Who can replace Xavi? A passing motif analysis of football players.

Javier López Peña ∗ Raúl Sánchez Navarro †

Abstract

Traditionally, most of football statistical and media coverage has been focused almost exclusively on goals and (occasionally) shots. However, most of the duration of a football game is spent away from the boxes, passing the ball around. The way teams pass the ball around is the most characteristic measurement of what a team’s “unique style” is. In the present work we analyse passing sequences at the player level, using the different passing frequencies as a “digital fingerprint” of a player’s style. The resulting numbers provide an adequate feature set which can be used in order to construct a measure of similarity between players. Armed with such a similarity tool, one can try to answer the question: ‘Who might possibly replace Xavi at FC Barcelona?’

1 Introduction

Association football (simply referred to as football in the forthcoming) is arguably the most popular sport in the world. Traditionally, plenty of attention has been devoted to goals and their distribution as the main focus of football statistics. However, shots remain a rare occurrence in football games, to a much larger extent than in other team sports.

Long possessions and paucity of scoring opportunities are defining features of football games. Passes, on the other hand, are two orders of magnitude more frequent than goals, and therefore constitute a much more appropriate event to look at when trying to describe the elusive quality of ‘playing style’. Some studies on passing have been performed, either at the level of passing sequences distributions (cf [6,9,13]), by studying passing networks [3,4,8], or from a dynamic perspective studying game flow [2], or passing flow motifs at the team level [5], where passing flow motifs (developed following [10]) were satisfactorily proven by Gyarmati, Kwak and Rodríguez to set apart passing style from football teams from randomized networks.

In the present work we elaborate on [5] by extending the flow motif analysis to a player level. We start by breaking down all possible 3-passes motifs into all the different variations resulting from labelling a distinguished node in the motif, resulting on a total of 15 different 3-passes motifs at the player level (stemming from the 5 motifs for teams). For each player in our dataset, and each game they participate in, we compute the number of instances each pattern occurs. The resulting 15-dimensional distribution is used as a fingerprint for the player style, which characterizes what type of involvement the player has with his teammates.

The resulting feature vectors are then used in order to provide a notion of similarity between different football players, providing us with a quantifiable measure on how close the playing styles between any two arbitrary players are. This is done in two different ways, first by performing a Clustering Analysis (with automatic cluster detection) on the feature vectors, which allow us to identify 37 separate groups of similar players, and secondly by defining a distance function (based on the mean features z-scores) which consequently is used to construct the distance similarity score.

As an illustrative example, we perform a detailed analysis of all the defined quantities for Xavi Hernández, captain of FC Barcelona who just left the team after many years in which he has been considered the flagship of the famous tiki-taka style both for his club

---

∗Kickdex Ltd, and Department of Mathematics, University College London. javier@kickdex.com
†Kickdex Ltd, and Universidad de Granada. snchznvrr@gmail.com
and for the Spanish national team. Using our data-based style fingerprint we try to address the pressing question: which player could possibly replace the best passer in the world?

2 Methodology

The basis of our analysis is the study of passing sub-sequences. The passing style of a team is partially encoded, from an static point of view, in the passing network (cf. [8]). A more dynamical approach is taken in [5], where passing subsequences are classified (at the team level) through “flow motifs” of the passing network.

Inspired by the work on flow motifs for teams, we carry out a similar analysis at the player level. We focus on studying flow motifs corresponding to sequences of three consecutive passes. Passing motifs are not concerned with the names of the players involved on a sequence of passes, but rather on the structure of the sequence itself. From a team’s point of view, there are five possible variations: ABAB, ABAC, ABCA, ABCB, and ABCD (where each letter represents a different player within the sequence).

