Bridging Arbitrary and Tree Metrics via Differentiable Gromov Hyperbolicity

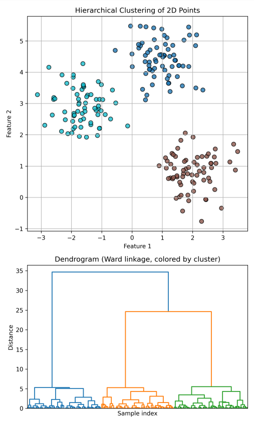
Pierre Houedry[1] Nicolas Courty[1] Florestan Martin-Baillon[2]
Laetitia Chapel[1] Titouan Vayer[3]

October 27, 2025

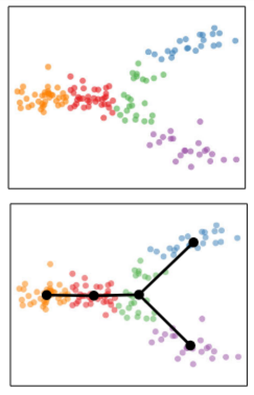
[1]IRISA, UMR 6074, CNRS, Université de Bretagne Sud
[2]IRMAR, UMR 6625, CNRS, Université de Rennes
[3]LIP, UMR 5668, CNRS, Univ. Lyon, ENS de Lyon, UCBL, CNRS, Inria

Introduction

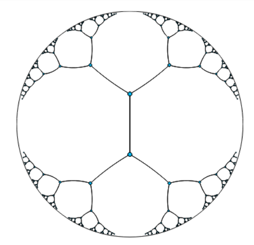
Motivation



Hierarchical Clustering

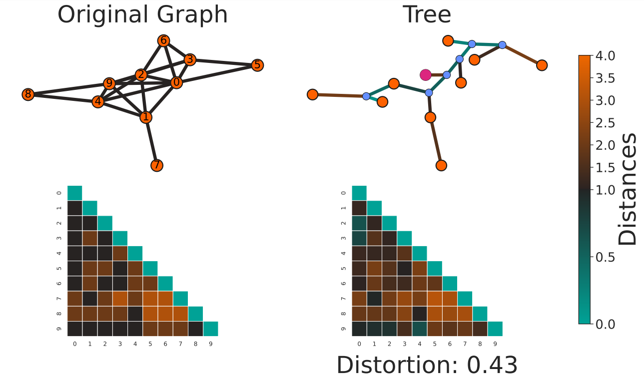


Single-cell trajectory inference (Street et al., 2018)



Tree Embedding (Nickel and Kiela, 2017)

Problem



A graph and a possible embedding of this graph into a tree.

Problem

Problem Statement

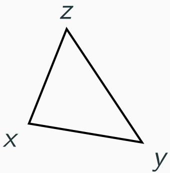
- *Given:* A finite metric space (X, d_X) (e.g. pairwise distances between n items)
- *Find:*
  1. A weighted tree T with vertex set $V(T)$
  2. An embedding $\Phi : X \hookrightarrow V(T)$
- *Objective:* Minimize the ℓ_∞ norm

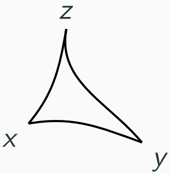
$$\max_{x,y \in X} |d_X(x, y) - d_T(\Phi(x), \Phi(y))|,$$

Background on δ-hyperbolicity

Geometric intuition

Triangles





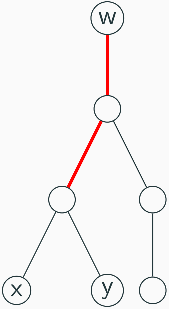
Geometry	Euclidean	Hyperbolic	Tree
Curvature	$K = 0$	$K < 0$	$K = -\infty$
δ-Hyperbolicity	$\delta = \infty$	$\delta > 0$	$\delta = 0$

Formal definition

Gromov Product (Gromov, 1987)

