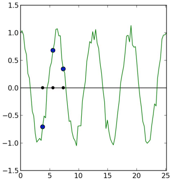
Quantum Computation in a Topological Data Analysis Pipeline

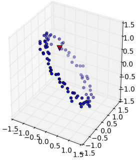
Jesse Berwald, Joel Gottlieb, **Elizabeth Munch**

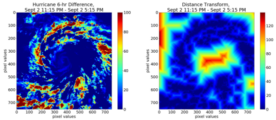
Michigan State University
::
Dept. of Computational Mathematics, Science, and Engineering
Dept. of Mathematics

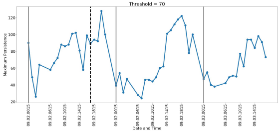
Qubits, Knoxville, TN - Sep 24, 2018

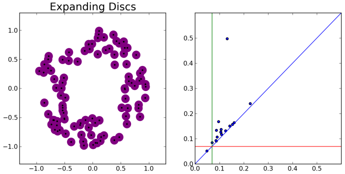
Topology in data

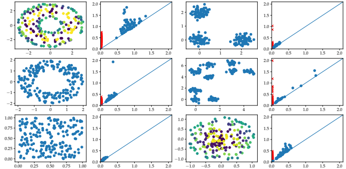


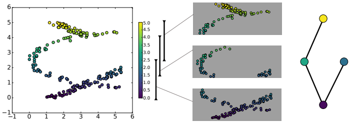


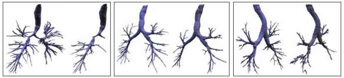












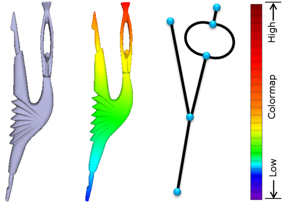
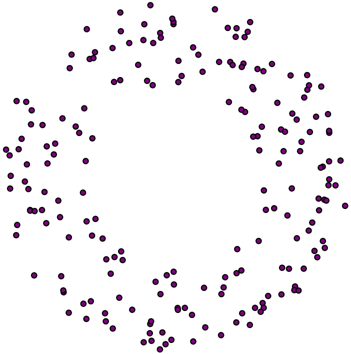
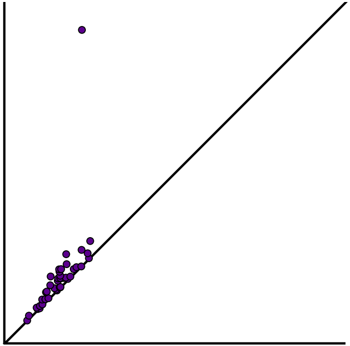


Image: Szymczak et al., Ma et al.

Large Data Sets



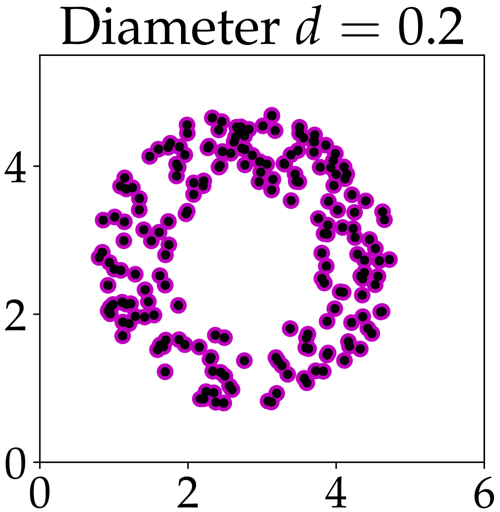


Main goal of Topological Data Analysis (TDA)

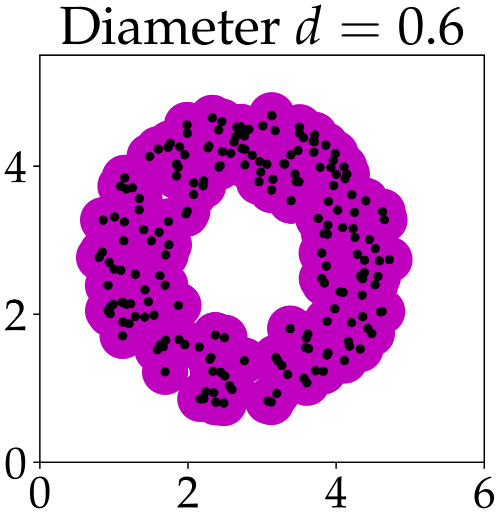
Find and quantify structure in noisy, complex data.

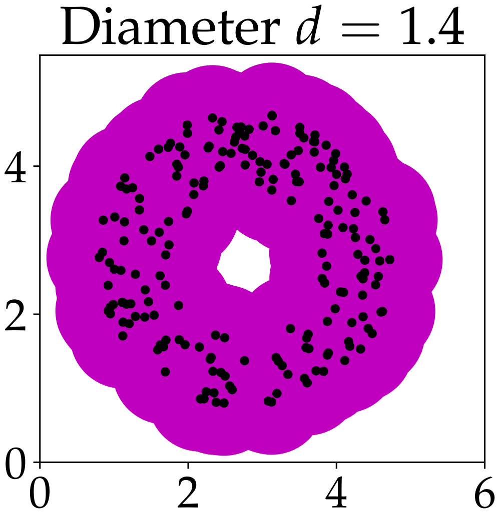
Section 1

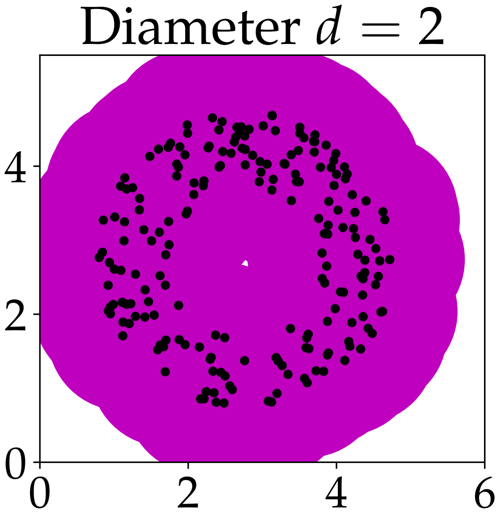
Persistent Homology and Wasserstein Distance


