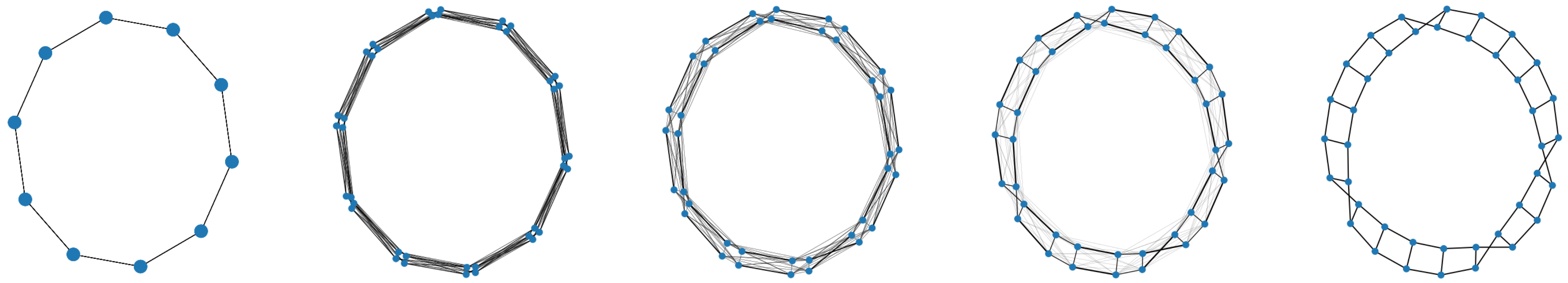
Gromov-Wasserstein Learning in a Riemannian Framework



Samir Chowdhury

UCD MADDD Seminar

December 08, 2020

Collaborators

- Nicolas Garcia Trillos (UW-Madison)
- Caleb Geniesse (Stanford)
- David Miller (Utah)
- **Tom Needham** (FSU)
- Manish Saggar (Stanford)
- Hongteng Xu (Duke/Infinia ML)
- Mengsen Zhang (Stanford)

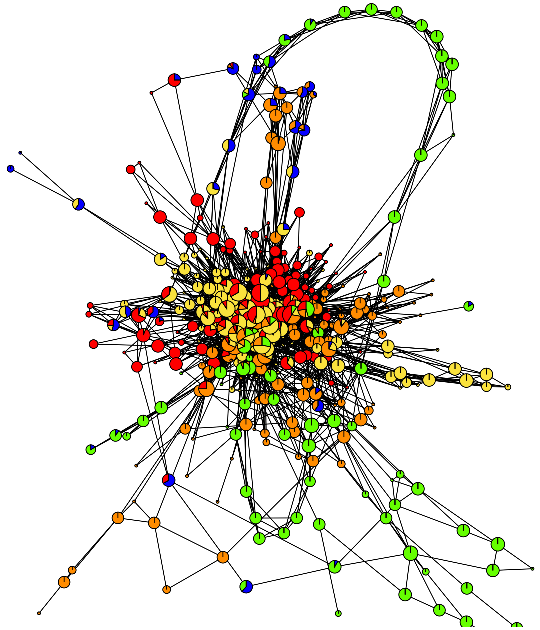
Papers:

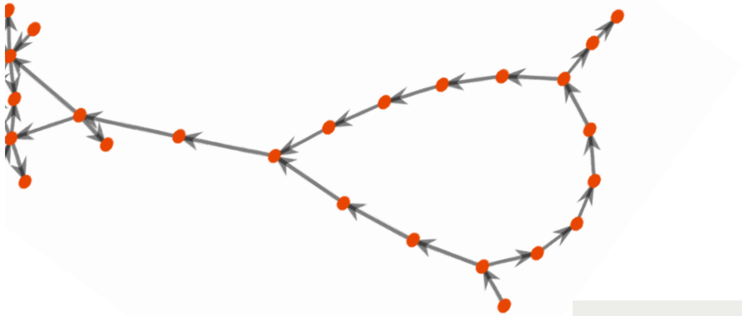
- *The Gromov-Wasserstein distance between networks and stable network invariants* (C., Mémoli 2019)
- *Gromov-Wasserstein Averaging in a Riemannian Framework* (C., Needham 2020)
- *Generalized Spectral Clustering via Gromov-Wasserstein Learning* (C., Needham 2020)
- *(Upcoming) Exploring the landscape of brain dynamics using topological data analysis* (Geniesse*,C.*, Saggar)
- *(Upcoming) The topology of time: transition networks in simulated and real neural dynamics* (Zhang*, C., Saggar)

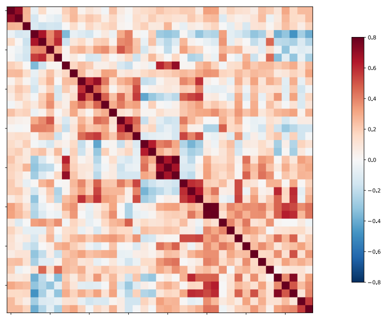
Code:

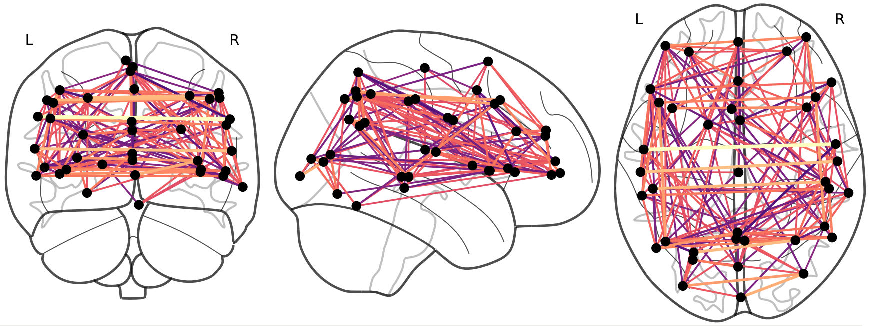
- https://github.com/trneedham/gromov-wasserstein-statistics
- https://github.com/trneedham/Spectral-Gromov-Wasserstein

Networks from data









Mathematical framework: Sturm's "space of spaces"

- Metric measure spaces equipped with GW distance
- Geodesics, spaces of directions, gradient flows
- General enough for many flavors of graph representations

Graph Learning:

- Graph Matching
- Graph Partitioning
- Graph Barycenters/Frechet means
- Graph PCA
  - Ideally all "without leaving graph space"

Contributions:

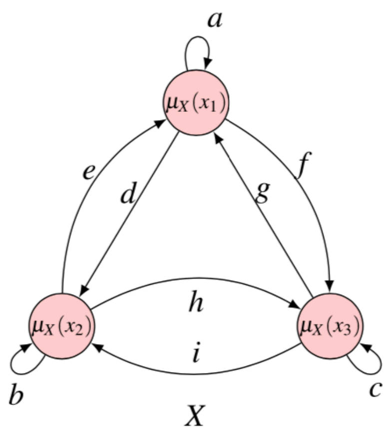
- Difficulties in adapting geometric statistics techniques (Pennec et al.) to Sturm's space of spaces
- (C., Needham '20) Heat kernel representations solve most difficulties

Problem setup

X	finite set
$\omega_X: X \times X \to \mathbb{R}$	edge weight function
μ_X	Borel probability measure
(X, ω_X, μ_X)	**network**

Example:
$G = (V, E)$ a (connected) graph

X	ω_X	μ_X
$X := V$	adjacency A^G distance d^G heat kernel $K^{G,t}$ $(\exp(-tL), L = D - A)$	uniform degree-based



Graph Matching problem (Umeyama 1988, Zaslavskiy-Bach-Vert 2009, many others):

G, H	graphs on n nodes
A^G, A^H	adjacency matrices
$\mathcal{P}$	permutation matrices on n nodes

Graph Matching problem: $\min_{P \in \mathcal{P}} \|A^G - PA^H P^T\|^2$

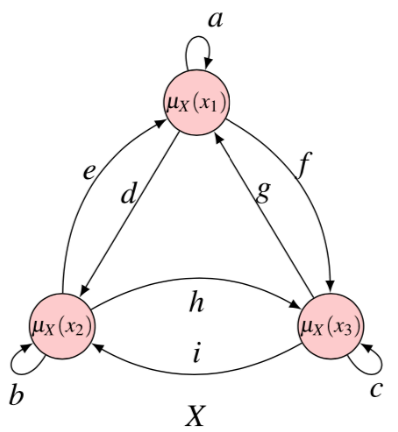
- When $|G| \neq |H|$, need to optimize over $|G| \times |H|$ matrices in $\{0,1\}$ with all row/column sums ≥ 1
- Carlsson, Mémoli, Ribeiro, Segarra 2013: Graph Matching $\approx$ Gromov-Hausdorff distance

Problem setup

X	finite set
$\omega_X: X \times X \to \mathbb{R}$	edge weight function
μ_X	Borel probability measure
(X, ω_X, μ_X)	**network**

Example:
$G = (V, E)$ a (connected) graph

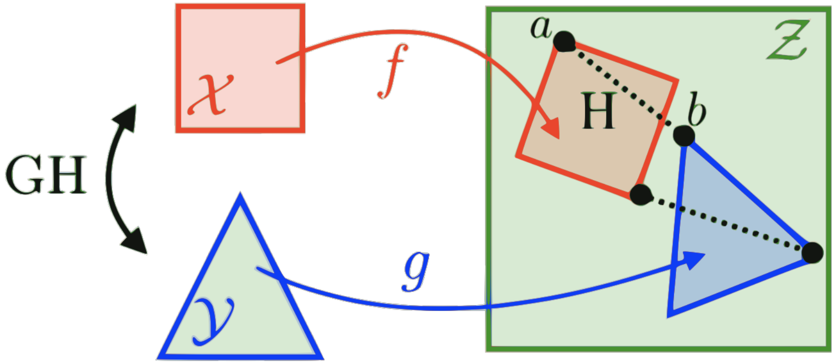
X	ω_X	μ_X
$X \coloneqq V$	adjacency A^G distance d^G heat kernel $K^{G,t}$ $(\exp(-tL), L = D - A)$	uniform degree-based



Gromov-Hausdorff distance between compact metric spaces:

$$d_{GH}((X, d_X), (Y, d_Y)) = \frac{1}{2} \inf \{ \sup_{(x,y),(x',y') \in R} |d_X(x, x') - d_Y(y, y')| : R \in \{0,1\}^{|X| \times |Y|}$$

with all row, column sums nonzero}



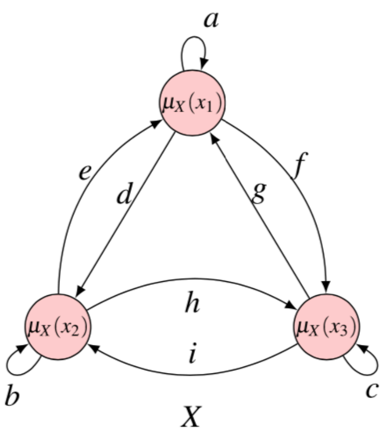
[Peyré, Cuturi
Computational Optimal Transport]

Problem setup

X	finite set
$\omega_X : X \times X \to \mathbb{R}$	edge weight function
μ_X	Borel probability measure
(X, ω_X, μ_X)	**network**

Example:
$G = (V, E)$ a (connected) graph

X	ω_X	μ_X
$X \coloneqq V$	adjacency A^G distance d^G heat kernel $K^{G,t}$ $(\exp(-tL), L = D - A)$	uniform degree-based



(Carlsson, Mémoli, Ribeiro, Segarra 2013) Generalized Gromov-Hausdorff distance between (finite) networks:

$$d_{\mathcal{N}}\big((X,\omega_X),(Y,\omega_Y)\big) = \frac{1}{2}\min\Big\{ \max_{(x,y),(x',y')\in R} |\omega_X(x,x') - \omega_Y(y,y')| : R \in \{0,1\}^{|X|\times|Y|}$$

with all row, column sums nonzero}

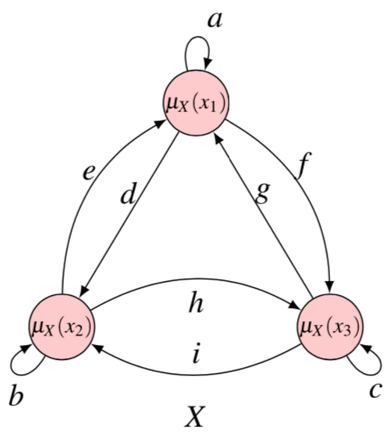
Hard combinatorial problem!

Problem setup

X	finite set
$\omega_X: X\times X \to \mathbb{R}$	edge weight function
μ_X	Borel probability measure
(X, ω_X, μ_X)	**network**

Example:
$G = (V, E)$ a (connected) graph

X	ω_X	μ_X
$X \coloneqq V$	adjacency A^G distance d^G heat kernel $K^{G,t}$ $(\exp(-tL), L = D - A)$	uniform degree-based



(Carlsson, Mémoli, Ribeiro, Segarra 2013) Generalized Gromov-Hausdorff distance between (finite) networks:

$$d_{\mathcal{N}}\big((X,\omega_X),(Y,\omega_Y)\big) = \frac{1}{2}\min\Big\{\max_{(x,y),(x',y')\in R} |\omega_X(x,x') - \omega_Y(y,y')|\colon R \in \{0,1\}^{|X|\times|Y|}$$

with all row, column sums nonzero}

Hard combinatorial problem!

