Motif Detection and Player Influence Within Football Matches

Mitchell vom Scheidt, Willem Strydom

Washington University in St. Louis, St. Louis, 63130

December 8, 2023

Abstract

Between 2016 and 2018 Real Madrid Club de Fútbol played countless matches. A quick search online can tell you the outcomes of those matches, but what about how they played those matches caused their continuation of being known as one of the most successful football clubs of all time? Given the second-by-second event data of 5 matches Real Madrid played during this period, we created a directed, weighted network that models the passes from one player in the starting lineup, to another. Using open-source data from statsbomb.com [1] [2] which tracks event data from countless football matches, we studied the connections between players, analyzing the network to better understand the strategy of Real Madrid, and how players work together in different patterns to win matches.

Our primary goal is to perform analysis to gain insights into whether or not the results of this network analysis would have beneficial applications to football clubs. Performing this analysis on your own team could give insights into how successfully a club's strategy is being implemented, and which players have the most impact. On the other hand, performing this analysis on your opponents in preparation for a match could be beneficial. A club could use it to help determine the best strategy to use during a game to help increase their chances of victory.

The way in which we have performed this analysis is by implementing various algorithms to determine if there are recurring patterns in the way multiple players pass between one another as well as investigating the role each player has on an individual level. These algorithms will include motif detection, to find recurring sub-graph patterns, as well as RolX analysis, to determine the role each player has and if multiple players have similar roles to each other.

GitHub Repository: https://github.com/cse416a-fl23/fp-mitch_willem

***Keywords*— Motif Detection**, **RolX Analysis**, **Complex Network Construction**

1 Introduction

There are countless ways a football team can configure their formation and strategy to help them gain an advantage over their opponents. How does that formation and strategy change when a team has possession, or are winning vs. losing? Everyone that follows football knows about these concepts, but how can we determine how these planned strategies come to fruition during the game? What effect do they have on how a team operates? Is there a specific player, or group of players that has more influence than another? These are some of the questions we have aimed to answer. Football is a meticulous game. There are many famous tactics and strategies in the game such as total football and tiki-taka. So, is there a way to analyze a football clubs match data to find out if they are following a particular strategy?

Using network science, we wanted to find out if there is a way to quantify these qualitative elements of football. Based on the discussion and questions above, this paper has two objectives. To determine if there are recurring sub-graph patterns, called motifs, across multiple matches for the same team, and if particular players on the same team share similar roles. The first will give us an interpretation of how players within a team collaborate, and the second will give show us on a lower level, how individual players impact the team as a unit.

2 Methodology

2.1 Data

We acquired our data from a python library created by StatsBomb [2], a database which provides data from throusands of football matches across all competitions. This data includes all events from particular matches including passes, shots, goals, substitutions, fouls, etc. all with timestamps. For this paper, we focused on the pass data from 5 particular matches. We looked at 5 matches Real Madrid played between 2016 and 2018. Specifically, these were:

- Real-Madrid vs Liverpool (Champions League final 2018-05-26)

  Match ID: 18245

- Real-Madrid vs Barcelona (La Liga 2018-05-06)

  Match ID: 9924

- Real-Madrid vs Barcelona (La Liga 2017-12-23)

  Match ID: 9736

- Real-Madrid vs Barcelona (La Liga 2017-04-23)

  Match ID: 276569

- Real-Madrid vs Barcelona (La Liga 2016-12-03)

  Match ID: 267076

For each match played, we constructed a complex network and performed our analysis. There were various considerations when constructing the networks. Firstly, each player in the starting XI is a node. We are uninterested in data from substitutes who may not have had a significant impact on the match. Secondly, we considered what constituted an edge and an increment to an edge weight. Our decision was to construct the network in such a way that for every 2 passes from one player to another a directed edge would be added between them with weight 1. If an edge already exists between two players, then we incremented the edge weight by 1. This way, we eliminated "odd-ball" passes between players such as an unintentional pass, or a "fluke", so to speak. This aided us in the latter stages of our analysis as it helped provide meaningful results when investigating sub-graph patterns. By this construction we had our complex networks in a form which is commonly referred to as a pass map.