Let (X, d_X) be a metric space and let $x, y, w \in X$. The *Gromov product* of x and y with respect to the basepoint w is defined as

$$(x|y)_w = \frac{1}{2}\left(d_X(x, w) + d_X(y, w) - d_X(x, y)\right).$$



Formal definition

δ-hyperbolicity and Gromov hyperbolicity (Gromov, 1987)

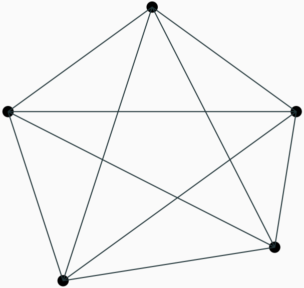
A metric space (X, d_X) is said to be δ-*hyperbolic* if there exists $\delta \geq 0$ such that for all $x, y, z, w \in X$, the Gromov product satisfies

$$(x|z)_w \geq \min\{(x|y)_w, (y|z)_w\} - \delta.$$

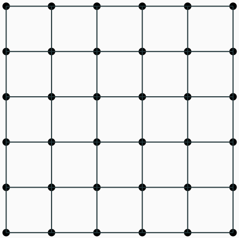
The *Gromov hyperbolicity*, denoted by δ_X, is the smallest value of δ that satisfies the above property. Consequently, every finite metric space (X, d_X) has a Gromov hyperbolicity equal to

$$\delta_X = \max_{x,y,z,w \in X} \left(\min\{(x|y)_w, (y|z)_w\} - (x|z)_w\right).$$

Some examples



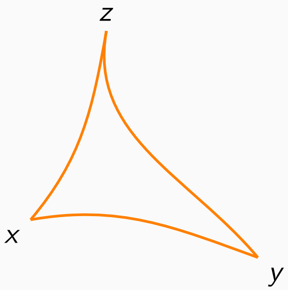
Complete graph K_5, $\delta_X = 0$



$6{\times}6$ grid graph, $\delta_X = 5$

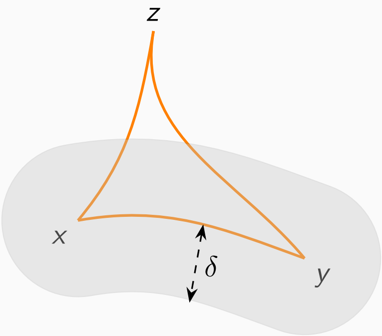
Two example graphs

Geometrical view of the definition



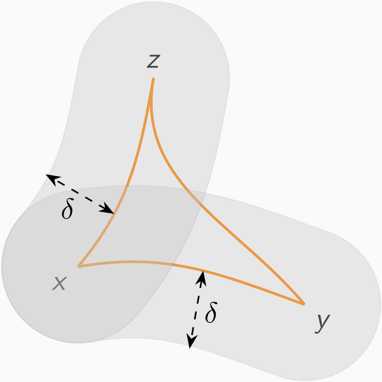
δ-thin triangle

Geometrical view of the definition



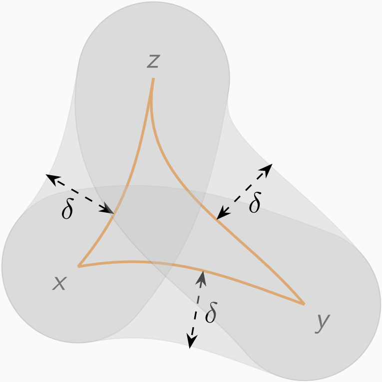
δ-thin triangle

Geometrical view of the definition



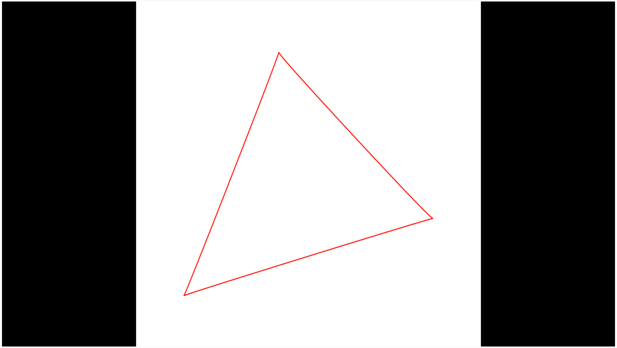
δ-thin triangle

Geometrical view of the definition



δ-thin triangle

Geometrical view of the definition



Tree Approximation of δ-Hyperbolic Spaces

Gromov's Theorem (Ghys et al., 1990, Ch. 2, §2, Thm. 12)

