Classifying Teams in the NBA with Player Behavioral Data

by

Colin Poler

Submitted to the Department of Mechanical Engineering
in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science in Mechanical Engineering

at the

Massachusetts Institute of Technology

June 2018

Signature redacted

Signature of Author: _____

Colin Poler
Department of Mechanical Engineering
May 11, 2018

Signature redacted

Certified by: _____

Peko Hosoi
Associate Dean of Engineering/Professor
Thesis Supervisor

Signature redacted

Accepted by: _____

Rohit Karnik
Professor of Mechanical Engineering
Undergraduate Officer

Classifying Teams in the NBA with Player Behavioral Data

by

Colin Poler

Submitted to the Department of Mechanical Engineering
on May 11, 2018 in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science in Mechanical Engineering

Abstract

I use SecondSpectrum play-by-play data from the 2016-2017 NBA season to assemble behavioral event data for each player. Behavioral data includes propensity to dribble/pass/shoot, and also the resulting quality of shot when players decide to shoot or make another pass. I apply a k-means clustering algorithm to cluster teams based on their starting lineup behavior data; the clusters show different team makeups within the behavioral data collected. In particular, the clustering identified pass-heavy vs dribble-heavy offenses, and good shot-decision making teams.

Thesis Supervisor: Peko Hosoi

Title: Associate Dean of Engineering/Professor

Table of Contents

Introduction

This paper attempts to classify NBA teams from the 2016-2017 season by behaviorally characterizing the players on the team. In particular, it focuses on the decision making trends of individual players on a team relative to other players in the NBA to differentiate different player behaviors. The paper then groups teams with similar compositions of player behaviors to suggest some classifications of team play style.

NBA teams are increasingly using analytics to guide play strategy, choose lineups and manage player health. For instance, analytics have shown that 3 point shots are more efficient at scoring than two-point jump shots.[1] However, most studies tend to analyze tactical choices or focus on individual player activity, and miss an opportunity to observe team dynamics that are critical in a passing-centric game like basketball.[2]

A study on the 2010 NBA playoffs introduced network analysis to examine the effect of passing probabilities on scoring. The results suggest that moving the ball more frequently and predictably to a single player (e.g. the best shooter) has a negative effect on scoring probability because the opposition can adjust their defense to counter.[3]

The network passing analysis led to a question: how do players decide when to shoot, or when to make an additional pass to reduce predictability? Of course, this question is in general nigh on impossible to answer, so this paper takes a statistical approach to suggest how different players' decisions vary, and how successful their decisions are.

---

[1] Kopf
[2] Fewell et al.
[3] Fewell et al.

This analysis is made possible by using data obtained from the Second Spectrum camera system installed on all NBA courts. The Second Spectrum system tracks the players and ball 25 times a second, then logs the players' passes, dribbles and shots.[4]

$$\text{EFG} = \frac{(\text{two-point shots made}) + 1.5(\text{three-point shots made})}{\text{shot attempts}}$$

Equation 1: Definition of Effective Field Goal Percentage. Shots are counted across the entire season for each player.

EFG is Effective Field Goal percentage, calculated as average number of points per shot divided by two; it indicates how effective a player's shots are accounting for the point premium of 3 point shots. [5]

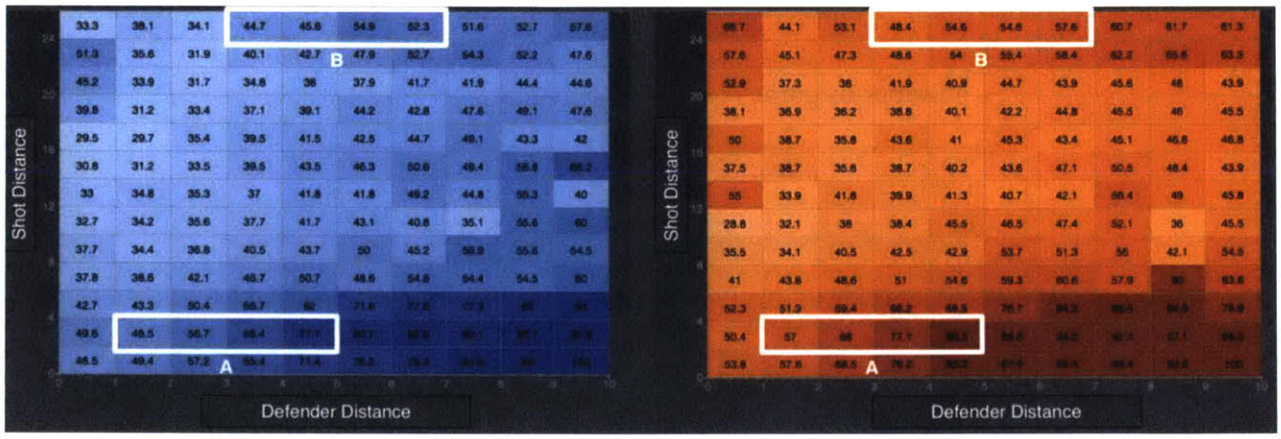

Figure 1: Computation of Effective Shot Quality from the Chang et al. paper. Shots taken off the dribble use the blue matrix on the left, and shots taken after catching a pass use the orange matrix on the right. The 'shot distance' between the player and the net and the 'defender distance' between the player and the nearest defender are used to look up the appropriate value.

ESQ is Effective Shot Quality, calculated for each shot as a function of the distance to the net, the distance of the nearest defender and whether the player was dribbling immediately prior. [6]

$$\text{EFG+} = \text{EFG} - \text{ESQ}$$

---

[4] McCann
[5] Chang et al.
[6] Chang et al.

Equation 2: Definition of EFG+: EFG is corrected for shot quality by subtracting ESQ.

EFG+ is a correction for EFG, computed as EFG minus the average ESQ of the shots made; it is a measure of a shooter's skill corrected for the difficult shots they are asked to make. [7]

Methods

The first step in the analysis was dividing all the logged events by possession, and classifying each event. Fortunately, the Second Spectrum data lists a time range for each possession. Each possession was associated with the actions that occurred during the listed time range, and these actions were processed with a finite state machine to determine when players make a pass, start to dribble or take a shot.[8]

Number of games	1,304
Number of players	398
Number of possessions	253,443
Number of holds	1,044,663

Table 1: A count of the number of games, players, possessions and holds that I processed from the SecondSpectrum data.

The next step was to visualize the data, and the first question was simply how often each player chooses to dribble, pass or shoot. From the possessions data compiled above, we can easily tally how often each player does each of passing, dribbling or shooting. Note that while there are three options with three associated probabilities, the probabilities must sum to 1, and therefore there are only two degrees of freedom. With this in mind, the propensities of each player to do a certain action can be scattered on a triangle, where each vertex of the triangle represents a player that

---

[7] Chang et al.
[8] Python code available at https://github.com/colinpoler/Classifying-NBA-17

would always take the same action. Because all the players clustered in the same area of the triangle, later plots of this style will be zoomed in to the region of interest.

