Observing Information: Applied Computational Topology.

Timothy Porter

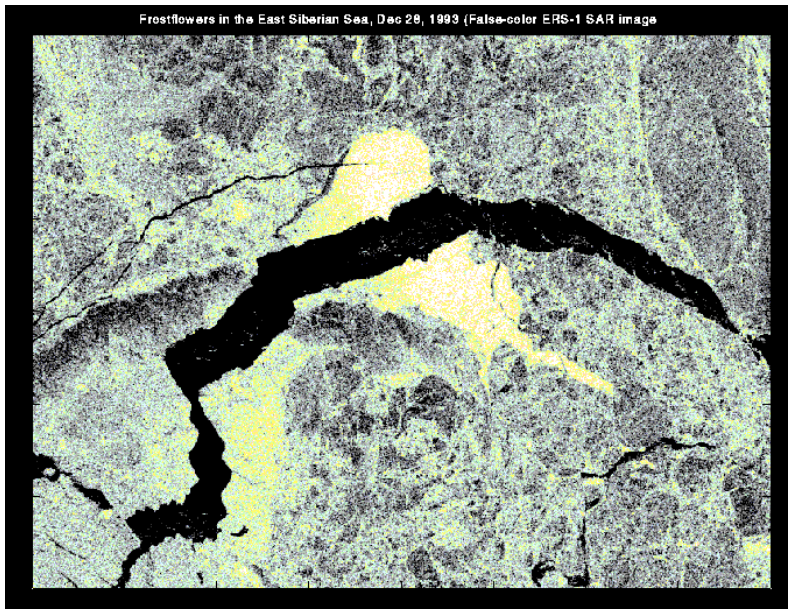
Bangor University, and NUI Galway

April 21, 2008

What is the 'geometric' information that can be gleaned from a data cloud?
Some ideas either already tried or 'under study':

- determining suspect patterns in medical scans, e.g. for the detection of possible cancers;
- finding significant geometric patterns within geological data;
- finding the type of microfossils in rock samples;
- determining flow patterns of pollutants in rivers, lakes and estuaries and comparing them, *in detail*, with theoretical predictions;

- finding active sites and voids in molecules for ‘designer drugs’;
- finding the size, shape and position of floes and leads in pack ice.



Learning to Crawl

Before attempting any of these deep scientific applications, we would need to learn how difficult it is to even crawl, let alone walk! We will only *crawl* in this lecture!

- The following is a noisy sample from a circle.

- Problem: develop automated methods to analyse the HOLE in the sample!

Spaces and Information.

Data is often looked at 'spatially', i.e. modelled by 'spaces' and spaces are made up of 'points'.

Points about 'points':

- Do 'spaces' really have 'points' or is that just a useful device for handling something else? What is the point of 'points'?
- 'Spaces' may correspond to some geometric object, but may also be used *just to organise data* which may not be spatial in essence. They may contain other measurements such as temperature, or discrete, perhaps 'yes/no', information, (see next frame!)
- In Physics, some of the problems of Quantum Relativity may be avoided by throwing out points and having a 'pointless model'!

Objects and attributes, Chu spaces, and Formal Context Analysis

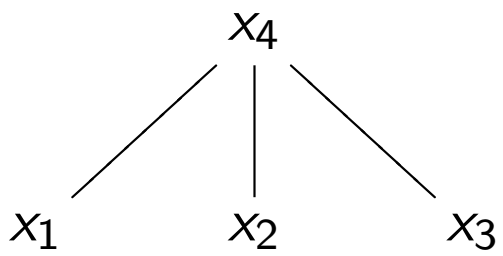
Example of data organisation: From any set of observed attributes, build 'spatial' objects that indicate the interrelationships and any hidden 'concepts'.
A Chu space, $\mathcal{C}$, is given as $\mathcal{C} = (O, \models, A)$, where O and A are sets, called the sets of *objects* and of *attributes* and $\models \subseteq O \times A$ is a relation: $o \models a$ reads 'object o has attribute a'.
The information in such a context has its own internal logical structure, from which some 'inferences' can be extracted. This is used in AI, in ontology and in natural language processing.

Example

Four objects and three attributes. The table says if an object x_i has attribute a_j or not. Of course, if it does, we put a 1 in the (i, j) position, and if not, we put 0.