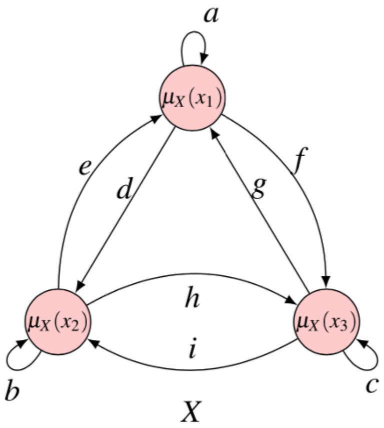
→ Convex relaxation

Problem setup

X	finite set
$\omega_X: X \times X \to \mathbb{R}$	edge weight function
μ_X	Borel probability measure
(X, ω_X, μ_X)	**network**

Example:
$G = (V, E)$ a (connected) graph

X	$\boldsymbol{\omega_X}$	$\boldsymbol{\mu_X}$
$X \coloneqq V$	adjacency A^G distance d^G heat kernel $K^{G,t}$ $(\exp(-tL), L = D - A)$	uniform degree-based

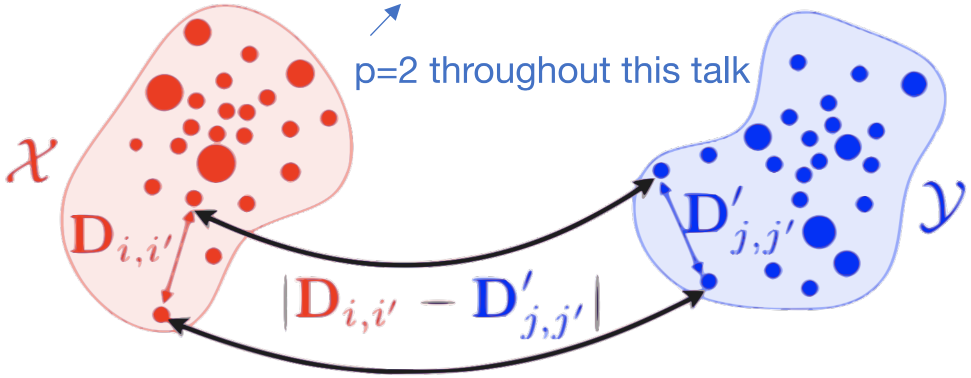


(Mémoli 2007) Gromov-Wasserstein distance between compact metric measure spaces:

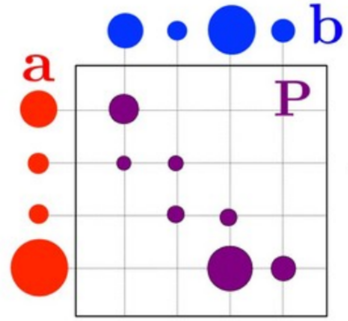
$$d_{GW}\big((X, d_X, \mu_X), (Y, d_Y, \mu_Y)\big) = \frac{1}{2}\min\ \{\|d_X - d_Y\|_{L^p(\mu\otimes\mu)}\colon\ \mu \in \Pi(\mu_X, \mu_Y), \mu\mathbf{1} = \mu_X,\ \mu^T\mathbf{1}\ = \mu_Y\}$$

p=2 throughout this talk

Gradient descent possible!



Coupling measures:



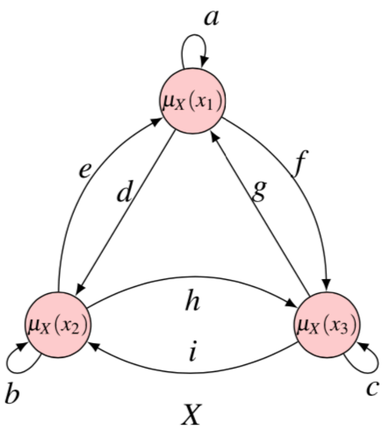
[Peyré, Cuturi
Computational Optimal Transport]

Problem setup

X	finite set
$\omega_X : X \times X \to \mathbb{R}$	edge weight function
μ_X	Borel probability measure
(X, ω_X, μ_X)	**network**

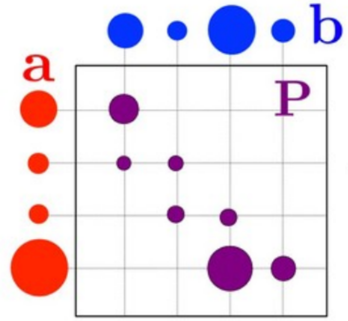
Example:
$G = (V, E)$ a (connected) graph

X	ω_X	μ_X
$X \coloneqq V$	adjacency A^G distance d^G heat kernel $K^{G,t}$ $(\exp(-tL), L = D - A)$	uniform degree-based

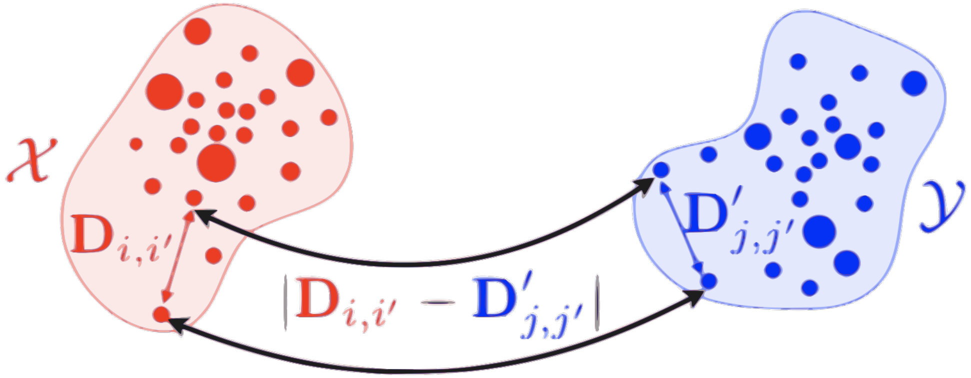


(Sturm 2012) Space of "almost"-metric measure spaces (satisfying triangle inequality a.e.) is a complete, geodesic space of nonnegative Alexandrov curvature → permits tangent spaces

Coupling measures:



Gradient descent possible!



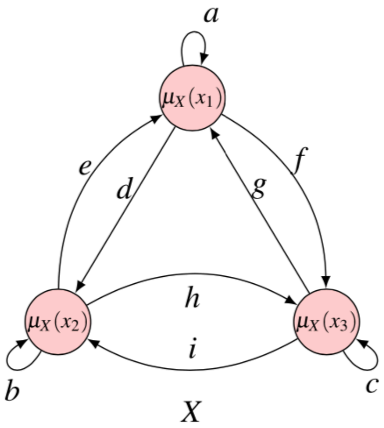
[Peyré, Cuturi
Computational Optimal Transport]

Problem setup

X	finite set
$\omega_X: X \times X \to \mathbb{R}$	edge weight function
μ_X	Borel probability measure
(X, ω_X, μ_X)	**network**

Example:
$G = (V, E)$ a (connected) graph

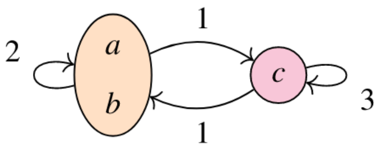
X	ω_X	μ_X
$X := V$	adjacency A^G distance d^G heat kernel $K^{G,t}$ $(\exp(-tL), L = D - A)$	uniform degree-based



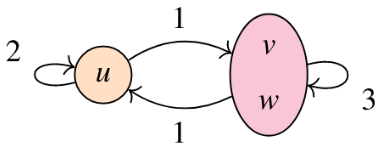
(C., Mémoli 2019) Gromov-Wasserstein distance defines a bona fide (pseudo)metric between networks:

$$d_{\mathcal{N}}\big((X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y)\big) = \frac{1}{2} \min \{\|\omega_X - \omega_Y\|_{L^p(\mu \otimes \mu)} : \mu \in \Pi(\mu_X, \mu_Y), \mu\mathbf{1} = \mu_X,\ \mu^T\mathbf{1} = \mu_Y\}$$

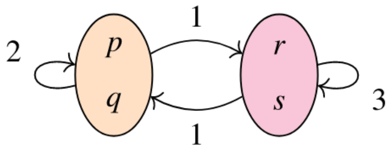
"Weak" isomorphism



$\omega_X(a,b,c) = \begin{pmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix}$, $\mu_X = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{2} \end{pmatrix}$



$\omega_Y(u,v,w) = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 3 \\ 1 & 3 & 3 \end{pmatrix}$, $\mu_Y = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$



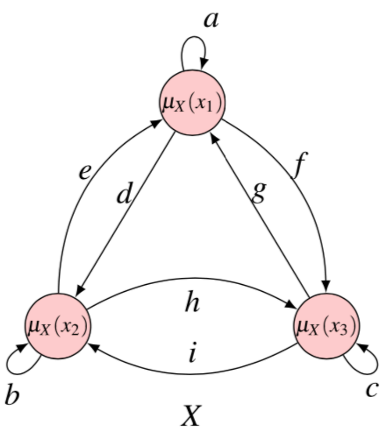
$\omega_Z(p,q,r,s) = \begin{pmatrix} 2 & 2 & 1 & 1 \\ 2 & 2 & 1 & 1 \\ 1 & 1 & 3 & 3 \\ 1 & 1 & 3 & 3 \end{pmatrix}$, $\mu_Z = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$

Problem setup

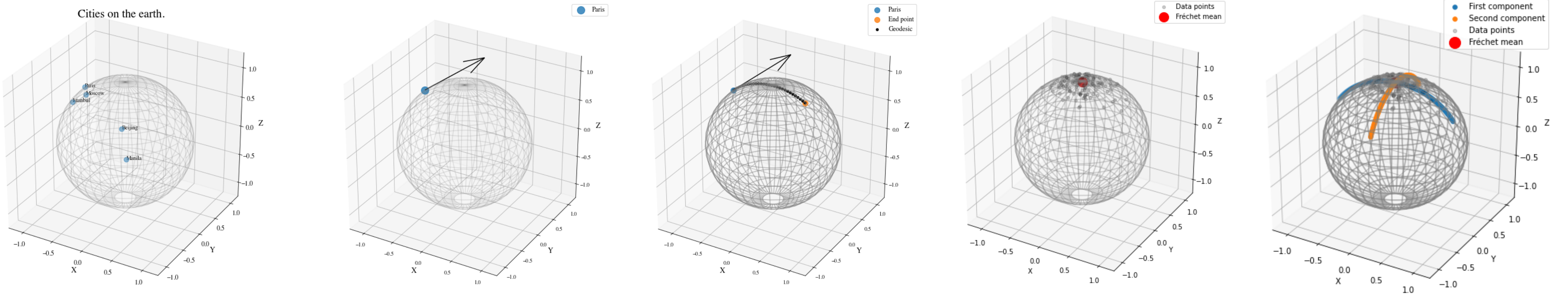
X	finite set
$\omega_X : X \times X \to \mathbb{R}$	edge weight function
μ_X	Borel probability measure
(X, ω_X, μ_X)	**network**

Example:
$G = (V, E)$ a (connected) graph

X	ω_X	μ_X
$X := V$	adjacency A^G distance d^G heat kernel $K^{G,t}$ $(\exp(-tL), L = D - A)$	uniform degree-based



(C., Needham 2019) Graph Learning framework = Graphs in GW space + GW Riemannian structures + “geomstats” in GW space



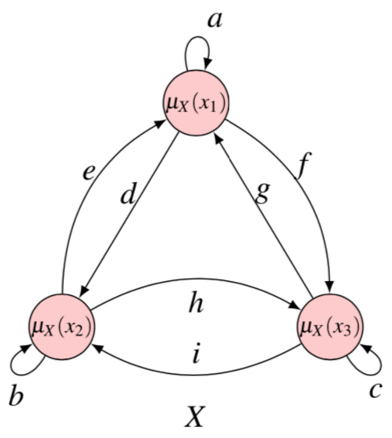
[Miolane et al., GeomStats INRIA-Stanford]

Problem setup

X	finite set
$\omega_X : X \times X \to \mathbb{R}$	edge weight function
μ_X	Borel probability measure
(X, ω_X, μ_X)	**network**

Example:
$G = (V, E)$ a (connected) graph

X	$\boldsymbol{\omega_X}$	$\boldsymbol{\mu_X}$
$X := V$	adjacency A^G distance d^G heat kernel $K^{G,t}$ $(\exp(-tL), L = D - A)$	uniform degree-based



But this is hard…

- "Sturm geodesics" between $(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y)$ occur over product space $X \times Y$, difficult to handle numerically
  - Exponential time/space complexity for computing barycenters

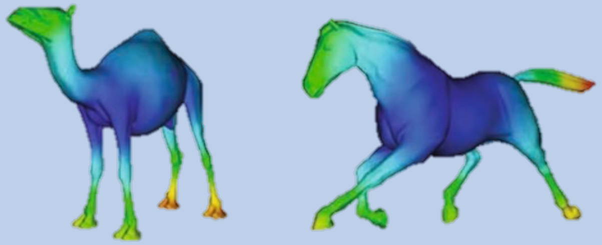
(C., Needham 2019) Compute geodesics over support of an optimal coupling $\mu \in \Pi(\mu_X, \mu_Y)$

- Sparse couplings support tractable computations
- Little known about structure of couplings except special cases (GW matching between points on $\mathbb{R}$ [Vayer et al., Sliced GW 2019] or subsets of $\mathbb{R}^n$ with rotationally invariant measures [Sturm 2020]

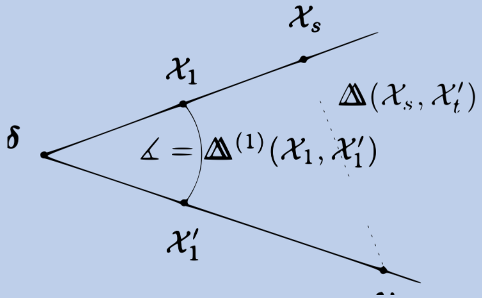
(C., Needham 2020) Networks represented by heat kernel yield sparse couplings with $o(n)$ nonzero entries

1. Problem setup

2. Gromov-Wasserstein distance and Sturm's constructions

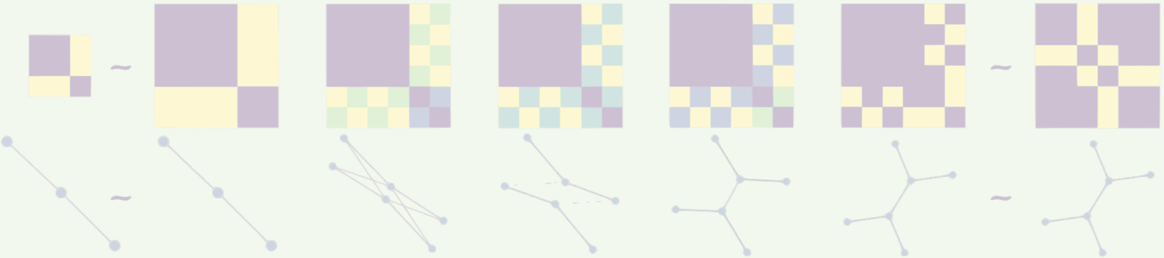


[Mémoli 2007]

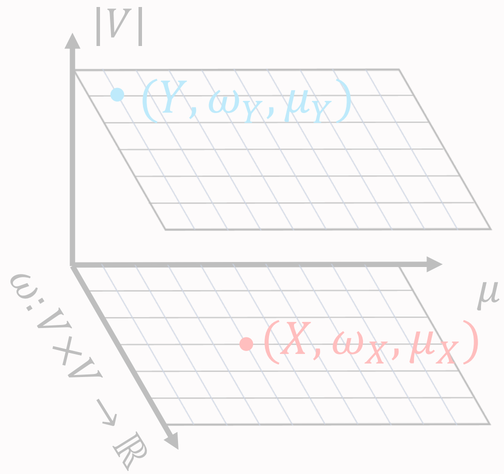


[Sturm 2012]