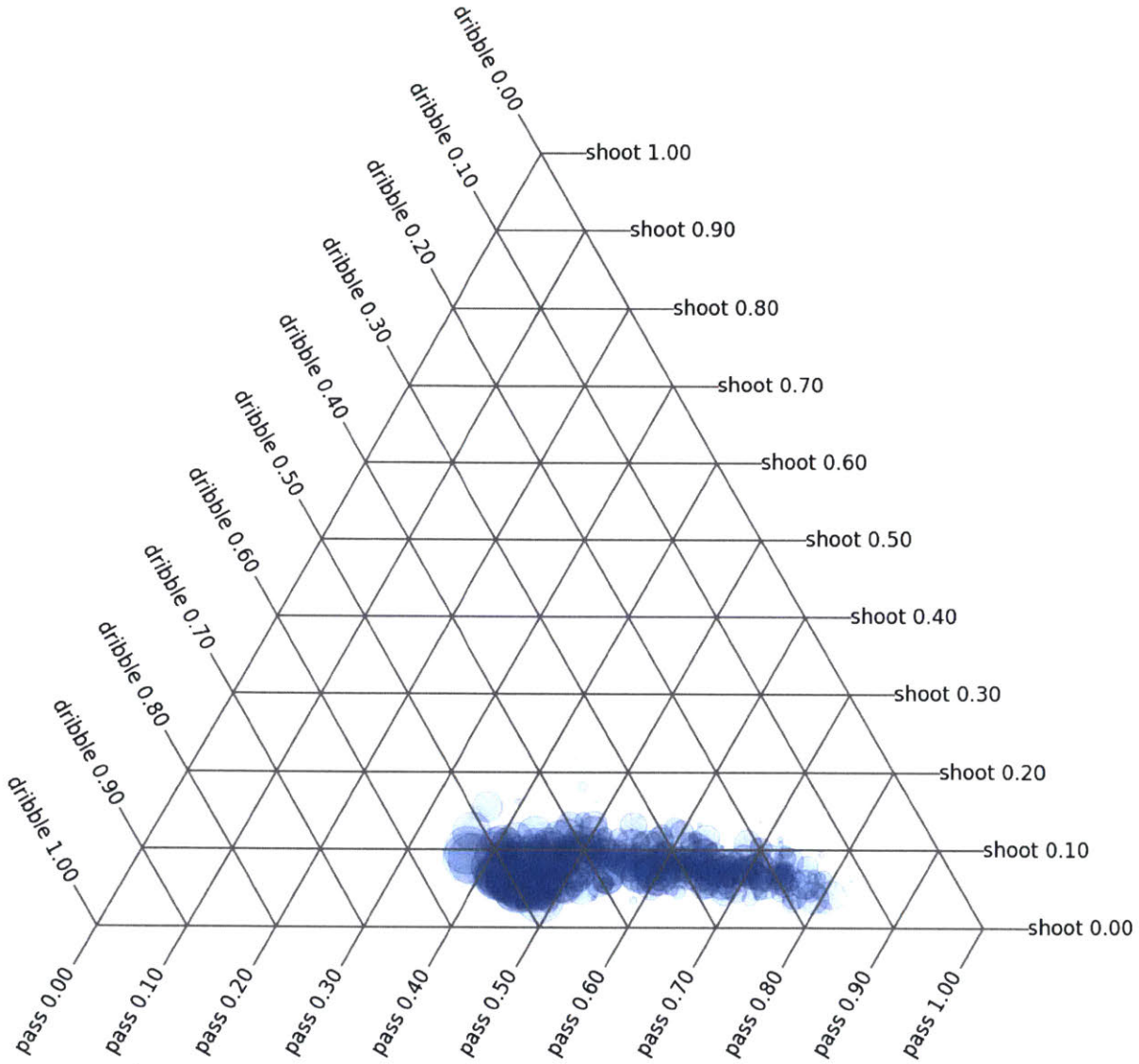

Figure 2: A scatter plot of each player's propensity to pass, dribble or shoot. Each circle represents a single player across all regular games in the season, and the position of the circle indicates the observed frequency of passing, dribbling or shooting. The size of each circle indicates how many of these actions were observed during the games.

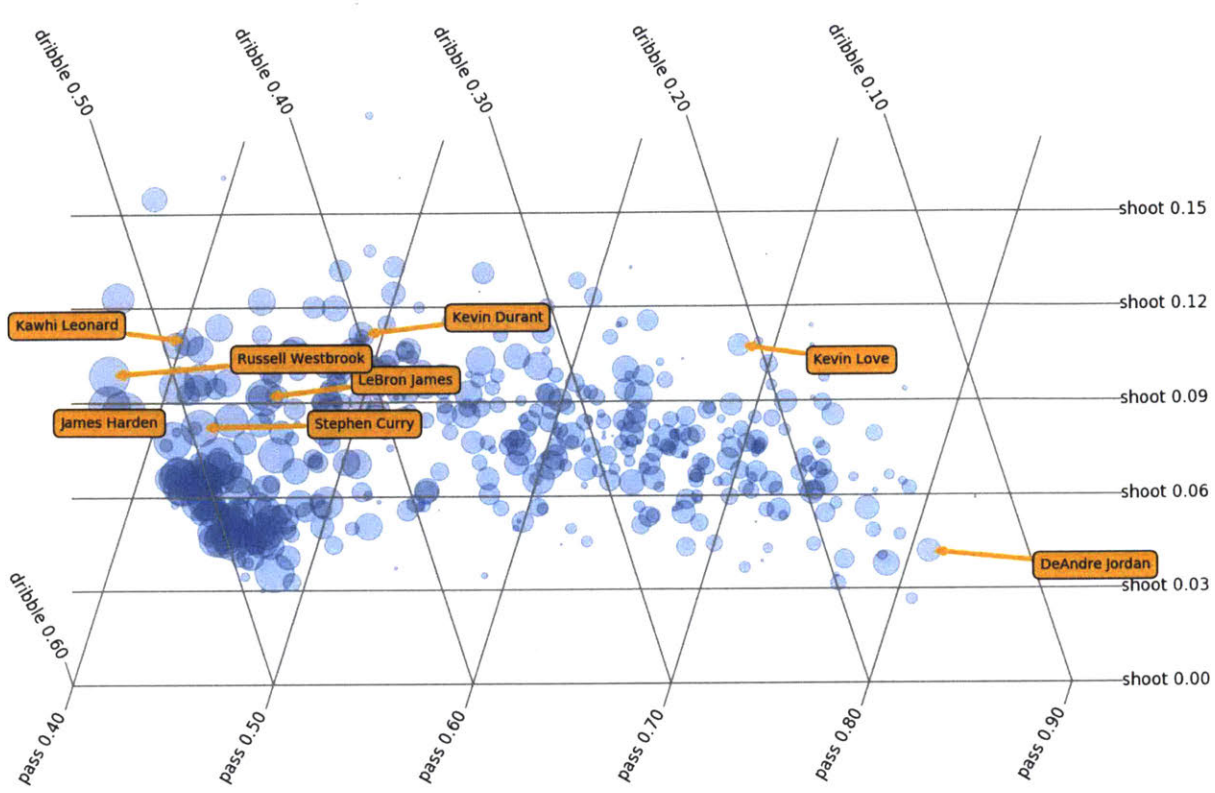

Figure 3: A zoomed-in version of the propensity to dribble, pass and shoot, with some example players labelled.

The second question was how effective each player is at discerning when an additional pass would increase shot quality. Looking at each possession, I record the player taking the shot and the resulting shot quality; I also record the player that had immediately passed to the shooter.[9] For each player, I computed the average shot quality for which the player himself is shooting, and the average shot quality for which the player he passed to immediately prior shoots. The resulting scatter plot is tightly clustered around the average shot quality of about 52, but the points do show a statistically significant spread.

---

[9] This analysis is unable to *directly* compare the shot quality that the player was facing versus the shot quality of their teammate; instead, it compares only the shot qualities where the player actually shoots against the shot qualities of the teammate when they choose to pass. Further analysis could attempt to correct this by computing the would-be shot quality when players decide *not* to shoot, and directly investigate the player's decision making.

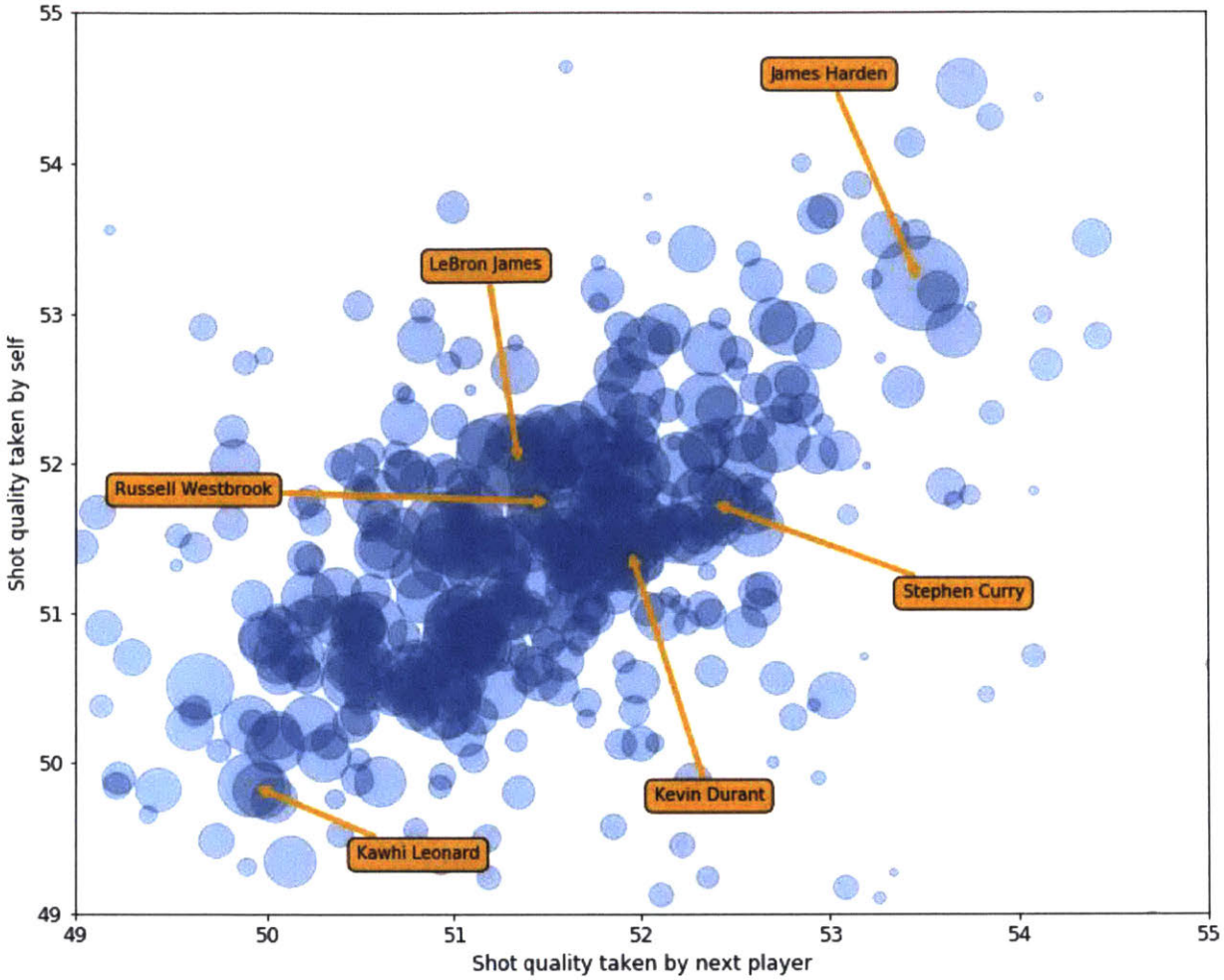

Figure 4: A scatter plot of shot-decisions for each player. Each circle represents a player over the entire regular season. The position of each player on the vertical axis represents the average shot quality for shots he takes himself. The position of each player on the horizontal axis represents the average shot quality for the shots taken by the player he passes to (when that player does shoot).