Diameter $d = 0.2$
4
2
0
0
2
4
6

Finding shape in a point cloud

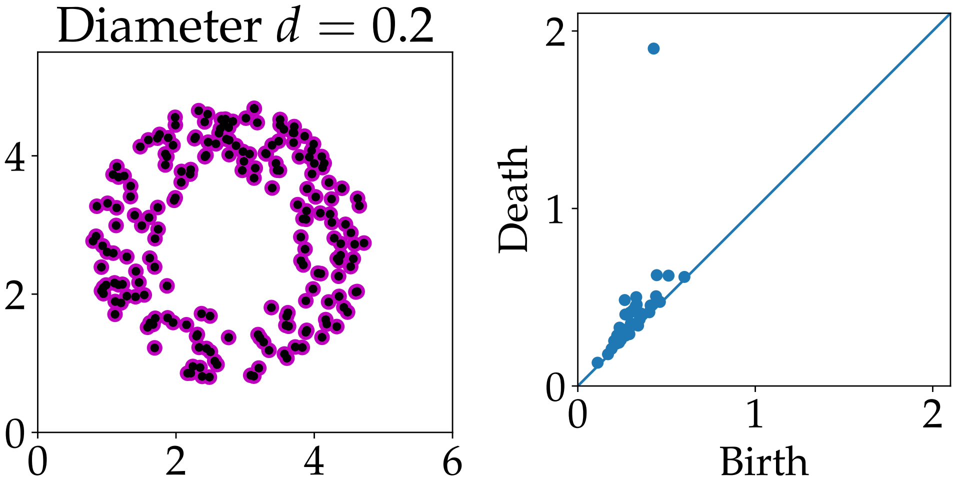



Diameter $d = 1.4$
4
2
0
0
2
4
6


Diameter $d = 2$
0
2
4
0
2
4
6

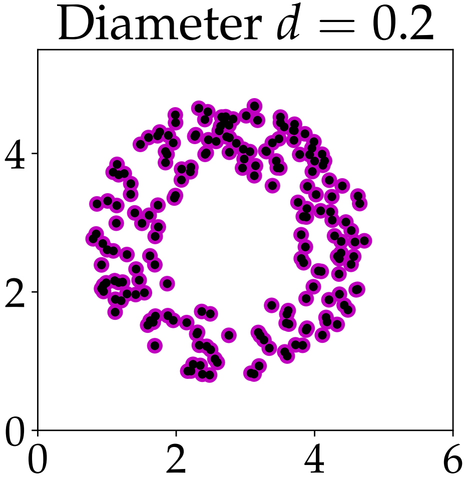
Understanding a persistence diagram

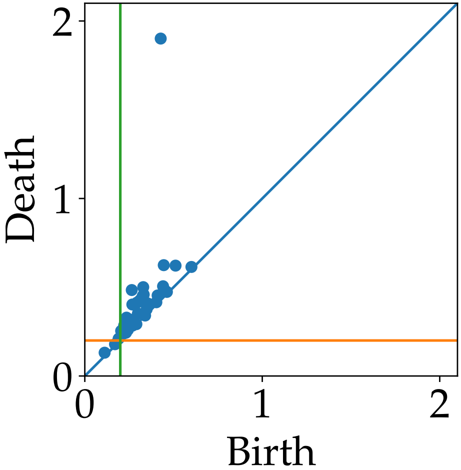
Annulus example



Understanding a persistence diagram

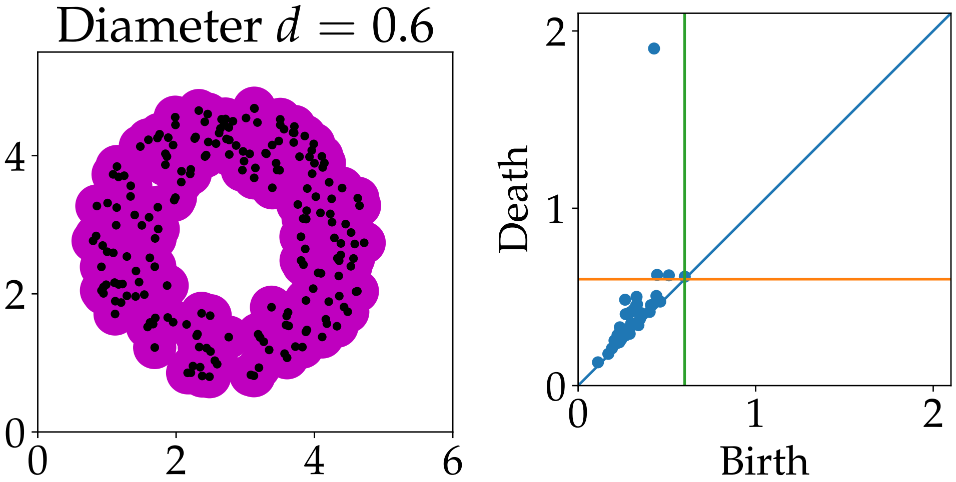
Annulus example





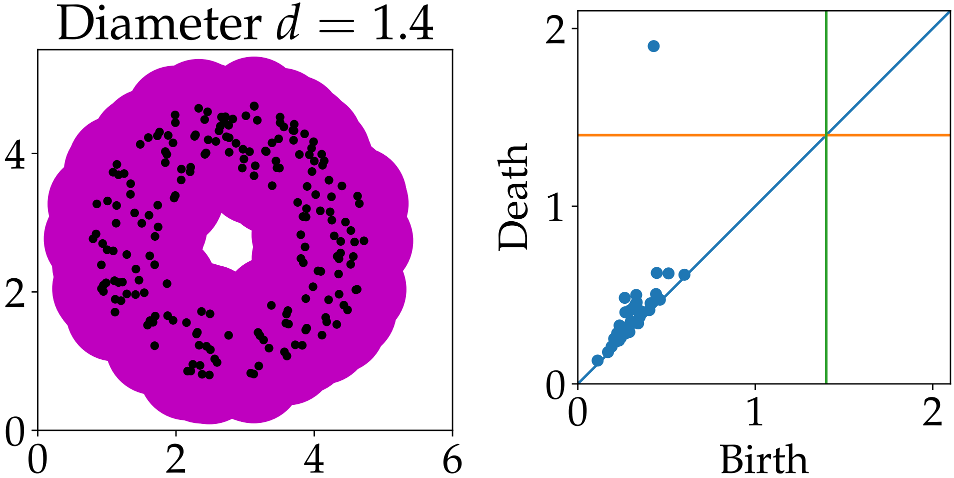
Understanding a persistence diagram

Annulus example