![Diagram of flow motifs](image)

Figure 1: The five team flow motifs

The situation is different when looking at flow motifs from an specific player’s point of view, as that player needs to be singled out within each passing sequence. Allowing for variation of a single player’s relative position within a passing sequence, the total number of motifs increases to fifteen. These patterns can all be obtained by swapping the position of player A with each of the other players (and relabelling if necessary) in each of the five motifs for teams. Adopting the convention that our singled-out player is always denoted by letter ‘A’, the resulting motifs can be labelled as follows (the basic team motif shown in bold letters):

| Motif | Flow Motifs |
|-------|-------------|
| ABAB  | BABA        |
| ABAC  | BABC, BCBA  |
| ABCA  | BACB, BCAB  |
| ABCB  | BACA, BCAC  |
| ABCD  | BACD, BCAD, BCDA |

When tracking passing sequences, we will consider only possessions consisting of uninterrupted consecutive events during which the ball is kept under control by the same team. As such, we will consider that a possession ends any time the game gets interrupted or an action does not have a clear passing target. In particular, we will consider that possessions get interrupted by fouls, by the ball getting out of play, whenever there is a “divided ball” (eg an aerial duel), by clearances, interceptions, passes towards an open space without a clear target, or by shots, regardless on who gets to keep the ball afterwards. The motivation for this choice is that we are trying to keep track of game style through controlled, conscious actions. It is worth noting that here we are using a different methodology from the one in [5] (where passes are considered to belong to the same sequence if they are separated by less than five seconds).

Our analysed data consists of all English Premier League games over the last five seasons (comprising a total of 1900 games and 1402195 passes), all Spanish Liga games over the last three season (1140 games and 792829 passes), and the last season of Champions League data (124 games and 105993 passes). To reduce the impact of outliers, we have limited our study to players that have participated in at least 19 games (half a season). In particular, this means that only players playing the English and Spanish leagues are tracked in our analysis. Unfortunately, at the time of writing we do not have at our disposal enough data about other European big leagues to make the study more comprehensive.

The resulting dataset contains a total of 1296 players. For each of the analyzed players, we compute the average number of occurrences of each of the fifteen passing motifs listed above, and use the results as the features vector in order to describe the player’s style. For some of the analysis which require making different types of subsequences com-
parable, we replace the feature vector by the corresponding z-scores (where for each feature mean and standard deviation are computed over all the players included in the study).

Our analysis uses raw data for game events provided by Opta. Data munging, model fitting, analysis, and chart plotting were performed using IPython [12] and the python scientific stack [7,11].

3 Analysis and results

Summary statistics and motifs distributions

A summary analysis of the passing motifs is shown in Table 1. Perhaps unsurprisingly, the maximum value for almost every single motif is reached by a player from FC Barcelona, the only exception being Yaya Touré. Figure 3 shows the frequency distributions for player values at every kind of motif, and the relative position of Xavi within those distributions.

| Motif | Mean | Std  | Max  | Player            |
|-------|------|------|------|-------------------|
| ABAB  | 0.33 | 0.31 | 3.56 | Dani Alves        |
| ABAC  | 1.52 | 1.30 | 8.71 | Thiago Alcántara |
| ABCA  | 0.90 | 0.73 | 5.99 | Xavi              |
| ABCB  | 1.53 | 1.08 | 7.69 | Sergio Busquets  |
| ABCD  | 6.03 | 3.62 | 25.53| Jordi Alba        |
| BABA  | 0.33 | 0.29 | 2.72 | Lionel Messi      |
| BABC  | 1.53 | 1.07 | 7.33 | Xavi              |
| BACA  | 1.51 | 1.28 | 8.94 | Thiago Alcántara |
| BACB  | 0.91 | 0.59 | 3.79 | Xavi              |
| BACD  | 6.01 | 4.17 | 27.21| Xavi              |
| BCAB  | 0.91 | 0.58 | 3.93 | Yaya Touré        |
| BCAC  | 1.52 | 1.08 | 6.83 | Jordi Alba        |
| BCAD  | 6.00 | 4.11 | 28.89| Xavi              |
| BCBA  | 1.53 | 1.03 | 8.29 | Sergio Busquets  |
| BCDA  | 6.01 | 3.47 | 23.64| Dani Alves        |

Table 1: Motif average values and players with highest values

We can see how Xavi dominates the passing, being the player featuring the highest numbers in five out of the fifteen motifs. Table 2 shows all the values and z-scores for Xavi. It is indeed remarkable that he manages to be consistently over four standard deviations away from the average passing patterns, and particularly striking his astonishing z-score of 6.95 in the ABCA motif, which corresponds to being the starting and finishing node of a triangulation. To put this number in context, if we were talking about random daily events, one would expect to observe such a strong deviation from the average approximately once every billion years.

