Enhancing Accessibility to Techniques in TDA

Ben Holmgren

September 15 2019

1 Introduction

Topological data analysis (TDA) is the application of techniques from topology in the computational analysis of various data. In TDA, one useful way to analyze data is by utilizing a technique known as persistent homology, which allows for the simultaneous study of homology in a topological space at multiple scales. In doing so, persistent homology can yield meaningful descriptors for a data set[5]. Recent advancements in persistent homology have necessitated an interface to apply theoretical algorithms on physical data. Due to the contributions of Dr. Fasy at Montana State and her collaborators, topological techniques- particularly those pertaining to persistent homology- are able to be applied and studied more collaboratively and efficiently with use of the R package TDA. More specifically, the R TDA package allows for the implementation of important functions in TDA, including distance and density estimators for point-cloud data, fundamental techniques in persistent homology and persistence diagram generators, bottleneck and Wasserstein distance functions, persistence landscapes and silhouettes, and bootstrap confidence sets for various estimations in TDA. Following its creation in 2013, the R TDA package has served as an important tool in TDA not only for Dr. Fasy and her collaborators, but for the wider international TDA community.

Due to the extremely broad scope of applications to TDA, a growing need for techniques in TDA is emerging across all facets of the data science community. As a result, a simultaneous need for comprehensive educational resources in TDA and its implementation has also arisen. The goal of this project is to help increase accessibility of techniques used in TDA through the R TDA package. I will work alongside both Dr. Fasy, the software's creator, and Dr. Millman, who is currently leading the development efforts of the R TDA package, to confront the challenges faced in the broader utilization of TDA.

2 Background

Thus far, the TDA package has successfully implemented an R platform for the efficient C++ libraries GUDHI, Dionysus, and PHAT. The TDA package is fully functional and available for download from CRAN [4]. It is equipped with robust vignettes and documentation built into the package. However, the R TDA package faces challenges in the onset of its application by new users. Because TDA is a uniquely applicable discipline, the algorithms implemented in the R TDA package could be made incredibly useful to virtually every imaginable academic discipline that deals with large amounts of 'point cloud' data. Furthermore, as a result of the pure math roots of topology, many of the techniques in TDA are severely underutilized simply because of the recency of developments in the computational side of topology, alongside a lack in communication from the TDA community to the rest of the data science world. We are concerned primarily in overcoming this hurdle. In order to do so, I will be working on a tutorial project which was born out of a need for improved documentation in the R TDA package last year. The tutorials are built in Jekyll, a tool which converts markdown documents into a website format, hosting them on github. The tutorials have begun with basic background knowledge in the techniques necessary to process and study data in a topological manner. I plan to expand on the tutorials for these fundamental techniques (primarily considering distance and density estimation), as well as to create a comprehensive foundation in the important filtrations used in TDA and in the study of persistent homology. Included in each tutorial for techniques in TDA will also be clear guidelines for their implementation within the R TDA package, with engaging examples.

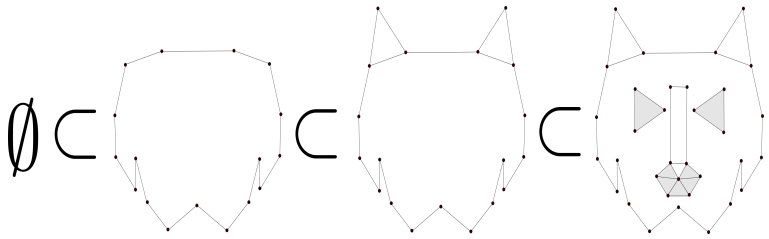

Figure 1: Bobcat shaped set of simplicial complexes from a preexisting TDA tutorial made in our group [6].

A logical progression from my previous work in the R TDA package, this project suits my needs perfectly as a developing researcher. By improving the accessibility of techniques in TDA, I am presented with an excellent opportunity to make a truly meaningful contribution to the TDA community without having taken graduate level courses in computational topology. This project will be largely oriented towards writing code, which I have the education and experience to effectively do. Strong communication skills will be of the utmost importance in creating quality materials to promote a wider understanding and application of TDA. Along with my abilities as a programmer, writing and communicating effectively are strengths of mine, and I believe that this will be a particular asset in this project. Perhaps most importantly, in the previous academic year I became familiar with nearly all of the functions in the TDA package as I worked to improve its documentation, and facilitated the beginning of the project I plan to work through in the coming academic year.

Finally, alongside being an ideal match for my skill set and simultaneously being worthwhile for the greater TDA community, my role in improving knowledge of TDA and of the R package also serves as a perfect stepping stone for my goals going forward. Few things have ever captured my imagination quite like computational topology, and looking ahead I hope to continue work in this field throughout my undergraduate and potentially graduate careers. TDA sits at the intersection of my two greatest academic passions in computer science and mathematics, and its applications are limitless. From improving road maps to correlating the patterns present in music to understanding the laws that govern the universe, topology is of unique importance to the future of humankind, and I am so incredibly excited by the prospect of taking part in that [2]. As I work on this project, I hope to gain greater knowledge of the algorithms prevalent in computational geometry and topology to allow me to continue to make meaningful contributions going forward.