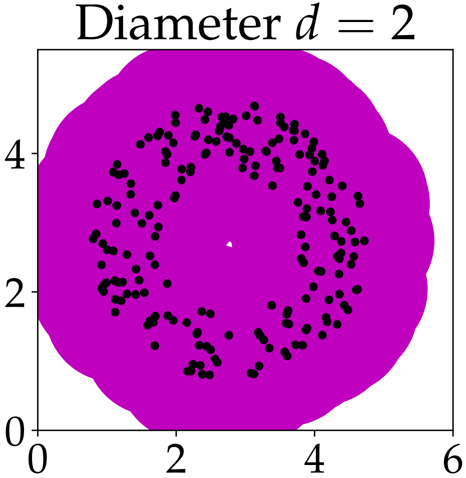
Understanding a persistence diagram

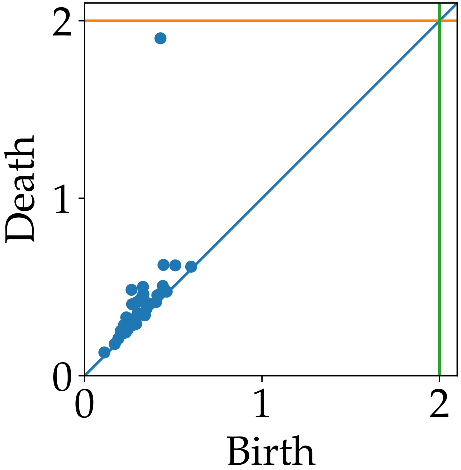
Annulus example



Understanding a persistence diagram

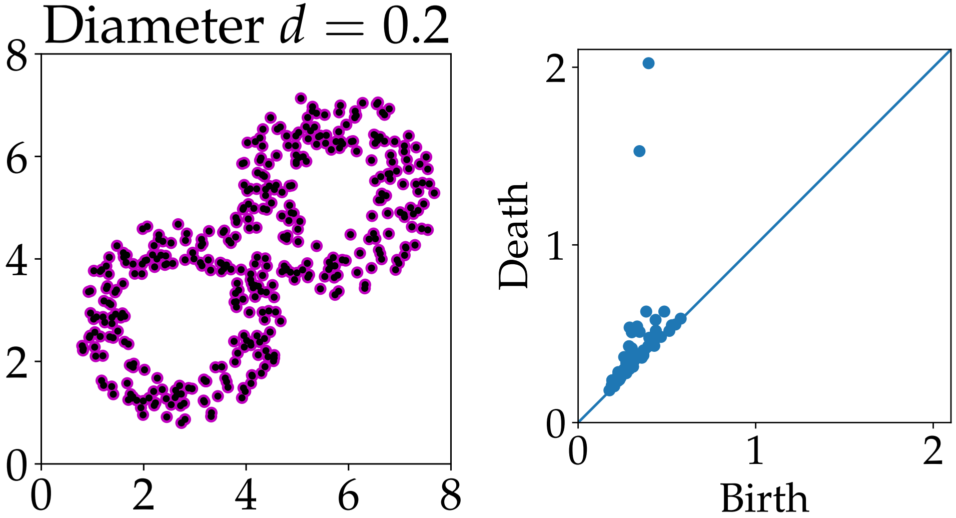
Annulus example





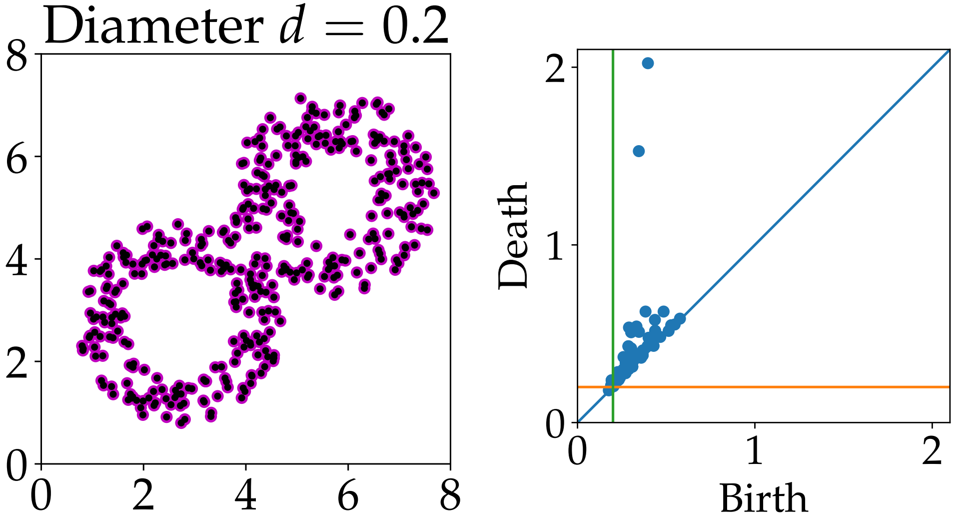
Understanding a persistence diagram

Double annulus example

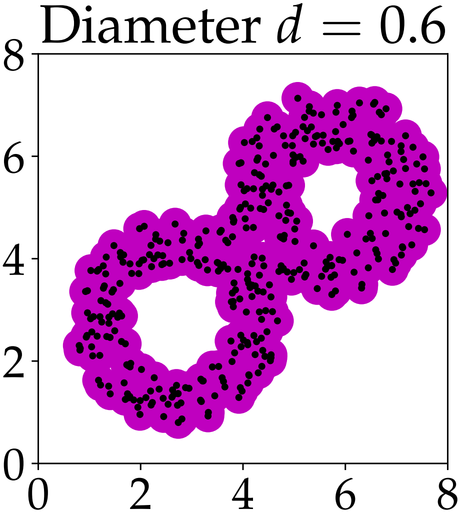


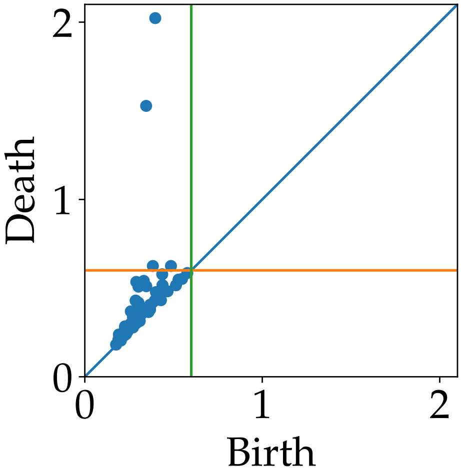
Understanding a persistence diagram

Double annulus example



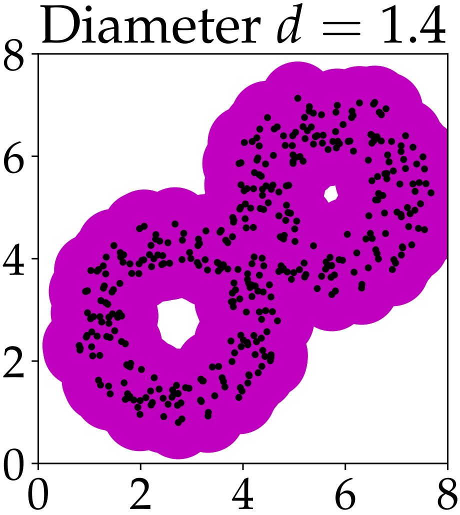