| Motif | Value | z-score |
|-------|-------|---------|
| ABAB  | 1.57  | 3.97    |
| ABAC  | 8.67  | 5.49    |
| ABCA  | 5.99  | 6.95    |
| ABCB  | 7.12  | 5.19    |
| ABCD  | 21.44 | 4.26    |
| BABA  | 1.71  | 4.71    |
| BABC  | 7.33  | 5.41    |
| BACA  | 8.58  | 5.51    |
| BACB  | 3.79  | 4.88    |
| BACD  | 27.21 | 5.08    |
| BCAB  | 3.27  | 4.06    |
| BCAC  | 6.78  | 4.86    |
| BCAD  | 28.89 | 5.57    |
| BCBA  | 7.08  | 5.40    |
| BCDA  | 23.03 | 4.90    |

Table 2: Motif values and z-scores for Xavi

Clustering and PCA

Using the passing motifs means as feature vectors, we performed some clustering analysis on our player set. The Affinity Propagation method with a damping coefficient of 0.9 yields a total of 37 clusters with varying number of players, listed in Table 7, where a representative player for every cluster is also listed. The explicit composition of each of the clusters of size smaller than 10 is shown in Table 8. Once again

1Touré did play for FC Barcelona, however, our dataset only contains games in which he played for Manchester City. On the opposite side, we only have data for Thiago Alcántara as a Barcelona player as our dataset does not include the German Bundesliga.

2From a very rigorous point of view, actual passing patterns are neither random nor normally distributed. Statistical technicalities notwithstanding, Xavi’s z-scores are truly off the charts!
we can observe how the passing style of Xavi is different enough from everyone else’s to the extent that he gets assigned to a cluster of his own!

Figure 2 shows the relative players feature vectors, plotted using the first two components of a Principal Component Analysis (after using a whitening transformation to eliminate correlation). The PC’s coefficients, together with their explained variance ratio, are listed in Table 3. After looking at Figure 2, one can think of the first principal component (PC 1) as a measurement of overall involvement on the game, whereas the second principal component (PC 2) separates players depending on their positional involvement, with high positive values highlight players playing on the wings and with a strong attacking involvement, and smaller values relate to a more purely defensive involvement. Special mention on this respect goes to Dani Alves and Jordi Alba, who in spite of playing as fullbacks display a passing distribution more similar to the ones of forwards than to other fullbacks. The plot also shows how Xavi has the highest value for overall involvement and a balanced involvement between offensive and defensive passing patterns.

Table 3: Principal Components and their explained variance

| Player distance and similarity |
|-------------------------------|
| The feature vector can be used in order to define a measure of similarity between players. Given a player $i$, let $v_i$ denote the vector of z-scores in passing motifs for player $i$. Our definition of distance between two players $i$ and $j$ is simply the Euclidean distance between the corresponding (z-scores) feature vectors:

$$d(i, j) := \|v_i - v_j\|_2 = \sqrt{\sum_{m \in \text{motifs}} (v_{i,m} - v_{j,m})^2}$$

This distance can be used as a measure of similarity between players, allowing us to establish how closely related are the passing patterns of any two given players. In more concrete terms, the coefficient of similarity is defined by

$$s(i, j) := \frac{1}{1 + d(i, j)}.$$ 

This similarity score is always between 0 and 1, with 1 meaning that two players display an identical passing pattern.

The reason for choosing z-scores rather than raw values is to allow for a better comparison between
different passing motifs, as using raw values would yield a distance dominated by the four motifs derived from ABCD, which show up in a frequency one order of magnitude higher than any other pattern. Table 4 shows a summary of the average and minimum distances for all the players in our dataset, showing that for an average player we can reasonably expect to find another one at a distance of 0.826 ± 0.5.

|          | Mean  | Closest |
|----------|-------|---------|
| Avg value| 4.471 | 0.826   |
| Std deviation | 1.800 | 0.500   |
| Min value  | 3.188 | 0.178   |
| Max value  | 19.960| 5.134   |

Table 4: Average and closest player distances.

