We have been investigating when topological persistence is robust to incomplete data. Specifically, the persistent homology of Vietoris-Rips complexes built from sparse weighted graphs. In the metric case we can compute the persistent homology of a set of points given a matrix of pairwise distances using state of the art software. However, computing the persistence of such a matrix requires working with a simplicial complex constructed from the complete metric graph of the point set. Our hope is that we can get some sense of how to work around missing data by identifying subsets of the complete graph with a persistence diagram close to that of the complete graph.

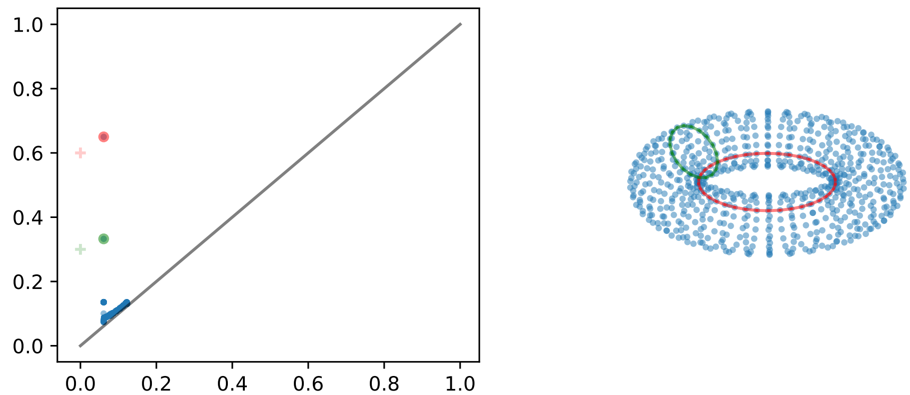

Figure 1: Persistent features of 722 uniformly spaced points sampled from $\mathbb{T}$. The green and red $+$ markers represent expected features at $r = 0.3$ and $R - r = 0.6$ respectively.

Let $\mathbb{T}_{r,R}$ denote the torus of revolution in $\mathbb{R}^3$ with minor radius r and major radius R. We consider the 1D persistent homology of uniformly and randomly spaced points sampled on $\mathbb{T}_{0.3,0.9}$. For the sake of notation we will let $\mathbb{T}$ denote $\mathbb{T}_{0.3,0.9}$.

As an initial test we removed edges at random. Figures 1 and 3 show the persistence diagrams and representative cycles of uniformly and randomly spaced points sampled from $\mathbb{T}_{0.3,0.9}$. Figures 2 and 4 show the effects of removing edges at random.

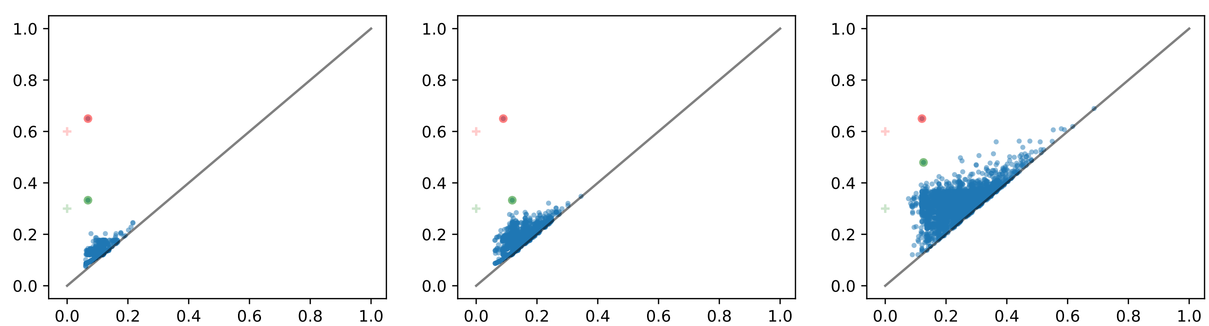

Figure 2: Persistence diagrams of 722 uniformly spaced points sampled from $\mathbb{T}$ with 25%, 50%, and 75% of the edges removed at random.

Note that in all cases the most prominent feature is retained. In both the uniform and random case we find that the second most prominent feature, representative of the minor radius can be identified up to 50% as a peak in the topological noise.