For each team, the players on the starting lineup were located on both of these plots and their positions recorded. I define the difference between two such teams located on the plots as sum of the distances between corresponding players, where players are chosen to correspond such as to minimize the sum.

$$\boldsymbol{correspondence\_distance}(s,t) = \min_{a \in A} \sum_{i} \|s_i - t_{a_i}\| \quad \text{for } A = \text{permutations}(1,2,3,4,5)$$

I define the mean of several teams to be a team of hypothetical players such that the sum of the distances to each of the teams to be averaged is minimized. This was computed by choosing positions for each player in advance, and finding the arithmetic mean of the positions of each player.

Input: *all teams* is a list of teams, each of which is a list of players' behavior (x,y)

groups ← n empty lists
for each *team* in *all teams*:
  assign *team* to a **random** *group* in *groups*

while iterations without **logging a fault** < 500:
  for each *group* in *groups*:
    if *group* is empty:
      log a fault
      assign 3 random *teams* to *group*
    prototype ← first team in group
    mean of *group* ← list of 5 copies of (0,0)
    for each *team* in *group*:
      reorder *team* to minimize $\sum_i \|prototype_i - team_i\|$
      mean of *group* ← *mean* of *group* + *team* / **length**(*group*)
    unassign each *team* in *group* from *group*

  for each *team* in *all teams*:
    choose *group* to minimize **correspondence_distance**(*mean* of *group*, *team*)
    assign *team* to *group*

Figure 5: Pseudocode for k-means clustering algorithm used to cluster teams.[10]

Finally, a k-means clustering algorithm was used to group and choose means for the clusters. In particular, the teams were initially randomly assigned to one of k groups. A mean is computed for each group, and the teams are reassigned to the group for which they are closest to the mean. If any group is empty, the mean is reassigned to the average of three random teams; an occurrence of this reassignment is called a *fault*. The above is repeated for many iterations, until no *fault* has

[10] Python code available at https://github.com/colinpoler/Classifying-NBA-17

occurred for 500 iterations. Several values of k were evaluated; a small value of k was selected such the average distance of each group from its mean would not decrease much for higher values of k, and such that the resulting clusters appeared to have some visual pattern. In particular, k=4 was chosen for both plot types.

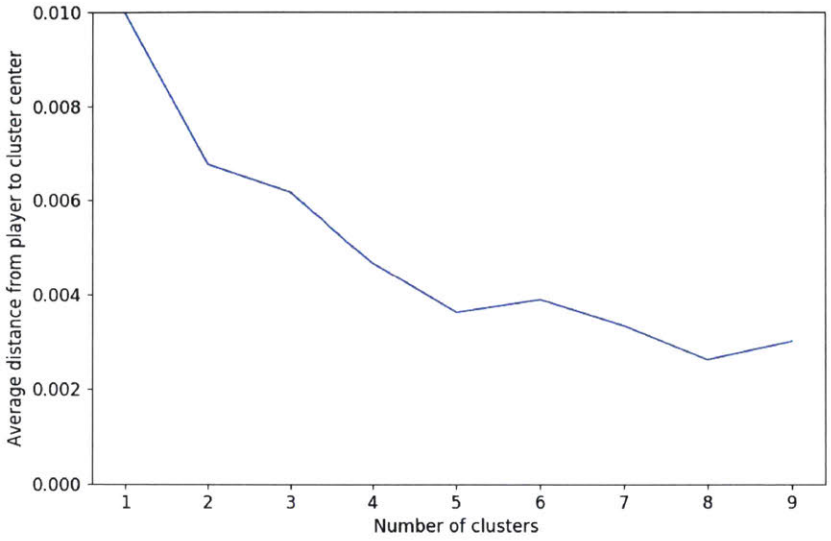

Figure 6: Convergence of clustering for the dribble/pass/shoot-propensity plot. The curve shows the average distance from players to their assigned mean location, revealing that there is not much improvement after 4 clusters.

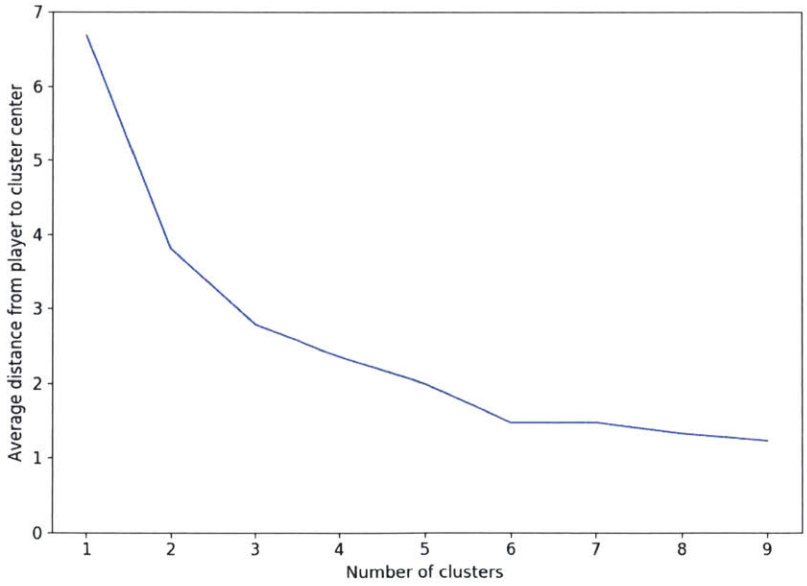

Figure 7: Convergence of clustering for the shot-decision plot. The curve shows the average distance from players to their assigned mean location, revealing that there is not much improvement after 4 clusters.

Results

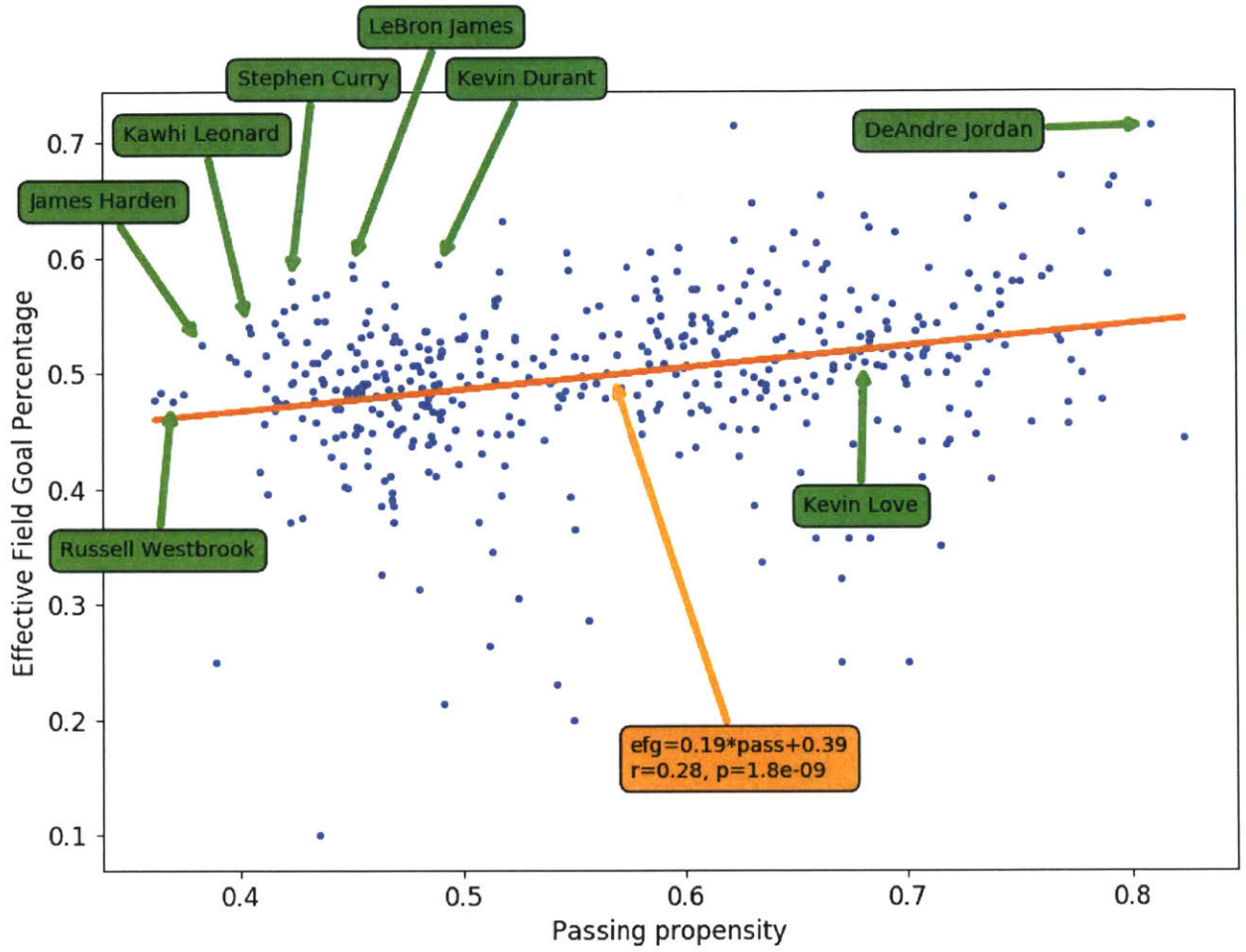

Figure 8: Correlation between passing propensity and Effective Field Goal Propensity (proportional to points scored per shot). A correlation coefficient of 0.28 is achieved, which is significant to 99% certainty. Causality could not be established.