$\mathcal{C}$	a_1	a_2	a_3
x_1	1	0	0
x_2	0	1	0
x_3	0	0	1
x_4	1	1	1

We can represent this by a graphical diagram, in this case the Hasse graph of the partially ordered set:



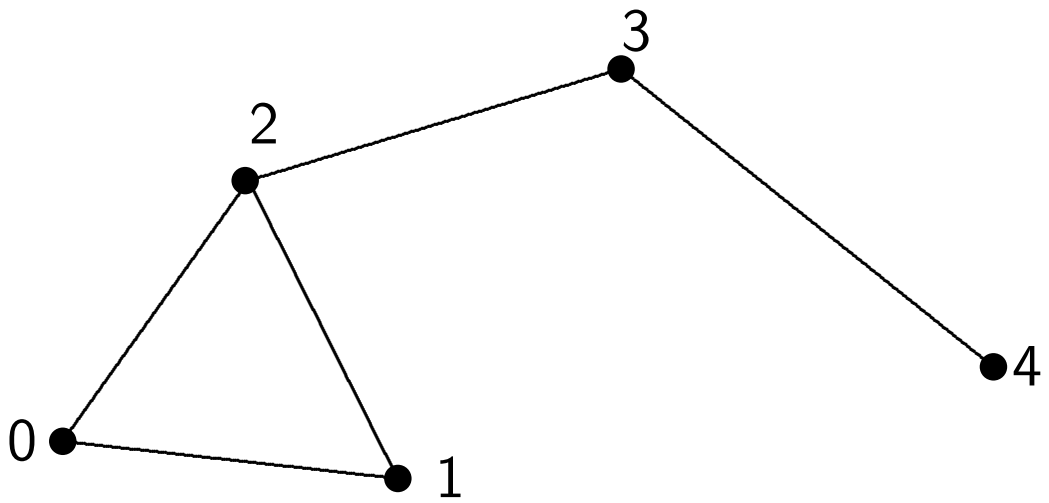
The relation is the partial order as shown. It is also the dual of the lattice of non-empty open sets of a three point discrete space, so is also spatial.

Graphs, Simplicial Complexes, Triangulations, etc.

- Graphs, or 'networks', are 1-dimensional diagrams extensively used to represent information.
- Graphs are specified by giving a set of vertices $V(K)$ and a set of edges $E(K)$ with incidence information.

Example of a graph

$V(K) = \{0, 1, 2, 3, 4\}$
$E(K) =$
$\{\{0\}, \{1\}, \{2\}, \{3\}, \{4\}, \{0, 1\}, \{0, 2\}, \{1, 2\}, \{2, 3\}, \{3, 4\}\}$
It looks like:

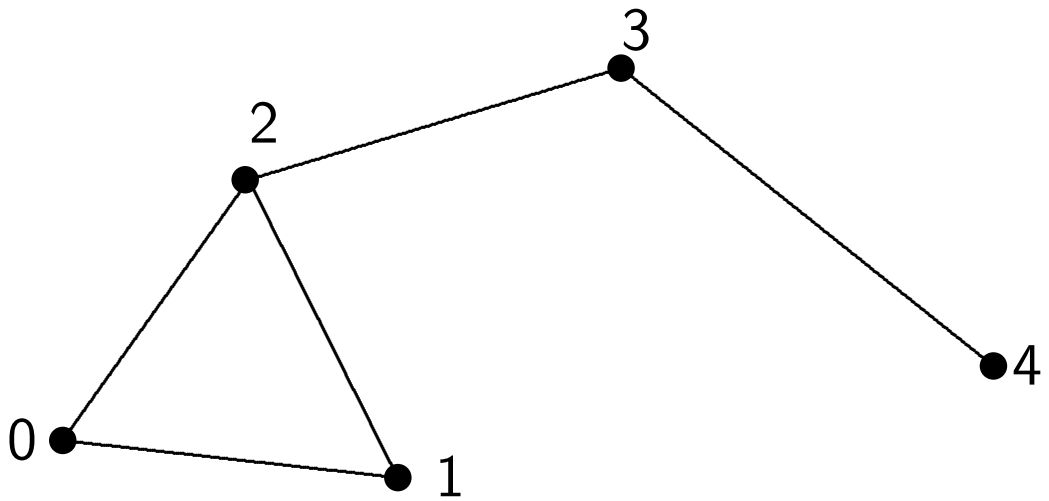


Simplicial Complexes

Graphs are not adequate to represent the multifaceted higher dimensional relations in data. A better combinatorial gadget for that is the simplicial complex:
A simplicial complex K is a set of objects, $V(K)$, called vertices and a set, $S(K)$, of finite non-empty subsets of $V(K)$, called simplices such that if $\sigma \subseteq V(K)$ forms a simplex, then any non-empty subset of σ does as well.
(So not just edges, possibly higher dimensional things as well.)

Example of a simplicial complex

$V(K) = \{0, 1, 2, 3, 4\}$
$S(K) =$
$\{\{0\}, \{1\}, \{2\}, \{3\}, \{4\}, \{0, 1\}, \{0, 2\}, \{1, 2\}, \{2, 3\}, \{3, 4\}, \{0, 1, 2\}\}$
It looks like:



BUT THIS TIME the triangle is meant to be filled in!

Triangulation.

- A triangulation (K, f) of a space X consists of a simplicial complex K and an identification of it as a realisation of a simplicial complex: $f : |K| \to X$.
  (We will usually confuse the geometric model $|K|$ with X and so will call X, itself, a polyhedron in this case.)
- We use triangulations to 'control' spaces, but are they something 'imposed on the space' or should be think of them as 'built' from the 'observations' of the 'space'? In other word, make the data 'king' not the space!
- Any sample of data points will give a polyhedron in various ways, **but** that polyhedron may be strongly dependent on the sample. That dependency needs more study.

Homology

By using some algebra, taking formal sums of simplices in a complex, one can get some computable algebraic and numerical invariants of the complex, for instance, homology groups, $H_i(K)$. These are typically vector spaces or similar structures, and their dimension tells one the number of holes of different dimension in the space.

e.g. $dim\, H_0(K)$ is the number of components of K; $dim\, H_1(K)$ is the number of 1-dimensional holes, so $dim\, H_1(circle)$ is 1, whilst $dim\, H_1(figure\ eight)$ is 2, and so on.

These dimensions are called the Betti numbers of K.

Čech and Vietoris: the observational solution from the 1920s and 1930s

Instead of a triangulation,

- Assume we are given an (open) cover $\mathcal{U}$ of our 'space' X, so $\mathcal{U}$ is a family of (open) sets, U, of X and for any $x \in X$ there is some $U \in \mathcal{U}$ that contains it.
- The 'observational' idea is that we probe X and each probe can measure things in a small patch. '*Physically, the idea is that what we actually observe are interactions between bounded regions of space-time.*'(Christensen-Crane,'04)
- To each such (open) covering we can attach two simplicial complexes one due to Vietoris (1927) the other to Čech (early 1930's).

Čech: the nerve of X

Denoted $N(X, \mathcal{U})$ or $N(\mathcal{U})$ if no confusion will arise.

- vertices, the open sets in $\mathcal{U}$,

and

- simplices, those finite families of (open) sets in $\mathcal{U}$ whose intersection is non-empty:
- i.e. $\{U_0, \ldots, U_n\} \subset \mathcal{U}$ is a simplex of $N(\mathcal{U})$ if and only if

$$\bigcap_{i=0}^{n} U_i \neq \emptyset.$$

- WE NEED A PICTURE so draw one on the board!

Vietoris complex of X

Denoted $V(X,\mathcal{U})$ or simply $V(\mathcal{U})$, the Vietoris complex reverses the roles of points and open sets:

- vertices are the points of X itself

and

- simplices : $\langle x_0, \ldots, x_n \rangle$ is an n simplex of $V(\mathcal{U})$ if there is a $U \in \mathcal{U}$ containing all the vertices x_i, $i = 0, \ldots, n$.

Dowker (1952)

Let $R \subset X \times Y$ be a relation. (In our topological case, $X = X$, $Y = \mathcal{U}$ and $xRU = x \in U$.) Any such relation determines two simplicial complexes:

1. $K = K_R$: - the set of vertices is the set, Y;
   p-simplex of K is a set $\{y_0, \ldots, y_p\} \subseteq Y$ such that there is some $x \in X$ with xRy_j for $j = 0, 1, \ \ldots, \ p$.
2. $L = L_R$: - the set of vertices is the set X;
   - a p-simplex of K is a set $\{x_0, \ldots, x_p\} \subseteq X$ such that there is some $y \in Y$ with x_iRy for $i = 0, 1, \ \ldots, \ p$.