Let (X, d_X) be a finite δ-hyperbolic metric space over n points. For every $w \in X$, there exists a finite metric tree (T, d_T), and an embedding $\Phi_w : X \to T$ such that:

Tree Approximation of δ-Hyperbolic Spaces

Gromov's Theorem (Ghys et al., 1990, Ch. 2, §2, Thm. 12)

Let (X, d_X) be a finite δ-hyperbolic metric space over n points. For every $w \in X$, there exists a finite metric tree (T, d_T), and an embedding $\Phi_w : X \to T$ such that:

1. Distance to the basepoint is preserved:

$$d_T(\Phi_w(x), \Phi_w(w)) = d_X(x, w).$$

Tree Approximation of δ-Hyperbolic Spaces

Gromov's Theorem (Ghys et al., 1990, Ch. 2, §2, Thm. 12)

Let (X, d_X) be a finite δ-hyperbolic metric space over n points. For every $w \in X$, there exists a finite metric tree (T, d_T), and an embedding $\Phi_w : X \to T$ such that:

1. Distance to the basepoint is preserved:

$$d_T(\Phi_w(x), \Phi_w(w)) = d_X(x, w).$$

2. For every pair of points $x, y \in X$, the tree-embedding Φ_w does not stretch distances and only contracts them by at most an additive factor of $2\delta \log_2(n-2)$.

$$d_X(x, y) - 2\delta \log_2(n-2) \le d_T(\Phi_w(x), \Phi_w(y)) \le d_X(x, y).$$

Tree Approximation of δ-Hyperbolic Spaces

Gromov's Theorem (Ghys et al., 1990, Ch. 2, §2, Thm. 12)

Let (X, d_X) be a finite δ-hyperbolic metric space over n points. For every $w \in X$, there exists a finite metric tree (T, d_T), and an embedding $\Phi_w : X \to T$ such that:

1. Distance to the basepoint is preserved:

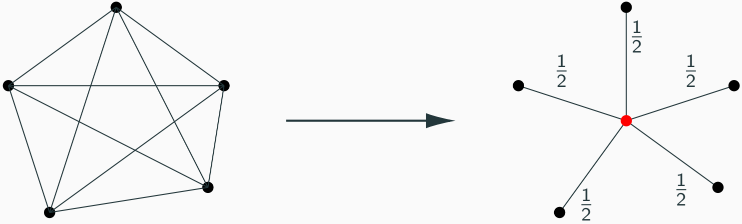
$$d_T(\Phi_w(x), \Phi_w(w)) = d_X(x, w).$$

2. For every pair of points $x, y \in X$, the tree-embedding Φ_w does not stretch distances and only contracts them by at most an additive factor of $2\delta \log_2(n-2)$.

$$d_X(x, y) - 2\delta \log_2(n-2) \leq d_T(\Phi_w(x), \Phi_w(y)) \leq d_X(x, y).$$

X can be embedded into a tree iff $\delta_X = 0$!!

Example



Computational Bottlenecks

- Gromov hyperbolicity requires checking **all point quadruples** $\Rightarrow$ naive cost is $\boldsymbol{O(n^4)}$.
- (Fournier et al., 2015) reduce base-point hyperbolicity to a $(\max, \min)$ matrix product $\Rightarrow$ Overall hyperbolicity in $\boldsymbol{O(n^{3.69})}$

Structural Pruning via Far-apart Pairs

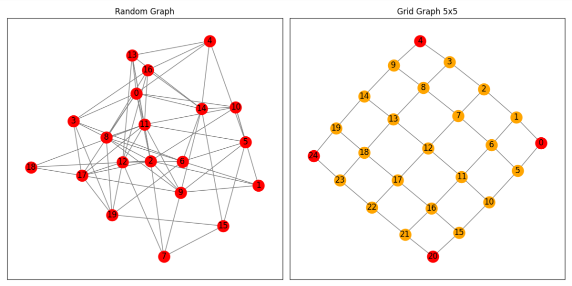


Illustration of far-apart pairs pruning

- Use far-apart vertex pairs to witness values.
- Prune by exploring only key quadruples.
- Delivers major speedups (Cohen et al., 2015; Coudert et al., 2022).

