M}_i$ is p_i. One assumes that the fibres $p_i^{-1}(z)$ for $z \in Z$ are computationally manageable (or tractable in the sense of the next section), as is the collection of functions $\{p_i\}$.

Subject to size assumptions on the cardinals $\underline{M}_i$ and the collection of functions p_i, the following discussion can be applied in such a setting.

One could even replace the sets $\underline{M}_i$ with metric spaces in the discussion that follows.

Write $\mathcal{L}_i$ for the image of the restricted function

$$p_i = \ell_i : \underline{M}_i \xrightarrow{C_i} \mathcal{T} \xrightarrow{\ell} \mathcal{L}.$$

The composite p_i restricts to a surjective function $p_i : \underline{M}_i \to \mathcal{L}_i$, and there is a commutative diagram of functions

$$\begin{array}{ccc} \underline{M}_i & \xrightarrow{p_i} & \mathcal{L}_i \\ \downarrow & & \downarrow \\ \sqcup_i \, \underline{M}_i & \xrightarrow[p]{} & \mathcal{L} \end{array}$$

in which the vertical maps are inclusions.

Each set $\underline{M}_i$ has a metric d with $d(x,y) = |y - x|$.

Suppose that r is a positive integer. Fix a word $v \in \mathcal{L}$, and suppose that $p_i^{-1}(w)_{\leq r}$ is the set of all elements $y \in p_i^{-1}(w)$ such that $d(x,y) \leq r$ for some $x \in p_i^{-1}(v)$. Then we have

$$p_i^{-1}(w)_{\leq r} = p_i^{-1}(w) \cap (\cup_{x \in p^{-1}(v)} \ [x - r, x + r])$$

in the set $\underline{M}_i$. The subsets $p_i^{-1}(w)_{\leq r}$ filter the fibre $p_i^{-1}(w)$. Observe that $p_i^{-1}(v)_{\leq r} = p_i^{-1}(v)$.

Set

$$d_i[r](v,w) = \sum_{x \in p_i^{-1}(v), y \in p_i^{-1}(w), d(x,y) \leq r} d(x,y), \tag{8}$$

and define

$$d[r](v,w) = \sum_i \ d_i[r](v,w)$$

for all $v, w \in \mathcal{L}$.

The number $d_i[r](v,w)$ is non-zero if and only if there are elements $x \in p_i^{-1}(v)$ and $y \in p_i^{-1}(w)$ such that $d(x,y) \leq r$, and $d[r](v,w) \neq 0$ if and only if $d_i[r](v,w) \neq 0$ for some i.

In particular, $d_i[r](v,v)$ is the sum of the distances $d(x,y)$ between $x, y \in p_i^{-1}(v)$ such that $d(x,y) \leq r$, and $d_i[0](v,v) = 0$. It follows that $d[r](v,v)$ can be non-trivial for $r > 0$, and $d[0](v,v) = 0$.

Take all elements x of the fibres $p_i^{-1}(v)$ and form all intervals $[x - r, x + r]$ in $\underline{M}_i$. The union

$$N_v[r] = \cup_i \ (\cup_{x \in p_i^{-1}(v)} \ p_i[x - r, x + r]) \subset \mathcal{L}. \tag{9}$$

is a neighbourhood of v in $\mathcal{L}$.

Observe that $N_v[0] = \{v\}$. Also, $N_v[r] \subset N_v[s]$ for $r \leq s$, and $\cup_r \ N_v[r] = \mathcal{L}$, so that the subsets $N_v[r]$ filter the set of words $\mathcal{L}$.

Subject to fixing a positive integer r, the set $N_v[r]$ is a neighbourhood for $v \in \mathcal{L}$, and the number $d[r](v,w)$ is the weight of $w \in N_v[r]$.

As in Section 5, the UMAP construction assembles the weighted neighbourhoods $(N_v[r], d[r])$, $v \in \mathcal{L}$, to form a the UMAP complex $V(\mathcal{L}, N[r])$, an ep-metric space $(\mathcal{L}, D[r])$, and a ray complex $R(N[r]) \subset V(\mathcal{L}, D[r])$, all of which compute the same clusters.

