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
April 20, 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 20, 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. — Date
Adviser

Job A. Nable, Ph.D. — Date
Adviser

Eden Delight P. Miro, Ph.D. — Date
Department Chair

Ralph Joshua Macarasig — Date
Panelist

Victor Andrew Antonio, Ph.D. — Date
Panelist

Acknowledgments

Our deepest gratitude is extended to our advisors, Dr. Clark Kendrick Go and Dr. Job Nable, whose guidance and support have been pivotal since the inception of our thesis idea through to the finalization of our manuscript. Their unwavering dedication and expertise have been the bedrock of this work.

We are also immensely thankful to our panelists, Dr. Victor Antonio and Mr. Macarasig. Their insightful critiques and thoughtful suggestions have been invaluable, significantly contributing to the depth and rigor of our research.

A special acknowledgment goes to Julia Abadilla and the many friends who graciously helped us with sourcing data.

And last but not the least, we also express our profound gratitude to our friends and families, who have provided us with the necessary support and encouragement throughout this academic journey.

Ad majorem Dei gloriam

Summary of the Thesis

Basketball is one of the most popular team sports, whose many competitive, well-organized and well-recorded leagues provide rich sources of data frequently used to analyze tactics, strategy, player performance and team composition. A consequence of basketball's team dynamics, comprising five active players per team on the court who may be replaced by substitute players on many occasions per game, is the formation of archetypes; players are categorized into positions based on roles they play, such as point-scoring, rebounding or defense. However, the evolution of game strategy has meant that the three traditional positions, namely—guard, forward, and center, are no longer sufficient to completely describe roles of players within the team.

Understanding the archetypes of basketball players offers an advantage to teams in strategizing against their opponents, and finding and fixing weaknesses in their team composition, whether by adding skills to current players or recruiting new talent. Usually a matter of subjective observation, especially in non-professional leagues such as the collegiate University Athletics Association of the Philippines (UAAP), we undertake the task of finding groups of players who fit in certain positions or archetypes more objectively with mathematical techniques from topological data analysis (TDA).

We use numerical data measuring player performance during the 84th season of the UAAP basketball tournament, with a high number of features per player corresponding to different player actions such as points, rebounds, assists, steals, blocks, turnovers, fouls and shot attempts. Using this high-dimensional dataset as a point cloud, we use Vietoris-Rips filtration to connect nodes representing players from progressively longer distances, discovering the presence of hole complexes that imply groupings or clusters of players. We study our dataset further using the Mapper algorithm which applies dimensionality reduction and clustering to more prominently display topological features as connections between nodes representing multiple players.

We are able to identify four subgroups of players not previously indicated in the five traditional basketball positions; paint dominators, stretch forwards, defensive rebounders, and scrappers. Furthermore, we discover two hole structures of a much larger prominence than others in the mapper plot, associating them with missing archetypes of players. These are three-point plus defensive specialists, and aggressive foul-drawers, whose attributes are currently underrepresented within the league, and thus could make a team to strategize against. We conclude that TDA techniques can identify archetypes of players beyond the traditional classifications and can provide advantageous insights for basketball teams.

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 [13]. Given this context, advanced analytics emerges as a crucial tool. It 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 [18] employed model-based clustering on data spanning the 2009-2010 to 2018-2019 National Basketball Association (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 [17] 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 [13] 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 (TDA) 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 [11] applied the same methodology to the EuroLeague. While Nathan Joel Diambra [5] used topology to delve 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 Asian basketball leagues like the University Athletic Association of the Philippines (UAAP) basketball championships 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 Topological Data Analysis (TDA) to player data from UAAP Season 85, which spanned from October 1, 2022, to December 19, 2022, 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 beyond the larger leagues.

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 Organization of the Thesis

Our thesis is organized into the following chapters: Chapter 2 provides a comprehensive review of the related literature. Chapter 3 defines some key concepts in topology, algebra, and homology, which serve as the mathematical foundation of our study. Chapter 4 discusses the dataset and details the data preprocessing methods employed. In Chapter 5, we explore persistent homology and present the results obtained from its application. Chapter 6 applies the mapper algorithm to visualize the dataset. Finally, Chapter 7 summarizes our findings and offers recommendations.

Chapter 2

Review of Related Literature

2.1 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.[37]

- 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 2000s. 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.

2.2 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, 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. [20]

2.3 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.

2.3.1 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 understand how the neo-position itself affects players, the study compares players of different skill levels who play the neo-position to those who play in traditional positions. Statistical metrics like win shares/48, assist percentage, usage rate, three-point attempts, and field goal attempts are used for the analysis. [26]

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. Players within the traditional position are then categorized by their level of play. 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 into 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.

2.3.2 Mega clustering

Hedquist, aimed to redefine NBA basketball positions through visualization and mega-cluster analysis. [13] 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." Nine "mega-clusters" were identified: 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.

2.4 Topological Data Analysis

2.4.1 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. [38]

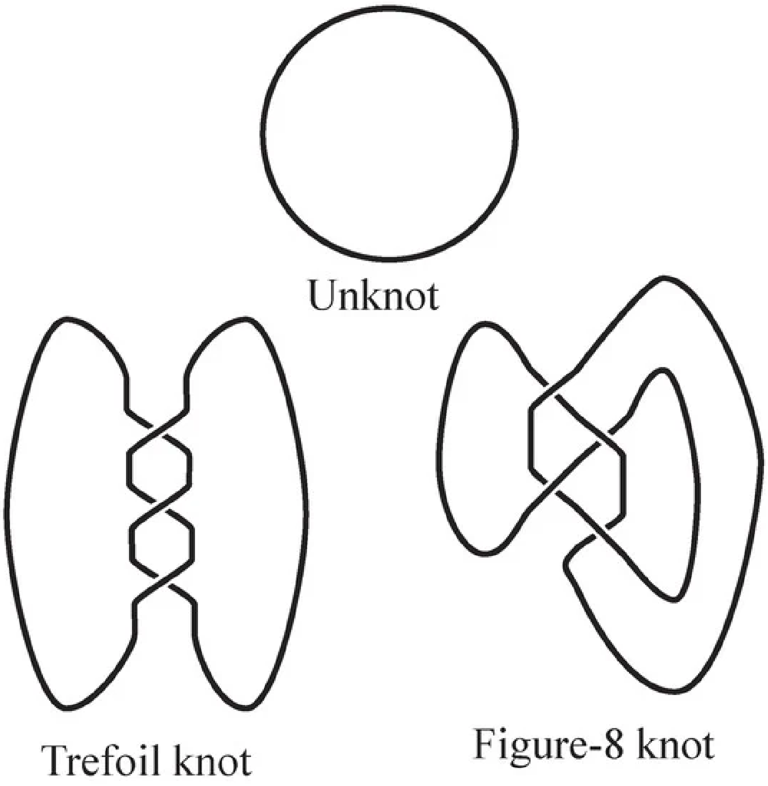


Figure 2.1: The unknot, the trefoil knot, and the figure-8 knot.

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. [24] Topology’s strength in pattern recognition via shape analysis hinges on three core concepts. Firstly, it operates in metric spaces where it is independent of the chosen coordinate system, allowing comparisons across different data platforms. Secondly, topology focuses on making it less sensitive to noise. Mathematically, this allows circles, ellipses, and hexagons to be considered equivalent. Finally, topology utilizes finite networks that capture essential features while discarding minor details. [24]

2.4.2 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. A 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 a way to grasp the structure of complex data. [4]

The value of topological data analysis is reflected across various fields as it has yielded promising results in highly different areas. Presented here is 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 [30]. In neuroscience, TDA has aided in tasks like identifying brain cavities [32], analyzing event-related fMRI data [8], and distilling complex neuroimaging data [10]. Additionally, it has shown promise in medical image processing for patient classification [9] and exploration of hyperspectral imaging data [7]. In the natural sciences, applications range from generating large-scale genomic recombination maps [3] to understanding biological aggregation models [36] and modeling the spread of Zika virus [22]. Its reach extends to physics (fast radio burst analysis) [25], engineering (aviation data exploration) [21], and signal processing (movie genre detection) [6]. Notably, it has also been applied in operations research, facilitating the identification of clusters in social networks [2] and the detection of structural information in manufacturing data [12].

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. However, domain knowledge is highly important as it can inform the selection of appropriate parameters that lead to more robust and meaningful conclusions. 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.

2.4.3 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 [24]. From the breast cancer datasets, NKI and GSE2034, they discovered that a sub-group of breast cancer patients with low ESR1 expression and elevated expression levels of genes in the immune pathways are easily detected using their methodologies that other clustering approaches could not easily do. The methodology was then applied to the 22 years of voting records from the members of the US House of Representatives to observe their voting behavior over the years. In this part, they showed that using principal component analysis has its limitations compared to the Mapper algorithm as it was only able to divide the Republicans and the Democrats. On the other hand, the Mapper algorithm not only divided the two but also was able to reveal subgroups within. 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 variance-normalized Euclidean distance metric and the principal and secondary SVD values as the filter for the Mapper Algorithm, 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.

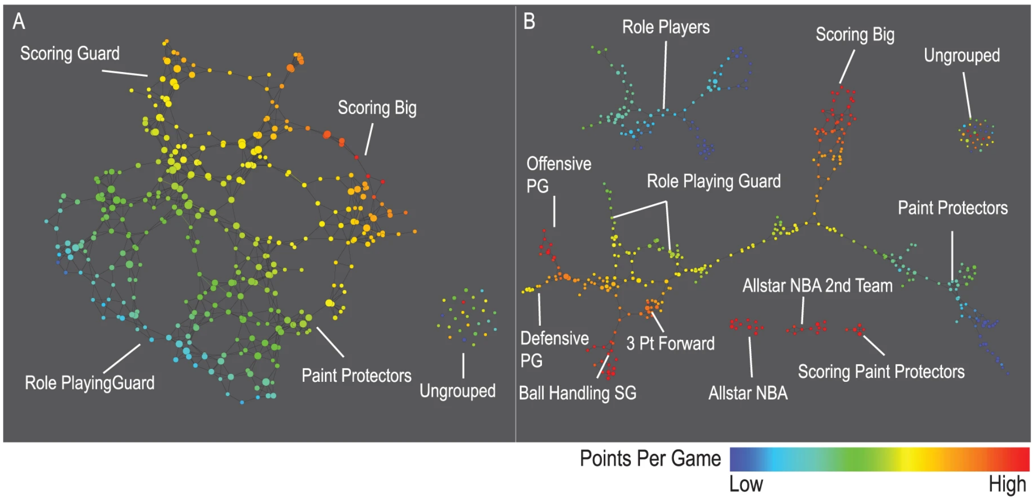


Figure 2.2: Two visualizations of NBA player clustering constructed using Singular Value Decomposition (SVD) filtering. Panels A and B used 20 intervals and 30 intervals resolution maps, respectively. Each panel utilizes both the principal and secondary SVD components to create the landscape. Overlap between adjacent intervals is set at 50%.

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

The pioneering study has catalyzed subsequent research endeavors. A notable example is the study by Albert Ratera Gispets which was able to discern the underlying subgroups of EuroLeague players from the 2019–2020 season by integrating statistical techniques with TDA [11]. 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 datasets that 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 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 [11].

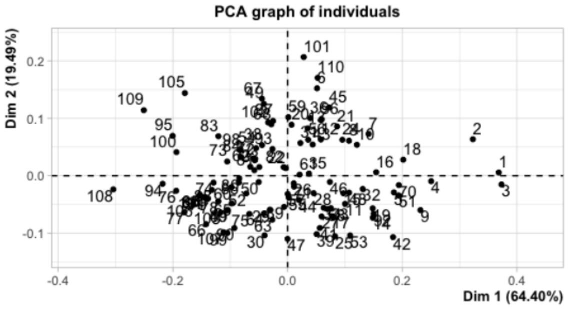


Figure 2.3: Distribution of Guards from EuroLeague 2019-2020 in R^2 using the first two principal components as axes which also keep 83.9% of the variance

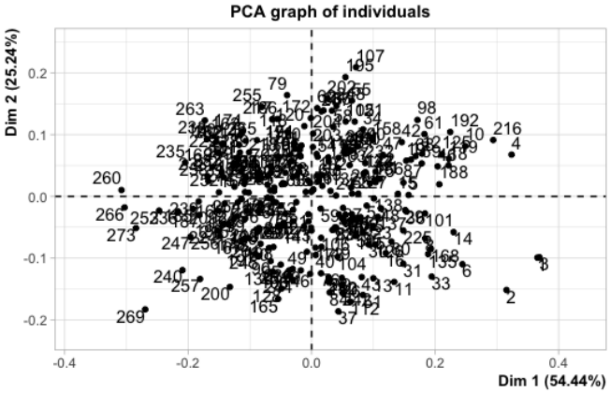
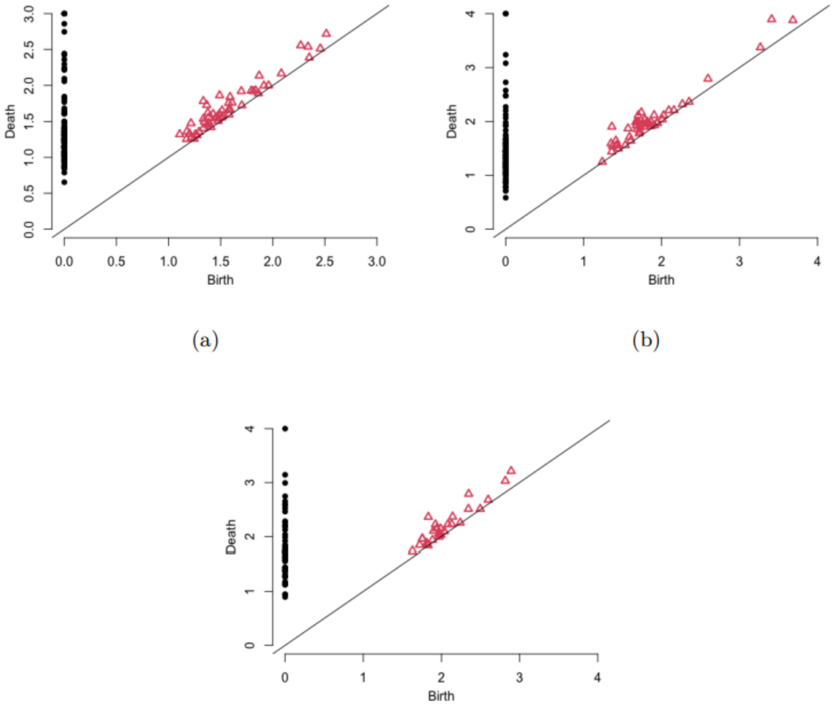


Figure 2.4: Distribution of All Players from EuroLeague 2019-2020 in R^2 using the first two principal components as axes which also keep 79.6% of the variance

The paper also used persistent homology to investigate the existence of latent subgroups of a specific position [27]. 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 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



(a)

(b)

Figure 2.5: Persistence Diagrams for guards (a), forwards (b), centers (c) from EuroLeague 2019-2020 normalized using variance normalized Euclidean distance computed using Vietoris-Rips complexes.

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.

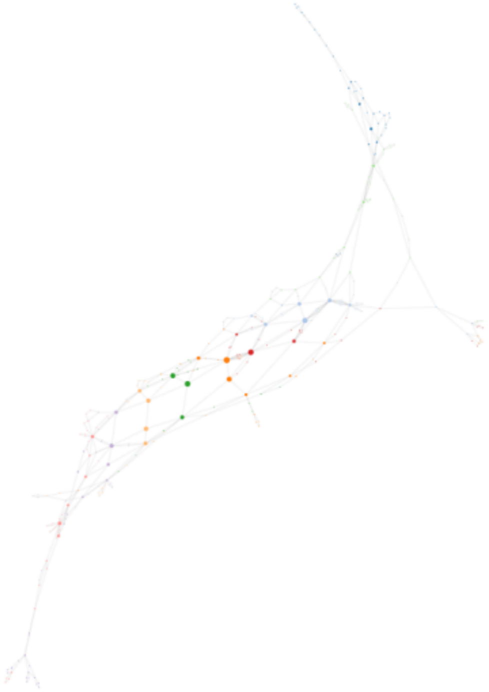


Figure 2.6: Shape of EuroLeague 2019-2020 generated using Mapper Algorithm with three filter functions and 20 intervals where nodes contain one or more players with similar player styles and nodes are connected if they share at least one player.

