Visualizing NBA Player Archetypes: CSE6242 - Team 135 Final Report

Matthew Chang, Benjamin Lewis, Richard Moss, Minkang Suk
Georgia Institute of Technology
{mkchang,blewis79,rmoss41,paulsuk}@gatech.edu

Introduction

There are 5 traditional basketball positions, but there have been numerous approaches to cluster NBA players by the roles they serve on the court. These approaches use small sets of data, relying on traditional ”box score” statistics which limit their analysis to certain aspects of the game.

Traditional position designations may no longer adequately capture the diverse skill sets and playing styles found in the modern game. Consequently, there is a growing interest in developing more sophisticated methods for clustering and classifying NBA players based on their on-court performance.

This paper aims to build upon the work of Muniz (2022) by introducing a refined clustering algorithm that takes advantage of larger and more diverse datasets, incorporating both traditional box score statistics and modern player tracking data. Our approach seeks to address some of the limitations identified in Muniz’s original methodology, including the need for a more balanced treatment of the various dimensions within the datasets.

To achieve these objectives, we first conduct a comprehensive literature review of existing clustering methods, examining their strengths and weaknesses in relation to the specific challenges presented by basketball data. We then present our proposed algorithm, which combines the strengths of several popular clustering techniques, such as hierarchical, with innovative approaches tailored to basketball-specific metrics.

To evaluate the effectiveness of our proposed algorithm, we apply it to a large dataset of NBA player performance data spanning multiple seasons, comparing the resulting clusters to those obtained through Muniz’s approach and other existing methods. We also assess the stability of our clusters across different seasons.

Our findings suggest that the refined clustering algorithm provides a more accurate and comprehensive representation of player roles in the NBA, offering valuable insights for coaches, general managers, and analysts seeking to understand and optimize team performance. By advancing the state of the art in player clustering and classification, this research contributes to the ongoing evolution of basketball analytics and the growing body of knowledge surrounding the game.

Stakeholders who would benefit from a deeper understanding of NBA player archetypes include general managers and scouts building cohesive rosters, players needing to market themselves by the value they bring, and sports betting and fantasy sports companies building models to predict and analyze players. This importance is highlighted by Berger and Daumann who discuss the importance of understanding player data to make draft decisions, as decisions based only on surface-level scouting often fail to evaluate player athleticism [16].

The main impact of this model is that by making player archetypes widely available, casual sports fans and those individuals hoping to innovate in the basketball industry will have access to both data typically sealed behind paywalls and proprietary data owned by teams. The benefits are also highlighted by Levine who used player classification to determine what makes a hall of fame player [8].

In the professional world, this impact can be measured via a sensitivity analysis whereby the performance of teams using the traditional player roles are compared to those using new archetypes to create rosters. In addition to a sensitivity analysis, we can use other quantitative measures to assess the quality of our clustering.

Problem Definition

We aim to categorize NBA players by their roles and skill set and develop an interactive tool to visualize player role comparisons.

Literature Review

Clustering Approaches

At the 2012 MIT Sloan Sports Analytics Conference, Muthu Alagappan used topological data analysis to cluster players based on their statistics [2], which was a seminal moment for redefining what modern NBA positions may look like. Many clustering approaches have since been used to group basketball players by their statistics. K-Means was used by Zhang et al [9] to identify different archetypes for guards (traditionally just defined as point and shooting guards), as well as by Patel [12] who additionally applied dimensionality reduction before performing the clustering. Kalman and Bosch use model-based clustering [13] to assign probabilities of cluster membership instead of doing hard assignment. Dehesa et al [5] uses a two-step clustering method, utilizing log-likelihood distance to classify players. While these prior models can inform and validate our approaches to player clustering, these suffer from use of smaller sets of data, often relying on traditional ”box score” statistics, which limit their analysis to certain components of the game, such as only modeling offense [13], or only categorizing guard players [9].

In 2022, Muniz proposed a novel clustering algorithm [11] which divides the dataset into multiple categories based on different aspects of basketball (such as scoring, passing, defence, hustle), and then uses PCA and k-means clustering to assign players to clusters for each skill. Using these clusters, players are then assigned similarity scores to finalize the player clusters. This novel approach is incredibly relevant to our project; some areas of improvement may be improving the similarity score metric, or balancing the dimensionality of the datasets.