3 Methods

In order to successfully create an accessible foundation for the understanding of important concepts in TDA, I will build off of the limited existing tutorials that have been created by the compTAG group at MSU using Jekyll. My schedule will be based roughly as follows, with an end goal being to create a robust network of tutorials to handle every remaining function in the R TDA package, as well as the materials necessary to give users the required background knowledge to get started.

To begin, I will build off of pre existing tutorials regarding basic functions in TDA dealing with distance and density estimation under the following timeline:
1. Fully address and finalize tutorials regarding kernel density and distance estimation: 1 week
2. Create tutorials for all functions regarding uniform shape construction: 2 weeks
3. Basic distance function: 1 week

Then, I will generate tutorials concerned with various techniques related to persistent homology. This is the bulk of the content in the R TDA package and is a central concern in TDA as a field. This will operate under the following timeframes:

4. Alpha shapes, complexes, and their filtration: 6 weeks
5. Finalize rips filtration tutorials: 2 weeks
6. Various other filtrations and persistent homology-related functions in the R package including the grid and fun filtrations and the functions related to filtration diagrams: 4 weeks

After covering the functions focused on persistent homology, I then plan to continue with persistence landscapes and silhouettes:

7. Persistence landscapes: 3 weeks
8. Silhouettes: 3 weeks

Next, I will cover the bootstrap functions included in the R package:

9. Bootstrap band: 1 week
10. Bootstrap diagram: 1 week
11. Multiplier bootstrap for persistence landscapes and silhouettes: 2 weeks

Finally, I will cover the functions pertaining to specific TDA distance measures:

12. Wasserstein distance: 1 week
13. Hausdorf distance confidence interval: 2 weeks
14. Bottleneck distance: 1 week

After following this schedule and completing tutorials for the central techniques in TDA, my work will be available online through the Montana State University compTAG group github, and will be available to anyone needing to learn about topological data analysis. I will also be setting aside time in March during the NCUR 2020 conference to hold public TDA tutorials for the many researchers who will be on MSU's campus. This should serve as an important opportunity both for myself to debut TDA tutorials on a relatively large scale, and for a wide selection of bright minds from around the country to learn about TDA and how computational topology could be of use in their respective fields.

4 Collaboration With Faculty Sponsor

Created by Dr. Fasy and her collaborators in order to compute increasingly complex topological data in a manner accessible for the wider TDA community, the R TDA package remains closely tied to Dr. Fasy's work and I will work with her to most effectively improve its documentation. Furthermore, with this project I am lucky enough to collaborate directly not only with Dr. Fasy but with Dr. Millman as well, whose work in TDA shares many similar interests and who has a wealth of experience in software development. I will work closely with both faculty, and will be participating in weekly seminars which address the important work in topology being done at Montana State and throughout the collective community. I will be presenting or co-presenting at least twice in these weekly seminars, and potentially conducting further supplementary research in

open problems in TDA and computational geometry as opportunities and new interests arise. Along with this, I will join group work sessions where both mentors will be available for questions. Otherwise, I plan to meet once each week with Dr. Millman and Dr. Fasy or as often as needed. In these meetings, we will discuss my progress, challenges I come across, the most effective methods for me to complete my project, and if necessary we will also be able to revisit my schedule in these meetings. As an end goal, I am seeking to improve collective knowledge of TDA and its implementation in the R package because of the massive potential that TDA has as a revolutionary tool in countless areas of study. Enhanced accessibility in TDA and to the R package will create a solid foundation for the betterment of Dr. Fasy and Dr. Millman's work as well as for the overall TDA community. With a multitude of exciting applications and new ideas originating constantly in TDA, the disconnect between its pure math foundations and the rest of the academic world must be overcome. I plan to do my part in realizing such a large task with this project throughout the academic year.

References

[1] Edelsbrunner, Herbert, Harer, John (2010). *Computational Topology: An Introduction.* Retrieved from *https://www.researchgate.net/publication/220692408ComputationalTopologyAnIntroduction*

[2] Fasy, B T (2017, August). *Research Statement.* Retrieved from *https : //www.cs.montana.edu/brittany/research/fasy brittany rsrch stmt.pdf*

[3] Fasy, B T, Kim, J, Lecci, F, Maria, C, Millman, D L, Rouvreau, V (2018). Introduction to the R package TDA. `The Comprehensive R Archive Network,` 1-24.

[4] Kim, J. (2018, August 6). *Statistical Tools for Topological Data Analysis.* retrieved from *https://www.rdocumentation.org/packages/TDA/versions/1.6.4*

[5] Weinberger, S.(2011). What is... Persistent Homology?. *American Mathematical Society,* 58(1), 36-39.

[6] compTAG at Montana State. "CompTAG/rpackage$_t$*utorials*." *GitHub*, 5*Aug*.2019, *github.com/compTAG/r*

[7] Jekyll. "https://jekyllrb.com/." *Jekyll*, 2019, https://jekyllrb.com/.