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 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 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 [5].

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

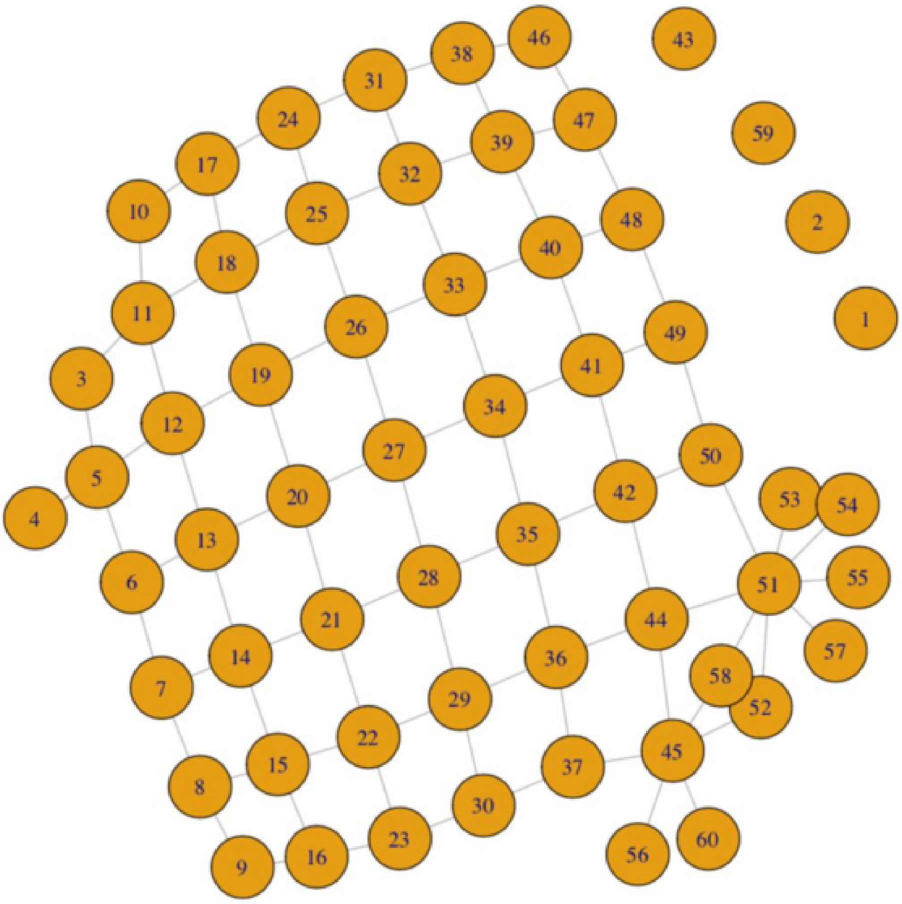


Figure 2.7: NCAA Division 1 Topological Mapping where each cluster grouped one or more players, examined individually and grouped by similar statistics

Chapter 3

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.

3.1 Topology

3.1.1 Metric Spaces

Definition 3.1.1. 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 $x = y$,
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 set X where the metric is defined as $d(x, y) = 1$ if $x \neq y$, and $d(x, x) = 0$ (This set is called a discrete metric space).

Open Sets and Closed Sets

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

Definition 3.1.2 (Ball)**.** Let X be a metric space. A *ball* B of radius r around a point $x \in X$ is $B(x, r) = \{y \in X \mid d(x, y) < r\}$. This ball contains all points whose distance is less than r from x.

Definition 3.1.3 (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. It can be easily shown that O is the union of balls around points $x \in O$.

Example 3.1.4. Let $X = \mathbb{R}$ with the standard Euclidean distance function $d(x, y) = |x - y|$. The interval $(0, 3/4)$ is open in X because, for any point z in $(0, 3/4)$, there is a radius $\epsilon > 0$ such that the ball $B(z, \epsilon) = \{y \in X \mid |y - z| < \epsilon\}$ is completely contained within $(0, 3/4)$.

Example 3.1.5. Let $X = \mathbb{R}$ with the Euclidean metric $d(x, y) = |x - y|$. The interval $[0, 1/2]$ is not open in X since no ball centered at 0 with a positive radius lies completely within $[0, 1/2]$, as it will always include points less than 0.

Note: The standard Euclidean metric $d(x, y) = |x - y|$ is commonly assumed for intervals on the real number line, which is why it is often omitted in textbooks.

Definition 3.1.6. A set C is *closed* if its complement $X - C$ is open.

3.1.2 Topological Spaces

Definition 3.1.7 (Topology)**.** A *topology* τ on a set X consists of subsets of X satisfying the following properties:

1. The empty set $\varnothing$ 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 3.1.8. Let $X = \{a, b, c\}$ and a proposed topology $\tau = \{\varnothing, X, \{a\}, \{b\}\}$. τ fails to meet property 2 of 3.1.7 as the union $\{a\} \cup \{b\} = \{a, b\}$ is not included in τ.

Example 3.1.9. Let X be a metric space. The set τ of all unions of open balls in X form a topology on X. This is called the metric topology.

Definition 3.1.10 (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: $\partial A = \overline{A} \setminus \mathring{A}$.

Definition 3.1.11 (Compact Set)**.** A subset K of a topological space X is called *compact* if every open cover of K has a finite subcover. This means that if K is covered by a collection of open sets $\{O_i\}_{i\in I}$, then there exists a finite subset J of I such that the sets $\{O_j\}_{j\in J}$ still cover K.

3.1.3 Homeomorphism

Definition 3.1.12 (Neighborhood)**.** Given a point x of X, a subset N of X is called a *neighborhood* of x if there exists an open set O such that $x \in O \subseteq N$.

Definition 3.1.13 (Continuous Function)**.** A continuous function $f : X \to Y$ between topological spaces is *continuous* if $f^{-1}(V)$ is open in X whenever V is open in Y. A function $f : X \to Y$ is also considered continuous if for any neighborhood V of Y, there exists a neighborhood U of X such that $f(U) \subseteq V$.

Remark 3.1.14. The composition of two continuous functions is also continuous.

Definition 3.1.15 (Homeomorphism)**.** A *homeomorphism* is a continuous function $f : X \to Y$ between two topological spaces X and Y such 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 3.1.16. Homeomorphism is the notion of equality in topology as it induces an equivalence relation among toplogical spaces.

Example 3.1.17. 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 3.1.18. 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.

Remark 3.1.19.

1. Continuous functions $f : X \to Y$ between topological spaces are important for several reasons. For instance, if X is a compact topological

space (a sort of finiteness condition), then its image $f(X)$ is compact. This result is the basis of the extreme value theorem in calculus. As another example of the importance of continuity, if X is connected (a condition that guarantees X is in "single piece"), then $f(X)$ remains connected. This result forms the basis of the intermediate value theorem in calculus.

2. Generic examples of continuous operations (functions) are translations, rotations, scalings (zoom in, zoom out), projections (shadows), chirping (circles into ellipses), and all sorts of transforms (Fourier Transform, Laplace Transform, etc.).

3. For most applications, and even theoretically, a homeomorphism is too restrictive. It is usually replaced by notions which are more relaxed but still preserve most important topological properties, such as homotopy.

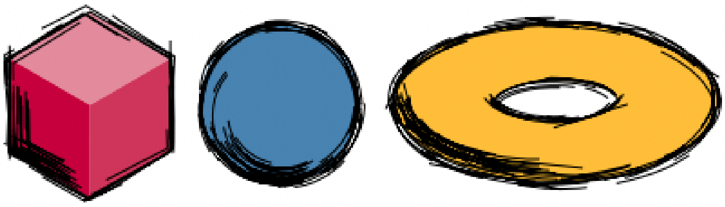


Figure 3.1: Examples of homeomorphic and non-homeomorphic objects. Homeomorphism implies that two objects can be continuously deformed into each other without tearing or cutting. In the illustration, the cube (left) can be transformed into the sphere (middle) through such deformations, indicating they are homeomorphic. In contrast, both the cube and sphere cannot be transformed into the torus (right) without introducing a discontinuity like a cut, due to the torus's central void, demonstrating non-homeomorphism [28].

3.2 Complexes

Consider a dataset with finitely many points. The geometric and topological features of the data come from the assumption that the data points are sampled from a manifold or a topological space.

The topological features will be obtained by constructing a simplicial representation coming from the data points. However, due to the possible presence of noise in the data, the philosophy of persistence will be used to get the true or persistent features. Thus, the need for the notion of simplicial complexes.

3.2.1 Simplicial Complexes

Definition 3.2.1 (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 3.2.2 (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 3.2.3 (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. Equivalently, a set of points $p_0, p_1, \ldots, p_k \in \mathbb{R}^d$ is *affinely independent* if

$$\alpha_1(p_1 - p_0) + \alpha_2(p_2 - p_0) + \cdots + \alpha_k(p_k - p_0) = 0$$

implies that

$$\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0.$$

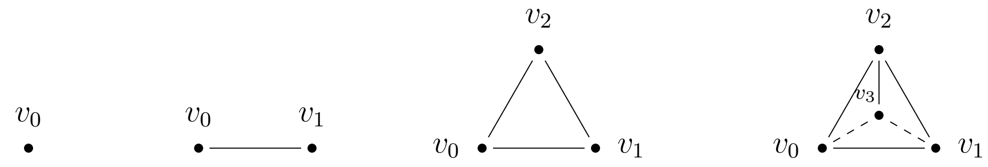
Definition 3.2.4 (Simplex)**.** A *k-dimensional simplex* σ in $\mathbb{R}^d$ is the convex hull of an affinely independent set of $k+1$ points, denoted by $S = \{v_0, v_1, \ldots, v_k\}$. These points are known as the vertices of the simplex. Thus, a *k-simplex* σ can be expressed as

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

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

Remark 3.2.5. 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 3.2.6 (Face)**.** Consider $S = \{v_0, v_1, \ldots, v_k\}$, a set of affinely independent vertices, which constructs a k-simplex, denoted as σ. A simplex

τ, resulting from a subset $T \subseteq S$, is termed a *face* of σ. We may refer to the set S by its vertices as $v_0v_1 \dots v_k$.

Definition 3.2.7 (Simplicial Complex)**.** A simplicial complex K 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.

Remark 3.2.8. The dimension of a simplex σ denotes $n+1$ vertices comprising the simplex, forming shapes such as edges, triangles, or tetrahedrons.

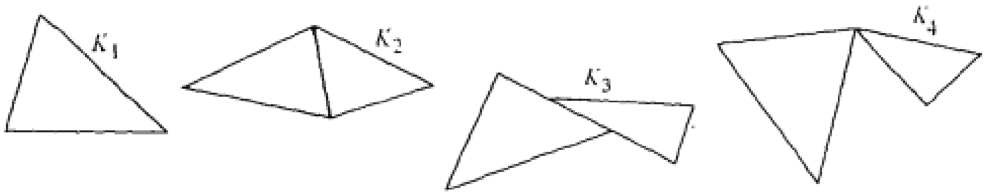


Figure 3.2: Simplicial Complex vs. Nonsimplicial Complex: K_1 and K_2 illustrate proper simplicial complexes, K_3 demonstrates an invalid intersection not forming a simplex, and K_4 shows a valid simplicial complex structure.

Example 3.2.9. In Figure 3.2, K_1 and K_2 are examples of simplicial complexes because they adhere to the definition that their simplices meet only at common faces. On the contrary, K_3 does not qualify as a simplicial complex since the intersection of its simplices (appearing to be a line segment) is not a common face, violating the definition. Lastly, K_4 returns to form as a valid simplicial complex with simplices intersecting only at shared vertices.

Definition 3.2.10 (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 3.2.11. 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. (The concept of a nerve and an open cover will be introduced in detail in Section 3.2.3.)

3.2.2 Subcomplexes and Closures

Definition 3.2.12 (Subcomplex)**.** Let K be a simplicial complex. A subset $L \subseteq K$ of simplices is called a *subcomplex* of K if for each simplex τ in L, if σ is a face of τ in K, then σ also belongs to L.

Definition 3.2.13 (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 3.2.14 (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 $\{\phi(v_0), \ldots, \phi(v_k)\}$.

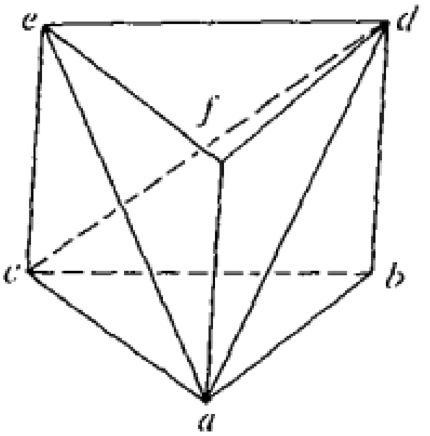


Figure 3.3: Geometric Realization of a Simplicial Complex

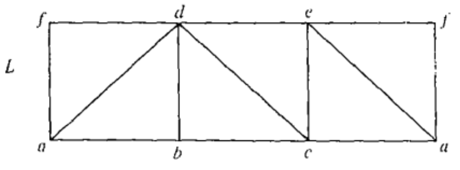


Figure 3.4: Abstract Simplicial Complex

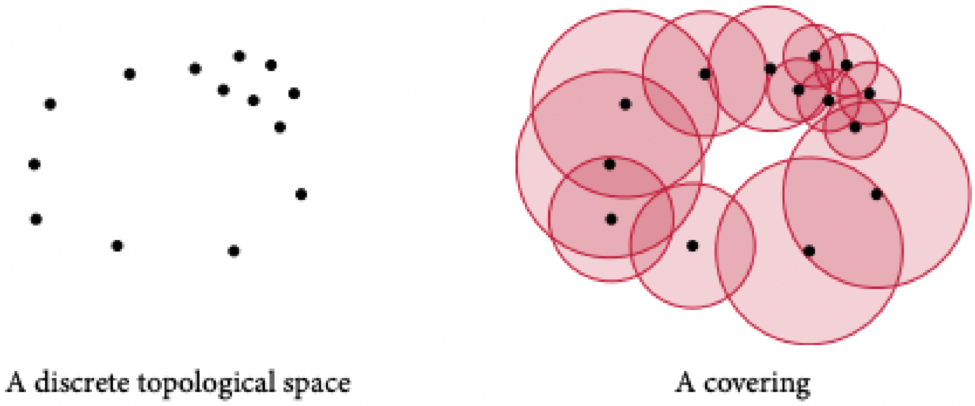
3.2.3 Nerves, Cover, and Complexes

Real-world data, such as a basketball dataset, doesn't come neatly packaged with this simplex structure. To analyze it with the tools of algebraic topology, we first need to shape our raw data into a form that fits this mathematical framework. To ensure our methods are sound and meaningful, we'll start with a broad idea called the nerve of a covering. This will lead us to two specific ways to model our data with simplicial complexes, letting us approximate and analyze the shape of the data in a rigorous way.

Definition 3.2.15 (Covering of a Topological Space)**.** Given a topological space X, an indexed family of sets $\mathcal{U} = \{U_i \mid i \in I\}$ is a cover of X if

$$X \subseteq \bigcup_{i \in I} U_i,$$

that is, the topological space is contained within the union of these sets.



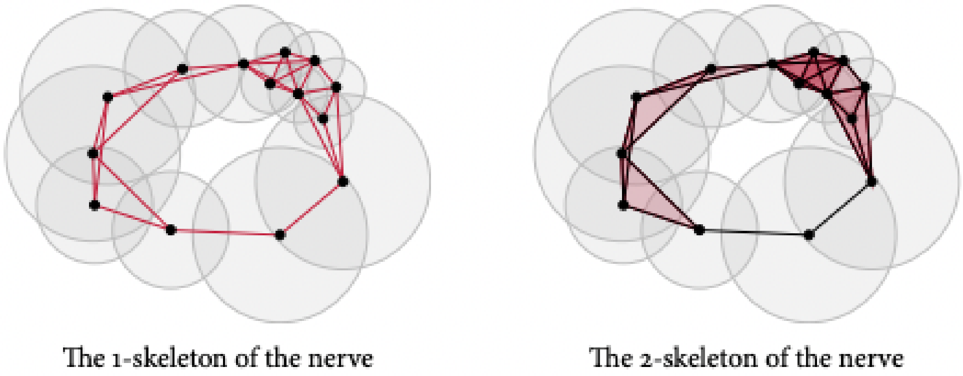
A discrete topological space

A covering

Figure 3.5: The figure illustrates the covering of a topological space. A set of the covering is represented as a disk.

