City, University of London

MSc in Data Science

Project Report

2024

Visual Analytics of Events in Football Games – Analysis Of Passing Strategy

A Case Study of Arsenal's Performance in the 2017-18 English Premier League Season

Nicholas Tsioras

Supervisor: Dr Gennady Andrienko

Submitted: 02/10/2024

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that related to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: Nicholas Tsioras

ABSTRACT

This study explores how data science techniques and algorithms can be used to provide an in-depth analysis of football events from matches and particularly to analyze the passing dynamics and skillsets of Arsenal FC during the 2017-18 English Premier League Season. Network Science was used to visualize the passing interactions between players in different game contexts, while factor analysis was applied to uncover underlying factors in players' passing statistics to help us gain a deeper understanding of how players are associated with the main passing factors and overall help us identify key players and identify players' strengths and weaknesses in passing. Passing event sequences were then grouped as possession episodes and unsupervised learning techniques such as Cluster Analysis and Topic Modeling were applied to group similar types of possessions together based on the distribution of the passing events, the spatiotemporal features of the episodes and the involvement of different player positions. This helped us identify common patterns and themes in Arsenal's possessions revealing their strategic nuances and technical challenges. This research uses multiple datasets from Figshares' dataset collection and combines them into a single, merged dataset that combines information about the events, the players and teams involved, alongside the matches and competitions in which the events occurred. The dataset was then filtered out to keep only events associated with Arsenal during the 2017-18 English Premier League season.

Keywords: Possession Episodes, Network Science, Cluster Analysis, Factor Analysis, Topic Modeling, English Premier League, Arsenal FC.

TABLE OF CONTENTS

Table of Contents – Images

Table of Contents - Figures

Table of Contents – Tables

1. Introduction and Objectives

1.1 Background and Challenges of Big Data Analytics in Football

1.1.1 History of Football Analytics

Although it may appear that the use of data analytics in football has been significant for only the last few years of the football modern era, the truth is that its roots go back to the 1950s. Big data analytics was first introduced by an Englishman war veteran and football enthusiast, Charles Reep during a match between Swindon Town and Bristol Rovers on March 18, 1950. Reep began to take notes of Swindon's style of play in multiple games and came up with fundamental discoveries about the positioning of players, and tactical patterns. His main focus was to identify passing plays that were more effective and resulted in more goals (Ritchie, 2020).

AC Milan made a more effective use of big data analytics in the late 1980s with the 'Mind Room' which was established by Dr. Bruno Demichelis and is a method which became well-known for combining player data with neurology and stress reduction therapy to enhance players' mental health and physical performance. This initiative was very effective and contributed significantly to the club's success, helping them win many trophies and reducing player injuries (Nassoori, 2022).

1.1.2 Evolution of Football Analytics and the Emerging Challenges of Big Data

Data analytics first emerged as a recognized profession in the late 1990s, during the rise of video content and analytics. During this period, many of the top-level European Clubs adopted systematic analysis and hired performance analysts who utilized video data extensively in their work. A great example of how data analytics started to play a significant role in coaching decisions was in 2001 when Alex Ferguson traded Jaap Stam to Lazio from Manchester United. Many football enthusiasts were baffled by the unexpected decision but Ferguson later admitted that match statistics played a major factor in his decision (Sekan, 2023).

An example of how big data analytics has been applied in recent years as a result of the tremendous advancement of modern technology, and the amount of data acquired from each training session and in multiple games is Liverpool FC to gain a competitive advantage over

their opponents and establish themselves as one of the best teams in Europe. Dr. Ian Graham, a key asset to Liverpool's analytics team, under the influence of the 'Moneyball' approach, popularized in Baseball, developed advanced mathematical models beyond the scope of traditional statistics to innovate their transfer strategy against wealthier clubs. Stats like expected goals (xG) and assists (xA), and metrics that analyze the defensive contributions, passing strategy, off-ball movements and the physical attributes of players were incorporated and were crucial for providing insights that could help players adapt to new roles that fit to Jurgen Klopp's 'gegenpressing' strategy. This methodical, data-centric approach allowed Liverpool to identify players who were undervalued by other teams but showed high potential according to their analytical models. Another way in which Liverpool FC employ advanced data analytics is through tracking systems, aerial cameras, and live video analytics to monitor real-time player and ball movements. These technologies offer data about player positioning, performance, and fatigue. Led by Ian Graham and supported by Michael Edwards, Liverpool's data analytics team primarily focuses on pitch control and strategic planning, integrating big data into player recruitment and match analysis (Dougramaji, 2023).

