Using Topology to Extract Insights from UAAP Basketball Data

An Undergraduate Thesis Presented to the
Faculty of the Department of Mathematics
Ateneo de Manila University
in Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Applied Mathematics with Specialization
in Data Science

Benjamin Louis Ang, Alexander Pino Jr. Dane Lauren Rosario
March 28, 2024

Table of Contents

Acceptance Page

The Faculty of the Department of Mathematics of Ateneo de Manila University accepts the undergraduate thesis entitled

Using Topology to Extract Insights from UAAP Basketball Data

submitted by Benjamin Louis Ang, Alexander Pino Jr. Dane Lauren Rosario and orally presented on April 18, 2024, in partial fulfillment of the requirements for the degree Bachelor of Science in Applied Mathematics with Specialization in Data Science.

Clark Kendrick C. Go, Ph.D. Adviser	Date
Job A. Nable, Ph.D. Adviser	Date
Eden Delight P. Miro, Ph.D. Department Chair	Date
Name Panelist	Date
Name Panelist	Date

Acknowledgments

For the people who supported us

Summary of the Thesis

We apply topological data analysis techniques to tabular datasets comprising several representative basketball leagues and tournaments played in the Philippines and abroad. We use average persistence scores to quantify diversity in player attributes within single teams and within leagues, and associate it with team performance and league parity. Furthermore, we use the Mapper algorithm to characterize the similarities between players in each league through an explainable graphical representation, identify clusters and holes of interest, and relate them to emerging trends in the evolution of basketball strategy and tactics.

Anti-Plagiarism Declaration

I declare that I have authored this thesis independently, that I have not used materials other than the declared sources or resources, and that I have explicitly marked all materials which have been quoted either literally or by content from the used sources.

Benjamin Louis Ang Student I.D. No. 200302	Date
Alexander Pino Student I.D. No. 204054	Date
Dane Lauren Rosario Student I.D. No. 204420	Date

Chapter 1

Introduction

1.1 Background of the Study

The traditional basketball system categorizes players into five distinct positions: point guard, shooting guard, small forward, power forward, and center. Yet, this system no longer fully encapsulates the breadth of player capabilities in today's game. The evolution of basketball, propelled by advances in nutrition, training, and strategy, has outgrown the rigidity of such classifications. Modern players often embody a blend of skills that defy these archaic confines, necessitating a reevaluation of how we categorize basketball talent [5]. Given this context, advanced analytics emerges as a crucial tool. It precisely analyzes performance data to reveal patterns and skills beyond traditional classifications. This approach allows for the identification of new, empirically-backed player archetypes that reflect the actual range of abilities in modern basketball more accurately. Consequently, advanced analytics not only questions traditional classifications but also enhances our understanding of basketball by offering a more detailed perspective on player roles.

Several studies were able to find subgroups within the traditional five positions using data science. Notably, Kalman and Bosch [10] employed model-based clustering on data spanning the 2009-2010 to 2018-2019 NBA seasons, identifying nine distinct clusters to redefine player positions. This study not only tracked the evolution of specific players' roles throughout their careers but also utilized regression and random forest models to predict lineup performance based on these newly defined clusters. Furthermore, Jyad [9] applied Principal Component Analysis (PCA) to the 2018-19 NBA season data, using hierarchical cluster analysis to categorize players into nine revised positions. Building on top of these studies, Hedquist [5] introduced visualization and mega cluster analysis to further redefine NBA positions, showcasing the utility of hierarchical clustering and long-term player tracking through mega-clustering across seasons.

Besides those, a group of researchers used Topological Data Analysis to find hidden subgroups. The first of which is Alagappan [1] who leveraged topological data analysis of player shot charts from the 2010-2011 NBA season to propose an expansion of player positions ”from 5 to 13.” This innovative classification provided a more nuanced understanding of player roles, offering significant insights into team composition, player management, and recruitment strategies. Inspired by Alagappan, Gispets [4] applied the same methodology on the EuroLeague. While Nathan Joel Diambra [3] used topology to delved deeper into the dynamics of NCAA Division I Men's Basketball.

The successful application of TDA in these studies highlights its versatility and effectiveness in dissecting the complex fabric of basketball across various leagues. Each application has shed light on distinctive player role dynamics, strategic nuances, and stylistic diversities inherent to different competitive environments.

Despite these advancements, a detailed exploration of player subgroups within the University Athletic Association of the Philippines (UAAP) Basketball League has yet to be undertaken. Given the unique context and competitive landscape of UAAP basketball, this thesis seeks to fill the existing void. By applying TDA to player data from UAAP Season 85, our objective is to uncover a richer, more detailed understanding of player roles within the Philippines' premier collegiate basketball league, thereby contributing to the broader discourse on basketball analytics.

1.2 Goal of the Study

Our study explores player dynamics within the UAAP basketball league, utilizing Topological Data Analysis (TDA) to uncover patterns not visible through traditional analysis methods. We aim to address three core questions:

1. How can TDA be used to identify and confirm the presence of player subgroups that extend beyond traditional basketball positions within the UAAP league?

2. In what ways can the Mapper Algorithm reveal potential strategic advantages by dissecting the league's player composition, particularly through the identification of outliers and regions of sparse data?

3. What distinct clusters of players exist within the UAAP league, and what unique attributes or strategic roles do these clusters embody?

Through these inquiries, we anticipate shedding light on the nuanced player roles and strategic dimensions that characterize the UAAP basketball landscape.

1.3 Review of Related Literature

1.3.1 Basketball Background

History of Basketball

In the winter of 1891, physical education instructor Dr. James Naismith invented basketball. Motivated by the need for an indoor game to keep his football players active, Naismith hung peach baskets at each end of the gymnasium and crafted rules emphasizing teamwork, skill, and strategy. Over time, these foundational principles laid the groundwork for the development of basketball into a globally recognized sport. [25]

Although the fundamental principles of basketball endure, the now-global sport has evolved significantly since its inception, bearing little resemblance to its 19th-century origins. One of the key drivers in this evolution was the introduction of the three-point line in the early 1980s. This innovation fundamentally altered the dynamics of the game, incentivizing players to shoot from long distances due to the extra points awarded for successful three-point shots. As a result, basketball strategies and playing styles have evolved to incorporate the three-point shot as a crucial element of offensive tactics, reshaping the way the game is played at all levels. [18]

Standard Positions and its Limitations

Basketball games feature two teams, each comprising five players on the court simultaneously. Traditionally, players are assigned positions and numbers corresponding to their roles: Point Guard (1), Shooting Guard (2), Small Forward (3), Power Forward (4), and Center (5). While teams can opt for variations, the standard lineup typically includes one player from each position, allowing for a flexible distribution of roles.

Ten basketball experts were assigned to rate the importance of nineteen performance criteria for different player positions in a study conducted by Trinicic and Dizdar. They established seven criteria for defensive performance (level of defensive pressure, defensive help, blocking shots, ball possession gained, defensive rebounding efficiency, transition defense efficiency, and playing multiple positions of defense) and twelve for offensive performance (ball control, passing skills, dribble penetration, outside shots, inside shots, free throws, drawing fouls and three-point plays, efficiency of screening, offense without the ball, offensive rebounding efficiency, transition offensive efficiency, playing multiple positions on offense). Strong agreement among the experts is observed with over 90% for all positions. Based on this, the study defined the specific roles of each position and identified how they are different in terms of which performance criteria are the most important.[23]

- The Point Guard has above-average importance for the level of defensive pressure, transition defense efficiency, ball control, passing skills, dribble penetration, outside shots, and transition offense efficiency;

- Shooting Guards, while sharing similarities with Point Guards have above-average importance for level of defensive pressure, transition defense efficiency, outside shots, dribble penetration, offense without the ball, and transition offense efficiency;

- Small Forwards exhibit versatility, contributing across various roles depending on team dynamics and game situations. As such, they prioritize transition defense efficiency, outside shots, dribble penetration, offense without the ball, free throws, and transition offense efficiency;

- Power Forwards, often larger versions of Small Forwards, specialize in defensive and offensive rebounding efficiency, inside shots, dribble penetration, efficiency of screening, and free throws;

- Centers, usually the tallest players on the team, prioritize defensive and offensive rebounding efficiency, inside shots, dribble penetration, efficiency of screening, drawing fouls and three-point plays, and free throws

The categorization of performance criteria in this study offers a valuable framework for evaluating basketball players. However, their descriptions for each position need an update to capture the nuances of the modern positionless game. Understandably, the study's categorization of player roles reflects the prevailing on-court landscape of the year 2000. The evolution of basketball towards a more positionless style, with players exhibiting a broader skillset, had yet to fully take hold at the time of publication. In today's game, many players defy traditional positional boundaries by showcasing a combination of skills that transcend conventional roles. For example, some centers possess perimeter shooting abilities comparable to guards, while some guards excel in rebounding and interior defense like forwards. Additionally, the rise of versatile "positionless basketball" strategies further blurs the lines between traditional positions, emphasizing adaptability and skill versatility over rigid positional assignments.

Overall, while the study's framework reflects traditional positions, its core criteria remain relevant for evaluating players in today's positionless basketball. By adapting the framework to assess emerging roles, we can gain valuable insights into player performance and their categorizations.

1.3.2 Sports Analytics in Basketball

In the realm of physical team sports, basketball, one of the most popular sports with global appeal and star talents, has lately also become one of the most analytics-driven sports throughout its competitive scene. Recent innovations in basketball strategy and tactics, such as the three-point shot-centered offense, originated from the results of data analysts employed in top basketball teams, and have proceeded to proliferate across all levels of play, even down to amateur leagues and training programs. In addition to on-court adjustments, sports analytics have also provided basketball teams with advanced decision-making tools for a large variety of tasks, from determining recovery timelines for injured players to scouting for potential player changes and acquisitions.

An often-cited case study of the interaction between data analytics and basketball is that of Daryl Morey, an award-winning executive who is currently, as of 2024, president of basketball operations at the Philadelphia 76ers of the National Basketball Association (NBA). His tenure as general manager of fellow NBA team, the Houston Rockets, between 2007 and 2020 was documented in the book The Undoing Project by Michael Lewis, highlighting his management style driven by using statistical analysis to overcome long-held biases in the intuition of basketball experts and ultimately increase performance. The Houston Rockets under Morey would become a perennial contender for the NBA championship throughout the 2010s and is credited for its evolution of basketball strategy and player development, particularly of its star player, James Harden, who was acquired in 2012 and became Most Valuable Player (MVP) of the entire league in 2018. [11]

1.3.3 Previous Research Into Updated Player Positions

Although the conventional classification of basketball players adheres to the standard five positions, recent years have seen a surge in research exploring 'advanced' player positions. Most of the researchers have used traditional methods like k-means and hierarchical clustering, but there is a growing trend toward utilizing topological data analysis to identify subgroups.

Statistical Analysis

Bruin Sports Analytics explored the concept of "neo-positions" in the NBA, emerging player roles that defy traditional positional boundaries using the lens of statistics. The article examines the prevalence and effectiveness of these new positions (point forwards, stretch bigs, and score-first point guards) at the team level. It analyzes how these roles might impact the future of NBA basketball. To isolate the impact of the neo-position itself, the study compares players within these roles (stars, mid-tier, and lower-level) to their traditionally positioned counterparts. Statistical metrics like Win Shares/48, assist percentage, usage rate, three-point attempts, and field goal attempts are used for the analysis. [15]

They begin by selecting a traditional position (e.g., center) and defining the neo-position qualitatively through "eye-test" analysis, focusing on skills that deviate from the norm (e.g., a shooting center). This test can also be used to supplement the analysis with domain knowledge to add depth to the analysis. Players within the traditional position are then categorized by playing time (star, mid-level, or low). Finally, the analysis hinges on identifying advanced statistics that quantify the neo-position's defining characteristics. This involves exploring relationships between metrics to create new composite measures. By comparing players categorized as neo-positions to their traditional counterparts across different playing time levels, the chosen statistics reveal the prevalence and effectiveness of this new role within the NBA. While this approach offers valuable insights, it relies on a subjective definition of the neo-position and requires careful selection of appropriate statistics for a nuanced analysis.