Understanding a persistence diagram
Double annulus example

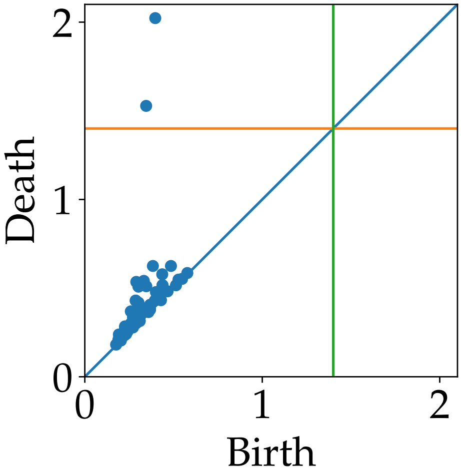


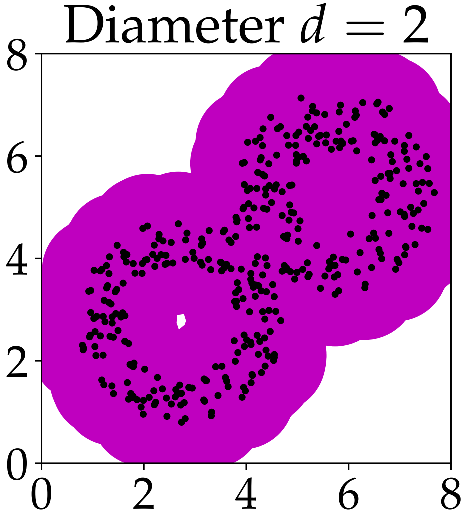


Understanding a persistence diagram

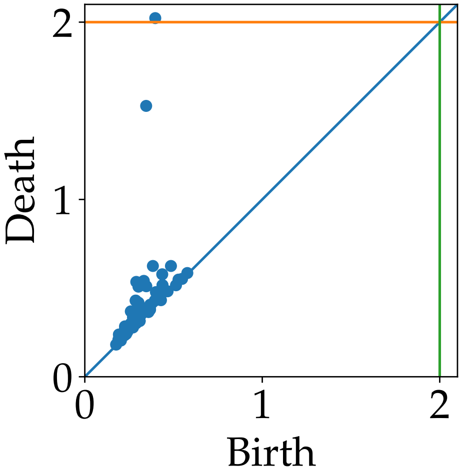
Double annulus example





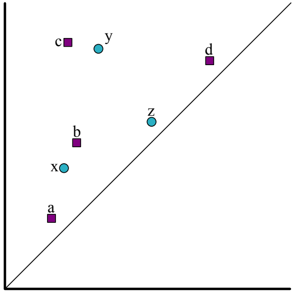


Diameter $d = 2$
0
2
4
6
8



Death
Birth
0
1
2

Wasserstein Distance on D_p

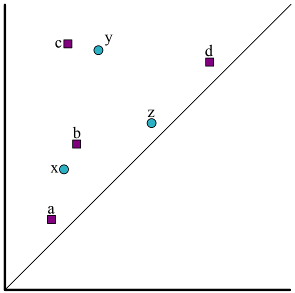


Definition

A distance on a set M is a function $d : M \times M \to \mathbb{R}_{\geq 0}$ such that

