Persistent homology of street tree locations in two cities

Adele Jackson

January 29, 2018

In this assignment, I investigated two datasets of street tree locations: one from New York City [3], and one from San Francisco [4]. These were both obtained from Kaggle, which hosts many public datasets.

I hypothesised that these point sets would closely map to the streets of their respective cities, and that their persistent homology would reflect the structure of the cities. I further thought that Manhattan's large number of regularly sized blocks would create a distinctive feature in the first dimension of the persistent homology, and so the persistent homology of these street tree samples would be drawn from different distributions.

I filtered the New York City data by borough to only include the data from Manhattan, to pick out this distinctive regularity of the block sizes. I also processed the San Francisco data to only include data from inner city San Francisco, as the dataset was too large to handle otherwise. I then removed all features from the data aside from latitude and longitude, and converted these polar coordinates into cartesian coordinates.

I then had two datasets of points in three dimensional space: one from New York with about 65,000 points, and the other from San Francisco with about 185,000 points. Plotting these points, as seen in Figure 1, shows they do indeed reflect the shape of the streets of the city. (The large white patches are parks, as there are no street trees where there are no streets.)

To construct $n = 10$ samples from these point sets, each containing $m = 1000$ points, I performed the following procedure. For the first sample, pick a base point at random, then take the m closest points to this set. For the next sample, pick a base point at least 450m away at random, then take the m closest points to this set by Euclidean distance, then repeat this for the remaining samples. The figure of 450m was calculated empirically for this value of m to minimise overlap of samples. In the majority of cases, the m points would be within 450m of the base point, and almost always most of them would be within this margin. Although a higher radius would have ensured no overlap, I found this favoured basepoints that were in extreme positions on the map too much, which gave samples that were not representative of the data.

I chose the number $m = 1000$ as a compromise between a large sample, and not sampling the majority of the dataset. The number of samples $n = 10$ was restricted by computing resources, as increasing the number of samples increases the running time as $O(n^3)$ when calculating the distance matrix for the samples.

To analyse the persistent homology, I calculated an alpha filtration of the data using Diode [1]. An alpha filtration was suitable for this dataset as it is a spatial interpretation of the data, that does not create simplices of higher dimension than the ambient space. This suited a view of street trees points as a sample from a map of the city.

I then calculated the persistence diagrams for all samples for homology 0 and 1 using Dionysus [2], which are shown in Appendix A attached after the report. Note that the first ten samples are from Manhattan, and the second ten are from San Francisco.

I noticed a gap in some of the first homology persistence diagrams for Manhattan, with very few births around 0.008. I plotted the barcodes for the first homology for all samples, attached in Appendix B. In most barcodes from Manhattan samples there is indeed a plateau around this point. To minimise the effect of points far from the diagonal, which would not capture the small

(a) San Francisco (b) Manhattan

Figure 1: Plots of street trees in the two cities.

scale structure of a block, I used Wasserstein distance with $q = 1$ to find the distance between persistence diagrams. Figure 2 is a colormap showing the distances between the sample points.

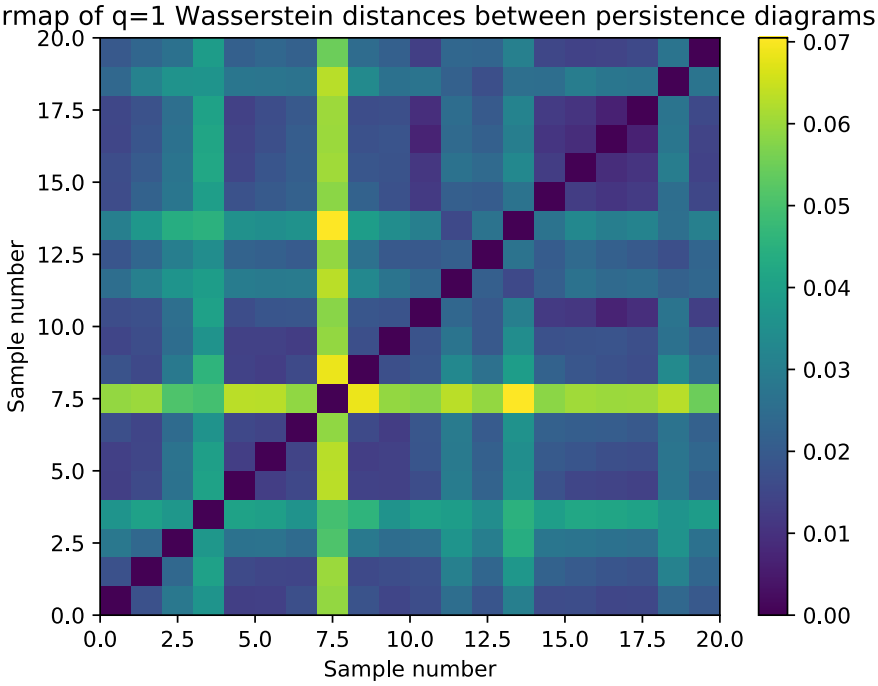


Figure 2: Colormap of Wasserstein distances, $q = 1$, between the persistence diagrams

To compute a p-value for the null hypothesis, I performed a permutation test on the partition of the samples. For the cost function, I used the sum of the squared distances between samples, and set the threshold p-value at the reciprocal of the number of permutations. For 1000 permutations, I obtained a result of $p \leq 0.001$, at the threshold. Thus, we can reject the null hypothesis and conclude that the two sample sets come from different distributions.

For the functional summary, I looked at the rank functions, again with an L_2 distance with a weight function that decays exponentially away from the diagonal. Figure 3 is a heat map of the distance from each sample's rank function to the average rank function for the two sample subsets. The first 10 samples, and first average function, are from Manhattan, and the others

from San Francisco. With 1000 permutations, the result is not statistically significant, as we get $p = 0.002$. As this is above the cutoff of 0.001, we cannot reject the null hypothesis.

The Wasserstein distance between diagrams has detected a significant difference between the two sample sets, so the city structures as reflected in the street tree locations are indeed significantly different. However, the functional summary did not detect a statistically significant difference. I would suggest that the rank function does not sufficiently emphasise points close to the diagonal, so is unsuited to this dataset.