Mega clustering

Hedquist [5], aimed to redefine NBA basketball positions through visualization and mega-cluster analysis. It leverages hierarchical clustering, a well-established technique for grouping similar data points. However, unlike traditional applications where individual data points are clustered, the mega-clustering approach focuses on pre-defined player clusters. Specifically, it applies hierarchical clustering to nine pre-identified player clusters obtained from each of the 20 NBA seasons under investigation. This methodology yields a total of 180 "objects" (9 clusters/season * 20 seasons) that are subsequently classified into nine overarching "mega-clusters." He identified nine "mega-clusters": score-first guards, pass-first guards, superstars, defensive big men, scoring big men, miscellaneous and transient players, and the bench role players.

While established clustering techniques like hierarchical clustering have proven valuable for player classification, this paper will explore the application of topological data analysis to identify potential subgroups within basketball player data. It offers advantages in analyzing high-dimensional performance metrics and revealing underlying structures within the data, potentially uncovering hidden subgroups that might be missed by traditional methods.

1.3.4 Topological Data Analysis

Topology

Topology focuses on identifying and characterizing intrinsic properties of geometric shapes that remain unchanged even when there are deformations. It allows for stretching, contracting, and bending of objects without tearing or gluing so it is described as a "rubber-sheet geometry." This enables the distinction between shapes based on their underlying topological structure. For instance, a square can be continuously deformed into a circle, highlighting its topological equivalence. In contrast, a figure-eight knot cannot be transformed into a circle without introducing tears, demonstrating their distinct topological nature. [24]

Recent years have witnessed a surge in adapting topological methods to analyze complex data, particularly large, high-dimensional datasets. This field, known as topological data analysis, leverages geometric approaches to identify patterns and shapes within the data. Unveiling these data "shapes" is crucial for extracting insights and identifying meaningful subgroups. [13] Topology's strength in pattern recognition via shape analysis hinges on three core concepts. Firstly, it operates in metric spaces where distances define a shape, independent of the chosen coordinate system, allowing comparisons across different data platforms. Secondly, topology focuses on properties that remain unchanged under minor deformations, making it less sensitive to noise. Mathematically, this allows circles, ellipses, and hexagons to be considered equivalent. Finally, topology utilizes compressed shape representations like triangulations (finite networks) that capture essential features while discarding minor details. [13]

Applications of Topological Data Analysis in Different Fields

Topological data analysis leverages topology by representing complex data as networks. Data points are represented as nodes in this network, and their relationships are depicted by the edges that connect the nodes. An intuitive map is created based on the similarity of the data points, and more similar data points are placed closer together which reveals underlying structures and patterns within the data. This allows for an understanding of high-dimensional data by reducing them to lower dimensions while preserving crucial topological features, providing an intuitive way to grasp the structure of complex data. [2]

The value of topological data analysis is reflected across various fields as it has yielded promising results in highly different areas. We will not undertake a complete overview or go extensively into individual study methodology due to the large body of research. Rather, we will present a high-level summary of the various domains in which it is being used to show the multidisciplinary nature of this technique.

Beyond its core application in mathematics, topological data analysis has fostered significant advancements in various scientific disciplines. In neuroscience, TDA has aided in tasks like identifying brain cavities, analyzing event-related fMRI data, and distilling complex neuroimaging data. Additionally, it has shown promise in medical image processing for patient classification and exploration of hyperspectral imaging data. In the natural sciences, applications range from generating large-scale genomic recombination maps to understanding biological aggregation models and quantifying pattern formations. Its reach extends to physics (fast radio burst analysis), engineering (aviation data exploration), and signal processing (movie genre detection). Notably, it has also been applied in operations research, facilitating the identification of clusters in social networks and the detection of structural information in manufacturing data. [19]

The aforementioned uses demonstrate how topological data analysis is developing into a potent instrument with uses that go well beyond its foundation in mathematics. This highlights its versatility, showcasing its effectiveness in diverse fields. One of its key strengths is its ability to extract meaningful information from complex, high-dimensional data. While the passage acknowledges the extensive research on specific topological data analysis methodologies, its focus is on showcasing the multidisciplinary nature of its applications. This approach emphasizes its potential as a generalizable technique that can be applied across a wide range of scientific disciplines. Furthermore, the inclusion of recent applications like fast radio burst analysis and social network clustering suggests topological data analysis is a continuously evolving field with the potential to tackle new and emerging challenges like the evolution of basketball.

Applications of Topological Data Analysis in Basketball

Topological data analysis identifies clusters by examining the geometric and topological properties of the dataset, unveiling hidden structures that may go unnoticed by conventional methods. An application of it is visualizing high-dimensional data. The pioneering study conducted by Lum demonstrated the effectiveness of the proposed method by applying it to three distinct datasets: gene expression from breast tumors, voting data from the U.S. House of Representatives, and National Basketball Association (NBA) player performance metrics [13] . However, for this section of the paper, we will be focusing on their third dataset. The last dataset that they conducted a study on are statistics of individual NBA players which includes rates (per minute played) of rebounds, assists, turnovers, steals, blocked shots, personal fouls, and points scored. Using a specific distance metric and filtering method for the Mapper Algorithm which were variance normalized Euclidean and principal and secondary SVD values, respectively, they were able to identify a wider range of playing styles. As a result, they were able to discover a more specific categorization of the traditional five positions into thirteen newly identified roles.

In each case, the technique showcased its prowess in analyzing complex, high-dimensional datasets, surpassing traditional clustering methods. The

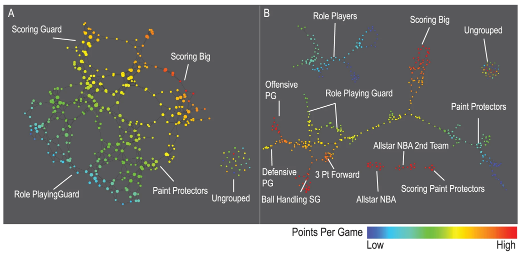


Figure 1.1: NBA Mapper Plot from [13]

method revealed more refined and nuanced stratifications within the datasets, providing deeper insights into the underlying structures. The versatility demonstrated across such diverse data types suggests the method's potential to enhance analytical approaches in various fields where conventional clustering may fall short.

Beyond its contribution to showing the multidisciplinary nature of topological data analysis, its merits also include promoting it to the field of sports which includes this paper. The pioneering study has catalyzed subsequent research endeavors. A notable example is the study by Albert Ratera Gispets which was able to the paper was able to discern the underlying subgroups of EuroLeague players from the 2019–2020 season by integrating statistical techniques with TDA [4]. His approach involved principal component analysis, persistent homology, and the application of Mapper.

While Principal Component Analysis (PCA) was initially applied to the entire dataset, it yielded results with limited interpretability in the context of player positioning. As a consequence, it was further investigated by applying it to filtered data sets which only included players from similar positions. Different players with the same playing style like Mike James (1) and Shane Larkin (3) were observed to be clustered together which can be used for

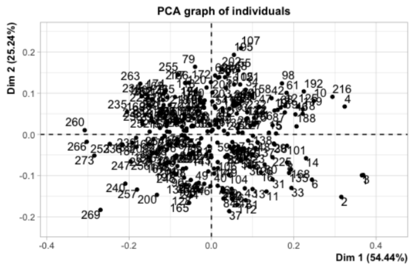


Figure 1.2: Distribution of All Players from [4]

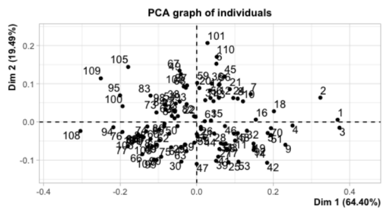


Figure 1.3: Distribution of Guards from [4]

finding player replacements. Despite being able to find subgroups, the paper insists that it was not enough to find new clusters that can modernize the traditional positions [4].

The paper also used Persistent Homology to investigate the existence of latent subgroups of a specific position [16]. The key here is that the existence of subgroups from the three positions can be determined by looking at the distribution of the 1-dimensional homologies in their respective persistent diagram. According to their findings, there does exist more than one group of guards, the number of subgroups for forwards is less than the guards as there are fewer 1-dimensional homologies, and there also exists more than one group of centers despite having fewer data points.

To achieve its key findings, the study primarily leverages the mapper

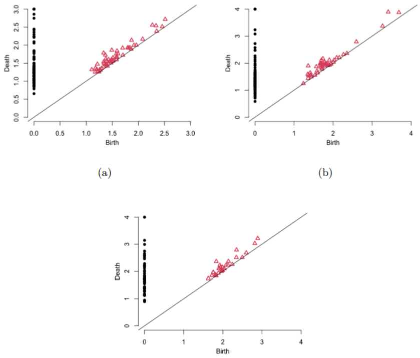


Figure 1.4: Persistent Diagrams for guards (a), forwards (b), centers (c) from [4]

algorithm. However, the inherent characteristic of the mapper algorithm is its sensitivity to filter function parameters. Consequently, the paper explores various configurations, resulting in multiple outputs from the algorithm. Two primary cases are investigated: one utilizing two filters and another employing three. Notably, the three-filter configuration yielded the most compelling results which used variance-normalized Euclidean distance as its distance metric and the first and second principal components from a Singular Value Decomposition (SVD) as filter functions.

In conclusion, the paper presented a compelling argument for the effectiveness of the Mapper Algorithm in identifying and clustering subgroups within traditional basketball positions. Furthermore, it highlights the significant impact of choosing appropriate distance metrics and filter functions on the results.

Building on the foundation laid by the aforementioned study, a subsequent research endeavor by Nathan Joel Diambra in 2018 titled “Using Topological Clustering to Identify Emerging Positions and Strategies in NCAA

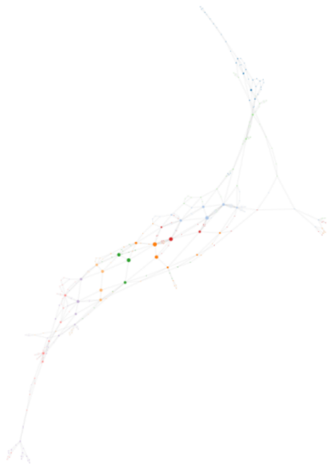


Figure 1.5: EuroLeague Mapper Plot with three filter functions from [4]

Men's Basketball" delved deeper into the dynamics of NCAA Division I Men's Basketball. This study not only displayed the significance of employing performance metrics but also took a step further by successfully identifying and delineating eight distinct positions within the college basketball landscape. These positions, including the Bench Warmer, Role Player, Rebounding Shot Blocker, Ball Handling Defender, Three Point Scoring Ball Handler, Three Point Scoring Rebounder, Close Range Dominator, and Point Producer, were identified based on nuanced statistical differentiation and topological data analysis.[3]

The expansion of this research into NCAA Division I Men's Basketball not only affirms the applicability of the methodology across diverse basketball leagues but also highlights the evolving nature of player roles and strategies within the collegiate setting. This shows that the replication of the methodology for identifying player subgroups across various basketball leagues holds inherent value, as it allows for a nuanced understanding of the diverse strate-

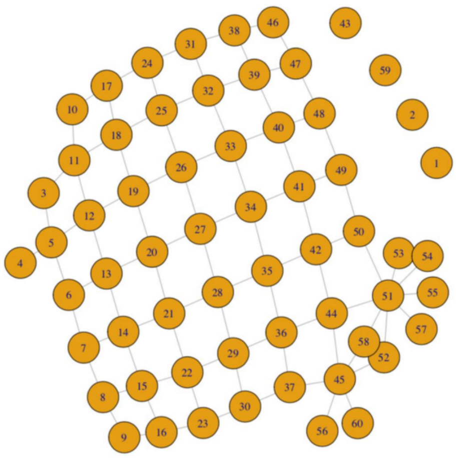


Figure 1.6: NCAA Division 1 Mapper Plot from [3]

gies, playing styles, and positional dynamics unique to each league.