2.2 Network Statistics

There are a number of statistics which give insight into the complex networks that we constructed. To start, since our network only has 11 nodes, it is feasible, with the absence of long run-times, to look at particular node centrality measures to gain some preliminary insights into both how our network construction works and which players are most central to it.

player	Degree	Betweeness	Closeness
Carlos Casimiro	1.5	0.014	0.91
Cristiano Ronaldo	1.6	0.020	0.83
Daniel Carvajal	1.1	0.0047	0.71
Francisco Suárez	1.6	0.017	0.83
Karim Benzema	1.8	0.029	0.91
Keylor Navas	0.7	0.0015	0.56
Luka Modrić	1.9	0.039	1.0
Marcelo de Vieira	1.6	0.016	0.83
Raphaël Varane	1.6	0.071	0.83
Sergio Ramos	1.8	0.066	0.83
Tonni Kroos	1.6	0.010	0.83

Table 1: Player Centrality measures vs Liverpool (Match ID: 18245)

Table 1 details three centrality measures we decided to utilize for one of the networks we constructed. Starting with closeness centrality, in general we saw that the values are fairly large. An interesting result is that central midfielder Luka Modrić has the highest closeness centrality with a perfect 1.0 - meaning that he made at least two passes to every player in the team during the game. This is a good sign that we have a reasonably constructed network, as you would expect a midfielder to make lots of passes. For betweeness centrality, we see that the

values are quite small. However, this is actually the result we would expect, since many players directly pass to the majority of the other players throughout the match. Therefore, it would be unlikely that a tertiary player lies on the shortest path between two players. Lastly, we see again that midfielder Luka Modrić has the highest degree centrality. This result concurs with our prior hypothesis that this should be the case, as you would expect a midfielder to be interacting the the majority of players on a team during a match.

Match ID	**267076**	**276569**	**9736**	**9924**	**18245**
Number of Nodes	11	11	11	11	11
Number of Edges	82	71	82	73	84
Avg. Node Degree	7.45	6.45	7.45	6.63	7.64
Average Clustering Coefficient	0.761	0.657	0.771	0.711	0.8289
Average Shortest Path Length	1.27	1.381	1.254	1.345	1.236
Graph Diameter	3	3	2	3	3

Table 2: Network Statistics for Real Madrid vs (Opponent)

Highlighting general network statistics for all five networks we constructed allowed us to determine what the effect of our complex network construction methods were. It is evident in table 2 that these statistics do not vary significantly between matches. There is an increase in clustering coefficient for the single game vs Liverpool which may be significant. However, in general, these statistics highlight the fact that our networks are fairly dense, and highly connected. This can be seen by the fact that the average node degree is roughly 7 which shows us that, on average, a player will pass to 7 out of 10 of other players at least twice throughout the course of a match. Additionally, the average shortest pass length is close to 1, which shows us that in the majority of cases, each node is connected to all other nodes directly.