The recent development of tracking technologies and data analysis are transforming sports science, particularly in analyzing team tactics and player performance optimization. Football analytics has seen a significant transformation due to the emergence of advanced big data technologies. Traditional methods of tactical analysis are heavily relied on observational data, often disregarding the contextual information which is crucial for a comprehensive understanding of the game dynamics. The use of sophisticated tracking technologies and the integration of extensive physiological data have substantially enriched the variety, volume and depth of data available, expanding the scope of analysis significantly. While this shift has addressed previous limitations due to data paucity, it has also introduced new challenges related to managing, processing, and extracting insights from the large volumes of data generated (Rein and Memmert, 2016).

The integration of big data in football analytics, has necessitated the development of new, advanced analytical techniques and theoretical models to guide data-driven tactical decision-making. Supervised and Unsupervised Learning Modeling techniques such as Machine Learning, cluster analysis, and artificial intelligence, have been heavily utilized in recent years and play pivotal roles in sports analytics as well as other domains such as medicine, finance, retail etc. However, the effective use of these analytical techniques requires overcoming

important organizational hurdles related to data governance, accessing cutting-edge technologies, and data processing and management (Rein and Memmert, 2016). Vermeulen and Yadavalli (2018), discuss the challenges associated with big data in sports analytics, particularly focusing on three main challenges:

- **Data Veracity:** Challenges concerning the accuracy and reliability of data extracted from wearable technology which may include noise and erroneous readings that could result to mislead research and decision-making processes. This indicates that advanced analytical methods and techniques need to be applied to handle and process the data effectively.
- **Data Privacy:** Challenges concerning the sensitivity of the data being accessed or misused. This challenge highlights the need for robust security measures to protect personal information of the athletes.
- **Athlete Autonomy:** This term refers to challenges concerning the excessive reliance of players on data for decision-making. This might limit the athlete's ability to trust their own judgments and instincts which could lead them to reduce their control and influence over their own sports strategies and performance.

The future of sports science appears to be on the brink of revolution, with big data technologies and advanced analytical methods at the forefront, promising new insights into data-driven tactical decision-making and player optimization. The challenges mentioned above, highlight the importance of addressing challenges regarding handling the depth of the data and the consideration of ethical and technical issues to ensure the responsible and valid use of big data in enhancing player performance.

1.2 Purpose of the Project and Objectives

1.2.1 Introduction

Football is often recognized as the 'beautiful game' and has become a complex interplay of strategies, skills, and teamwork. As technology continues to evolve, more sophisticated methods for data collection have emerged and therefore the role of analytics in football is constantly growing. The ability of sensing technologies to capture records of real-life spatiotemporal events and statistics of passing, shooting, positioning, etc. has enabled coaches, analysts, and clubs to utilize football analytics as an indispensable tool making football an interesting sport from a scientific point of view.

1.2.2 Importance of Analyzing Passing Dynamics for Tactical Understanding

One of the most critical aspects of football is the passing dynamics, which can significantly impact the outcome of matches and can help coaches and players to understand the overall tactics of their opponents. In the modern era of the game, many coaches see football as a possession-based game, meaning that coaches prioritize ball control with high passing accuracy while aiming to keep possession of the ball for extended periods of time by constantly finding numerical superiority across different areas of the pitch. Coaches and analysts, constantly seek to find ways to gain competitive advantage of their opponents. One very effective way of doing that is to analyze and understand the passing network and patterns in different contexts.