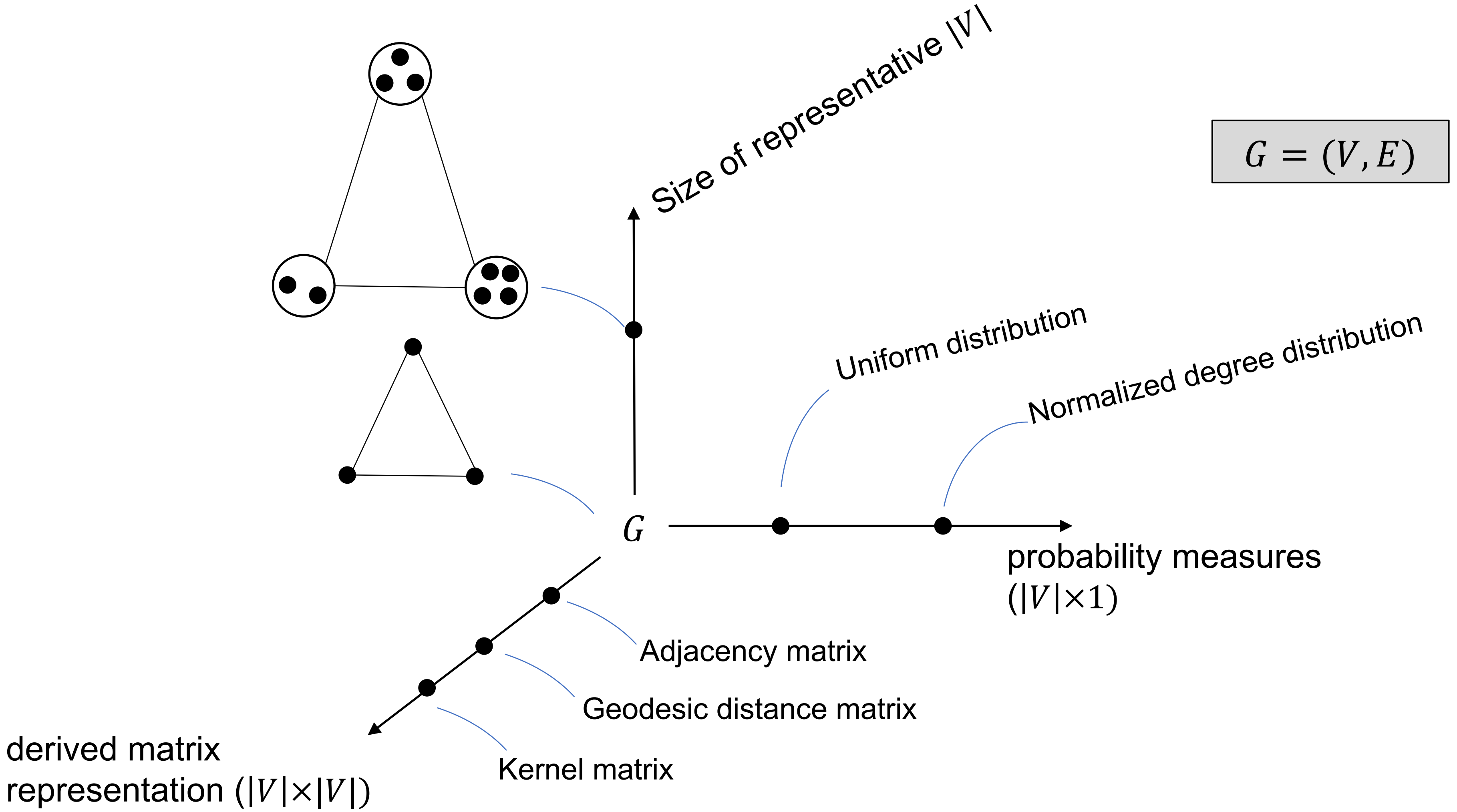
3. Statistical learning in the Riemannian framework