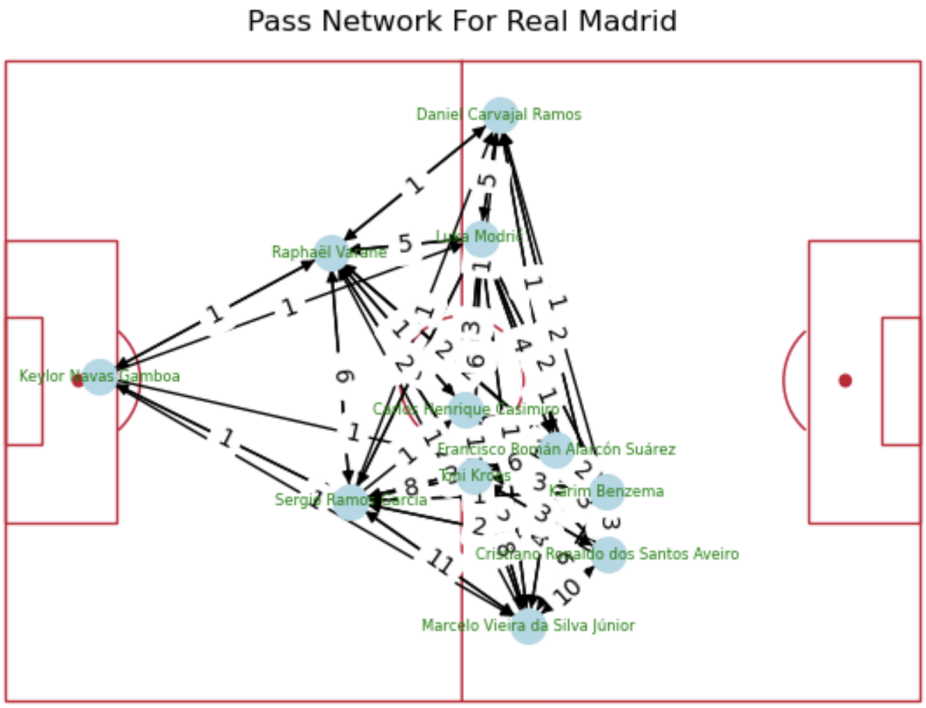


Figure 1: Real Madrid Pass Map from Match ID 18245

Lastly, one of the complex networks we constructed for this analysis can be visualized above in figure 1. It is from the match with match ID 18245 between Real Madrid and Liverpool and was created by using the methods detailed previously in section 2.1. To aid thee visualization, although unimportant as it relates to our analysis, each node is placed at the player's average position when they pass the ball. This particular visualization shows that Real Madrid favored the right side of the pitch when they had the ball, as well as the closeness of some players to each other. Additionally, we can infer than Real Madrid played in a 4-3-3 formation, and that their fullbacks liked to transition into more attacking roles when they had possession.

3 Experimental Results

3.1 Motif Detection

As we were interested in determining whether or not we can use network analysis to determine if there are certain sub-graph patterns that occur frequently in our network, we decided to conduct motif detection analysis. We first counted the occurrences of certain motifs (induced sub-graphs) and then used the configuration model [3] as a random model to standardize the results [4]. This allowed us to visualize whether or not particular induced sub-graph patterns are either over or under-represented in our real world networks. Specifically, a value greater than 0 indicates that a motif is over-represented in our real world network in comparison to the random model, and a value less than 0 indicates that a motif is under-represented in our real world network in comparison to the random model. Figure 2 below details the specific motif's we decided to investigate the occurrences of. Specifically, we were interested in looking at all the possible induced sub-graphs of size 3.

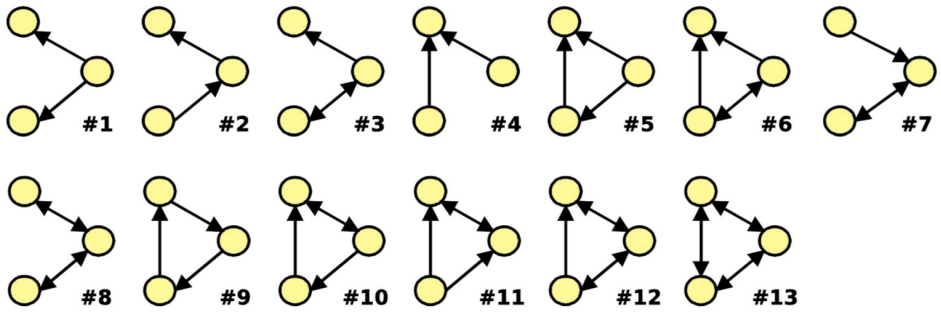


Figure 2: All Induced Sub-Graphs of Size 3

As visualized in figure 3, we implemented a motif detection algorithm for two teams. While we were only interested in the results of the Real Madrid motif detection results, we needed to conduct a proof of concept to enable us to be confident that our results provide meaningful insight into a teams tactics and strategy. Hence, we looked at five matches Liverpool played between 2015 and 2016 in addition to the matches, stated previously, that Real Madrid played. It is clear that fully connected sub-graphs (M13) of size 3 are highly over-represented in our real world networks. This result is true for both Real Madrid and Liverpool. This could indicate that this level of over-representation is common for all football clubs. However, when looking at the occurrences of M6, M7, and M8 we see a variation in the over-representation of motif's between the two clubs. This could lead us to the conclusion that Real Madrid likes to pass the ball in a way outlined motif #8 in figure 2 more than Liverpool does. Likewise, the same can be said about motif #6 for Liverpool.