NBA Datasets

In 1991, Dean Oliver developed a new scoring method for basketball game that identifies valuable statistics for teams and players that go beyond ”value approximation” methods. The new basketball scoring method generates new statistics, such as pace of the team, offensive/defensive rating, and effective FG%. [4]. Additionally, spatially referenced game data has been made available with the convergence of computing. [6] This data allows NBA analysts to look further than the simple traditional metrics and create advanced metrics. All of these metrics are potentially important features in differentiating NBA player archetypes.

Player Analysis

Shea uses polar coordinates to classify players into a “periodic table of elements” with 2-3 players occupying a typical cluster. [14] Both the selection of variables and clusters is arbitrary in this approach, relying upon trial and error. We plan to use PCA to select variables as well as K-means to generate clusters.

NBA Metric Visualization

Goldsberry’s visualisations stand out as easily interpretable to a layperson. [6] As they are static images, we seek to make them interactive. Additionally, Shea and Baker create visualizations for a single player and then combine them to view an entire team at a glance [15]. Ensuring that our individual data points have context will contribute to a more rich visualization overall.

Proposed method

We built off of Muniz’s novel CD-kMeans clustering approach [10] and supplemented this approach in 3 main ways:

Dataset Augmentation

We performed augmentation of the dataset used in the clustering algorithm with additional data sources since the existing clutch dataset only contained five features. In the initial exploratory data analysis to expand the clutch dataset, we found that was was limited. Due to the competitive nature of the NBA, most of the latest shot tracking data is proprietary and not shared publicly. However, we were able to obtain an additional dataset that supplements a gap in the original paper’s clutch dataset. The clutch metric is defined as a game within 5 points with 5 minutes left. There is a leverage metric for end of game situations that is more sensitive than the above rough definition. This added another distinguishing factor for player clustering.

The high leverage dataset wass pulled by leveraging the PBP Stats API [1]. Using the API endpoint, we pulled every player from 2014 to 2023 for clutch situations, which amounts to around half a million new data points. In this API, a clutch situation is defined as NBA player performance statistics in high and very high leverage situations as outlined in [3], which is very similar to the existing clutch dataset definition. We pulled the data from the API endpoint into a pandas dataframe and filtered out the features with a null rate greater than 30 percent. Since PCA is completed in the algorithm state, we decided to leave features with sufficient data and let the algorithm perform the feature selection. We then generated a player and season based key id that maps to the original clutch dataset. Finally we merged the play-by-play high leverage dataset into the clutch dataset and replaced the original clutch dataset with the augmented clutch dataset in model training.

Algorithm Refinement

Muniz's algorithm follows these steps:

0. Perform PCA for Dimension Reduction
1. Perform k-Means Clustering
2. Build Weighted Network, $\mathcal{WN}(k)$
3. Perform the CD Algorithm on Each $\mathcal{WN}(k)$

Muniz conducts a thorough sensitivity analysis covering alternative choices that were the basis for our approach and innovations:

- PCA and data selection: We included additional data points covering high leverage situations, which provides a larger set of players that the original clutch metrics. In Muniz' sensitivity analysis, including more players from the clutch dataset led to a higher modularity score. We found the same result by increasing the number of players in the clutch cluster with augmented high leverage data.
- Clustering: We used k-means for clustering. The same k is chosen for all datasets on each run to prevent a combinatorial search of using multiple k values across different datasets. The micro clusters can be considered a first pass at clustering and represent the 6 core datasets we used: Scoring, Passing, Defense, Rebounding, Clutch, and Misc.
- Weighted Network: To build the weighted network, each player is assigned an edge based on the number of times they appear in a similar cluster generated in the previous step. That edge is then scaled by the number of total datasets in which each player appeared. The maximum value of an edge is 1 if a player appears in the same cluster as another player across all six datasets. The minimum value of an edge is 1/6, which occurs when a player appears in only one cluster across the six datasets. The author considered different binning strategies and found that a four-bin strategy in Table 1 led to the best separation among groups.
- Community Detection: We used the Louvain method to ensure clusters were meaningfully sized and representative of players within them. The Louvain algorithm, considered state-of-the-art in weighted network clustering, optimizes modularity. Modularity is defined as the relative density of edges inside communities with respect to edges outside communities. These final clusters are considered macro clusters and are the basis of our player archetypes.