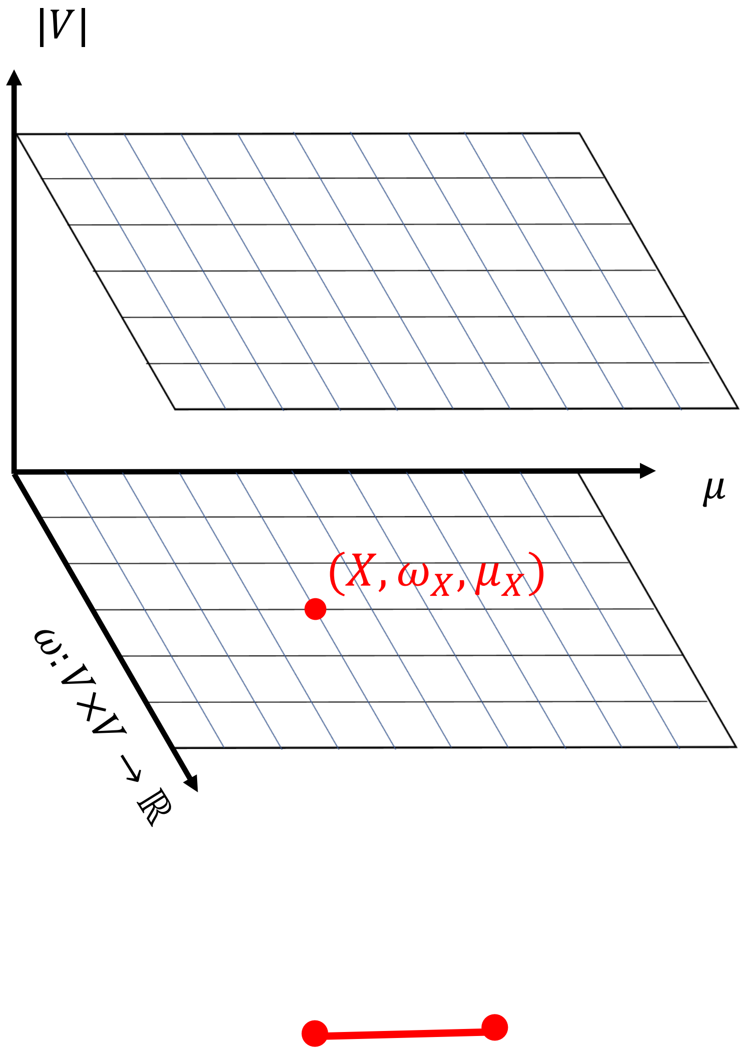
4. Future directions

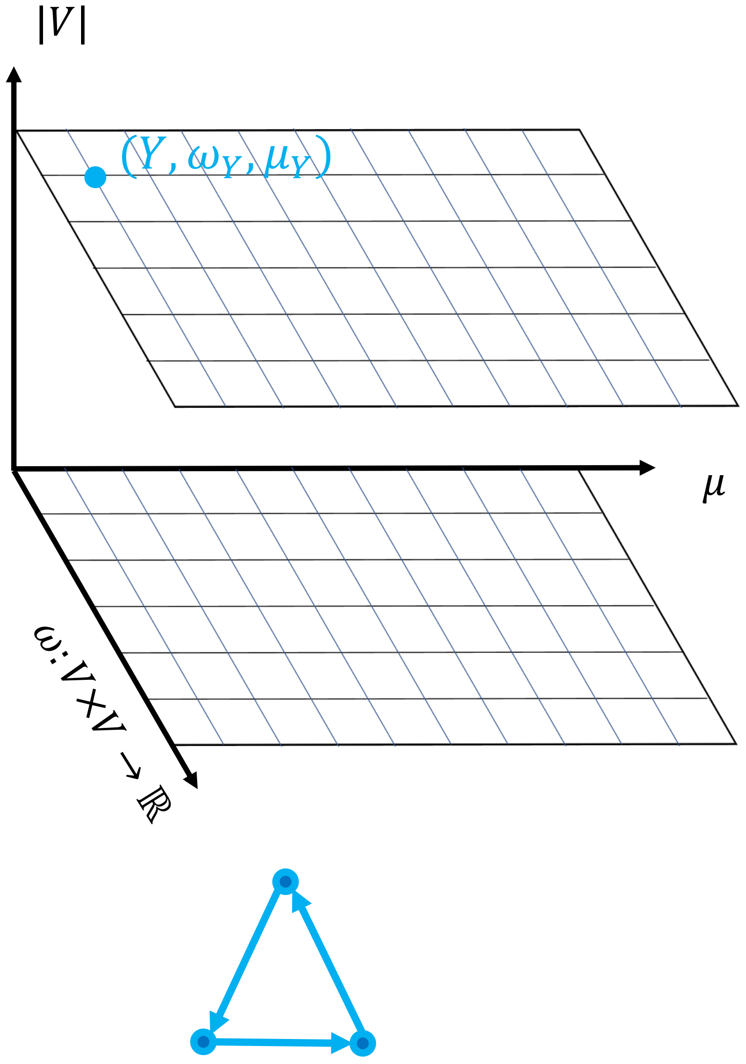


A coordinate system for graphs I

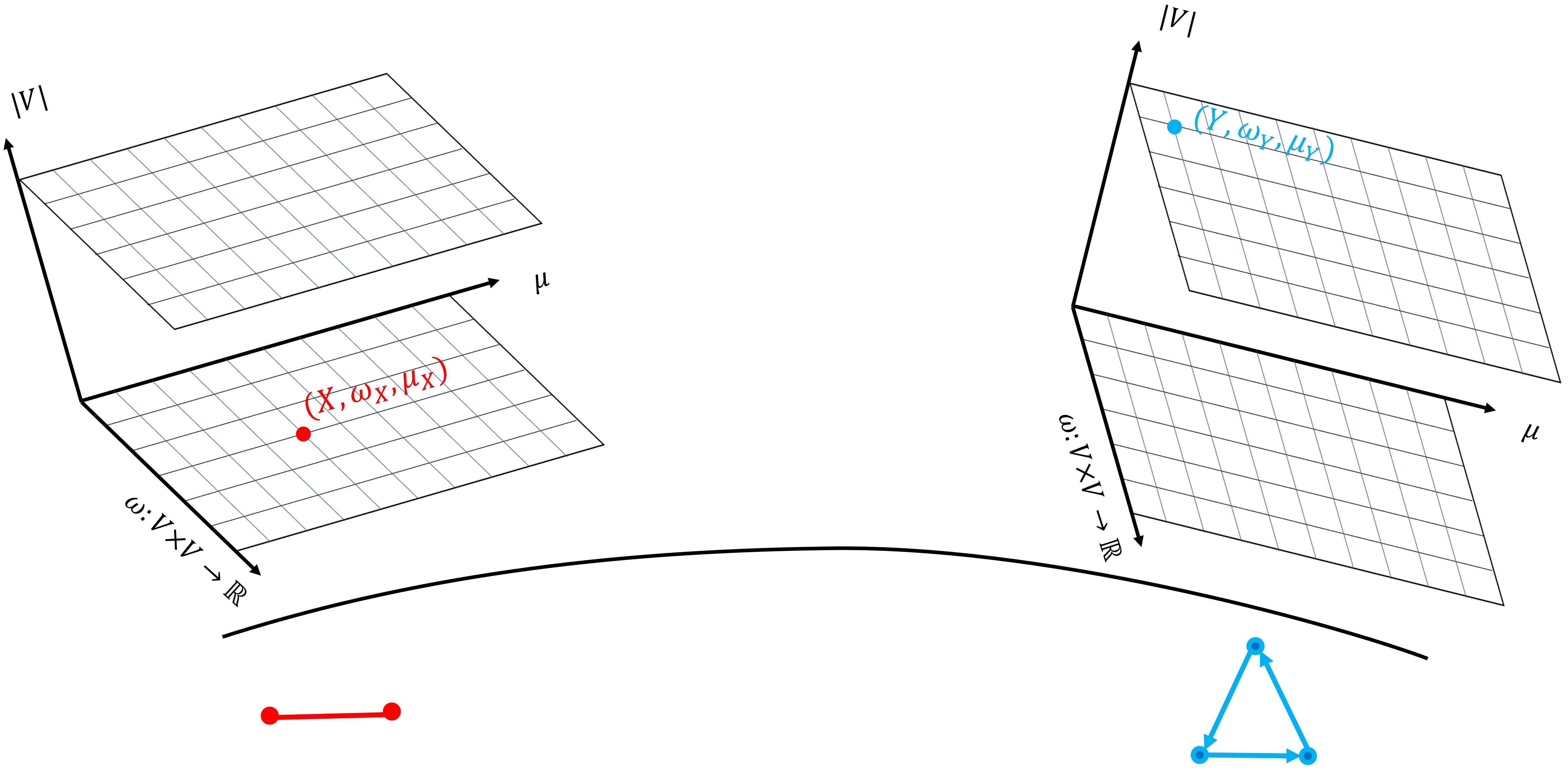


A coordinate system for graphs II

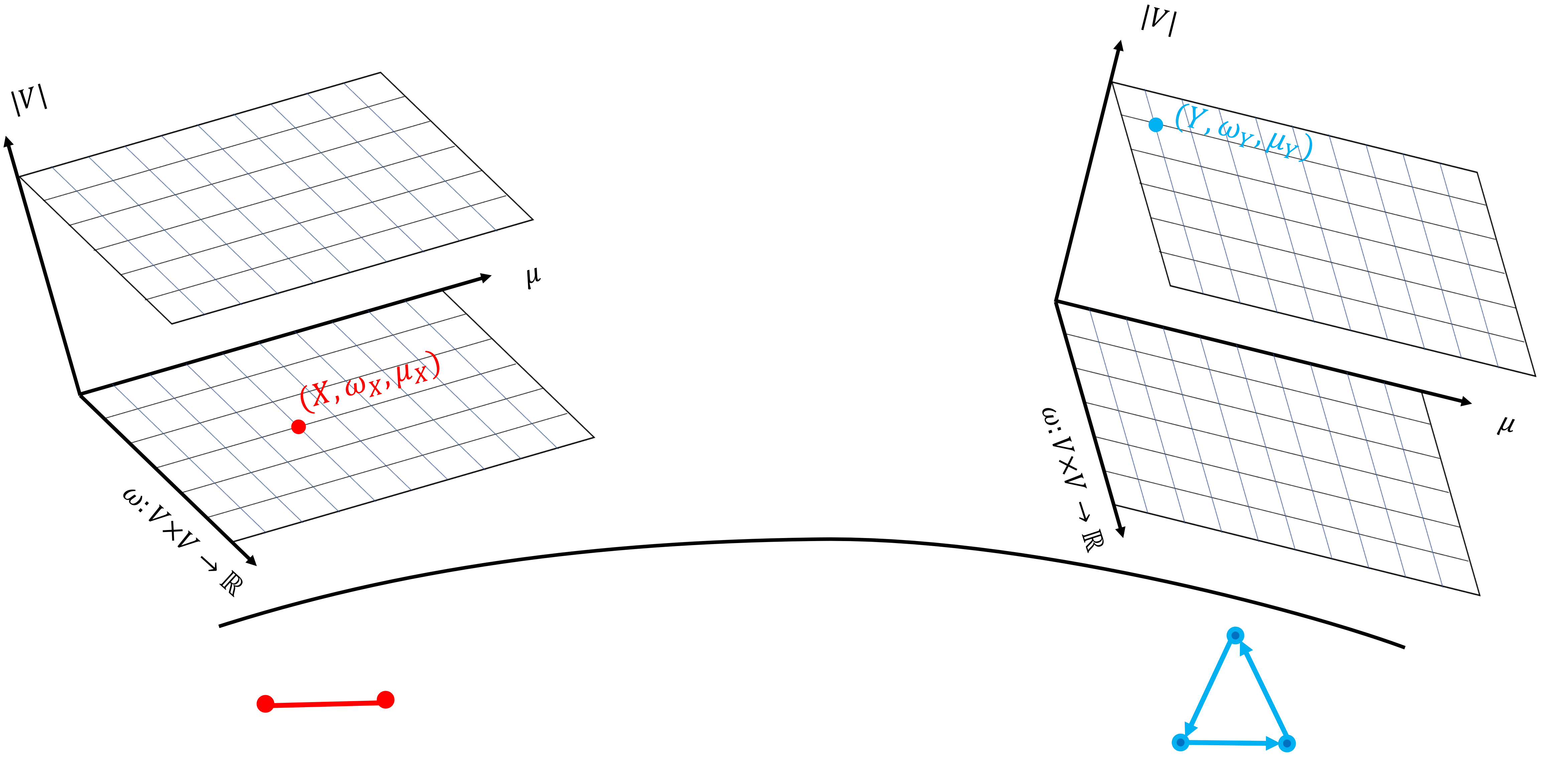




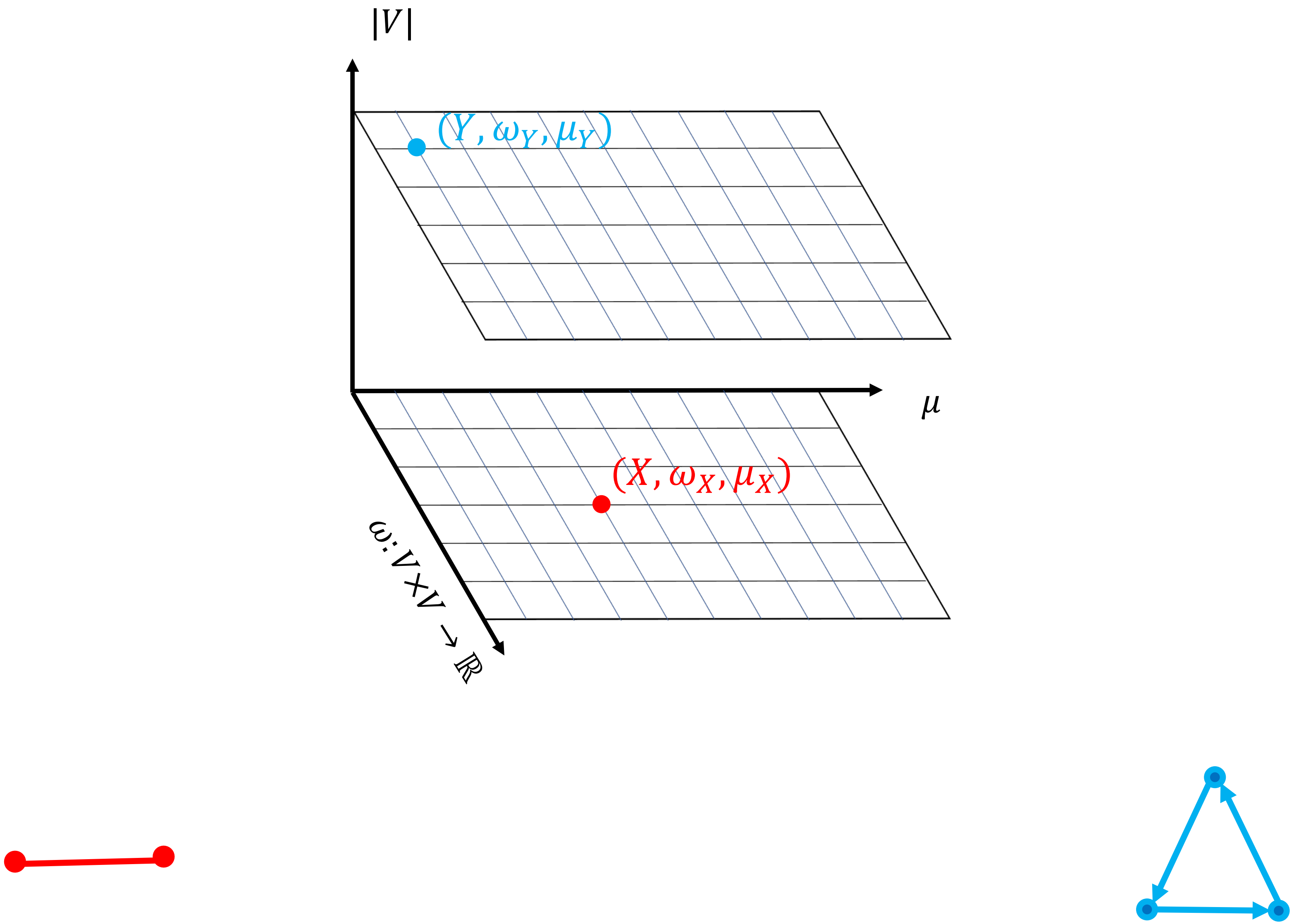
A coordinate system for graphs II



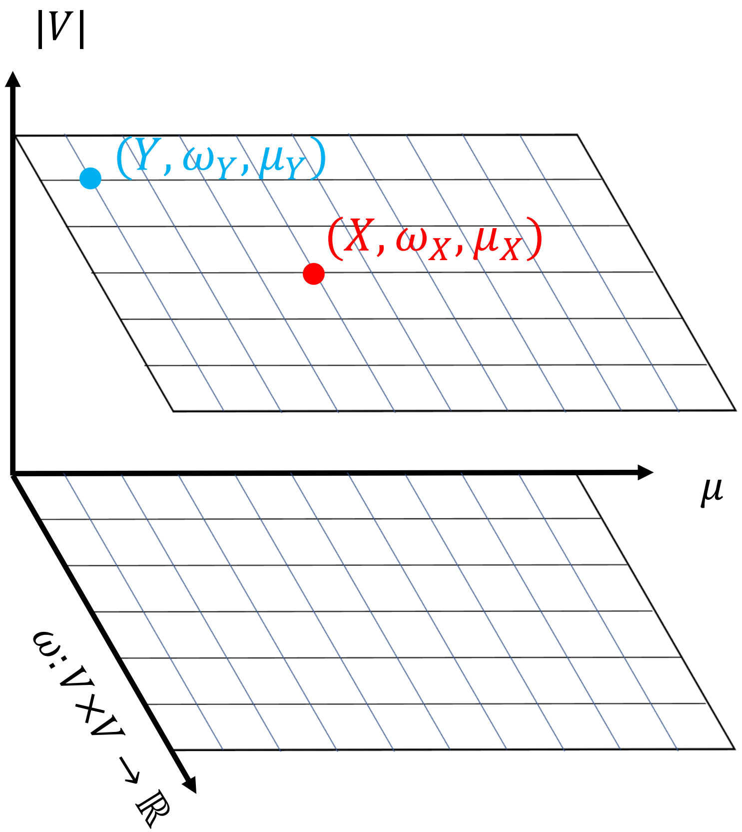
A coordinate system for graphs II



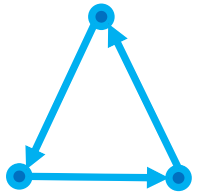
A coordinate system for graphs II



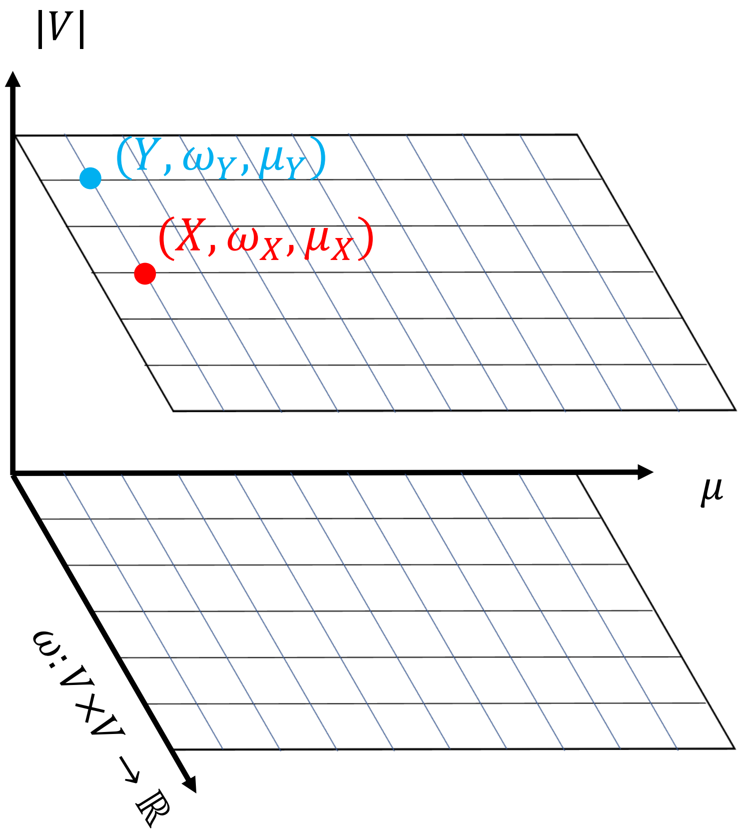
A coordinate system for graphs II



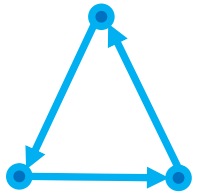
Blow-up



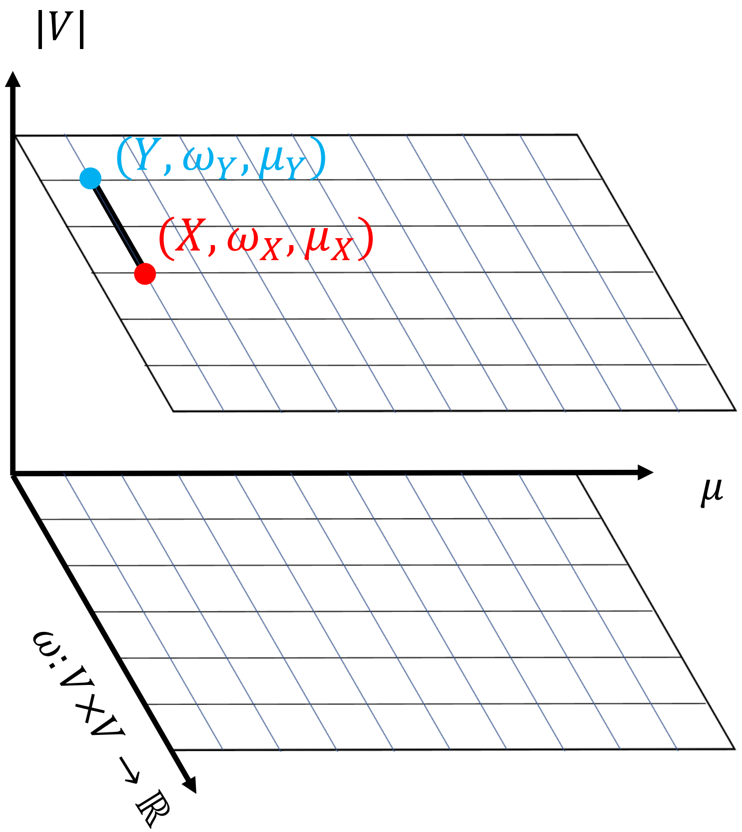
A coordinate system for graphs II



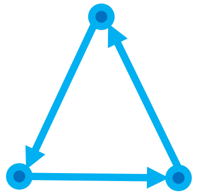
Align



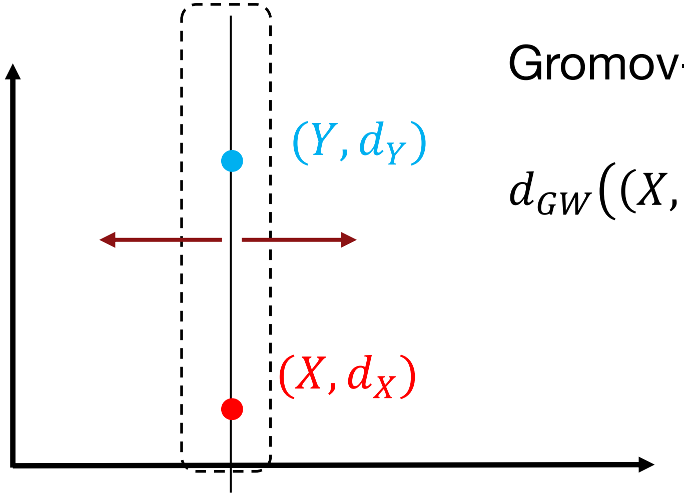
A coordinate system for graphs II



Frobenius product



A coordinate system for graphs II



$d: V \times V \to [0, \infty)$
metric functions

Gromov-Wasserstein distance between compact metric measure spaces:

$$d_{GW}\big((X, d_X, \mu_X), (Y, d_Y, \mu_Y)\big) = \frac{1}{2} \min \{\|d_X - d_Y\|_{L^p(\mu \otimes \mu)}: \mu \in \Pi(\mu_X, \mu_Y), \mu\mathbf{1} = \mu_X,\ \mu^T\mathbf{1} = \mu_Y\}$$

(Sturm 2012) Space of metric measure spaces is not complete under d_{GW}

- completion is the space of "almost" metric measure spaces that satisfy triangle inequality a.e.
- Riemannian structures (including exponential maps) in ambient L^2 space where triangle inequality is removed altogether
- Ambient L^2 space is nonnegatively curved

Gromov-Wasserstein distances for network comparison

Gromov-Wasserstein distance between compact metric measure spaces:

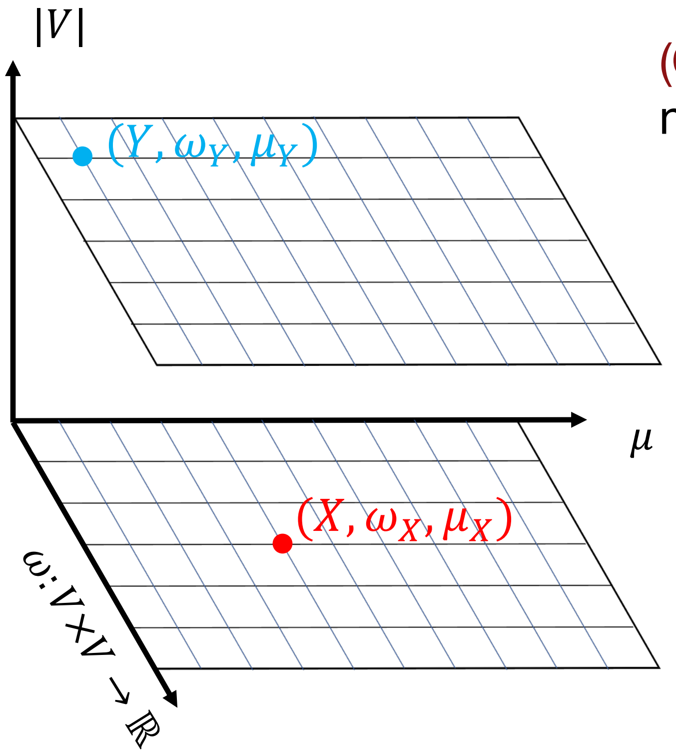
$$d_{GW}\big((X, d_X, \mu_X), (Y, d_Y, \mu_Y)\big) = \frac{1}{2}\min\ \{\|d_X - d_Y\|_{L^p(\mu\otimes\mu)}\colon\ \mu \in \Pi(\mu_X, \mu_Y), \mu\mathbf{1} = \mu_X,\ \mu^T\mathbf{1} = \mu_Y\}$$

Isomorphism structure:
$d_{GW}(X, Y) = 0 \iff X, Y$ isometric
as metric measure spaces

$$\begin{tikzcd} & (Z, \mu_Z) \arrow[dl, "f"'] \arrow[dr, "g"] & \\ X & & Y \end{tikzcd}$$

- f, g bijective, Borel mble
- $\|f^* d_X - g^* d_Y\|_\infty = 0$
- $f_\# \mu_Z = \mu_X,\ g_\# \mu_Z = \mu_Y$

Gromov-Wasserstein distances for network comparison



(C., Mémoli 2019) Generalized Gromov-Wasserstein distance between compact measure networks:

$$d_{\mathcal{N}}\big((X,\omega_X,\mu_X),(Y,\omega_Y,\mu_Y)\big) = \frac{1}{2}\min\,\{\|\omega_X - \omega_Y\|_{L^p(\mu\otimes\mu)}\colon\ \mu \in \Pi(\mu_X,\mu_Y), \mu\mathbf{1} = \mu_X,\ \mu^T\mathbf{1} = \mu_Y\}$$

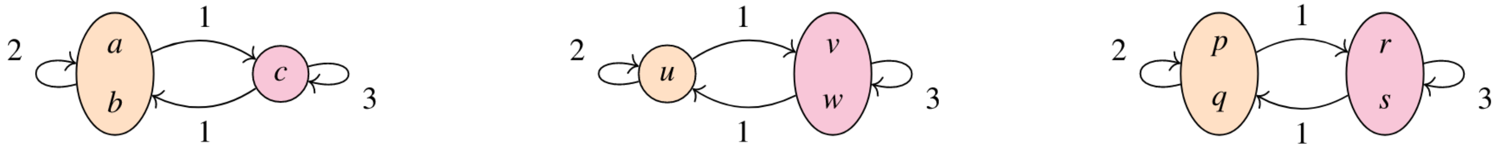
"Weak" isomorphism structure:

$d_{\mathcal{N}}(X,Y) = 0 \iff X, Y$ related by a "tripod" ($X \cong^w Y$)

- Any choice of ω gives a metric modulo this isomorphism structure

$$\begin{tikzcd} & (Z,\mu_Z) \arrow[dl, "f"'] \arrow[dr, "g"] & \\ X & & Y \end{tikzcd}$$

- f, g Borel
- $\|f^*\omega_X - g^*\omega_Y\|_\infty = 0$
- $f_\# \mu_Z = \mu_X,\ g_\# \mu_Z = \mu_Y$



$\omega_X(a,b,c) = \begin{pmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 3 \end{pmatrix}$, $\mu_X = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{2} \end{pmatrix}$ $\omega_Y(u,v,w) = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 3 \\ 1 & 3 & 3 \end{pmatrix}$, $\mu_Y = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$ $\omega_Z(p,q,r,s) = \begin{pmatrix} 2 & 2 & 1 & 1 \\ 2 & 2 & 1 & 1 \\ 1 & 1 & 3 & 3 \\ 1 & 1 & 3 & 3 \end{pmatrix}$, $\mu_Z = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}$

GW as a nonconvex quadratic optimization problem

Networks: $(X, p), \ (Y, q)$

Couplings/transport plans: $\mathcal{C}(p,q) \coloneqq \{C \in \mathbb{R}^{n\times m} : C\mathbf{1} = p, C^T\mathbf{1} = q, C \succcurlyeq 0\}$

Gromov-Wasserstein (GW) problem:

$$d_{GW}(X,Y) \coloneqq \frac{1}{2}\min_{C\in\mathcal{C}(p,q)} \left(\sum_{i,j,k,l} \left|X_{ik} - Y_{jl}\right|^2 C_{kl}C_{ij}\right)^2$$

Matrix notation: write (X, ω_X, μ_X) as (X, p)

Complexity: $O(n^3 \log(n))$

After unrolling into $(nm \times 1)$-dimensional vectors, the problem reads:

minimize $\langle C, JC \rangle$

subject to coupling and nonnegativity constraints

Gradient of map $C \mapsto \langle C, JC \rangle$ (after reshaping into matrix form):

$$(J + J^T)C$$

GW as a nonconvex quadratic optimization problem

Networks: $(X, p),\ (Y, q)$

Couplings/transport plans: $\mathcal{C}(p,q) \coloneqq \{C \in \mathbb{R}^{n\times m} : C\mathbf{1} = p, C^T\mathbf{1} = q, C \succcurlyeq 0\}$

Gromov-Wasserstein (GW) problem:

$$d_{GW}(X,Y) \coloneqq \frac{1}{2} \min_{C\in\mathcal{C}(p,q)} \left(\sum_{i,j,k,l} \left|X_{ik} - Y_{jl}\right|^2 C_{kl} C_{ij} \right)^2$$

Matrix notation: write (X, ω_X, μ_X) as (X, p)

Complexity: $O(n^3 \log(n))$

After unrolling into $(nm \times 1)$-dimensional vectors, the problem reads:

minimize $\langle C, JC \rangle$

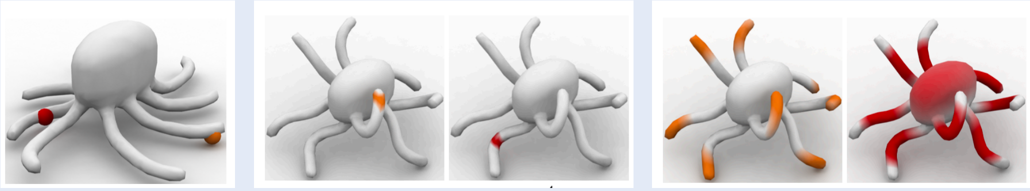
subject to coupling and nonnegativity constraints

Gradient of map $C \mapsto \langle C, JC \rangle$ (after reshaping into matrix form):

$$(J + J^T)C$$

Regularization techniques speed up computations at the expense of losing sparse couplings:

Solomon, Peyré, Kim, Sra SIGGRAPH 2016



Source Sparse coupling Non-sparse coupling

Sturm's constructions

[Sturm 2012. *The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces*]

$(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y)$ measure networks, C optimal measure coupling

"product geodesic": $\gamma(t) \coloneqq (X \times Y, \Omega_t, C)$,

$$\Omega_t\big((x,y),(x',y')\big) \coloneqq (1-t)\omega_X(x,x') + t\omega_Y(y,y')$$

Sturm's constructions

[Sturm 2012. *The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces*]

$(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y)$ measure networks, C optimal measure coupling

Attach a copy of Y to each point of X to resize

"product geodesic": $\gamma(t) := (X \times Y, \Omega_t, C)$,

Align terminal network to initial network

$$\Omega_t\big((x, y), (x', y')\big) := (1 - t)\omega_X(x, x') + t\omega_Y(y, y')$$

After resizing and realigning, take **linear** combination

Sturm's constructions

[Sturm 2012. *The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces*]

$(X, \omega_X, \mu_X), (Y, \omega_Y, \mu_Y)$ measure networks, C optimal measure coupling

Attach a copy of Y to each point of X to resize

"product geodesic": $\gamma(t) \coloneqq (X \times Y, \Omega_t, C)$,

Align terminal network to initial network

$$\Omega_t\big((x,y),(x',y')\big) \coloneqq (1-t)\omega_X(x,x') + t\omega_Y(y,y')$$

After resizing and realigning, take **linear** combination

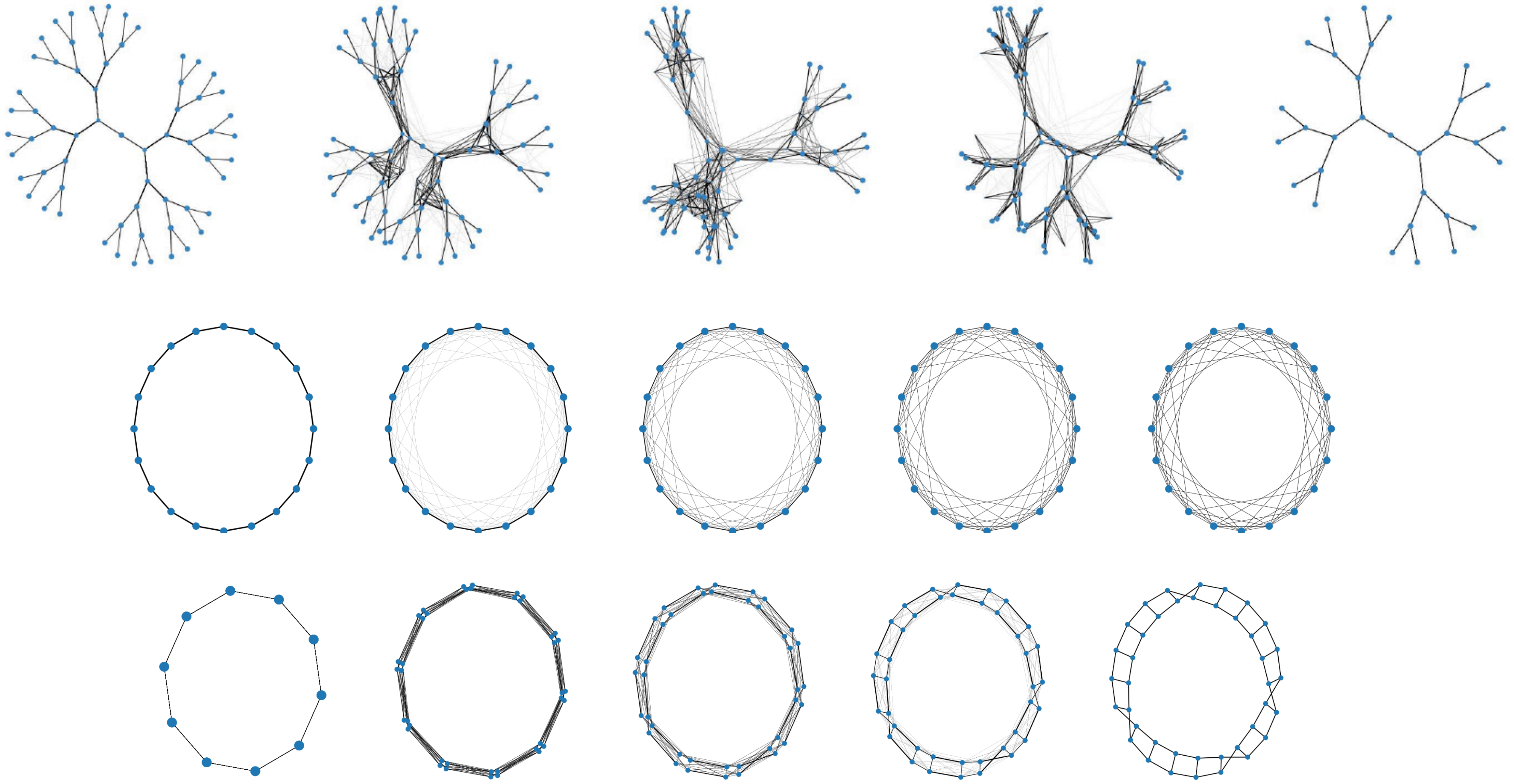
Weak isomorphism structure $\Longrightarrow$ $\Omega_0 \cong^w X, \quad \Omega_1 \cong^w Y$

- $v \coloneqq \Omega_1 - \Omega_0$
- To go from X to Y ``via vector addition'', take: $\boxed{\Omega_0 + v = \Omega_1}$
- Exponential map: vector addition modulo isomorphism

Formal definitions of tangent cones and exponential map in [Sturm 2012]

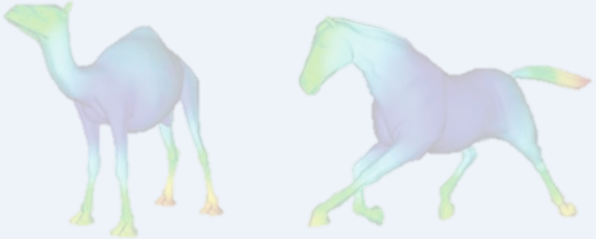
- Also [C., Needham 2020] for related details

Visualizations of GW geodesics

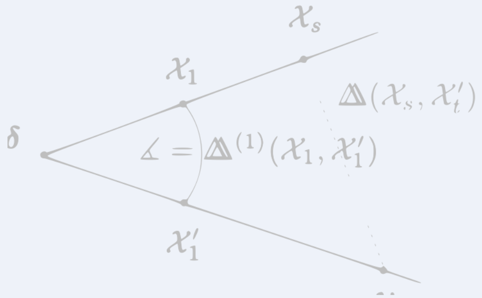


1. Problem setup

2. Gromov-Wasserstein distance and Sturm’s constructions

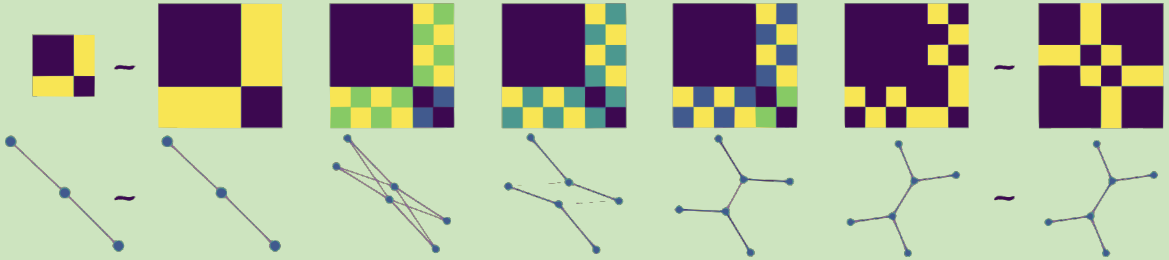


[Mémoli 2007]

