DSC190 (Sp'21) – Potential Project Topics

1 Reminder of timelines

Please choose your project topics by Week 6 (by May 8th). At that time, undergraduate students need to send me an email informing the topic that she/he will likely to work on – it is possible to change later as long as there is a good reason. By May 8th, graduate student should send me a one-page proposal on what you plan to do, motivation and some action plan.

Always feel free to come talk to me if you have questions or trouble in identifying a topic. The standard office hour is Tue@10am. You can also make an appointment with me.

2 Potential project topics for students from DSC 190

Below I will provide some potential topics. However, you are also free to come up with your own ideas. If you hope to have more challenging topics, you can come talk to me or you can also check out the list below for DSC291. In these projects, part of the task may involve downloading and processing some raw data as well. Also note that for these projects, you can use existing tools and do not need to implement, say persistence computation, from scratch.

T1 Download some neuron cell data sets from either Neuromorpho.org (or from a github that I will provide). Each neuron cell is represented by a tree. Compute suitable persistence diagrams (as topological summaries) for these neuron trees, and compare / cluster them via some clustering method. You can try different distance metrics for persistence diagram summaries.

T2 Download some surface models from AIMShape repository (*http://visionair.ge.imati.cnr.it/ontologies/shapes/*). Treat them as point clouds or as just graph. Compute the persistent homology profiles and compare/ cluster them. You can try different distance metrics for persistence diagram summaries.

T3 Topological analysis of time-series data. Given a time-series data, one can embed it as a point clouds data via say the Taken's embedding. one can then perform topological analysis to the datasets. Download some time-series traffic data from *https://pems.dot.ca.gov/?dnode=Clearinghouse*. Compute and compare these time-series data via their persistence profiles.

T4 Download some protein structures from protein data bank . Compute suitable persistence diagrams (as topological summaries) for these atomic structures. Compare / cluster them via some clustering method. You can try different distance metrics for persistence diagram summaries.

Alternatively, you can focus on a few molecules and provide detailed topological analysis. I think they have structures for some SARS-CoV-2 Spike and Antibodies available (*http://pdb101.rcsb.org/motm/256*). You can study their topological profiles.

T5 Computing topological profiles for some (2D or 3D) imagine data, which can be rock micro-images or images for CNT-composite. You can try different ways to produce filtrations and thus persistence diagrams. You can also try both local and global ones as motifs.

T6 We have high dimensional point cloud data (which comes from single-cell RNAseq data). Compute its topological persistence profile for suitable filtration. Or, compute graph skeleton of it via discrete Morse algorithm. Or other topological analysis you can propose.

3 Potential project topics for students from DSC 291

Ultimately, I hope that the class project can encourage you to explore topological data analysis methods (whether in applications or theoretically). You can develop a project topics in three different ways:

(Option 1) Talk to me and we can find a suitable topic for your project that can fit your background or your current research interests better. For example, you can have a topic to use topological methods to analyze the datasets from your own research. Come to the office hour (10am everything Tuesday) or send me an email to make an appointment to discuss.

(Option 2) I have a list potential topics below and you can choose one there. Feel free to discuss with me for details about any specific topics. [For students with mathematics background and may be more interested in theoretical / mathematical aspect of TDA, there are some survey topics I provide. Given the short period of time, it would be hard to have a theoretical research topic as a course project. However, if you are really interested in carrying out such a direction, we can discuss.]

(Option 3) Check out recent TDA related papers in NeurIPS, ICML or ICLR from 2019-2020 (or 2021 for ICLR). You can then propose to improve one such paper, or use it for your data (in a non-trivial manner).

A list of potential topics for (Option 2).

(T1) Graph classification. Using topological persistence-based summaries to classify graphs. Such summaries can then be combined with kernel methods; e.g, as in [4], or with neural networks; e.g, as in [1, 3]. There are general benchmark datasets available (can be found from the above references and the github for their codes). One can also try on chemical graphs, which are actually *geometric graphs* (namely, nodes are atoms and thus have geometric information associated with them). There are databases for such chemical graphs available as well.