With the volume and variety of event-based data increasing rapidly in the field of sports, coaches and analysts can capture the complexities of sequential passing interactions in real time. The purpose of this project is to explore the details of the passing dynamics and patterns of a single team in different game contexts. By leveraging the power of data science and advanced analytical and modeling techniques, this project aims to provide a comprehensive review of how the passing behaviours change over time and over different contexts and how effective they are. For this project, we will utilize a publicly available dataset from figshare, which contains detailed statistics of events from multiple games played in a single season across the top 5 European Leagues and Major International Football Tournaments such as the European Championship and the World cup. For this project, we will be analyzing the spatiotemporal passing events of Arsenal across multiple games of the 2017-18 English Premier League Season. We will be combining multiple datasets that capture various aspects of the events, such as the pass types, success rates, spatial coordinates, and temporal dynamics alongside data about the players involved in the passing sequences, and contextual information about the matches played. By focusing on a single team, we can apply a more precised and focused analysis of the passing dynamics with greater context.

1.2.3 Project Objectives for Analyzing Football Passing Dynamics

In this project, we aim to address a main question that explores the potential and ability of data science techniques to unravel the passing patterns and dynamics, as well as evaluating player's skillsets in passing. The primary question of our project is:

Can Data Science Techniques Uncover Hidden Patterns and Dynamics in Football, and Provide Insights into the Skillsets of Players in Passing?

To answer the main question posed in this project, several objectives need to be completed.

- **Identify the roles of key players that influence a team's playing style and their interactions with other players.** Identify the common factors that define unique passing styles and measure the association of players within each factor to find versatile and less versatile players. Find the passing interactions of players to identify key players that influence the team's passing strategy.

- **Analyze the positional tendencies of players across different game situations.** Analyzing the passing interactions and the positional movements of players to gain deeper insights into team dynamics and the specific roles and contributions of individual players.

- **Identify distinct passing patterns and dynamics across various game scenarios.** Use contextual information about the passing events such as the opponent strength, venue and current score to visualize how the distinct passing patterns change in different contexts. This will help us identify the main factors that influence the passing dynamics.

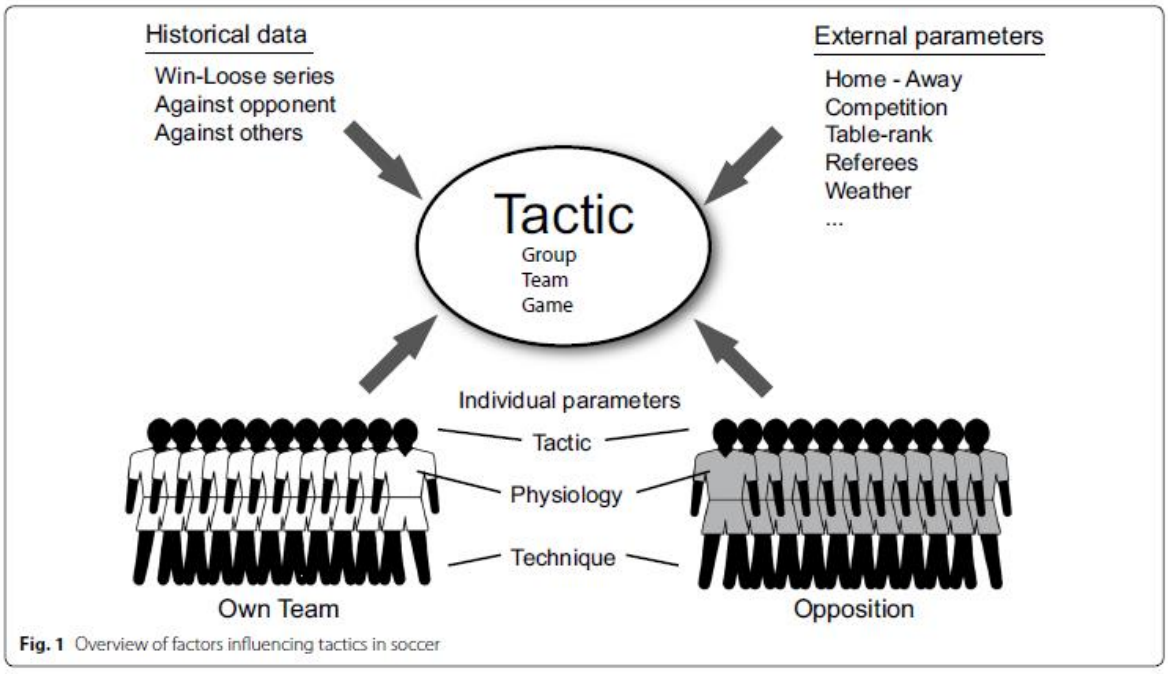


Image 1. Overview of the main factors influencing tactics in football

Image Source: Rein and Memmert, 2016.