Restatement of the problem

Problem Statement

- *Given:* A finite metric space (X, d_X) (e.g. pairwise distances between n items)
- *Find:*
  1. A weighted tree T with vertex set $V(T)$
  2. An embedding $\Phi : X \hookrightarrow V(T)$
- *Objective:* Minimize the ℓ_∞ norm

$$\max_{x,y \in X} |d_X(x, y) - d_T(\Phi(x), \Phi(y))|,$$

Existing Methods Comparison

Method	HCC[1]	Gromov[2]	TR[3]	NJ[4]	LT[5]
Complexity	$O(n^2 \log n)$	$O(n^2)$	$O(n^2)$	$O(n^3)$	$O(\|E\|)$
Differentiable	✗	✗	✗	✗	✗
Requires a root	✓	✓	✗	✗	✓
Deterministic	✓	✓	✗	✓	✓
Non graph metrics	✓	✓	✓	✓	✗

Comparison of existing methods.

[1] (Yim and Gilbert, 2023)
[2] (Gromov, 1987)
[3] (Sonthalia and Gilbert, 2020)
[4] (Saitou and Nei, 1987)
[5] (Chepoi et al., 2012)

Our Approach: Differentiable Hyperbolicity

- We introduce a **smooth, differentiable surrogate** of hyperbolicity.
- Enables **gradient-based optimization**.
- We propose a **batched approximation scheme**:
  - **Scalable** computation
  - **Independent of graph size**

DeltaZero

Our approach

The set of n-points metrics:

$$\mathcal{D}_n := \left\{ \boldsymbol{D} \in \mathbb{R}_+^{n \times n} : \boldsymbol{D} = \boldsymbol{D}^\top,\ D_{ii} = 0,\ D_{ij} \le D_{ik} + D_{kj} \right\}$$

is a closed, convex, polyhedral cone. Each metric d_X is represented by $\boldsymbol{D}_X \in \mathcal{D}_n$.

Objective: Find a metric $\mathbf{D}_{X'}$ that is:

- **Close** to $\mathbf{D}_X$ (via ℓ_∞ distortion),
- Does not stretch $\mathbf{D}_X$,
- **Tree-like** (via low hyperbolicity $\delta_{X'}$).

Our approach

The set of n-points metrics:

$$\mathcal{D}_n := \left\{ \boldsymbol{D} \in \mathbb{R}_+^{n\times n} : \boldsymbol{D} = \boldsymbol{D}^\top,\ D_{ii} = 0,\ D_{ij} \leq D_{ik} + D_{kj} \right\}$$

is a closed, convex, polyhedral cone. Each metric d_X is represented by $\boldsymbol{D}_X \in \mathcal{D}_n$.

Objective: Find a metric $\mathbf{D}_{X'}$ that is:

- **Close** to $\mathbf{D}_X$ (via ℓ_∞ distortion),
- Does not stretch $\mathbf{D}_X$,
- **Tree-like** (via low hyperbolicity $\delta_{X'}$).

Lagrangian type problem

$$\min_{\substack{\mathbf{D}_{X'} \in \mathcal{D}_n \\ \mathbf{D}_{X'} \leq \mathbf{D}_X}} \mathcal{L}_X(\mathbf{D}_{X'}, \mu) := \mu \|\mathbf{D}_X - \mathbf{D}_{X'}\|_\infty + \delta_{X'}.$$

Ideal world...