- $d(x, y) \geq 0$ and $d(x, y) = 0$ iff $x = y$
- $d(x, y) = d(y, x)$
- $d(x, y) + d(y, z) \geq d(x, z)$

Wasserstein Distance on D_p



Definition

A distance on a set M is a function $d : M \times M \to \mathbb{R}_{\geq 0}$ such that

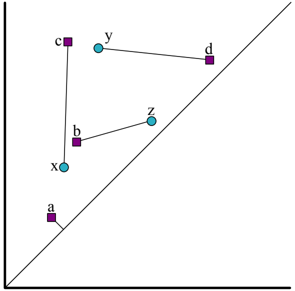
- $d(x,y) \geq 0$ and $d(x,y) = 0$ iff $x = y$
- $d(x,y) = d(y,x)$
- $d(x,y) + d(y,z) \geq d(x,z)$

Wasserstein distance for diagrams

Given diagrams X and Y, the distance between them is

$$d_p(X,Y) = \inf_{\varphi: X \to Y} \left(\sum_{x \in X} \|x - \varphi(x)\|^p \right)^{1/p}$$

Wasserstein Distance on D_p



Definition

A distance on a set M is a function $d : M \times M \to \mathbb{R}_{\geq 0}$ such that

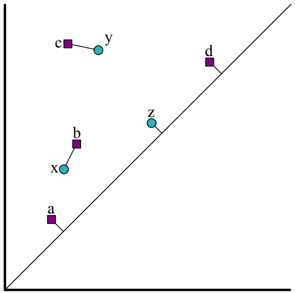
- $d(x,y) \geq 0$ and
  $d(x,y) = 0$ iff $x = y$
- $d(x,y) = d(y,x)$
- $d(x,y) + d(y,z) \geq d(x,z)$

Wasserstein distance for diagrams

Given diagrams X and Y, the distance between them is

$$d_p(X,Y) = \inf_{\varphi : X \to Y} \left(\sum_{x \in X} \|x - \varphi(x)\|^p \right)^{1/p}$$

Wasserstein Distance on D_p



Definition

A distance on a set M is a function $d : M \times M \to \mathbb{R}_{\geq 0}$ such that

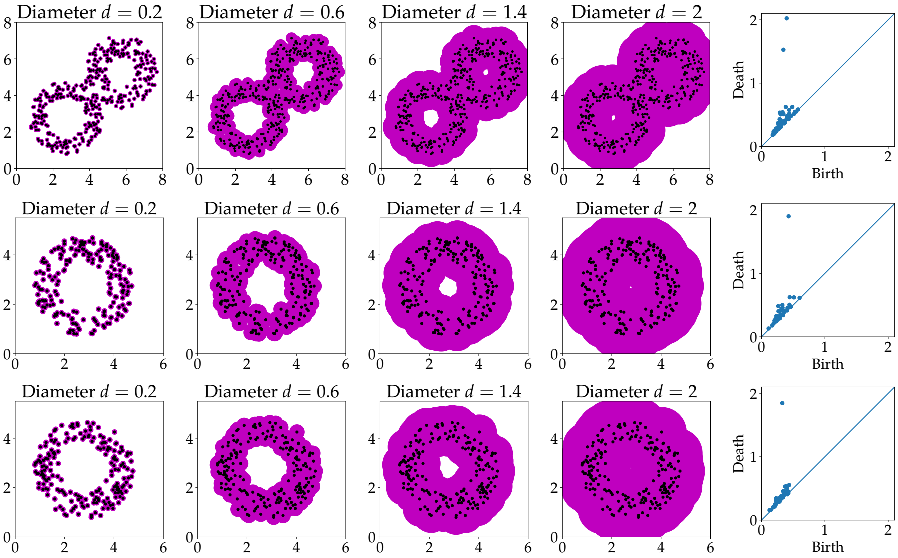
- $d(x, y) \geq 0$ and $d(x, y) = 0$ iff $x = y$
- $d(x, y) = d(y, x)$
- $d(x, y) + d(y, z) \geq d(x, z)$

Wasserstein distance for diagrams

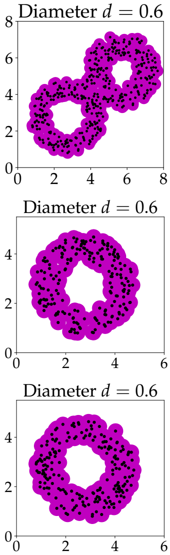
Given diagrams X and Y, the distance between them is

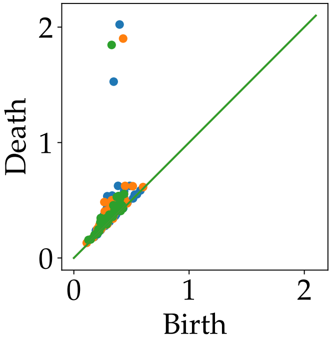
$$d_p(X, Y) = \inf_{\varphi : X \to Y} \left(\sum_{x \in X} \| x - \varphi(x) \|^p \right)^{1/p}$$

Example



Diameter $d = 0.2$
Diameter $d = 0.6$
Diameter $d = 1.4$
Diameter $d = 2$
Death
Birth
Diameter $d = 0.2$
Diameter $d = 0.6$
Diameter $d = 1.4$
Diameter $d = 2$
Death
Birth
Diameter $d = 0.2$
Diameter $d = 0.6$
Diameter $d = 1.4$
Diameter $d = 2$
Death
Birth