An immediate application of this is to find out, for a given player, who is his closest peer, which will be the player displaying the most similar passing pattern. Table 5 shows the minimum distances to the ten bottom players (the ones with the smallest minimum distance, hence easier to replace) and the top 10 players (the ones with the highest minimum distance, thus harder to replace). Once again, we can see how the top 10 players are dominated by FC Barcelona players.

![Table 5: Players minimum distance (bottom 10 and top 10)](image)

Note that in some cases, the closest peer for a player happens to play for the same team, as it is the case for Jordi Alba, whose closest peer is Dani Alves. We decided against filtering closest player to search in team as it would make the analysis overly complicated due to constant player movement between teams.

Previous table shows that Xavi is amongst the hardest players to find a close replacement for. Table 6 shows the 20 players closest to Xavi. Among those, no one has a similarity score higher that 18.2%, and only ten players have a score higher than 10%.

![Table 6: Distances and similarity scores of the 20 players closest to Xavi](image)

4 Conclusions and future work

We have shown how the flow motif analysis can be extended from teams to players. Although there is an added level of complexity raising from the increasing of the different motives, the resulting data does a good job classifying and discriminating players. Clustering analysis provides a reasonable grouping of players with similar characteristics, and the similarity score provides a quantifiable measure on how
similar any two players are. We believe these tools can be useful for scouting and for early talent detection if implemented properly.

For future work, we plan to expand our dataset to cover all the major European leagues over a longer time span. A larger dataset would allow us to measure changes in style over a player’s career, and perhaps to isolate a team factor that would allow to estimate what would be a player’s style if he were to switch teams. Another interesting thing to explore would be the density of each of the passing motifs according to pitch coordinates.

Coming back to our motivating question, who can replace Xavi at Barcelona? Amongst all the ten players that showing a similarity score bigger than 10, three are already at Barcelona (Busquets, Iniesta and Jordi Alba), and another three used to play there but left (Touré, Alcántara and Fàbregas). Arteta, Carrick and Cazorla are all in their thirties, ruling them out as a long-term replacement, and Toni Kroos plays for Barcelona arch-rivals Real Madrid, making a move quite complicated (although not impossible, as current Barcelona manager Luis Enrique knows very well), the only choices for Barcelona seem to be either to recover Alcántara or Fàbregas, or to reconvert Iniesta to play further away from the opposition box. A bolder move would be the Dutch rising star, Daley Blind (who used to play as a fullback, but has been tested as a midfielder over the last season in Van Gaal’s Manchester United), hoping that the young could rise to the challenge.

Xavi’s passing pattern stands out in every single metric we have used for our analysis. Isolated in his own cluster, and very far away from any other player, all data seems to point out at the fact that Xavi Hernández is, literally, one of a kind.

| Representative Player | Cluster size |
|-----------------------|--------------|
| Xavi                  | 1            |
| Dani Alves            | 2            |
| Thiago Alcántara      | 4            |
| David Silva           | 4            |
| Gerard Piqué          | 5            |
| Bacary Sagna          | 6            |
| Isco                  | 8            |
| Chico                 | 10           |
| Mahamadou Diarra      | 12           |
| Jonny Evans           | 12           |
| Jordan Henderson      | 15           |
| Andreu Fontás         | 17           |
| Christian Eriksen     | 17           |
| Hugo Mallo            | 18           |
| Victor Wanyama        | 19           |
| César Azpilicueta     | 19           |
| Alberto Moreno        | 20           |
| Gareth Bale           | 20           |
| Fran Rico             | 30           |
| David de Gea          | 36           |
| Antolin Alcaraz       | 36           |
| Phil Jagielka         | 39           |
| Sebastian Larsson     | 40           |
| Liam Ridgewell        | 41           |
| Emmerson Boyce        | 44           |
| Nyom                  | 46           |
| John Ruddy            | 48           |
| Adam Johnson          | 52           |
| Richmond Boakye       | 57           |
| Chechu Dorado         | 61           |
| Manuel Iturra         | 62           |
| Loukas Vyntra         | 62           |
| Kevin Gameiro         | 72           |
| Borja                 | 73           |
| Rubén García          | 85           |
| Gabriel Agbonlahor    | 90           |
| Steven Fletcher       | 113          |

