The Choice and Impact of Metric in Persistent Homology

Ethan Scheelk
Macelseter College
1600 Grand Avenue, Saint Paul, MN 55105
escheelk@macalester.edu

Persistent homology has an implicit assumption of a metric. We analyze metric choices and how this impacts the analysis of data sets by comparing the bottleneck distances between the applications of these difference metrics. We do this through the use of real and contrived point cloud data sources and we analyze these with Python packages `RIPSER` and `Persim`. We find that the bottleneck distance between distributions is a poor comparison between different metrics, yielding high distances even when the topological results are the same.

Introduction

Persistence Homology and Topological Data Analysis are as new math as math can get. These techniques analyze the shape of data by filtering features through an increasing epsilon radius, connecting points in a cloud to make a series of simplicial complexes. These complexes give rise to interesting topological properties, where Persistence Homology hopes to glean some sort of meaning. Positions measured within physical space have a regular notion of distance. But how do you measure the distance between the voting patterns of elected representatives, or the genome of a deadly virus? Data interpreted in this way do not have a traditional concept of distance. So how does the choice of metric impact Persistent Homology?

As [**Carlsson(2009)**] claims, data is being created at an unfathomable rate. This data is also frequently of a greatly higher dimension than is possible to graphically represent. Persistent Homology is a way to do this. The data, as some dimensional point cloud, can be thought to represent some geometric object with some amount of offset from noise.

Persistent homology is elected to try to solve the problem of interpretability of data sets that are large and of high dimension.

Persistent Homology

Modern, powerful computing has given modern topolgists expand their capabilities interpretive abilities into a field known as Persistent Homology. Essentially, persistent homology is the viewing of the same point cloud data set at differing spatial resolutions. This 'resolution' is modified through the use of an increasing distance threshold ϵ.

The space must first be represented as a simplicial complex, where features (simplices) are points, lines, filled triangles, filled tetrahedrons, and so on. Specifically, the most implemented methods use Vietoris-Rips Complexes, where an n-complex is made if all the points are pairwise adjacent. Adjacency is determined by checking if the distance between points is less then ϵ. Thus, increasing ϵ induces the creation of more simplices and is the filtering of the space to the next ϵ-step.

The metric provided will dictate how readily two points will connect.

Metrics

A metric is a function that gives a distance and it must follow the follow rules:

1. $d(\vec{x}, \vec{x}) = 0$
2. If $\vec{x} \neq \vec{y}$, then $d(\vec{x}, \vec{y}) > 0$ **(Positivity)**
3. $d(\vec{x}, \vec{y}) = d(\vec{y}, \vec{x})$ **(Symmetry)**
4. $d(\vec{x}, \vec{z}) \leq d(\vec{x}, \vec{y}) + d(\vec{y}, \vec{z})$ **(Triangle Inequality)**.

Metrics can be distances between points, vectors, or distributions. In this case, all the metrics we consider are metrics between points. In exception is the bottleneck distance, which we use to find the distance between persistences.

A metric is necessary for persistent homology in order to justify the inclusion of two points into a simplex.

Euclidean When topology is concerned with open sets and 'balls', the Euclidean metric is the typically assumed metric. This metric is the usual distance function used in most of mathematics, and reality. Otherwise known as *the* distance function (mathematicians know better than use that definite article), it is also called the Pythagorean. In words, it is the square root of the sum of the squared differences in each dimension; as an image, it is a circle in $\mathbb{R}^2$ (Figure 1); and as an equation,

$$d(\vec{p}, \vec{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}. \tag{1}$$

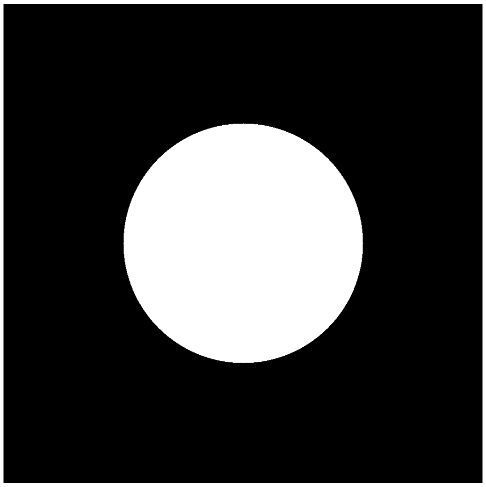


Figure 1 All the points a distance of one or less from the origin, the unit ball, given the Euclidean metric, in $[-2, 2] \times [-2, 2]$.

