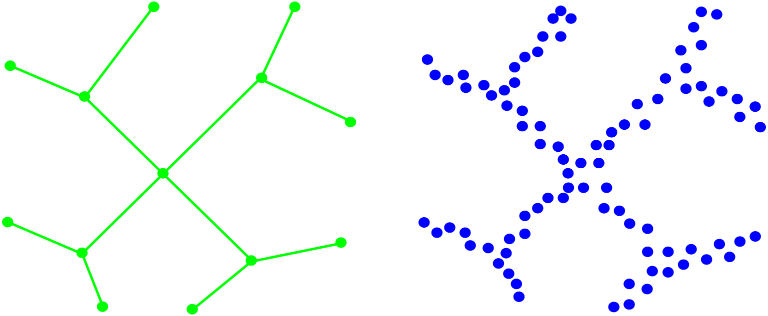
Persistent homotopy types

of noisy samples of graphs in the plane

Vitaliy Kurlin, http://kurlin.org
Durham University, UK

Noisy point clouds around graphs



Problem : given only a blue point cloud $C \subset \mathbb{R}^2$ around a green planar graph $\Gamma \subset \mathbb{R}^2$, detect a likely *structure* of Γ (e.g. the homotopy type of Γ) under some conditions when C is close to Γ.

Related work on noisy data

Metric graph reconstruction from noisy data. Aanjaneya, Chazal, Chen, Glisse, Guibas, Morozov. Int J Comp Geometry Appl, 2011.

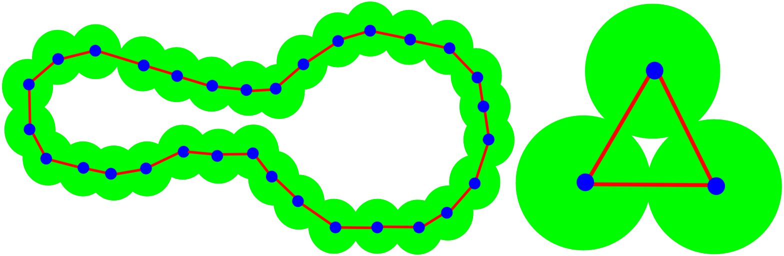
Input: a large metric graph Y (the shortest path distance) approximating an unknown graph X.

Output: a small metric graph $\hat{X}$ close to X.

Proved: $\hat{X}$ is almost isometric to X if Y is close enough to X and edges of X are not too short.

Complexes associated to a cloud

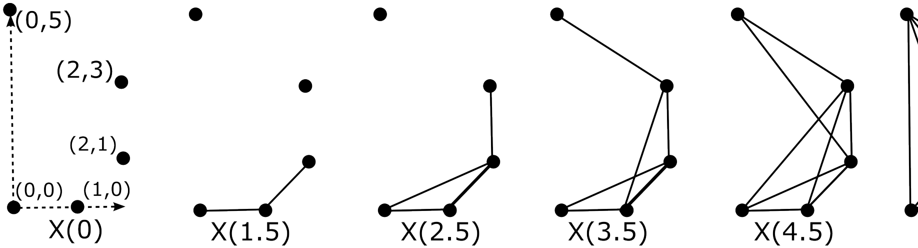
Def : for a cloud $C \subset \mathbb{R}^m$ and $\varepsilon > 0$, the Čech complex $\text{Čh}(\varepsilon)$ has vertices from C, simplices spanned by vertices $v_1, \ldots, v_k$ if $\cap_{i=1}^k B_\varepsilon(v_i) \neq \emptyset$.



The Vietoris-Rips complex $\mathrm{VR}(\varepsilon)$ has simplices spanned by $v_1, \ldots, v_k$ if distances $d(v_i, v_j) \leq \varepsilon$.

1-skeleton depending on ε

1-dimensional skeleton $X(\varepsilon)$ of Čh and VR for the cloud of 5 points $C \subset \mathbb{R}^2$ on the left picture.