Chapter 2

Mathematical Framework

In this chapter, we establish the mathematical underpinnings crucial for our exploration into topological data analysis. We begin with the fundamentals of topology, the bedrock of this field, which equips us with the abstract language necessary for describing various notions of space, continuity, and deformation. Topology's focus on properties preserved under continuous transformations makes it the ideal framework for studying the shapes and features within data.

Building upon this topological base, we advance to complexes and algebra, which translate these abstract concepts into algebraic forms—enabling precise computation and manipulation. We culminate with homology, a powerful tool that bridges the gap between qualitative geometric intuition and quantitative algebraic invariance.

2.1 Topology

2.1.1 Metric Spaces

Definition 2.1.1. A metric space is a set M together with a "distance" function $d : M \times M \to \mathbb{R}$ called a metric, satisfying the following properties for all $x, y, z \in M$:

1. $d(x, y) \geq 0$ (non-negativity),
2. $d(x, y) = 0$ if and only if $x = y$ (identity of indiscernibles),
3. $d(x, y) = d(y, x)$ (symmetry),

4. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality).

Examples of metric spaces include:

- The set of real numbers $\mathbb{R}$ with the metric $d(x, y) = |x - y|$.
- The Euclidean space $\mathbb{R}^n$ with the metric $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, where $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$.
- A discrete space X where the metric is defined as $d(x, y) = 1$ if $x \neq y$, and $d(x, x) = 0$.

Open Sets and Closed Sets (in a metric space)

Now that we have a notion of distance, we can define what it means to be an open set in a metric space.

Definition 2.1.2 (Ball)**.** Let X be a metric space. A *ball* B of radius r around a point $x \in X$ is $B = \{y \in X \mid d(x, y) < r\}$.

We recognize that this ball encompasses all points whose distance is less than r from x.

Definition 2.1.3 (Open Set)**.** A subset $O \subseteq X$ is *open* if for every point $x \in O$, there is a ball around x entirely contained in O.

Example 2.1.4. Let $X = [0, 3/4]$. The interval $(0, 1/4)$ is open in X.

Example 2.1.5. Let $X = \mathbb{R}$. The interval $[0, 1/2)$ is not open in X.

Definition 2.1.6. A set C is *closed* if $X - C$ is open.

2.1.2 Topological Spaces

Definition 2.1.7 (Topology)**.** A *topology* τ on a set X consists of subsets of X satisfying the following properties:

1. The empty set and the space X are both sets in the topology.
2. The union of any collection of sets in τ is contained in τ.
3. The intersection of any finitely many sets in τ is also contained in τ.

Members of τ are called *open sets* in X.

Example 2.1.8. Investigating a set X comprising three elements $\{a, b, c\}$ and a proposed topology $\tau = \{\varnothing, X, \{a\}, \{b\}\}$, it becomes evident that τ fails to meet one of the key criteria for a topology. The principle that the union of any two elements within a topology should also belong to the topology is violated here, as the union $\{a\} \cup \{b\} = \{a, b\}$ is not included in τ, thereby indicating that τ cannot be a valid topology for X.

Example 2.1.9. In a scenario where a set X has a defined topology τ, the elements of τ are recognized as open sets. This implies that for a set X with an inherent metric, or distance measure, the collection of all such open sets as per the metric definition constitutes a valid topology on X, commonly referred to as the Euclidean topology. This concept highlights the interplay between metric spaces and topological spaces.

Definition 2.1.10 (Interior, Closure, and Boundary)**.** Given a topological space X and a subset $A \subseteq X$:

- The *interior* of A, denoted by $\mathring{A}$, is the union of all open sets contained in A.

- The *closure* of A, denoted by $\overline{A}$, is the intersection of all closed sets containing A.

- The *boundary* of A, denoted by ∂A, is given by the set difference between the closure and the interior: $\partial A = \overline{A} \setminus \mathring{A}$.

2.1.3 Homeomorphism

Definition 2.1.11 (Continuous Function)**.** A function $f : X \to Y$ between topological spaces is *continuous* if $f^{-1}(V)$ is open in X, whenever V is open in Y.

Definition 2.1.12 (Neighborhood)**.** Given a point x of X, we call a subset N of X a *neighborhood* of x if we can find an open set O such that $x \in O \subseteq N$.

1. A function $f : X \to Y$ is continuous if for any neighborhood V of Y there is a neighborhood U of X such that $f(U) \subseteq V$.

2. A composition of 2 continuous functions is continuous.

Definition 2.1.13 (Homeomorphism)**.** A *homeomorphism* is a function $f : X \to Y$ between two topological spaces X and Y that:

- is a bijection, meaning it is both injective (one-to-one) and surjective (onto), ensuring each element of X is paired with a unique element of Y and every element of Y is matched with some element of X, and

- has a continuous inverse function f^{-1}, which means that f^{-1} is also continuous, preserving the topological structure in both directions.

Remark: Homeomorphism is the notion of equality in topology as it induces an equivalence relation among toplogical spaces.

Example 2.1.14. Every open interval within the real numbers $\mathbb{R}$ can be transformed into any other open interval through a homeomorphism. For instance, take $X = (-2, 2)$ and $Y = (1, 4)$. A mapping $f : X \to Y$ given by $f(x) = \frac{3}{4}x + 2.5$ is both bijective and continuous. Its inverse function $f^{-1}(y) = \frac{4}{3}y - \frac{10}{3}$ maintains continuity, illustrating the concept of homeomorphism.

Example 2.1.15. From a topological perspective, a circle $S^1 = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$ and a square $T = \{(x, y) \in \mathbb{R}^2 \mid \max(|x|, |y|) = 1\}$ are indistinguishable. This equivalence is established through a mapping $f : S^1 \to T$ defined by $f(x, y) = \left(\frac{x}{\max(|x|,|y|)}, \frac{y}{\max(|x|,|y|)}\right)$, which is continuous, one-to-one, and onto, with a continuous reverse mapping. Hence, topologically, the circle and square are identical.

2.2 Complexes

2.2.1 Simplicial Complexes

Definition 2.2.1 (Affine Combination)**.** Let d be a positive integer, and consider the points $p_0, p_1, \ldots, p_k \in \mathbb{R}^d$. If $x_0, x_1, \ldots, x_k$ are real numbers satisfying $\sum_{i=0}^{k} x_i = 1$, then the expression $\sum_{i=0}^{k} x_i p_i$ represents an *affine combination* of the points p_i. Specifically, if all x_i are nonnegative, the combination is termed a *convex combination.*

Definition 2.2.2 (Convex Hull)**.** The *convex hull* of a set of points $p_0, p_1, \ldots, p_k \in \mathbb{R}^d$ is the collection of all possible convex combinations of these points.

Definition 2.2.3 (Affinely Independent Set)**.** A subset $S \subseteq \mathbb{R}^d$ is *affinely independent* if no point in S can be written as an affine combination of the other points in S.

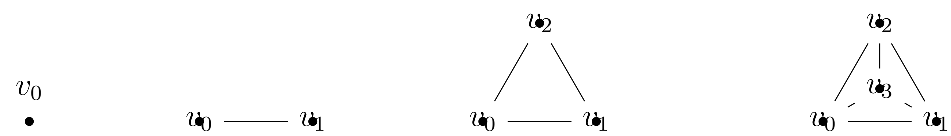
Definition 2.2.4 (Simplex)**.** A *p-dimensional simplex* σ in $\mathbb{R}^d$ is the convex hull of an affinely independent set of $k+1$ points, denoted by $S = \{v_0, v_1, \ldots, v_k\}$. These points are known as the vertices of the simplex. Thus, a *k-simplex* σ can be expressed as

$$\sigma = \left\{ \sum_{i=0}^{k} x_i v_i \;\middle|\; \sum_{i=0}^{k} x_i = 1, x_i \geq 0 \right\}.$$

Furthermore, a k-simplex is recognized to have a dimension of k.

Remark 2.2.5. Examples of low-dimensional simplices include:

- A 0-simplex is a point.
- A 1-simplex is a line segment.
- A 2-simplex is a triangle, including its interior.
- A 3-simplex is a tetrahedron, a three-dimensional figure with triangular faces.



0-simplex $\{v_0\}$ 1-simplex $\{v_0, v_1\}$ 2-simplex $\{v_0, v_1, v_2\}$ 3-simplex $\{v_0, v_1, v_2, v_3\}$

Definition 2.2.6 (Simplicial Complex)**.** A simplicial complex is a fundamental concept in algebraic topology. Let K be a simplicial complex, which is a set of simplices satisfying the following properties:

1. If σ is a simplex in K, then every face of σ is also in K.
2. The intersection of any two simplices in K is either empty or a face of each.

The dimension of a simplex n denotes $n + 1$ vertices comprising the simplex, forming shapes such as edges, triangles, or tetrahedrons.

add diagram

Definition 2.2.7 (Geometric Simplical Complex)**.** A *geometric simplicial complex* K is a finite set of simplices such that for any simplex $\sigma \in K$, the faces of σ are also in K. Additionally, for any two simplices $\sigma, \tau \in K$, their intersection is either empty or a face of both σ and τ. Note that if k is the maximum dimension of all simplices in K, then we say that K is of dimension k or K is a simplicial k-complex.

Definition 2.2.8 (Abstract Simplical Complex)**.** An *abstract simplicial complex* is a collection of sets closed under the operation of taking non-empty subsets. Each set in the collection is a simplex, and if a set is in the collection, then so are all its non-empty subsets.

Remark 2.2.9*.* Examples of abstract simplicial complexes include:

- The collection of subsets of vertices that form a graph.

- The nerve of an open cover in a topological space.

Definition 2.2.10. Given two simplices σ and τ of a simplicial complex K, we say that σ is a *face* of τ, denoted $\sigma \leq \tau$, whenever every vertex of σ is also a vertex of τ.

2.2.2 Subcomplexes, Closures, and Filtrations

Definition 2.2.11 (Subcomplex)**.** Let K be a simplicial complex. A subset $L \subseteq K$ of simplices is called a *subcomplex* of K if it satisfies the following property: for each simplex τ in L, if σ is a face of τ in K, then σ also belongs to L.

Definition 2.2.12 (Closure)**.** The *closure* of a collection of simplices K' in a simplicial complex K is defined to be the smallest subcomplex $L \subseteq K$ satisfying $K' \subseteq L$.

Definition 2.2.13. Let K be a simplicial complex; a *filtration* of K (of length n) is a nested sequence of subcomplexes of the form

$$F_1K \subseteq F_2K \subseteq \cdots \subseteq F_{n-1}K \subseteq F_nK = K.$$

In general, the dimensions of the intermediate F_iK are not constrained by i. On the other hand, in order to have a well-defined notion of length, we require $F_iK \neq F_{i+1}K$ for all i.

Definition 2.2.14 (Geometric Realization)**.** Let $\phi : K_0 \to \mathbb{R}^n$ be any function that sends the vertices of K to points in $\mathbb{R}^n$. The geometric realization of K with respect to ϕ is the union

$$|K|_\phi = \bigcup_{\sigma \in K} |\sigma|_\phi,$$

where for each $\sigma = \{v_0, \ldots, v_k\}$ in K, the set $|\sigma|_\phi \subseteq \mathbb{R}^n$ is the geometric simplex spanned by the points $\{\phi(v_0), \ldots, \phi(v_k)\}$.