You can also pretend to be looking at a sphere from any angle and Figure 1 is also an accurate depiction of how the Euclidean metric looks in $\mathbb{R}^3$. But beyond this, it is hard to illustrate spheres in higher dimensions.

Euclidean distance acted as the prototype and originator of the properties of metric spaces.

City Block Though you could theoretically calculate the distance from one point in a city to another via the euclidean distance, given that you are likely not walking through buildings, it would be more accurate for you to calculate your distance as the number of square blocks you walk in one direction in addition to the number of blocks you walked in the other. Though, perhaps you wish taxi drivers calculated this distance with the euclidean metric. Hence, it is also called the Taxicab metric as well as the Manhattan metric.

This is the city block metric. Distance is the sum of the absolute differences between all coordinates for two points. More formally,

$$d(\vec{p}, \vec{q}) = \sum_{i=1}^{n} |p_i - q_i|. \tag{2}$$

America's uniform, new, and car-centric architecture inspired this metric, since it would be difficult to come up with a Paris metric. Observe in Figure 2 that the metric forms a diamond. This diamond is also a circle, where the circumference is $8r$, so $\pi = 4$.

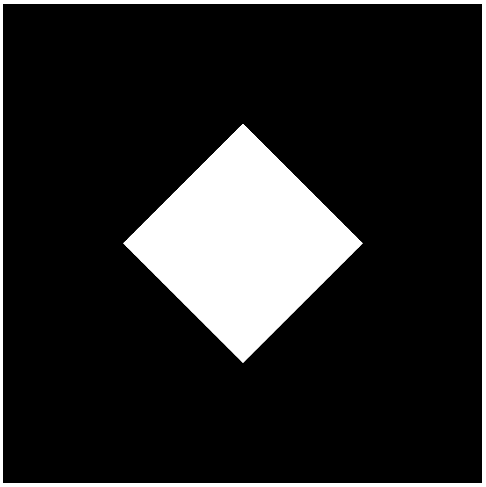


Figure 2 The unit ball in the City Block metric in $[-2, 2] \times [-2, 2]$.

Chebyshev If you've ever considered or desired measuring distance by the number of moves it would require the King in chess to move from one square on a chessboard to another square, then this is the perfect metric for you. Or perhaps you have an interest in warehouse crane logistics, where x and y axes can move independently.

In the Chebyshev metric, diagonal moves have the same cost as a horizontal or vertical move. Hence, the distance from the origin to the point $(1, 1)$ is 1. This is represented in Figure 3, where the metric forms a square of radius one centered at the origin. In higher dimensions, this metric is a cube (or hypercube) with each face perpendicular to its coordinate axis.

The Chebyshev metric is represented as

$$[h] d(\vec{p}, \vec{q}) = \max_{i}^{n} |p_i - q_i|. \tag{3}$$

In effect, considering all axes, Chebyshev distance is the farthest distance you would have to walk along any axis.

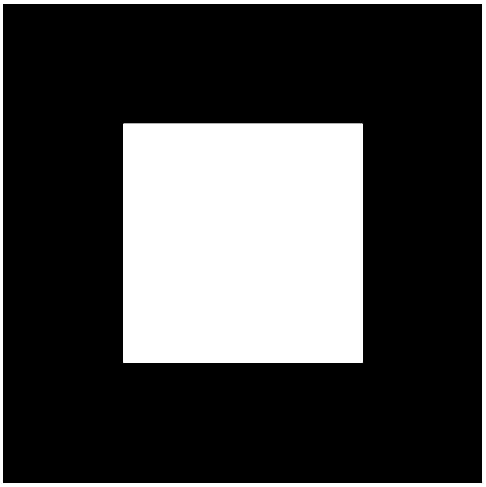


Figure 3 The unit ball, given the Chebyshev metric, in $[-2, 2] \times [-2, 2]$.

Minkowski / P-Norm The three discussed metrics have something in common, and that is the absolute difference. In reality, they come from the same set of distance metrics called the Minkowski metrics, or otherwise known as P-Norms.

The Minkowski metric, for a choice of p, is as follows:

$$d(\vec{p}, \vec{q}) = \left(\sum_{i=1}^{n} |p_i - q_i|^p \right)^{\frac{1}{p}}. \tag{4}$$

It follows that equation 2, the city block metric, is the result of setting $p = 1$ and equation 1, the euclidean metric, when $p = 2$. What is not immediately obvious is that $\lim_{p \to \infty} \left(\sum_{i=1}^{n} |p_i - q_i|^p \right)^{\frac{1}{p}} = \max_i^n |p_i - q_i|$ is Equation 3, the Chebyshev metric.