Minimizer + Gromov Embedding	**Gromov Embedding**
$\mathbf{D}_{X^*} = \underset{\substack{\mathbf{D}_{X'} \in \mathcal{D}_n \\ \mathbf{D}_{X'} \leq \mathbf{D}_X}}{\arg\min} \ \mathcal{L}_X(\mathbf{D}_{X'}, \mu),$	
$(T^*, d_{T^*}) = \Phi_w(\mathbf{D}_X^*)\,.$	$(T, d_T) = \Phi_w(\mathbf{D}_X)$
$d_{T^*}(\Phi_w(x), \Phi_w(y)) \leq d_X(x, y)$	does not stretch distances
$d_X(x, y) - C_\mu \leq d_{T^*}(\Phi_w(x), \Phi_w(y))$	only contracts them by at most $2\delta \log_2(n-2)$

In particular if $\mu \geq \frac{1}{2\log_2(n-2)}$ then we have $C_\mu < 2\delta \log_2(n-2)$.

How to make the objective function differentiable?

Objective: $\min_{\substack{\mathbf{D}_{X'} \in \mathcal{D}_n \\ \mathbf{D}_{X'} \leq \mathbf{D}_X}} \mathcal{L}_X(\mathbf{D}_{X'}, \mu) := \mu \|\mathbf{D}_X - \mathbf{D}_{X'}\|_\infty + \delta_{X'}.$

How to make the objective function differentiable?

Objective: $\min_{\substack{\mathbf{D}_{X'} \in \mathcal{D}_n \\ \mathbf{D}_{X'} \leq \mathbf{D}_X}} \mathcal{L}_X(\mathbf{D}_{X'}, \mu) := \mu \|\mathbf{D}_X - \mathbf{D}_{X'}\|_\infty + \delta_{X'}.$

Non-differentiable		**Differentiable**
$\|\mathbf{D}_X - \mathbf{D}_{X'}\|_\infty$	$\rightsquigarrow$	$\|\mathbf{D}_X - \mathbf{D}_{X'}\|_2^2$

How to make the objective function differentiable?

Objective: $\min_{\substack{\mathbf{D}_{X'} \in \mathcal{D}_n \\ \mathbf{D}_{X'} \leq \mathbf{D}_X}} \mathcal{L}_X(\mathbf{D}_{X'}, \mu) := \mu \|\mathbf{D}_X - \mathbf{D}_{X'}\|_\infty + \delta_{X'}$.

Non-differentiable		**Differentiable**
$\|\mathbf{D}_X - \mathbf{D}_{X'}\|_\infty$	$\rightsquigarrow$	$\|\mathbf{D}_X - \mathbf{D}_{X'}\|_2^2$
$\mathbf{D}_{X'} \leq \mathbf{D}_X$	$\rightsquigarrow$	✗

How to make the objective function differentiable?

Objective: $\min_{\mathbf{D}_{X'} \in \mathcal{D}_n} \mathcal{L}_X(\mathbf{D}_{X'}, \mu) := \mu \|\mathbf{D}_X - \mathbf{D}_{X'}\|_\infty + \delta_{X'}$.

Non-differentiable		**Differentiable**
$\|\mathbf{D}_X - \mathbf{D}_{X'}\|_\infty$	$\rightsquigarrow$	$\|\mathbf{D}_X - \mathbf{D}_{X'}\|_2^2$

How to make the objective function differentiable?

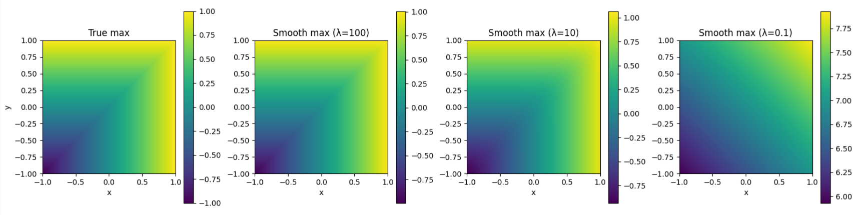
Objective: $\min_{\mathbf{D}_{X'} \in \mathcal{D}_n} \mathcal{L}_X(\mathbf{D}_{X'}, \mu) := \mu \|\mathbf{D}_X - \mathbf{D}_{X'}\|_\infty + \delta_{X'}$.

