A Discussion on the Effectiveness of TDA on Prediction with Application to Professional Counter-Strike Teams

Keaton Asbach

December 2024

0.1 Introduction

Topological Data Analysis (TDA) is an emerging and versatile tool in the field of data analysis, offering unique ways to understand complex data structures. As a growing branch of mathematics, TDA has already demonstrated its value in diverse applications, such as capturing intrinsic clustering in data sets, enabling better stratification, and automated prediction of manufacturing productivity. Methods like Persistent Homology make TDA not only insightful but also efficient, often saving time and resources compared to traditional approaches. This essay explores the practical utility of TDA in prediction and decision-making, assessing whether it provides a more effective framework than conventional data analysis methods for deriving actionable insights. We will also apply TDA to a certain data set in an attempt to predict the leading Counter-Strike "Team of the Year".

0.2 Mapper Algorithm

To begin, we will explore a specific TDA algorithm known as the "Mapper Algorithm", and its use case in automatic prediction of manufacturing productivity. The goal is to translate the data into an interactive, graphical representation, which enables us to find new patterns and connections in said data. Wei Guo and Ashis G. Banerjee [GB17] describe the problem as being "data rich but information poor", in an effort to justify TDA's use in this scenario. Before applying this mapper algorithm, we must formulate the problem mathematically. The way in which this is achieved is as follows: "we suppose there are m process variables and N measurements. Each measurement is, thus, represented by am $m-$dimensional vector $x_i \in \mathbb{R}^m, i = 1, 2, \ldots, N$. The data is then assembled into a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times m}$. Each column denotes a process variable, which is measured by one sensor operating alone or by a synergistic operation of a few sensors. The latter case is known as data fusion, which provides a wide sensing range of parameters, and is, hence, more reliable for data analysis. For each row, the measurement is either spatially-sampled or temporally-sampled. For instance, in the semiconductor manufacturing environment, electronic wafer map data collected from inline measurements are sampled spatially across the surface of the wafer for defect inspection." Having formulated the problem mathematically, we can move to understanding the Mapper Algorithm and interpreting its results.

Given a data matrix $\mathbf{X}$, the setup of the mapper algorithm goes as follows:

1. Set resolution parameters: a number of intervals l and overlap percentage p, where $p \in (0, 100)$.

2. Compute the pairwise distance matrix $\mathbf{D} = [d(\mathbf{x}_i, \mathbf{x}_j)] \in \mathbb{R}^{N \times N}$ based on the distance metric chosen.

3. Select a filter function $f : X \to \mathbb{R}^n$ to stratify the data, where X is our topological space.

There are a number of filter functions to guide a clustering algorithm, namely the Gaussian kernel density estimator, eccentricity filter, and eigenvectors of graph Laplacians to name a few. Additionally, we could take the projection found by dimensionality reduction techniques as the filter function. In [GB16] the "choice of filter function is the 2-D projection found by the multidimensional scaling (MDS) method. MDS in this case attempts to embed the data such that the pairwise distances in the high-dimensional space are preserved in the 2-D Euclidean space. Accordingly, the 2-D embedding coordinates denoted by $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_N$, are the minimizers of a loss function, σ, defined as"

$$\sigma(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_N) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} (\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|_2 - d(\mathbf{x}_i, \mathbf{x}_j))^2$$

After formulating the problem and attacking the data with our filter function, we begin the data processing and prediction models. As a result, a topological network in the form of a simplicial complex can be recovered as seen in 1, with each node representing a cluster and edges indicating at least one measurement in common. From here, more intricate and valuable insights can be drawn (namely the subgroups circled in 1), enabling us to form better inferences and actionable insights courtesy of the data.