In order to create a visual intuition for this, consider Figure 4. As p increases, the extent reaches closer and closer to the corners. At $p = 5$, the shape is practically a squircle [**Parker(2021)**].

This visualization uses p-values of 10,000, 5, 3, 2.5, 2, 1.5, 1, 0.75, and 0.50.

With $p < 1$, it's no longer a metric, but a quasi-metric, since the triangle inequality is broken, but that is not relevant right now. The relationship here is that the smaller the value of p, the farther everything else is.

Bray-Curtis The Bray-Curtis Metric, also know as the Bray-Curtis Dissimilarity, as well as the Sorensen metric is frequently used as normalization in ecology [**Teknomo(2011)**]. If all coordinates are positive, the distance is between zero and one. The function is as follows:

$$d(\vec{p}, \vec{q}) = \frac{\sum_{i=1}^{n} |p_i - q_i|}{\sum_{i=1}^{n} |p_i + q_i|}. \tag{5}$$

The function is undefined for two zero coordinates.

Hamming Hamming distance is a measure of difference between strings and binary numbers and can be extended to a measure of how many components of a vector differ from each other.

So, comparing two positions, the hamming distance is increased for each component that differs. The hamming metric may lead to very high distances due to the

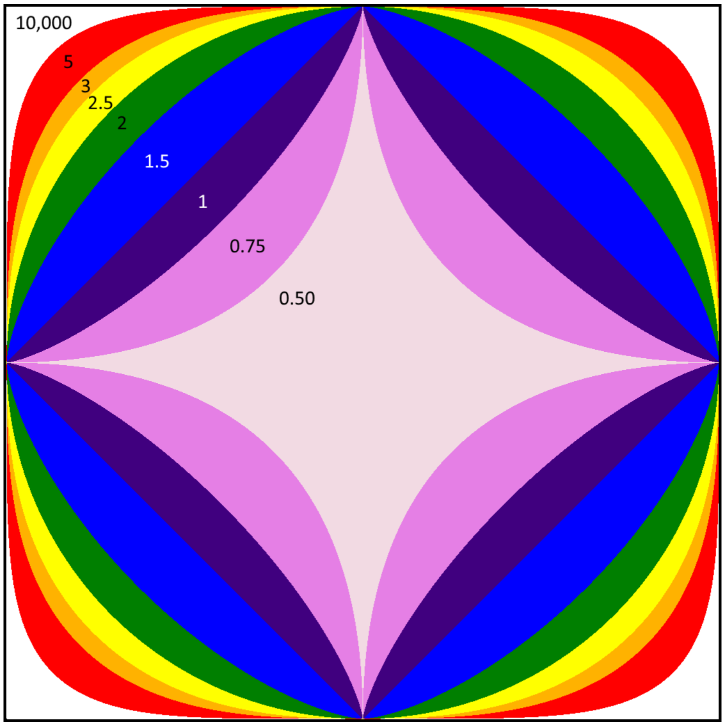


Figure 4 Labelled unit balls given Minkowski metric with varying p. Each section has its p value labeled. Graphed in $[-1,1] \times [-1,1]$.

imprecise and dubious nature of floating point number equality.

Methods

For our persistent homology, we primarily used RIPSER implemented within Python 3 [**Tralie et al.(2018)**].

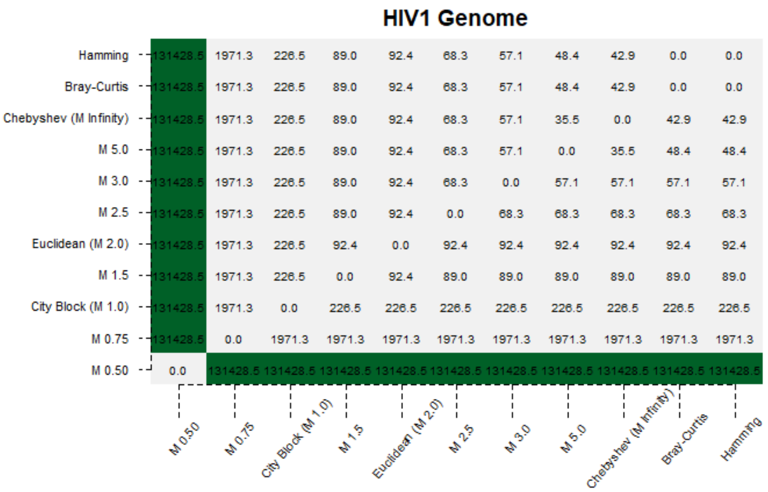
Using the RIPSER package, we are able to read in either point clouds or precompiled distance matrices and create the persistence diagrams, where points are saved as a birth time and a death time. Then we can use the Persim package to calculate the Bottleneck distance. The Bottleneck compares two distributions of points and returns the maximum distance between matchings.