Although coverings can consist of an infinite number of sets, in practice, we often assume the covering is finite for simplicity and computational feasibility. Without the finiteness assumption, certain topological properties may be more difficult to study, and computational methods, such as those used in analyzing point cloud data, may not be practical. We will later on see how to obtain coverings automatically for data from $\mathbb{R}^n$. In any case, the assumption that such coverings exist for the point cloud data means that the points are assumed to be sampled from a manifold or a topological space. If the coverings are obtained using a metric, then the points are assumed to be sampled from a metric space.

Definition 3.2.16 (Nerve of a Covering)**.** Consider a set $\mathcal{U}$ that covers a topological space with open sets. The *nerve* of $\mathcal{U}$ is defined as the collection of all non-empty subsets whose common intersection is non-empty. In



The 1-skeleton of the nerve

The 2-skeleton of the nerve

Figure 3.6: The nerve of a covering is discerned by inspecting overlaps among k-sized subsets. The analysis reveals a concentration of 2-simplices in the topological space's upper right region.

mathematical terms, this is expressed as:

$$\text{Nerve}\,\mathcal{U} := \{U_i \subseteq \mathcal{U} \mid \bigcap U_i \neq \varnothing\}$$

In line with this, any subset τ of a given set σ within the nerve also belongs to the nerve of $\mathcal{U}$. Hence, the structure of the nerve $\mathcal{U}$ fulfills the criteria for an abstract simplicial complex as specified in Definition 3.2.15.

Example 3.2.17. Given a covering of a topological space, as in Figure 3.5, we may obtain a simplicial complex from its nerve by checking all subsets for common intersections. The 1-skeleton is given by all pairs of cover sets that intersect, while the 2-skeleton contains all triples of cover sets that intersect, and so on (See Figure 3.6).

Definition 3.2.18 (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 3.2.19 (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.

Example 3.2.20. Let Σ be the complex consisting of a 2-simplex σ and its faces. The subdivision K of Bd σ indicated in Figure 3.7 can be extended to a subdivision Σ' of Σ by forming the cone $w * K$, where w is an interior point of σ; the subdivision Σ' is said to be obtained by "starring K from w." This method of subdividing complexes will prove very useful.

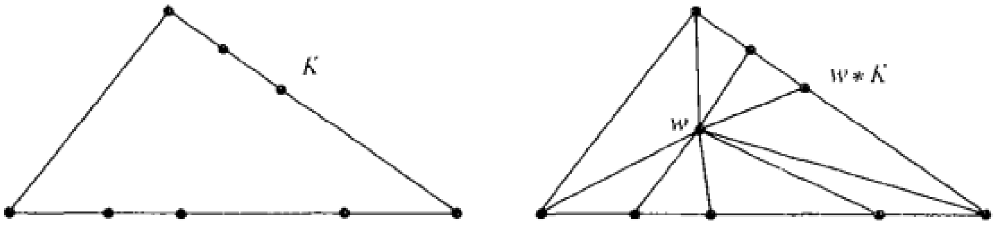


Figure 3.7: Subdividing the Complexes using the "Starring" Method

Definition 3.2.21 (Simplicial Filtration)**.** 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 3.2.22 (Čech Complex)**.** Let P be a finite set of points in $\mathbb{R}^n$, and $B(x, r)$ denote a ball 3.1.2 with center $x \in \mathbb{R}^n$ and radius $r \in \mathbb{R}$. The *Čech complex* of P and r is the nerve of this family of balls:

$$\mathcal{C}(r) := \{\sigma \subseteq P \mid \bigcap_{x \in \sigma} B(x, r) \neq \varnothing\}$$

The aforementioned definition is a particular case of the one presented in Definition 3.2.16, describing the nerve of a covering. Essentially, the Čech complex serves as the Euclidean counterpart to this nerve, specifically for coverings created by spheres in Euclidean space.

Definition 3.2.23 (Diameter)**.** Consider a finite topological space X with a given metric d. The *diameter* of X is defined as the greatest distance between any two points in X, mathematically expressed as:

$$\text{diam} X := \sup\{d(x, y) \mid x, y \in X\}$$

Given the finiteness of X, there will always be a maximum distance

that is calculable within X. This concept is extendable to its subsets as well.

Definition 3.2.24 (Vietoris–Rips Complex)**.** For a given distance threshold r and a set P of finite points, the Vietoris–Rips complex, denoted as $\mathcal{V}(r)$, consists of all point subsets within P where every pair of points is no more than r units apart:

$$\mathcal{V}(r) := \{\sigma \subseteq P \mid \text{diam } \sigma \leq r\}$$

A set of points $\sigma = \{v_0, \ldots, v_k\}$ forms a simplex in $\mathcal{V}(r)$ precisely when every two points in σ are at most r units away from each other. This method is widely used for constructing simplicial complexes out of discrete spatial data [28].

We will write $\mathcal{C}_\epsilon$ and $\mathcal{V}_\epsilon$ to refer to the Čech complex and the Vietoris–Rips complex of scale ϵ.

Remark 3.2.25. The Čech and Vietoris-Rips complexes provide a crucial framework within which we can study topological spaces via algebraic tools. In algebraic terms, the Čech and Vietoris-Rips complexes allow us to translate geometric information into algebraic data, which can then be analyzed and manipulated to reveal the intrinsic structure of a dataset. This algebraic data takes the form of groups and related constructs that can be studied systematically. The two complexes are also of importance for the next section as they lay the groundwork for discussing homology groups, which are a type of algebraic invariant that can be used to classify topological spaces.

3.3 Algebra

3.3.1 Groups and Related Concepts

Definition 3.3.1 (Group)**.** A *group* $(G, +)$ is a set G with a binary operation $+$ such that:

[28]

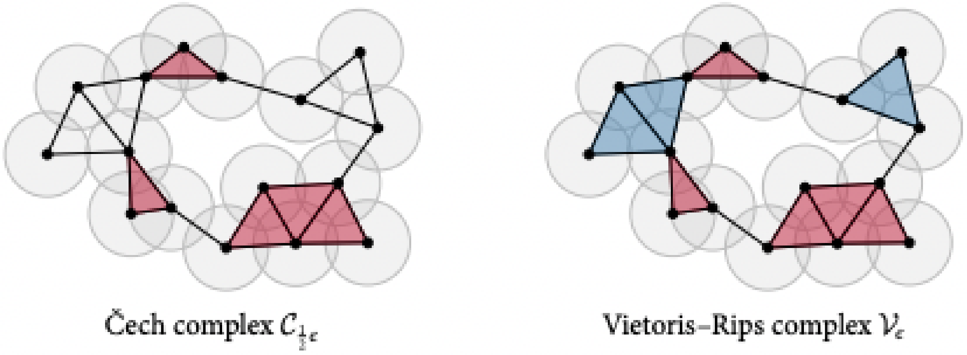

Čech complex $\mathcal{C}_{\frac{1}{2}\varepsilon}$
Vietoris–Rips complex $\mathcal{V}_{\varepsilon}$

Figure 3.8: Both complexes are shown for the same value of the scale parameter. The Čech complex contains a triangle for each subset of three balls with a non-empty intersection. By contrast, the Vietoris–Rips complex contains a triangle whenever three balls have a pairwise intersection.

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

Definition 3.3.3 (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 3.3.4 (Homomorphism and Isomorphism)**.** A function $f : G \to H$ between two groups is a(n):

- *Homomorphism* if $f(a + b) = f(a) + f(b)$ for any $a, b \in G$.
- *Isomorphism* if G and H are homomorphic and bijective; we say they are *isomorphic*, denoted $G \cong H$.

Definition 3.3.5 (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 3.3.6 (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.

Remark 3.3.7.

1. Typical cyclic groups are $\mathbb{Z}$ and the finite group $\mathbb{Z}_n$ already mentioned above.

2. The set of real numbers $\mathbb{R}$ under addition is an abelian group. The set $\mathbb{R}^\times$ consisting of the nonzero real numbers is a group under multiplication, with 1 as the identity.

3. The set of two by two matrices with nonzero determinant is a group under matrix multiplication.

4. The mapping $\phi : \mathbb{Z} \longrightarrow \mathbb{Z}_n$, defined by $\phi(k) =$ remainder when k is divided by n is a homomorphism of groups, whose kernel is the subgroup $n\mathbb{Z}$ consisting of all multiples of n. The quotient group $\mathbb{Z}/n\mathbb{Z}$ is a group isomorphic to $\mathbb{Z}_n$.

3.3.2 Rings

Definition 3.3.8 (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 $a, b \in R$. A *unity* in R is an element 1 such that $1 \cdot a = a \cdot 1 = a$ for any $a \in R$. It is easily shown that a unity is unique, if it exists.

Definition 3.3.9 (Field)**.** A *field* is a commutative ring $(R, +, \cdot)$ with unity 1 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 3.3.10 (R-Module Commutative Ring)**.** Given a commutative ring $(R, +, \cdot)$ with unity 1, 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 3.3.11 (Vector Space)**.** Let $(F, +, \cdot)$ be a field. An F-module V is called a vector space over F provided $+$ is vector addition and $\cdot$ induces scalar multiplication.
Note: In the context of vector spaces, the uniqueness condition typically refers to the unique representation of each vector as a linear combination of basis vectors, which is a consequence of linear independence of the basis.

3.4 Homology

3.4.1 Chain Complexes and Homology Groups

Definition 3.4.1 (Chain Group of a Simplicial Complex)**.** Let K be a simplicial complex. The p-th chain group, denoted by C_p, consists of all sums of p-simplices from K. Formally, elements of C_p look like $\sum_j c_j \sigma_j$ with each

σ_j being a simplex in K and c_j either 0 or 1, indicating whether a simplex is included in the sum.

Remark 3.4.2. The chain group allows us to define a group structure to the simplicial complex, which permits us to define boundaries and cycles in a formal mathematical sense.

Definition 3.4.3 (Simplicial Chain)**.** The elements of the p^{th} chain group C_p are referred to as *simplicial chains*. A simplicial chain is thus a linear combination of p-simplices. Since C_p is a group, we can perform the addition of two different simplicial chains. For two simplicial chains a and b, their addition over $\mathbb{Z}_2$ coefficients is equivalent to computing their *symmetrical difference*, given by:

$$a + b = (a \cup b) \setminus (a \cap b)$$

Definition 3.4.4 (Boundary Homomorphism)**.** Let K be a simplicial complex. The p-th boundary homomorphism ∂_p maps a simplex $\sigma = \{v_0, \ldots, v_p\}$ in K to a formal sum representing its boundary. This sum is formed by alternately excluding each vertex v_i from σ and combining the results with alternating signs:

$$\partial_p \sigma = \sum_i (-1)^i \{v_0, \ldots, \hat{v}_i, \ldots, v_p\}$$

Here, $\hat{v}_i$ means that the vertex v_i is omitted. The function ∂_p links the p-simplices to their boundaries, effectively relating p-chains with $(p-1)$-chains.

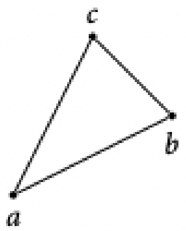


Figure 3.9: Calculating the boundaries of a 2-simplex (a triangle) and its 1-simplices (edges) [28].

Example 3.4.5. The boundary of the triangle in Figure 3.9 is non-zero. We have $\partial_2(a, b, c) = \{b, c\} + \{a, c\} + \{a, b\}$. The set of edges, on the other hand, does not have a boundary, i.e., $\partial_1(\{b, c\} + \{a, c\} + \{a, b\}) = \{c\} + \{b\} + \{c\} + \{a\} + \{b\} + \ a\} = 0$, because the simplices cancel each other out in $\mathbb{Z}_2$.

Definition 3.4.6 (Chain Complex of a Simplicial Complex)**.** The chain complex of an n-dimensional simplicial complex K is the sequence of chain groups, connected with the corresponding boundary homomorphisms:

$$0 \longrightarrow C_n \xrightarrow{\partial_{n+1}} C_{n-1} \xrightarrow{\partial_n} \cdots \xrightarrow{\partial_3} C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \longrightarrow 0$$

Definition 3.4.7 (Cycle and Boundary Groups)**.** Given a chain group C_p, the p-th cycle group is denoted as

$$Z_p := \ker \partial_p,$$

which encompasses all p-simplices and simplicial chains that lack a boundary, hence each element of Z_p is known as a p-cycle. Similarly, the p-th boundary group is given by

$$B_p := \mathrm{im}\partial_{p+1},$$

where it includes the boundaries of all $(p+1)$-dimensional simplices. This definition necessitates a dimensional shift because a simplex of higher dimension is essential for a chain to act as a boundary. Therefore, the simplicial chains that are in B_p are referred to as bounding cycles.

Definition 3.4.8 (Homology Group)**.** For a given chain group C_p with cycle subgroup Z_p and boundary subgroup B_p, the p-th homology group is given by the formula:

$$H_p = Z_p/B_p = \mathrm{ker}\partial_p/\mathrm{im}\partial_{p+1},$$

where the division symbol represents the formation of a quotient group. The homology group H_p consists of equivalence classes of cycles. If two elements, a and b in Z_p, differ by a boundary element c from B_p, they are treated as indistinct in H_p. This is based on the understanding that differences that are purely boundaries do not alter the intrinsic characteristics of the cycles.

The elements of H_p, known as homology classes, essentially capture the 'shape' of cycles ignoring the boundaries, offering a way to classify features of the space that persist beyond mere topological boundaries.

Definition 3.4.9 (Betti Numbers)**.** The p-th Betti number β_p is the algebraic rank of the p-th homology group, i.e., $\beta_p = \mathrm{rank}H_p$.

Figure 3.10 illustrates the Betti geometric interpretation of Betti numbers. They are used to count the number of independent k-dimensional holes in a topological space. Here, β_0 denotes the number of connected components, β_1

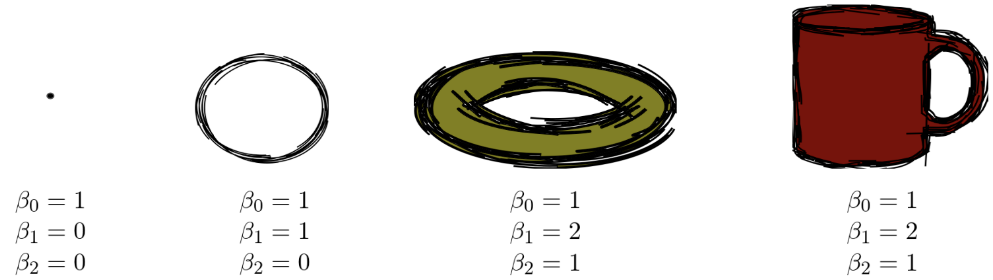


Figure 3.10: Betti numbers of common objects. A k-th dimensional betti number corresponds to the number of k-dimensional holes in a topological space. Here, β_0 denotes the number of connected components, β_1 denotes the number of one-dimensional or "circular" holes, and β_2 denotes the number of two-dimensional "voids" or "cavities" [28].

denotes the number of one-dimensional or "circular" holes, and β_2 denotes the number of two-dimensional "voids" or "cavities".

- The first object is a single point, representing a space with one connected component and no holes or voids, hence $\beta_0 = 1$, $\beta_1 = 0$, and $\beta_2 = 0$.

- The second object is a single loop, representing a space with one connected component and a one-dimensional hole, thus $\beta_0 = 1$, $\beta_1 = 1$, and $\beta_2 = 0$.

- The third object is a torus, which has one connected component, two one-dimensional holes (one in the center of the torus and one around it), and one void in the middle, leading to $\beta_0 = 1$, $\beta_1 = 2$, and $\beta_2 = 1$.

- The fourth object is a mug, which like the torus, has one connected component, two one-dimensional holes (one formed by the handle and the other around the mug's body), and one void inside the mug, giving $\beta_0 = 1$, $\beta_1 = 2$, and $\beta_2 = 1$.

Theorem 3.4.10 (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.*

Remark 3.4.11. This is the basis of the computations of the β_i via Gauss-Jordan elimination.

Remark 3.4.12. The converse of the previous theorem is not true. That is, if X and Y are from the same homology groups, we cannot conclude that X and Y are homeomorphic.