It can be hard to manually find a good value of ε.

Capturing a homotopy type

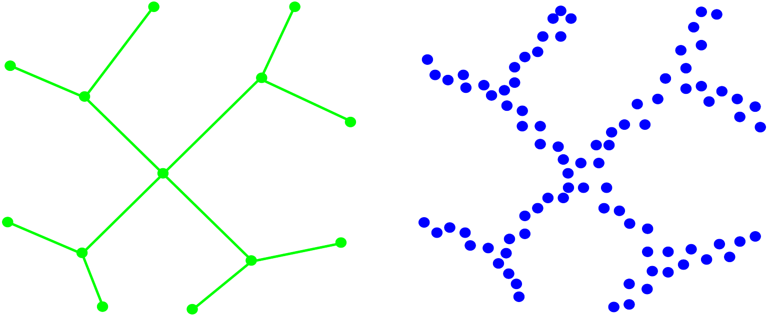
Nerve lemma for a point cloud $C \subset \mathbb{R}^m$ says: its abstract Čech complex $\check{\mathrm{C}}\mathrm{h}(\varepsilon)$ has the *homotopy type* of the ε-offset $C^\varepsilon = \cup_{a \in C} B_\varepsilon(a) \subset \mathbb{R}^m$.

The complex $\mathrm{VR}(\varepsilon)$ is built from the graph $X(\varepsilon)$. Also $\check{\mathrm{C}}\mathrm{h}(\varepsilon) \subset \mathrm{VR}(2\varepsilon) \subset \check{\mathrm{C}}\mathrm{h}(2\varepsilon)$ for any $\varepsilon > 0$.

$\check{\mathrm{C}}\mathrm{h}(\varepsilon), \mathrm{VR}(\varepsilon)$ have high-dimensional simplices even for $C \subset \mathbb{R}^2$, witness complexes are simpler.

Parameter-less reconstruction

Our aim is to reconstruct Γ from a close sample without user-defined parameters when possible.



Simplest case: reconstructing isolated vertices is equivalent to clustering a given cloud $C \subset \mathbb{R}^2$.

Persistence-based clustering

Persistence-based clustering in Riemannian manifolds. Chazal, Guibas, Oudot, Skraba. Proceedings Sympos Comp Geometry 2011.

ToMATo: Topological Mode Analysis Tool.

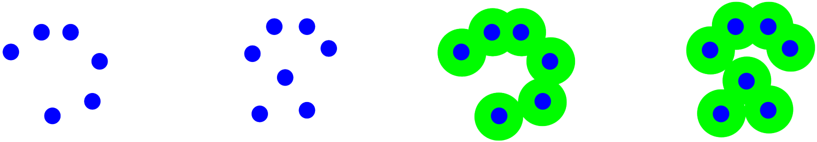
Input: neighborhood graph (Rips with fixed ε), density estimator f, threshold τ for peaks of f.

Proved: there is a range of τ when $\#$clusters$=$ $\#$peaks with a high probability.

Single edge clustering

$C \subset \mathbb{R}^2$, 1-dimensional skeleton $X(\varepsilon)$ evolves:

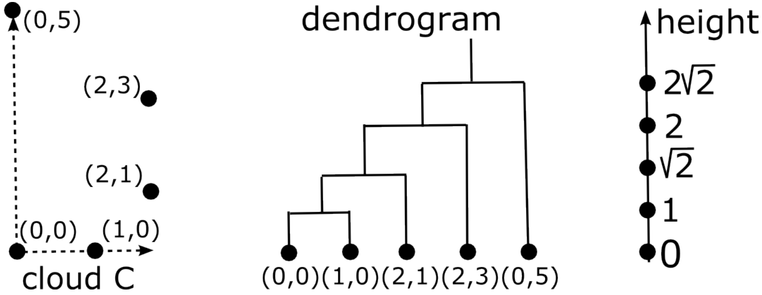
persistent components



Persistent connect. components of $X(\varepsilon)$ living over a long interval of ε are likely *clusters* of C.