We then create a distance matrix, looping over each persistence. We use R and RStudio to visualize this data in a heatmap.

Data Sets and Analysis

We utilize four data sources of varying types. We wish to use point clouds from real sources as well as from a contrived example with expected results. Since distance between coordinates is frequently (but not always) nonsensical for these sources, the application of a variety metrics will give us some clue as to how the choice of metric impacts topological data analysis. Our point cloud choices are inspired primarily by [**Otter et al.(2015)**].

HIV Genome Inspired by [**Otter et al.(2015)**], we analyze the difference that metric makes on sequences of the HIV1 Genome. This is inspired by the analysis performed in [**Chan et al.(2013)**] and the data is sourced from [**Los Alamos National Laboratory(2021)**].



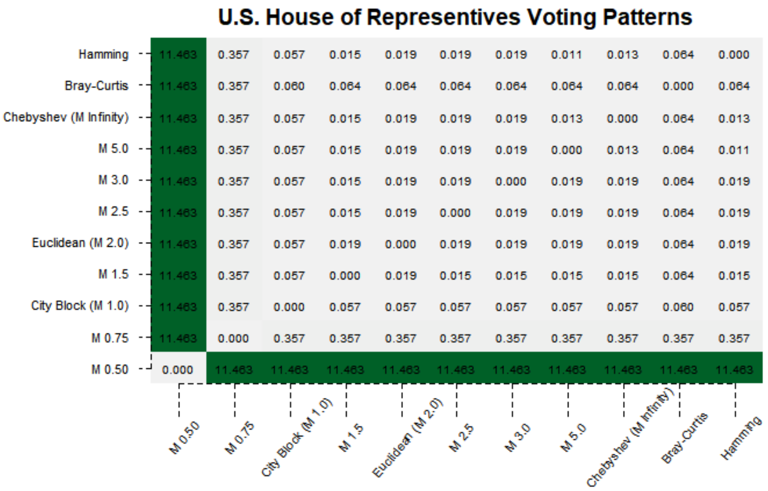
HIV1 Genome

	M 0.50	M 0.75	City Block (M 1.0)	M 1.5	Euclidean (M 2.0)	M 2.5	M 3.0	M 5.0	Chebyshev (M Infinity)	Bray-Curtis	Hamming
Hamming	131428.5	1971.3	226.5	89.0	92.4	68.3	57.1	48.4	42.9	0.0	0.0
Bray-Curtis	131428.5	1971.3	226.5	89.0	92.4	68.3	57.1	48.4	42.9	0.0	0.0
Chebyshev (M Infinity)	131428.5	1971.3	226.5	89.0	92.4	68.3	57.1	35.5	0.0	42.9	42.9
M 5.0	131428.5	1971.3	226.5	89.0	92.4	68.3	57.1	0.0	35.5	48.4	48.4
M 3.0	131428.5	1971.3	226.5	89.0	92.4	68.3	0.0	57.1	57.1	57.1	57.1
M 2.5	131428.5	1971.3	226.5	89.0	92.4	0.0	68.3	68.3	68.3	68.3	68.3
Euclidean (M 2.0)	131428.5	1971.3	226.5	92.4	0.0	92.4	92.4	92.4	92.4	92.4	92.4
M 1.5	131428.5	1971.3	226.5	0.0	92.4	89.0	89.0	89.0	89.0	89.0	89.0
City Block (M 1.0)	131428.5	1971.3	0.0	226.5	226.5	226.5	226.5	226.5	226.5	226.5	226.5
M 0.75	131428.5	0.0	1971.3	1971.3	1971.3	1971.3	1971.3	1971.3	1971.3	1971.3	1971.3
M 0.50	0.0	131428.5	131428.5	131428.5	131428.5	131428.5	131428.5	131428.5	131428.5	131428.5	131428.5

Figure 5 Distance matrix between metrics on HIV1 Genome.

