Project Proposal: Identifying Defensive Coverage Gaps in NFL Tracking Data Using Topological Data Analysis

Arjun Mahesh

1 Motivation / Research Question

I would like to study defensive pass coverage patterns in American football through TDA. In a passing play, defenders' positions on the field create spatial patterns. Defender = A player on the team defending against the offense, attempting to prevent successful passes and gains. Coverage = The defensive strategy and positioning used to defend against pass plays. Defenders position themselves spatially to cover receivers and prevent completions. Some coverages leave "holes" or open spaces, while others are tight. My research question is: *Can persistent homology quantify these coverage gaps, and do such topological features correlate with play outcomes (completed or incomplete passes)?* Understanding defensive coverage is important to team strategy and TDA can capture the shape of complex data like player formations. This article describes how TDA was used to cluster basketball players into 13 "new" positions, finding groupings that traditional positions missed. This paper shows how persistent homology on team roster data found that certain topological features of player skill distribution correlated with better team offense. In soccer, a study used TDA for soccer scouting and showed that it can capture nonlinear relationshpis in player data that conventional clustering might miss.
Defensive coverage data, the on-field coordinates of 11 defenders (and receivers), is inherently geometric. Traditional stats like "yards of separation" reduce this to a single nummber, whereas persistent homology can quantify the shape of the entire formation, identifying multiple disconnected groups or a large void in coverage. A "hole" in a defender arrangement is a 1-dimensional homology feature, indicating an open passing lane. TDA provides a systematic way to detect and measure such features across scales. It could offer coaches a novel quantitative view of where and when coverage breaks down, complementing existing metrics.

2 Data and Resources

I will leverage the NFL Big Data Bowl 2021 dataset which provides detailed tracking for all passing plays in the 2018 NFL regular season. This public dataset (available via Kaggle) includes the x,y coordinates of every player (and the football) at 10 frames per second throughout each play. It also comes with play-level context (down, distance, outcome, etc.). The size is substantial (over 17,000 pass play sequences, from 253 games), but it is well-structured in CSV files by week. I will also use relevant resources for reference and validation. For example, if available, I might use game film or diagrams to qualitatively interpret what certain persistent homology features correspond to (though primary analysis will be data-driven). Additionally, I will consult prior literature for guidance (e.g. definitions of zone vs. man coverage formations) to help inform our interpretation of results.

3 Methods

My approach centers on computing and analyzing persistent homology of the defensive coverage formations.

3.1 Point Cloud Representation

For each passing play (at the moment of ball release), we represent the positions of defensive players (and possibly receivers) as a set of points in $\mathbb{R}^2$ (the field plane). This point cloud encodes the spatial structure of the coverage. We may include only defenders, or defenders and intended receiver, depending on what yields more insight (this will be experimented with).

3.2 Persistent Homology Calculation

Using a Vietoris–Rips filtration on each point cloud, we will compute persistent homology (primarily in dimensions 0 and 1). Dimension 0 (H_0) features correspond to connected components—essentially clusters of players. Dimension 1 (H_1) features correspond to loops or holes in the coverage. We anticipate H_1 is especially relevant as an indicator of an open gap in the defense formation. We will employ Python libraries such as Ripser (ripser.py) or GUDHI to efficiently compute persistence diagrams for each play's point set. The result for each play will be a persistence diagram (or barcode) summarizing the topological features of that defensive formation across different spatial scales.

3.3 Topological Feature Extraction

We will extract quantitative summaries from the persistence diagrams to facilitate comparison and modeling. For example, we might record the number of significant H_1 features and their lifetimes (how persistent the coverage holes are). We can also compute Bottleneck or Wasserstein distances between diagrams of different plays to measure similarity in coverage topology. If needed, we might vectorize diagrams into persistence images or landscapes for statistical analysis.