Diameter $d = 0.6$
Diameter $d = 0.6$
Diameter $d = 0.6$


Death
Birth

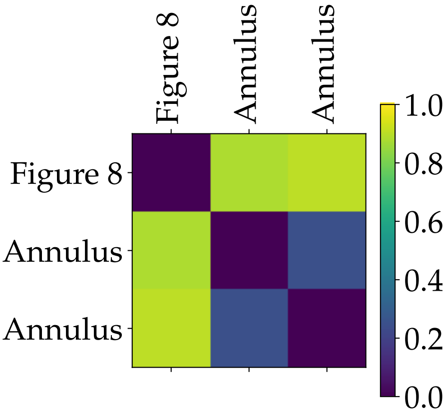
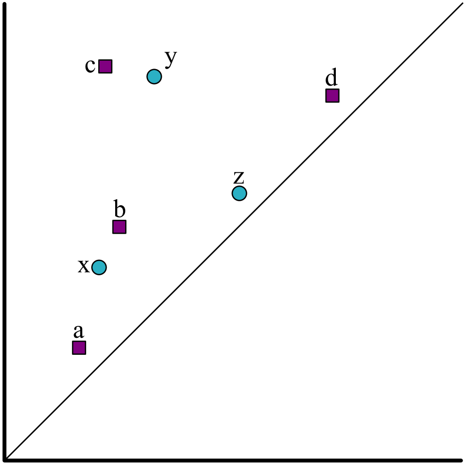
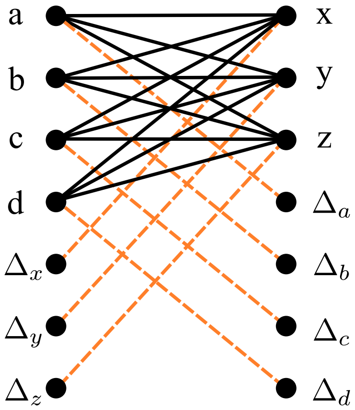


Figure 8
Annulus
Annulus
Figure 8
Annulus
Annulus
1.0
0.8
0.6
0.4
0.2
0.0

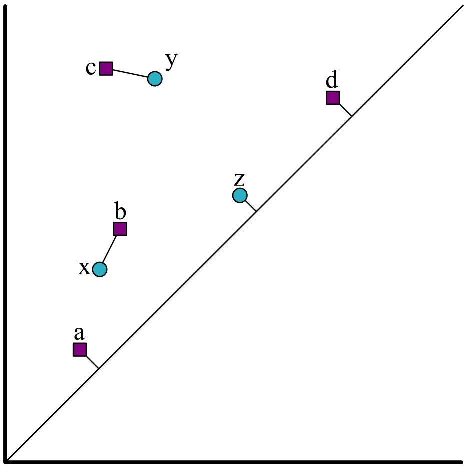
Computation of Wasserstein Distance

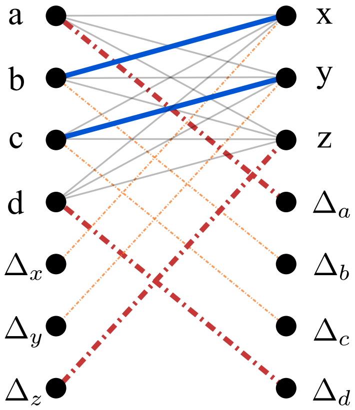




$$\omega(u, v) = \|u - v\|^p$$

Computation of Wasserstein Distance



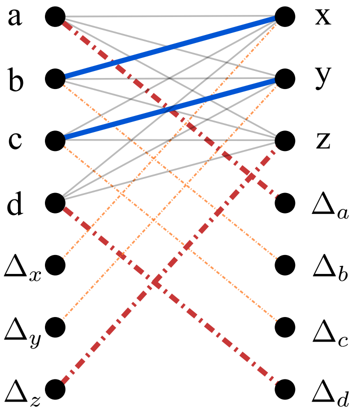


$$\omega(u, v) = \|u - v\|^p$$

Matchings

Definition

A subset of edges $\mathbf{x} \subseteq E$ is a **matching** if every vertex is adjacent to **at most one** edge in $\mathbf{x}$.



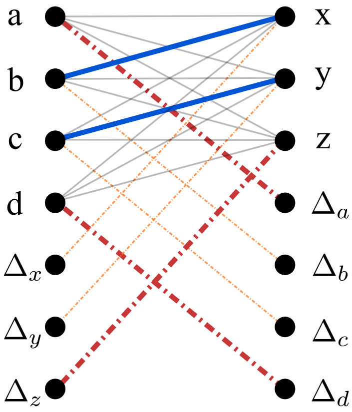
Matchings

Definition

A subset of edges $\mathbf{x} \subseteq E$ is a **matching** if every vertex is adjacent to **at most one** edge in $\mathbf{x}$.

Definition

A matching $\mathbf{x} \subseteq E$ is a **maximal matching** if it **is not contained** in a larger matching $\mathbf{y}$.



Definition

A subset of edges $\mathbf{x} \subseteq E$ is a **matching** if every vertex is adjacent to **at most one** edge in $\mathbf{x}$.

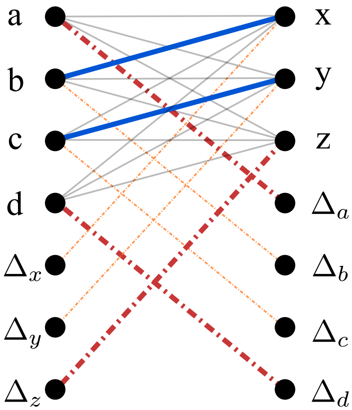
Definition

A matching $\mathbf{x} \subseteq E$ is a **maximal matching** if it **is not contained** in a larger matching $\mathbf{y}$.