The unique reason for picking this source is the lack of concept of distance between genomes. In the roadmap for persistent homology [**Otter et al.(2015)**], the authors used the hamming metric to find the difference between genomes. This seems like a reasonable choice, given that hamming distance in strings is the minimum number of modifications to move from one string to another. Genomes, of course, are represented as strings of A, T, C, and G.

The distance matrix is shown in Figure 5.



U.S. House of Representives Voting Patterns

	M 0.50	M 0.75	City Block (M 1.0)	M 1.5	Euclidean (M 2.0)	M 2.5	M 3.0	M 5.0	Chebyshev (M Infinity)	Bray-Curtis	Hamming
Hamming	11.463	0.357	0.057	0.015	0.019	0.019	0.019	0.011	0.013	0.064	0.000
Bray-Curtis	11.463	0.357	0.060	0.064	0.064	0.064	0.064	0.064	0.064	0.000	0.064
Chebyshev (M Infinity)	11.463	0.357	0.057	0.015	0.019	0.019	0.019	0.013	0.000	0.064	0.013
M 5.0	11.463	0.357	0.057	0.015	0.019	0.019	0.019	0.000	0.013	0.064	0.011
M 3.0	11.463	0.357	0.057	0.015	0.019	0.019	0.000	0.019	0.019	0.064	0.019
M 2.5	11.463	0.357	0.057	0.015	0.019	0.000	0.019	0.019	0.019	0.064	0.019
Euclidean (M 2.0)	11.463	0.357	0.057	0.019	0.000	0.019	0.019	0.019	0.019	0.064	0.019
M 1.5	11.463	0.357	0.057	0.000	0.019	0.015	0.015	0.015	0.015	0.064	0.015
City Block (M 1.0)	11.463	0.357	0.000	0.057	0.057	0.057	0.057	0.057	0.057	0.060	0.057
M 0.75	11.463	0.000	0.357	0.357	0.357	0.357	0.357	0.357	0.357	0.357	0.357
M 0.50	0.000	11.463	11.463	11.463	11.463	11.463	11.463	11.463	11.463	11.463	11.463

Figure 6 Distance matrix between metrics on House of Representatives.

U.S. House of Representatives Roll-Call Voting Records This data source creates a network of the 435 members of the 104th United States House of Representatives, as created by [**Waugh et al.(2011)**]. The original source is from [**Voteview, UCLA(2023)**]. Each representative is a node and the relationship between nodes is a ratio of how often they voted the same on bills.

The distance matrix between metrics is shown in Figure 6.

Stanford Dragon The Standford Dragon is a common 3D modeling object, 3D scanned from an original model [**Stanford University(1996)**]. The data source we use is 2000 random points from the dragon model, as sourced from [**Otter et al.(2015)**]. The dragon is seen in Figure 7.

Figure 7 A 3D Graphics Scene with the Stanford Dragon

In this case, we would expect the euclidean metric to be best model for distance, since these are points describing an object in 3-Dimensional space. Our metric distances are shown in Figure 8

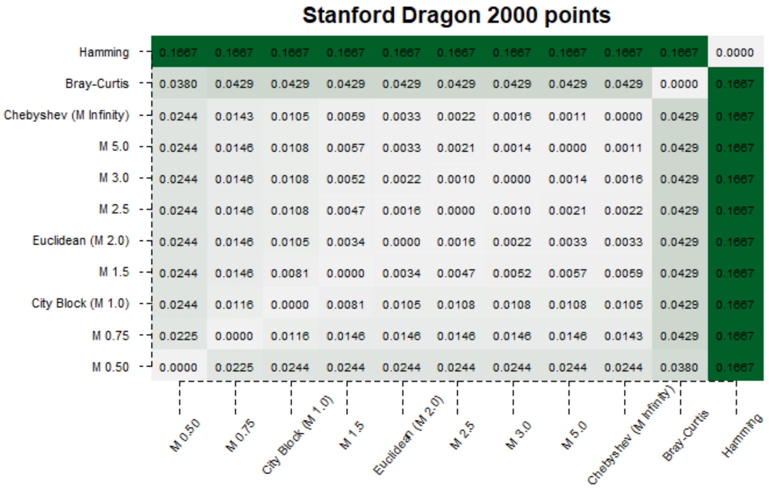
Torus of genus 1 in the 400th dimension This is a contrived data source, where 400 dimensional points create a torus. This point cloud is sourced from Dr. Lori Ziegelmeier at Macalester College. In this case, we have expected results of one connected component, two holes, and one internal void. How will different metrics impact this data set?

The distance matrix result between all the metrics is shown in Figure 9.