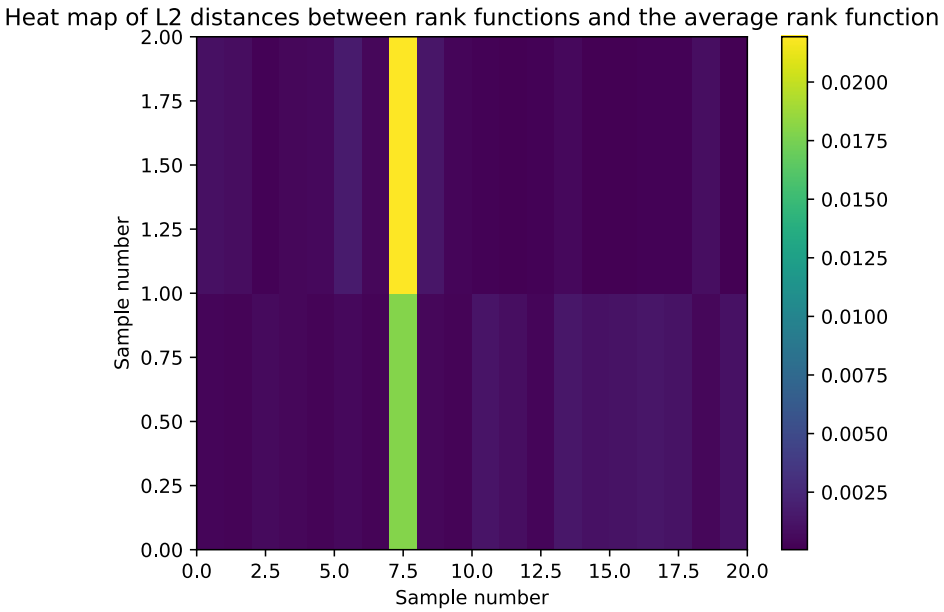


Figure 3: Colormap of L_2 distances from sample rank functions to the average rank functions for each sample subset.

The code used to sample and process this data can be found at `https://github.com/adelejackson/street-tree-homology`.

References

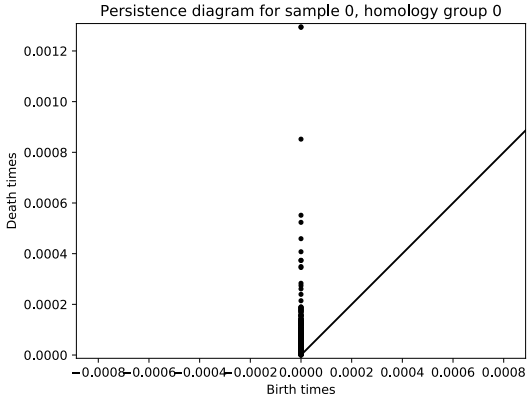
[1] Morozov, D. (2017). Diode [Python package.] Retrieved from `https://github.com/mrzv/diode`

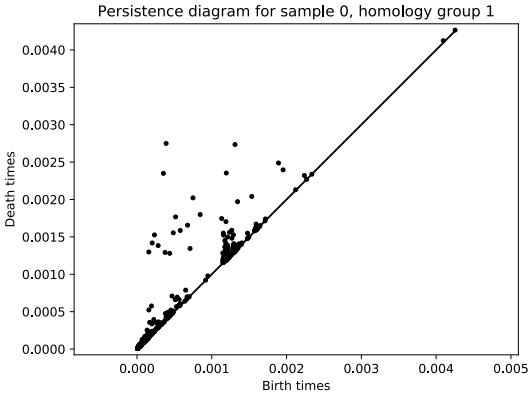
[2] Morozov, D. (2017). Dionysus [Python package.] Retrieved from `https://github.com/mrzv/dionysus`

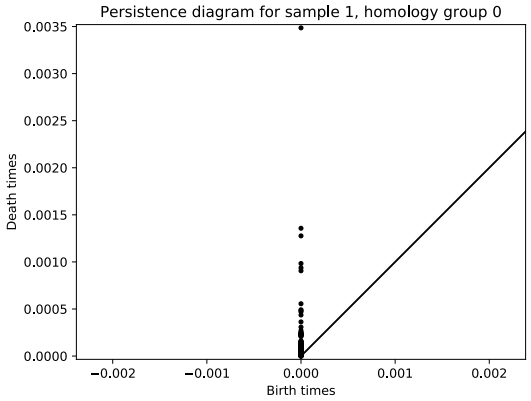
[3] New York City Department of Parks and Recreation. (2015). *TreesCount! 2015 Street Tree Census*. Retrieved from `https://www.kaggle.com/nycparks/tree-census`

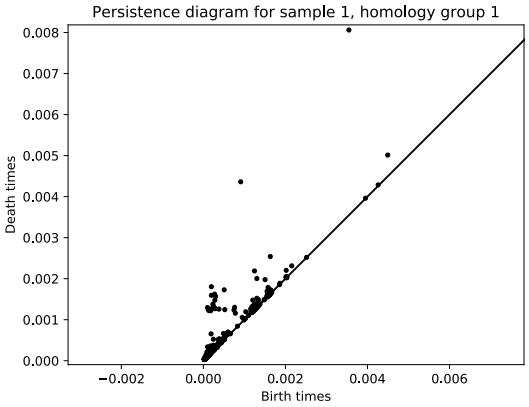
[4] San Francisco Public Works. (2016). *San Francisco Street Trees*. Retrieved from `https://www.kaggle.com/jboysen/sf-street-trees`