Definition 2.2.15 (Simplicial Map)**.** A simplicial map $f : K \to L$ is an assignment $K_0 \to L_0$ of vertices to vertices which sends simplices to simplices. So for each simplex $\sigma = \{v_0, \ldots, v_k\}$ of K, the image $f(\sigma)$ is $\{f(v_0), \ldots, f(v_k)\}$ which must be a simplex of L.

Definition 2.2.16 (Barycentric Subdivision)**.** The barycentric subdivision of K is a new simplicial complex $Sd\ K$ defined as follows; for each dimension $i \geq 0$, the i-dimensional simplices are given by all sequences

$$\sigma_0 < \sigma_1 < \cdots < \sigma_{i-1} < \sigma_i$$

of (distinct) simplices in K ordered by the face relation.

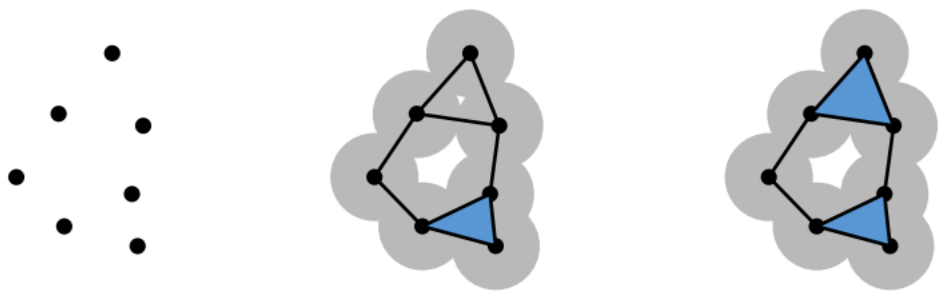
Definition 2.2.17 (Vietoris-Rips Filtration)**.** Let (M, d) be a finite metric space. The Vietoris-Rips filtration of M is an increasing family of simplicial complexes $VR_\epsilon(M)$ indexed by the real numbers $\epsilon \geq 0$, defined as follows: a subset $\{x_0, \ldots, x_k\} \subseteq M$ forms a k-dimensional simplex in $VR_\epsilon(M)$ if and only if the pairwise distances satisfy $d(x_i, x_j) \leq \epsilon$ for all i, j.

Definition 2.2.18 (Čech Filtration)**.** Let M be a finite subset of a metric space (Z, d). The Čech filtration of M with respect to Z is the increasing family of simplicial complexes C_ϵ indexed by $\epsilon \geq 0$ defined : a subset

Figure 2.1: On the left is the Čech complex and on the right is the Vietoris-Rips complex of the point set P.

$\{x_0, \ldots, x_k\} \subseteq M$ forms a k-dimensional simplex in $C_\epsilon(M)$ if and only if there exists some z in Z satisfying $d(z, x_i) \leq \epsilon$ for all i.

Example 2.2.19. Let $P \subseteq \mathbb{R}^2$ with the Euclidean distance metric be given below. A Čech complex and a Vietoris-Rips complex of P are illustrated below [14].



P $\quad$ $C^r(P)$ $\quad$ $VR^r(P)$

The Čech complex on the left depicts the union of intersections formed by overlapping neighborhoods around each point in P, resulting in a connected structure between the points. The Vietoris-Rips complex on the right connects points within a certain range directly, forming a graph-like structure that highlights the immediate neighbors of each point without considering the intersecting neighborhoods.

2.3 Algebra

2.3.1 Groups and Related Concepts

Definition 2.3.1 (Group)**.** A *group* $(G, +)$ is a set G with a binary operation $+$ such that:

- Closure: For any $a, b \in G$, the result $a + b \in G$.
- Associativity: For any $a, b, c \in G$, we have $(a + b) + c = a + (b + c)$.
- Identity: There exists an element $0 \in G$ such that $a + 0 = 0 + a = a$ for any $a \in G$.
- Inverses: For each $a \in G$, there exists an element $-a \in G$ such that $a + (-a) = (-a) + a = 0$.

If the operation $+$ also satisfies $a + b = b + a$ for any $a, b \in G$, then G is called an *abelian group*, or *commutative group*.

Definition 2.3.2 (Subgroup)**.** A subset $H \subseteq G$ is a *subgroup* of $(G, +)$, denoted $(H, +)$, if it also forms a group under the operation $+$.

Definition 2.3.3 (Cosets and Quotient Groups)**.** Given a group $(G, +)$ and a subgroup H, for any $a \in G$, the sets

- $aH = \{a + h \mid h \in H\}$, the *left coset*, and
- $Ha = \{h + a \mid h \in H\}$, the *right coset*,

are called the *cosets* of H in G. For an abelian group G, left and right cosets are identical.

The *quotient group* G/H is defined as the set of all cosets of H in G.

Definition 2.3.4 (Homomorphism and Isomorphism)**.** A function $f : G \to H$ between two groups is a:

- *Homomorphism* if $f(a + b) = f(a) + f(b)$ for any $a, b \in G$.
- *Isomorphism* if it is a bijective homomorphism, implying G and H are *isomorphic*, denoted $G \cong H$.

Definition 2.3.5 (Kernel and Image)**.** For a homomorphism $f : G \to H$, the *kernel* is $\ker(f) = \{a \in G \mid f(a) = 0_H\}$, and the *image* is $\mathrm{im}(f) = \{b \in H \mid \exists a \in G \text{ with } f(a) = b\}$.

Definition 2.3.6 (2.36. Cyclic Group)**.** Let n be a positive integer. The *cyclic group* $\mathbb{Z}_n$ consists of the set $\{0, 1, \ldots, n-1\}$ with addition modulo n as the operation. This group is closed under addition, and every element $a + b$ results in a value within the set after modulo n reduction.

2.3.2 Ring Theory

Definition 2.3.7 (Ring)**.** A *ring* $(R, +, \cdot)$ is a set R equipped with two operations: addition $(+)$ and multiplication $(\cdot)$. This structure satisfies the following:

- $(R, +)$ forms an abelian group under addition.
- Multiplication is associative: $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.
- Multiplication is distributive over addition: $a \cdot (b + c) = a \cdot b + a \cdot c$ and $(a + b) \cdot c = a \cdot c + b \cdot c$.

The ring $(R, +, \cdot)$ is *commutative* if $a \cdot b = b \cdot a$ for all $a, b \in R$.

Definition 2.3.8 (Field)**.** A *field* is a commutative ring $(R, +, \cdot)$ in which every nonzero element has a multiplicative inverse: for any nonzero $a \in R$, there exists an $a^{-1} \in R$ such that $a \cdot a^{-1} = 1$.

Definition 2.3.9 (Commutative Ring)**.** Given a commutative ring $(R, +, \cdot)$, an *R-module M* is an abelian group under addition that allows 'scaling' by elements of R via multiplication. It satisfies the conditions:

- $a \cdot (x + y) = a \cdot x + a \cdot y$,
- $(a + b) \cdot x = a \cdot x + b \cdot x$,
- $1 \cdot x = x$,
- $(a \cdot b) \cdot x = a \cdot (b \cdot x)$,

for all $a, b \in R$ and $x, y \in M$.

Definition 2.3.10 (2.40. Vector Space)**.** A *vector space* over a field $(R, +, \cdot)$ is a set V that acts like an R-module where the operations of addition and scalar multiplication satisfy vector space axioms. A finite set of vectors $\{b_1, b_2, \ldots, b_k\}$ *generates* V if every element $v \in V$ can be uniquely expressed as a linear combination of these vectors. If so, the set forms a *basis* of V, and its size k is known as the *dimension* of V.

2.4 Homology

2.4.1 Orientations and Boundaries

Definition 2.4.1 (Face of a Simplex)**.** Let K be an oriented simplicial complex and let $\sigma = (v_0, \ldots, v_k)$ be an oriented k-simplex in K. For each i in

$\{0, 1, \ldots, k\}$, the i-th face of σ is the $(k-1)$-dimensional simplex

$$\sigma_{-i} = (v_0, \ldots, v_{i-1}, \widehat{v_i}, v_{i+1}, \ldots, v_k)$$

obtained by removing the i-th vertex.

Definition 2.4.2 (Algebraic Boundary)**.** Let σ be a k-dimensional oriented simplex. The algebraic boundary of σ is the linear combination

$$\partial_k \sigma = \sum_{i=0}^{k} (-1)^i \sigma_{-i},$$

where σ_{-i} denotes the i-th face of σ as in Definition 3.3.

2.4.2 Chain Complexes

Definition 2.4.3 (Chain Group)**.** For each dimension $k \geq 0$, the k-th chain group of K is the vector space $C_k(K)$ over $\mathbb{F}$ generated by treating the k-simplices of K as a basis.

Definition 2.4.4 (Boundary Operator)**.** For each dimension $k \geq 0$, the k-th boundary operator of K is the $\mathbb{F}$-linear map $\partial_k : C_k(K) \to C_{k-1}(K)$ which sends each basis k-chain to the $(k-1)$-chain

$$\partial_k^K(\sigma) = \sum_{i=0}^{k} (-1)^i \sigma_{-i}.$$

Definition 2.4.5 (Chain Complex)**.** A chain complex $(C_\cdot, d_\cdot)$ over the field $\mathbb{F}$ is a collection of $\mathbb{F}$-vector spaces C_k (indexed by integers $k \geq 0$) and $\mathbb{F}$-linear maps $d_k : C_k \to C_{k-1}$ which satisfy the condition $d_k d_{k+1} = 0$ for all k.

2.4.3 Homology Proper

Definition 2.4.6 (Homology and Simplicial Homology Groups)**.** The k-th *homology group* $H_k(C., d.)$ of a chain complex $(C., d.)$ for $k \geq 0$ is the quotient vector space defined by $H_k(C., d.) = \ker\, d_k / \mathrm{img}\; d_{k+1}$, which captures the k-dimensional cycles that are not boundaries of $(k+1)$-chains. For a simplicial complex K, endowed with a field $\mathbb{F}$, the *simplicial homology groups* $H_k(K)$ or $H_k(K; \mathbb{F})$ are similarly constructed and reflect the k-dimensional holes within K.

Remark 2.4.7. These groups are particularly meaningful in topological data analysis as they provide an algebraic snapshot of the topology of K, and they are finite-dimensional provided that K is a finite complex.

Definition 2.4.8 (Betti Number)**.** The k-th *Betti number* of a simplicial complex K, denoted $\beta_k(K; \mathbb{F})$, is the rank of the k-th simplicial homology group, given by the dimension of $H_k(K; \mathbb{F})$ over a field $\mathbb{F}$:

$$\beta_k(K; \mathbb{F}) = \dim H_k(K; \mathbb{F}).$$

The Betti number reflects the number of k-dimensional holes.

Theorem 2.4.9 (Invariance of Homology Groups)**.** *Let $f : X \to Y$ be a homeomorphism between topological spaces. Then the induced homomorphism $f_* : H_p(X) \to H_p(Y)$ is an isomorphism for all p. This implies that homology groups are topological invariants.*

Theorem 2.4.10 (Rank-Nullity for Homology Groups)**.** *For a chain complex*

C, the pth homology group H_p satisfies the rank-nullity theorem:

$$\text{rank } Z_p - \text{rank } B_p = \text{rank } H_p,$$

where Z_p is the group of cycles and B_p is the group of boundaries within the chain complex C.