Looking first at passing propensity for all players, there is a fairly significant correlation between passing propensity and effective field goal percentage: higher passing propensity is associated with higher effective field goal percentage. Interestingly, all the MVP candidates for this season had low passing propensity, but were high outliers in effective field goal percentage.

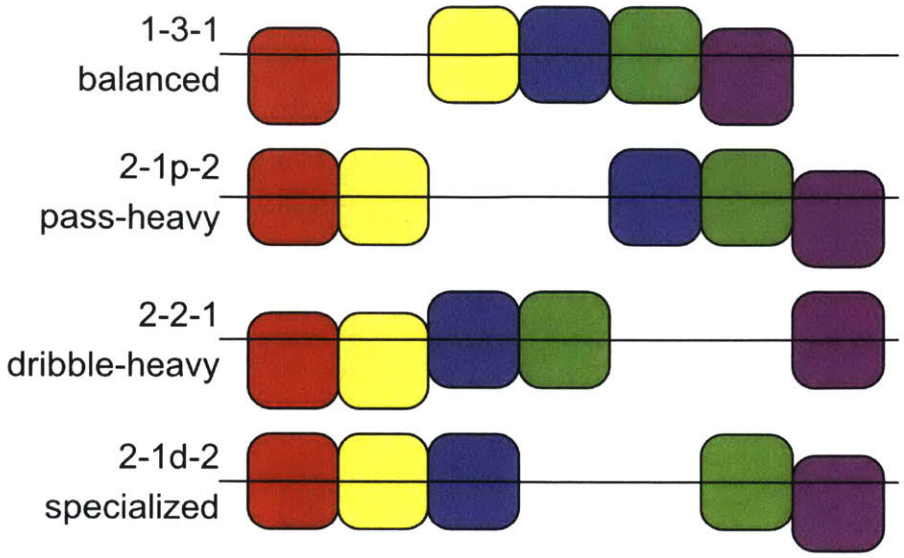

Figure 9: A comparative schematic of the clusters found in the dribble/pass/shoot-propensity plots. The colors show the relative positions of the clusters along the dribble-pass axis, and the clusters sitting lower on the line show a lower propensity to shoot.

Then, I present the results of clustering teams on the dribble/pass/shoot-propensity plots. Four clusters are found. Cluster 1-3-1 is 'balanced' in that most players on these teams choose to dribble or pass very near the center of the distribution. Cluster 2-1p-2 is 'pass-heavy' in that 3 of 5 players on these teams are very likely to pass instead of dribble. Cluster 2-2-1 is 'dribble-heavy' in that 4 of 5 players are relatively likely to dribble instead of pass. Cluster 2-1d-2 is 'specialized' in that 3 players are relatively likely to dribble, and the other 2 players are relatively likely to pass.

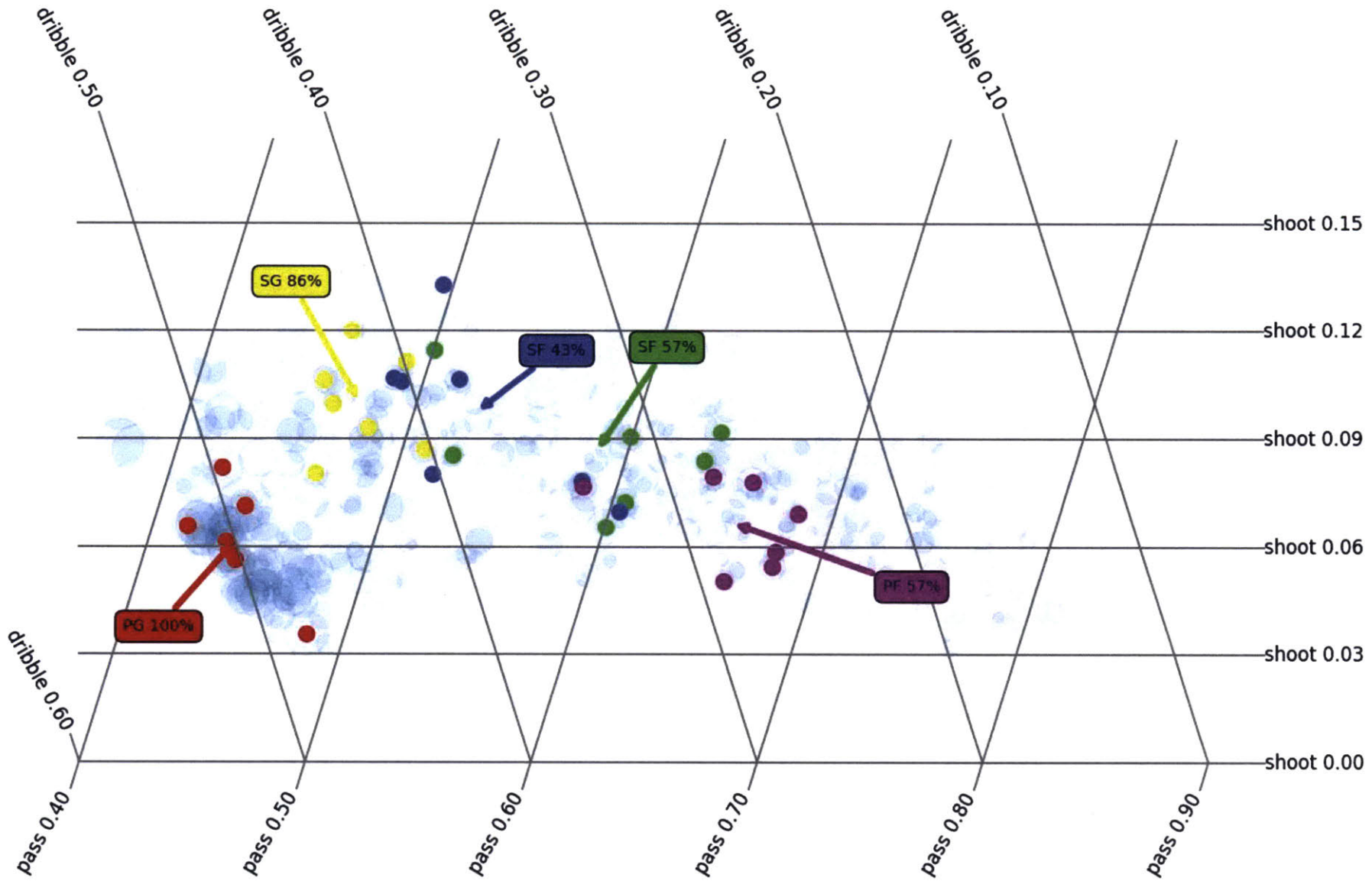

Figure 10: Cluster 1-3-1 is 'balanced' in that most players on these teams choose to dribble or pass very near the center of the distribution. This cluster consisted of: Golden State Warriors, Philadelphia 76ers, Memphis Grizzlies, Orlando Magic, Boston Celtics, Denver Nuggets, Detroit Pistons. The average points-per-possession was 1.09.

