PERSISTENCE-BASED SUMMARIES FOR METRIC GRAPHS

Radmila Sazdanović

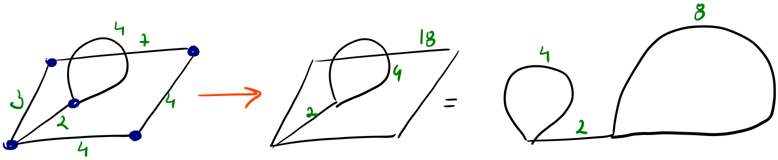
joint work with Ellen Gasparovic, Maria Gommel, Emilie Purvine, Bei Wang, Yusu Wang, and Lori Ziegelmeier

NC State University

Topological Methods in Brain Network Analysis
BIRS 11 May 2017

Metric graphs

- Input: Weighted graph $G = (V, E, L)$ with a weight function $L : E \rightarrow R_{\geq 0}$
- Output: metric graph (G, d_G)
  - homeomorphic to a 1-dim stratified space
  - Every point of every edge is a point in this space-it is infinite!
  - Embedding does not matter.



Motivation

Why metric graphs

- Data often comes from a hidden space that is graph-like: cosmic networks, road maps,...
- Graphs are often simplest meaningful way to represent non-linear structure of the data: internet, social networks, structural and functional connectome, etc.

(Metric) Graph comparison

- Graph isomorphism: conjecturally NP complete (for metric graphs there is a problem of noise and small deformations)
- Distances: discrete or not-computable (Gromov-Hausdorff compares metric graphs as metric spaces and it is NP-hard even to approximate within a constant factor!)

TDA TO THE RESCUE

QUALITATIVE SUMMARIES OF METRIC GRAPHS

- Relate properties of a metric graph G and the homology of an associated complex.
- Identify which topological properties of a graph are contained in the persistence diagram
- Give complete characterization of the 1-dimensional persistence diagrams for metric graphs with the Čech complex construction in terms of graph properties.
- Goal: provide powerful insights to understanding underlying data.

TDA to the rescue

Persistance-based distances

- Construct a continuous distance on the space of metric graphs, stable under metric perturbations (noise and small deformations) using persistence of an associated simplicial complex
- Compare discriminative powers of distances based on the bottleneck distance between persistent diagrams obtained from different constructions:
  - Čech and
  - persistence distortion (PD) distance [Dey, Shi, Wang]
- Polynomial-time computable

PRIOR RESULTS

CYCLE GRAPHS

Nerve complex for a finite collection of arcs on a circle has the homotopy type of a

- point
- odd-dimensional sphere
- wedge sum of spheres with the same even dimension

[Adamaszek, Adams, Florian, Peterson, Previte-Johnson]

COROLLARY

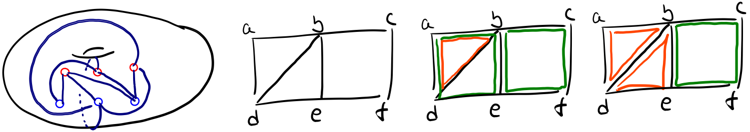
The 1-dimensional persistence diagram of a circle split into arcs of total length equal to ℓ consist of at most one bar

- $[0, \ell/4)$ *for the Čech complex, and*
- $[0, \ell/6)$ *for the Vietoris-Rips complex.*

TOPOLOGICAL GRAPH THEORY

THE GENUS OF A GRAPH

- the minimal integer n such that the graph can be drawn without crossing itself on an oriented surface of genus n.
- β_1: number of cycles in a basis for the first homology



SHORTEST SYSTEM OF LOOPS

- cycles that are shortest non-trivial paths from a vertex to itself
- shortest representatives of homology classes in lexicographical order

INTRINSIC ČECH FILTRATION

Let (G, d_G) be a metric graph.