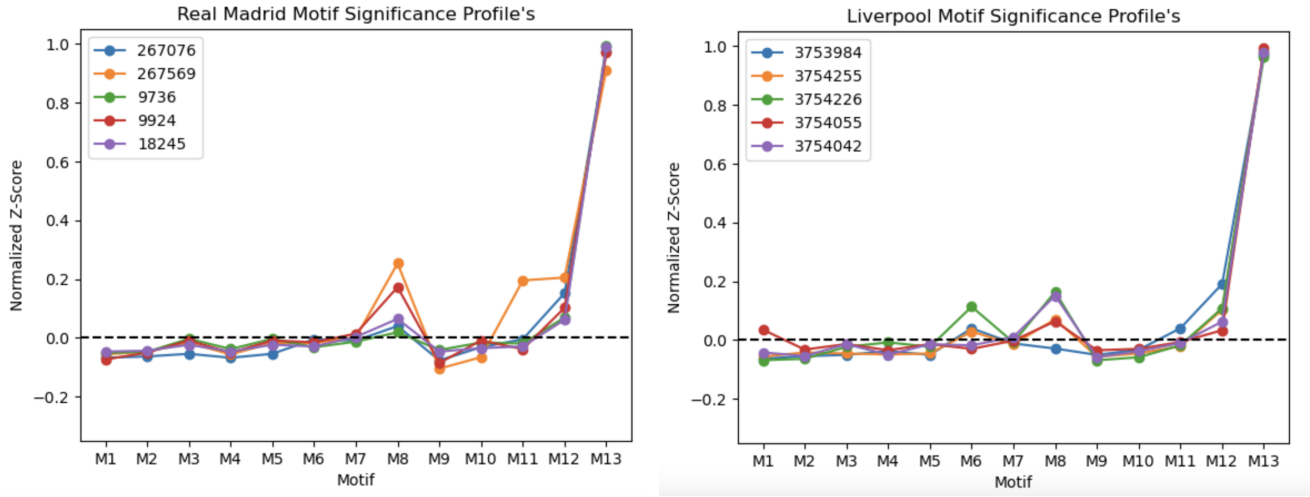
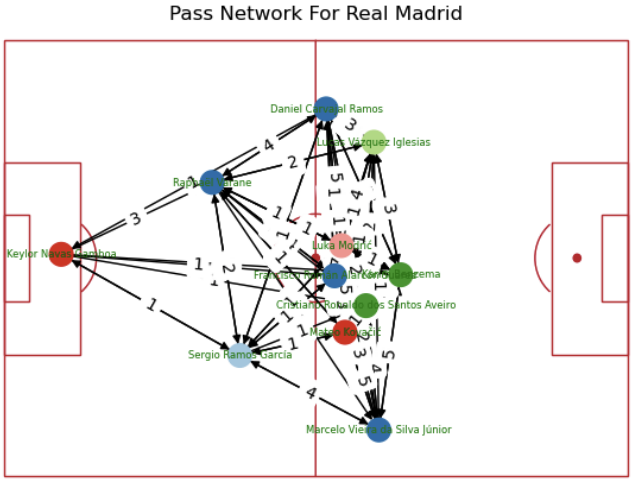


Figure 3: Motif Detection Comparison Between Real Madrid & Liverpool

Additionally, to make sure these results are concrete we conducted the same analysis for five matches Napoli played in Serie A during 2015-2016. The results of this analysis can be seen in figure 6 in Appendix A. We can see that the level of motif representation in the Napoli networks is similar to that of Real Madrid, suggesting that they might have a similar playing style in terms of passing.

3.2 RolX Analysis

As we are interested in investigating how each players contributes to the team in order to determine how the team as a unit operates we decided to perform RolX analysis instead of using alternate measures such as node centrality. RolX is an algorithm which extracts binary, structural features of a network and uses a node-feature matrix to assign each node a different, or equal role depending on how similar their features are [5]. As a result of this analysis, figure 4 provides a network visualization of the role's assigned to each node (player) as well as a summary of those roles on Real-Madrid for the match Real-Madrid vs Barcelona (La Liga 2016-12-03) with match ID 267076.



Player	Role
Cristiano Ronaldo dos Santos Aveiro	4
Daniel Carvajal Ramos	1
Francisco Román Alarcón Suárez	1
Karim Benzema	4
Keylor Navas Gamboa	6
Lucas Vázquez Iglesias	2
Luka Modrić	5
Marcelo Vieira de Silva Júnior	1
Mateo Kovacić	6
Raphaël Varane	1
Sergio Ramos García	0

Figure 4: RolX Analysis on Real Madrid Players During Match ID 267076

Through this analysis it is evident that the RolX algorithm sorted the players into 7 different roles. Through the visualization it is clear that there is a distinction between players and their positions. Note, two nodes which have similar, yet still distinct, roles are represented through shading (ie light vs. dark blue). We see similarities between defenders, as well as similarities between attackers. This highlights that the way these players interact with the whole team in terms of passing are similar, and thus play in a similar way to one another. For example, attackers might have a high in-degree but a low out-degree. In addition to this iteration of the algorithm, figure 7 in Appendix B highlights the RolX analysis of another match Real Madrid played. Showing that the same players/positions in Real Madrid have similar roles across games.