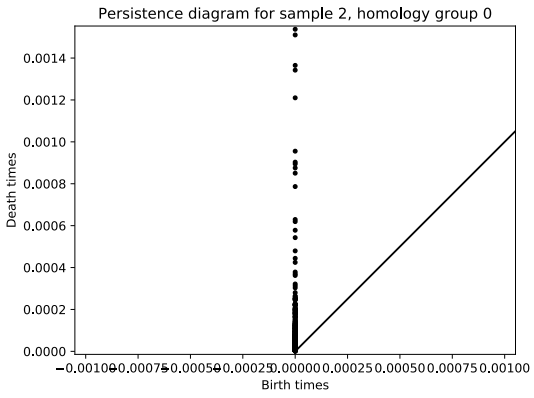
A Persistence diagrams for all samples

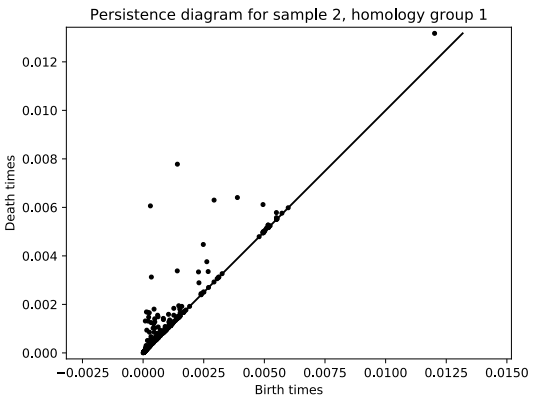


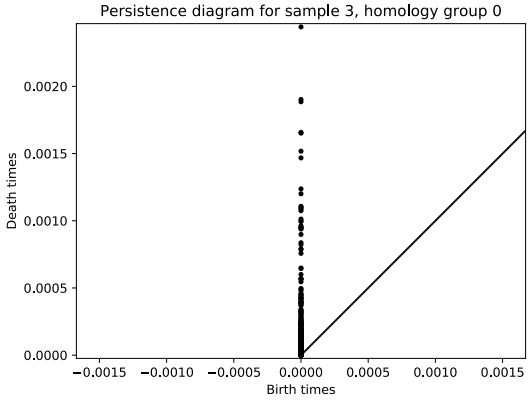


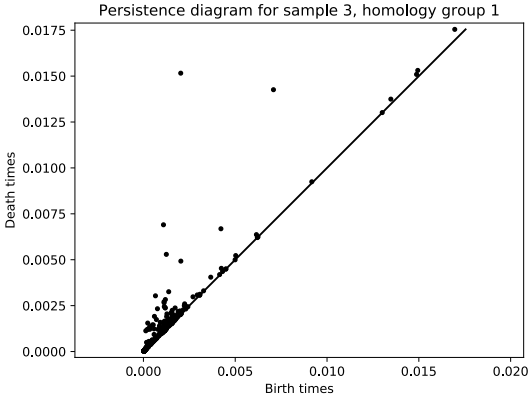


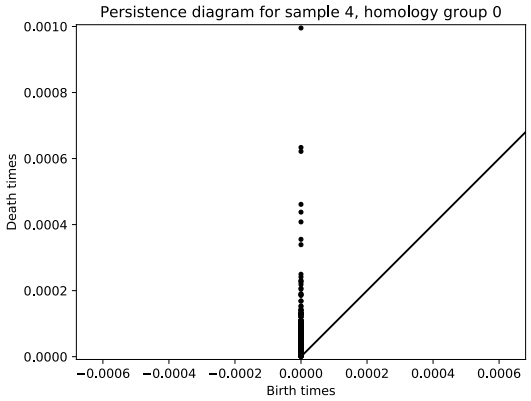


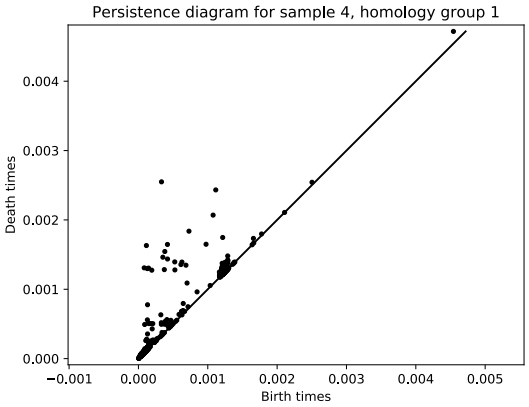


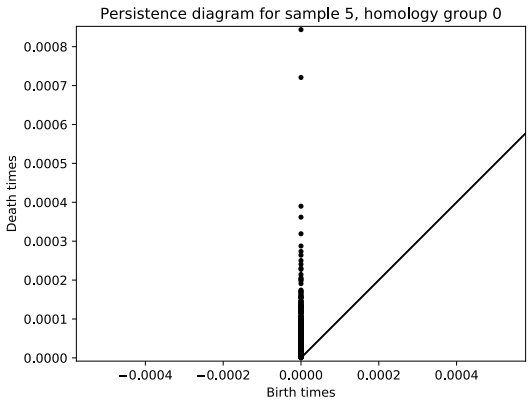


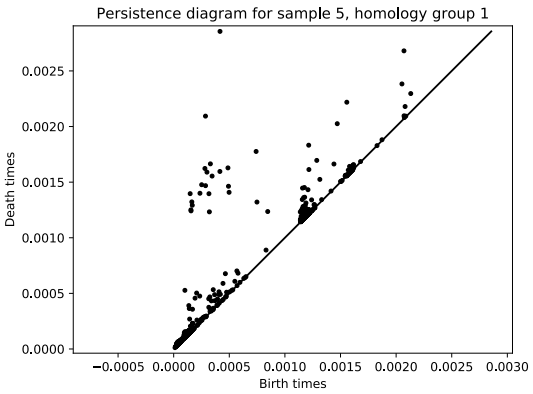


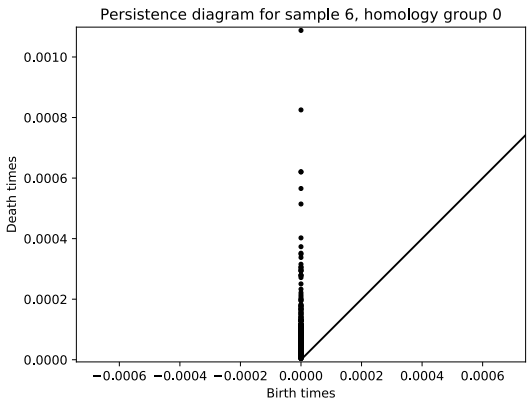


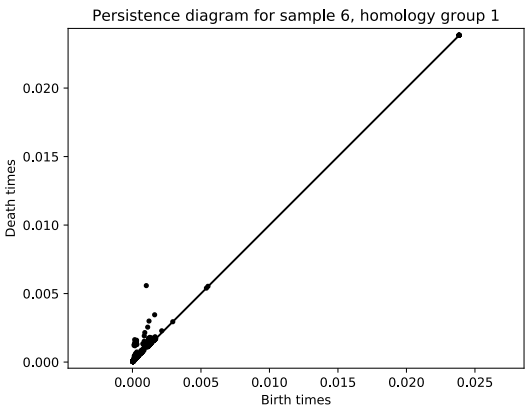

Persistence diagram for sample 4, homology group 0
Death times
Birth times

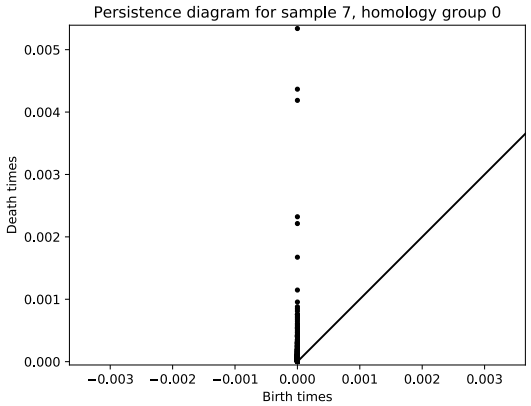

Persistence diagram for sample 4, homology group 1
Death times
Birth times

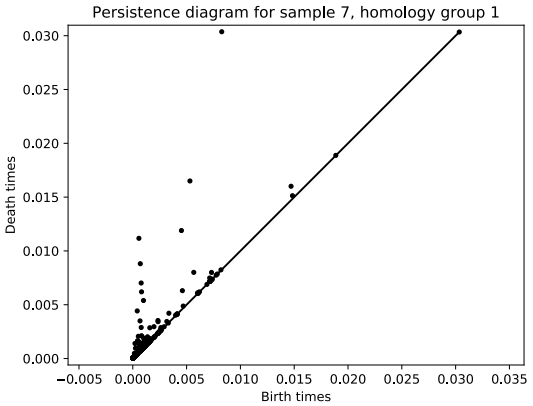

Persistence diagram for sample 5, homology group 0
Death times
Birth times

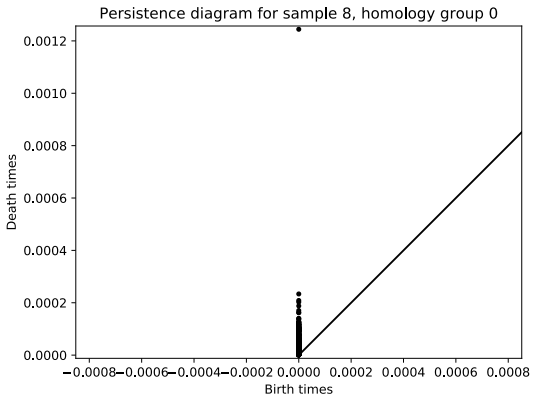

Persistence diagram for sample 5, homology group 1