Theorem 2.4.11 (Stability of Betti Numbers)**.** *Betti numbers are stable under small perturbations in the Gromov-Hausdorff sense. Specifically, if two topological spaces X and Y are close in the Gromov-Hausdorff distance, then their Betti numbers are the same.*

Theorem 2.4.12 (Persistence Invariance)**.** *Given a finite sequence of topological spaces connected by inclusions $X_1 \subset X_2 \subset \ldots \subset X_n$, their persistence diagrams are invariant under homeomorphisms of the spaces. That is, if $f_i : X_i \to Y_i$ are homeomorphisms, then the persistence diagrams of $\{X_i\}$ and $\{Y_i\}$ are equal.*

Chapter 3

Data Preparation

3.1 Dataset

The chosen dataset for this study is the 85th season of the University Athletics Association of the Philippines's (UAAP) Men's Basketball tournament, which took place between October and December 2022. The dataset, comprising 127 rows representing the complete rosters of 8 teams, was A total of 46 features included in the dataset, with 41 being numerical. Some of the more important features are the following: Name, Team, Position, PPG (Points Per Game), APG (Assists Per Game), RPG (Rebounds Per Game), SPG (Steals Per Game), BPG (Blocks Per Game), TOPG (Turnovers Per Game), +/- (Plus-Minus, the player's impact on the game score while they are on the court), FG% (Field Goal Percentage), 3PT% (Three Point Percentage), FT% (Free Throw Percentage), MPG (Minutes Per Game)

3.2 Programming Language

Python became the main programming language for this research due to its versatility and libraries suitable for advanced data analysis. Specifically, Python supports the implementation of the Mapper algorithm, a crucial part of our methodology, through dedicated packages like Giotto. With its accessibility, widespread support, and numerous data analysis tools, Python ensures the efficient development and implementation of our analytical framework, highlighting its pivotal role in the successful execution of this research.

3.2.1 Library

Giotto-TDA [**giotto**] is a high-performance full-featured library for topological data analysis and topological machine learning in Python with several key advantages:

- Complete feature set: giotto-tda provides implementations for many topological data analysis techniques, including both Persistent Homology using Vietoris-Rips persistence, and the Mapper algorithm.

- Integration with scikit-learn: giotto-tda is compatible with sklearn functions such as PCA and DBSCAN, as well as using a similar pipeline object further making it more convenient to code.

- Mapper Interactive Plotter: this feature allows us to conveniently test different hyperparameters for using the Mapper algorithm on our data.

3.2.2 Data Preprocessing

Before going into the analysis, a comprehensive preprocessing phase was undertaken to ensure the quality and suitability of the basketball data. The initial step involved thorough data cleaning, addressing any missing values and anomalies within the dataset. To ensure the statistical significance of the analysis, players with insufficient playing time were systematically excluded from the dataset, establishing a threshold to focus the study on contributors who actively participated in the games during the tournament. Following this, normalization and standardization techniques were applied to the numerical features for a consistent scale across different variables. Additionally, the dataset's structure was refined to enhance its compatibility with the algorithms employed in the study. This meticulous preprocessing phase laid the groundwork for a robust and reliable analysis of the UAAP Men's Basketball data, minimizing the impact of potential data artifacts on the outcomes of the investigation.

Chapter 4

Persistent Homology

4.1 Background on Persistent Homology

Persistent homology is a method topological data analysis (TDA) that examines the evolution of topological features across varying scales or parameters. It builds upon the foundational concept of a simplicial complex K 2.2.6.

To understand the dynamic nature of data, we consider not just a single simplicial complex, but a sequence of complexes that unfold over a parameter—typically representing a scale or a progression through time. This sequence is known as a *simplicial filtration* 2.2.13:

$$\mathcal{F} = \{K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K\},$$

where each K_i is a subcomplex 2.2.11 of K corresponding to a particular parameter value. The filtration is constructed by first defining a monotonic function $f : K \to \mathbb{R}$ that assigns a real number to each simplex, capturing its 'birth time' within the dataset. For any value $a \in \mathbb{R}$, the sub-level set $K(a) = f^{-1}((-\infty, a])$ is the collection of all simplices born on or before time a, forming a subcomplex.

With this sequence, one can observe how the topology of the data changes as one progresses through the filtration. The homology groups 2.4.6 at each stage, $H_k(K_i)$, highlight the k-dimensional holes present within K_i. These groups, and their associated Betti numbers 2.4.11, serve as topological signatures of the data.

As we move through the filtration, we track the persistence of these homological features. The homomorphisms 2.3.4 induced by inclusion maps between consecutive simplices in the filtration,

$$H_k(K_i) \to H_k(K_{i+1}),$$

allow us to see which features persist and which dissipate as the parameter a increases. The concept of an algebraic boundary operator, ∂_k, is central to this analysis. It maps k-chains to their $(k-1)$-dimensional boundaries, thereby distinguishing between trivial cycles and non-trivial features.

Persistence homology translates these observations into a powerful visual tool: persistence diagrams and barcodes. Each barcode or point in the persistence diagram represents a homological feature, with its length or coordinates encoding the feature's birth and death within the filtration. The stability of these diagrams under small perturbations, as confirmed by the Stability of Betti Numbers theorem 2.4.11, is crucial when dealing with noisy data, making persistent homology a robust tool in data analysis.

Remark 4.1.1. The simplicial filtration provides a granular view of the topological landscape of a dataset, highlighting not just the features themselves, but the scale at which they are relevant. This multiscale perspective is indispensable in identifying the inherent geometric and topological structures within complex data.

4.2 Results and Analysis

We adopt the persistent homology methodology outlined by Gispets [4]. He created persistence diagrams using Vietoris-Rips complexes for the complete player dataset, as well as for individual player categories: guards, forwards, and centers. However, due to the limited number of centers in the UAAP dataset, we were unable to obtain a persistence diagram for the centers.

This methodology enables us to examine the distribution and the relationship of 0-dimensional and 1-dimensional homological features to the persistence diagrams' diagonal. The primary aim is to uncover latent groupings within the data that these features may indicate.

The results for the entire dataset are shown in 4.1. A point close to the diagonal indicates a feature that "lives" only for a brief range of the filtration parameter- suggesting that the data points forming this component are very close to each other.

Taking a look at the red dots, which represent the connected components, we see that some are very far from the diagonal, indicating a low birth, but a high death value. This indicates that there are connected components that are distanced away from the rest. In simple terms, some players are very "far" away from the rest.

Now, for the H1 components, It's noteworthy that most 1-dimensional homological features are distanced away from the diagonal, which indicates the presence of significant holes in the dataset. This implies that there exist subgroups in the dataset, which is to be expected since we are looking at the persistence diagram of the players regardless of position.

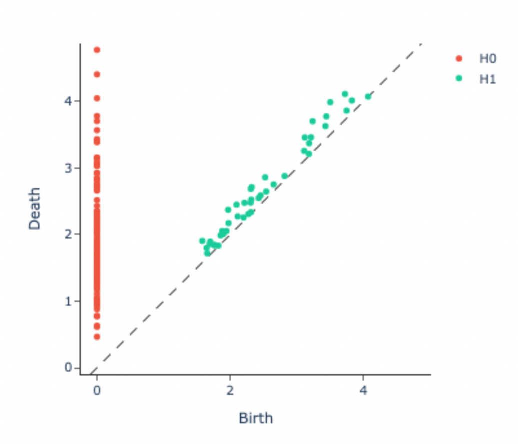


Figure 4.1: Persistence Diagram for the Whole Dataset

Now for the persistence diagram of the guards 4.2, we see that numerous red dots are clustering close to the diagonal, suggesting many transient connected components that form and merge as the scale increases, potentially representing noise or less significant groupings. Notably, there's a red dot with a relatively low birth and high death value, far from the diagonal. This indicates a more substantial and possibly well-separated cluster in the data.

Meanwhile, for the green dots, most are closer to the diagonal, which implies that loops or holes present among the Guards' data are not persistent and fill in relatively quickly. There are only a few holes that persist over a significantly long timeframe.

The persistence diagram of the forwards 4.3 features several connected components near the diagonal, with one dot standing out with higher persistence. The features are generally more spread out than in the Guards' diagram, which could indicate more variation in the clustering of the Forwards' data points.

Again, we see green dots near the diagonal. However, compared to the Guards, the green dots for the Forwards are spread out over a wider range, suggesting slightly more persistent loops or holes.

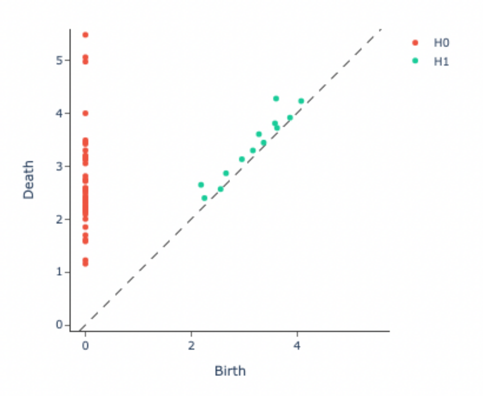


Figure 4.2: Persistence Diagram for the Guards

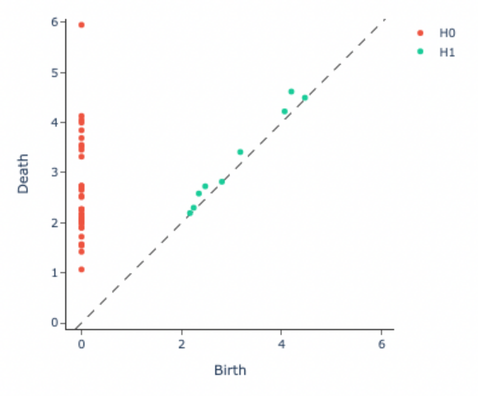


Figure 4.3: Persistence Diagram for the Forwards

Chapter 5

Mapper

5.1 Background

The Mapper algorithm, introduced by Singh et al. in 2007, serves the purpose of extracting global features from datasets with high dimensions. This algorithm facilitates the representation of point clouds through simplicial complexes, offering a simplified description that abstracts from precise distances or angles, and individual data points. The result of the Mapper algorithm is a simplicial complex, presenting a concise and comprehensive global representation of the original data.[20]

5.1.1 Topological Background

Recent studies [16] that have applied the Mapper Algorithm in their datasets used the theoretical background on topology from the paper that proposed Mapper as a new method that extracts descriptions of high dimensional data sets through simplicial complexes.[20]

In this paper, we will be using the same background [20] so we begin with some definitions. We start by letting $U = (U_\alpha)_{\alpha \in A}$ be a finite covering of a space X and the *nerve* of the covering U to be the simplicial complex $N(U)$. To construct a map from X to a specific space denoted by $N(U)$, we can utilize a concept called a partition of unity. A partition of unity subordinate to the covering U is a collection of real-valued functions $\{\phi_\alpha\}_{\alpha \in A}$, one for each open set U_α in the cover. These functions satisfy the following properties:

1. $0 \leq \phi_\alpha(x) \leq 1$ for all $\alpha \in A$ and $x \in X$.

2. $\sum_{\alpha \in A} \phi_\alpha(x) = 1$ for all $x \in X$.

3. The closure of the set $\{x \in X \mid \phi_\alpha(x) > 0\}$ is contained in the open set U_α.

Now, let there be a simplex with vertices $\{v_0, v_1, \ldots, v_k\}$ with a one-to-one correspondence between points inside the simplex and ordered k-tuples of real numbers $(r_0, r_1, \ldots, r_k)$ satisfying $0 \leq r_i \leq 1$ for all i (0 to k) and $\sum_{i=0}^{k} r_i = 1$. This mapping is called barycentric coordinatization, and the numbers r_i are the barycentric coordinates of a point. Now, for any point $x \in X$, let $T(x) \subseteq A$ be the set of all indices α such that x belongs to the open set U_α. We define $\rho(x) \in N(U)$ to be the point in the simplex spanned by the vertices $\alpha \in T(x)$. The barycentric coordinates of this point in the simplex are $(\phi_{\alpha_0}(x), \phi_{\alpha_1}(x), \ldots, \phi_{\alpha_l}(x))$, where $\{\alpha_0, \alpha_1, \ldots, \alpha_l\}$ is an ordering of the elements in $T(x)$. This map ρ can be shown to be continuous and provides a partial way to represent points in X using coordinates within the simplicial complex $N(U)$.