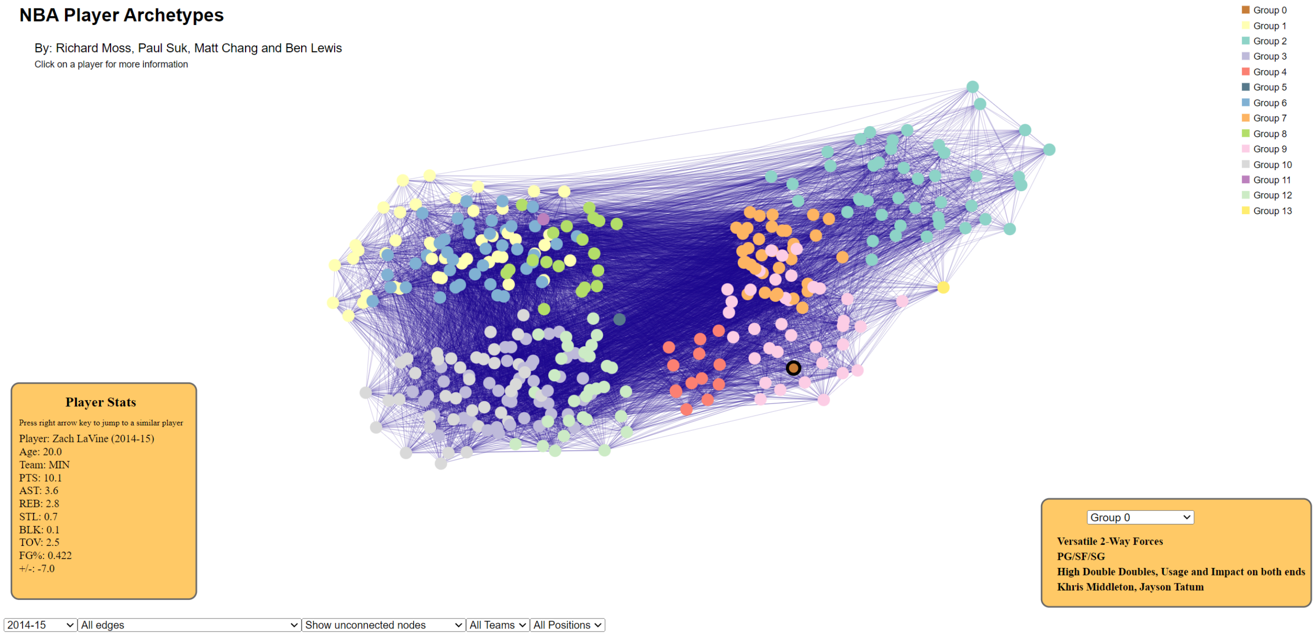
Table 1. Arc weights based on co-occurrence percentages

Percentage of Time Co-occurring	Arc Weight
$[0, 25\%)$	0
$[25, 50\%)$	1
$[50, 75\%)$	2
$[75, 100\%)$	3

Visualization Improvement

We utilized d3.js to create an interactive visualization shown in Figure 1. The tool itself allows adhoc analysis of player groups. Its portability and ease of use allows it to be shared without necessary knowledge of the underlying algorithms or assumptions.

The d3 visualization uses a force directed graph to separate the player clusters based on their similarity across micro clusters. It gives the user the ability to


NBA Player Archetypes

By: Richard Moss, Paul Suk, Matt Chang and Ben Lewis

Click on a player for more information

Player Stats

Press right arrow key to jump to a similar player

Player: Zach LaVine (2014-15)
Age: 20.0
Team: MIN
PTS: 10.1
AST: 3.6
REB: 2.8
STL: 0.7
BLK: 0.1
TOV: 2.5
FG%: 0.422
+/-: -7.0

Group 0

Versatile 2-Way Forces
PG/SF/SG
High Double Doubles, Usage and Impact on both ends
Khris Middleton, Jayson Tatum

2014-15 | All edges | Show unconnected nodes | All Teams | All Positions

Figure 1. 2014-2015 Player Archetype Network

filter the nodes based on the season, player team, and player position. The edges can also be filtered to identify stronger player connections and to hide nodes with no connections. These filters can be used simultaneously to get more comprehensive results, as shown in the bottom left of Figure 1.

To improve interactivity - we include the player card in the bottom left of Figure 1 that includes their name and per game pace adjusted statistics when you select a node in the main visualization. From that player card, users can press the right arrow key to jump to another player with a high similarity score. This allows users to quickly jump between similar players and evaluate them. The user can also press enter at any time to jump to any random node. Selecting a player using any of these methods require that the node be visible based on the applied filters.