- For any point $x \in G$ define $B(x, \epsilon) := \{y \in |G| : d_G(x, y) \leq \epsilon\}$
- Covering $U_\epsilon := \{B(x, \epsilon) : x \in |G|\}$ and let
- C_ϵ denote the nerve of U_ϵ
- *Intrinsic Čech filtration* is the set of inclusions

$$\{\mu_\epsilon^c : C_\epsilon \hookrightarrow C_{\epsilon'}\}_{\forall 0 \leq \epsilon \leq \epsilon'}.$$

- *Intrinsic Čech persistence diagram $Dg_* IC_G$*, is obtained from the induced persistence module

$$\{\mu_\epsilon^h : H_*(C_\epsilon) \to H_*(C_{\epsilon'})\}_{\forall 0 \leq \epsilon \leq \epsilon'}$$

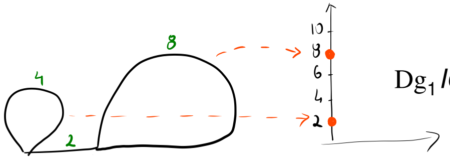
Characterization of the 1-dim persistence diagrams

Theorem (GGPSWWZ '17)

Let G be a metric graph of genus g with

- *a shortest cycle basis $\beta = \{\gamma_1, \ldots, \gamma_g\}$*
- *with cycles γ_i of length ℓ_i for $1 \le i \le g$ such that $\ell_1 \le \ldots \le \ell_g$*

Then the 1-dimensional intrinsic Čech persistence diagram of G, $\mathrm{Dg}_1 IC(G)$, consists of the following collection of points on the y-axis:



$$\mathrm{Dg}_1 IC(G) = \left\{ \left[0, \frac{\ell_i}{4}\right) : 1 \le i \le g \right\}$$

Proof of the main theorem

- For small $\delta > 0$, C_δ has the same homotopy type as G.
- C_δ^0 inherits metric from G
- All γ_i's are born at δ (consider it to be 0)
- Each γ_i must die at $\frac{\ell_i}{4}$ or earlier since $\forall x, y, z \in \gamma_i$, $B\left(x, \frac{\ell_i}{4}\right) \bigcap B\left(y, \frac{\ell_i}{4}\right) \bigcap B\left(z, \frac{\ell_i}{4}\right) \neq \emptyset$.

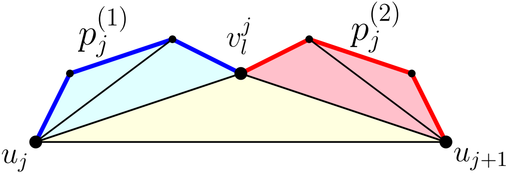
Need to show

A No other cycles are created in C_ϵ, $\epsilon > \delta$ due to interference from other cycles: β spans 1-dim persistence

B For $i = 1, \ldots, g$, $[\gamma_i]$ does not die before $\epsilon = \dfrac{\ell_i}{4}$: β is linearly independent

Proof of Part A

The map $\mu^h_\epsilon : H^{(1)}_\delta \to H^{(1)}_\epsilon$ is surjective, has a right inverse up to homotopy.



In other words, there exists a combinatorially defined map $\rho : C^{(1)}_\delta \to C^{(1)}_\epsilon$ such that for every $[\eta] \in H^{(1)}_\epsilon$

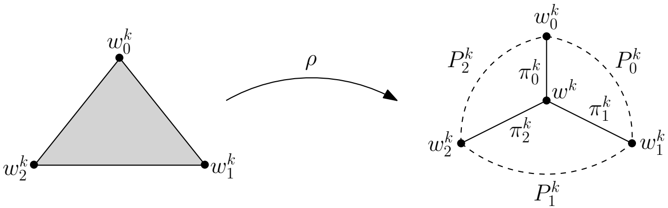
$$\mu^h_\epsilon([\rho(\eta)]) = [(\mu^c_\epsilon \circ \rho)(\eta)] = [\eta] \in H^{(1)}_\epsilon$$

Proof of Part B

For any $i = 1, \ldots, g$, the set

$$\{[\mu_\epsilon^c(\gamma_i)], [\mu_\epsilon^c(\gamma_{i+1})], \ldots, [\mu_\epsilon^c(\gamma_g)]\}$$

is a basis for $H_\epsilon^{(1)}$ where $\frac{\ell_{i-1}}{4} \leq \epsilon < \frac{\ell_i}{4}$ and $\ell_0 = 0$.