Dendrogram of clustering

Def : a hierarchical clustering produces nested partitions represented by the dendrogram: each internal node is a cluster merged from smaller 2+ clusters at the node's children.



Choosing a distance threshold

Multivariate data analysis using persistence-based filtering and signatures. Rieck, Mara, Leitte. IEEE Trans Vis Comp Graphics 2012.

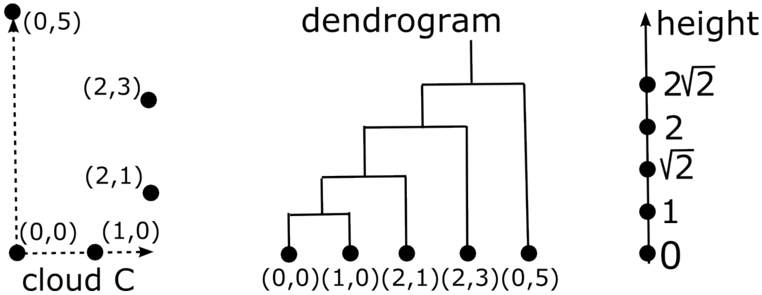
The distance threshold ε for clusters is from the dendrogram of the single link clustering.

Input: $k = \#$neighbors in a density estimator.

No guarantees given when $\#$clusters is correct.

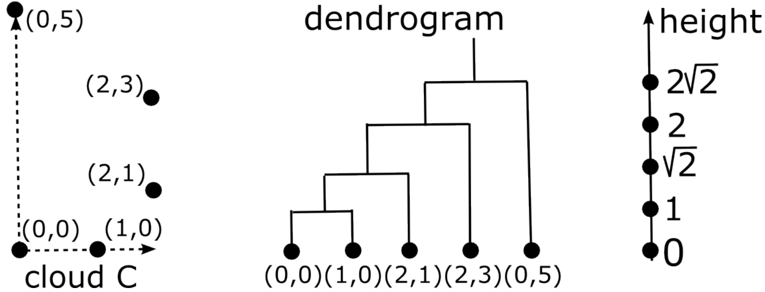
Persistent clusters

Def: in a general dendrogram, clusters merge at $n-1$ crit. heights $0 = h_0 < h_1 < \cdots < h_{n-1}$. A partition with the *longest life span* $s = h_i - h_{i-1}$ is persistent. If $i = 1$, take 1 cluster instead of n.



Associated probability

For $s = h_i - h_{i-1}$, the probability $P = \dfrac{s}{h_{n-1}}$.



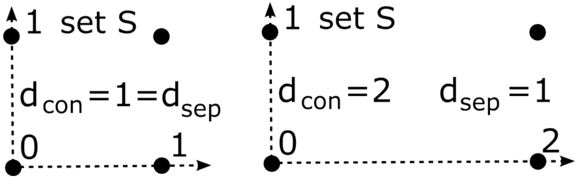
1st result: 1 cluster, $P = \frac{1}{2\sqrt{2}} \approx 35\%$.

2nd result: 2 clusters, $P = \frac{2\sqrt{2}-2}{2\sqrt{2}} \approx 30\%$.

3 clusters: $\frac{2-\sqrt{2}}{2\sqrt{2}} \approx 20\%$. 4 clusters: $\frac{\sqrt{2}-1}{2\sqrt{2}} \approx 15\%$.

Well-disconnected sets

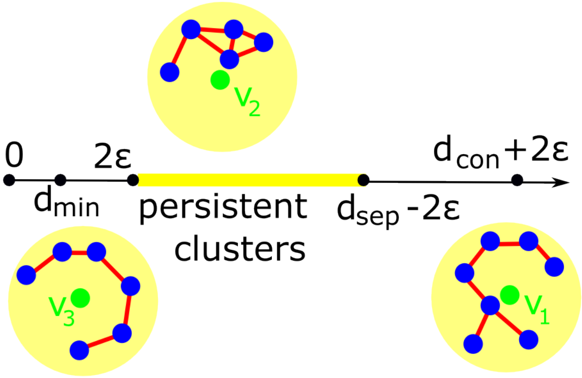
Def: for a triangulable set $S \subset \mathbb{R}^m$, consider the minimum distance $d_{sep}(S)$ between any connected components of S. Let $d_{con}(S) =$ min distance when $\frac{1}{2}d_{con}$-offset of S is connected.