Additionally, we include the dropdown in the bottom right of Figure 1 that includes a name for the archetype, positions, a brief description, and representative players.

Experiments

Scalability Evaluation: How does the project scale to include new seasons?

The expectation is to refresh the generated visualization on an annual basis. The number of players in each season would remain relatively static and new season performance metrics will be generated for said players. Therefore the data will scale linearly annually. As the current season progresses, the NBA data visualization tool would refresh on a monthly basis. There needs to be enough new game data from the players to justify a refresh. To make sure that the project will be able to scale accordingly, we broke it down into subsections to ensure there are no bottlenecks at each stage.

- Data Pipeline - We successfully executed the data pipeline for each season on a relatively linear scale, with each season comprising approximately 130 MB of play-by-play data. Consequently, the runtime and storage requirements for refreshing the data should not pose significant concerns for our research.
- Model Training - The model requires approximately one hour for training and must be retrained for each season. We anticipate mini-

mal variation in players' archetypes between seasons, enabling us to utilize the hyperparameter grid search space from the previous season. This approach will reduce the runtime following the initial model training. As visualizations are not updated in real time, the model training should be capable of accommodating refresh rates on a daily or weekly basis.

- Data Visualization. - The interactive force graph data visualization employs a collection of pre-computed graphs, incorporating filter parameters to facilitate user navigation. Specifically, the data visualization tool utilizes the season as a filter parameter to display NBA archetypes exclusively for a particular season. As each season progresses, we will augment the visualization tool with an updated set of pre-computed graph data, using the Python notebooks included in the software package.

Player Archetype Accuracy: How closely do the player archetypes match our intuition?

This can be evaluated by our group (with basketball knowledge) as well as other NBA data communities (e.g. Reddit) or as the project is scaled, it can be distributed to media members with knowledge of advanced statistics for futher input.

A fundamental validation step involves examining the correlation between player categorization and the positions of players within each community. Generally, it is anticipated that players belonging to a community will possess similar positions. Although communities may further delineate a specific position or comprise players from two closely related positions, it is expected that most communities will not typically encompass players from all positions or positions with vastly divergent roles.

A more comprehensive assessment would be to calculate a mean Average Precision approach from user feedback. The correct classification for each player is subjective, so this consensus method could reduce bias in player archetype classification evaluation. The method would include creating a feature in the visualization that would allow users to report players they feel are misclassified. Taking this feedback data, we can calculate the precision from each user and create a mean Average Precision of the model with the formula below where C is the number of classes, N is the feedback, and TP_i is true positive for class i and FP_i is false positive for class i.

$$\text{mAP} = \frac{1}{C}\sum_{c=1}^{C}(\frac{1}{N}\sum_{i=1}^{N}\frac{TP_ic}{FP_ic + TP_ic})$$

It is advised to use this metric with caution as the single source of model evaluation as human subjective judgement should not be taken as the ground truth as the model could have predictive power that is not apparent to human judgement.

Usability: How intuitive and useful is the tool?

The effectiveness of the visualization is assessed by its intuitiveness in terms of operation and interpretation of the metrics. Evaluating this aspect can be achieved by soliciting feedback from both casual users and dedicated NBA fans. In the alpha testing phase, we present a summary of the results obtained from our team, an external UX designer, and additional individual testers:

- Intuitiveness - The visualization is one page and easy to navigate. Some areas of improvements are:
  - The right arrow is used to jump between players but users can't go back with left arrow
  - Hard to know when selecting a new node if unfamiliar with player stat card
  - Yellow is hard to see; consider more accessible colors
  - Initally, the drop downs were below the fold on certain browsers
- Usefulness

  The tool seems easy to to explore. Areas of improvement include:
  - Overall performance. Reduce the frequency of calculations to speed up rendering.
  - Provide group labels in the network - looking back and forth between group numbers

on the legend and group description card provides less affordance compared to including them directly in the visualization

- Add move details/better mouse over on player hover - the network is dense and makes it hard to see individual player names

After fine tuning even further, the plan is to perform beta testing to known NBA fans on the r/NBA subreddit. With beta testing, we can expand to evaluate quantitative UX metrics such as User Retention, Time-on-Task and Customer Satisfaction.

Conclusion

In this paper, we propose a new way to interactively visualize NBA player archetypes. To evaluate goodness of fit, we provide several metrics that establish criteria to evaluate players. Our interactive tool allows a human element to evaluate the networks rapidly. With an extensive series of parameters, one is able to construct individual networks that match their intuition. We hope that the end user is able to more tangibly understand and appreciate player clusters using our tool rather than getting specific numbers. The visualization allows us to understand the data in a better way than simply viewing it in a static table.