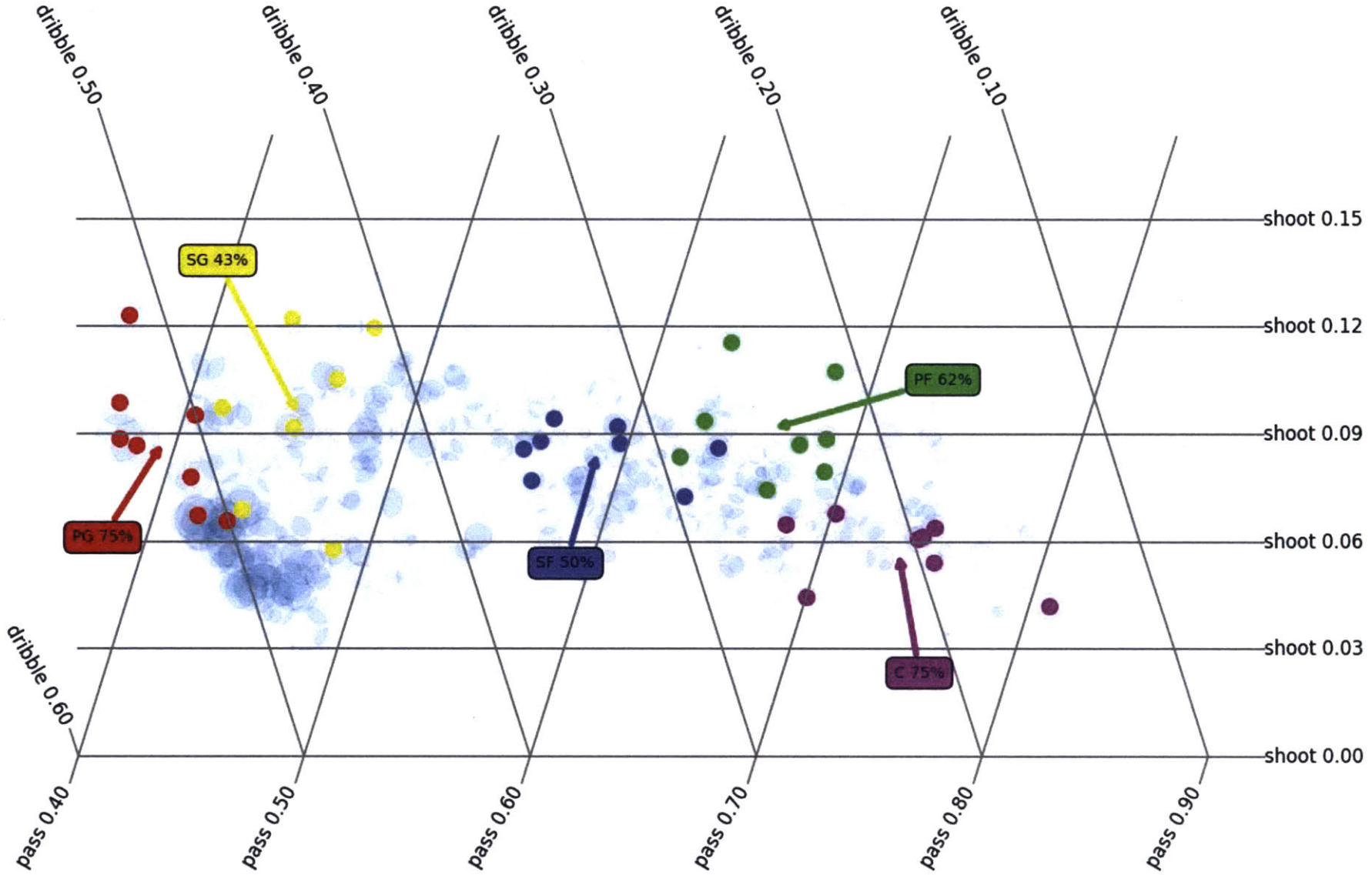

Figure 11: Cluster 2-1p-2 is 'pass-heavy' in that 3 of 5 players on these teams are very likely to pass instead of dribble.This cluster consisted of: Houston Rockets, Cleveland Cavaliers, Oklahoma City Thunder, New York Knicks, Toronto Raptors, Washington Wizards, Los Angeles Clippers, Portland Trail Blazers. The average points-per-possession was 1.12.

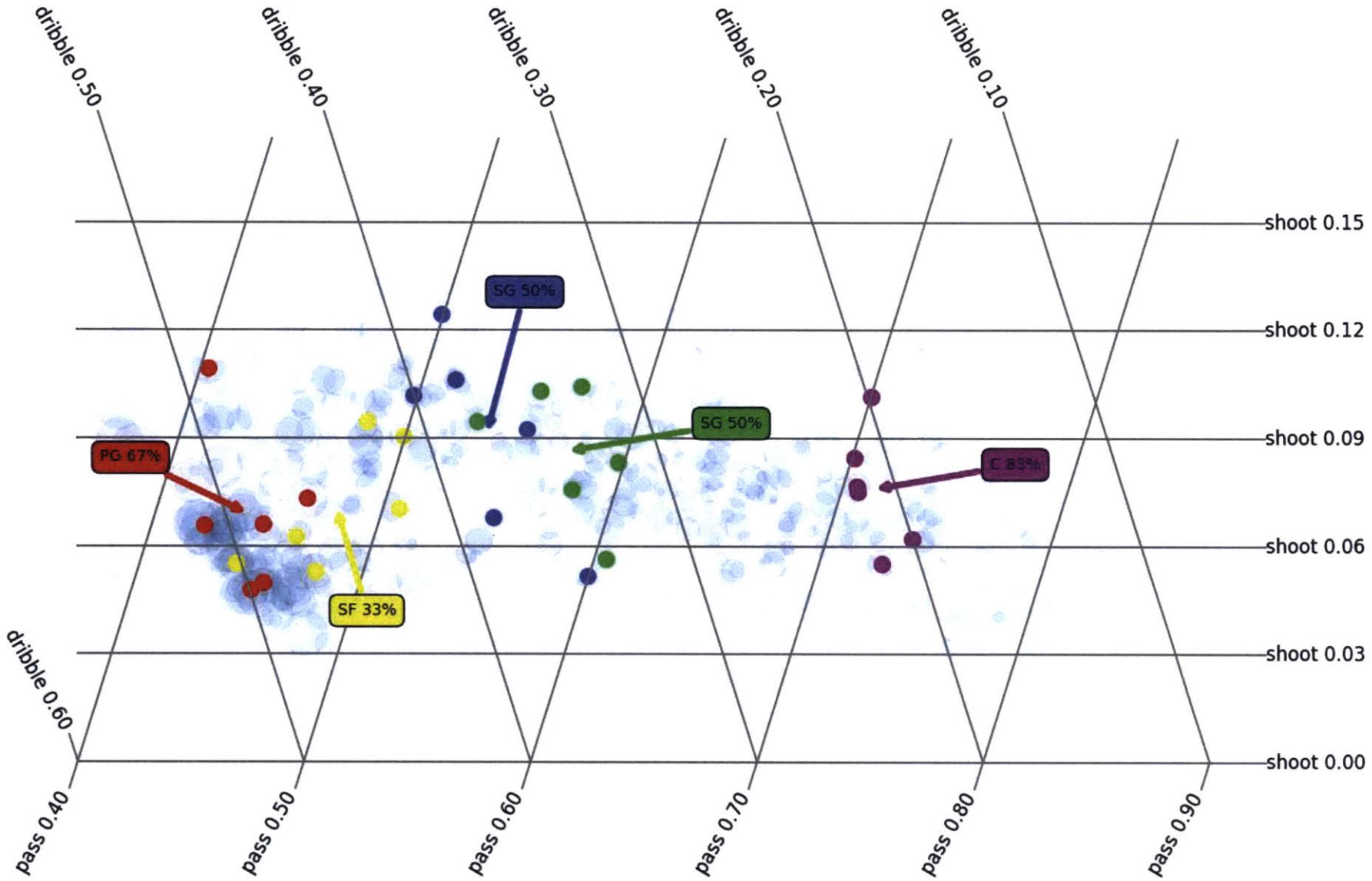

Figure 12: Cluster 2-2-1 is 'dribble-heavy' in that 4 of 5 players are relatively likely to dribble instead of pass.This cluster consisted of: Milwaukee Bucks, Sacramento Kings, Atlanta Hawks, Los Angeles Lakers, Dallas Mavericks, San Antonio Spurs. The average points-per-possession was 1.08.