Further, we also explored showing the actual homology of this torus for each metric, shown in Figure 10

Results

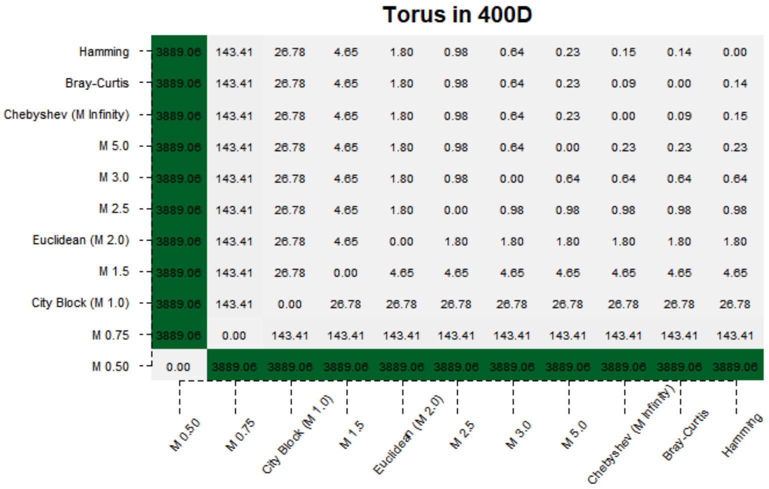
For every distance matrix, there is a common pattern. If you pick an equality result (where the distance is 0, to the same metric), if you move down the chart, staying in the same column, the distance increases only. If you move up, it stays the same size, especially within the Minkowski family. In all cases, the Hamming metric or the Minkowski ($p = 0.5$) are the metrics furthest away from every other metric.



Stanford Dragon 2000 points

	M 0.50	M 0.75	City Block (M 1.0)	M 1.5	Euclidean (M 2.0)	M 2.5	M 3.0	M 5.0	Chebyshev (M Infinity)	Bray-Curtis	Hamming
Hamming	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	0.0000
Bray-Curtis	0.0380	0.0429	0.0429	0.0429	0.0429	0.0429	0.0429	0.0429	0.0429	0.0000	0.1667
Chebyshev (M Infinity)	0.0244	0.0143	0.0105	0.0059	0.0033	0.0022	0.0016	0.0011	0.0000	0.0429	0.1667
M 5.0	0.0244	0.0146	0.0108	0.0057	0.0033	0.0021	0.0014	0.0000	0.0011	0.0429	0.1667
M 3.0	0.0244	0.0146	0.0108	0.0052	0.0022	0.0010	0.0000	0.0014	0.0016	0.0429	0.1667
M 2.5	0.0244	0.0146	0.0108	0.0047	0.0016	0.0000	0.0010	0.0021	0.0022	0.0429	0.1667
Euclidean (M 2.0)	0.0244	0.0146	0.0105	0.0034	0.0000	0.0016	0.0022	0.0033	0.0033	0.0429	0.1667
M 1.5	0.0244	0.0146	0.0081	0.0000	0.0034	0.0047	0.0052	0.0057	0.0059	0.0429	0.1667
City Block (M 1.0)	0.0244	0.0116	0.0000	0.0081	0.0105	0.0108	0.0108	0.0108	0.0105	0.0429	0.1667
M 0.75	0.0225	0.0000	0.0116	0.0146	0.0146	0.0146	0.0146	0.0146	0.0143	0.0429	0.1667
M 0.50	0.0000	0.0225	0.0244	0.0244	0.0244	0.0244	0.0244	0.0244	0.0244	0.0380	0.1667

Figure 8 Distance matrix between metrics on Stanford Dragon.



Torus in 400D

	M 0.50	M 0.75	City Block (M 1.0)	M 1.5	Euclidean (M 2.0)	M 2.5	M 3.0	M 5.0	Chebyshev (M Infinity)	Bray-Curtis	Hamming
Hamming	3889.06	143.41	26.78	4.65	1.80	0.98	0.64	0.23	0.15	0.14	0.00
Bray-Curtis	3889.06	143.41	26.78	4.65	1.80	0.98	0.64	0.23	0.09	0.00	0.14
Chebyshev (M Infinity)	3889.06	143.41	26.78	4.65	1.80	0.98	0.64	0.23	0.00	0.09	0.15
M 5.0	3889.06	143.41	26.78	4.65	1.80	0.98	0.64	0.00	0.23	0.23	0.23
M 3.0	3889.06	143.41	26.78	4.65	1.80	0.98	0.00	0.64	0.64	0.64	0.64
M 2.5	3889.06	143.41	26.78	4.65	1.80	0.00	0.98	0.98	0.98	0.98	0.98
Euclidean (M 2.0)	3889.06	143.41	26.78	4.65	0.00	1.80	1.80	1.80	1.80	1.80	1.80
M 1.5	3889.06	143.41	26.78	0.00	4.65	4.65	4.65	4.65	4.65	4.65	4.65
City Block (M 1.0)	3889.06	143.41	0.00	26.78	26.78	26.78	26.78	26.78	26.78	26.78	26.78
M 0.75	3889.06	0.00	143.41	143.41	143.41	143.41	143.41	143.41	143.41	143.41	143.41
M 0.50	0.00	3889.06	3889.06	3889.06	3889.06	3889.06	3889.06	3889.06	3889.06	3889.06	3889.06

