University of Oxford

DEPARTMENT OF
STATISTICS

Topological Data Analysis of Temporal Networks

by

Dimitri Lozeve

St Catherine's College

A dissertation submitted in partial fulfilment of the degree of Master of Science in Applied Statistics

Department of Statistics, 24–29 St Giles,
Oxford, OX1 3LB

September 2018

Declaration of authorship

This my own work (except where otherwise indicated).

Date Signature

Abstract

Abstract here

Acknowledgements

Thank you!

Contents

List of Figures

List of Tables

1 Introduction

2 Topological Data Analysis and Persistent Homology

2.1 Homology

Our goal is to understand the topological structure of a metric space. For this, we can use *homology*, which consists in associating for a metric space X and a dimension i a vector space $H_i(X)$. The dimension of $H_i(X)$ will give us the number of i-dimensional components in X: the dimension of $H_0(X)$ is the number of path-connected components in X, the dimension of $H_1(X)$ is the number of holes in X, and the dimension of $H_2(X)$ is the number of voids.

Crucially, these vector spaces are robust to continuous deformation of the underlying metric space (they are *homotopy invariant*). However, computing the homology of an arbitrary metric space can be extremely difficult. It is necessary to approximate it in a structure that would be both combinatorial and topological in nature.

2.2 Simplicial Complexes

In order to understand the topological structure of a metric space, we need a way to decompose it in smaller pieces which, when assembled, conserve the overall organisation of the space. For this, we use a structure called a *simplicial complex*, which is a kind of higher-dimensional generalization of graphs.

The building blocks of this representation will be *simplices*, which are simply the convex hull of an arbitrary set of points. Examples of simplices include single points, segments, triangles, and tetrahedrons (in dimensions 0, 1,, 2, and 3 respectively).

Definition 2.1 (Simplex)**.** The *k-dimensional simplex* $\sigma = [x_0, \dots, x_k]$ is the convex hull of the set $\{x_0, \dots, x_k\} \in \mathbb{R}^d$, where $x_0, \dots, x_k$ are affinely independent. $x_0, \dots, x_k$ are called the *vertices* of σ, and the simplices defined by the subsets of $\{x_0, \dots, x_k\}$ are called the *faces* of σ.

We then need a way to combine these basic building blocks meaningfully so that the resulting object can adequately reflect the topological structure of the metric space.

Definition 2.2 (Simplicial complex)**.** A *simplicial complex* is a collection K of simplices such that:

- any face of a simplex of K is a simplex of K
- the intersection of two simplices of K is either the empty set or a common face or both.

Using these definitions, we can define homology on simplicial complexes.

2.3 Filtrations

If we consider that a simplicial complex is a kind of "discretization" of a metric space, we realise that there must be an issue of *scale*. For our analysis to be invariant under small perturbations in

the data, we need a way to find the optimal scale parameter to capture the adequate topological structure, without taking into account some small perturbations, nor ignoring some important smaller features.

The ideal solution to these problems is to consider all scales at once: this is the objective of *filtered simplical complexes*.

Definition 2.3 (Filtration)**.** A *filtered simplicial complex*, or simply a *filtration*, K is a sequence $(K_i)_{i\in I}$ of simplicial complexes such that:

- for any $i, j \in I$, if $i < j$ then $K_i \subseteq K_j$,
- $\bigcup_{i\in I} K_i = K$.

2.4 Persistent Homology

We can now compute the homology for each step in a filtration. This leads to the notion of *persistent homology*, which gives us all the information necessary to establish the topological structure of the metric space at multiple scales.

Definition 2.4 (Persistent homology)**.** The *p-th persistent homology* of a simplicial complex $K = (K_i)_{i\in I}$ is the pair $(\{H_p(K_i)\}_{i\in I}, \{f_{i,j}\}_{i,j\in I, i\leq j})$, where for all $i \leq j$, $f_{i,j} : H_p(K_i) \mapsto H_p(K_j)$ is induced by the inclusion map $K_i \mapsto K_j$.

The functions $f_{i,j}$ allow us to link generators in each successive homology space in the filtration. Since each generator correspond to a topological feature (connected component, hole, void, etc, depending on the dimension p), we can determine whether it survives in the next step of the filtration. We can now determine when each feature is born and when it dies (if it dies at all). This representation will be dependent on the choice of basis for each homology space $H_p(K_i)$. However, by the Fundamental Theorem of Persistent Homology, we can choose base vectors in each homology space such that the collection of half-open intervals is well-defined and unique. This construction is called a *barcode*.

2.5 Topological summaries: barcodes and persistence diagrams

In order to interpret the results of the persistent homology computation, we need to compare the output for a particular data set to a suitable null model. For this, we need some kind of a similarity measure between barcodes and a way to evaluate the statistical significance of the results.

One possible approach for this is to define a space in which we can project barcodes and study their geometric properties. *Persistence diagrams* are an example of such a space.

Definition 2.5 (Persistence diagrams)**.** A *persistence diagram* is the union of a finite multiset of points in $\bar{\mathbb{R}}^2$ zith the diagonal $\Delta = \{(x, x) \mid x \in \mathbb{R}^2\}$, where every point of Δ has infinite multiplicity.

The diagonal Δ is added to facilitate comparisons between diagrams, as points near the diagonal correspond to short-lived topological feature, thus likely to be caused by small perturbations in the data.

We can now define several distances on the space of persistence diagrams.

Definition 2.6 (Wasserstein distance)**.** The *p-th Wasserstein distance* between two diagrams X and Y is

$$W_p[d](X, Y) = \inf_{\phi: X\mapsto Y} \left[\sum_{x\in X} d\left(x, \phi(x)\right)^p\right]$$

for $p \in [1, \infty)$, and

$$W_\infty[d](X, Y) = \inf_{\phi: X \mapsto Y} \sup_{x \in X} d\left(x, \phi(x)\right)$$

for $p = \infty$, where d is a distance on $\mathbb{R}^2$ and ϕ ranges over all bijections from X to Y.

Definition 2.7 (Bottleneck distance)**.** The *bottleneck distance* is defined as the infinite Wasserstein distance with d the uniform norm: $d_B = W_\infty[L_\infty]$.

Since the bottleneck distance is by far the most commonly used, we will focus on it in the following. It is symmetric, non-negative, and satisfies the triangle inequality. However, it is not a true distance, as it is fairly straightforward to come up with two distinct diagrams at bottleneck distance zero, even on multisets not touching the diagonal Δ.

2.6 Stability

3 Temporal Networks