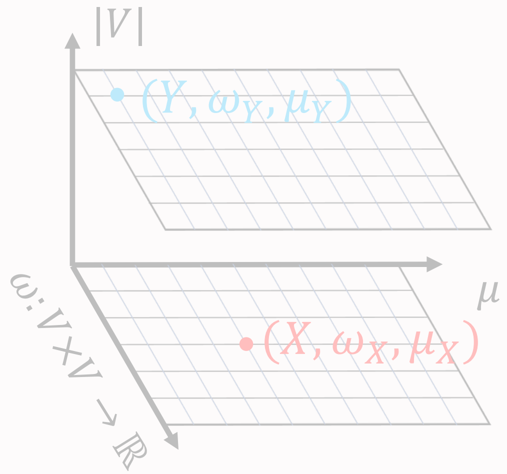


[Sturm 2012]

3. Statistical learning in the Riemannian framework



4. Future directions

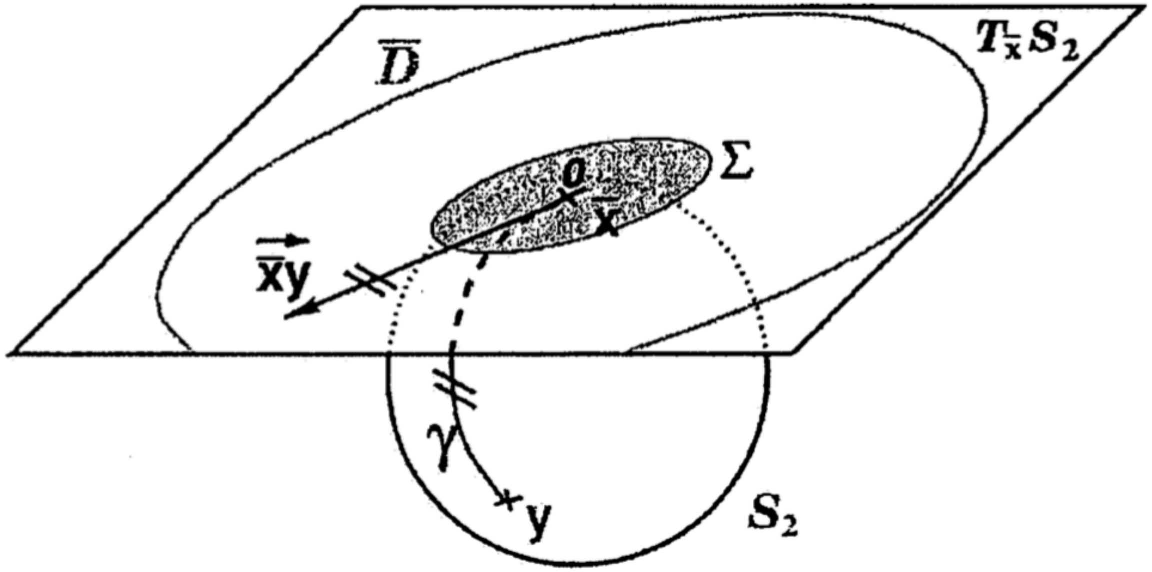


Iterative averaging

[Pennec 2006. *Intrinsic statistics on Riemannian manifolds*]

Pennec (2006): Start with a "seed" point x on a manifold and points $y_1, y_2 \dots y_n$

- Use log maps to lift geodesics $x \to y_i$ to vectors in T_x
- Average in T_x
- Exp down to manifold, iterate
- Each iterate is a gradient descent step for the Fréchet functional



$\{(X_i, \omega_i, \mu_i)\}_{i=1}^n$ a collection of finite measure networks. Given (X_0, ω_0, μ_0), define the Fréchet functional:

$$F(X_0) := \frac{1}{n}\sum_{i=1}^{n} d_{\mathcal{N}}(X_0, X_i)^2$$

- Theorem (C., Needham 2019): The Fréchet functional on $\mathcal{N}$ is differentiable, and its gradient descent steps are given by the log-average-exp iterative scheme

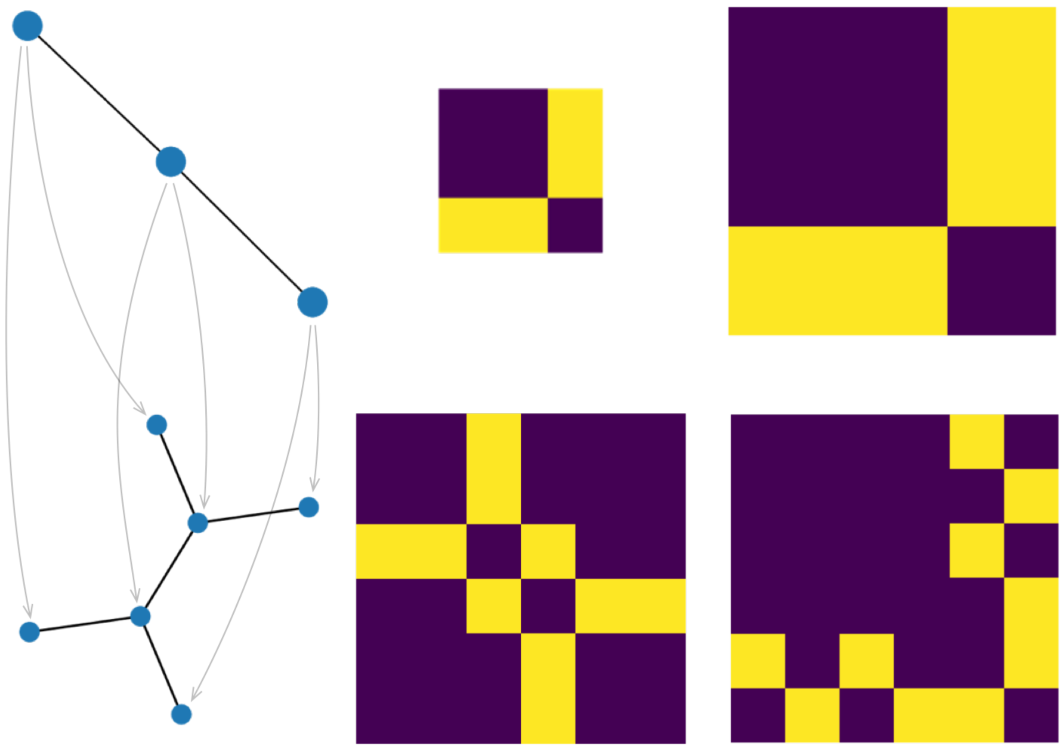
Sparse geodesics

[C., Needham 2020. *Gromov-Wasserstein Averaging in a Riemannian Framework*]

- Product geodesics incur exponential cost for any iterative method

- Instead, use sparsity of optimal couplings to only blow-up points as needed
  - Poses challenge for entropic regularization
  - Gradient descent to get local optima of GW cost

- This yields sparse geodesics and computationally tractable exp maps

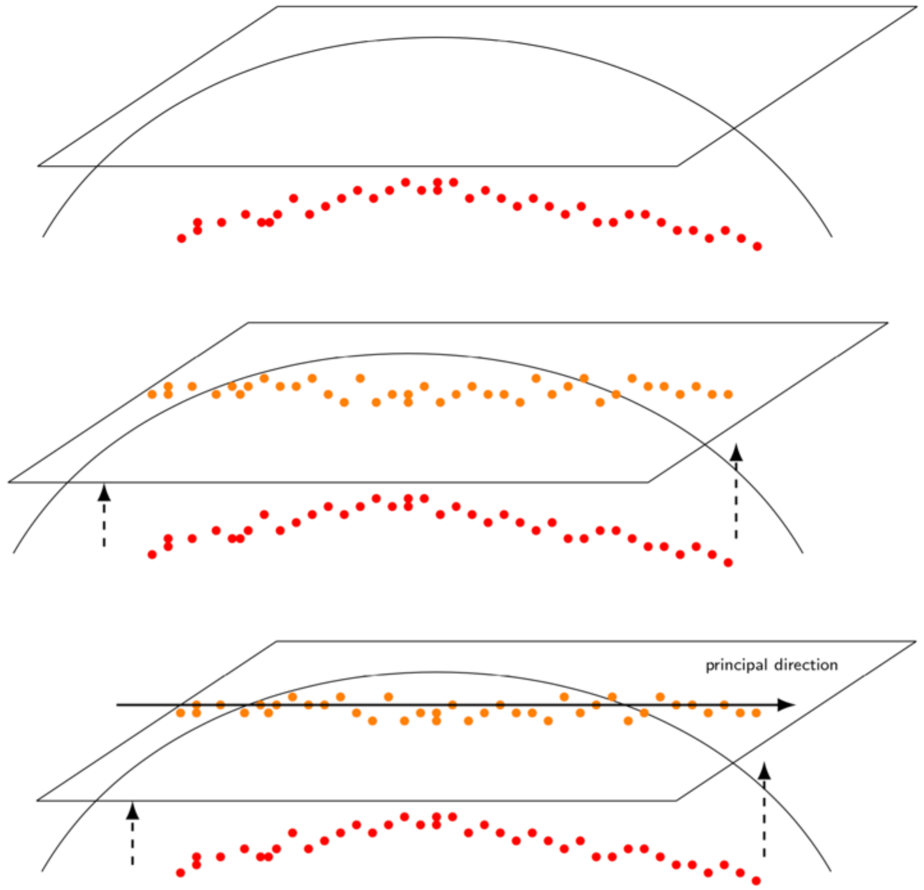


Proof of concept: Tangent PCA

Goal: Verify that sparse geodesics + exp map makes sense

Proof of concept: Tangent PCA

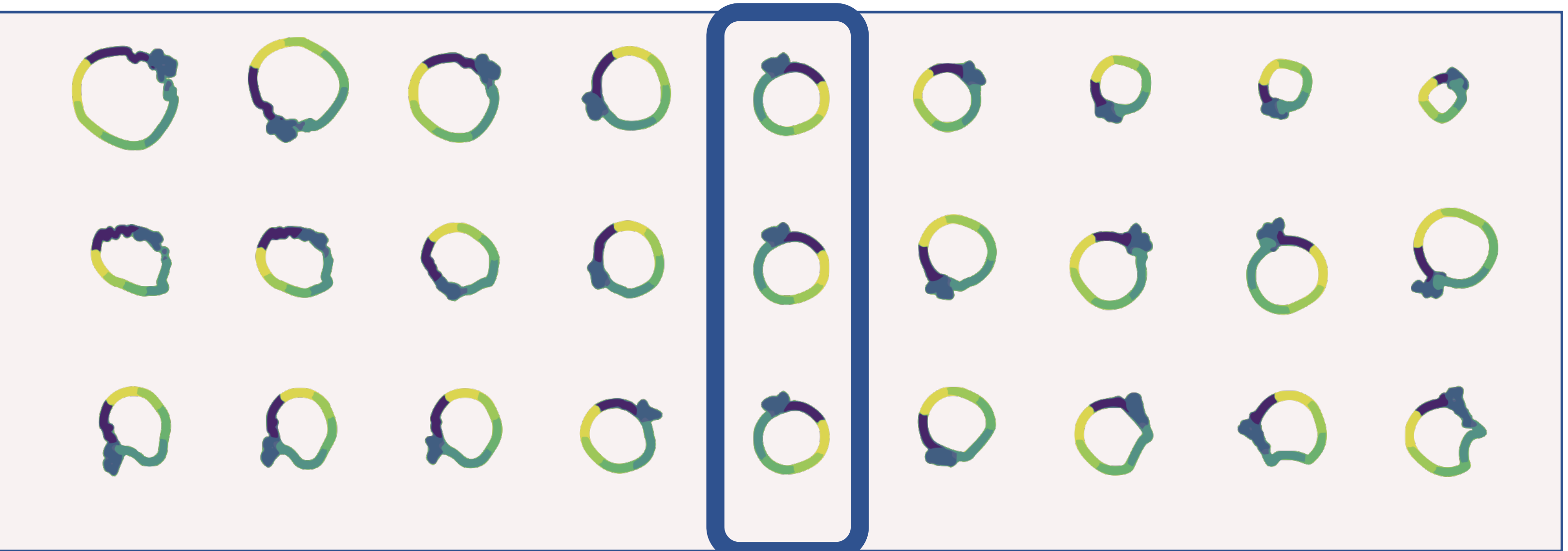
Goal: Verify that sparse geodesics + exp map makes sense