The set S is well-disconnected if $d_{con} < 2d_{sep}$.

Finding persistent clusters

Claim: if a cloud C is ε-close to a set $S \subset \mathbb{R}^m$ and $d_{con}(S) + 8\varepsilon \leq 2d_{sep}(S)$, then the persistent clusters of C correctly detect components of S.



Sharp condition on persistence

Example: $S = \{0, 1, 2\} \subset \mathbb{R}$, $d_{sep} = 1 = d_{con}$.

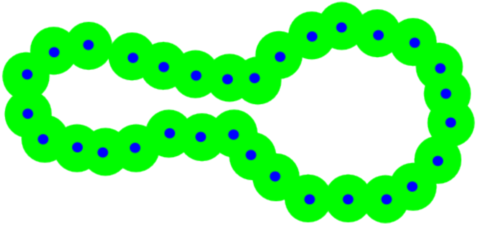
Take ε-close cloud $C = \{-\varepsilon, \varepsilon, 1-\varepsilon, 2+\varepsilon\}$.

Crit. heights: $h_1 = 2\varepsilon$, $h_2 = 1 - 2\varepsilon$, $h_3 = 1 + 2\varepsilon$.

To get 3 clusters $\{\pm\varepsilon\} \cup \{1-\varepsilon\} \cup \{2+\varepsilon\}$, we need $h_2 - h_1 = 1 - 4\varepsilon > h_3 - h_2 = 4\varepsilon$, so $\varepsilon < \frac{1}{8}$.

Distance function of a cloud

Def : for a compact set (e.g. a cloud) $C \subset \mathbb{R}^m$, define $d_C : \mathbb{R}^m \to \mathbb{R}$, $d_C(a)$ is the distance from $a \in \mathbb{R}^m$ to the closest point from the set $C \subset \mathbb{R}^m$

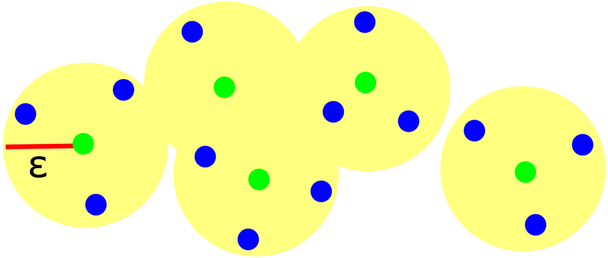


A sublevel set $d_C^{-1}[0, \varepsilon]$ is the union of balls with the radius $\varepsilon > 0$ and centers at the points of C.

The distance between clouds

Def : the distance between clouds $C, C' \subset \mathbb{R}^2$ is

$$d(C, C') = ||d_C - d_{C'}|| = \sup_{a \in \mathbb{R}^2} |d_C(a) - d_{C'}(a)|.$$



Geometrically, $d(C, C')$ is the smallest $\varepsilon > 0$ such that $C' \subset \cup_{a \in C} B_\varepsilon(a)$ and $C \subset \cup_{a \in C'} B_\varepsilon(a)$.

