PassFlow: Soccer Passing Route Visualizer

Team 096

Enoch Anim-Koranteng, Serdar Aydinoglu, Jonathan Eaton,
Wayne Fong, Vigneshwar Perumal, and Jackson Schieber

Introduction

There are few activities that garner more attention from the world than sports. The most popular and influential of them all is soccer (football). For example, the sport is so influential that it makes up 1.4% of Spain's GDP (Fernández, Menchu 2018). In order to justify that valuation, teams have huge pressure and incentive to refine every aspect of their performance. The most prevalent and important action in the sport is passing. Therefore, this paper discusses an initiative to create an app which allows a user to easily, interactively, and effectively visualize goal-to-goal passing routes for the entire field with weighted probabilities. The application highlights the most probable route calculated from historical team passing trends using Markov chain analysis. Furthermore, the user is enabled to 'block' a chosen route causing the app to recalculate a new expected pass sequence. Nodes (players) are located by average historical player position. The opponent's goal serves as a 12th node to represent when a player is more likely to take a shot than pass the ball again. The app could later be adjusted to add a number of additional features; an example would be recommending passing routes to maximize expected goals. The app should be productive to coaches, training staff, and soccer analysts because of its strong visualization and interactive and dynamic functionality.

Problem Statement

Presently, much of the soccer community is stuck on overly simplistic analytics like number of assists or pass completion percentage; the more complex analytics content is often abstract and bereft of visualization or application. Therefore, there is a need to produce a sophisticated analysis that is easily comprehensible and functional so as to effectively impact coaching and training preparation.

Literature Survey

As mentioned, passing is one of the most critical elements of any soccer (football) match, and researchers have found a wide variety of ways to quantify pass quality and effectiveness (Xiaoguang et al., 2001; Kite & Nevill, 2017). Goes et al. (2019) used field position, pass length, pass angle, and pass velocity to predict how much a given pass will disrupt the defense, while Bravo et al. (2021) focused on a more common measure of pass effectiveness—expected goals (EG). They used k-means to cluster similar types of passes and track the EG value for each pass. Bransen et al. (2019) took a similar approach by measuring passing statistics of possession sequences using K-means. Value of the passes was calculated by subtracting the expected goal value at the beginning from the end of the sequence. These articles inform the team how its tool ought to rate and advise passing routes, but lack strong focus on predicting historical/probable passing routes. James et al. (2013) provide even deeper insight into the use of unsupervised clustering algorithms such as K-Means, K-nearest neighbor and principal component clustering. Anzer and Bauer (2022) used machine learning to assess the difficulty of a pass, differentiating an effective and difficult pass from common and general passes. Their model leverages gradient boosting algorithms to derive success probability of a pass based on the circumstances of the player. Chawla et al. (2017) used a different method to classify pass quality. They compared three classification algorithms including multinomial logistic regression, support vector machine, and RUSBoost. In addition, they calculated the dynamic area of influences of each player called the dominant region. It was established that the isomorphism of passing networks and social media networks allows better quantification of passing (Zhipei & Bingyu, 2021). Current methods include passing evaluation systems that can quantify the passing ability of players based on metrics such as degree centrality and

closeness centrality. These papers serve as inspiration for more exotic ways to apply statistical methods to analyze player passing routes and to predict success; however there is still lack of focus on analyzing unique team passing tendencies, which is our emphasis. We also expanded by adding interactive visualization.

While traditional statistical methods such as clustering and regression have helped analyze game outcomes through passing statistics, the advent of tracking data has helped teams track more detailed metrics such as how fast a player arrived at the ball or into position and where players tended to gravitate (Memmert et al., 2016). Due to the variation in teams performance across seasons, Markov models and dynamic systems theory are increasingly being used to understand individual and team dynamics in sports settings (Shafizadeh et al., 2013; Van Roy et al. 2021; Bukiet et al. 1997). Similarly, network sciences are helping create passing networks that can be analyzed at different scales, such as space, time and dynamics (Buldú et al., 2018). These approaches offer excellent examples of analysis of spatial data and behavioral habits unique to a team, and they showcase the use of Markov models to determine likely outcomes; however, they do not directly apply these methods to modeling passing effectiveness. We have furthered their research by applying similar methods to model passing trends.