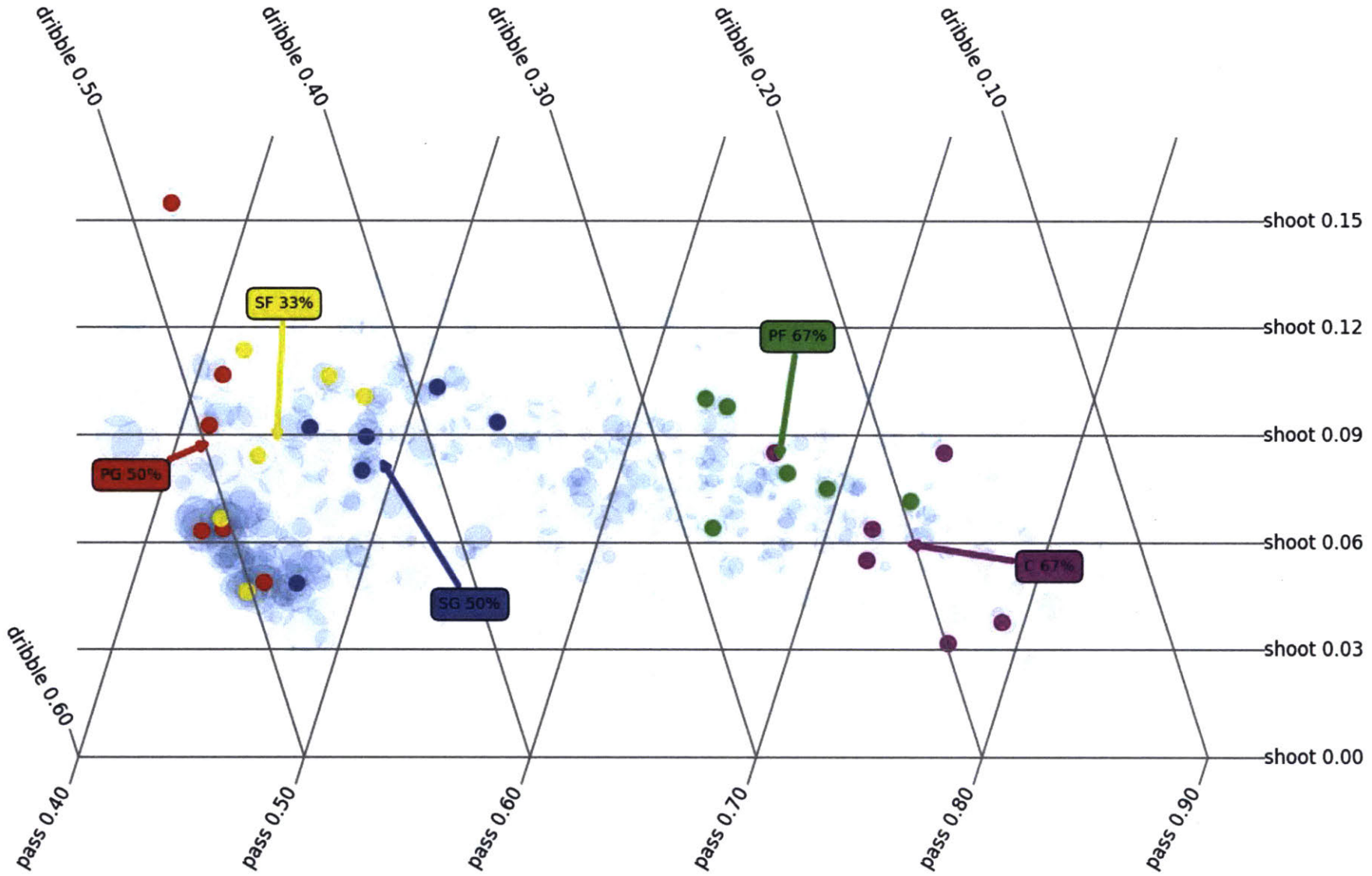

Figure 13: Cluster 2-1d-2 is 'specialized' in that 3 players are relatively likely to dribble, and the other 2 players are relatively likely to pass. This cluster consisted of: Minnesota Timberwolves, Phoenix Suns, Chicago Bulls, Indiana Pacers, Utah Jazz, Miami Heat. The average points-per-possession was 1.13.

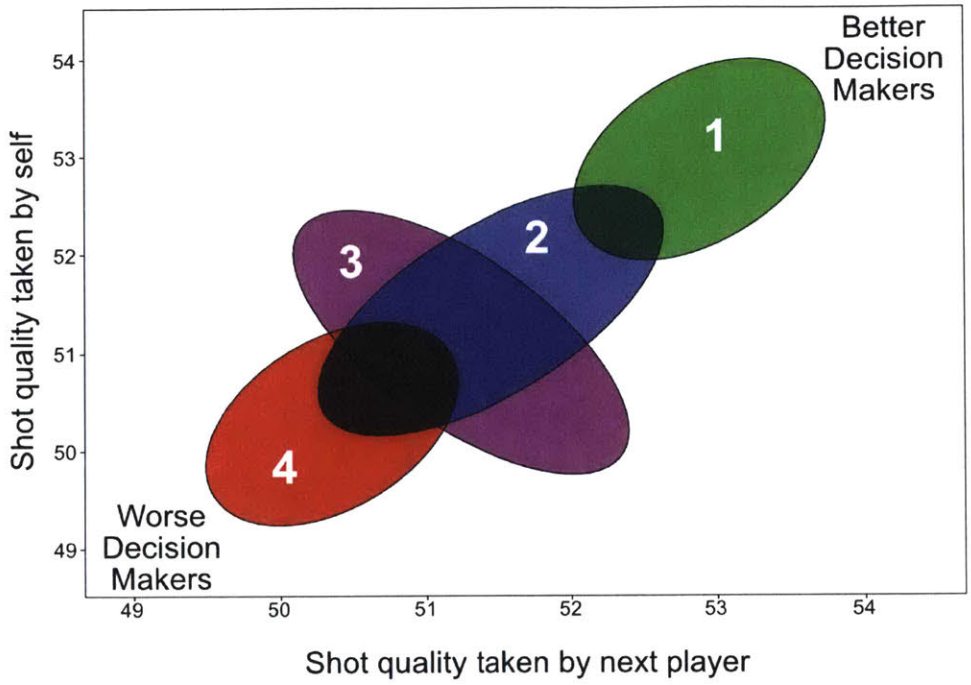

Figure 14: A comparative schematic of the clusters found in the shot-decision plots. The ellipses show the relative positions of the teams clustered. Each ellipse shows a different rough shape in the plot space.

Then I present the results of clustering on the shot-decision plot. Cluster 1 teams are centered on the top right, which means the players are skilled at discriminating when to shoot and when to pass. Cluster 2 teams are nearer the center, and have a variety of players more likely to pass or to shoot. Cluster 3 teams are near the center, and have a variety of skill levels in pass-shot decisions. Cluster 4 teams are near the bottom left, indicating that the players aren't discriminating well when to shoot or pass well.

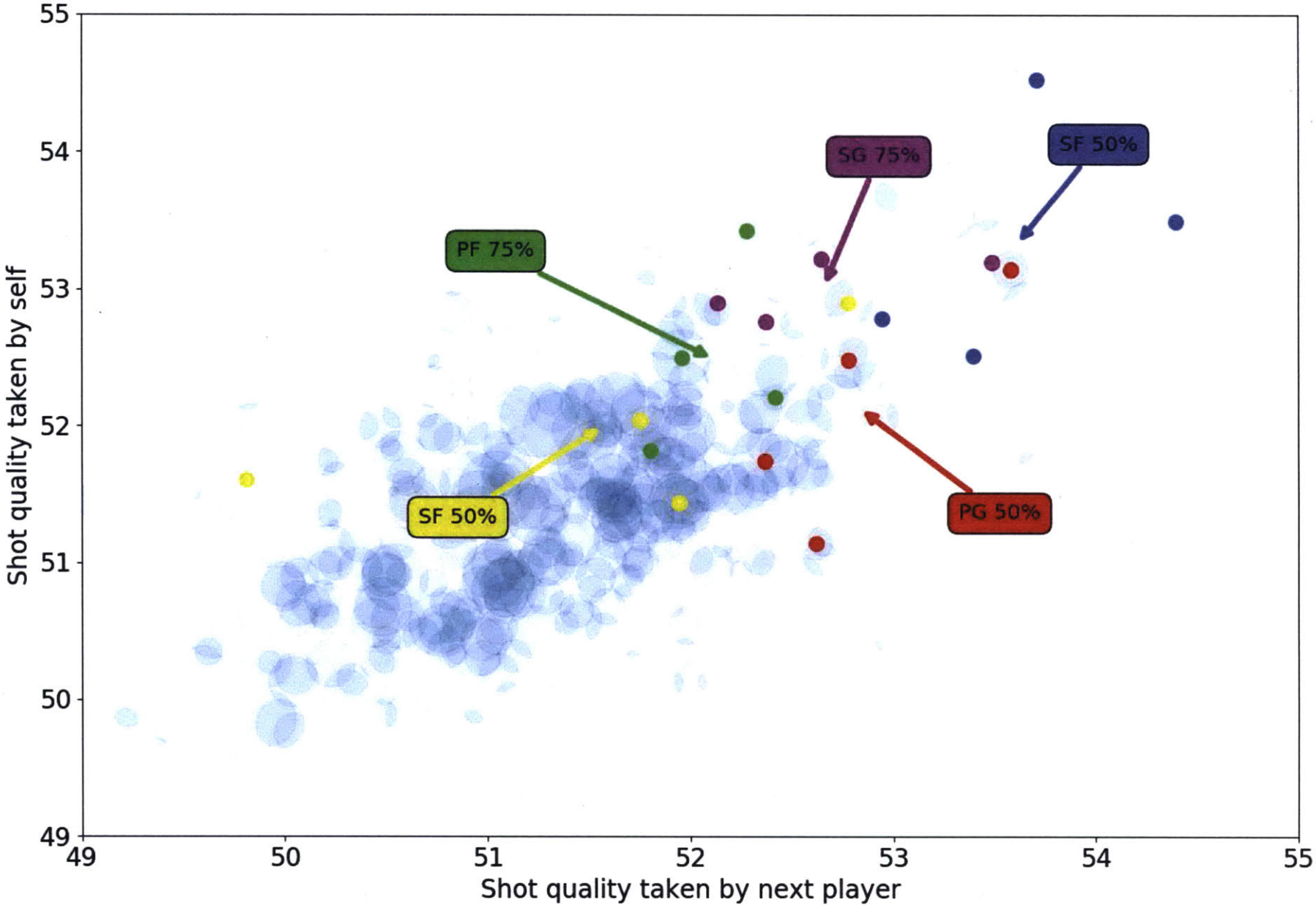

Figure 15: Cluster 1 from the shot-decision plots. This cluster shows teams whose players are in the top right corner, indicating that players are skilled at choosing which shots to take. This cluster consisted of: Golden State Warriors, Philadelphia 76ers, Houston Rockets, Milwaukee Bucks. The average points-per-possession was 1.11.