$\rho(t_k) = \sum_{n=0}^{n=2} \pi_n^k + P_n^k - \pi_{(n+1 \mod 3)}^k$ is the sum of three cycles, each of length at most $\epsilon + \epsilon + 2\epsilon = 4\epsilon < \ell_i$.

SUMMARY

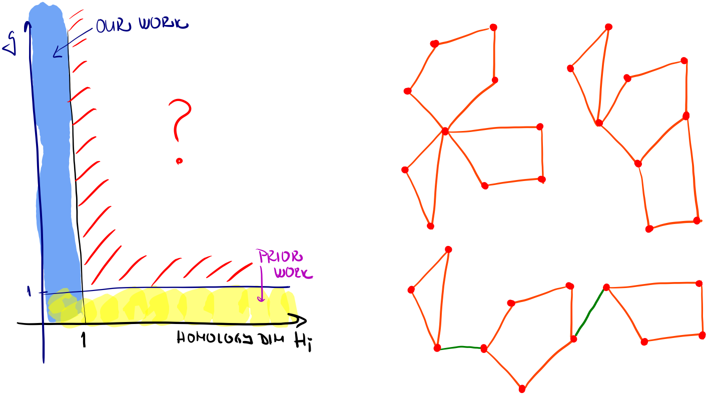


FIGURE: What is missing? Graphs that we can not distinguish...

Intrinsic Čech distance

Intrinsic Čech distance $d_{IC}(G_1, G_2)$

Let d_B denore the bottleneck distance between the two intrinsic Čech persistence diagrams in dimension 1. Then

$$d_{IC}(G_1, G_2) := d_B(Dg_1 IC_{G_1}, Dg_1 IC_{G_2}),$$

[Chazal Cohen, Steiner, Guibas, Memoli, Oudout 2009]

Modified bottleneck distance $\delta(D_1, D_2)$

Given persistence diagrams D_1 and D_2 let the distance between $(x, y) \in D_1$ and $(x', y') \in D_2$ be $|x - x'| + |y - y'|$.

- If a point (x, y) is matched to a point on a diagonal, it contributes $y - x$ to $\delta(D_1, D_2)$
- δ is within a factor of 2 of the standard bottleneck d_B

Intrinsic Čech distance

Theorem (GGPSWWZ '17)

Let $D_1 := \{(0, a_i)\}_{i=1}^{s}$ and $D_2 := \{(0, b_j))\}_{j=1}^{t}$ be persistence diagrams with $a_1 \leq \cdots \leq a_s$, $b_1 \leq \cdots \leq b_t$ and, WLOG, $s \leq t$. Let $A = \{a'_1, a'_2, \ldots a'_t\}$ and $B = \{b_1, \ldots b_t\}$, where A consists of $t - s$ zeroes at the beginning, followed by the sequence $a_1, a_2, \ldots, a_s$.

Then $\delta(D_1, D_2) = d_B(D_1, D_2) = \max_{i=1}^{t} |a'_i - b_i|$.