8 Sampling

Suppose that the universal data set $\mathcal{U}$ has an ep-metric space structure, but with no other information.

In this case, one approximates (or discovers) a neighbourhood N_x for a given point $x \in \mathcal{U}$ with a brute force method that is based on sampling techniques and construction of k-complete neighbourhoods within samples.

Suppose that Z is a randomly chosen subset of $\mathcal{U}$ (a sample), and that Z is tractable in the sense that there is a cardinality bound $|Z| \leq M$, where data sets of size at most M can be analyzed by available computational devices. We assume that $x \in Z$.

For such a subset Z the distance function $d_x : Z \to [0, \infty)$, with $d_x(z) = d(x, z)$, can be computed, and the image $d_x(Z)$ of d_x defines a tractable subset of the interval $[0, \infty)$. The set Z is a disjoint union of fibres

$$Z = \sqcup_{s \in d_x(Z)} \ d_x^{-1}(s)$$

of the distance function d_x.

Suppose that k is a fixed choice of positive integer with $k \leq |Z|$.

The element x has a uniquely defined k-complete neighbourhood N in Z, as in Section 2, which is the smallest complete neighbourhood $Z(x, s)$ such that $|Z(x, s)| \geq k$. The neighbourhood N is a union of fibres $d_x^{-1}(t)$ for smallest values of t,

This construction can be repeated, in parallel, for an appropriately sized collection of samples $Z_1, \ldots, Z_p$ that contain x, with distance functions $d_x : Z_i \to [0, \infty)$. Each sample Z_i has a uniquely defined (and computable) k-complete neighbourhood N_i of x, and the k-complete neighbourhood N of x in the union $\cup_i \ Z_i$ is a k-complete neighbourhood of x in the smaller object $\cup_i \ N_i$, by Lemma **??**.

There are various ways to invoke the samples Z_i:

1) Starting with a k-complete neighbourhood N_x of x in a sample Z_x, choose samples Z_y for each $y \in N_x$, with associated k-complete neighbourhoods $N_y \subset Z_y$. The union $\cup_{y \in N_x} \ N_y$ contains a k-complete neighbourhood N'_x, which is potentially a better approximation of a k-complete neighbourhood of x in the universe $\mathcal{U}$.

This sequence of steps is an analogue of the k-nearest neighbour algorithm of [**?**].

2) The determination of a k-complete neighbourhood N of x in V for some $V \subset \mathcal{U}$ can be extended to larger subsets of $\mathcal{U}$, subject to computational constraints, by adding more tractable samples to V. This is again a simple application of Lemma **??**.

3) If $Z_1, \ldots, Z_p$ is a tractable collection of tractable samples in $\mathcal{U}$, then we can find a k-complete neighbourhood N_y in $Z = \cup_i \ Z_i$ for any $y \in Z$. The corresponding subcomplexes $V(N_y) \subset V(Z)$ and $R(N_y) \subset V(N_y)$ determine filtered subcomplexes

$$\cup_{y \in Z} \ R(N_y) \subset \cup_{y \in Z} \ V(N_y),$$

which lead to a UMAP-style analysis that computes the clusters of $V(Z)$, and approximates the clusters of all $V(\mathcal{U})$.

The sampling technique displayed here is completely brute force. It only approximates clusters and neighbouhoods of points, and does not speak to the entire data set $\mathcal{U}$.

The method can be refined in the presence of global constraints, such as the local uniformity assumption of [**?**] that produces sets of k-nearest neighbours up to a probability estimate.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Wei Dong, Charikar Moses, and Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th International Conference on World Wide Web, WWW'11*, pages 577–586, New York, 2011. ACM.

[3] J.F. Jardine. Stability for UMAP. Preprint, arXiv: 2011.13430 [math.AT], 2020.

[4] Leland McInnes. UMAP Documentation. github.com/lmcinnes/umap, 2023.

[5] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint, arXiv: 1802.03426 [stat.ML], 2020.

[6] Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, editors, *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association, 2007.