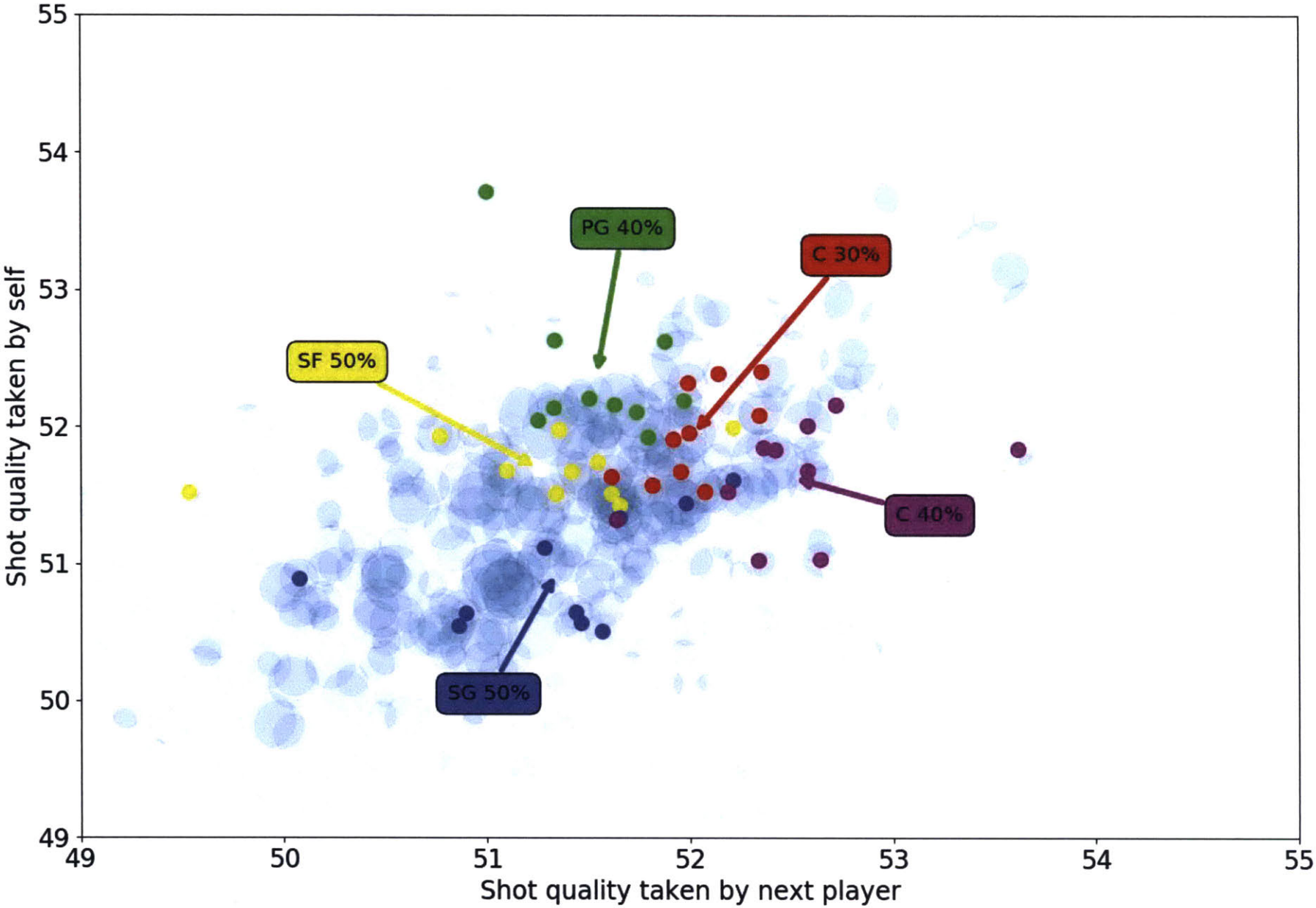

Figure 16: Cluster 2 from the shot-decision plots. This cluster shows teams whose players are near the middle, but some players focus on taking good shots while others focus on making good passes. This cluster consisted of: Memphis Grizzlies, Orlando Magic, Boston Celtics, Denver Nuggets, Cleveland Cavaliers, Oklahoma City Thunder, New York Knicks, Sacramento Kings, Atlanta Hawks, Minnesota Timberwolves. The average points-per-possession was 1.08.

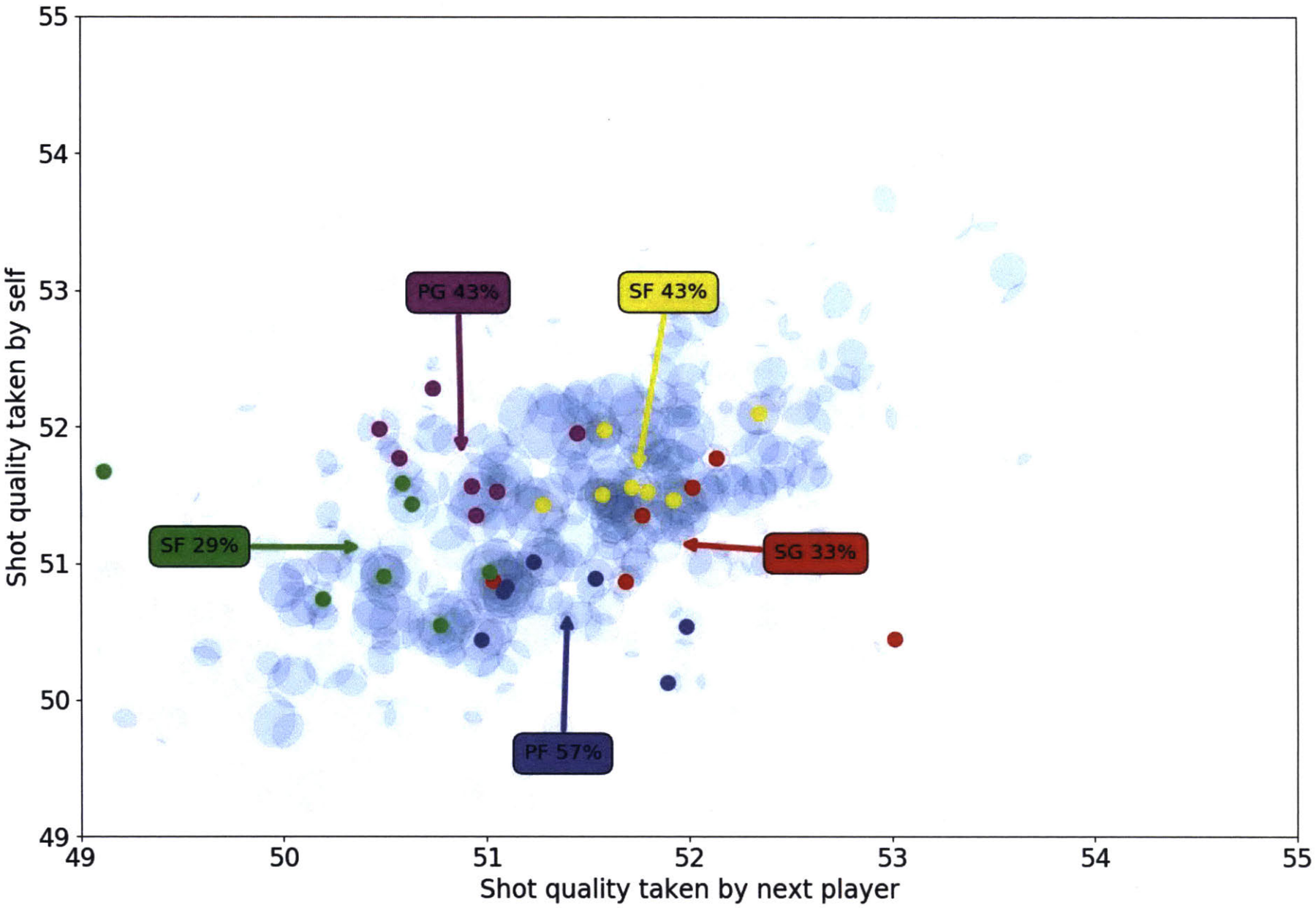

Figure 17: Cluster 3 from the shot-decision plots. This cluster shows teams whose players are near the middle, but some players are better decision makers than others. This cluster consisted of: Toronto Raptors, Washington Wizards, Los Angeles Clippers, Los Angeles Lakers, Phoenix Suns, Chicago Bulls, Indiana Pacers. The average points-per-possession was 1.12.