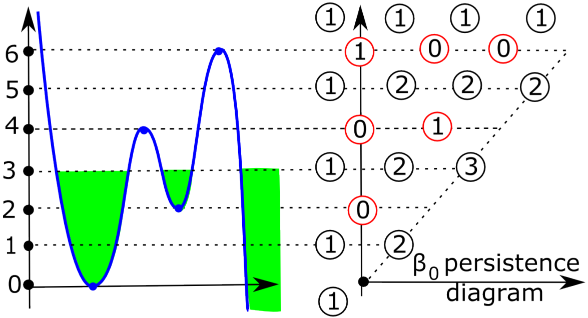
Persistent homology theory

Def : for a cloud $C \subset \mathbb{R}^2$, complexes $\{\mathrm{VR}(\varepsilon)\}$ with inclusions $\mathrm{VR}(\varepsilon) \subset \mathrm{VR}(\varepsilon')$ for any $\varepsilon < \varepsilon'$ lead to the persistence space $\{H_k(\mathrm{VR}(\varepsilon))\}$ with coefficients in a field F and induced linear maps $\varphi_k(\varepsilon, \varepsilon') : H_k(\mathrm{VR}(\varepsilon)) \to H_k(\mathrm{VR}(\varepsilon'))$ for $\varepsilon < \varepsilon'$.

$f : M \to \mathbb{R}$, take sublevels $M(\varepsilon) = f^{-1}(-\infty, \varepsilon]$. Let $0 < \varepsilon_1 < \cdots < \varepsilon_m$ be all *critical* values when $V(\varepsilon_i - \delta) \to V(\varepsilon_i + \delta)$ aren't isomorphisms, small δ. Let $t_0 < \varepsilon_1 < t_1 < \varepsilon_2 < \cdots < t_{m-1} < \varepsilon_m < t_m$.

Persistence diagrams

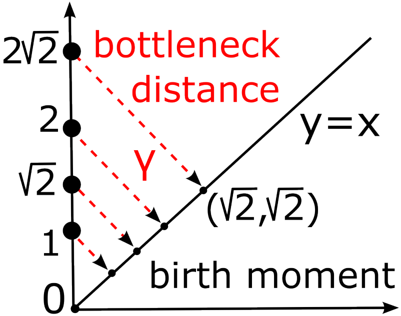
Def : the persistence diagram of $\{V(\varepsilon)\}$ is the set of $(\varepsilon_i, \varepsilon_j) \in \mathbb{R}^2$ for all $i < j$ with multiplicities $\mu_{ij} = \beta(i-1, j) - \beta(i, j) + \beta(i, j-1) - \beta(i-1, j-1)$, where $\beta(i, j) = \mathrm{rank}\,(\mathrm{image}\,(V(t_i) \to V(t_j)\,)\,)$.



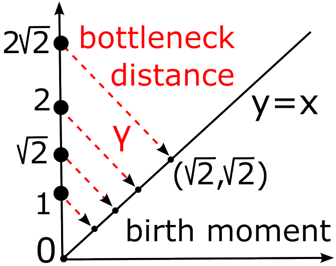
Distance between diagrams

Let P be $\{(x,x) \in \mathbb{R}^2\} \cup \{$a finite set of points$\}$.

Def : $d_B(P,Q) = \inf_\gamma \sup_{a \in P} |a - \gamma(a)|$ over all 1-1 maps $\gamma : P \to Q$ is the bottleneck distance.



Stability of persistence

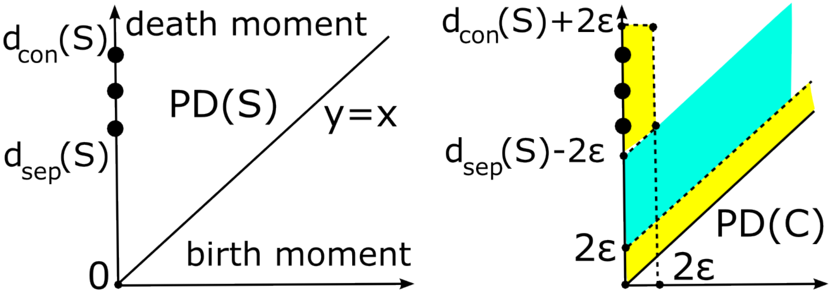


Stability of Persistence Diagrams. Edelsbrunner, Cohen-Steiner, Harer. Discr. Comp. Geometry 2007. **Proved**: $d_B(D(f), D(g))| \leq ||f - g||_\infty$.