Another area of research is studying pass routes and patterns. Malqui et al. (2019) created a clustering algorithm to group sequences of four passes into eight major passing patterns. Kawasaki et al. (2019) and Trequattrini et al. (2015) used similar clustering methods to build passing networks. By analyzing the density and distance of nodes, they determined the best pass positions, directions, and most efficient pass pairings. Gabriel Anzer et al.(2022) used semi-supervised graph neural networks to detect tactical patterns of soccer players on the field to yield an impact on the game. Using machine learning and visualization techniques they identified patterns that lead to goal scoring and possession. These papers offer well rounded information on formations and passing effectiveness. They also offer inspiration for creating effective visualizations; however, they focus on clustering algorithms while we focus on different methods that will be more specific to each team's passing tendencies.

Ultimately every paper mentioned so far falls short of the group's objectives because they do not allow the user to easily, interactively, and effectively visualize goal-to-goal passing routes for the entire field with weighted probabilities, which is what the group hopes to have achieved.

Proposed Method

Data Preparation

Our application relied on StatsBomb's open source soccer data. The statsbombpy package in Python enabled us to pull the data for a wide range of soccer tournaments and seasons. We were primarily using the events data, which showed game details like passes and shots, along with the involved player's location at the time. To start, we looked at Spain's La Liga 2017-2018 season; however, we expanded that to include 16 additional La Liga seasons. We originally intended to expand to all 43 competitions included in the StatsBomb open data, but we ran into RAM limitations and chose to focus on all available La Liga seasons.

During a given season, teams typically have 20-30 players with some amount of playing time. To simplify our app and make it more accessible, we analyzed only the most common starting 11 players for the season. We then calculated passes made and received for each of those players and the average location of each player. This essentially gave us a graph with the in/out-degree for each of the top 11 players. That data then gets fed into our model for further processing.

Modeling and Analysis

The application's main objective was to determine the most likely passing route from the goalkeeper to the opponent's goal without encountering any interceptions. Initially, we considered the 11 players on the field as nodes and the opponent's goal as the 12th node. If the goalkeeper is node 1,

then the starting and ending nodes are 1 and 12 respectively, and our application will compute the most probable path between these two nodes. To apply the Markov chain model, we constructed a state transition matrix that is a 12 by 12 probability matrix containing the likelihood of one player passing the ball to another. Each row of the matrix adds up to 1, representing the probabilities of a player passing to any other player on the field.

Since our aim was to identify the passing sequence that was not intercepted, we filtered for successful passes from the event data. Additionally, to incorporate the opponent's goal into the passing sequence, we added shot data to the filtered event data. The resulting smaller dataset comprised the origin and destination nodes and locations, as well as the number of successful passes from the origin to the destination. This dataset formed the primary input for our model and was used to create the state transition matrix.

Another input was the initial distribution vector, which typically contained the probabilities of each node initiating the sequence. However, in our case, we were using the goalkeeper as the fixed initial node, resulting in a vector of length 12, with all probabilities equal to zero except for the goalkeeper's probability of starting the sequence, which was 1.

Once we had the state transition matrix and initial probability vector, we used the power iteration method to determine the state transition probabilities for each step of the passing sequence. Each iteration step identified the next node with the highest probability of having the ball, which became the initial node for the subsequent step.

We were also interested in assessing the significance of each player involved in the sequence. To accomplish this, we examined each player on the route and calculated their total probability of receiving the ball (in degree). Next, we adjusted the radius of the nodes to correspond with the total probability associated with each player.

To prevent the occurrence of infinite loops within the model, a preference was given to progressive passes. To illustrate this issue, suppose that the model identifies player A as the most probable recipient of a pass from player B, and player B as the most probable recipient of a pass from player A. In such a scenario, if either player A or player B were to receive the ball, the model would enable these two players to repeatedly exchange passes with each other. To avoid this situation, the model prioritized passes that advance the ball towards the goal. If either the pass from A to B or from B to A failed to meet this criterion, the model selected the next highest probability pass option.

Another feature of the model was to enable users to block a player, which served the purpose of identifying the next most probable scenario in the event that a player on the most likely path is blocked. To achieve this, when a user selects a node on a given path, the corresponding node ID is transmitted to the model. Subsequently, the model set all row and column values that correspond to the selected node to zero, indicating that the associated player cannot receive the ball. Next, the model normalized the rows of the state transition matrix and employed the same iterative method used in determining the most probable route to generate the next most possible path.

