

STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-2125

GUNNAR CARLSSON
ANN AND BILL SWINDELLS PROFESSOR
DEPARTMENT OF MATHEMATICS
(650) 723-2224
gunnar@math.stanford.edu

October 24, 2017

**Report on *Sur les propriétés métriques et statistiques des descripteurs
topologiques pour les données géométriques* by M. Carrière**

This dissertation studies a very interesting and important problem within topological data analysis. I will first summarize the context and then discuss the contributions made in the present dissertation.

Topological data analysis is a relatively new area of inquiry within mathematics and computer science. It has been underdevelopment over the last 15-20 years. Its goal is to provide methods for understanding the “shape” of data sets, suitably interpreted. A useful way to define a shape on a finite set of points is through a *metric* or *distance function*, which is typically encoding information about similarity of data points. There have been two separate directions of inquiry within the subject.

- **Measuring shape:** The field of algebraic topology within pure mathematics defines algebraic invariants of shapes (perhaps very high dimensional) that capture the presence of gross features, such as loops, spheres, etc.. There are variants that capture other things, such as the presence of flares etc.. One of the achievements in the subject has been to extend these methods (which apply to situations where we have complete information about the shapes that we are dealing with) to situations where we only have information about a sample from an underlying space. This extension is referred to as *persistent homology*. While the output of standard

homology consists of numbers (so-called Betti numbers), which roughly count the instances of certain patterns occurring in the shape), the new methodology produces outputs that are called bar codes, or persistence diagrams. They consist of collections of intervals, each capturing a particular feature, which an extent that is indicative of the scale of the feature. This more continuous representation allows one to infer the presence of these features in an underlying space from which the points have been sampled. Many variants of these constructions have been developed, and have been found to be extremely useful for many different scientific problems.

- **Representing shape:** It is also useful to deal directly with the shape of the data set. A natural class of representations of shapes consists of *simpli-
cial complexes*, which are collections of points, edges, triangles, tetrahedra, and higher dimensional analogues, with suitable intersection properties. The simplest examples are graphs, in the computer science sense. One of the main threads in ordinary topology is the approximation of spaces by such complexes. Indeed, this is essential for the definition and computation of the homological invariants mentioned above. A family of methods which has long been used in data analysis is described as *cluster analysis*. In this case, the complexes have only points, with no edges or higher dimensional objects. There are many constructors for complexes based on distance functions, including Čech, Vietoris-Rips, alpha shapes, witness complexes, etc.. However, many of these create very high dimensional complexes, which is very problematic for dealing directly with the shape. There is one method, Mapper, which explicitly controls the dimension of the complex, and it has been found to be extremely useful for many applied problems. It has been used to discover the decomposition of type 2 diabetes and asthma into smaller different disease forms, which will lead to much more targeted methods for treatment. It has also been used to map the state space of subjects undergoing infection by a particular disease,

allowing for different approaches to the understanding of the function of the immune system. There are many other examples. Once constructed, the Mapper model complex can be laid out on the screen using standard layout algorithms, and it is possible to interact with it in numerous ways, including the selection of subsets, the encoding by coloring of function values on the data set, finding the variables that characterize the various regions in the data set, etc.

One of the problems of the Mapper construction is that it involves parameters and can sometimes exhibit instability with respect to the choice of these parameters. This dissertation is a step toward the solution of this problem. It uses methods from both threads above. Specifically, it does four things.

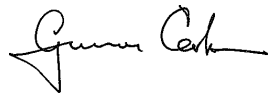
1. It constructs a pseudometric on objects closely related to the Mapper construction, namely *Reeb graphs*. Mapper (in its one-dimensional form) is a discretized version of Reeb graphs. The pseudometric is obtained by computing persistence diagrams for the Reeb graph, and using known distance functions on the set of persistence diagrams. The reason it is a pseudometric is that it is quite possible for somewhat different Reeb graphs to have identical persistence diagrams. However, it is understood that large changes in the Reeb graph will be captured in the persistence diagrams.
2. The relationship between Mapper (applied to a space) and its idealized version, a Reeb graph, is analyzed, and this analysis enables the extension of the pseudometric for Reeb graphs to Mapper outputs. This allows for the proof of a certain kind of stability theorem for Mapper, which is a very important contribution. This is an excellent result which has appeared in the proceedings of the Symposium on Computational Geometry, the premier computer science conference for computational geometry.
3. Prove a convergence theorem for Mapper applied to point clouds within

a space to the Mapper construction applied to the full space. This “completes” the circle, and permits the statistical analysis of the construction, in terms of various kinds of features.

4. Analyze kernel methods for the space of persistence diagrams. This is important in that it will make possible better inference methods for the study of persistence diagrams, and therefore of shapes.

This dissertation is an outstanding piece of work. It is clearly of theoretical importance, and I think is also extremely likely that it will be of importance in the applied world. It may over time suggest methods for modifying the construction so that it will be stable, and will give quantifiable methods for understanding the presence of features complex data sets. These are both very desirable goals, in that they will allow users to more quickly gain confidence in the results they observe. I strongly recommend its acceptance.

Sincerely yours,

A handwritten signature in black ink, appearing to read "Gunnar Carlsson". The signature is fluid and cursive, with a large initial 'G' and a distinct 'C'.

Gunnar Carlsson

Ann and Bill Swindells Professor, Emeritus

Department of Mathematics

Stanford University