Any ε-perturbation of a point cloud $C \subset \mathbb{R}^2$ deforms the persistence diagram by at most ε.

Stable persistent clusters

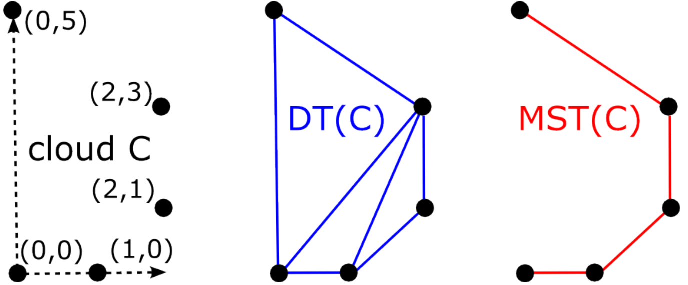
All components of $S \subset \mathbb{R}^m$ live from 0. Any noise of a cloud C can appear only in yellow areas.



Correct #clusters in the range $[2\varepsilon, d_{sep}(S) - 2\varepsilon]$, longest when $2\varepsilon \leq d_{sep} - 4\varepsilon \geq d_{con} - d_{sep} + 4\varepsilon$.

Delaunay triangulation and MST

For a cloud $C \subset \mathbb{R}^2$, a Delaunay triangulation *DT* has no point of C inside the circumcircle of any triangle. A minimum spanning tree *MST* has vertices at C and minimum total length.



How to find persistent clusters

Fact: for a cloud C of n points, $MST \subset DT$ can be found in $O(n \log n)$-time using $O(n)$ space.

Idea: critical heights in single link clustering are the lengths of $n-1$ edges in $MST(C)$, which can be sorted in $O(n \log n)$ time to find the longest life span and a few alternatives.

So $MST(C)$ contains *all* 0*-dim persistence* of $X(\varepsilon)$, no need to try many threshold values ε.

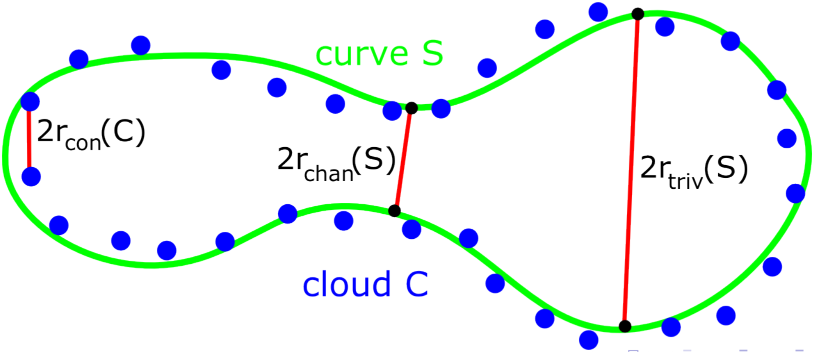
Critical radii for β_1

Def: for a triangulable set $S \subset \mathbb{R}^m$, consider

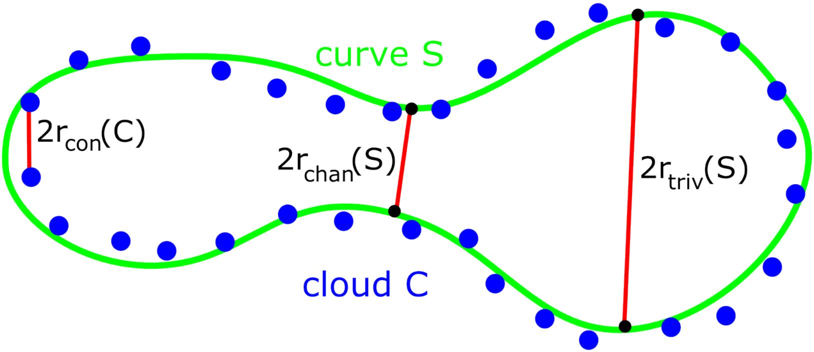
$r_{chan}(S) = \min \varepsilon$ when $\beta_1(S^\varepsilon)$ starts changing.