Death times
Birth times

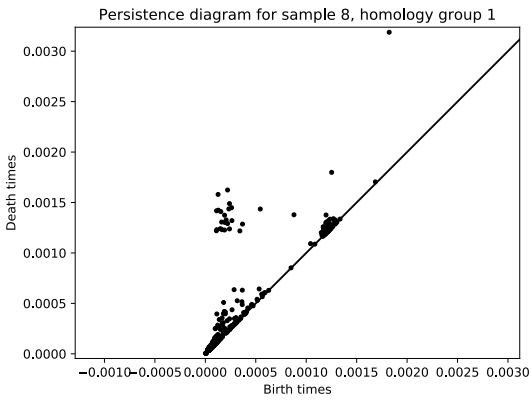

Persistence diagram for sample 6, homology group 0
Death times
Birth times

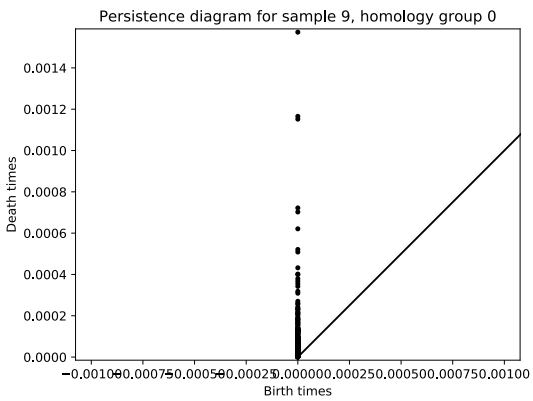

Persistence diagram for sample 6, homology group 1
Death times
Birth times

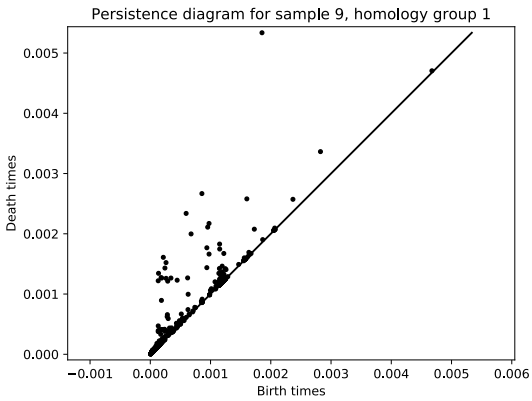

Persistence diagram for sample 7, homology group 0
Death times
Birth times

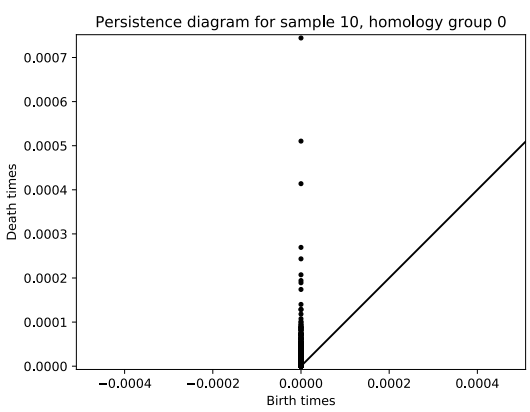

Persistence diagram for sample 7, homology group 1
Death times
Birth times

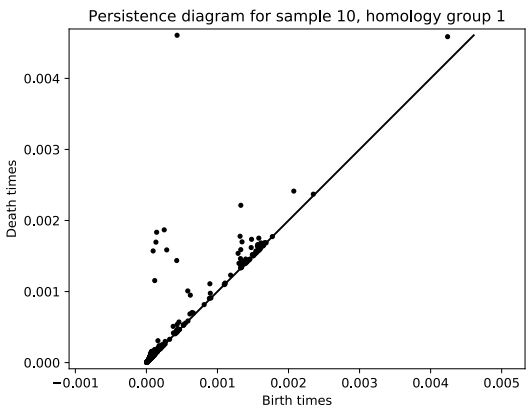

Persistence diagram for sample 8, homology group 0
Death times
Birth times

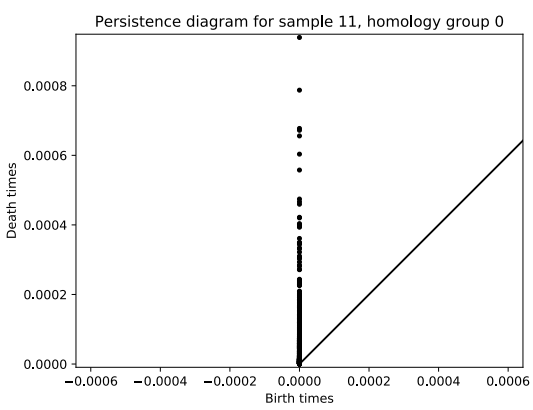

Persistence diagram for sample 8, homology group 1
Death times
Birth times

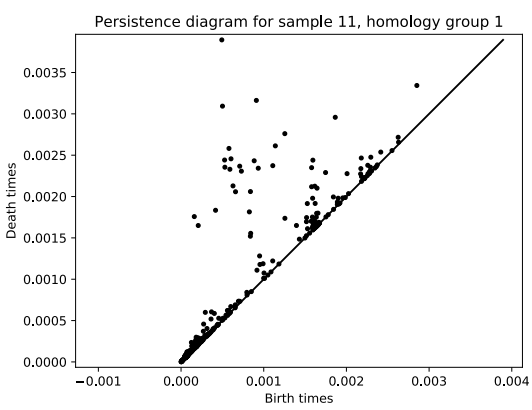

Persistence diagram for sample 9, homology group 0
Death times