Non-differentiable		**Differentiable**
$\|\mathbf{D}_X - \mathbf{D}_{X'}\|_\infty$	$\rightsquigarrow$	$\|\mathbf{D}_X - \mathbf{D}_{X'}\|_2^2$
$\delta_{X'}$	$\rightsquigarrow$	?

Smooth min-max

We introduce a differentiable surrogate using log-sum-exp:

$$\mathrm{LSE}_\lambda(\boldsymbol{x}) = \frac{1}{\lambda} \log \left(\sum_i e^{\lambda x_i} \right)$$



Smooth max visualization

Smooth Hyperbolicity

True $\delta_{\mathbf{X}}$ $\quad \max\left(\min\left\{(x|y)_w, (y|z)_w\right\} - (x|z)_w\right)$

Smooth $\delta_{\mathbf{X}}^{(\lambda)}$ $\quad \mathrm{LSE}_\lambda \left\{\mathrm{LSE}_{-\lambda}((x|y)_w, (y|z)_w) - (x|z)_w\right\}.$

Smooth Hyperbolicity

True $\delta_{\mathbf{X}}$	$\max(\min\{(x\|y)_w, (y\|z)_w\} - (x\|z)_w)$
Smooth $\delta_{\mathbf{X}}^{(\lambda)}$	$\mathrm{LSE}_\lambda\{\mathrm{LSE}_{-\lambda}((x\|y)_w, (y\|z)_w) - (x\|z)_w\}$.

Bounds

$$\delta_X - \frac{\log 2}{\lambda} \leq \delta_X^{(\lambda)} \leq \delta_X + \frac{4 \log n}{\lambda}.$$

$\delta_X^{(\lambda)}$ still requires $O(n^4)$ operations !

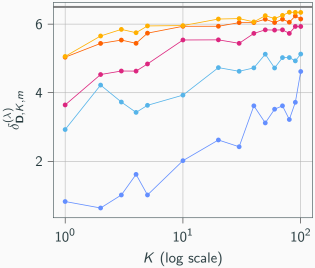
Batched Approximation for Scalability

- We sample K subsets $X_m^1, \ldots, X_m^K \subset X$, each of size m.
- Compute local estimates $\delta^{(\lambda)}_{X_m^i}$ and aggregate:

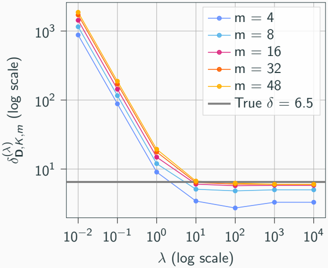
$$\delta^{(\lambda)}_{X,K,m} = \mathrm{LSE}_\lambda\left(\delta^{(\lambda)}_{X_m^1}, \ldots, \delta^{(\lambda)}_{X_m^K}\right)$$

- Reduces complexity to $O(K \cdot m^4)$.
- Caveat: Rare, high-curvature configurations may be missed in random batches $\Rightarrow$ potential underestimation.

In practice



Evolution of the mean $\delta^{(\lambda)}_{\boldsymbol{D},K,m}$ as a function of the number of batches K, with $\lambda = 1000$.



Estimation of $\delta^{(\lambda)}_{\boldsymbol{D},K,m}$ as a function of λ, with $K = 50$.

Mean of $\delta^{(\lambda)}_{\boldsymbol{D},K,m}$ over 5 runs, computed on the $\mathrm{CS\text{-}P{\scriptsize H}D}$ dataset for different batch sizes m.