Theorem 3.4.13 (Rank-Nullity for Homology Groups)**.** *For a chain complex C, the pth homology group H_p satisfies the rank-nullity theorem:*

$$rank\ Z_p - rank\ B_p = rank\ H_p,$$

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

Theorem 3.4.14 (Stability of Betti Numbers)**.** *Betti numbers are stable under small perturbations in the Gromov-Hausdorff sense.*

Chapter 4

Data Preparation

4.1 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 comprised of 127 rows representing the complete rosters of 8 teams and 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).

Table 4.1: Label Explanations for individual UAAP Player Tables

Label	**Name**	**Explanation**
Team	Team	The team for which the player currently plays listed as a three to four-letter abbreviation (Ex: ADMU = Ateneo de Manila University)
Pos	Position	One of five standard player positions: 'PG' = Point Guard, 'SG' = Shooting Guard, 'SF' = Small Forward, 'PF' = Power Forward, 'C' = Center
PPG	Points Per Game	Average number of points scored per game
APG	Assists Per Game	Average number of assists made per game
RPG	Rebounds Per Game	Average number of rebounds collected per game
SPG	Steals Per Game	Average number of steals per game
BPG	Blocks Per Game	Average number of blocks per game
TOPG	Turnovers Per Game	Average number of turnovers per game
+/-	Plus-Minus	The point differential when the player is playing. For example, if a player's ± is +5, the team outscored opponents by 5 points while the player was on the court.
FG%	Field Goal Percentage	Percentage of attempted shots made
3PT%	Three-Point Percentage	Percentage of three-point shots made
FT%	Free Throw Percentage	Percentage of free throws made
MPG	Minutes Per Game	Average number of minutes played per game

4.2 Programming Language

Python was 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 and persistent homology 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.

4.2.1 Library

Giotto-TDA [33] 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.

4.2.2 Data Preprocessing

Before going into the analysis, a comprehensive preprocessing phase was undertaken to ensure the quality and suitability of the basketball data.

Column Names

To improve data coherence and for convenience, we renamed several columns.

- ’Name’: Renamed to ’Player’ for clarity.
- ’TOPG’: Renamed to ’TOV’ to reflect the standard abbreviation for turnovers.
- ’FPG’: Renamed to ’PF’ to represent fouls committed.
- ’MPG_x’: Renamed to ’MPG’ for brevity.
- ’3PT FGs %’: Renamed to ’3P%’ for compactness.
- ’2PT FGs %’: Renamed to ’FG%’ to signify overall field goal

Feature Decomposition

Columns representing attempts and made shots were separated for better analysis of shooting efficiency. This involved splitting:

- ’3PT FGs M/A’ into ’3PM’ (made three-point shots) and ’3PA’ (attempted three-point shots).
- ’2PT FGs M/A’ into ’2PM’ (made two-point shots) and ’2PA’ (attempted two-point shots).
- ’FTs M/A’ into ’FTM’ (made free throws) and ’FTA’ (attempted free throws).

Feature Creation

Using the decomposed features from the previous step, two additional columns were derived by summing ’3PM’ with ’2PM’ and ’3PA’ with ’2PA’, respectively.

- ’FGM’: Total field goals made
- ’FGA’: Total field goals attempted

In addition to that, We created two new columns to make the rebounding data uniform with the other columns by dividing the total defensive rebounds (DEF) and offensive rebounds (OFF) by the games played (GP).

- ’DRB’: Defensive rebounds per game.
- ’ORB’: Offensive rebounds per game.

Data Cleaning

Following the feature engineering and decomposition steps, any remaining unnecessary columns and duplicate rows were removed to streamline the data for analysis.

4.2.3 Principal Component Analysis

Principal component analysis (PCA) is a common dimentionality-reducing function, used to project a point cloud into a lower-dimensional space which makes some techniques feasible or more interpretable. One example is in the Mapper algorithm to be applied in the following chapters, where PCA allows the dataset to be divided into regions which are then clustered into nodes. However, the results of PCA itself often also reveal insights regarding the structure of the data and the relationships between its features.

From a set of data points $\mathbf{x}_i \in \mathbb{R}^f$, PCA obtains a sequence of vectors $\mathbf{w}_j \in \mathbb{R}^f$ that successively capture the variance, or shape, of the dataset, such that $\sum_{i=1}^{n} \mathbf{w}_i \cdot \mathbf{x}_i$ already approximates the shape of the data well at low n, and only performs better with increasing n.

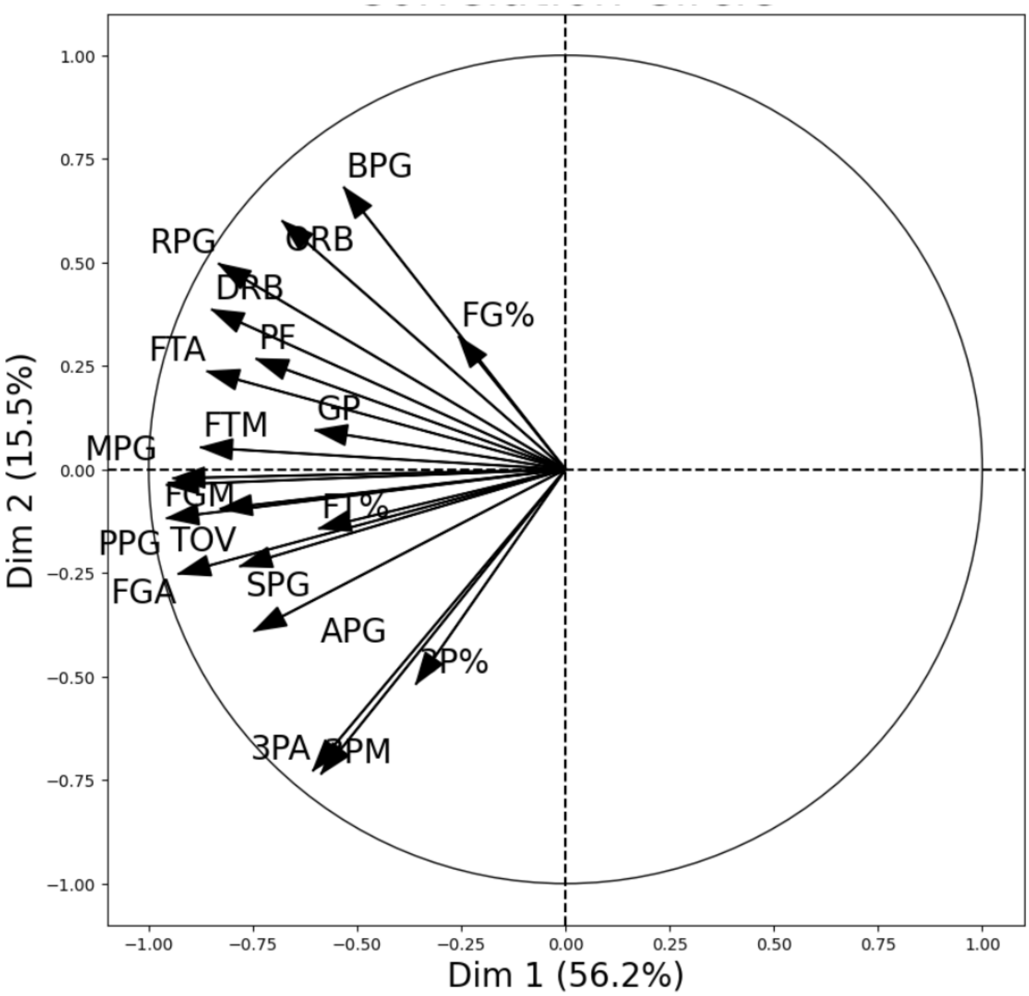


Figure 4.1: Circle of Correlations between dataset features and PCA dimensions.

With the 2-dimensional PCA, we discovered that 56.2% of the variance of the dataset is accounted for by the first dimension, and 15.5% by the second dimension. As seen in the circle of correlations figure, the first dimension correlates negatively with all features of the dataset, and the second dimension is the one which divides the features into two groups.

We interpret the first dimension as a measure of the usage rate of players, as evidenced by minutes per game (MPG) being the feature most correlated to the first dimension and having almost zero correlation to the second. A higher MPG is associated with an increase in all of a player's statistics because it relates to more time spent in-game to perform the game actions associated with those statistics, such as field goal attempts, points, rebounds and others.

Meanwhile, the second dimension positively correlates with features such as blocks per game (BPG) and rebounds per game (RPG), and negatively correlates with features such as three point shots attempted (3PA) and made (3PM). An interpretation for this dimension could be the stature of the player, separating the shorter players who usually play around the edges of the court and the taller players who usually play close to the rim.

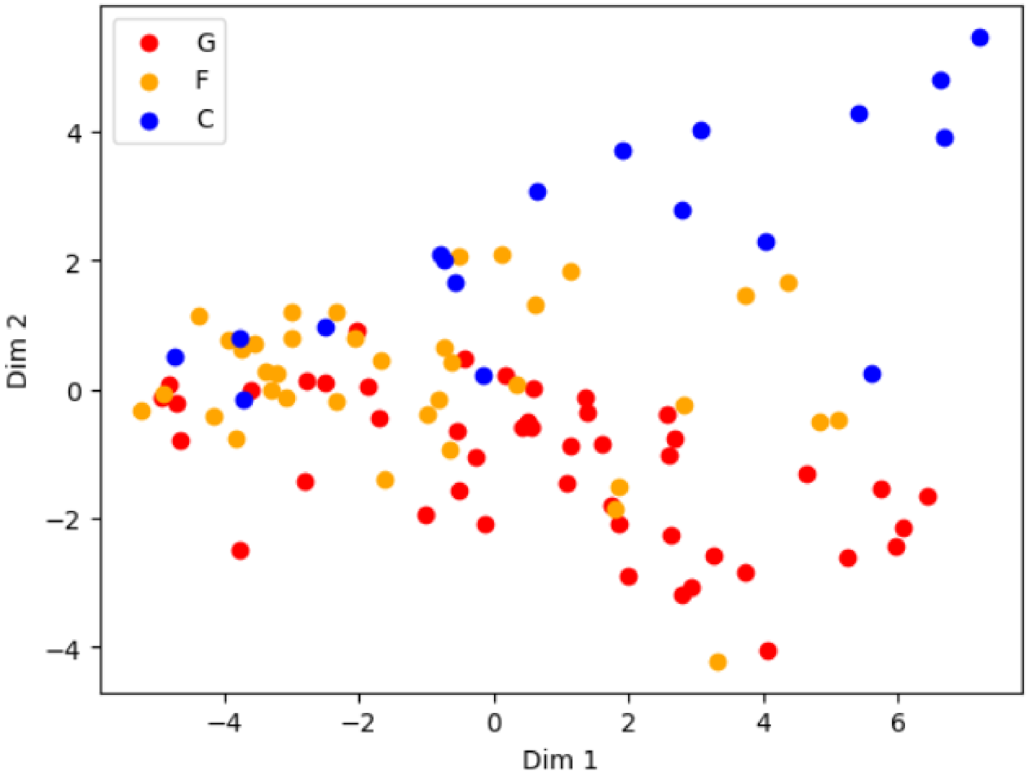


Figure 4.2: Plot of all players based on the values of 2-dimensional PCA.

The results above are supported by the projection of all players to the 2-dimensional PCA space, where guards (often shorter players) are negative in dimension 2, and centers (often taller players) are positive. In the negative portion of dimension 1, the players of low usage are mixed through all positions because of the low values in their statistics. It can also be seen that from this plot, it is still difficult to create clusters or separate different types of players with boundaries. This shows that PCA is not fully enough to characterize and classify the distribution of players in the league, motivating our usage of TDA techniques to further explore the dataset.

Chapter 5

Persistent Homology

5.1 Background on Persistent Homology

Persistent homology is a method of 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 as defined in Definition 3.2.7.

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*:

$$\varnothing = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K, \tag{5.1.0.1}$$

where each K_i is a subcomplex 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. For a monotonic function on a simplicial complex, the sublevel sets of function f are simplicial subcomplexes. Finally, we arrange them as an increasing sequence, yielding equation 5.1.0.1.

The nesting relation of simplicial complexes in a filtration induces an inclusion map, which in turn induces a homomorphism between the corresponding homology groups. We denote this homomorphism by

$$f_i^{j,p} : H_p(K_i) \to H_p(K_j),$$

where p is the dimension and we have $i \leq j$ to ensure a proper ordering. Hence, a filtration gives rise to a sequence of homology groups that are connected via the functions $f_i^{j,p}$, i.e.

$$0 = H_p(K_0) \xrightarrow{f_0^{1,p}} H_p(K_1) \xrightarrow{f_1^{2,p}} \dots \xrightarrow{f_{n-1}^{n,p}} H_p(K_{n-1}) \xrightarrow{f_{n-1}^{n,p}} H_p(K_n) = H_p(K),$$

where p again denotes the dimension of the homology groups. As we apply the filtration, or as we sift through the data step by step, the structure of the homology groups shifts. We might see some features within the groups disappear, while new features may emerge. The systematic approach of filtration allows us to sort these features by the point in the process at which they show up or fade away. With reference to the earlier explanations in Definition 3.4.7 and Definition 3.4.8, we can contextualize these definitions in terms of persistence homology.

Definition 5.1.1 (Persistent Homology Group)**.** Given two indices $i \leq j$, the p^{th} persistent homology group $H_p^{i,j}$ is defined as

$$H_p^{i,j} = Z_p(K_i)/(B_p(K_j) \cap Z_p(K_i)),$$

which means $H_p^{i,j}$ includes all homology classes in K_i that remain when we move to K_j. Using the notation previously introduced, we can write $H_p^{i,i} = H_p(K_i)$.

Definition 5.1.2 (Persistence)**.** Consider a homology class c that appears within the complex K_i and vanishes in K_j. Let the scales at which c appears and vanishes be denoted by y_i and y_j, respectively. The *persistence* of the class c, denoted by pers c, is the interval length

$$\text{pers } c := y_j - y_i$$

over which c exists. For persistence, we use the real numbers $\mathbb{R}$ with the assumption that the scales are real values. If a homology class c persists throughout the entire process without vanishing, its persistence is defined as infinite, labeling c as an *essential homology class.* The persistence measure can thus assume any value in the extended real number line $\mathbb{R}_\infty = \mathbb{R} \cup \{\infty\}$, accommodating both finite and infinite lifetimes of features.

Remark 5.1.3. A persistence interval defined as $[y_i, y_j]$ is the interval at which a particular homology group is alive.

Remark 5.1.4. The persistence of a homology class serves as a relevance criterion. For instance, a low persistence value indicates a small-scale feature.

Definition 5.1.5 (Persistent Betti Number)**.** Given two indices $i \leq j$, the p-th persistent Betti number is defined as

$$\beta_p^{i,j} := \text{rank} H_p^{i,j},$$

And so, persistent homology is all about tracking the topological invariants through the sequence of homology groups as defined in Definition 5.1. The next section will discuss ways to visualize these.

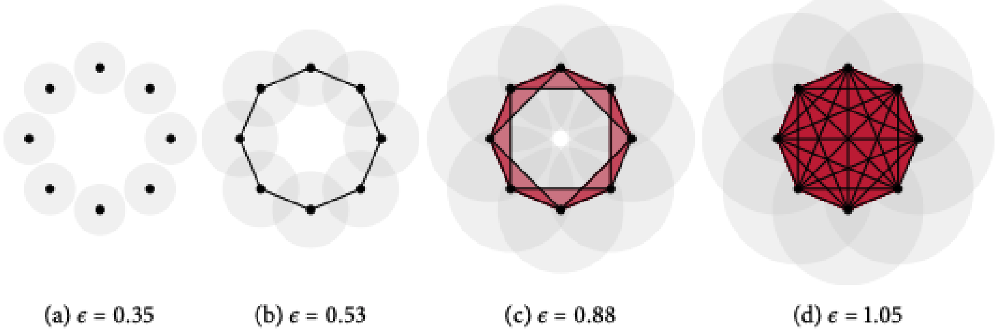


Figure 5.1: Analyzing the first Betti number β_1, we observe that $\beta_1 = 1$ for complexes (b) and (c). For $\varepsilon \leq 0.50$, the value of β_1 is zero, as illustrated by (a). Conversely, when $\varepsilon \geq 1.0$, we also have $\beta_1 = 0$ because the circular gap is filled and, as demonstrated in (d), all potential simplices have been included. The fundamental concept behind persistent homology is to acknowledge that within the interval $\varepsilon \in [0.50, 1.0)$, the Betti number β_1 remains equal to one. This signifies the endurance of the hole across these scales.