Birth times

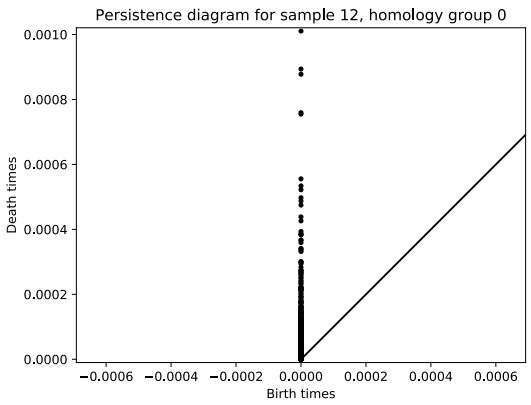

Persistence diagram for sample 9, homology group 1
Death times
Birth times

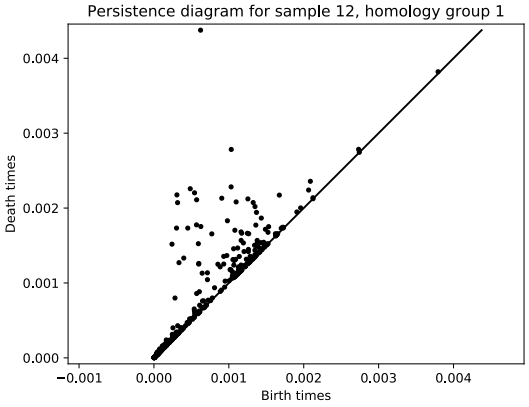

Persistence diagram for sample 10, homology group 0
Death times
Birth times

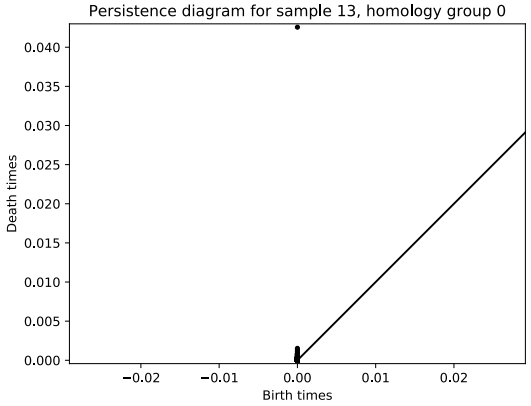

Persistence diagram for sample 10, homology group 1
Death times
Birth times

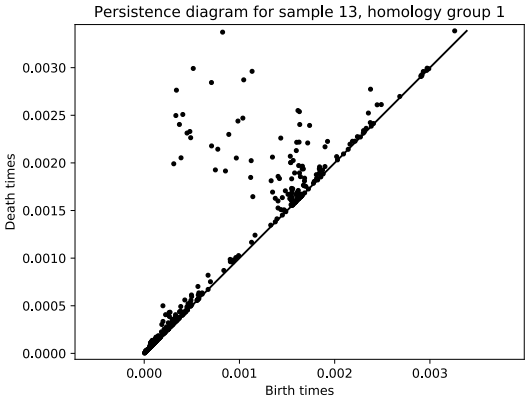

Persistence diagram for sample 11, homology group 0
Death times
Birth times

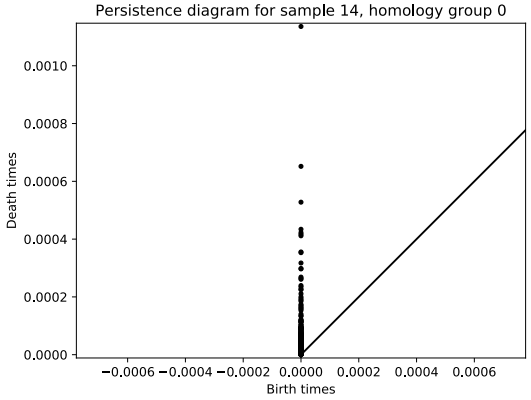

Persistence diagram for sample 11, homology group 1
Death times
Birth times

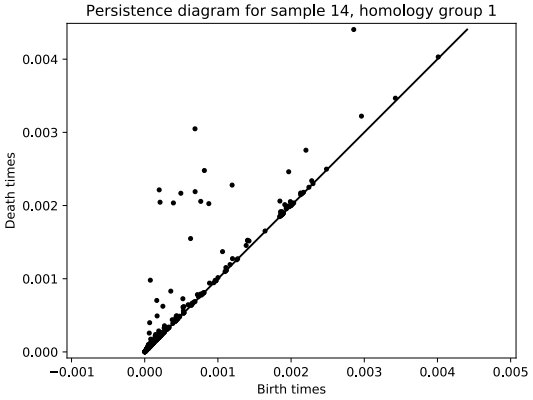

Persistence diagram for sample 12, homology group 0
Death times
Birth times

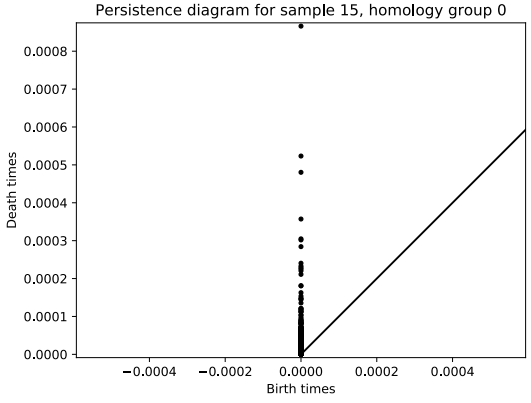

Persistence diagram for sample 12, homology group 1
Death times
Birth times