In general, no 'spatial' context is needed (cf. extensive applications in **AI**) and any metric data is not used. We have:

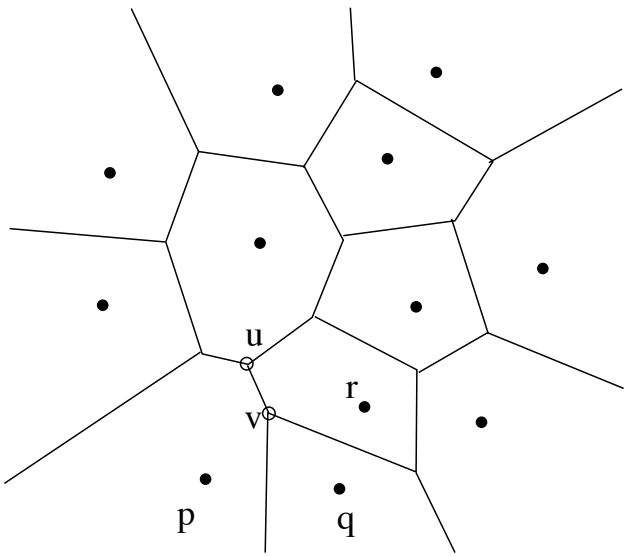
Theorem: The homology of the two complexes is the same.

so we can use either.

Computational substitutes in the metric case.

For the last part of the talk, we will concentrate on the metric case and **Topological Data Analysis**.
We had these two complexes from 'classical algebraic topology'.
They are usually not computationally feasible as such. Various replacements are used. They exploit the metric structure of much data.
Usual assumption: *The data is sampled from some 'idealised' subspace X of some $\mathbb{R}^n$*, (but both the ambient and intrinsic metrics may be used).
We need to recall Voronoi diagrams and the related Delaunay triangulations given by the sample.

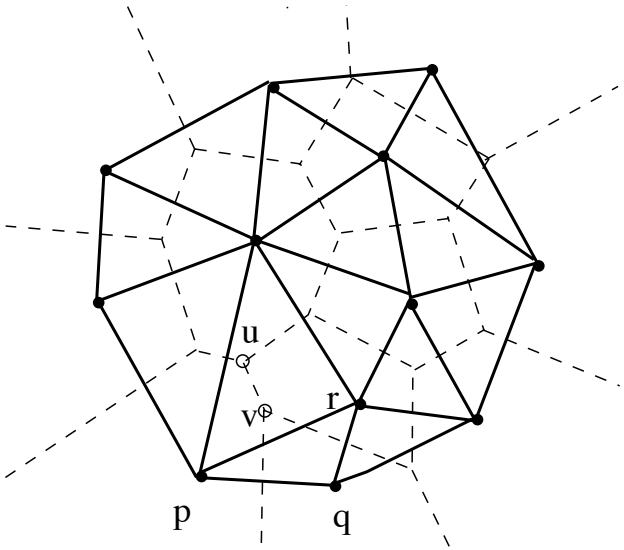
Voronoi diagrams. Let P be a set of data points in $\mathbb{R}^n$. The *Voronoi diagram* of P denoted V_P is a collection of *Voronoi cells* V_p, one for each point $p \in P$, where V_p is the set of all points in $\mathbb{R}^n$ that are closer or at least equidistant to p than to any other point in P. [1]



[1] Play the Voronoi game at http://www.voronoigame.com/VoronoiGameApplet.html

Delaunay triangulation. There is an associated *dual* structure to Voronoi diagram V_P, called the *Delaunay triangulation* denoted D_P. Formally, we define D_P as a simplicial complex where

$$D_P = \{\sigma \mid \bigcap V_p \neq \emptyset \text{ where } p \text{ is any vertex of } \sigma\}.$$



Simplicial Complex Approximations:

$X \subset \mathbb{R}^n$, a subspace; $Z \subset X$ a finite set of sample points. Need:

1. A construction $\mathcal{S} = \mathcal{S}(Z)$ of a simplicial complex depending on Z and possibly on additional parameters, but not depending on X itself;
2. A *similarity result* comparing X with $\mathcal{S}(Z)$ under reasonable conditions on Z as a sample of X, and for some choice of values for the additional parameters.