Suppose we have a space X equipped with a continuous map $f : X \to Z$ to a parameter space Z. Additionally, assume Z has a covering $U = \{U_\alpha\}_{\alpha \in A}$. The preimages under f Tf the open sets in U also form an open covering of X as f is continuous. We denote this preimage covering as $f^{-1}(U) = \{f^{-1}(U_\alpha)\}_{\alpha \in A}$. Now, for each element α in the indexing set A, we can further decompose the preimage $f^{-1}(U_\alpha)$ into its path-connected components. Let's denote this decomposition as:

$$f^{-1}(U_\alpha) = \bigcup_{i=1}^{j_\alpha} V^{(\alpha,i)}$$

Here, j_α represents the number of path-connected components within $f^{-1}(U_\alpha)$. Finally, we denote the collection obtained by decomposing each preimage set $f^{-1}(U_\alpha)$ as $\mathcal{U} = \{V^{(\alpha,i)}\}$, which forms an open covering of X derived from the original covering U of Z.

We can further extend the concept of coverings to maps between coverings. Consider two coverings $U = \{U_\alpha\}_{\alpha \in A}$ and $V = \{V_\beta\}_{\beta \in B}$. A map of coverings from U to V is a function $f : A \to B$ satisfying the already stated conditions. Given a map of coverings $f : A \to B$, we can induce a corresponding map of simplicial complexes, denoted by $N(f) : N(U) \to N(V)$. This map acts on vertices by simply applying the function f itself. For a family of coverings $U_i, i = 0, 1, \ldots, n$ and maps of coverings $f_i : U_i \to U_{i+1}$ for each index i, we can construct a diagram of simplicial complexes connected by simplicial maps:

$$N(U_0) \xrightarrow{N(f_0)} N(U_1) \xrightarrow{N(f_1)} \ldots \xrightarrow{N(f_{n-1})} N(U_n)$$

Given a space X with a map $f : X \to Z$ (parameter space Z) and coverings $U \to V$, a corresponding map of coverings $\hat{U} \to \hat{V}$ exists on X. Since $U \subseteq V$, $f^{-1}(U) \subseteq f^{-1}(V)$. Consequently, each connected component of $f^{-1}(U)$ is contained within a unique connected component of $f^{-1}(V)$. Formally, $\hat{U} \to \hat{V}$ maps $U_\alpha(i)$ to the unique $V_{f(\beta)}(j)$ such that $U_\alpha(i) \subseteq V_{f(\beta)}(j)$.

5.1.2 Pipeline

We start with a point cloud X. The first step involves grouping data points based on both their physical closeness and a user-defined "lens function." A filter function $f : X \to \mathbb{R}^k$ assigns a real-valued vector to each point in the input data space X. This allows researchers to analyze information depending on their research question. This function serves to encapsulate the key characteristics of each data point. Common selections for the filter function includes distance metrics, density estimators, or alternative feature extraction techniques. For our study, we used 2-dimensional Principal Component Analysis (PCA) as the filter function.

Figure 5.1: Point Cloud Illustration from [12]

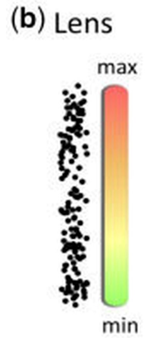


Figure 5.2: Filter Function Illustration from [12]

After the mapping, the cover $\mathcal{U} = \{U_i\}_{i=1}^n$, a collection of overlapping sets that cover the input data space, is then constructed. Each U_i is an open set in X such that $\cup_{i=1}^n U_i = X$. Within each overlap region U_{ij}, data points are clustered based on their filter function values. This clustering step groups together points that are similar according to the filter function.

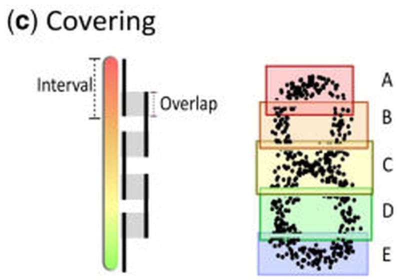


Figure 5.3: Covering Illustration from [12]

Clustering is conducted afterward before it is converted into a graph. In this paper, we will be using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) as our clustering algorithm as it can uncover clusters of diverse shapes and sizes in large datasets, even in the presence of noise and outliers.

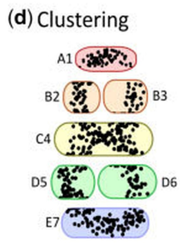


Figure 5.4: Clustering Illustration from [12]

Lastly, the output of the Mapper algorithm is a simplicial complex, which is a combinatorial object representing the topology of the data. It consists of vertices, edges, triangles, and higher-dimensional simplices, where each simplex corresponds to a cluster of data points.

However, the implementation of Mapper needs careful consideration of

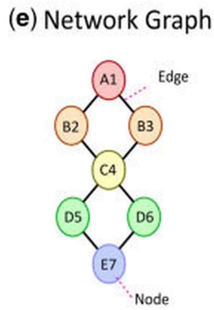


Figure 5.5: Mapper Output Illustration from [12]

several parameters as suggested in [13]. This is because the final geometric representation is highly sensitive to the chosen distance metric, lens function, clustering algorithm, and the parameters defining the intervals used (number and overlap percentage).[12]

5.1.3 Application of Mapper to the Dataset

We used the `make_mapper_pipeline` function of Giotto-TDA to create a data pipeline using 2-dimensional Principal Component Analysis (PCA) as the filter function, a cube-based cover with 10 intervals and 50% overlap fraction, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for clustering. Implementations of the PCA and DBSCAN algorithms were taken from the Scikit-Learn library which Giotto-TDA was designed to be compatible with. A `MapperInteractivePlotter` object was created using the data pipeline and each of the five datasets, and the adjustable graph was generated with the method `MIP.plot`. From this interface, the output of the Mapper algorithm can be changed by adjusting hyperparameters such as the cardinality of the cover and the overlap between each cover set. Lastly, we use the `color_data` argument in `MIP.plot` to represent position designations to generate insights related to basketball's positional trends and team performance.

5.2 Results and Analysis

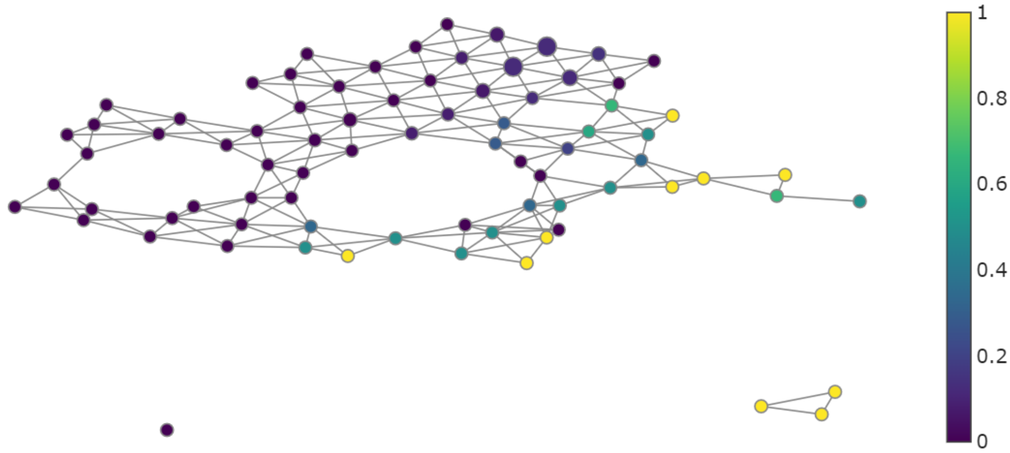


Figure 5.6: Mapper graph of the UAAP dataset, colored by proportion of centers in each node.

As shown in the figure above, the form of the graph generated by the Mapper algorithm for the UAAP dataset is of a monolithic graph with two hole structures, and two outlier graphs.

Outliers

The first outlier consists of a single node containing a single player, Isaiah Blanco of De La Salle University. He played one minute in only one game, but is noted for having one three-point shot made of one attempted for a 100% three-point accuracy. This factor separates him from the other players with low Games Played (GP); indeed, of six players with less than or equal to four Games Played, only Blanco and JC Fetalvero made any three-point shot.

Table 5.1: Outlier vs. League Average (Games Played $\leq$ 4)

Statistic	**Isaiah Blanco**	**League Average (GP $\leq$ 4)**
Games Played (GP)	1	3.17
Minutes Per Game (MPG)	1.49	4.10
Points Per Game (PPG)	3.0	1.08
Field Goals Made (FGM)	1.0	0.38
Field Goals Attempted (FGA)	2.0	1.26
Field Goal Percentage (FG%)	50.0	23.88
Three-Point Field Goals Made (3PM)	**1.0**	0.04
Three-Point Field Goals Attempted (3PA)	**1.0**	0.38
Three-Point Field Goal Percentage (3P%)	**100.0**	3.33
Free Throws Made (FTM)	0.0	0.33
Free Throws Attempted (FTA)	0.0	0.50
Free Throw Percentage (FT%)	0.0	37.50
Offensive Rebounds (ORB)	0.0	0.39
Defensive Rebounds (DRB)	0.0	0.33
Rebounds Per Game (RPG)	0.0	0.80
Assists Per Game (APG)	0.0	0.17
Steals Per Game (SPG)	0.0	0.12
Blocks Per Game (BPG)	0.0	0.08
Turnovers (TOV)	0.0	0.47
Personal Fouls (PF)	0.0	0.50

The second outlier consists of three nodes containing four players, which are Ange Kouame, Adama Faye, Michael Phillips, and Malick Diouf. All centers, they have been noted to be the most impactful players in the league with several Most Valuable Player and Mythical Five awards among them [7], [21]. Three of them are Foreign Student-Athletes (FSAs), who are students of fully foreign ancestry playing in the collegiate league under special limits. Particularly, since 2014, only one FSA is allowed per team in the league [22]. The non-FSA player, Michael Phillips, is also of foreign ancestry, being Filipino-American. In addition, all four of these players are of African descent.

In terms of statistics, the features by which the outlier players most exceed the league average are in blocks per game (BPG) at 654.10% above on average, and rebounding skills, particularly defensive rebounds per game (DRB) at 274.76% above, over offensive rebounds at only 248.23% above. However, the outliers struggle in three points made and attempted, being 63.52% and 63.97% below league average, respectively.

Table 5.2: Outliers vs. League Average and League Average Center

Statistic	**Node Avg**	**League Avg**	**% Diff**	**C Avg**	**% Diff**
GP	12.75	11.17	14.11%	12.56	1.55%
MPG	26.39	14.07	87.53%	16.47	60.17%
PPG	10.90	4.80	127.17%	6.21	75.65%
FGM	4.18	1.73	142.05%	2.42	73.03%
FGA	8.44	4.64	81.93%	5.03	67.69%
FG%	49.12	35.60	38.01%	44.36	10.75%
3PM	***0.18***	***0.49***	***-63.52%***	***0.14***	***23.16%***
3PA	***0.66***	***1.83***	***-63.97%***	***0.58***	***14.15%***
3P%	17.86	17.35	2.93%	9.42	89.66%
FTM	2.30	0.86	168.48%	1.23	87.16%
FTA	4.39	1.33	228.82%	2.20	99.31%
FT%	53.75	52.92	1.57%	55.76	-3.60%
ORB	**3.54**	**1.02**	**248.23%**	**2.11**	**68.24%**
DRB	**7.66**	**2.04**	**274.76%**	**3.67**	**108.91%**
RPG	**11.32**	**3.06**	**269.99%**	**5.78**	**96.01%**
APG	1.80	1.18	52.88%	0.82	118.92%
SPG	1.05	0.61	72.25%	0.55	90.91%
BPG	**2.00**	**0.27**	**654.10%**	**0.86**	**132.26%**
TOV	2.20	1.09	101.75%	1.10	100.00%
PF	2.55	1.37	86.19%	1.81	40.80%

Hole 1

The first hole consists of a cycle of 9 nodes that generally connect to their neighboring nodes but not to their non-neighboring nodes. The majority of the players (18 of 23) represented in these nodes are guards, but a substantial portion (5 of 23) are forwards.