Final objective

Non-differentiable		Differentiable	Cost
$\|\mathbf{D}_X - \mathbf{D}_{X'}\|_\infty$	$\rightsquigarrow$	$\|\mathbf{D}_X - \mathbf{D}_{X'}\|_2^2$	$O(n^2)$
$\delta_{X'}$	$\rightsquigarrow$	$\delta^{(\lambda)}_{X,K,m}$	$O(Km^4)$
$\mathcal{L}_X(\mathbf{D}_{X'}, \mu)$	$\rightsquigarrow$	$\mu\|\boldsymbol{D}_X - \mathbf{D}_{X'}\|_2^2 + \delta^{(\lambda)}_{X,K,m}$	$O(Km^4 + n^2)$

Final Algorithm

DeltaZero

Require: A metric space (X, d_X), root $w \in X$, learning rate ϵ, batches K, size m, scale λ, regularization μ, steps T

Initialize $\boldsymbol{D}_0 = \boldsymbol{D}_X$
for $t = 0$ to $T - 1$ **do**
  Sample K batches of m points
  Compute $\delta^{(\lambda)}_{\boldsymbol{D}_t, K, m}$
  $\boldsymbol{G}_t = \nabla L_X(\boldsymbol{D}_t)$
  $\tilde{\boldsymbol{D}}_{t+1} = \textsc{AdamStep}(\boldsymbol{D}_t, \boldsymbol{G}_t, \epsilon)$
  $\boldsymbol{D}_{t+1} = \textsc{FloydWarshall}(\tilde{\boldsymbol{D}}_{t+1}) \quad \Rightarrow O(n^3)$
end for
return $\textsc{GromovEmbed}(\boldsymbol{D}_T, w)$

$$\text{Total cost of } \mathbf{O}\left(\mathbf{T}(\mathbf{K}\mathbf{m}^4 + \mathbf{n}^3)\right)$$

Experiments: Setup

Setup: 100 random roots for pivot-based methods, mean and std. of distortion reported. DeltaZero hyperparameters selected via grid search:

- $\epsilon \in \{0.1, 0.01, 0.001\}$,
- $\mu \in \{0.1, 0.01, 1.0\}$,
- $\lambda \in \{0.01, 0.1, 1.0, 10.0\}$,
- $K \in \{100, 500, 1000, 3000, 5000\}$,
- $T = 1000$,
- $m = 32$.

Experiments: Unweighted Graphs

Table 1: ℓ_∞ error on unweighted graphs (lower is better). Best result in bold, second-best underlined.

Dataset	C-ELEGAN	CS PhD	CORA	AIRPORT	WIKI
n	452	1025	2485	3158	2357
Diameter	7	28	19	12	9
NJ	<u>2.97</u>	16.81	13.42	4.18	6.32
TR	5.90 ± 0.72	21.01 ± 3.34	16.86 ± 2.11	10.00 ± 1.02	9.97 ± 0.93
HCC	4.31 ± 0.46	23.35 ± 2.07	12.28 ± 0.96	7.71 ± 0.72	7.20 ± 0.60
LT	5.07 ± 0.25	25.48 ± 0.60	<u>7.76</u> ± 0.54	<u>2.97</u> ± 0.26	<u>4.08</u> ± 0.27
Gromov	3.33 ± 0.45	<u>13.28</u> ± 0.61	9.34 ± 0.53	4.08 ± 0.27	5.54 ± 0.49
DeltaZero	**1.87** ± 0.08	**10.31** ± 0.62	**7.59** ± 0.38	**2.79** ± 0.15	**3.56** ± 0.20
Improvement (%)	43.8%	22.3%	2.3%	6.0%	12.7%

Experiments: General Metrics

Table 2: ℓ_∞ error on general metrics (lower is better). Best result in bold, second-best underlined.