This section was based on chapter 4 of Persistent Homology in Multivariate Data Visualization [28].

5.2 Visualizing Persistent Homology

5.2.1 Persistence Diagrams

A persistence diagram acts as a graphical representation of persistence intervals. For each interval of the same dimension, represented by $[c, d]$, a corresponding point (c, d) is plotted in the Euclidean plane $\mathbb{R}^2$. If $d = \infty$, the point is plotted just above the top edge of the diagram to indicate unbounded persistence. With increasing filtration values, these points accumulate above the diagonal line, filling up that region in the diagram. For example, Figure 5.2 illustrates such a diagram for the 1-dimensional persistent homology of an artificially created torus.

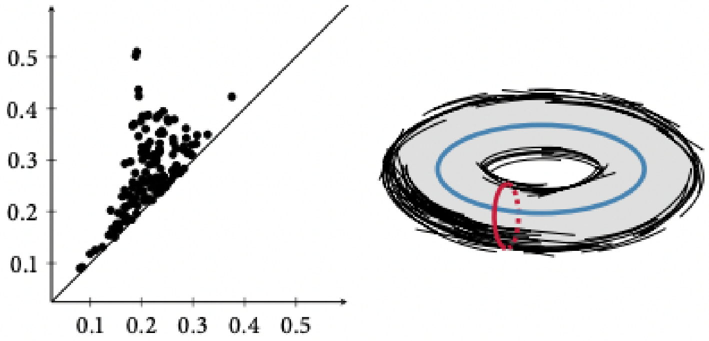


Figure 5.2: In the persistence diagram representing the 1-dimensional persistent homology for a synthetic torus, there's a group of points near the diagonal that emerge due to the process of sampling. Apart from these, there are two distinct points, born at $\epsilon = 0.2$ and dies at $\epsilon = 0.5$ that represent the torus' two circular features, illustrated on the right side [28].

While the persistence diagram does an excellent job providing a compact summary of the topological features, it suffers from some drawbacks. Firstly, persistence diagrams tend to appear cluttered and suffer from overplotting. Secondly, analyzing persistence diagrams requires having to estimate the distances of point to the diagonal. Lastly, persistence diagrams are not the most intuitive visualization for showing persistence intervals.

In the next section, we introduce persistence barcodes, which improves displaying the scale information of each persistence interval.

5.2.2 Persistence Barcodes

Persistence barcodes contain each interval $[c, d]$ as an interval in the plane. The individual intervals are then stacked on top of each other. This stacking of the intervals make it so that it becomes easy to measure and compare the lengths of lines. Thus, the viewer is immediately drawn towards the largest scale information. However, this may be misleading if topological features

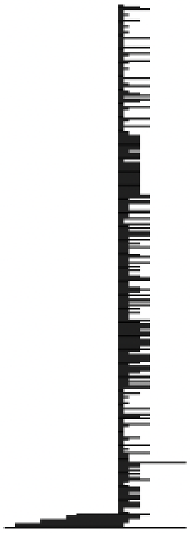


Figure 5.3: The persistence barcode of the H_1 betti numbers of a synthetic torus.

exist at different scales whose relative differences are large.

Another drawback of the barcode is its scalability issues. Even with datasets that have just a few hundred points, the barcode can become excessively large. For instance, Figure 5.3 shows a straightforward barcode of the 1-dimensional persistent homology from a synthetic torus dataset. The two lengthy bars signify the two generators in the first dimension. All other bars result from the sampling method and, although they provide insights into the construction of the torus—indicating it was sampled in circular slices—they aren't as significant as the two actual generators.

5.3 Results and Analysis

We adopt the persistent homology methodology outlined by Gispets [11]. He first limited the dataset to only contain points per game, rebounds per game, assists per game, steals per game, blocks per game, and turnovers per game. Then, using those features, he created persistence diagrams using Vietoris-Rips complexes for the complete player dataset, as well as for individual player categories: guards, forwards, and 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.

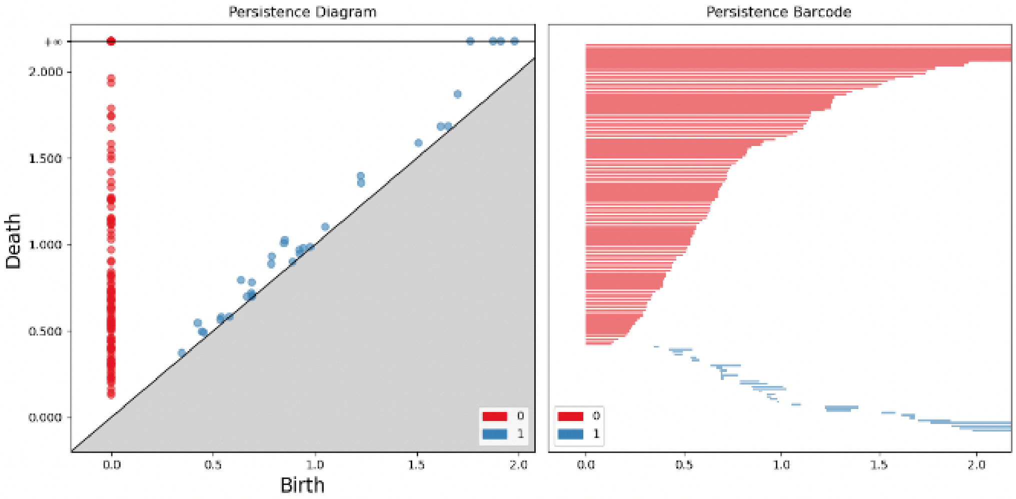


Figure 5.4: Persistence Diagram and Barcode for the Whole Dataset: The persistence diagram illustrates the birth and death of the H_0 and H_1 Betti numbers, represented as red and blue dots, at varying ϵ. As we can see, there are four β_1 that persists. This indicates that there are holes in the dataset.

The results for the entire dataset are shown in Figure 5.4. 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 all are born at the same time, but die at varying ϵ. 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. This hints that there are outliers in the UAAP season 85 dataset.

Now, for the H_1 components, It's noteworthy that many 1-dimensional homological features are distanced from the diagonal, which indicates the presence of significant holes in the dataset. Moreover, three of them are even persist all throughout. 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. A subgroup would imply that there are clusters in the dataset.

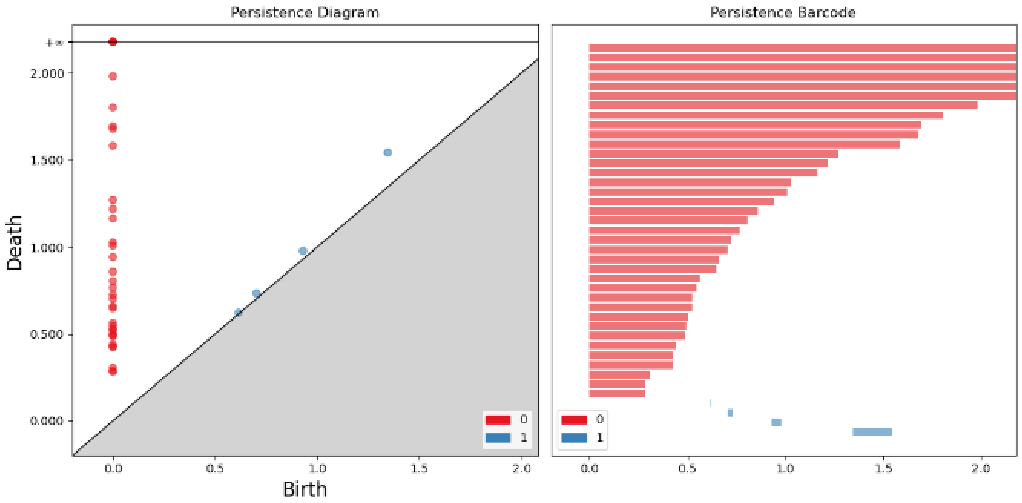


Figure 5.5: Persistence Diagram and Barcode for the Guards: All of the H_0 live near the diagonal, which indicate that there are no holes when the dataset is limited to the guards.

In examining the persistence diagram of the guards, shown in Figure 5.5, we observe that the majority of H_1 values are transient, suggesting a lack of significant topological features, such as holes, within this group of players.

Contrastingly, the persistence diagram for the forwards, presented in Figure 5.6, similarly indicates fleeting H_1 values, implying an immediate closure of any gaps that may arise.

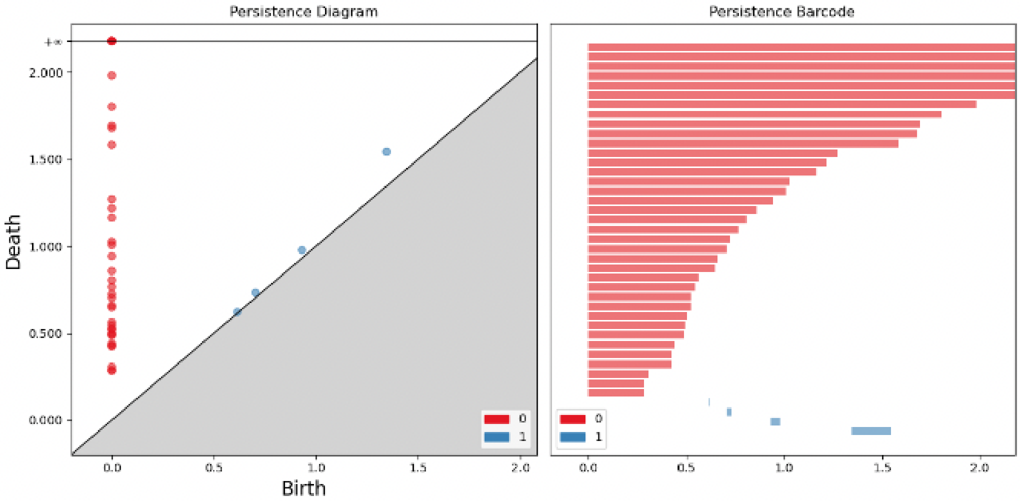


Figure 5.6: Persistence Diagram and Barcode for the Forwards: All of the H_1 live near the diagonal, which indicate that there are no holes when the dataset is limited to the forwards.

Most intriguing is the persistence diagram for the centers, depicted in Figure 5.7. Here, a notable persistence of a H_1 value can be seen. Since there are holes in the persistence diagram among the centers, this could indicate that there are subgroups within the centers in the UAAP that will stay separated from the main bunch of points in the point cloud.

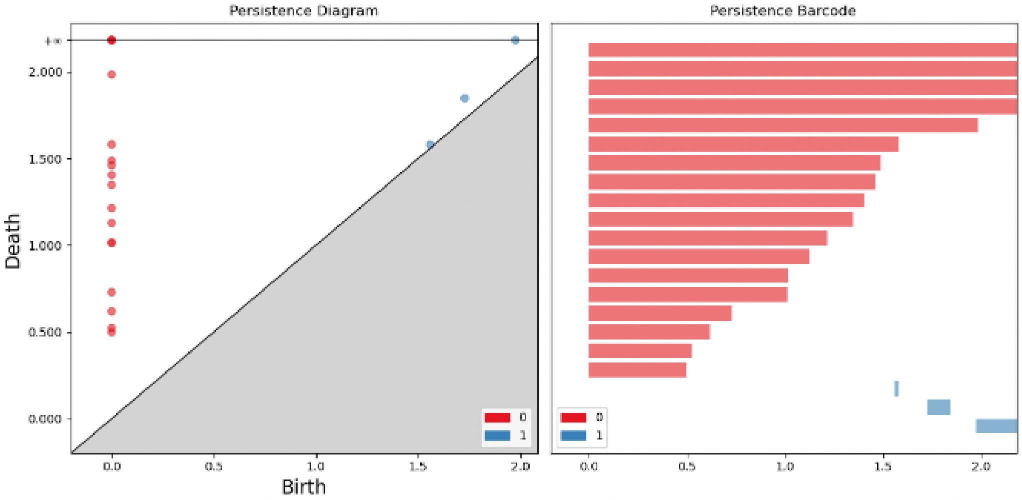


Figure 5.7: Persistence Diagram and Barcode for the Centers: One of the β_1 persists, which indicate that there is a hole among the centers.

Chapter 6

Mapper

6.1 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. [31]

6.1.1 Topological Background

Throughout the papers which have been utilizing it until today, the Mapper algorithm has been relatively unchanged in its implementation and theoretical framework based in topology,and specifically, simplicial complexes. [27] In this paper, we will be using the same background [31] 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* 3.2.16 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.

Definition 6.1.1 (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.

Definition 6.1.2 (Barycentric Coordinates)**.** 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 $i = 0, 1, ..., k$ 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.

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 $\alpha \in T(x)$.

Theorem 6.1.3 (Barycentric Coordinatization)**.** *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 $N(U)$.

Definition 6.1.4 (Decomposition of a Covering)**.** 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 $f^{-1}(U_\alpha)$ of 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.

Definition 6.1.5 (Map of Simplicial Complexes)**.** 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. For a family of coverings $U_i, i = 0, 1, \dots, n$ 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:

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

Theorem 6.1.6. *Given a space X with a map $f : X \to Z$ (parameter space Z) and and the induced map of 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 $f^{-1}(V)$, again by continuity of f. This means that $U_\alpha(i)$ is mapped to the unique $V_{f(\beta)}(j)$ such that $U_\alpha(i) \subseteq V_{f(\beta)}(j)$.

6.1.2 Pipeline

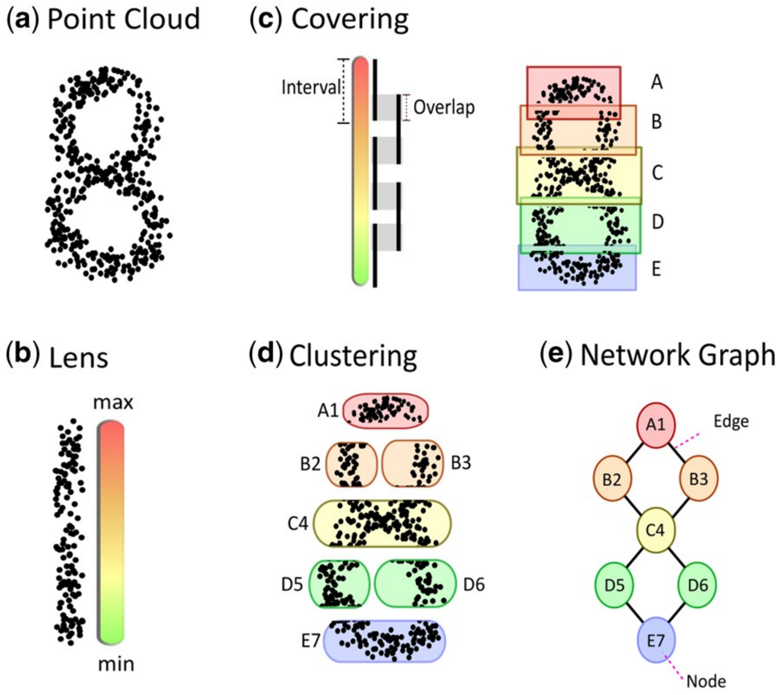


Figure 6.1: Visual Representation of the Mapper Pipeline: (a) A point cloud representing data, (b) A filter (lens) applied to assign a real-valued function to each data point, (c) The covering of the filter range with overlapping intervals, (d) Clustering of data points within each interval, and (e) The final network graph where nodes represent clusters and edges represent the non-empty intersections between clusters in the covering.[23]

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 or filter 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.

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 $\mathbb{R}^k$ such that all points in $f(X)$, the projection of X, lie in $\cup_{i=1}^{n} U_i$. 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.

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.

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 several parameters as suggested in [24]. 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).[23]

6.1.3 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 function, 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. A `MapperInteractivePlotter` object was created using the data pipeline mentioned previously, and an 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.

6.2 Results and Analysis

6.2.1 Hyperparameter Analysis

To investigate the general effect of varying hyperparameters on the Mapper graph, we measure its graph density. Proposed by Lawler [19] in 1976, the density or average degree is the average number of edges connected to each vertex or node, and is computed as follows:

$$\text{Density} = \frac{2|E|}{|V|},$$

where $|E|$ is the total number of edges and $|V|$ is the total number of vertices or nodes in the graph. The range of this measure is $[0, \infty)$.