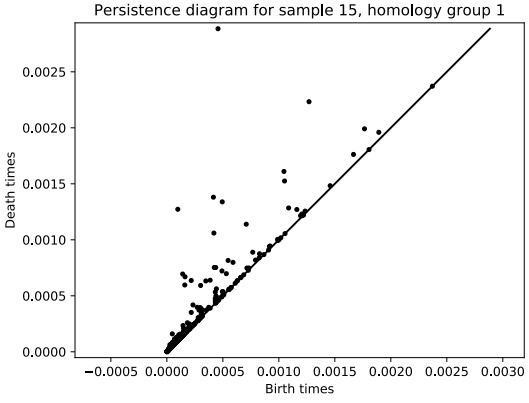

Persistence diagram for sample 13, homology group 0
Death times
Birth times


Persistence diagram for sample 13, homology group 1
Death times
Birth times


Persistence diagram for sample 14, homology group 0
Death times
Birth times


Persistence diagram for sample 14, homology group 1
Death times
Birth times


Persistence diagram for sample 15, homology group 0
Death times
Birth times


Persistence diagram for sample 15, homology group 1
Death times
Birth times

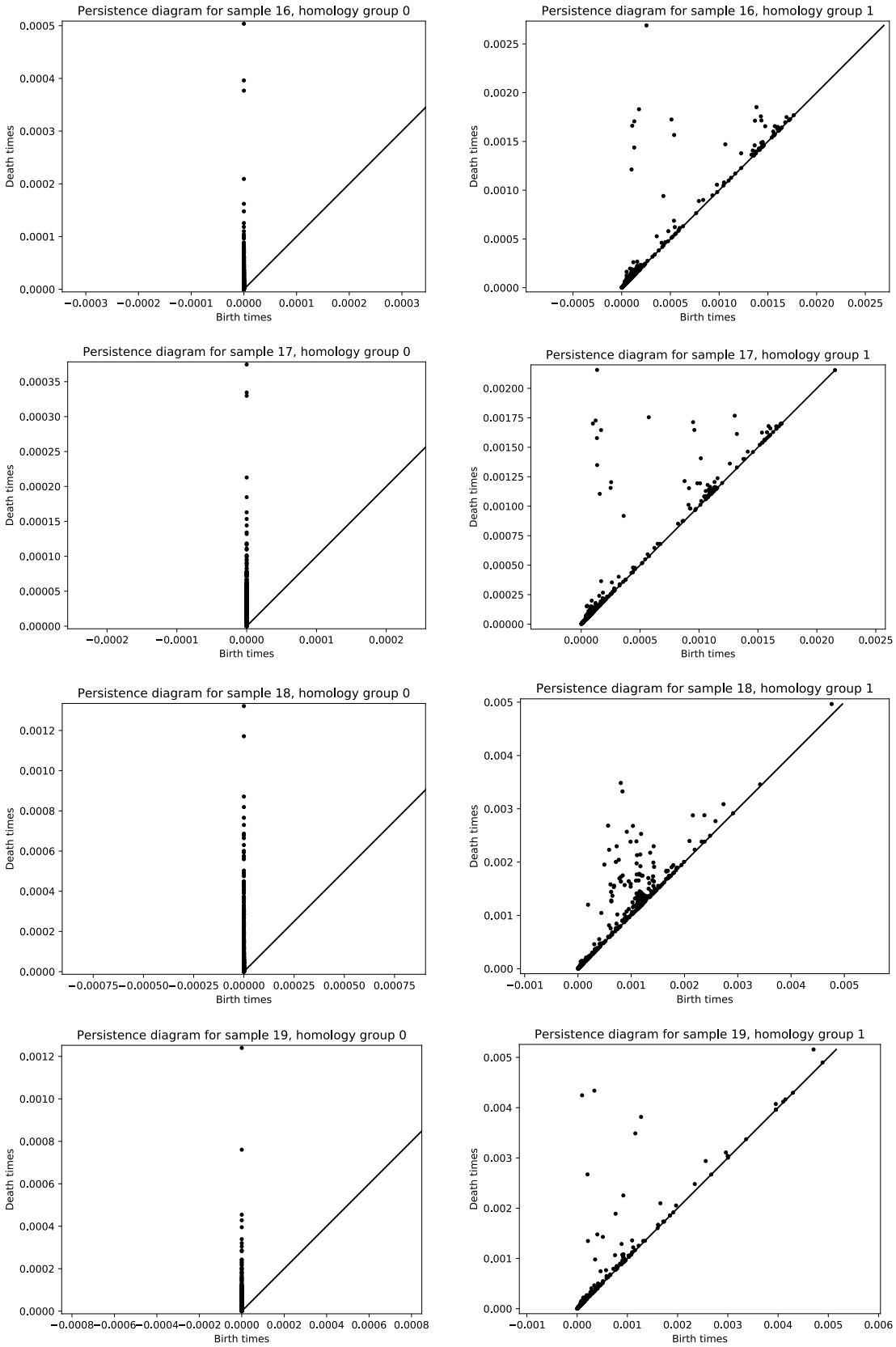
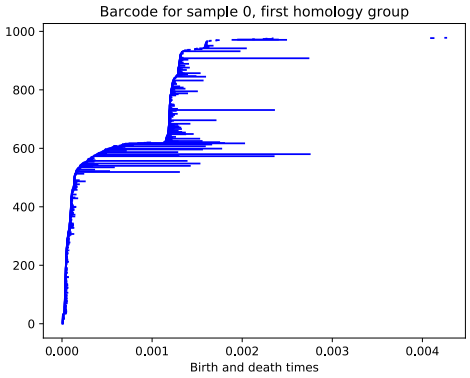
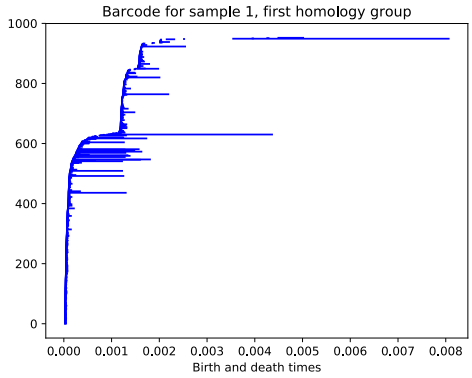
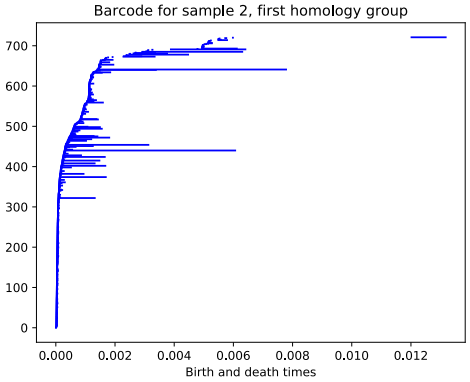


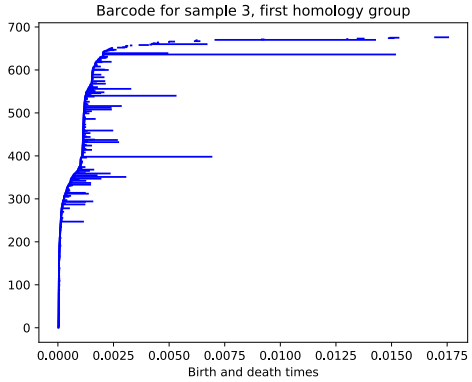
Figure 4: Persistent homology diagrams for all samples, dimensions 0 and 1. Samples 0-9 are from Manhattan; Samples 10-19 from San Francisco.