Implications: An exciting implication is the ability to project future stars based on their membership in groups. Shai Gilgeous-Alexander, relative unknown in the 2019-2020, is in that season's "superstar" cluster with 2019 NBA champion and Finals MVP Kawhi Leonard. Gilgeous-Alexander had his breakout season this year and is a finalist for the Most Improved Player Award and is very likely to make All-NBA (awarded to the best 15 players in the league). As the NBA becomes increasingly position-less, that fact being often mentioned by members of the media on broadcast, the visualization lets us observe the movement of players in different positions season-by-season. As the overall structure of the graph and the cluster membership of players by position change we expect to be able to draw some conclusions about how to build teams in this new era of basketball.

Limitations: It is essential to acknowledge that our evaluation is restricted to the available data. There are countless player situations that do not occur in any game within our dataset. Generating metrics based on simulated or hypothetical situations would necessitate the development of a deterministic or probabilistic model of basketball, which is beyond the scope of this study. Moreover, the analysis is subject to the inherent variability of made baskets. Two shots with the same expected value can result in either a make or a miss. Additionally, players may experience prolonged cold or hot streaks throughout a season, where their performance deviates significantly from their career averages.

Potential Future extensions: The proliferation of NBA data allows us to imagine many future directions. The well documented limits of the box score could hopefully give way to more robust player tracking data. There is promising research done with nba2vec [7], which creates embeddings for individual NBA plays and actions. Representing players by learned features from play-by-play data rather than hand-picked heuristics or aggregate statistical measures could lead to better approximations of similarity.

These embeddings could also be used for NCAA or international players as a more accurate way to scout their capabilities. It would be easier to capture player tendencies with the new model rather than adjusting their statistics for the level of competition. Some college statistics such as Rebound and Block rate correlate to NBA production, but many others such as scoring do not.

All team members have contributed a similar amount of effort.

References

[1] *PBP Stats API*. 3

[2] Muthu Alagappan. Redefining the positions in basketball. *MIT Sloan Sports Analytics Conf.*, 2012. 2

[3] Darryl Blackport. Leverage stats on pbpstats.com. 2023. 3

[4] Yago Colás. *Mapping the Geography of the NBA*. University of Nebraska Press, 2020. 2

[5] Rubén Dehesa Suances, Alejandro Vaquera, Bruno Gonçalves, Nuno Mateus, Miguel Ruano, and Jaime Sampaio. Key game indicators in nba players' performance profiles. *Kinesiology*, 51:92–101, 03 2019. 2

[6] Kirk Patrick Goldsberry. *Sprawlball: A visual tour of the New Era of the NBA*. Houghton Mifflin Harcourt, 2019. 2

[7] Webster Guan, Nauman Javed, and Peter Lu. Nba2vec: Dense feature representations of nba players, 2023. 6

[8] Graydon R. Levine. All-nba team voting patterns: Using classification models to identify how and why players are nominated. 2019. 1

[9] An Liu Pingping Guo Cong Liu Libao Zhang, Faming Lu. Application of k-means clustering algorithm for classification of nba guards. *International Journal of Science and Engineering Applications*, 2016. 2

[10] Megan Muniz and Tülay Flamand. Sports analytics for balanced team-building decisions. *Journal of the Operational Research Society*, 2022. 2

[11] Megan Muniz and Flamand Tulay. A weighted network clustering approach in the nba. *Journal of Sports Analytics*, 2022. 2

[12] Riki Patel. Clustering professional basketball players by performance. *UCLA*, 2017. 2

[13] Jonathan Bosch Samuel Kalman. Nba lineup analysis on clustered player tendencies: A new approach to the positions of basketball modeling lineup efficiency of soft lineup aggregates. *MIT Sloan Sports Analytics Conference*, 2020. 2

[14] Stephen Shea. *Basketball Analytics: Spatial Tracking*. Createspace Independent Publishing Platform, 2014. 2

[15] Stephen M. Shea and Christopher E. Baker. *Basketball analytics: Objective and efficient strategies for understanding how teams win*. Advanced Metrics, 2013. 2

[16] Frank Daumann Tobias Berger. Jumping to conclusions – an analysis of the nba draft combine athleticism data and its influence on managerial decision-making. *Sport, Business and Management*, 2021. 1