Definition

The **cost** of a matching $\mathbf{x} \subseteq E$ is the sum of the weights,

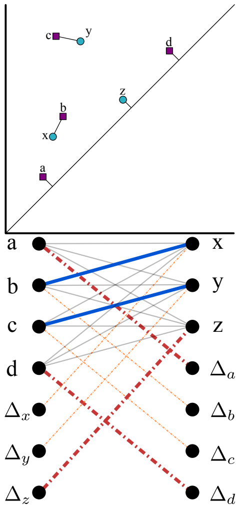
$$C_p(\mathbf{x}) = \sum_{e \in \mathbf{x}} \omega(e) = \sum_{(u,v) \in \mathbf{x}} \|u - v\|^p.$$



Equivalence

Theorem (Edelsbrunner et al 2010)

$$d_p(X, Y) = \varepsilon$$
iff
$$\varepsilon^p = \min\{C(\mathbf{x}) \mid \mathbf{x} \textit{ is a maximal matching}\}.$$



Equivalence

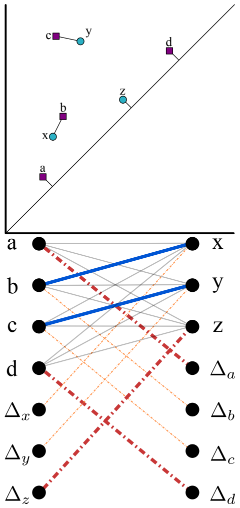
Theorem (Edelsbrunner et al 2010)

$$d_p(X, Y) = \varepsilon$$
iff
$$\varepsilon^p = \min\{C(\mathbf{x}) \mid \mathbf{x} \textit{ is a maximal matching}\}.$$

Definition

A min-cost maximal matching (MCMM) is a maximal matching $\mathbf{y}$ for which

$$C(\mathbf{y}) = \min\{C(\mathbf{x}) \mid \mathbf{x} \text{ is a maximal matching}\}$$



Section 2

A Qubo for Wasserstein distance

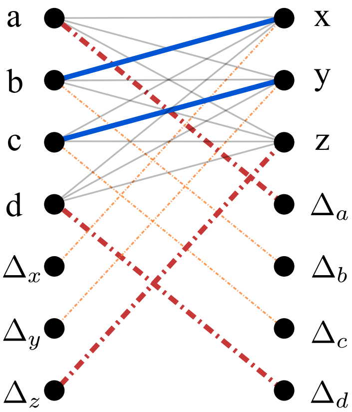
The variables

$$\mathbf{x} \in (\mathbb{Z}_2)^{nm+n+m}$$

Binary variables: $\mathbf{x} = \{x_{u,v} \mid (u,v) \in E\}$

$\Updownarrow$ $\Updownarrow$

Sets of edges $\mathbf{x} \subseteq E$



$$\omega(u,v) = \|u - v\|^p$$

The variables

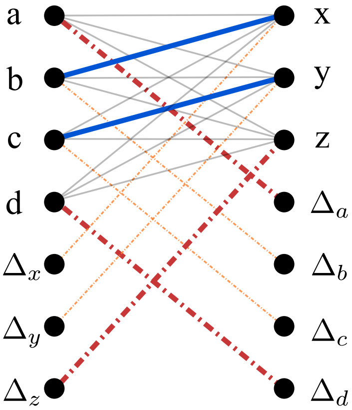
$$\mathbf{x} \in (\mathbb{Z}_2)^{nm+n+m}$$

Binary variables: $\mathbf{x} = \{x_{u,v} \mid (u,v) \in E\}$

$\Updownarrow$ $\Updownarrow$

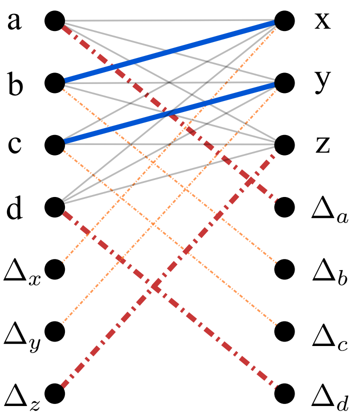
Sets of edges $\mathbf{x} \subseteq E$

Warning: Not just maximal matchings



$$\omega(u,v) = \|u - v\|^p$$

The QUBO



$$\omega(u,v) = \|u - v\|^p$$

$$F_c(\mathbf{x}) = \sum_{(u,v)\in E} \omega(u,v) x_{u,v}$$

$$F_U(\mathbf{x}) = B \sum_{u \in X \subset U} \left(1 - \sum_{\substack{v \in V \\ (u,v)\in E}} x_{u,v} \right)^2$$

$$F_V(\mathbf{x}) = B \sum_{v \in Y \subset V} \left(1 - \sum_{\substack{u \in U \\ (u,v)\in E}} x_{u,v} \right)^2$$

$$H = F_c + F_U + F_V.$$

Correctness

Theorem (Berwald, Gottlieb, EM, 2018)

Assume $B > B^* := \max_{(u,v)\in E(\widetilde{G})} \omega(u,v)$.

Then $\mathbf{x}$ *is a solution which minimizes* H
if and only if
$\mathbf{x} \subseteq E$ *is a MCMM of* $\widetilde{G}$.

Correctness

Theorem (Berwald, Gottlieb, EM, 2018)

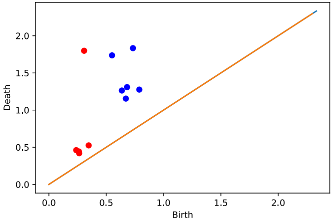
Assume $B > B^* := \max_{(u,v)\in E(\widetilde{G})} \omega(u,v)$.