Table 7: Affinity propagation cluster sizes and representative players.
| Size | Players |
|------|---------|
| 1    | Xavi    |
| 2    | Dani Alves, Jordi Alba |
| 4    | David Silva, Lionel Messi, Samir Nasri, Santiago Cazorla |
| 4    | Andrés Iniesta, Cesc Fàbregas, Thiago Alcántara, Yaya Touré |
| 5    | Daley Blind, Gerard Piqué, Javier Mascherano, Sergio Busquets, Toni Kroos |
| 6    | Adriano, Angel Rangel, Bacary Sagna, Gaël Clichy, Marcelo, Martín Montoya |
| 8    | Emre Can, Isco, James Rodríguez, Juan Mata, Maicon, Mesut Özil, Michael Ballack, Ryan Mason |
| 10   | Ashley Williams, Carles Puyol, Chico, Marc Bartra, Marcos Rojo, Michael Carrick, Mikel Arteta, Nemanja Matic, Paul Scholes, Sergio Ramos |
| 12   | Dejan Lovren, Garry Monk, John Terry, Jonny Evans, Ki Sung-yueng, Matija Nastasic, Michael Essien, Morgan Schneiderlin, Nabil Bentaleb, Per Mertesacker, Roberto Trashorras, Vincent Kompany |
| 12   | Aaron Ramsey, Alexandre Song, Fernandinho, Gareth Barry, Jerome Boateng, Jonathan de Guzmán, Leon Britton, Luka Modric, Mahamadou Diarra, Mamadou Sakho, Steven Gerrard, Xabi Alonso |
| 15   | Ander Herrera, Eric Dier, Frank Lampard, Ivan Rakitić, Jamie O’Hara, Jordan Henderson, Michael Krohn-Dehli, Rafael van der Vaart, Rafinha, Sascha Riether, Scott Parker, Seydou Keïta, Steven Davis, Vassilíki Abou Diaby, Wayne Rooney |

Table 8: Affinity propagation clustering: Composition of small clusters
Figure 3: Passign motifs distributions
References

[1] C. Anderson and D. Sally The Numbers Game: Why everything you know about football is wrong. Penguin UK, 2013

[2] D.R. Brillinger A potential function approach to the flow of play in soccer, Journal of Quantitative Analysis in Sports, 3 (2007), DOI: jqas.2007.3.1.1048

[3] Carlos Cotta, Antonio M. Mora, Cecilia Merelo-Molina, and Juan Julián Merelo. FIFA World Cup 2010: A Network Analysis of the Champion Team Play. Complex Systems in Sports Workshop (CS-Sports 2011), August 2011.

[4] J. Duch, J. S. Waitzman, and L. A. N. Amaral. Quantifying the performance of individual players in a team activity. PloS One, 5(6):e10937, 2010.

[5] L. Gyarmati, H. Kwak and P. Rodríguez Searching for a Unique Style in Soccer. http://arxiv.org/abs/1409.0308.

[6] M. Hughes and I. Franks Analysis of passing sequences, shots and goals in soccer Journal of Sports Sciences, 23 (2005) 509–514

[7] J.D. Hunter Matplotlib: A 2D Graphics Environment Computing in Science & Engineering 9, 90 (2007)

[8] J. López Peña and H. Touchette A network theory analysis of football strategies. In Sports Physics. École Polytechnique Univ. Press, 519–530.

[9] J. López Peña A Markov model for association football possession and its outcomes. http://arxiv.org/abs/1403.7993.

[10] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon Network Motifs: Simple building blocks of complex networks Science, 298 (5594) 824–827, 2002

[11] T.E. Oliphant Python for Scientific Computing Computing in Science & Engineering 9, 90 (2007)

[12] F. Pérez and B. E. Granger IPython: A System for Interactive Scientific Computing Computing in Science and Engineering, 9 (2007) 21–29 DOI: 10.1109/MCSE.2007.53.

[13] C. Reep and B. Benjamin Skill and chance in Association Football J. of the Royal Stat. Soc. A, 131 (1968) 581–585.