The model required two CSV files that are generated from the main event data of StatsBomb. These CSV files encompass data for all 17 seasons that are currently available. Once these files were input into the model, they were filtered based on the selection made by the user through the dropdown menu. The model subsequently executed and computed the optimal route, and saved three CSV files in the application data folder, which in turn are used to feed d3 and update the visual output.

Visualization

We initially experimented in R using the r2d3 package to create a football field. The r2d3 package has the ability to parse d3 scripts directly in R through RShiny. But to realize the full potential of our project and to increase our visualization capabilities we shifted to using Python, HTML5 and D3.js. The Python and Javascript code were incorporated together in Flask. For our initial testing, we had our football framework and static plots of players along with passes they have made to players for one specific game. The test game included 14 players (starting 11 with 3 substitutions), which we used to make sure that our visualization is populating the players and their passing routes.

In subsequent iterations of the app, we focused only on the most common starting 11 players over the course of an entire season. The app highlights the nodes and edges involved in the most probable route to goal, and the nodes are sized based on degree. The width of the graph edge corresponds to the pass completion percentage. The app also displays a dropdown menu allowing the user to choose between the 17 La Liga seasons. Hovering over a node will change the node color and display the player's name (see Exhibit 2).

Users can also block a player, eliminating them from calculation of the most probable route. The app will then recalculate the next most probable route excluding the chosen player. This gives the user more versatility in how they interact with the data, and it enables coaches and staff to formulate defensive plans. As shown in Exhibits 2 and 3, when the user selects center defender Javier Mascherano, he is removed from the path, and the new most probable passing route follows the right wing of the pitch.