Corollary (GGPSWWZ '17)

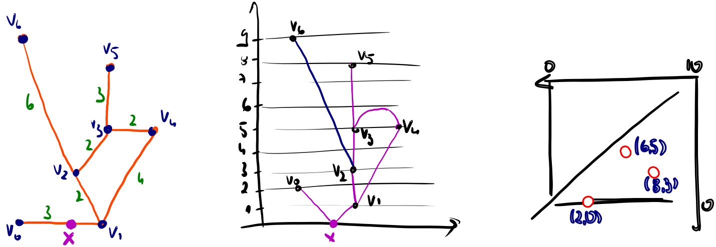
The distance between 1-dim intrinsic Čech persistence diagrams $Dg_1 IC_{G_1} = \left\{ \left(0, \frac{\ell_i}{4}\right) \right\}_{i=1}^{s}$ and $Dg_1 IC_{G_2} = \left\{ \left(0, \frac{m_j}{4}\right) \right\}_{j=1}^{t}$ associated to metric graphs G_1 and G_2 is

$$d_{IC}(G_1, G_2) = d_B(Dg_1 IC_{G_1}, Dg_1 IC_{G_2}) = \max_{i=1}^{t} \frac{|\ell'_i - m_i|}{4}.$$

Persistence with respect to intrinsic distance d_G

Persistence diagram $Dg_0(d_G, x)$

Given a metric graph G, fix a base point $x \in |G|$ and consider connected components of the superlevel sets $G \setminus B(x, \epsilon)$ with respect to the geodesic intrinsic distance.



PD distance d_{PD}

PD distance $d_{PD}(G_1, G_2)$

The persistence distortion distance between G_1 and G_2 is

$$d_{PD}(G_1, G_2) = \max \left\{ \max_{x \in G_1} \min_{y \in G_2} d_B\left(Dg_0(d_{G_1}, x), Dg_0(d_{G_2}, y)\right), \max_{y \in G_2} \min_{x \in G_1} d_B\left(Dg_0(d_{G_2}, y), Dg_0(d_{G_2}, y)\right) \right\}$$

$d_{PD}(G_1, G_2)$ is the Hausdorff distance between collections of 0-dimensional persistence diagrams with respect to all possible pairs of baspoints, as subspaces of the space of persistence diagrams equipped with the bottleneck distance.

Proposition (Dey, Shi, Wang)

$d_{PD}(G_1, G_2) \leq 6 d_{GH}(G_1, G_2)$

d_{PD} vs. d_{IC}?

Conjecture (GGPSWWZ '17)

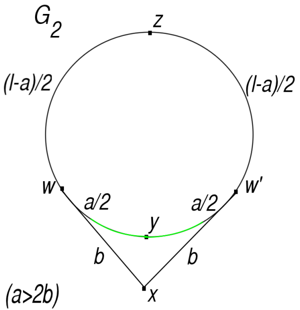
For any two metric graphs G_1 and G_2, $d_{IC}(G_1, G_2) \leq \frac{1}{2} d_{PD}(G_1, G_2)$.

Theorem (GGPSWWZ '17)

In the case of

- G_1, G_2 are metric trees; or
- G_1 is a single cycle of length ℓ and G_2 is the graph with two cycles shown in Figure with a is slightly larger than $2b$.

we have equality.



PROPOSITION

Let G_1 be a graph containing a single cycle of length ℓ and G_2 a graph containing a single cycle of length m with $\ell \geq m$. Then:

- $d_{IC}(G_1, G_2) = \dfrac{\ell - m}{4}$
- $d_{PD}(G_1, G_2) = \dfrac{\ell - m}{2}$, *and therefore* $d_{IC}(G_1, G_2) = \frac{1}{2} d_{PD}(G_1, G_2)$.

THEOREM

Let G_1 be a bouquet of circles and G_2 any metric graph. Then

$$d_{IC}(G_1, G_2) \leq \frac{1}{2} d_{PD}(G_1, G_2).$$

To be continued...

- What does the higher persistence diagram know about the underlying topology of a graph?
- Can we use these persistence summaries to distinguish common types of graph motifs?
- Characterize PD persistence?
- The geodesic persistence diagram of f_x is stable w.r.t. small perturbation of its geometric realization under the Gromov-Hausdorff distance.
- Is it possible to develop a persistence-distortion for the combinatorial graphs that would be stable with respect to some appropriate notion of perturbation?

Thank you!