Updating the R Package TDA

Ben Holmgren

September 16, 2018

1 Introduction

In the field of topological data analysis (TDA), one useful way to analyze topological 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 lead to the detection of high persistence features in a data set [7]. Recent advancements in persistent homology have necessitated an interface between the wider topological community and the algorithms needed to continue progress in the field. Due to the contributions of Dr. Fasy at Montana State and her collaborators, persistent homology is able to be 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 the distance function, the distance to a measure function, the k-nearest neighbor density estimator, the kernel density estimator, and the kernel distance. 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. While a vital resource, the R TDA package could be greatly improved. Increasing use of the R TDA package has revealed that the package lacks important documentation and often doesn't adequately account for gaps in theoretical and technological capabilities of users. Regularly, users of the package possess a sophisticated theoretical understanding of topology but are limited by lagging abilities in computation which slows the contribution of new discoveries in the field.

The goal of this project is to help bridge the gap between computational and theoretical abilities of R TDA package users by adding important documentation to the package to streamline its use and to clarify its application. 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 and has extensive experience leading large development efforts throughout his career. Successful completion of the project will allow for greater progress in TDA by making the technology a useful asset rather than a potential distraction from the important research taking place. Furthermore, I also plan to streamline the addition of new algorithms to the package by creating the infrastructure to check additional code for bugs before it is uploaded. Creation of these testing mechanisms will greatly improve the viability of the package going forward as developments continue to be made in TDA and new algorithms are developed [2].

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. However, only minimal documentation is available to the public for use of the R TDA package [4]. As the TDA package only grows in relevance, shortcomings in documentation become increasingly harmful. Currently, the TDA package is experiencing 197 direct monthly downloads which is a good indicator of its importance to the TDA community. As new users continue to implement the package, the importance of proper documentation only continues to grow [3].

This project is perfectly suited to my strengths as a developing researcher. In fact, by improving the R TDA package, I am presented with an excellent opportunity to make a truly meaningful contribution to the field of TDA without having 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. Furthermore, I believe that my ability to quickly learn and apply new concepts in computer science will allow me to quickly learn and understand R, and will be crucial for the success of this project. Strong communication skills will be of the utmost importance in creating quality documentation and improving the user experience in the TDA package. Along with my abilities as a programmer, writing and communicating effectively are particular strengths of mine, and I believe that this will be a particular asset in this project.

Finally, alongside being an ideal match for my skill set and simultaneously being worthwhile for the greater TDA community, my role in improving the TDA 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 even graduate career. 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 both the proficiency in R and the theoretical background knowledge in topology to allow me to further my understanding of topology and to continue to make meaningful contributions going forward.

3 Methods

In order to most effectively improve user experience in the R TDA package, I will need to add documentation to explain the implementation of important functions. I will accomplish this by first gaining a background in R [5]. I will refer to R coursework from John's Hopkins University to do so, using the following timeline as a guide:

1. Data Types and Basic Operations: 9/16 to 9/17
2. Subsetting: 9/18 to 9/19
3. Reading and Writing Data: 9/20 to 9/22
4. Control Structures and Functions: 9/23 to 9/25
5. Scoping Rules and Debugging Tools: 9/26 to 9/29
6. Simulation and R Profiler: 9/30 to 10/02

Following my establishment of a foundation in R, I will need to understand the integration of C++ and R to be effective in my project, since the R TDA package is mostly written in C++ [8], [6]. To do so, I will cover the Rcpp library and C++ templates under the timeline:

7. Simplifying and integrating C++ and R with Rcpp library: 10/03 to 10/06
8. C++ templates 10/07

With a necessary background in R and C++ established, I will begin adding documentation to the R TDA package. I will first address this by adding documentation to the fundamental TDA functions in the package. Conveniently, beginning with these functions will allow me to systematically understand the actual role of each base function before I attempt to add documentation to more abstract algorithms. I will add documentation to the fundamental TDA functions in the timeline:

9. The Distance Function and Distance to Measure Function: 10/07 to 10/22
10. The k-nearest neighbor density estimator, Gaussian Kernel Density Estimator, and Kernel distance estimator: 10/22 to 11/21

After adding documentation to each of the major functions in the TDA package, I then plan to add documentation to functions which specifically apply to persistent homology in the timeline:

11. The gridDiag function and the ripsDiag function: 11/21 to 12/14
12. The alphaComplexDiag function and the alphaShapeDiag function: 01/09 to 02/05
13. The ripsFiltration, alphaComplexFiltration, and alphaShapeFiltration functions: 02/05 to 02/19
14. The bottleneck and wasserstein wrappers: 02/19 to 03/01
15. landscape and silhouette functions: 03/02 to 03/16
16. multipBootstrap and maxPersistence functions 03/01 to 03/11
17. maxPersistence: 03/11 to 03/21

Once I complete this fourth phase of my project, I will finally add documentation to the areas of the R package which concern Density Clustering. I plan to spend an additional two weeks on this phase, which will take place in the timeframe:

18. clusterTree function: 03/21 to 04/10

Finally, with the documentation complete for each of the functions included in the TDA package, I will spend the remainder of my time adding tools for debugging additional code which will be added to the package, as well as preparing a poster to present my work in the MSU Student Research Celebration. Both tasks will be completed simultaneously in the timeline:

19. Provide mandatory testing of code before uploading it to the R package: 04/10 to 05/01
20. Create a poster to present my work: 04/10 to 04/17

Expected Project Outcome

Upon completing my improvements to the R TDA package, I expect to have added documentation for all of the aforementioned functions which will be available both on the official R documentation website and in R to be called upon by the user when needed. Along with this, infrastructure will be available to test the functionality of new code being uploaded to the package. Both projects will be important strides in making the R TDA package more accessible to researchers in the TDA community.

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. Along with this, I will join group work sessions where both mentors will be available for questions, and otherwise 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. Lastly, I plan to attend weekly sessions created by Dr. Millman for anyone involved in topology to code in R, at which I will be able to ask further questions and collaborate with other TDA researchers at MSU.

As an end goal, I am seeking to improve the R TDA package because it remains of the utmost importance for the work of both Dr. Fasy and Dr. Millman as well as the international TDA community. With a multitude of ongoing projects in TDA, the R package is a vital tool to ensure continued progress. My work in improving the package will make these projects and the influx in additions to the package sustainable. I will work towards much greater clarity in the function of the package, while simultaneously improving the mechanisms for its growth so that the work of my sponsors and their collaborators can be more effectively centered on TDA going forward.

References

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

[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] R Programming Lecture Materials. (2018). Retrieved from http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/rprog/coursePage/lectureNotes/

[6] Templates. (2000-2017). Retrieved from http://www.cplusplus.com/doc/oldtutorial/templates/

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

[8] Wickham, Hadley. *High performance functions with Rcpp.* Retrieved from http://adv-r.had.co.nz/Rcpp.html