These two following hyperparameters apply specifically to the cube-based cover function.

Overlap Fractions

The `overlap_frac` value plays a crucial role in the final shape of the Mapper graph as it determines the overlap between cover sets. To understand the impact of this parameter, we experimented by testing the range of overlap values from 0 to 1 with a step of 0.05. This allowed us to identify the ideal overlap that would best capture the shape of the league.

In our initial investigation, we tested the extremes of the overlap parameters. This allowed us to observe its influence on the clustering outcomes. We discovered that high overlap values (greater than 0.65) resulted in excessive intersections. This led to the formation of large nodes within the clusters, rendering the clustering ineffective as even players with no similarities are grouped which hinders the identification of distinct groups.

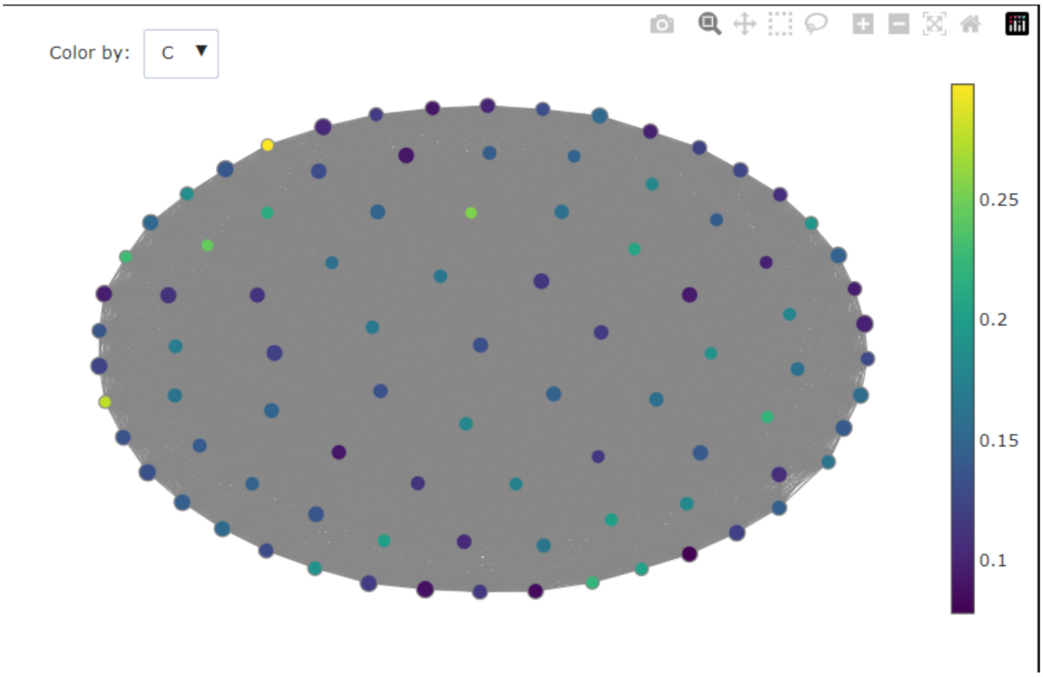


Figure 6.2: Mapper Plot with overlap_frac = 0.95

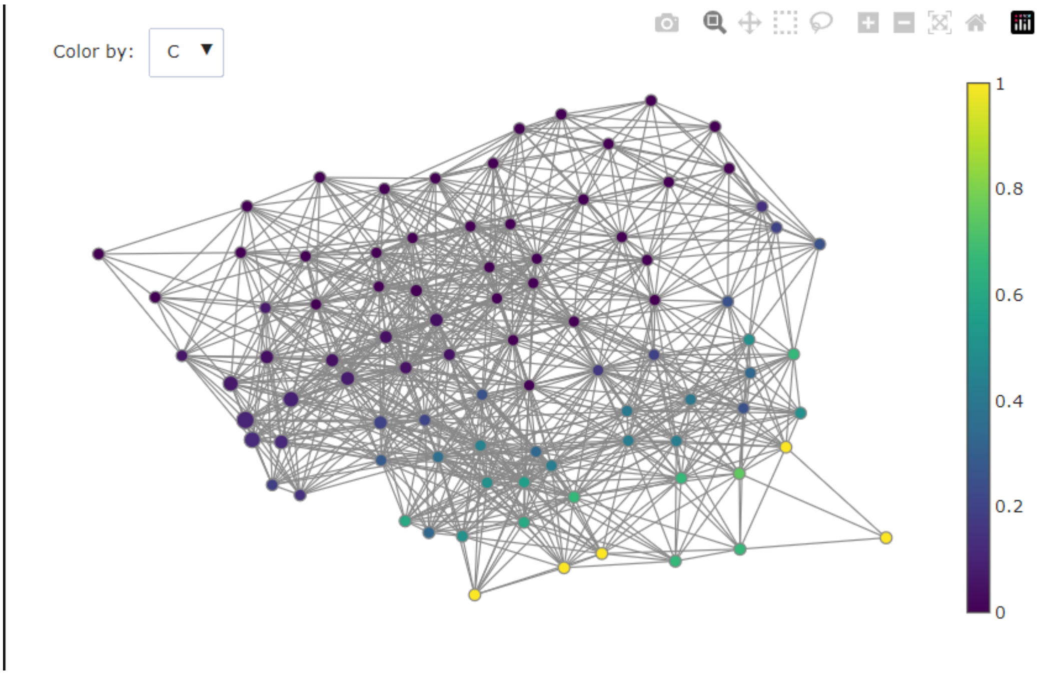


Figure 6.3: Mapper Plot with overlap_frac = 0.7

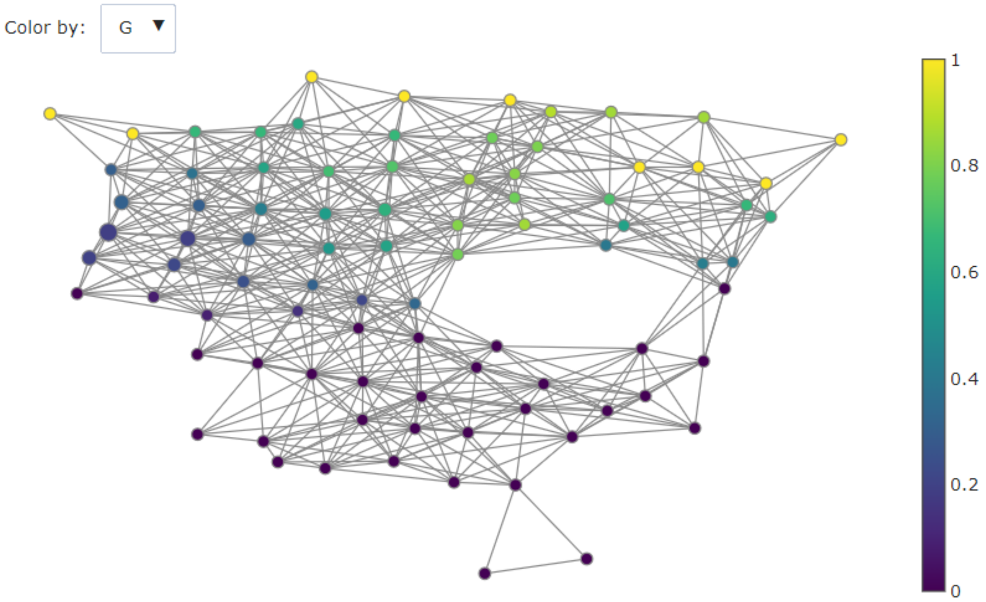


Figure 6.4: Mapper Plot with overlap_frac = 0.65

Conversely, using low overlap values (less than 0.3) produced minimal connections. This resulted in nodes containing a small number of players. Such limited connections make it challenging to establish well-defined clusters which hinders the clustering process.

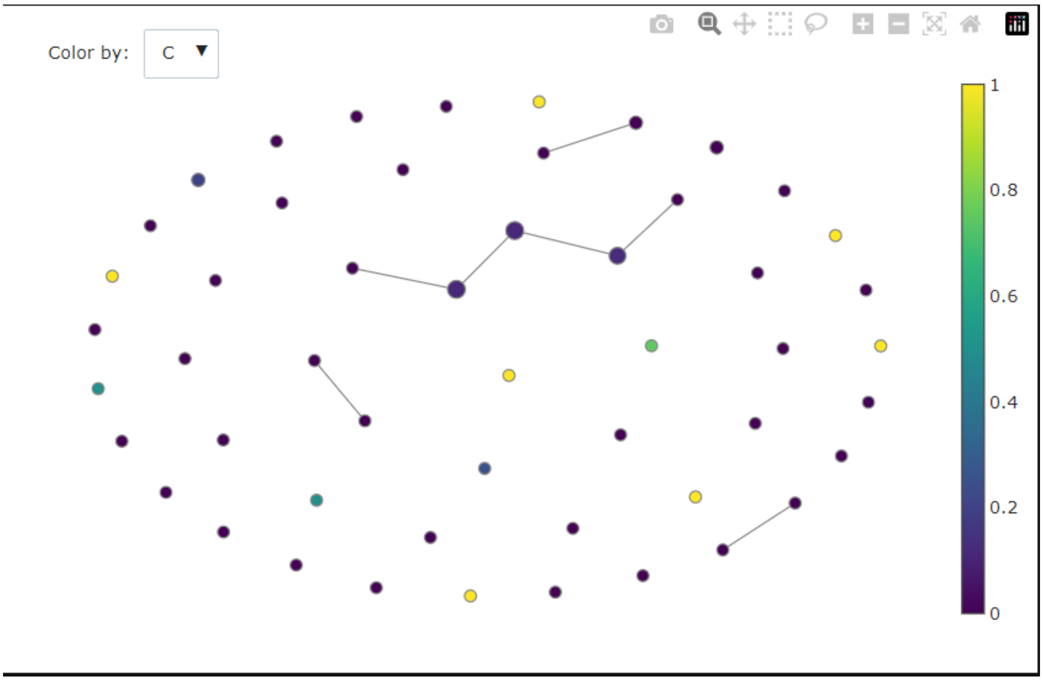


Figure 6.5: Mapper Plot with overlap_frac = 0.05

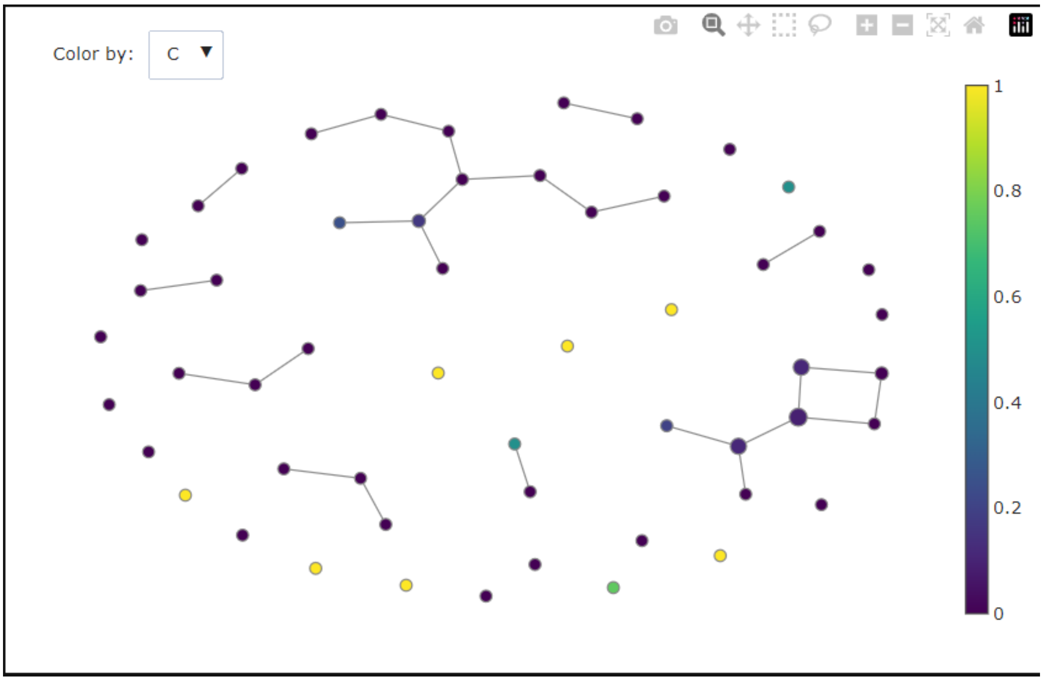


Figure 6.6: Mapper Plot with overlap_frac = 0.15

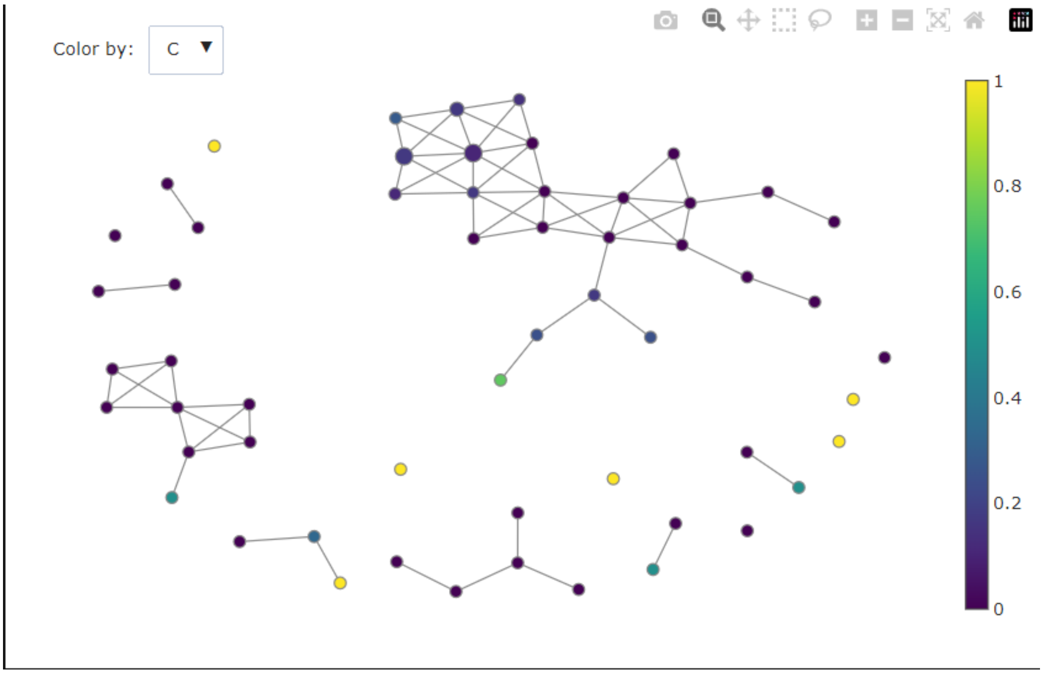


Figure 6.7: Mapper Plot with overlap_frac = 0.3

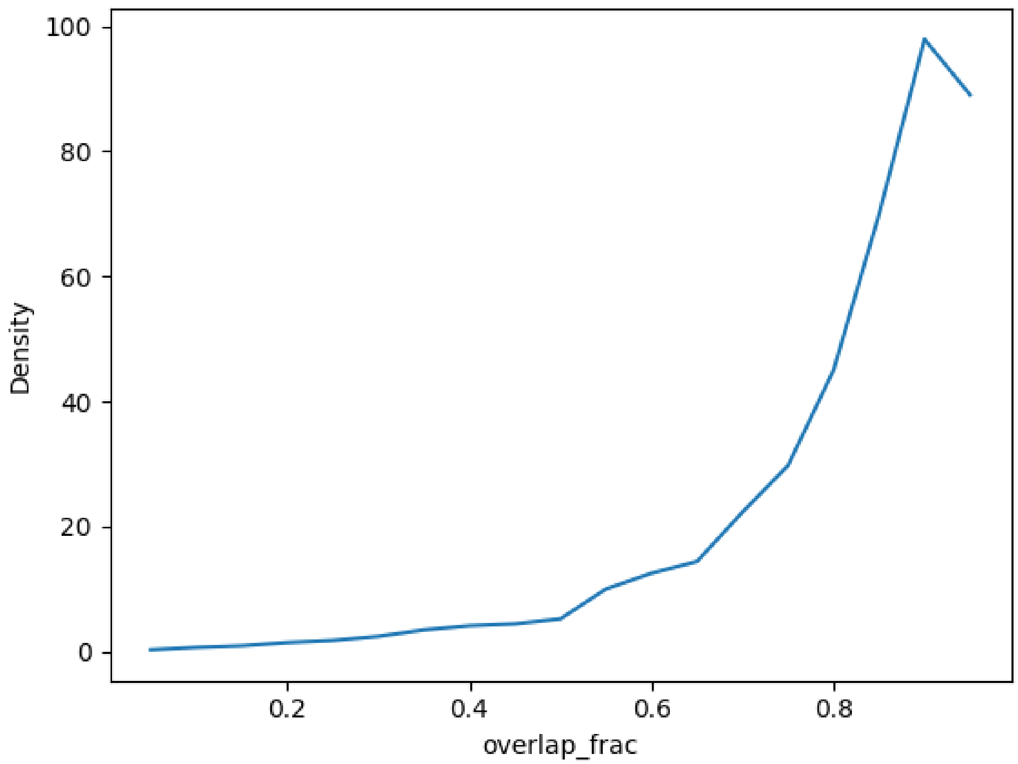


Figure 6.8: Density of Mapper graph at varying `overlap_frac`

Our subjective analysis is confirmed by our plot in Figure 6.8, which shows a distinct change in the growth rate at an overlap value of 0.5. Above 0.5, the density grows at a much faster rate than before. In general, the shape of the plot is that of exponential growth.