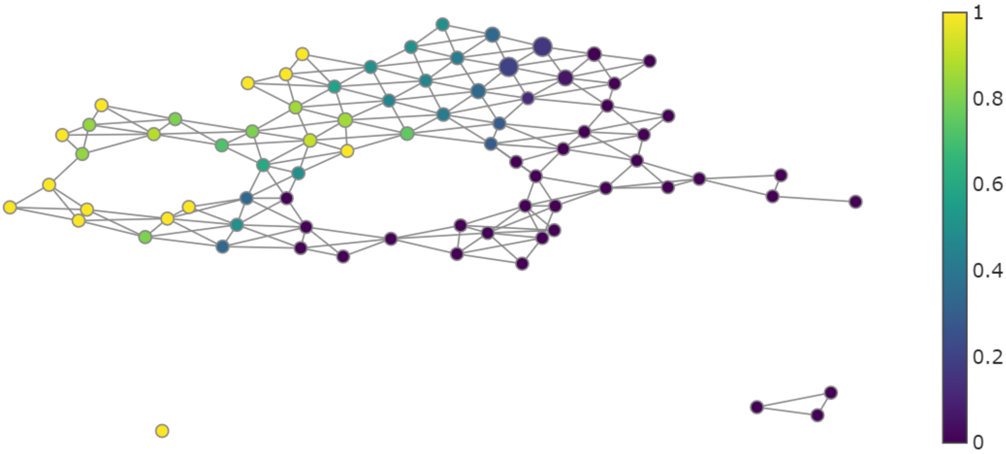


Figure 5.7: Mapper graph of the UAAP dataset, colored by proportion of guards in each node. Hole 1, on the left, is mainly surrounded by nodes representing guards.

We analyze the difference in variance between the players in the node and the whole league. There is inherently a decrease in variance expected since the nodes were formed by clustering, but the features with an increase or least decrease in variance denotes the dispersed features which separate nodes across each other along the hole. The top six features sorted by difference in variance are assists per game (APG), free throws made (FTM), turnovers per game (TOV), three points attempted (3PA) and made (3PM), and steals per game (SPG).

Incidentally, with the exception of TOV and FTM, which are replaced with field goals attempted (FGA) and points per game (PPG), four of these six features are those by which the mean of players in the node exceed the league average the most. In particular, the mean 3PM and 3PA of the players in the node exceed the league average by 186.42% and 164.14%, and the league average for guards by 72.06% and 63.96%, respectively. Meanwhile, the features by which the hole average player lags the league average player is in blocks per game (BPG) and field goal accuracy (FG%), by 36.07% and 3.18%, respectively. Despite this, the BPG of the hole average player is actually 85.71% better than the league average for guards.

Table 5.3: Hole 1 vs. League average

Statistic	**Hole Avg**	**Hole Variance**	**League Avg**	**% Diff**
GP	12.57	3.08	11.17	12.45%
MPG	23.05	6.90	14.07	63.83%
PPG	9.27	7.74	4.80	93.28%
FGM	3.17	1.04	1.73	83.46%
FGA	9.10	5.49	4.64	96.09%
FG%	34.47	42.68	35.60	-3.18%
3PM	**1.40**	**0.26**	**0.49**	**186.42%**
3PA	**4.84**	**2.82**	**1.83**	**164.14%**
3P%	29.33	32.35	17.35	69.06%
FTM	**1.51**	**0.57**	**0.86**	**76.49%**
FTA	**2.09**	**1.01**	**1.33**	**56.62%**
FT%	72.03	180.73	52.92	36.11%
ORB	1.05	0.28	1.02	3.46%
DRB	2.72	0.90	2.04	33.25%
RPG	3.76	1.88	3.06	22.87%
APG	**2.76**	**1.57**	**1.18**	**134.49%**
SPG	**1.22**	**0.20**	**0.61**	**100.43%**
BPG	0.17	0.06	0.27	-36.07%
TOV	**1.95**	**0.56**	**1.09**	**78.63%**
PF	1.79	0.33	1.37	30.79%

This characterization of the hole average player matches that of the "3-and-D" specialist player which was developed and is now common in the NBA. According to Joseph [8], the 3-and-D player is a type of player whose responsibilities are almost entirely limited to perimeter shooting (i.e., 3PA and 3PM) on the offensive end, and exceptional defense (i.e., SPG and BPG) on the other end. They may also be expected to take secondary ball distribution roles, but this is not a priority, which could correlate to a moderate increase in APG with a corresponding increase to TOV because of an undeveloped, high assist-to-turnover ratio.

Hole 2

The second hole, meanwhile, consists of a cycle of 11 nodes. The number of guards (10) and forwards (9) represented in the hole are roughly equal; additionally, the hole comprises 4 center players.

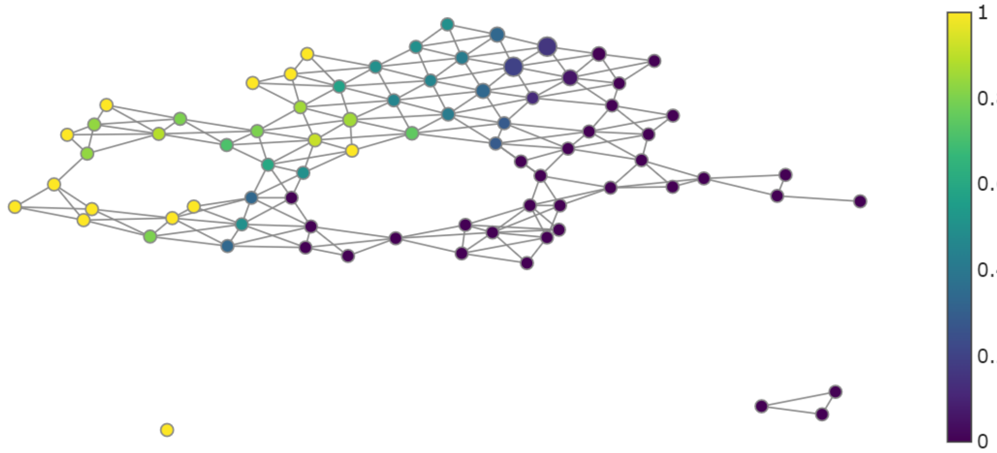


Figure 5.8: Mapper graph of the UAAP dataset, colored by proportion of forwards in each node. Hole 2, on the right, is mainly surrounded by nodes representing a mix of guards and forwards.

By difference of variance, the top six features are similar to, or related to, those of hole 1; field goals made (FGM), free throws attempted (FTA), FTM, PPG, SPG and FGA. However, FTA and SPG are only fourth and fifth of the top six features sorted by percentage difference to league average, with the more prominent features being offensive rebounds (ORB), rebounds (RPG), and defensive rebounds (DRB) per game. Minutes per game (MPG) at sixth, and personal fouls per game (PF) at seventh, are also substantially higher than league average, but in general, these differences are less intense (35-50%) than those of hole 1 (90-200%). The features by which the hole 2 average player are most behind the league average player are, interestingly, the defining features presented by hole 1, being 3PA, 3.91% less, and 3PM, 12.33% less than league average.

Table 5.4: Hole 2 vs. League average

Statistic	**Node Avg**	**Node Variance**	**League Avg**	**% Diff**
GP	13.58	0.51	11.17	21.56%
MPG	19.19	25.79	14.07	36.41%
PPG	6.19	9.15	4.80	28.95%
FGM	2.29	1.49	1.73	32.61%
FGA	5.78	6.45	4.64	24.54%
FG%	38.74	103.72	35.60	8.84%
3PM	0.43	0.11	0.49	-12.33%
3PA	1.76	1.25	1.83	-3.91%
3P%	20.64	145.72	17.35	18.99%
FTM	1.16	0.42	0.86	35.01%
FTA	1.85	0.86	1.33	39.01%
FT%	61.59	190.59	52.92	16.38%
ORB	**1.52**	**0.37**	**1.02**	**49.37%**
DRB	**2.86**	**1.34**	**2.04**	**39.76%**
RPG	**4.39**	**2.53**	**3.06**	**43.34%**
APG	1.48	0.58	1.18	25.28%
SPG	**0.85**	**0.16**	**0.61**	**38.76%**
BPG	0.28	0.06	0.27	6.83%
TOV	1.42	0.38	1.09	29.92%
PF	**1.85**	**0.27**	**1.37**	**35.08%**

The statistics described by the hole 2 average player signify a player with strengths particularly in rebounding, stealing and drawing free throws. In particular, the higher prominence of offensive rebounding (49.37% over league average) against defensive rebounding (39.76% over league average) signifies a player more deeply involved in the offensive end of the court, being able to hustle, or chase the rebound after a teammate misses a shot. Furthermore, a high number of free throw attempts and personal fouls is associated with a type of player called a "foul-drawing player", known for being aggressive on both ends of the court to force opponents to make illegal contact moves, which result in free throw attempts. Herring [6] uses the example of Trae Young, a guard player of short stature in the NBA, citing situations where "he initiated contact by pressing his shoulder into the defender trying to keep him out of the paint, or abruptly hit the brakes after dribbling past a wing stopper, causing that man to crash into him as a result." In increasing the number of free throw attempts by forcing opponents to commit fouls, aggressive players like Young are simultaneously more likely to be called for their own fouls.

Hole Filling

Roehm [17] uses the example of Kevin Durant, characterizing him as a "hole-filling" player. Plotting the roster of the NBA's Golden State Warriors during the the 2016 and 2017 season by total rebound percentage (TRB%) and three-point attempt rate (3PAr), he discovers a hole with cycle count of 4 present in the 2016 season. But, with the addition of Kevin Durant to the roster in 2017, the hole is reduced to a cycle count of 3.

Based on this method, we measure the effect of including the missing archetypes discovered through the hole analysis by applying the Mapper algorithm with the same parameters on an augmented UAAP dataset where the means of each hole are appended as synthetic player data.

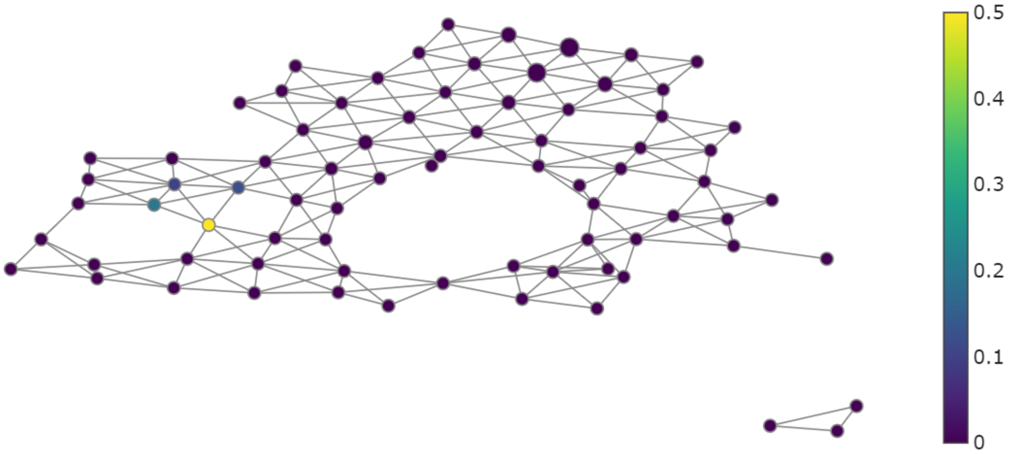


Figure 5.9: Mapper graph of the UAAP dataset, where the synthetic hole-filling player of hole 1 is highlighted in yellow.