Exhibit 1: Most probable passing route for FC Barcelona, 2014/2015

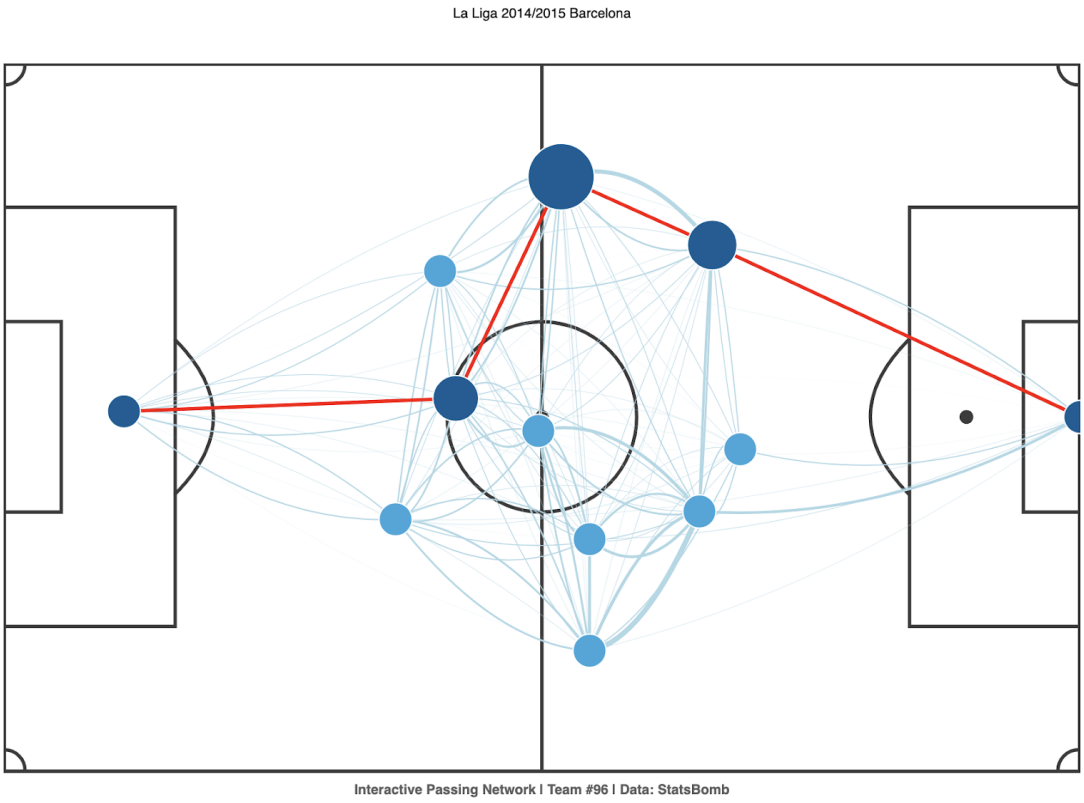


Exhibit 2: Node selection in PassFlow

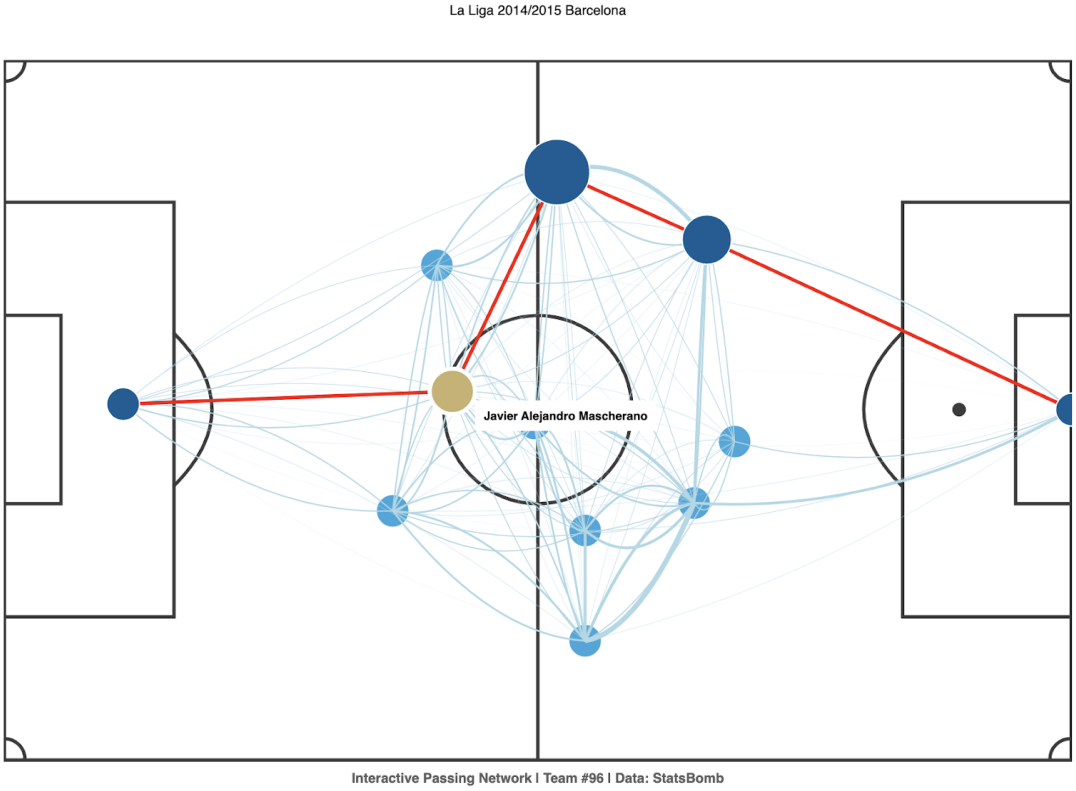
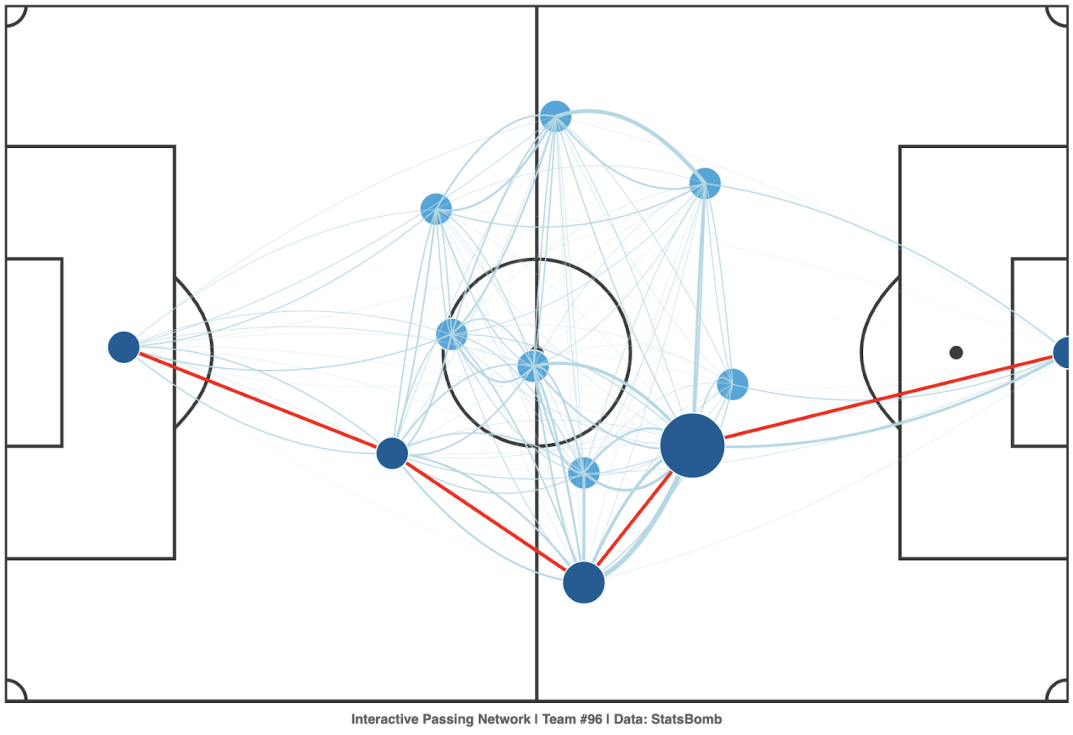