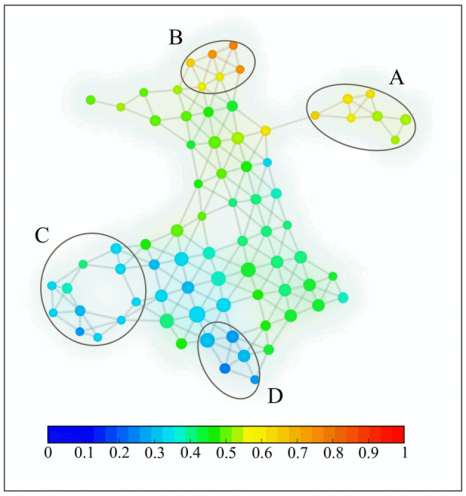

Figure 1: 2-D Embedding

0.3 Applying TDA to Counter-Strike Team Data

In my study of TDA, I had originally wanted to apply what I had learned to some data set, however I had struggled to get the software (Ripser) working correctly. In an effort to perform this data analysis, I searched online until I came across this website (live.ripser.org), able to compute persistence barcodes, which is exactly what I was looking for. The data I had gathered is akin to athlete/player data, for a specific video game called Counter-Strike. HLTV.org has a list of Top 10 teams of each year going back to 2018, and I wished to analyze data from last year to see if there was a trend or commonality between last years' Top 1 team and the data we have from this year. My goal is to predict who will be crowned the best team of 2024.

The way in which I hoped to achieve and compare TDA between these different teams is by creating persistence barcodes. In short, a persistence barcode is a way to visualize the homology groups of our data set – in other words, giving us a way to see the number of connected components, holes, etc. Using live.ripser.org, I was able to create said persistence barcodes for 5 teams using each of their 5 players data (#1 from 2023, along with current standing #1, #2, #3, and #4 of this year). My goal was to see if I could find a matching barcode from last years #1 team in this years team data, or at least some similarity amongst the barcodes. I wished to mathematically quantify the difference between each barcode, so the method I used was to treat the length of each persistence interval as an observation with 5 coordinates, then calculate the distance between 2023's #1 Team vs. the collected teams for 2024, to see if there was a team that stood out as being relatively "close" to 2. It would follow that a smaller value may indicate a closer correlation, hence enabling us to determine our closest match. I'll walk through the calculations for just one comparison, namely

Team Vitality 2023 vs. Natus Vincere 2024

$$48.4379 = \sqrt{(141.183 - 124.526)^2 + \cdots + (47.4507 - 28.6551)^2}$$

Our remaining 3 teams compared against Vitality in 2023 are as follows: G2 = 22.1321, Vitality = 18.5422, Spirit = 43.6299. Using this method, we are able to see how Team Vitality in 2024 appears to be our front runner for best team of the year.