Figure 9 400D Genus 1 Torus Metric Distances.

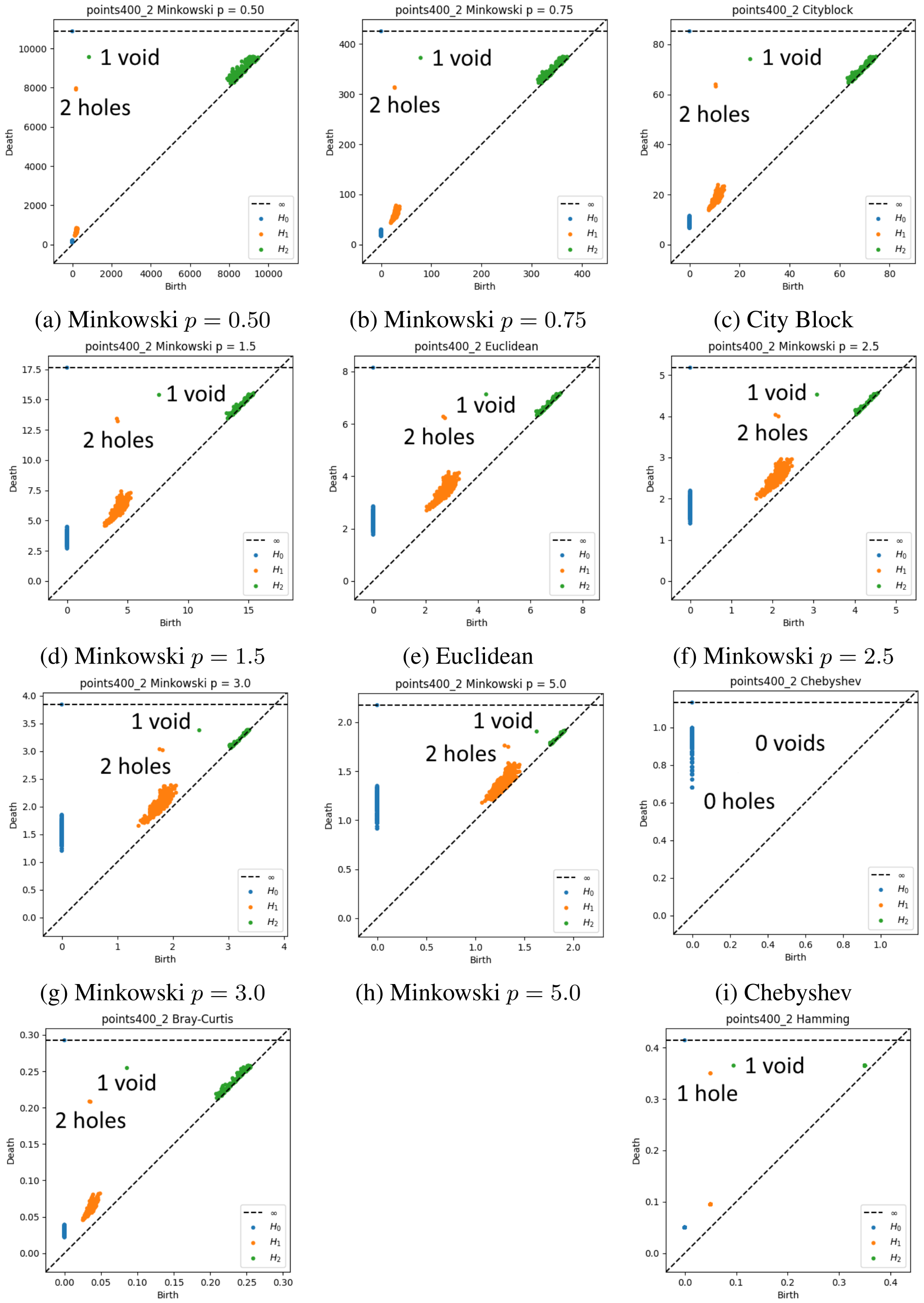
There is some sort of appearance of a ‘local minima’ effect, but the human brain identifies 200% of the patterns there are. So we are not likely to put any weight on this.