(T2) Graph motif generation and bag of motifs. The goal of this project is to explore how to use the topological profiles of local graph motifs to help analyze (clustering or classifying) graphs. Similar to bag of motifs for sentences, here one can imagine to have a bag of topoogical-profiles of graph motifs, and use such motifs to represent a graph.

(T3) Tracking topological summaries for simulation data of Carbon-nanotube (CNT) composite data. Here the input is a collection of CNT-composite simulation data. A single simulation data can be thought of as a sequence $P_1, P_2, \ldots, P_n$, where each P_i is the atomic structure of a specific input CNT composite at the i-th time step of the simulation. The goal of this project is to track certain topological summaries through time (i.e, for P_i's for increasing is). For example, one simple approach would be to compute a certain persistence diagram summary for each P_i based on some descriptor function, and track it as i increases using the so-called *vineyard* [2].

(T4) Rock imaging data analysis: segmentation and motif distributions. There are some 3D micro-images of rocks, which contains grains (kind of forming voids) and the pore-space between the grains. Provide topological analysis of such images, for example, compute distributions of local persistence

summaries (to capture the grains), or segment the pore-networks (e.g, using discrete Morse based graph reconstruction).

If you have other types imaging data, you can also perform analysis for them.

(T5) Using NNs to predict persistence diagrams? In class we learned that the general persistence diagrams still take cubic time to compute in the worst case, which can expensive. Can we train a neural network to compute persistence diagrams; or an easier task will be to compute persistence images (which itself is an image)? The precise architecture of the NN will depend on the type of data. The easiest to handle would be 3D images – it may not make much sense for 2D images as computing persistence for a 2D image can be done in near linear time. More interesting datasets would be (i) graph data sets: in which case you will need a graph neural network) or (ii) point clouds in 2D/3D: in which case you can use pointNet, or still graph neural network for some proximity graph, or even treating point clouds as volume metric data (i.e, images).

(T6) Exploring the space behind high dimensional RNASeq data. We have some single cell RNAseq data (which are like gene expression data, essentially, you can think of these as high dimensional point clouds). Provide topological study of the space that these points are sampled from: e.g, using the Mapper methodology (which we will briefly describe in class).

(T7) Using topological tools to understand the behavior of neural networks: For example, given a neural network, we can view it as a weighted graph. Track the topological profile of it as one trains this neural network. Alternatively, given multiple trained neural netwoks for some simple classification problems (e.g, CNNs to classify MNIST images), compute the topological profiles of these networks and compare them.

(T8) Survey topic 1: Path homology for directed graphs.

(T9) Survey topic 2: Optimal (persistent) generating cycles and localized homology.

References

[1] M. Carrière, F. Chazal, Y. Ike, T. Lacombe, M. Royer, and Y. Umeda. Perslay: a neural network layer for persistence diagrams and new graph topological signatures. In *Proc. 23rd Internat. Conf. Artificial Intelligence Stat. (AISTATS)*, volume 108, pages 2786–2796, 2020.

[2] D. Cohen-Steiner, H. Edelsbrunner, and D. Morozov. Vines and vineyards by updating persistence in linear time. In *Proc. 22nd Annu. Sympos. Comput. Geom. (SoCG)*, pages 119–126, 2006.

[3] K. Kim, J. Kim, M. Zaheer, J. Kim, and L. W. Frédéric Chazal. PLLay: Efficient topological layer based on persistent landscapes. In *Proc. 33rd Annu. Conf. Advances Neural Info. Processing Sys. (NeurIPS)*, 2020.

[4] Q. Zhao and Y. Wang. Learning metrics for persistence-based summaries and applications for graph classification. In *Proc. 33rd Annu. Conf. Neural Info. Processing Sys. (NeuRIPS)*, pages 9855–9866, 2019.