A typical additional parameter might be some notion of *sampling scale* $R \geq 0$. This can sometimes be interpreted as an amount of blurring or 'fuzziness' applied to Z. Varying R and / or the sample Z, we hope to capture 'qualitative' information on the idealised X. Often for two values $R \leq R'$, the constructions will give nested simplicial complexes,

$$\mathcal{S}(Z, R) \subseteq \mathcal{S}(Z, R').$$

We will quickly look at some examples.

The Čech complex.

This replaces the arbitrary open cover of the nerve construction, by little open balls around data points. The radius used gives a nesting parameter, R:

- Vertex set : all data points in Z;
- Parameter: $R > 0$, nested;
- Definition: the p-simplex $\sigma = [z_0, z_1, ..., z_p]$ belongs to Čech(Z, R) if and only if the closed Euclidean balls $B(z_j, R/2)$, $j = 0, 1, \ldots, p$ have non-empty common intersection.
  (so we are using classical Čech with nice round open discs as the open sets.)

The Rips complex.

This is a variant of the Čech complex which is easier to calculate.

- Vertex set : all data points in Z;
- Parameter: $R > 0$, nested;
- Definition: the p-simplex $\sigma = [z_0, z_1, ..., z_p]$ belongs to Rips(Z, R) if and only if for every edge $[z_j, z_k]$, $0 \leq j < k \leq p$, we have $||z_j - z_k|| \leq R$.
  (so each 'edge' is of length less than R.)

The α-shape complex, $A(Z, R)$: (Edelsbrunner)

- Vertex set : all data points in Z;
- Parameter: $R > 0$, nested;
- If V_z is the closed Voronoi cell of z, then define the α-cell for z to be the convex set $\alpha(z, R) = B(z, R) \cap V_z$;
- the p-simplex $\sigma = [z_0, z_1, ..., z_p]$ belongs to $A(Z, R)$ if and only if the α-cells, $\alpha(z_j, R/2)$, $j = 0, 1, \ldots, p$ have non-empty common intersection.

α-complex shape technology (developed by Geomagic), was used in the examination of the damaged re-entry shield tiles of the space shuttle, Endeavour. They allowed accurate 3-D reconstruction of the damaged tiles from scanned data and so helped to assess the extent of the damage prior to re-entry.

Persistent homology

Any of these approximations can give a homology and Betti numbers (for that construction and that value of R):
$\beta_i(Z,R) = rank H_i(\mathcal{S}(Z,R))$.
If the construction is a 'nested' one then if $R \leq R'$, we have complexes,

$$\mathcal{S}(Z,R) \subseteq \mathcal{S}(Z,R')$$

and induced maps

$$H_i(\mathcal{S}(Z,R)) \to H_i(\mathcal{S}(Z,R')).$$

Algebraically we can compute persistent Betti numbers $\beta_i(R,R')$ for every pair $R \leq R'$.

Interpretation

Intuitively $\beta_i(R, R')$ counts the number of i-dimensional holes in $\mathcal{S}(Z, R)$ which remain open when we thicken the complex to $\mathcal{S}(Z, R')$.

Produce bar codes or interval graphs.

For each dimension i get a set of closed intervals above an axis parametrised by R.

Long intervals correspond to large holes and thus to genuine features. Small intervals are usually regarded as 'noise'.

Where to now?

Experiments and theory so far (mostly by the Stanford and Duke research teams) have looked at feature detection and topological invariants from very large data sets, both artificial and 'real', but always with the assumption that a polyhedron underlies the data.

Plans

1: to see if it is possible to detect non-polyhedral behaviour in artificial data generated, initially, from fractal spaces such as the dyadic solenoid and the Menger cube.

2. Another area is that of data evolving in time. This should be modelled by 'spaces evolving through time'! The algebra needed is harder and may be a stiff challenge, but the resulting problem is potentially very useful and interesting. It relates to many areas of Theoretical Computer Science and Physics.

The End

T.P., Galway, April 2008