0.3.1 Persistence Barcodes

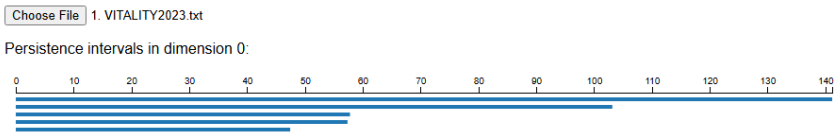

Figure 2: #1 Team of 2023 - Vitality

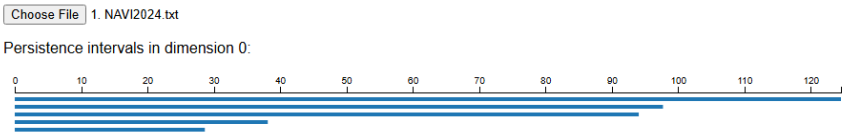

Figure 3: Current #1 Team in the world - Natus Vincere

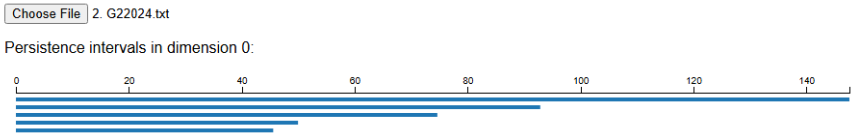

Figure 4: Current #2 Team in the world - G2 2024

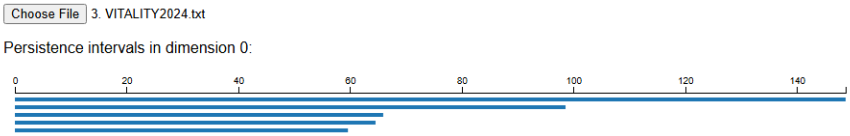

Figure 5: Current #3 Team in the world - Team Vitality

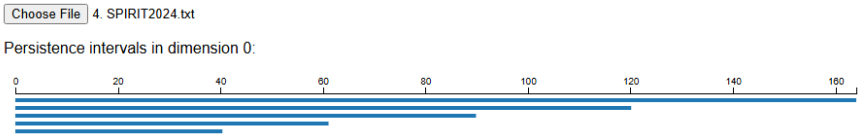

Figure 6: Current #4 Team in the world - Team Spirit

0.4 Shortcomings and non-TDA analysis

After analyzing what has been computed, it's perhaps no surprise that we found Team Vitality from 2024 to be most akin to Team Vitality from 2023. Despite their numerous roster changes, keeping a core 3/5 players, in-game system, teamwork, and support staff, lends itself to being quite similar. I would however like to discuss how this can guide our prediction, and with the

use of non-TDA insights, help us to view the scene holistically and come to our anticipated conclusion.

There are a number of important factors not included in our analysis, namely Big Tournament Placings, Prize Money, and length of time in the Top 10 just to name a few. Taking a look at these, coupled with our calculated "Ratings" for each team to guide our decision making process, we can see a couple teams begin to stand out amongst the rest.

1. Natus Vincere: 48.4379 "Rating", 4 S-Tier trophies, $1,622,500 in prize money, and peaking at #1 in the world for a staggering 18 weeks.
2. G2: 22.1321 "Rating", 3 S-Tier trophies, $1,331,000 in prize money, staying consistently around #5, peaking at #2 for the last number of months.
3. Team Vitality: 18.5422 "Rating", 1 S-Tier trophy, $904,000 in prize money, peaking at #1 at the beginning of the year, staying consistent around #3.
4. Team Spirit: 43.6299 "Rating", 3 S-Tier trophies, $1,419,500 in prize money, peaking at #1 for a month, hovering consistently around #2-4.

0.5 Who is the best team of 2024?

Taking everything into account, the race for the #1 Team of 2024 seems to be a two horse race between Natus Vincere and G2. Truthfully, it comes down to how you value trophies, prize money, and all the other factors that can go into determining a teams success. The players also play a huge role, as by consensus, G2 has 2 of the top 5 most individually strong players, which helps bring their team into the conversation for best in the world. Natus Vincere on the other hand, is a more well rounded unit, with every player contributing significantly in their own way. To personally reach a conclusion for best team of the year, I think it strongly depends on the ongoing Perfect World Shanghai Major. The winner will receive 1 S-Tier trophy, $500,000 USD, and will likely close out the year as the #1 team. Since Natus Vincere has been eliminated from the tournament, the choice is clear that G2, should they win the major, will be the best team of 2024. In the case they don't win, and Spirit or Vitality win, they have a strong stake to the #1 Team of the Year.

0.6 Conclusion

TDA serves as a powerful tool for interpreting complex data, acting more as a translator than a source of direct conclusions or predictions. By transforming messy, noisy, and intricate data sets into more comprehensible forms, it bridges the gap between raw data and actionable information. While TDA itself doesn't directly yield conclusions, it simplifies the process, saving time and effort for its users. This makes it an invaluable resource for extracting meaningful insights and improving efficiency of predictions.

Bibliography

[GB16] Wei Guo and Ashis G. Banerjee. Toward automated prediction of manufacturing productivity based on feature selection using topological data analysis. In *2016 IEEE International Symposium on Assembly and Manufacturing (ISAM)*, pages 31–36, 2016.

[GB17] Wei Guo and Ashis G. Banerjee. Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs. *Journal of Manufacturing Systems*, 43:225–234, 2017. High Performance Computing and Data Analytics for Cyber Manufacturing.