Dataset	ZEISEL	IBD
n	3005	396
Diameter	0.87	0.99
NJ	0.51	<u>0.90</u>
TR	0.66 ± 0.10	1.60 ± 0.22
HCC	0.53 ± 0.07	1.25 ± 0.11
LT	–	–
Gromov	<u>0.43</u> $\pm$ 0.02	1.01 ± 0.04
DeltaZero	**0.24** $\pm$ 0.00	**0.70** $\pm$ 0.03
Improvement (%)	44.1%	22.2%

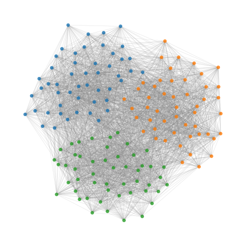
Toy Application: Hierarchical Clustering

Setup:

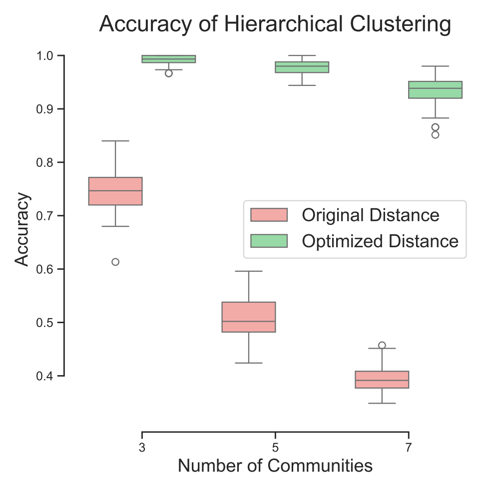
- Stochastic Block Model with $N = 3, 5, 7$ communities, 50 nodes each
- $p_{in} = 0.6$, $p_{out} = 0.2$

Procedure:

1. Compute shortest-path distances and optimize them using DeltaZero ($\mu = 1$, $\lambda = 100$).
2. Apply Ward's linkage on both original and optimized distance matrices.



Toy Application: Hierarchical Clustering



Accuracy of hierarchical clustering with varying number of communities (3, 5, 7). Each setting is repeated 30 times.

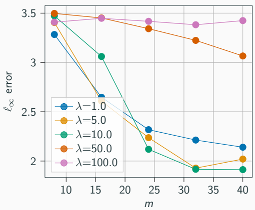
Future work

- Further theoretical investigations to:
  - Reduce the computational cost of $\delta^{(\lambda)}_{\boldsymbol{D},K,m}$.
  - Reduce the number of hyperparameters.
- Explore applications, such as:
  - Hierarchical clustering on real-world datasets,
  - Phylogenetic tree reconstruction,
  - Single-cell trajectory inference,
  - Hyperbolicity-aware learning ?

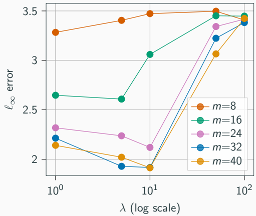
Thanks for listening!

Any questions?

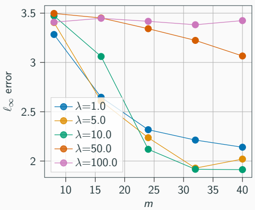
Sensitivity Analysis



Effect of the distance regularization coefficient μ on ℓ_∞ distortion.



Impact of the log-sum-exp scale λ on ℓ_∞ error across various batch sizes m.



Impact of the batch size m on ℓ_∞ error for multiple λ values

Sensitivity analysis of optimization hyperparameters on the C-ELEGAN dataset. In each plot, non-varied hyperparameters are set to their optimal values from a prior grid search. Results are averaged over 5 runs, and distortion values averaged over 100 root samples per run.

References

Chepoi, V., Dragan, F. F., Estellon, B., Habib, M., Vaxès, Y., and Xiang, Y. (2012). Additive Spanners and Distance and Routing Labeling Schemes for Hyperbolic Graphs. *Algorithmica*, 62(3-4):713–732.

Cohen, N., Coudert, D., and Lancin, A. (2015). On computing the gromov hyperbolicity. *ACM J. Exp. Algorithmics*, 20.

Coudert, D., Nusser, A., and Viennot, L. (2022). Enumeration of far-apart pairs by decreasing distance for faster hyperbolicity computation. *ACM J. Exp. Algorithmics*, 27.

Fournier, H., Ismail, A., and Vigneron, A. (2015). Computing the Gromov hyperbolicity of a discrete metric space. *Information Processing Letters*, 115(6):576–579.

Ghys, E., De La Harpe, P., Oesterlé, J., and Weinstein, A. (1990).