Number of Intervals

The `n_intervals` hyperparameter controls the number of divisions of the filter projection space created by the cover function. Since we are using a cubical cover, the number of cover sets generated is the square of this hyperparameter.

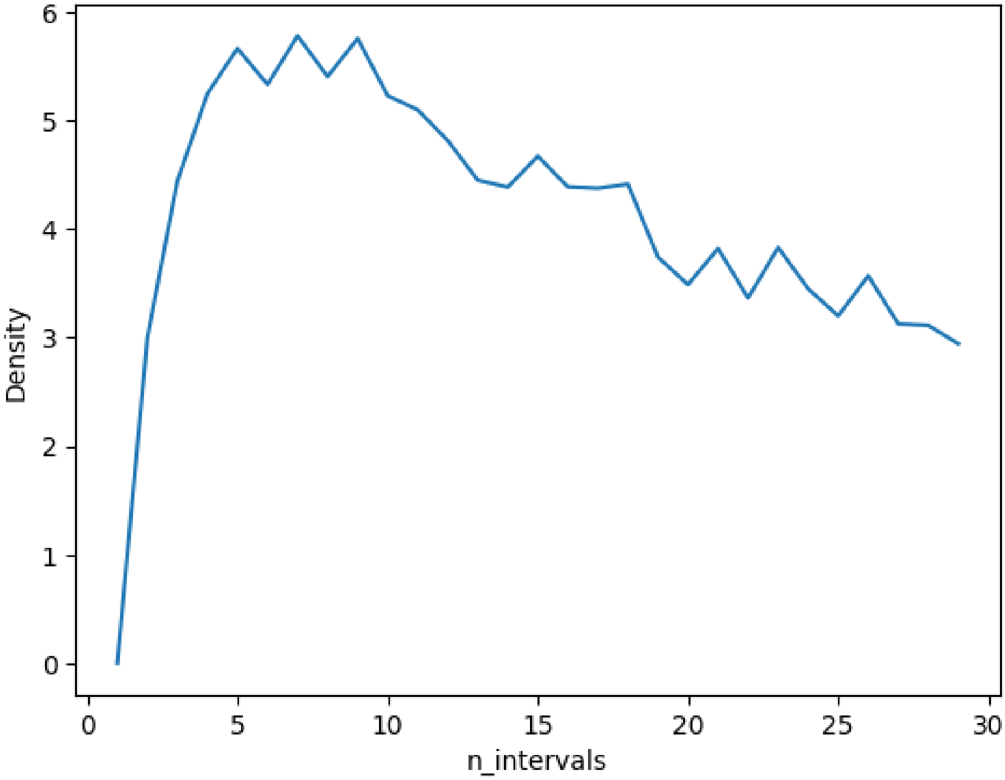


Figure 6.9: Density of Mapper graph at varying `n_intervals`

We see that adding more intervals after 10 begins to decrease the density measure, showing that the returns of information (edges and vertices) per added interval begins to diminish at this point.

Based on this assessment, for our analysis of topological structure, we choose a `n_intervals` of 10 and a `overlap_frac` of 0.5.

However, we also notice that at higher counts of intervals, the Mapper graph begins to split from one monolithic graph to an increasing number of isolated graphs. While topological structures such as holes are less prominent, this set-up is more convenient for identifying subgroups of players.

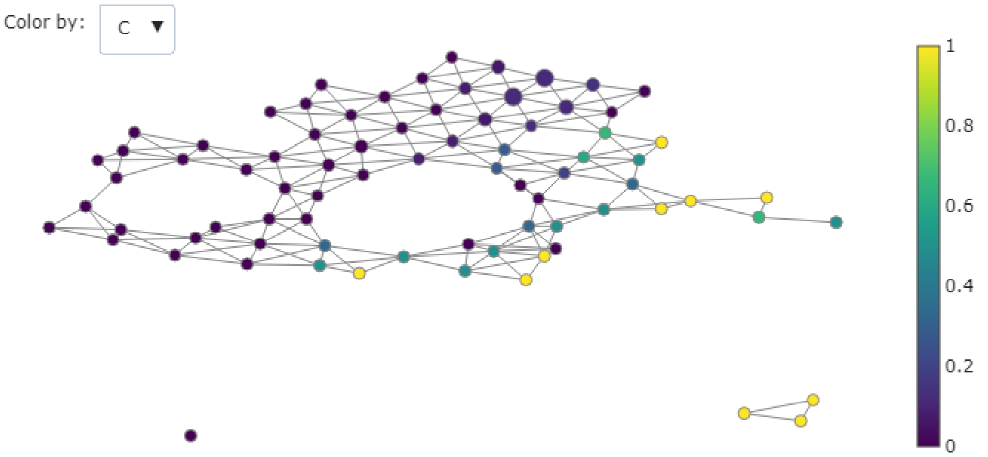


Figure 6.10: Mapper plot with 10 intervals shows clearly-defined hole structures.

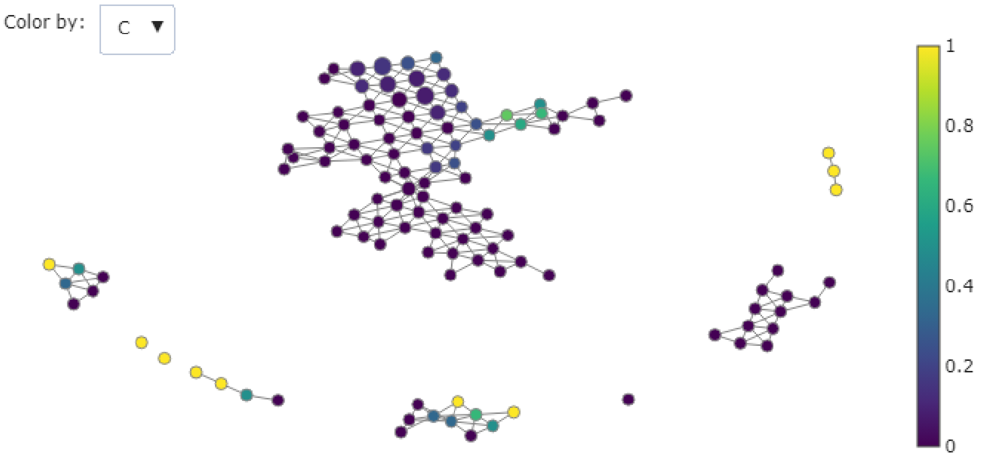


Figure 6.11: Mapper plot with 15 intervals shows more complicated isolated graphs.

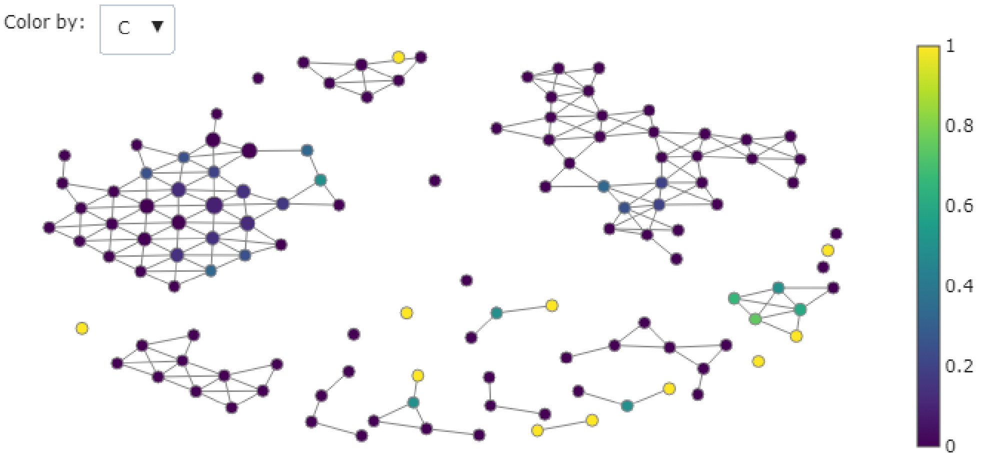


Figure 6.12: Mapper plot with 20 intervals no longer shows a prominent monolith graph, instead being composed of many smaller isolated graphs of varying sizes.

6.2.2 Identifying Subgroups

For this section, we go into the application of the Mapper Algorithm to uncover the subgroups within the UAAP Men's Basketball League (Season 85) players. As seen in the previous sections, we explored various parameters to create different visualizations that capture the inherent groupings within the data. We determined that using `n_intervals = 15` and `overlap_frac=0.5` created the most informative shape that captures the player subgroups.

Through the Mapper algorithm with these optimized parameters, we were able to generate a visual representation of the player data, highlighting the inherent groupings or subgroups present within the dataset. For uncategorized players, we relied on statistics and domain knowledge to see where the ungrouped player fits unless they were statistical outliers themselves. The subsequent sections will provide detailed definitions and characteristics associated with each of these identified subgroups.

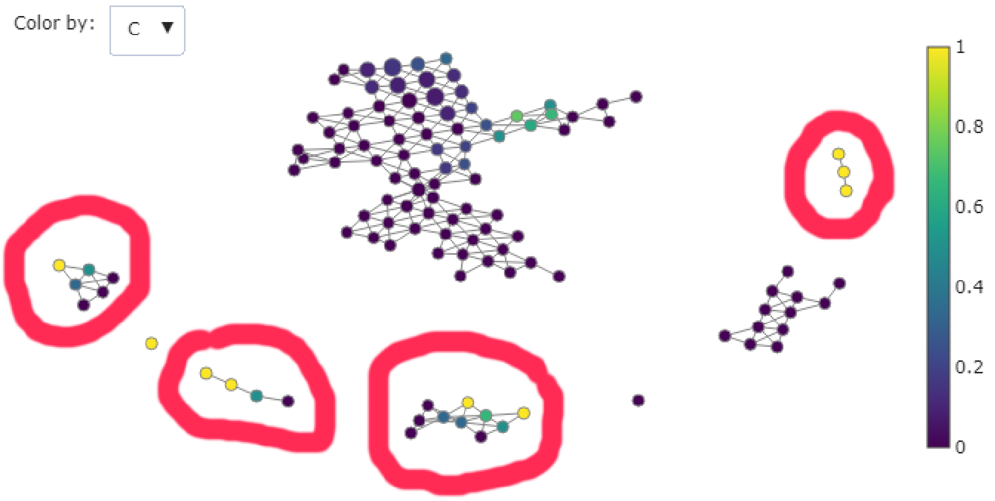


Figure 6.13: Mapper plot with 15 intervals, colored by proportion of centers (0% centers is blue, 100% centers is yellow). The subgroups of interest, comprised of forwards and centers are highlighted; meanwhile, the remaining subgroup on the lower right is composed of guards.

Paint Dominators

This subgroup represents big men who dominate on both ends of the court, with a focus on scoring and defense as this cluster comprises standout foreign-student athletes such as Ange Kouame, Malick Diouf, and Faye Adama.

They dominate on offense as they average more points (PPG) than the league average with efficiency(FG%, FGM/FGA). This can be attributed to their physical advantage which allows them to overpower opponents near the basket. A surprising statistic is their three-point percentage (3P%) which is above average compared to the league, despite fewer attempts (3PA), adding another dimension to their offensive skill and allowing them to have decent assist numbers (APG). A common tactic in basketball to deter opposing centers is to foul them as it is a common stereotype that big men are below-average free-throw shooters. But this subgroup seems to have developed measures against that tactic as they excel at drawing fouls and converting free throws (FTM/FTA, FT%) rendering it useless.

Their impact on the defensive end is equally impressive. They boast superior rebounding (ORB/DRB/RPG) with 281.16% greater rebounding numbers than the league average. Once again, their physical stature and athleticism allow them to secure rebounds more consistently. They further solidify their defensive presence with elite shot-blocking ability (BPG), exceeding the league average by a staggering 700%. This skill set significantly disrupts the opponent's scoring at the rim.

Despite their strengths, there are areas for improvement. Their turnover rate (TOV) is higher than the league average, suggesting a need for better ball security as centers do not usually handle the ball. Additionally, they commit fouls at a higher rate than average (PF), fitting their stature and aggressiveness on court. While their aggressive style contributes to their dominance, minimizing these fouls would further optimize their on-court impact. Interestingly, they play fewer games (GP) compared to the average player. This could be due to load management strategies implemented to prevent injuries or foul trouble for these key players.

Stretch Forwards

This subgroup comprises players who challenge the conventional perception of big players in the context of basketball in the Philippines due to their abilities on offense. Classified as a forward or a center, Kevin Quiambao, Karl Tamayo, and Zav Lucero exhibit a unique blend of offensive versatility and playmaking ability that was unprecedented for players their size in the past.

Their extended playing time (MPG) surpasses the league average, solidifying their roles as crucial starters or key contributors for their respective teams. Offensively, they demonstrate above-average scoring prowess (PPG, FGM/FGA), complemented by a field goal percentage that marginally exceeds the norm. However, the most intriguing aspect of their skillset lies in their unexpected marksmanship (3PM/3PA, 3P%). Despite their size and traditional big-man designation, they possess surprisingly good three-point shooters. Both their attempts and success rate from beyond the arc are significantly higher than the league average, transforming them into a potent threat from long range. Furthermore, their ability to draw fouls and convert free throws at a well-above-average clip (FTM/FTA, FT%) suggests a mastery of utilizing their size to gain an advantage and penetrate the defense.

Their rebounding numbers (ORB/DRB/RPG) could use some work as it fall below expectations for lengthy players who can take advantage of their sizes to get rebounds. However, they effectively compensate for this with their above-average assist ability (APG), showcasing a surprising capacity to create scoring opportunities for teammates. This additional dimension adds depth and dynamism to their offensive repertoire. Defensively, they disrupt the opponent's offense with a high steal rate (SPG) and contribute with a decent shot-blocking presence (BPG).

This subgroup transcends the limitations of traditional big men in the context of the Philippines where they are expected to play with their backs to the basket. Their offensive versatility is highlighted by their strong three-point shooting and surprising playmaking ability, skills that are expected from a guard. While their rebounding falls short of expectations for their size, they contribute significantly with steals. To fully maximize their skillset, reducing turnovers remains a key area for improvement. Nevertheless, their durability allows them to be impactful players with extended court presence throughout the season. This style of play has proven remarkably successful, with two players within this subgroup achieving the prestigious honor of Most Valuable Player and all three of them being previous Mythical Team awardees.

Defensive Rebounders

Another subgroup of the bigs are the defensive rebounders. While they do not dominate the low post like the Paint Dominators, they make it up with their relentless rebounding and disruptive defense.

Averaging around 6.9 points per game (PPG), this subgroup is only a bit above the league average by a few points. Their high field goal percentage (FG%) highlights efficient scoring near the basket which is supported by their field goal attempts and makes (FGM, FGA). On the other hand, their three-point attempts and percentages (3PA, 3P%) are below average, but their ability to draw fouls and convert free throws at a slightly above-average clip (FTM/FTA, FT%) adds another dimension to their offensive repertoire. Despite not dominating on the offensive end, the Defensive Rebounders still excel at utilizing their size and positioning to score effectively near the basket for easy points.

