CS 6170 Computational Topology: Topological Data Analysis **University of Utah**
Spring 2017 **School of Computing**

Lecture 25: April 11, 2017

Lecturer: Prof. Bei Wang <beiwang@sci.utah.edu> *Scribe: Qi WU*

Brief Today's lecture Prof. Bei talked about two research presentations related to Topological Data Analysis (TDA). The first presentation came from Jennifer Gamble at noodle.ai which is a company focusing on offering pioneering business solutions in Enterprise Artificial Intelligence. The second presentation came from Prof. Erica Flapan, Pomona College, Claremont, CA.

25.1 Topological Data Analysis

What is TDA

Assume we have a point cloud X inside a high dimensional space R^D, e.g. $X \subseteq R^D$, TDA studies how to represent/project the set X in a low dimensional space without losing its main features.

Ways to Describe Data Using Topological Summary

- Persistence Homology
- Mapper
- Euler Characteristic / Morse Smale Complex (will be taught soon)

Start with Data

There are two problems we must consider before doing topological data analysis:

- How to clean up the data? (since table form data is usually imperfect, including both data noises and data missing)
- How to normalize[1] the data? (since we don't always have an metric[2] intrinsically defined)

[1] Different data dimensions usually have different value scales. Data normalization means to scale values in different dimension to an unified scale such as [0, 1] or by their standard deviations

[2] The distance function to measure the similarity between points inside the dataset

Prediction Task

Currently there are two different ways for data prediction:

- (linear) regression / parametric modeling: $Y = \beta X + \epsilon$
- machine learning: $Y \rightarrow$ [redacted] $\rightarrow X$

The motivation of using TDA in prediction task is that, we want to better understand our data. However we usually need to do exploratory data analysis (because a good parameter setting is not always obvious), and define metric function (based on the distance or similarity between a collection of shapes) if the data is not a point cloud data.

Sometimes instead of applying TDA directly, we can also use TDA to interpret outputs from traditional data analyzing methods such as machine learning ($ML \rightarrow TDA$).

Example: Random Forest

Brief Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. [RF2017]

Ways to Generate Trees

- Random subset: subsets of dimensions
- Bootstrap: subset of data

When we have a random forest, we can embed the tree into a spacial domain, and use mapper to analysis the decision tree. (see slides for detailed plots)

Example: Network Analysis

Network Non-weighted and undirected graph

- node level analysis: level of impacts to the network
- intermediate level analysis:
  - community / modularity analysis [3]
  - core-periphery decomposition

[3] Modularity is one measure of the structure of networks or graphs. It was designed to measure the strength of division of a network into modules (also called groups, clusters or communities). [M2017]

How to apply TDA?

- think of the graph as a 1-skeleton.
- define higher dimensional simplicial complex in natural way.

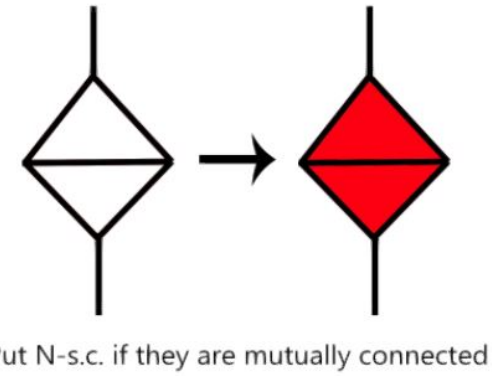

Figure 25.1: Define simplicial complex in network

- use node dominance collapse (while preserving homology).

  Definition 25.1. *if $N(v) \subseteq N(v)$, then v is dominated by w, and we can collapse edges between v and w. This process can be iterated for many times until there is no edge that can be collapsed anymore. The simplified network is called* **core-network**

  It is believed that the network flow (i.e. results from max-flow-min-cut) will not change too much after simplification. It is also possible that multiple core-networks could be obtained based on different ordering schemes.

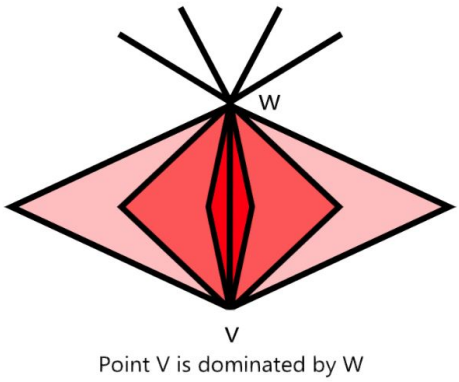

Figure 25.2: node dominance collapse

methods for using TDA on large dataset

1. sparsification[4]
2. sampling
3. parallel/distributed processing
4. sketching/approximation

25.2 Topological Complexity in Protein Structure

See research by Erica Flapan: `http://pages.pomona.edu/~elf04747/`.

References

[RF2017] WIKIPEDIA: THE FREE ENCYCLOPEDIA "Random forest", *Wikimedia Foundation, Inc*, Tue. 11 Apr. 2017.

[M2017] WIKIPEDIA: THE FREE ENCYCLOPEDIA "Modularity (networks)", *Wikimedia Foundation, Inc*, Tue. 11 Apr. 2017.

[4] Approximating a given graph by a graph with fewer edges or vertices is called sparsification