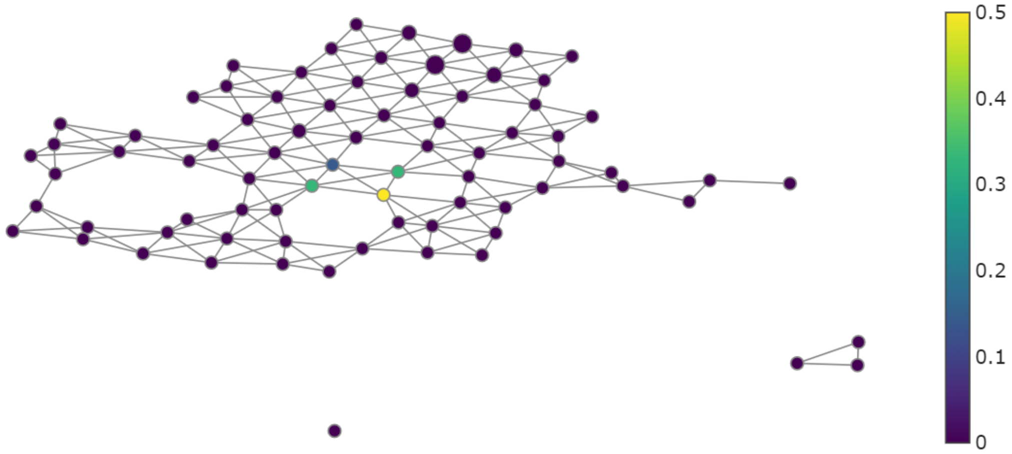


Figure 5.10: Mapper graph of the UAAP dataset, where the synthetic hole-filling player of hole 2 is highlighted in yellow.

From a cycle count of 9, hole 1 decreases to a cycle count of 6, and from a cycle count of 11, hole 2 decreases to a cycle count of 6. This supports the missing archetype players being sufficiently different from the players comprising the hole nodes. In the context of basketball, these players fill an underrepresented niche in the league. Therefore, a team that includes players fitting these missing archetypes could gain an advantage by introducing this new niche which is presently unaccounted for by other teams' analysis, strategy and tactics.

Chapter 6

Conclusions and Recommendations

6.1 Conclusions

We successfully extracted insights from the UAAP dataset by analyzing outliers and holes in a graph generated by the Mapper algorithm. We discover two outliers which represent a single player with a standout three-point shot, and four high-output centers with Most Valuable Player-class skills. Then, we discover missing archetypes of players through holes present in the Mapper graph, being the "3-and-D", three-point shooters with stealing and assisting skills, and "foul drawing" players with a particular skill in offensive rebounding and getting free throws by forcing fouls. Using hole-filling analysis, we discover that including these types of players in the dataset successfully reduces the cycle count of each hole, proving that they fill underrepresented niches in the UAAP basketball league.

6.2 Research Recommendations

By following these recommendations, stakeholders can not only address the current gaps identified in the UAAP dataset but also enhance the strategic depth and competitiveness of the league. Future studies may focus on the practical application of these insights in real-world scenarios and the development of predictive models for player performance and team success.

6.3 Recommendations for Further Work

Throughout this study, our investigation into the UAAP dataset has been driven by a rigorous application of the Mapper algorithm to analyze player archetypes within collegiate basketball. Our findings have revealed both outliers and holes that signify standout individual performances and underrepresented player types, respectively. While our analysis has succeeded in extracting valuable insights from the dataset at hand, it is important to acknowledge the inherent limitations that have shaped our conclusions.

One of the primary constraints encountered in this research was the limited scope of the dataset. The dataset's coverage, both in terms of the number of seasons and the breadth of player statistics, inherently restricted our ability to fully explore the dynamics and evolution of player roles within the UAAP. Additionally, the absence of certain advanced metrics, which could provide deeper insights into player contributions and team strategies, was a notable limitation.

Moreover, our analytical approach, while robust, opens the door to further exploration using more advanced statistical techniques and machine learning models. The potential for temporal analysis and predictive modeling was not fully realized within the confines of this study but represents a promising avenue for future research.

Given these considerations, the forthcoming recommendations are crafted with an eye towards not only addressing the gaps identified through our analysis but also expanding the horizons of future research in sports analytics within the UAAP and beyond. By enhancing data collection practices, integrating sophisticated analytical methods, and applying the insights gained to strategic decision-making, we can further enrich our understanding of the complex interplay between player characteristics, team dynamics, and league-wide trends.

6.3.1 Enhancing Data Collection and Quality

- To capture a more holistic understanding of player performance, it is advisable to broaden the dataset. Including a wider range of seasons and integrating granular player statistics will help identify subtle trends and the emergence of new player archetypes.

- Integrating advanced metrics such as Player Efficiency Rating (PER), Win Shares, plus-minus ratings, and other defensive statistics can provide a more complex view of player contributions that may not be apparent through traditional statistics.

6.3.2 Advanced Analytical Techniques

- A temporal analysis can reveal how player roles and effectiveness change over seasons, during playoff pressures, or even within a single game. This can inform strategic decisions in real time.

- Predictive modeling could be employed to forecast player development trajectories and team performance, which would be instrumental in strategic planning and long-term team development.

6.3.3 Strategic Implications

- Training programs could be customized to address the skills gaps identified in the analysis. Developing underrepresented skills in players can enhance their versatility and value.

- Recruitment strategies may need adjustments to fill the identified player archetype gaps, leading to a balanced and strategically diverse team.

6.3.4 Comparative and Cross-disciplinary Analysis

- Comparing the UAAP with other basketball leagues using the Mapper algorithm could highlight unique attributes of the league and offer broader insights.

- Applying the analytical framework to other sports may reveal underlying principles of team dynamics and player roles that transcend basketball.

Bibliography

[1] Muthu Alagappan. *From 5 to 13: Redefining the Positions of Basketball – Sloan Sports Conference*. https://web.math.utk.edu/~fernando/Students/GregClark/pdf/Alagappan-Muthu-EOSMarch2012PPT.pdf. Accessed: [April 1, 2024]. 2012.

[2] Liang Cheng. "The Application of Topological Data Analysis in Practice and Its Effectiveness". In: *E3S Web of Conferences*. Vol. 214. EDP Sciences. 2020, p. 03034.

[3] Nathan Joel Diambra. "Using topological clustering to identify emerging positions and strategies in NCAA men's basketball". In: (2018).

[4] Albert Ratera Gispets. "Extracting insights from the shape of EuroLeague data using statistics and topology". Treball final de grau. Grau de Matemàtiques. Barcelona: Universitat de Barcelona, June 2021.

[5] Alexander L. Hedquist. "Redefining NBA Basketball Positions Through Visualization and Mega-Cluster Analysis". Available at DigitalCommons@USU. MS Thesis. Utah State University, Aug. 2022. URL: https://digitalcommons.usu.edu/etd/8602.

[6] Chris Herring. *Trae Young and the Art of Drawing Fouls in the NBA Playoffs*. Accessed: 2024-03-28. 2021. URL: https://www.si.com/nba/2021/06/10/trae-young-hawks-nba-playoffs-foul-drawing.

[7] JR Isaga. *UP's Diouf wins MVP as Katipunan dominates UAAP Season 85 awards*. Accessed: 2024-03-28. 2022. URL: https://www.rappler.

com/sports/uaap/malick-diouf-mvp-up-ateneo-dominate-season-85-men-basketball-awards/.

[8] Adi Joseph. *'3-and-D': The specialist's path to a long NBA career*. Accessed: 2024-03-28. 2014. URL: https://www.usatoday.com/story/sports/nba/2014/11/26/three-and-d-specialists-kyle-korver-garrett-temple-martell-webster-willie-green-the-next-bruce-bowen/70123886/.

[9] A. Jyad. "Redefining NBA Player Classifications Using Clustering". In: *Towards Data Science* (Nov. 2020). Accessed: [Insert access date here]. URL: https://towardsdatascience.com/redefining-nba-playerclassifications-using-clustering-36a348fa54a8.

[10] S. Kalman and J. Bosch. "NBA Lineup Analysis on Clustered Player Tendencies: A New Approach to the Positions of Basketball & Modeling Lineup Efficiency". In: *MIT Sloan Sports Conference*. Accessed: [Insert access date here]. 2020. URL: https://www.sloansportsconference.com/research-papers/nba-lineup-analysis-on-clustered-player-tendencies-a-new-approach-to-thepositions-of-basketball-modeling-lineup-efficiency.

[11] Michael Lewis. *The Undoing Project*. W. W. Norton Company, 2016.

[12] Ciara F Loughrey et al. "The topology of data: opportunities for cancer research". In: *Bioinformatics* 37.19 (2021), pp. 3091–3098.

[13] P. Y. Lum et al. "Extracting insights from the shape of complex data using topology". In: *Scientific Reports* 3 (2013), p. 1236. DOI: 10.1038/srep01236. URL: https://www.nature.com/articles/srep01236.

[14] Ralph Joshua P. Macarasig et al. "Topological Data Analysis of Collective Behavior in Public Transportation". Undergraduate Thesis. Ateneo de Manila University, 2022.

[15] Ved Phadke and Ollie Pai. *NBA's 3-point revolution: How 1 shot is changing the game*. https://www.bruinsportsanalytics.com/post/neo_positions. Accessed: 2024-03-24. 2021.

[16] Albert Ratera Gispets. "Extracting insights from the shape of EuroLeague data using statistics and topology". In: (2021).

[17] Jerome Roehm. *An Application of TDA to Professional Basketball*. Presented by Applied Algebraic Topology Network. https://www.youtube.com/watch?v=-cfp-tH-vIM. Accessed: access-date. 2021.

[18] John Schuhmann. *NBA's 3-point revolution: How 1 shot is changing the game*. https://www.nba.com/news/3-point-era-nba-75. Accessed: 2024-02-26. 2021.

[19] Christopher Shultz. "Applications of Topological Data Analysis in Economics". In: *Available at SSRN 4378151* (2023).

[20] Gurjeet Singh, Facundo Mémoli, Gunnar E Carlsson, et al. "Topological methods for the analysis of high dimensional data sets and 3d object recognition." In: *PBG@ Eurographics* 2 (2007), pp. 091–100.

[21] The Game Team. *The Top Players Of The UAAP Season 86 Men's Basketball Tournament*. Accessed: 2024-03-28. 2023. URL: https://thegame-onemega.com/the-top-players-of-the-uaap-season-86-mens-basketball-tournament/.

[22] Reuben Terrado. *Source: UAAP set to restrict number of foreign players to only one per school.* Accessed: 2024-03-28. 2023. URL: https://www.spin.ph/basketball/uaap-men/source-uaap-set-to-restrict-number-of-foreign-players-to-only-one-per-school-v02.

[23] S Trninic and D Dizdar. "System of the Performance Evaluation Criteria Weighted per Positions in the Basketball Game". In: *Hrčak: časopis za sport i zdravstvenu zaštitu* (2000). URL: https://hrcak.srce.hr/file/15447.

[24] University of Waterloo. *What is Topology?* https://uwaterloo.ca/pure-mathematics/about-pure-math/what-is-pure-math/what-is-topology. Accessed: 2024-03-24.

[25] *Where Basketball was Invented: The History of Basketball.* https://springfield.edu/where-basketball-was-invented-the-birthplace-of-basketball. Accessed: 2024-02-26.