After establishing that the RolX algorithm provided meaningful results we wanted to determine whether the results obtained were the same for every team, or whether they give relevant insight into a teams unique strategy and tactics. For this reason we ran the RolX algorithm on Real Madrid's opponent, Barcelona, in the same game. The results can be seen below in figure 5. It is important to note that if two nodes share the same color between these two networks, it does not mean they have the same role. The node coloring is independent within each iteration of the algorithm. You can see that there are 2 distinct roles assigned to players, and that every player has either the same, or similar role. This coincides with the strategy Barcelona is famous for, named tiki-taka, a play style in which possession of the football coupled with one-touch passing is heavily valued [6].

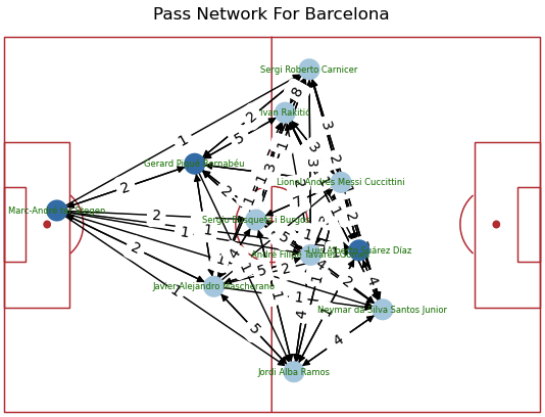


Figure 5: RolX Analysis on Barcelona Players During Match ID 267076

Similarly to the motif detection analysis, to ensure that our results were concrete and meaningful, we conducted the RolX analysis on a match which Napoli played. This analysis can be see in figure 8 in Appendix C. In this case, the algorithm found 4 different roles, differing from both the Real Madrid and Barcelona analysis. The algorithm clearly separated the attackers from the rest of the team in terms of roles, suggesting that they play in a varying style to the rest of the team.

Hence, we observed that the RolX algorithm provides valuable, meaningful results regarding a team's play style and the role that each player undertakes.

4 Discussion

This project was designed as a feasibility test for using advanced network analysis techniques in order to analyze football data, and overall was successful in demonstrating that employing algorithms such as RolX, Modularity Based Clustering, and Motif Detection can give insights into football match data. The results of Motif Detection and RolX analysis are particularly encouraging for several reasons. From motif detection, we observe a highly significant increase in the number of M13 fully connected sub graphs. This could be explained by the use of "triangle passing" - a tactic that is commonly used by high level teams to maintain possession and quickly move the ball in the oppositions half of the field. From RolX analysis, we see a clear separation of players into roles which align very well with their actual roles given by the teams starting lineup. We see in figure 4 that the true starting lineup is matched fairly well by the RolX analysis, where the 3 attacking players are given similar (green) roles, Midfielders are given red, and wingers can be seen in the blue roles.

The implementation we have demonstrated could be used on any football match with the appropriate data, namely pass data which includes the recipient and initiator respectively. In our creation of the graph visualization, we used the mean x and y coordinates of each player as their respective position's, but this is not necessary for the creation of best partitions, motif detection, and RolX analysis. This type of analysis therefore has great potential to be scaled and used to take a look at potentially any teams play-style.

There are also various elements of the graph creation that could be tweaked and played with in order to accommodate different analysis types. One choice that we made which could have been changed would be the decision to take the edge $weight$ to be $\lfloor \frac{weight}{2} \rfloor$. This was done to remove long ball passes from the network, which we took to not be representative of the general playing structure of a team. There are of course many other ways to do, or not do this, and each could result in different analysis outcomes. Another graph creation choice was to include only the starting XI players in a graph and to ignore substitutes. Including substitutes could be an interesting way to compare how two players fill the same theoretical role in a team, and is an area of potential further analysis.

Although the results we found are meaningful, there are still additions to our analysis which could be interesting to explore in more detail. The biggest influence in the results of this type of analysis is how the network is constructed. In the complex networks we constructed, we only consider the passes between players. In would be interesting to see how the results of the analysis would change if our methods of network creation changed. For example, considering shots on target, goals, and possession time as node attributes which could be used to extend the functionality of the Motif detection and RolX algorithms to produce meaningful results.