Considering Figure 10 and the results of the Torus shown in Figure 9, there is an interesting comparison between what the matrix is saying versus the actual shown results in the persistence diagrams. Nearly every metric gives the same result of 2 holes and 1 void. Chebyshev fails, but this may be due to the construction of the point cloud. It would require further testing.

As the Minkowski p increases, the birth of holes incrementally moves to the right within the lifetime of the system. Interestingly, the only other major difference is the scale of the axes, which we believe to be the major factor for the large distances from the bottleneck function. In fact, this seems to tell us that the bottleneck distance is highly ineffective when comparing between two different metrics: axis scale is a greater contributor to the distance between persistences than the persistences having the same result.

Conclusion

There is no ‘best’ metric for persistence homology and these distance matrices are hard to interpret. However, when compared also with the actual persitence diagrams we must come to the conclusion that the bottleneck distance, which gives us these



(a) Minkowski $p = 0.50$ (b) Minkowski $p = 0.75$ (c) City Block

(d) Minkowski $p = 1.5$ (e) Euclidean (f) Minkowski $p = 2.5$

(g) Minkowski $p = 3.0$ (h) Minkowski $p = 5.0$ (i) Chebyshev

(j) Bray-Curtis (k) Hamming

Figure 10 Persistence Diagrams for 400-Dimensional Genus 1 Torus

results, is inappropriate for comparing the results of persistent homology. An ideal distance function between persistences would be one that did not care about axis scale and would instead be more impacted by the presence and the noise and errors of the actual features–those being holes and voids.

Limitations Though we compared the homologies of the metrics for the torus, a more thorough analysis would include measuring the homology for each metric and for each data set and comparing the number of components, holes, etc.

Further Work We would like to work more in the future to consider the Wasserstein distance function as well as other possible distance functions between distributions, or perhaps devising our own.

REFERENCES

Carlsson(2009). Gunnar Carlsson. 2009. Topology and Data. *Bulletin of The American Mathematical Society - BULL AMER MATH SOC* 46 (04 2009), 255–308. https://doi.org/10.1090/S0273-0979-09-01249-X

Chan et al.(2013). Joseph Minhow Chan, Gunnar Carlsson, and Raul Rabadan. 2013. Topology of viral evolution. *Proceedings of the National Academy of Sciences* 110, 46 (2013), 18566–18571. https://doi.org/10.1073/pnas.1313480110 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1313480110

Los Alamos National Laboratory(2021). Los Alamos National Laboratory. 2021. HIV Databases. https://www.hiv.lanl.gov/content/index.

Otter et al.(2015). Nina Otter, Mason Porter, Ulrike Tillmann, Peter Grindrod, and Heather Harrington. 2015. A roadmap for the computation of persistent homology. *EPJ Data Science* 6 (06 2015). https://doi.org/10.1140/epjds/s13688-017-0109-5

Parker(2021). Matt Parker. 2021. What is the area of a Squircle? https://youtu.be/gjtTcyWL0NA.

Stanford University(1996). Stanford University. 1996. The Stanford 3D Scanning Repository. http://graphics.stanford.edu/data/3Dscanrep/.

Teknomo(2011). K. Teknomo. 2011. Similarity Measurement. *http://people.revoledu.com/kardi/tutorial/Similarity* (2011). https://cir.nii.ac.jp/crid/1570009750841011584

Tralie et al.(2018). Christopher Tralie, Nathaniel Saul, and Rann Bar-On. 2018. Ripser.py: A Lean Persistent Homology Library for Python. *The Journal of Open Source Software* 3, 29 (Sep 2018), 925. https://doi.org/10.21105/joss.00925

Voteview, UCLA(2023). Voteview, UCLA. 2023. UCLA presents Voteview.com beta. https://voteview.com/.

Waugh et al.(2011). Andrew Scott Waugh, Liuyi Pei, James H. Fowler, Peter J. Mucha, and Mason A. Porter. 2011. Party Polarization in Congress: A Network Science Approach. arXiv:0907.3509 [physics.soc-ph]