Exhibit 3: Alternate route after blocking a node



Experiments and Evaluation

PassFlow was designed and tested on standard quad-core PCs. The backend code to extract and manipulate the data was created in Google Colab, and there are no special computing requirements to install and run our code.

The app is designed to show the user the most probable passing route to the opponent's goal, enabling the user to see strengths and weaknesses in a team's passing schema. As such, there are no standard measurements of accuracy that apply (e.g., R-squared, MAPE, etc.). However, robust testing showed that our Markov model was correct in the displayed pass probabilities given the initial state.

Usability surveys on our application were done to determine ease of use and interactivity. Friends were invited to test our interface and given a set of survey questions. Questions included the following: On a scale of 1 to 10, how simple was it to use the application? How easy is it to understand what is displayed? Is the information provided clear?

After the first round of surveys, the app was well-received. The visualization was intuitive and accessible for those with a basic understanding of soccer. Respondents said the visualization was high-quality, and they were intrigued to see which players were key to the most probable passing route. However, there was some confusion regarding the alternative route function. Some did not know they could click on the player to block and were focused solely on the dropdown menu. Therefore, we added some instructions to clear up options for the user.

Several respondents gave additional feedback that was helpful but not possible at this time due to time and data constraints. However, future iterations of the app could incorporate some of the following suggestions. One suggestion was to expand the possible data, including data from women's leagues. Another suggestion was to use conditional probabilities to allow the user to see how the team performed against different formations, thus allowing coaches to see what defensive formations are most disruptive. One respondent said they would like to see additional data for each player, such as foot dominance, position, age, etc. They also stated that more control over the lineup (i.e., substituting players instead of only seeing the most common starting 11) could be beneficial. We had considered this previously, but chose not to implement it in order to streamline the user interface.

While the example routes shown in Exhibits 1-3 are straightforward, some routes may involve more backtracking, making it difficult to follow the edge direction. One user suggested adding arrows or another indicator to represent the direction of the pass. They also requested a dropdown or radio buttons to allow the user to see the next most probable routes for deeper analysis. A further suggestion was to show the breakdown of pass types, such as ground pass versus aerial pass, when selecting an edge.

Conclusions and Discussion

The app enables users to interactively and effectively visualize goal-to-goal passing routes based on weighted probabilities and a Markov chain algorithm approach. From the reviewed literature, this approach is unique, as many existing approaches primarily use clustering algorithms or focus on other measures like expected goals and assists. Also, the added functionality of allowing the user to exclude a player from a passing route is a novel interactive feature.

A limitation of our work was our app only includes 17 FC Barcelona seasons. We were unable to gain access to the full StatsBomb API which includes data across 90 global leagues. Access to more data would allow us to customize our app for specific customers as well as provide deeper insights across teams and leagues. The app would allow coaches, training staff, and soccer analysts to perform in-depth analysis on probable passing routes and spatial locations specific to an opposing team based on their historical data rather than synthetic or simulated data. Hence, the app will customize the visualization and analysis to a specific opposing team to potentially identify a winning formation against it.

