Computational topology - group project

A generic TDA pipeline

Introduction: Modern tools of data analysis are mostly tailored for use in Euclidean spaces. The output of persistence algorithm, the persistence diagrams, however are not vectors. In this project you will test various vectorization techniques for persistence diagrams, use the output to further process the result, and observe how the choice of vectorization effects the final result.

Goal: Test the standard TDA pipeline on a dataset. Given a collection of pointclouds, process each of them as follows:

1. Build a filtration;
2. Compute persistent homology;
3. Map it into a Euclidean space.

Then use a clustering to partition the resulting representation into different classes.

Detailed description:

1. Choose a collection of pointclouds in a metric space. Our goal is to try to find different classes.
2. Choose a filtration construction to be used. Then compute persistent homology for each pointcloud.
3. Choose a vectorization technique. This will allow you to represent each persistence diagram as a point in some Euclidean space. Some of these techniques include:
   - http://www.lix.polytechnique.fr/Labo/Ovsjanikov.Maks/papers/perslocsig.pdf
   - Persistence Images https://jmlr.org/papers/volume18/16-337/16-337.pdf

   You should feel free to adapt, simplify, or modify these techniques in any way you see fit.
4. Analyze the single obtained pointcloud (PCA, clustering, etc.). Try to use the result to partition the original pointclouds into clusters.

Results: The report should include a description and justification of techniques, pseudocodes, methods of computation, results of experiments, and division of work.

Data: Number of children per woman
https://ourworldindata.org/grapher/fertility-and-wanted-fertility
Choose several countries from at least 4 different continents, at least 10 countries in each continent (where available).
Some ideas for point coordinates: year, latitude/longitude of the capital city, number of people (changes yearly), GDP of the country, energy consumption of the country, average years of schooling...
How do data from different continents compare? How does data from rich countries compare to poor countries?

You are encouraged to take the initiative and possibly implement your own ideas on the theme of the project: perhaps thinking of new vectorisation techniques, pre- or post-processing steps, etc.