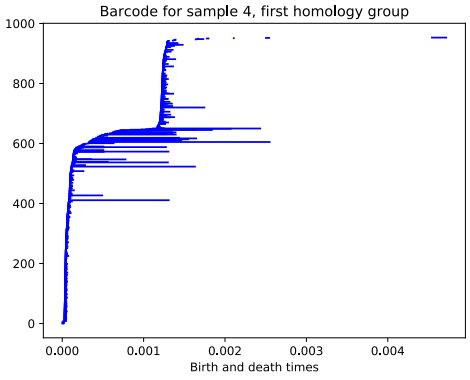
B Barcode diagrams for first persistent homology for all samples

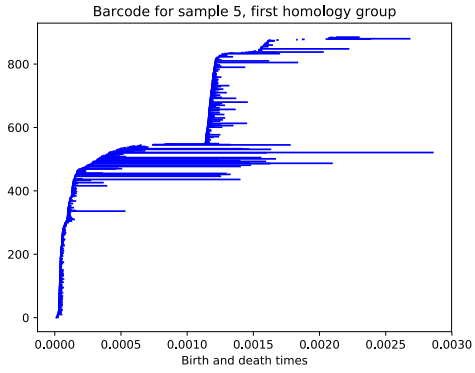


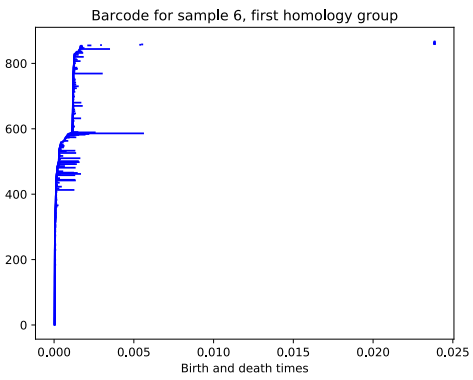


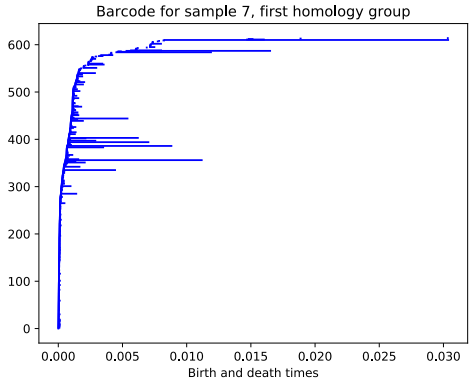


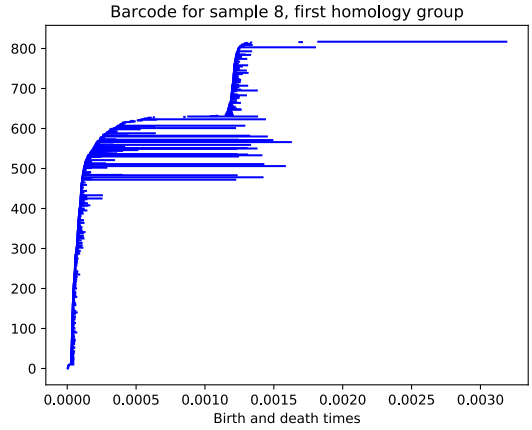


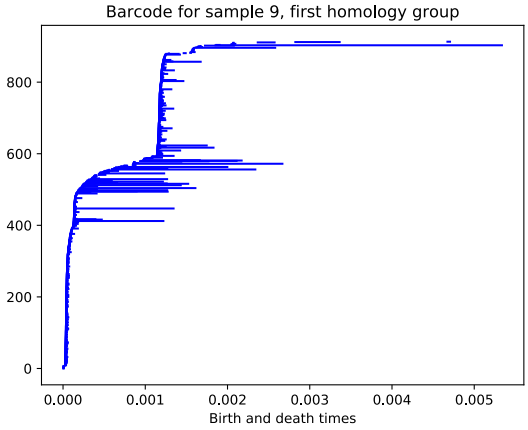


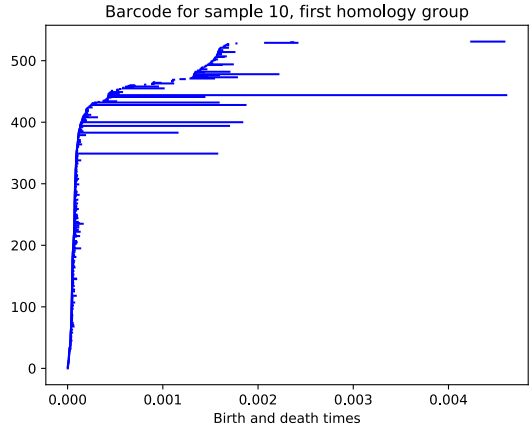


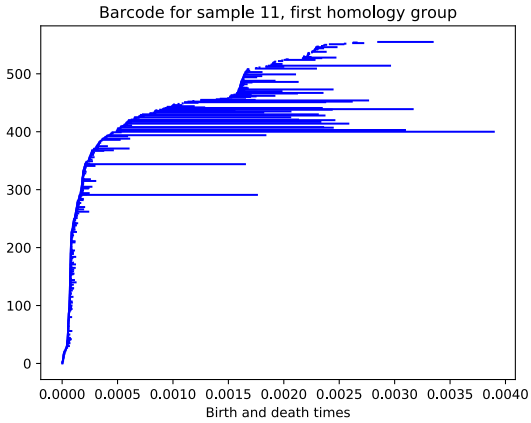


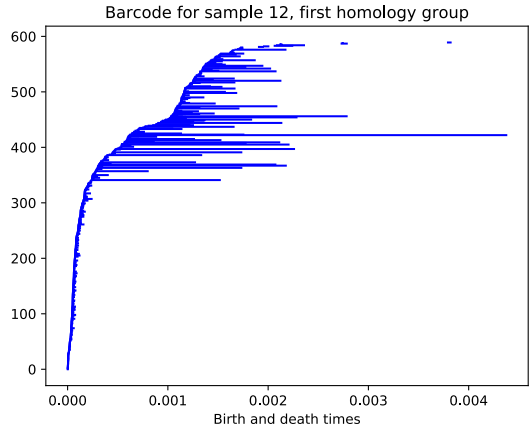


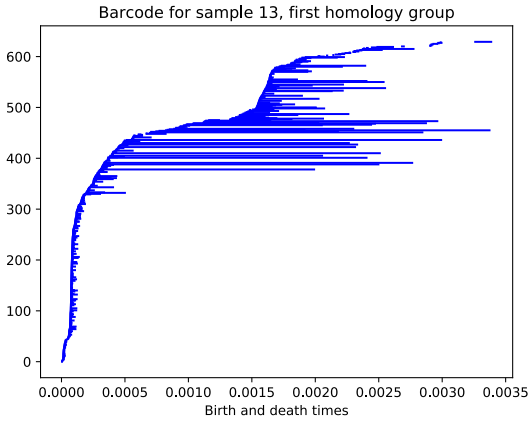

Barcode for sample 8, first homology group
0 100 200 300 400 500 600 700 800
0.0000 0.0005 0.0010 0.0015 0.0020 0.0025 0.0030
Birth and death times