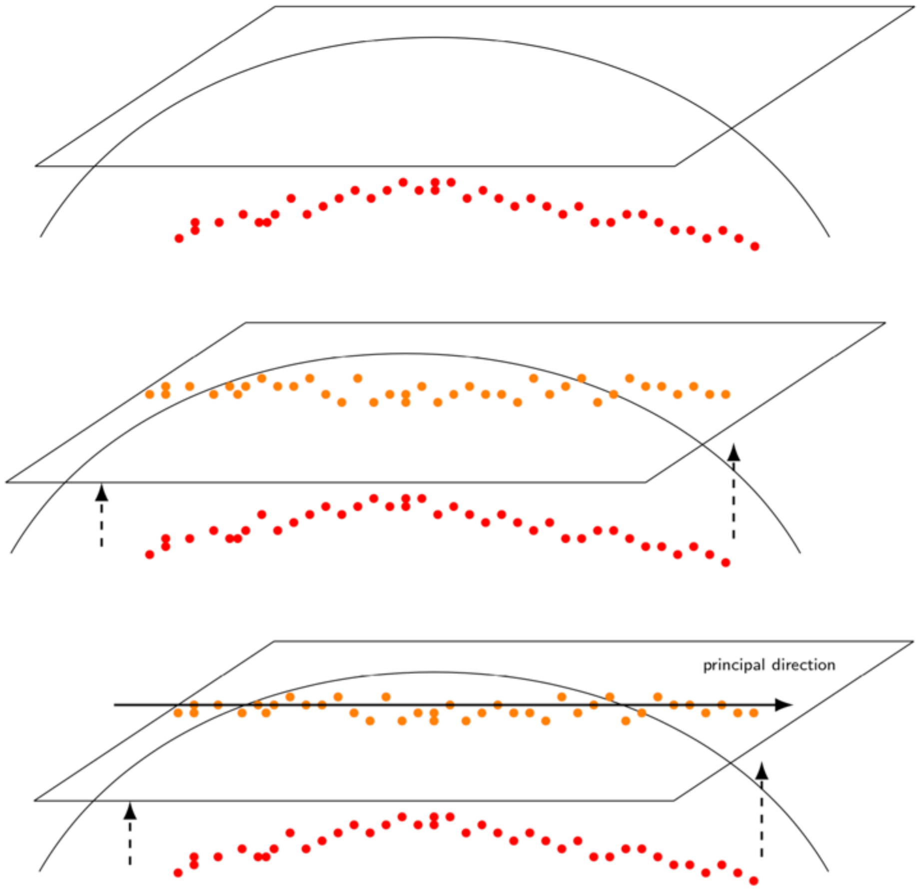


Proof of concept: Tangent PCA

Goal: Verify that sparse geodesics + exp map makes sense

First three principal directions

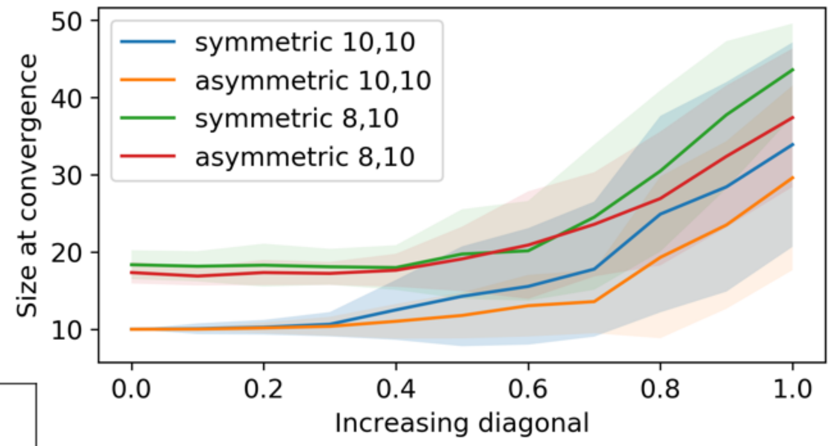


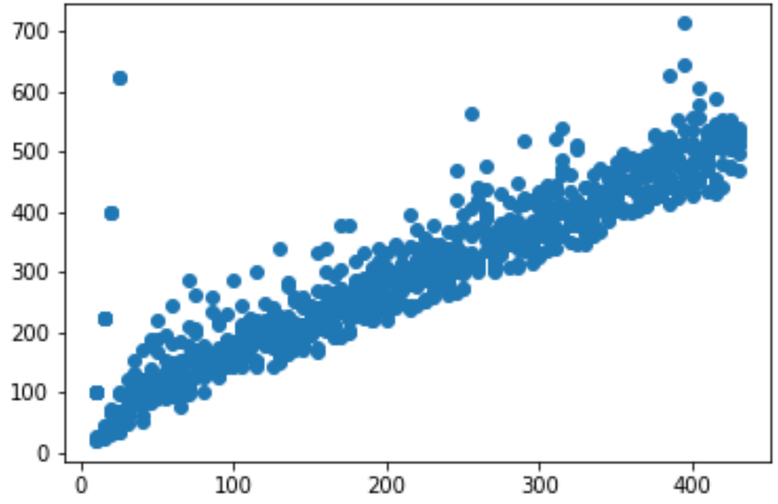


When can things go wrong?

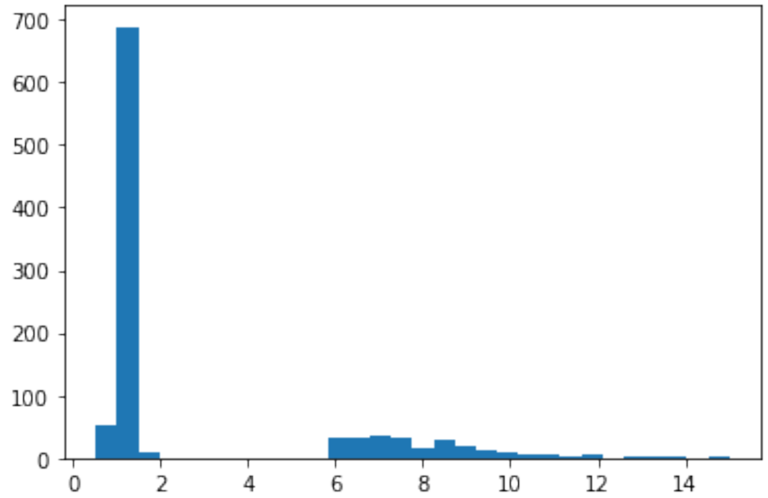
- If optimal couplings are dense, then iterative exp map computations become impractical
- Necessary to identify classes of ω for which optimal couplings remain sparse
- For graphs with adjacency loss and geodesic distance loss, we empirically observe sparsity

Completely random matrices with random diagonal entries show excessive blow-ups





For random measure networks with adjacency loss, couplings remain sparse and grow linearly with size of network



IMDB network, geodesic distance loss: Histogram shows support size of optimal coupling divided by total size of networks being compared

Return to “Spectral GW”

[Mémoli 2011. *A spectral notion of Gromov-Wasserstein distance and related methods]*

Using the “any $\omega : V \times V \to \mathbb{R}$ gives a (pseudo)metric” approach, consider the following:

- G, H two graphs, $t > 0$
- $K^{G,t}, K^{H,t}$ heat kernels (exponentials of graph Laplacians)
- Form the GW loss using $||K^{G,t} - K^{H,t}||_{L^p(\mu \otimes \mu)}$
  - Simplified discrete analogue of the Spectral Gromov-Wasserstein distance introduced in [Mémoli 2011]: we don’t need to worry about blowups of the heat kernel, and we do not optimize over t
  - [Mémoli 2011] related to earlier work of Reuter et al 2006 (”Shape-DNA”), Kasue-Kumura (1994)

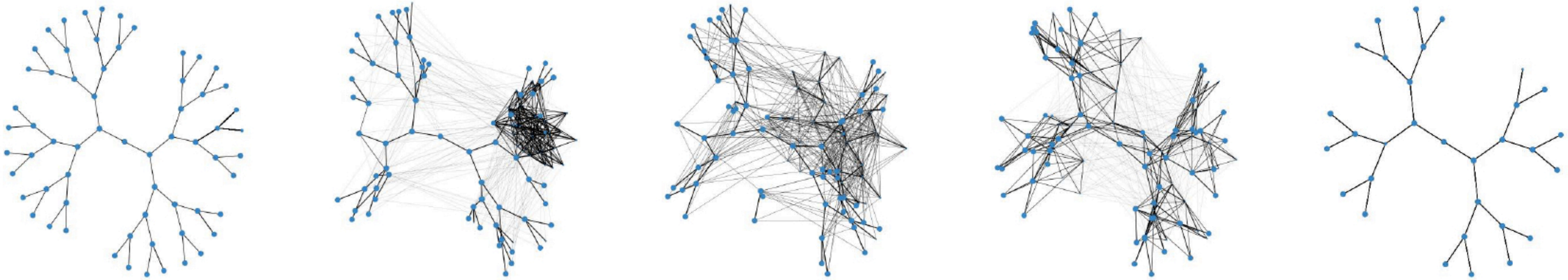
Theorem (C., Needham 2020): For spectral loss, number of nonzero entries in optimal coupling is $o(n)$.
[GW problem becomes maximization of convex function, use KKT conditions + dimension counting]

- Consequence: No blowups in iterative log-exp maps

Observation: Running MCMC to sample different initial couplings for gradient descent suggests that the loss landscape for spectral loss is much nicer than for adjacency loss

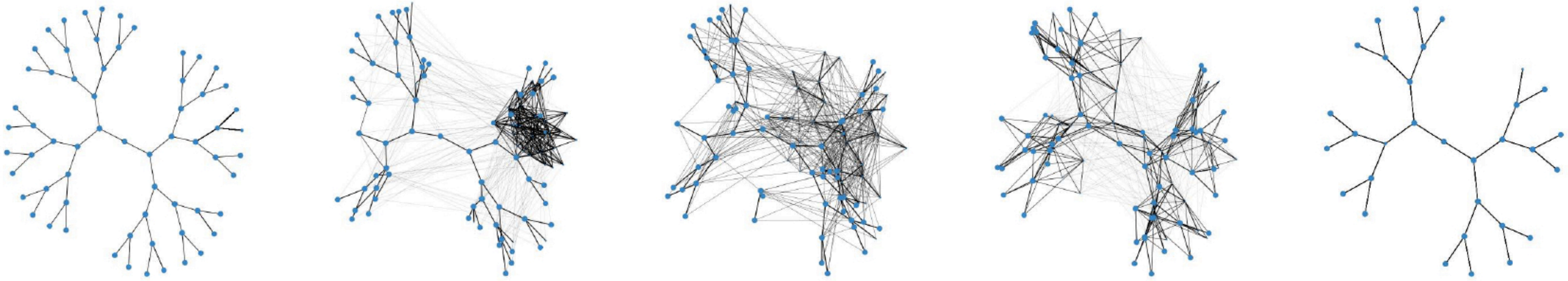
Geodesics via adjacency loss and spectral loss