All team members contributed a similar amount of effort.

Works Cited

Anzer, G., & Bauer, P. (2022). Expected passes: Determining the difficulty of a pass in football (soccer) using spatio-temporal data. Data Mining and Knowledge Discovery, 36(1), 295-317.

Anzer, G., Bauer, P., Brefeld, U., & Fassmeyer, D. (2022). Detection of tactical patterns using semi-supervised graph neural networks. In MIT Sloan Sports Analytics Conference (Vol. 16, pp. 1-3).

Bransen, L., Van Haaren, J., & van de Velden, M. (2019). Measuring soccer players' contributions to chance creation by valuing their passes. Journal of Quantitative Analysis in Sports, 15(2), 97-116. https://doi.org/10.1515/jqas-2018-0020

Bravo, A., Karba, T., McWhirter, S., & Nayden, B. (2021). Analysis of individual player performances and their effect on winning in college soccer. SMU Data Science Review, 5(1), 114-128. Available at: https://scholar.smu.edu/datasciencereview/vol5/iss1/8.

Bukiet, B., Harold, E. R., & Palacios, J. L. (1997). A Markov chain approach to baseball. Operations Research, 45(1), 14-23.

Buldú, J. M., Busquets, J., Martínez, J. H., Herrera-Diestra, J. L., Echegoyen, I., Galeano, J., & Luque, J. (2018). Using Network Science to Analyse Football Passing Networks: Dynamics, Space, Time, and the Multilayer Nature of the Game. Frontiers in Psychology, 9, 1900. doi: 10.3389/fpsyg.2018.01900

Chawla, S., Estephan, J., Gudmundsson, J., & Horton, M. (2017). Classification of passes in football matches using spatiotemporal data. ACM Transactions on Spatial Algorithms and Systems (TSAS), 3(2), 1-30.

Fernández, Menchu. "'Football Could Be the 17th Largest Global Economy.'" *Diario AS*, AS, 17 June 2018, https://en.as.com/en/2018/06/17/soccer/1529259985_901247.html.

Goes, F. R., Kempe, M., Meerhoff, L. A., & Lemmink, K. A. (2019). Not every pass can be an assist: a data-driven model to measure pass effectiveness in professional soccer matches. Big data, 7(1), 57-70.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Kawasaki, T., Sakaue, K., Matsubara, R., & Ishizaki, S. (2019). Football pass network based on the measurement of player position by using network theory and clustering. International Journal of Performance Analysis in Sport, 19(3), 381-392.

Kite, C. S., & Nevill, A. (2017). The Predictors and Determinants of Inter-Seasonal Success in a Professional Soccer Team. Journal of human kinetics, 58, 157–167.

Malqui, J. L. S., Romero, N. M. L., Garcia, R., Alemdar, H., & Comba, J. L. (2019). How do soccer teams coordinate consecutive passes? A visual analytics system for analyzing the complexity of passing sequences using soccer flow motifs. Computers & Graphics, 84, 122-133.

Memmert, D., Lemmink, K. A. P. M., Sampaio, J., & Williams, A. M. (2017). Current approaches to tactical performance analyses in soccer using position data. Sports medicine, 47(1), 1-10. https://doi.org/10.1007/s40279-016-0562-5.

Qu Xiaoguang, Sun Hao, & Chen Jian. (2001). Research on the Technical and Tactical Characteristics of Chinese, American and Norwegian Women's Football Teams. China Sports Technology, 37(8), 26-29.

Shafizadeh, M., Sproule, J., & Gray, S. (2013). The emergence of coordinative structures during offensive movement for goal-scoring in soccer. International Journal of Performance Analysis in Sport, 13. https://doi.org/10.1080/24748668.2013.11868675

Trequattrini, R., Lombardi, R., & Battista, M. (2015). Network analysis and football team performance: a first application. Team Performance Management.

Van Roy, M., Robberechts, P., Yang, W. C., De Raedt, L., & Davis, J. (2021). Learning a Markov Model for Evaluating Soccer Decision Making. In Reinforcement Learning for Real Life (RL4RealLife) Workshop at ICML 2021.

Zhipei, Z. & Bingyu, P. (2021). Evaluating Player's Passing Ability based on Passing Network. School of Sports Engineering, Beijing Sport University.