Another Implementation of this analysis could be to analyze data on a much larger scale. For example, investigating every game in a season from a particular football league and determining the average motif occurrences from every team in that league. Those results could then be used to give definitive results as to what a clubs playing philosophy is as well and possibly leading to explanations for a clubs performance in that season. Further adaptations to this analysis could involve prediction models with machine learning to make predictions on how football clubs will perform in the coming years.

References

[1] "Statsbomb." https://statsbomb.com.

[2] "Statsbomb data." https://github.com/statsbomb/open-data.

[3] "Directed configuration model." https://networkx.org/documentation/stable/reference/generated/networkx.generators.degree_seq.directed_configuration_model.html.

[4] R. Milo, S. Shen-Orr, S. Itzovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple buildingblocks of complex networks,"

[5] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li, "Rolx: Structural role extraction mining in large graphs,"

[6] "What is tiki-taka? how tactics made famous by barcelona and spain work." https://www.goal.com/en-us/news/what-is-tiki-taka-barcelona-spain-tactics/5f3qumd4uank198jwik1ww8mr.

Implementation

Our Implementation can be found here: https://github.com/cse416a-fl23/fp-mitch_willem

Appendix

A:

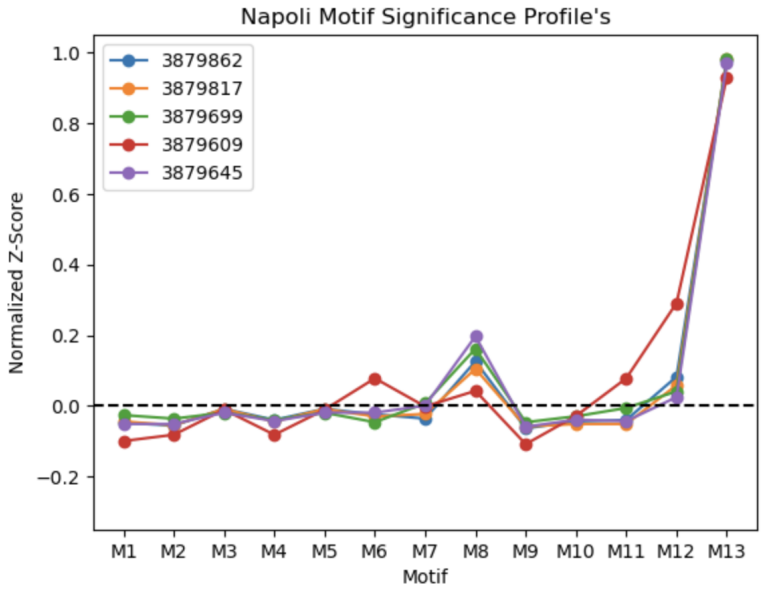


Figure 6: Napoli Motif Detection

B:

Figure 7: RolX Analysis on Real Madrid Players During Match ID 18245

C:

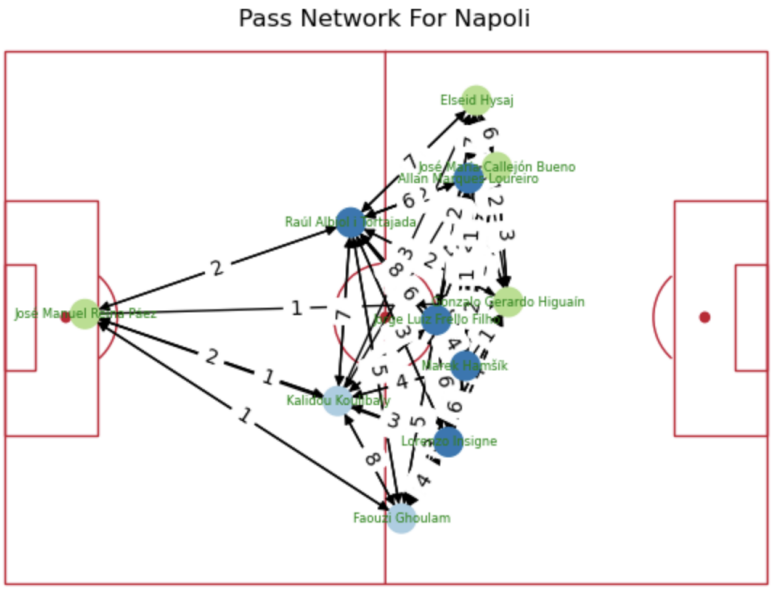


Figure 8: RolX Analysis on Napoli Players During Match ID 3879862