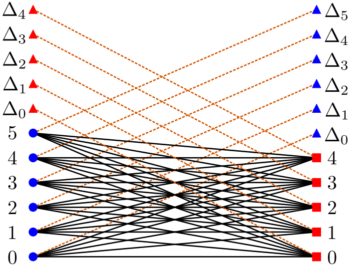
Then $\mathbf{x}$ *is a solution which minimizes* H
if and only if
$\mathbf{x} \subseteq E$ *is a MCMM of* $\widetilde{G}$.

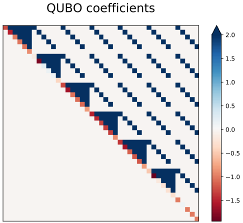
Proof sketch

- For every subset $\mathbf{z}$, there is a matching $\mathbf{y}$ with $H(\mathbf{y}) < H(\mathbf{z})$.
- For every non-maximal matching $\mathbf{y}$ there is a maximal matching $\mathbf{x}$ with $H(\mathbf{x}) < H(\mathbf{y})$.
- Maximal matchings have $H(\mathbf{x}) = C(\mathbf{x})$

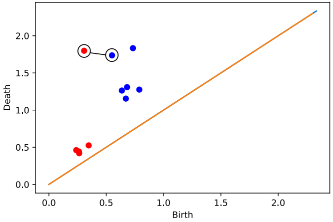
Experiment


Death
Birth
0.0
0.5
1.0
1.5
2.0

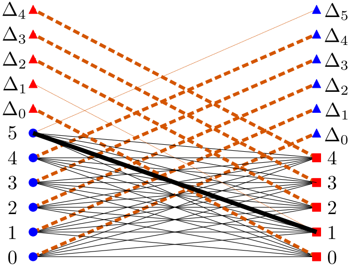

Δ_4
Δ_3
Δ_2
Δ_1
Δ_0
5
4
3
2
1
0
Δ_5
Δ_4
Δ_3
Δ_2
Δ_1
Δ_0
4
3
2
1
0


QUBO coefficients
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

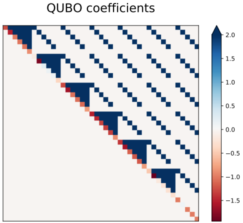
Experiment



Death
Birth
0.0
0.5
1.0
1.5
2.0

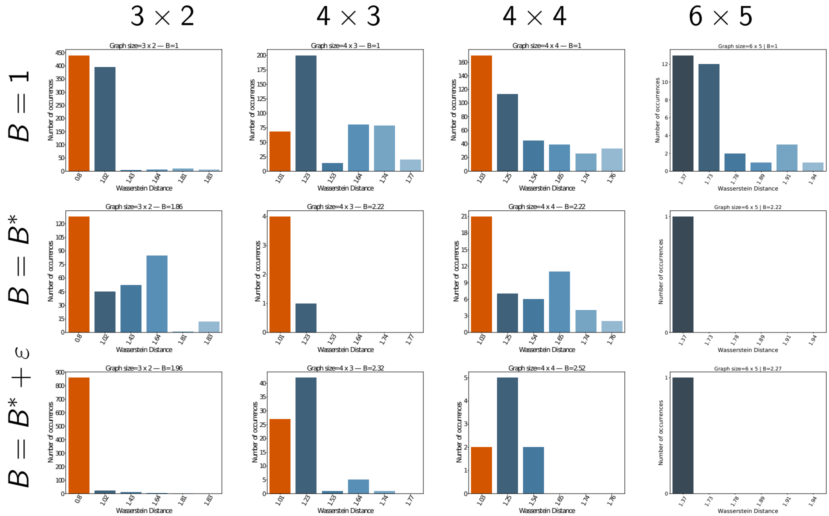


Δ_4
Δ_3
Δ_2
Δ_1
Δ_0
5
4
3
2
1
0
Δ_5
Δ_4
Δ_3
Δ_2
Δ_1
Δ_0
4
3
2
1
0



QUBO coefficients
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5

Size:


3×2
4×3
4×4
6×5
$B = 1$
$B = B^*$
$B = B^* + \varepsilon$
Graph size=3 x 2 — B=1
Graph size=4 x 3 — B=1
Graph size=4 x 4 — B=1
Graph size=6 x 5 | B=1
Graph size=3 x 2 — B=1.86
Graph size=4 x 3 — B=2.22
Graph size=4 x 4 — B=2.22
Graph size=6 x 5 | B=2.22
Graph size=3 x 2 — B=1.96
Graph size=4 x 3 — B=2.32
Graph size=4 x 4 — B=2.52
Graph size=6 x 5 | B=2.27
Number of occurrences
Wasserstein Distance

Future work

- Exactly what collection of B gives good solutions?

Future work

- Exactly what collection of B gives good solutions?
- Bottleneck distance

Future work

- Exactly what collection of B gives good solutions?
- Bottleneck distance
- Why do the tests fail for larger sized problems?
  - Long chains of physical qubits?

Future work

- Exactly what collection of B gives good solutions?
- Bottleneck distance
- Why do the tests fail for larger sized problems?
  - Long chains of physical qubits?
- Other places in the persistent homology pipeline that can be swapped out for QC
  - Flag/Clique complexes
  - Computation of full persistent homology
    (Betti numbers: *Lloyd et al 2016, Siopsis 2018, Dridi Alghassi 2015*)
  - Multiparameter persistence

Thank you!

Collaborators:

Jesse Berwald

Joel Gottlieb

Teaspoon code available:
github.com/lizliz/teaspoon

Relevant papers

- J. Berwald, J. Gottlieb, EM. *Computing Wasserstein Distance for Persistence Diagrams on a Quantum Computer*. arXiv:1809.06433, 2018
- EM. *A User's Guide to Topological Data Analysis*. Journal of Learning Analytics, 2017

elizabethmunch.com
muncheli@egr.msu.edu

CMSE
Dept. of Computational Mathematics, Science, and Engineering