We then assumed the diagram of the underlying space is known - that is, in the case of $\mathbb{T}$ we

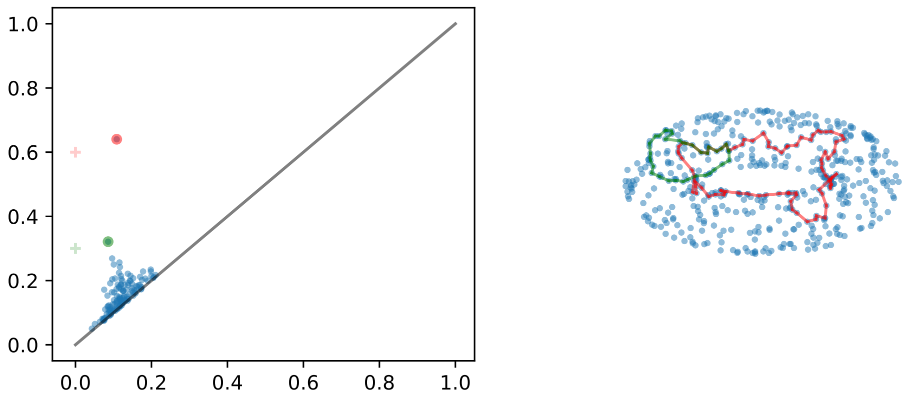

Figure 3: Persistent features of 500 randomly sampled points from $\mathbb{T}$. The green and red + markers represent expected features at $r = 0.3$ and $R - r = 0.6$ respectively.

know $r = 0.3$ and $R = 0.9$. Figures 5 and 6 show the effect of only keeping edges $\{p, q\}$ such that $|d(p, q)| < 0$, $|d(p, q) - r| < \epsilon$, or $|d(p, q) - (R - r)| < \epsilon$ for decreasing ϵ.

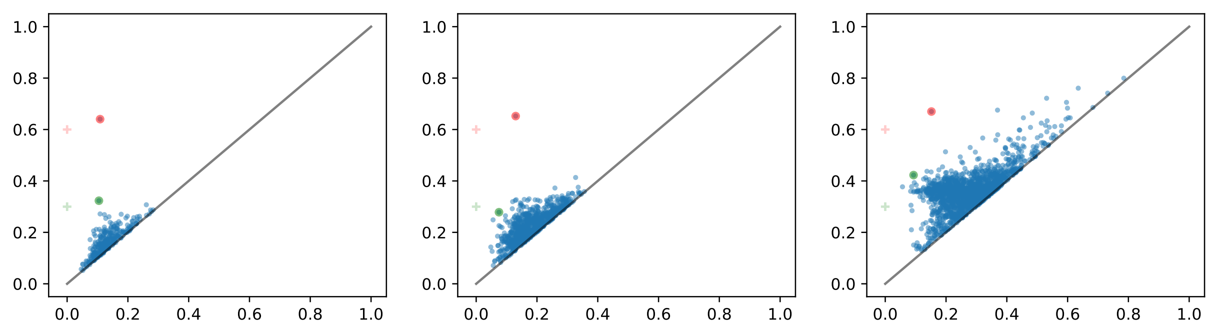

Figure 4: Persistence diagrams of 500 randomly sampled points from $\mathbb{T}_{3,9}$ with 25%, 50%, and 75% of the edges removed at random.

Figure 5 shows the persistence diagrams of four decrasing ϵ values constructed from 36%-18% of the total number of edges in the 722 point uniform sample. The full diagram of the uniformly spaced sample took 64 seconds. For $\epsilon \approx 0.06$, the maximum value for which both features appear, computing the diagram with 20% of the edges took 7.8 seconds.

Figure 6 shows the persistence diagrams of four decrasing ϵ values constructed from 43%-24% of the total number of edges in the random sample. The full diagram of the random spaced sample took 28 seconds. For $\epsilon \approx 0.09$, the maximum value for which both features appear, computing the diagram with 24% of the edges took 6.7 seconds.

In both cases we see that for the last ϵ value the feature whose death depends on the presence of the longest edges does not die, resulting in a diagram which is not immediately representative of the underlying space. One direction for future would be to work around this issue however, this experiment is useful primarily as a proof-of-concept. In practice the diagram of the underlying space is often not known.