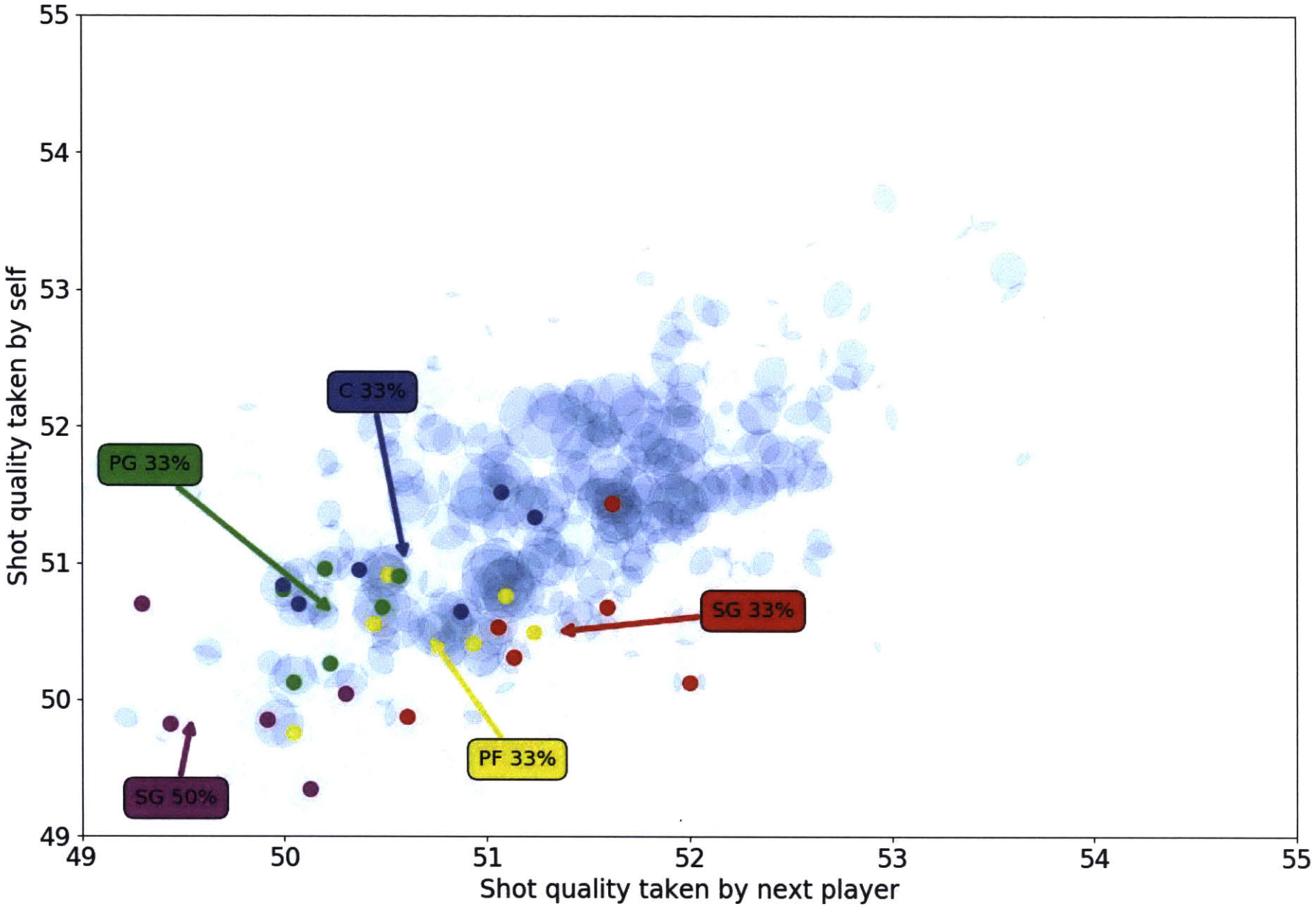

Figure 18: Cluster 4 from the shot-decision plots. This cluster shows teams whose players are in the bottom-left corner, indicating they are not discriminating well between when they should pass or not. This cluster consisted of: Detroit Pistons, Portland Trail Blazers, Dallas Mavericks, San Antonio Spurs, Utah Jazz, Miami Heat. The average points-per-possession was 1.13.

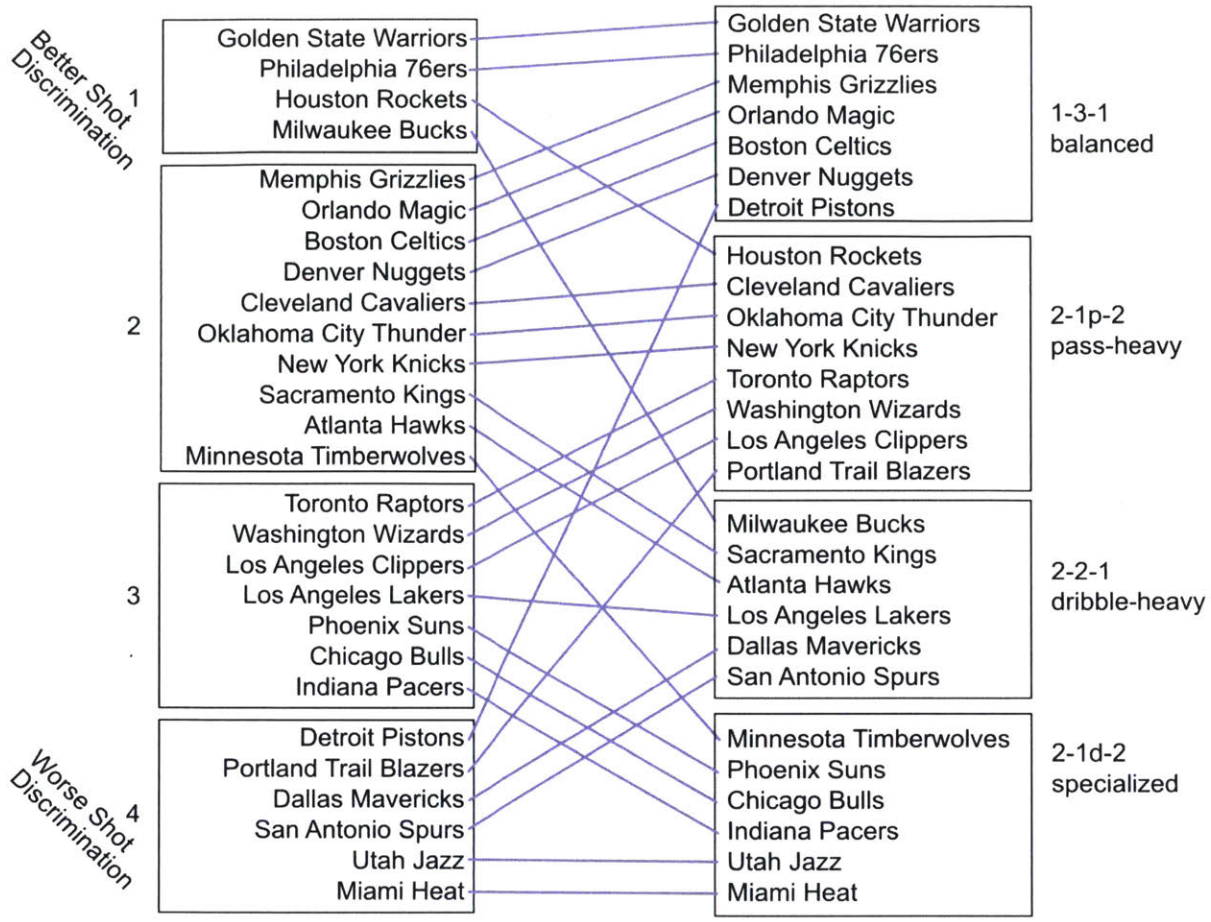

Figure 19: A comparison of the shot-decision and dribble/pass/shoot-propensity clusterings. The lists were manually arranged to maximize correlation between the two clusterings, i.e. minimizing the length of the blue lines between the clusterings.

Finally, we can compare the two clusterings. We can see that the teams with better shot discrimination tend to be those that are either 'balanced' or 'pass-heavy'. We can also see that the 'specialized' teams tend to have worse shot discrimination. 'Dribble-heavy' teams have a wide variable level of shot discrimination.

Conclusions

This analysis accumulated SecondSpectrum event data to generate behavioral player data, and classify NBA teams based on player composition. The analysis for both kinds of clustering that there are roughly four types of NBA teams, and a comparison of these clusterings shows that better shot discrimination is associated with teams that are 'balanced' or 'pass-heavy'.

References

Chang, Yu-Han, Rajiv Maheswaran, Jeff Su, Sheldon Kwok, Tal Levy, Adam Wexler, and Kevin Squire. "Quantifying Shot Quality in the NBA." MIT Sloan: Second Spectrum, Inc, 2014.

Fewell, Jennifer, Dieter Armbruster, John Ingraham, Alexander Petersen, and James Waters. "Basketball Teams as Strategic Networks." *PLOS One*, November 6, 2012. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0047445.

Kopf, Dan. "Data Analytics Have Made the NBA Unrecognizable." *Quartz*, October 18, 2017. https://qz.com/1104922/data-analytics-have-revolutionized-the-nba/.

McCann, Zach. "Player Tracking Transforming NBA Analytics." *Tech - ESPN Playbook*, May 19, 2012. http://www.espn.com/blog/playbook/tech/post/_/id/492/492.

Poler, Colin. "Classifying-NBA-17." Github Repository, May 11 2018. https://github.com/colinpoler/Classifying-NBA-17.