This subset has above-average rebounding statistics in all three categories (ORB, DRB, and RPG). Their most significant contribution is their dominance in rebounding, which gives their teams additional possessions. Moreover, their superior shot-blocking percentage (BPG) greatly outperforms the league average. In addition to their average steals per game (SPG), they make it difficult for the opposition to score at the rim and make them adjust their offensive game plans.

But this aggressive style of play comes at a cost as their turnover rate (TOV) is higher than the league average. In addition to that, they also commit fouls at a higher rate than average. While this aggressive approach can be beneficial in certain situations, minimizing these fouls would further optimize their on-court impact and prevent them from spending time on the bench due to foul trouble. This can be a point of improvement as they participate in a bit more games compared to the average.

Scrappers

This subgroup can be characterized by their tenacity on the court if we look at the statistics that they excel at, particularly at rebounding and disrupting the offense of the opposing team.

While size helps in increasing these statistics, it is effort that is more important. As a result, compared to the league average, their playing time (MPG) is significantly higher, suggesting a pivotal role as starters or key contributors (GP) for their respective teams. Despite averaging fewer points per game (PPG) than the league average, their remarkable field goal percentage (FG%) highlights efficient scoring within the paint area. This is further supported by their solid shot attempts (FGM, FGA) and success rate. Their reliance on interior scoring is undeniable, evident in their limited three-point attempts (3PA) and below-average success rate from beyond the arc (3PM, 3P%). However, they excel at drawing fouls and converting free throws at a significantly higher clip than the average player (FTM/FTA, FT%). This proficiency in getting to the free throw line suggests a strong ability to utilize their size and aggressiveness to create scoring opportunities, even if those opportunities don't always come from three-point range.

Beyond paint scoring, this subgroup dominates rebounding (ORB/DRB/RPG), arguably their most impactful contribution, generating valuable second chances. While their playmaking is decent (APG), their defensive presence shines with a high steal rate (SPG) and elite shot-blocking (BPG), disrupting the opponent's offense. However, turnovers (TOV) and a high foul rate (PF) require improvement. Despite these areas, their extended playing time (GP) highlights their durability and value throughout the season.

6.2.3 Analysis of Topological Structures

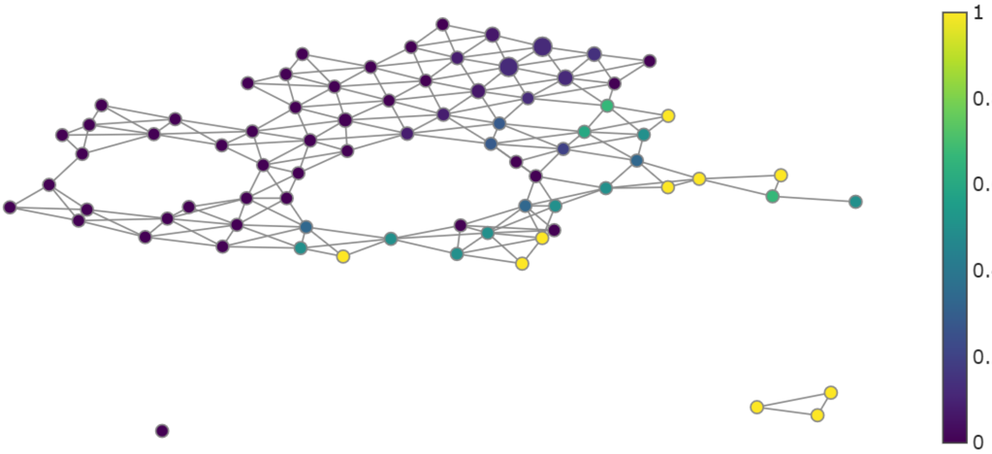


Figure 6.14: Mapper graph of the UAAP dataset, colored by the proportion of centers in each node. The darker each node is, the more concentrated it is with players that are labeled centers.

As shown in the figure, the form of the graph generated by the Mapper algorithm for the UAAP dataset using 10 intervals is of a monolithic graph with two hole structures, and two isolated sets of outlier nodes.

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 6.1: Outlier vs. League Average (Games Played ≤ 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 [15], [34]. 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 [35]. 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 6.2: 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.

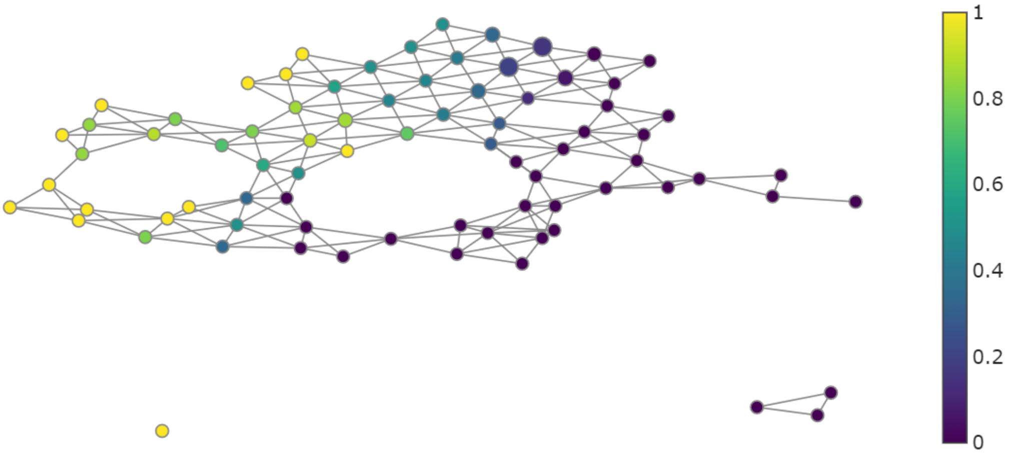


Figure 6.15: Mapper graph of the UAAP dataset, colored by proportion of guards in each node. Hole 1, on the left, is mainly surrounded by nodes representing guards.

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 6.3: 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 [16], 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.

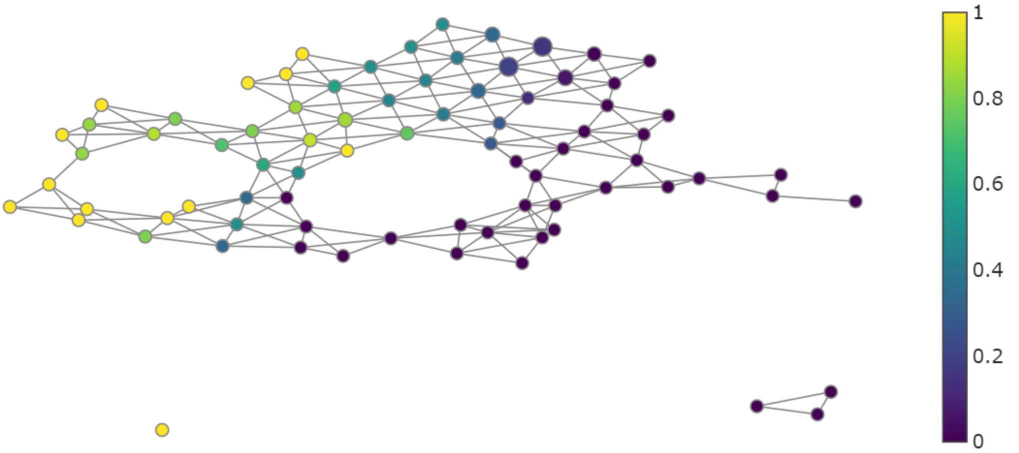


Figure 6.16: Mapper graph of the UAAP dataset, colored by proportion of forwards in each node. Hole 2, on the right, is mainly surrounded by nodes representing a mix of guards and forwards.

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 6.4: 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 [14] 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 [29] 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.

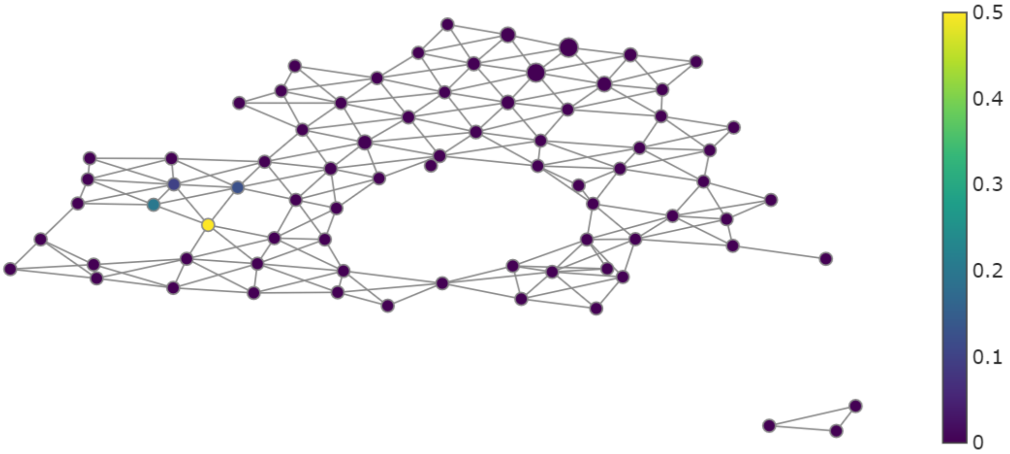


Figure 6.17: Mapper graph of the UAAP dataset, where the synthetic hole-filling player of hole 1 is highlighted in yellow.

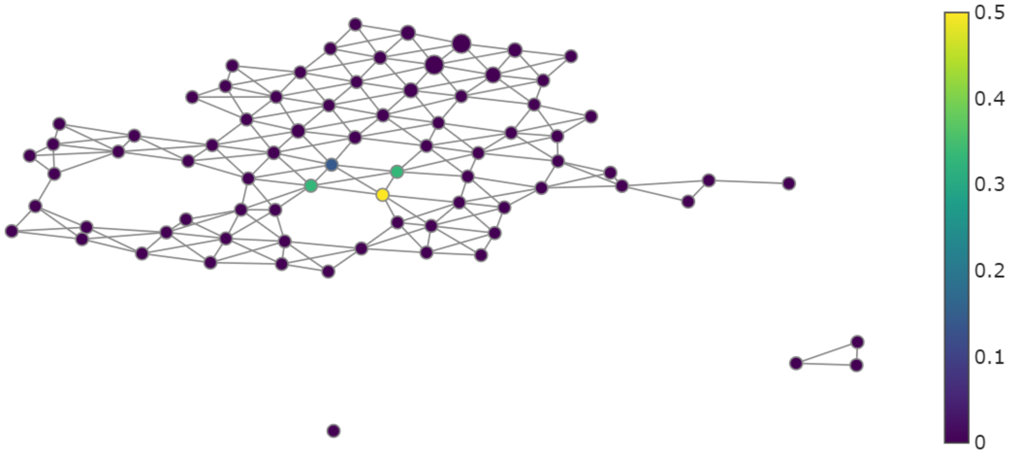


Figure 6.18: Mapper graph of the UAAP dataset, where the synthetic hole-filling player of hole 2 is highlighted in yellow.

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 7

Conclusions and Recommendations

7.1 Conclusions

We extracted insights from the UAAP dataset by first using persistent homology to identify the shape of the UAAP Season 85 dataset by computing homology groups across the simplicial filtration. Specifically, we were able to confirm the presence of persistent 1-dimensional holes for the whole dataset. Moreover, we are even able to find persistent 1-dimensional holes when we limit the players to the centers.

Then we analyzed outliers and holes in a graph generated by the Mapper algorithm. We can identify four subgroups of players as isolated networks of nodes in a Mapper graph, classifying players of the forward and center position into paint dominators, stretch forwards, defensive rebounders, and scrappers. The claim of the existence of an outlier subgroup among the centers based on their persistence diagram was also proved when the Mapper algorithm was applied to the dataset. A subgroup of centers that were statistically above the rest was discovered which is the paint dominators who are composed of high calibre players who contribute on both ends of the floor. Furthermore, using a Mapper graph with different parameters optimizing for topological structure, 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 this 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.

Table 7.1: Summary of Conclusions based on different Mapper graphs

n_intervals	**overlap_frac**	**Features**	**Conclusions**
10	Varying		Overlap fraction increases density exponentially
Varying	0.5		Density peaks at 5 to 10 intervals, - higher interval counts show more isolated graphs and less hole structures
15	0.5	7 outliers	4 player subgroups discovered - Paint dominators, stretch forwards, defensive rebounders, scrappers
10	0.5	2 holes, 2 outliers	2 missing player archetypes discovered - ”3-and-D”, foul drawers

7.2 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.

7.3 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 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 toward 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.

7.3.1 Enhancing Data Collection and Quality

1. 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.

2. 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.

7.3.2 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.

7.3.3 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.

7.3.4 Analysis of Other Seasons

- Tracking the evolution or the existence of specific subgroups in different seasons can give valuable insights into how the league changes
- Adding a time component and observing similar players across different seasons can create a detailed player report on how improvements or regressions occur.

7.3.5 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] Khaled Almgren, Minkyu Kim, and Jeongkyu Lee. "Extracting knowledge from the geometric shape of social network data using topological data analysis". In: *Entropy* 19.7 (2017), p. 360.

[3] Pablo G Camara et al. "Topological data analysis generates high-resolution, genome-wide maps of human recombination". In: *Cell systems* 3.1 (2016), pp. 83–94.

[4] 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.

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

[6] Pratik Doshi and Wlodek Zadrozny. "Movie genre detection using topological data analysis". In: *Statistical Language and Speech Processing: 6th International Conference, SLSP 2018, Mons, Belgium, October 15–16, 2018, Proceedings 6*. Springer. 2018, pp. 117–128.

[7] Ludovic Duponchel. "Exploring hyperspectral imaging data sets with topological data analysis". In: *Analytica chimica acta* 1000 (2018), pp. 123–131.

[8] Cameron T Ellis et al. “Feasibility of topological data analysis for event-related fMRI”. In: *Network Neuroscience* 3.3 (2019), pp. 695–706.

[9] Kathryn Garside et al. “Topological data analysis of high resolution diabetic retinopathy images”. In: *PloS one* 14.5 (2019), e0217413.

[10] Caleb Geniesse et al. “Generating dynamical neuroimaging spatiotemporal representations (DyNeuSR) using topological data analysis”. In: *Network neuroscience* 3.3 (2019), pp. 763–778.

[11] 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.

[12] Wei Guo and Ashis G Banerjee. “Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs”. In: *Journal of Manufacturing Systems* 43 (2017), pp. 225–234.

[13] 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`.

[14] 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`.

[15] 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/.

[16] 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/.

[17] 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.

[18] 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.

[19] E.L. Lawler. *Combinatorial Optimization: Networks and Matroids.* Holt, Rinehart and Winston, 1976. ISBN: 9780030848667. URL: https://books.google.com.ph/books?id=YvlQAAAAMAAJ.

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

[21] Max Z Li, Megan S Ryerson, and Hamsa Balakrishnan. "Topological data analysis for aviation applications". In: *Transportation Research Part E: Logistics and Transportation Review* 128 (2019), pp. 149–174.

[22] Derek Lo and Briton Park. "Modeling the spread of the Zika virus using topological data analysis". In: *PloS one* 13.2 (2018), e0192120.

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

[24] 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`.

[25] Jeff Murugan and Duncan Robertson. "An introduction to topological data analysis for physicists: From LGM to FRBs". In: *arXiv preprint arXiv:1904.11044* (2019).

[26] 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.

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

[28] Bastian Alexander Rieck. "Persistent Homology in Multivariate Data Visualization". Inaugural-Dissertation zur Erlangung der Doktorwürde der Naturwissenschaftlich-Mathematischen Gesamtfakultät. PhD thesis. Heidelberg: Ruprecht-Karls-Universität Heidelberg, 2017.

[29] 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.

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

[31] 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.

[32] Ann E Sizemore et al. "The importance of the whole: topological data analysis for the network neuroscientist". In: *Network Neuroscience* 3.3 (2019), pp. 656–673.

[33] Guillaume Tauzin et al. *giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration.* 2020. arXiv: `2004.02551 [cs.LG]`.

[34] 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/`.

[35] 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`.

[36] Chad M Topaz, Lori Ziegelmeier, and Tom Halverson. “Topological data analysis of biological aggregation models”. In: *PloS one* 10.5 (2015), e0126383.

[37] 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`.

[38] 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.