Let $r_{triv}(S) = \min \varepsilon$ when $\beta_1(S^\varepsilon) = 0$ after that.

$r_{con}(C) = \min \varepsilon$ when $X(\varepsilon)$ becomes connected.



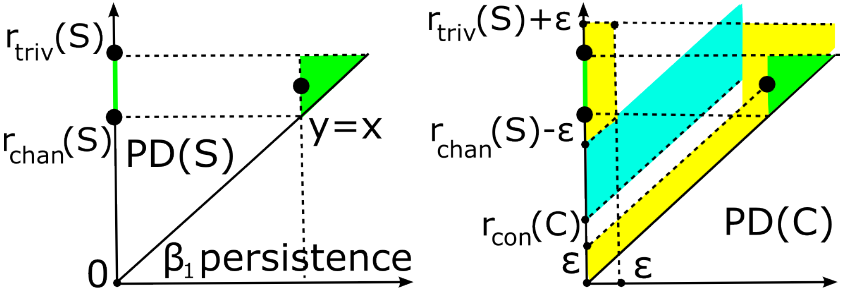
Existence of persistent β_1



Claim: if a cloud C is ε-close to a set $S \subset \mathbb{R}^m$,
$r_{triv}(S) + r_{con}(C) + 3\varepsilon \leq 2r_{chan}(S) \geq 4r_{con}(C) + 2\varepsilon$,
then $\beta_1(S) = \beta_1(\check{C}h_2(\varepsilon))$ with longest life span.

β_1 with the longest life span

Any noise of C can appear only in yellow areas.



Correct β_1 in $[r_{con}(C), r_{chan}(S) - \varepsilon]$, longest life span if $r_{con} \leq r_{chan} - \varepsilon - r_{con} \geq r_{triv} - r_{chan} + 2\varepsilon$.

Reeb graph of a height function

Def: for $f : X \to \mathbb{R}$, the Reeb graph $R_f(X)$ is the quotient $X/\sim$, where $a \sim b \Leftrightarrow a, b$ are in the same connected component of $f^{-1}(c)$.

Data skeletonization via Reeb graphs.
Ge, Safa, Belkin, Wang. NIPS 2011.

Proved: if a complex $K \sim$ deform retracts to ε-close graph G and $4\varepsilon <$ min edge length of G, there is a 1-1 map between loops of $R_f(K)$, G.

Persistent β_1 of Reeb graphs

Difficulty: for complexes $K_1 \subset \cdots \subset K_m$, Reeb graphs $R_f(K_i)$ aren't a filtration, even zigzag.

Reeb Graphs: Approximation and Persistence. Dey, Wang. Discrete Comp Geometry 2012.

Proved: all persistent β_1 of $R_f(K_i)$ can be found in $O(n^4)$ time, $n =$ size of the 2-skeleton of K_m.

Plane shadow of Rips complex

Vietoris-Rips complexes of planar point sets.
Chambers, de Silva, Erickson, Ghrist.
Discrete Computational Geometry 2010.

Proved: for a point cloud $C \subset \mathbb{R}^2$, the projection to the shadow: $\mathrm{VR} \to S(\mathrm{VR}) \subset \mathbb{R}^2$ respects π_1.

For a cloud of n points, can we find all persistent β_1 of the shadows $S(\mathrm{VR}(\varepsilon))$ in $O(n \log n)$ time?

Future work and problems

- Topology Analyzer Java applet on graph reconstruction at http://kurlin.org
- reconstructing *topological types* of graphs
- detecting homotopy types of noisy graphs by using plane shadows of Rips complexes
- statistics of persistent clusters or Betti numbers for randomly generated clouds
- automatic choice of a *density threshold* to find persistent clusters with long life spans