3.4 Comparative Analysis

We will investigate how these topological features relate to outcomes and strategies: **Correlation with play outcome:** We will group plays by whether the pass was completed or not (and possibly big gains vs. short gains) and compare their persistence features. Our hypothesis is that successful pass plays often coincide with a prominent persistent hole in the defense (i.e., one large H_1 class), whereas incompletions or interceptions might show tighter coverage (no large holes). We can use statistical tests or a simple classifier to see if topological features have predictive power for completion. For example, a logistic regression could use the size of the largest H_1 bar as a feature to predict completion. **Cluster analysis of coverage patterns:** Using the distances between persistence diagrams, we will perform clustering or dimensionality reduction on the plays. This could reveal distinct types of coverage shapes. We might discover, for instance, one cluster of plays where two separate defender clusters (H_0) indicate double coverage on multiple receivers, versus another cluster where defenders form a ring-like formation leaving a central hole (high H_1). We will examine if these clusters align with known defensive strategies (e.g., man-to-man coverage might produce many small clusters each around a receiver, whereas zone coverage might produce a loop enclosing an area). **Team or Player tendencies:** If time permits, we will aggregate results per team or per defensive unit. For example, some teams might consistently produce certain topological

signatures (like always one big hole, or none at all). We can compute average persistence features for each team's plays to see if that correlates with defensive effectiveness rankings. This could yield insights like "Team X often left a large coverage gap (loop) which opponents exploited, reflected in a high average H_1 lifespan for their plays." Such insight could be valuable for scouting.

3.5 Software and Tools

My implementation will be in Python. I will use simple libraries including Pandas/Numpy for data handling, scikit-learn for clustering or classification, and Ripser/GUDHI (with possibly the Giotto-TDA framework) for computing persistent homology and creating persistence diagrams. Visualization libraries (matplotlib/plotly) will help in plotting barcodes and any Mapper graphs if we explore that. Throughout the method development, I will validate my approach on toy examples (for instance, constructing a contrived formation of points with a known hole to ensure our pipeline detects the hole). I will also be mindful of computational complexity. If computing diagrams for all $\sim$17k plays is too slow, we might use random subsampling or focus on a single week to iterate, then scale up once optimized.

4 Expected Outcomes

By the end of the project, I expect to have the following outcomes and findings:

4.1 Visualization of Topological Features

Persistence diagrams and barcodes that visually illustrate the coverage patterns. For example, we might show a typical persistence diagram for a play with a blown coverage (where we expect a long-lived H_1 feature) versus a diagram for a well-covered play (showing only short or no H_1).

4.2 Quantitative Insights

i anticipate discovering that plays with successful passes often have noticeable topological "holes" in the defender formation. If this hypothesis holds, the persistence diagrams of completed passes will frequently contain an H_1 point far from the diagonal (i.e., a long persistence), corresponding to a durable open gap. In contrast, unsuccessful passes might have only short-lived H_1 features or none at all, indicating tighter coverage. We will quantify this by comparing metrics like the maximum H_1 bar length across groups. We expect a statistically significant difference, supporting the idea that coverage topology influences play success.

4.3 Clustered Coverage Types

I hope to identify and describe a few archetypal coverage patterns. For example, we might label one cluster of plays as "Tight Man Coverage" (characterized by many connected components in H_0, each around a receiver, and no H_1 holes) and another as "Zone Coverage with Gap" (characterized by defenders forming a perimeter with a clear void in the middle, i.e., one significant H_1). These findings can be cross-checked with actual play annotations if available (to see if they correspond to known defensive calls).

5 Timeline

Oct 19-20 Background research on TDA methods and review of the NFL Big Data Bowl 2021 dataset
Oct 20-27 Data preprocessing and initial construction of point clouds and filtrations
Oct 28-Nov 8 Compute persistence diagrams; begin exploratory analysis of defensive shape
Nov 9-15 Analyze topological features against defensive outcomes; develop comparison metrics (e.g., bottleneck distances)
Nov 16-22 Visualizations, clustering, and any classification modeling
Nov 23-29 Finalize results, draft slides, prepare for presentation

6 *AI Acknowledgement*

Acknowledgment of AI assistance: This document was partially drafted and revised with the help of ChatGPT (OpenAI, 2025) to improve clarity, organization, and formatting consistency, mostly concerning Latex formatting.