Adjacency loss

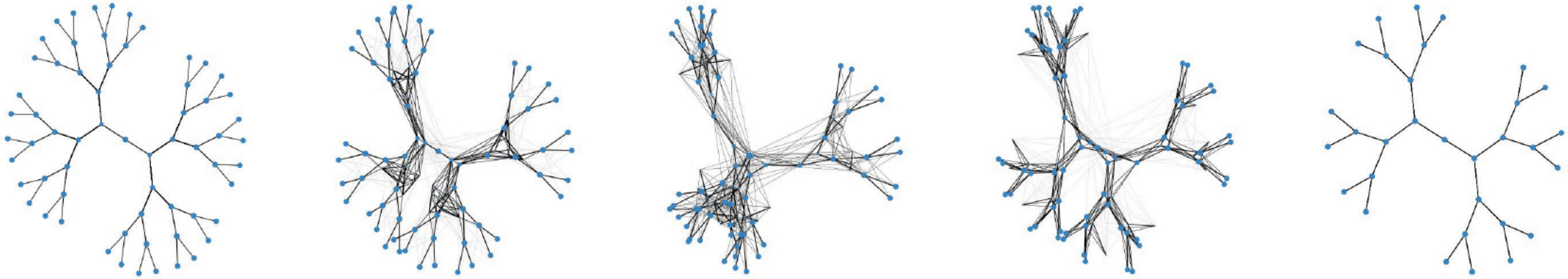


Geodesics via adjacency loss and spectral loss

Adjacency loss

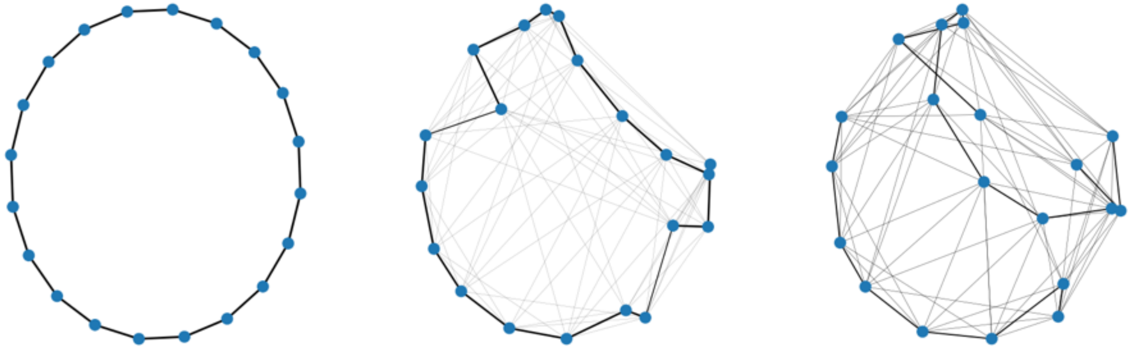


Spectral loss

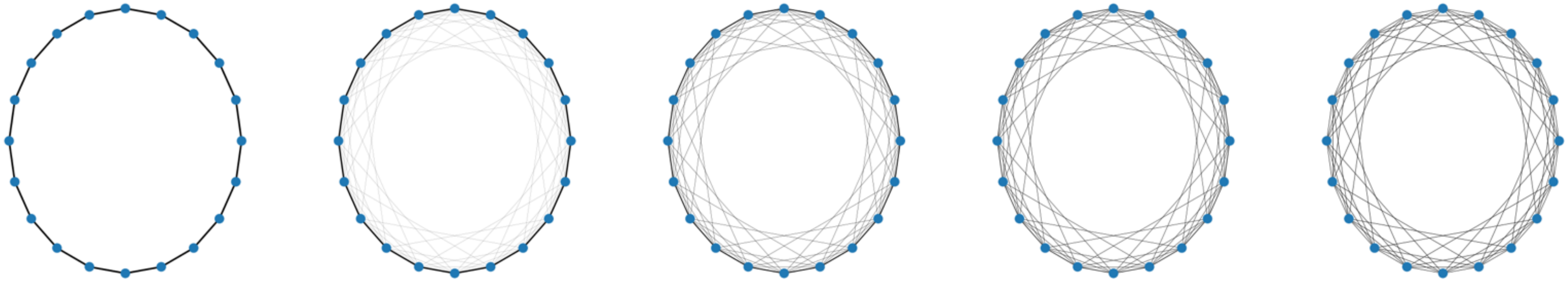


Geodesics via spectral loss for different choices of t

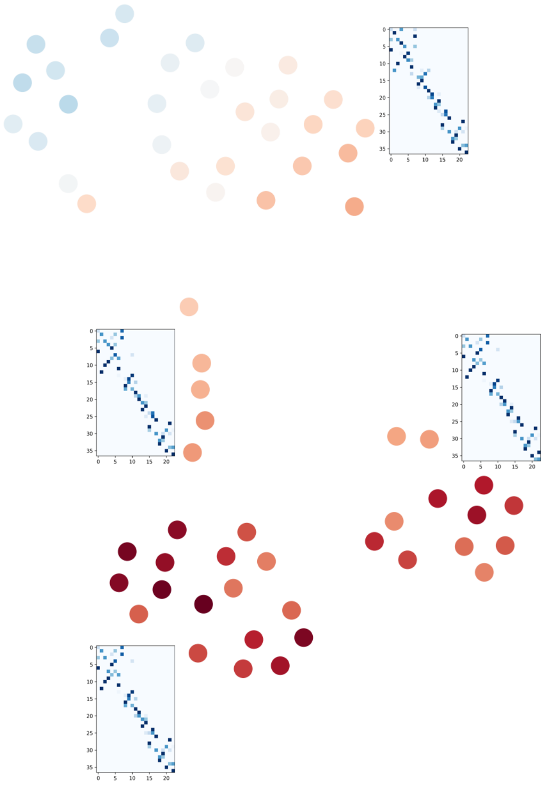
Spectral loss, $t = 2$

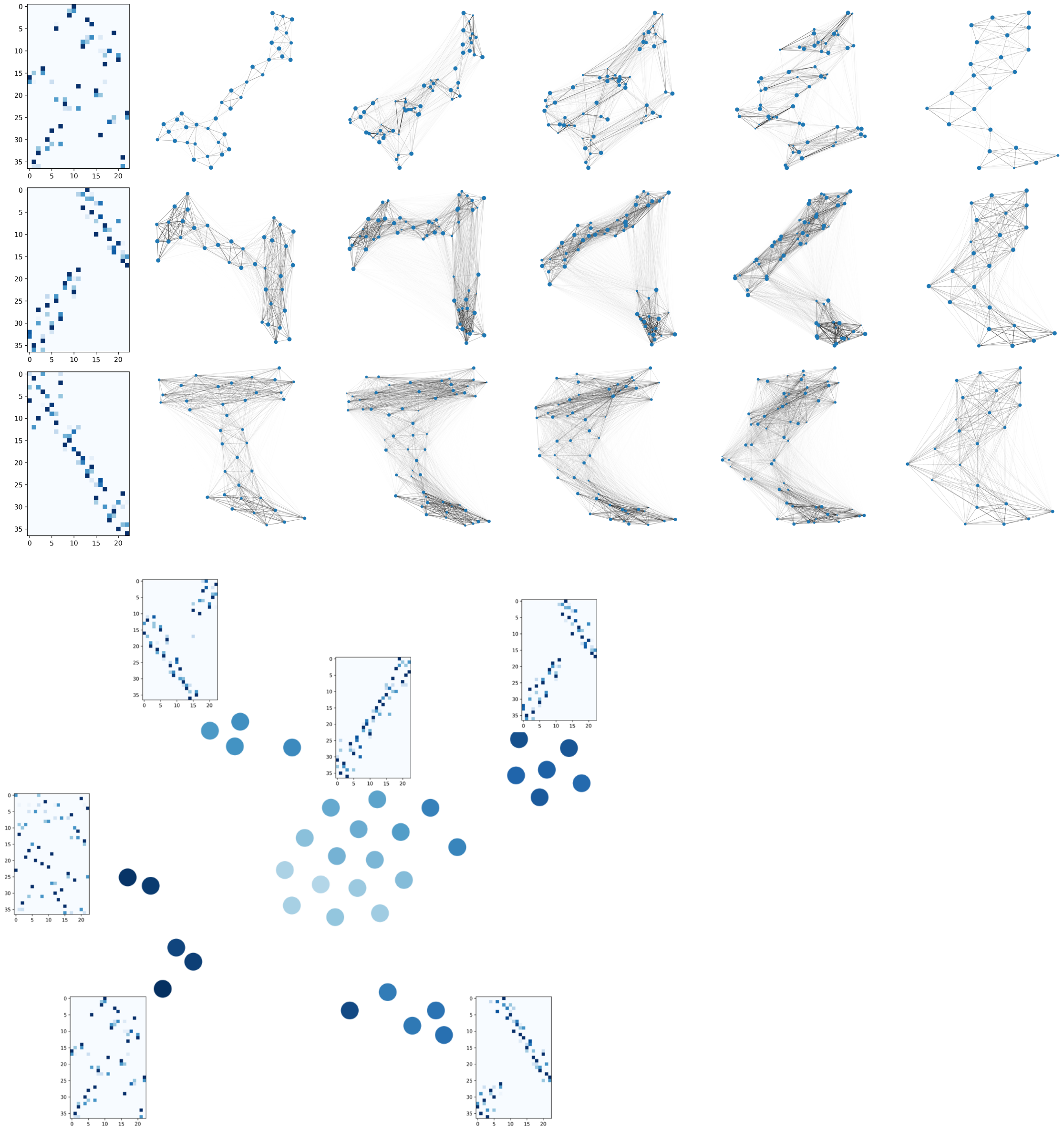


Spectral loss, $t = 10$

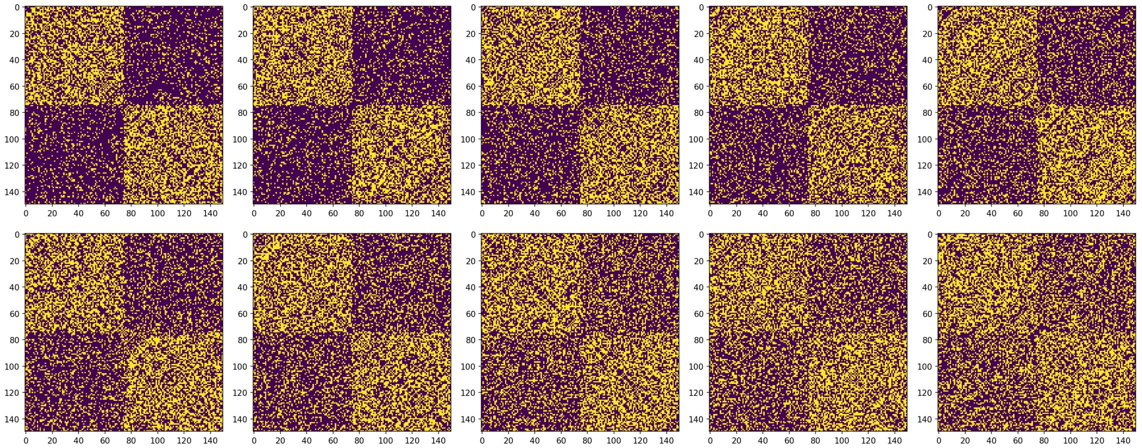


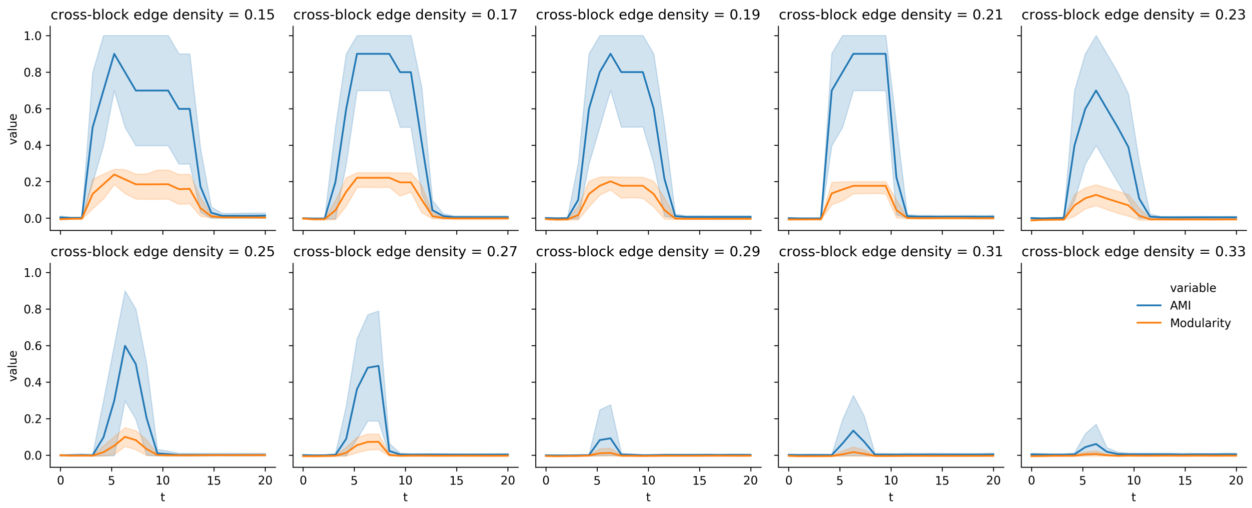
Clusters of scale parameters t possibly capture multiscale features





Heuristic for choosing the scale parameter t in graph partitioning





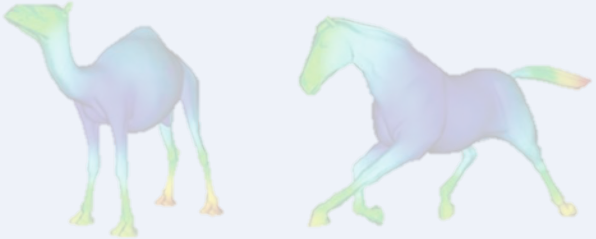
Graph partitioning results for real data

- [Xu et al 2019] produced a GW-based graph partitioning procedure using adjacency matrix representations and showed superior performance compared to various benchmarks
- Spectral loss provides improved scores and ~10x speedup in the “small” graph regime (~2000 nodes)
- Caveat: $O(n^3)$ cost of eigendecomposition, lack of sparsity is a bottleneck for spectral loss in the large graph regime

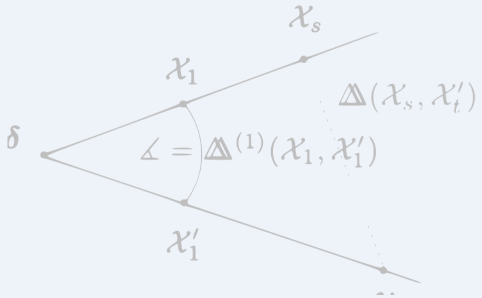
Dataset		Fluid	FastGreedy	Louvain	Infomap	GWL	SpecGWL
Wikipedia	sym, raw	—	0.382	0.377	0.332	0.312	**0.442***
	sym, noisy	—	0.341	0.329	0.329	0.285	**0.395**
	asym, raw	—	—	—	0.332	0.178	**0.376**
	asym, noisy	—	—	—	**0.329**	0.170	0.307
EU-email	sym, raw	—	0.312	0.447	0.374	0.451	**0.487**
	sym, noisy	—	0.251	0.382	0.379	0.404	**0.425**
	asym, raw	—	—	—	**0.443**	0.420	0.437
	asym, noisy	—	—	—	0.356	**0.422**	0.377
Amazon	raw	—	0.637	0.622	**0.940**	0.443*	0.692
	noisy	0.347	0.573	**0.584**	0.463	0.352	0.441
Village	raw	—	**0.881**	**0.881**	**0.881**	0.606*	0.801*
	noisy	—	0.778	**0.827**	0.190	0.560	0.758

1. Problem setup

2. Gromov-Wasserstein distance and Sturm's constructions

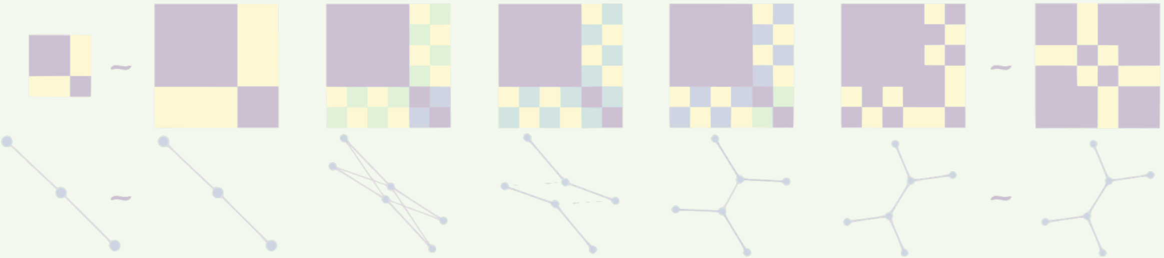


[Mémoli 2007]

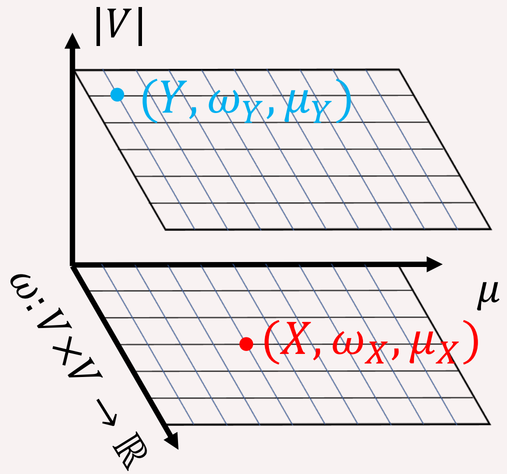


[Sturm 2012]

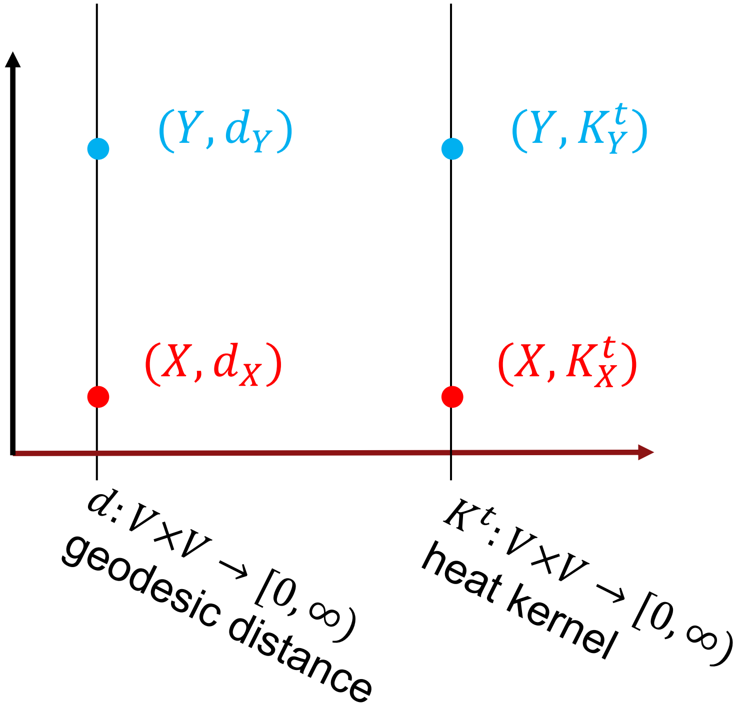
3. Statistical learning in the Riemannian framework



4. Future directions



Future directions



Questions and connections:

- What are other useful classes of functions $V \times V \to \mathbb{R}$? Currently known benefits of each:
  - Adjacency: sparse representations [Xu et al 2019]
  - Heat kernel: more global structure, faster on the order of graphs with a few thousand nodes
  - Distance: established asymptotics [Weitkamp et al 2020]
- Extensions: more learning tasks, comparing data across modalities
  - Application: neurobiological insights across populations
- Many statistical questions remain for the GW framework

Thank you!