Moving forward we would like to use these observations to explore ways to manipulate and simplifies input such as distance matrices common among state of the art persistence software to

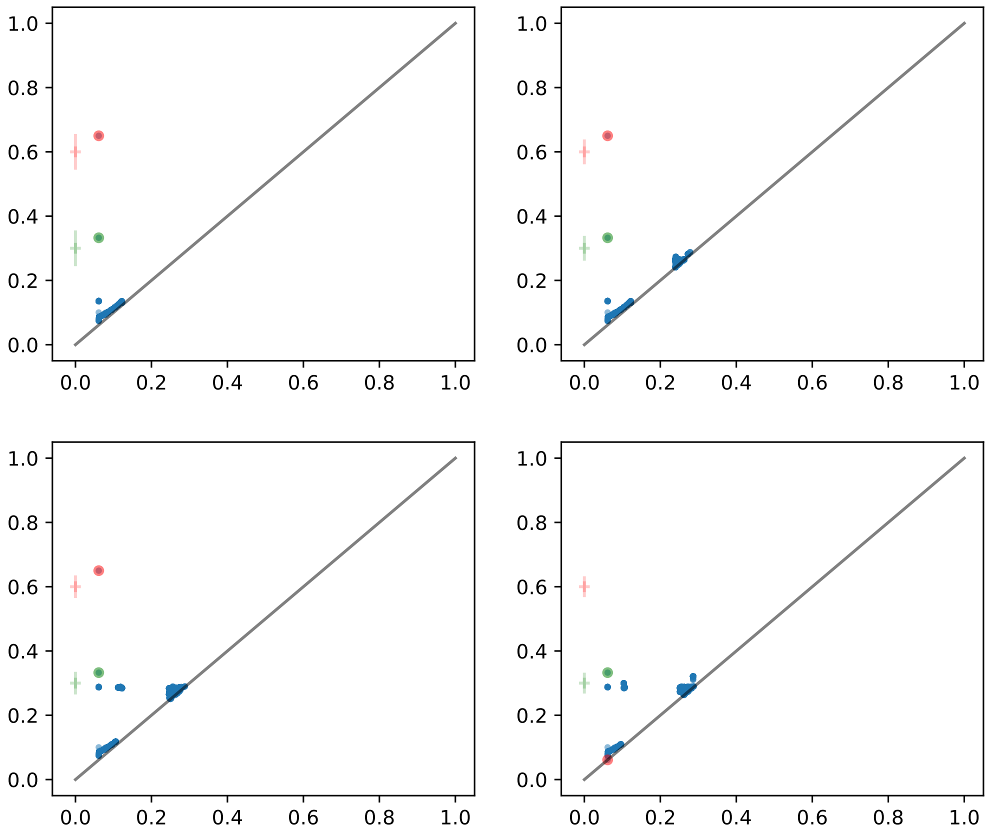

Figure 5: Persistence diagrams of 722 uniformly spaced points sampled from $\mathbb{T}$ with $\epsilon \approx 0.1, 0.07, 0.6, 0.55$ respectively.

implement approximation schemes such as sparse filtrations and geometric spanners. Our hope is that methods that are robust to the absence and manipulation of metric information can be naturally applied to non-metric data.

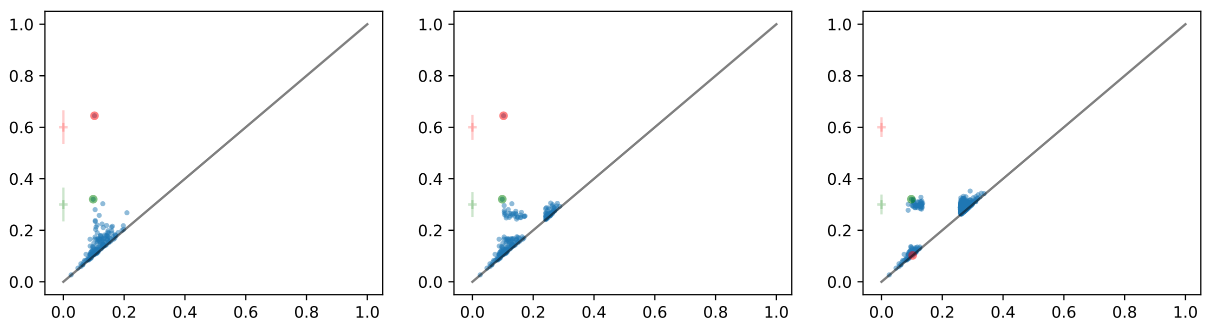

Figure 6: Persistence diagrams of 500 points randomly sampled from $\mathbb{T}$ with $\epsilon \approx 0.123, 0.087, 0.068$ respectively.