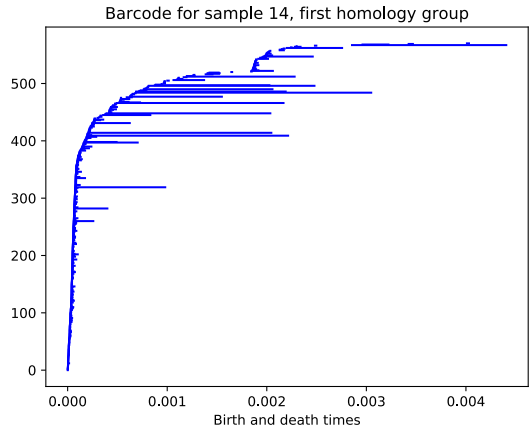

Barcode for sample 9, first homology group
0 200 400 600 800
0.000 0.001 0.002 0.003 0.004 0.005
Birth and death times

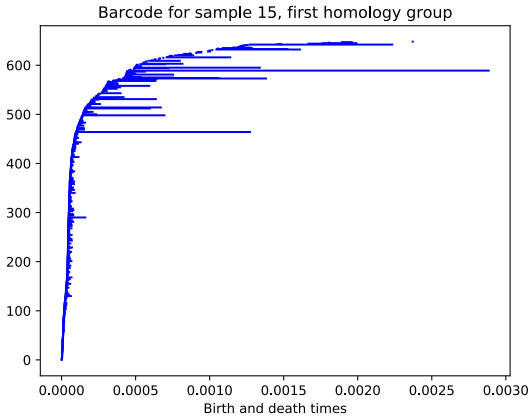

Barcode for sample 10, first homology group
0 100 200 300 400 500
0.000 0.001 0.002 0.003 0.004
Birth and death times


Barcode for sample 11, first homology group
0 100 200 300 400 500
0.0000 0.0005 0.0010 0.0015 0.0020 0.0025 0.0030 0.0035 0.0040
Birth and death times


Barcode for sample 12, first homology group
0 100 200 300 400 500 600
0.000 0.001 0.002 0.003 0.004
Birth and death times


Barcode for sample 13, first homology group
0 100 200 300 400 500 600
0.0000 0.0005 0.0010 0.0015 0.0020 0.0025 0.0030 0.0035
Birth and death times


Barcode for sample 14, first homology group
0 100 200 300 400 500
0.000 0.001 0.002 0.003 0.004
Birth and death times


Barcode for sample 15, first homology group
0 100 200 300 400 500 600
0.0000 0.0005 0.0010 0.0015 0.0020 0.0025 0.0030
Birth and death times

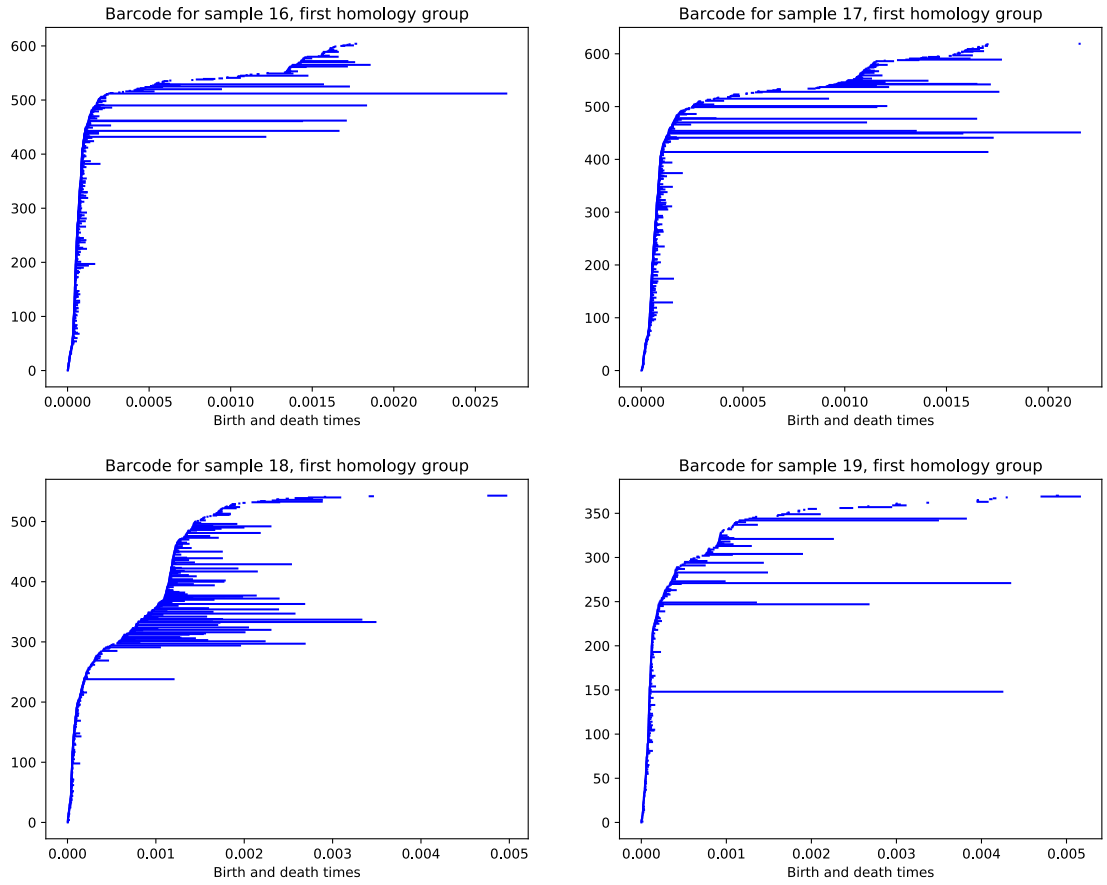


Figure 5: First dimension barcodes for all samples. Samples 0-9 are from Manhattan; Samples 10-19 are from San Francisco.