- **Uncover underlying themes in football passing styles that reveal distinctive possession strategies across various game situations.** Find themes (topics) in ball possession and see how they occur in different game contexts.

- **Find the on-ball passing patterns and themes that are most relevant to success.** Study the outcomes of unique passing patterns and themes. Use measurements such as possessions leading to shots, or leading to non-accurate passes and interceptions to define the success of these possessions. Analyzing the distribution of possession patterns and themes across wins, draws and losses in football matches will also uncover their relevance to success.

1.3 Approaches for Completing the Objectives

- **Identify the roles of key players that influence a team's playing style and their interactions with other players.** Network graphs of passing interactions between players with adjusted weights of links between them combined with factor analysis of passing events will help us gain a deeper understanding of the unique roles players have in the team's passing strategy and also highlight each player's strengths and weaknesses.

- **Analyze the positional tendencies of players across different game situations.** Player heatmaps that show the main positions where players attempt passes will help us understand their roles and movements during ball possession. To study better the positional changes of players in unique scenarios, we will be comparing the average X and Y coordinates of passes in different game contexts (against different level of opposition, home vs away etc.). Heatmaps will be generated to show the areas where players are more active and contribute to the team's passing strategy.

- **Identify distinct passing patterns and dynamics across various game scenarios.** The first step is to group sequential passing events into possession episodes until the team loses possession. After defining the possession episodes, the next step is to extract information about the types of passes in each possession, the player positions involved, and summary statistics of the passes. Multi-Dimensional scaling, Principal Component Analysis and Clustering algorithms will be used to identify similar possessions and separate the regular

and occasional possessions. Contextual information about the possession clusters will help us analyze when and in what circumstances does the team use each cluster.

- **Uncover underlying themes in football passing styles that reveal distinctive possession strategies across various game situations.** Topic modeling is one very effective way to analyze and identify underlying themes (topics) of football possessions. The next step will be analyzing the occurrence of each topic in different game scenarios to see how various contexts influence the occurrence of the main passing themes. Temporal analysis of the topic occurrences over different time segments of games and over the course of the whole season will help us understand changes in passing tactics over time.

- **Find the on-ball passing strategies that are most relevant to success.** Using the tag_description column which has information about the outcomes of passes, for each possession we will define the outcome of the last event as the outcome of the whole possession. This feature will add contextual information to the possession clusters and will help us understand which types of possessions are more and less effective. Tags like shots will indicate successful possessions where the team was able to distribute the ball well and create a significant chance, while other tags like interceptions and non-accurate passes indicate possessions where the team struggled to move the ball forward to the final third.

1.4 Beneficiaries

This study is beneficial for numerous professionals in the field of football who can exploit data analytics to gain a competitive advantage over their opponents. There are three main types of professionals who can massively benefit from this project:

- **Team Coaches and Analysts.** With football analytics becoming more important, team coaches and analysts can gain deeper insights into the passing dynamics and strategies of their opponents. This will help them identify their opponent's strengths and weaknesses to refine their tactics, and optimize their own player's performance. A primary example of how coaches and analysts can benefit from the implementation of advanced data science techniques to gain insights into team strategies and player roles is detailed in the paper 'An Improved Passing Network for Evaluating Football Team Performance which was published by Zhou et al. (2023). This study introduces an advanced passing network model

which incorporates a coordinate-based system to evaluate each pass's potential contribution. By calculating the start points and endpoints of passes, while also assigning weights to pass interactions of players, this paper was able to uncover the passing dynamics and roles of Everton players in different games during the 2017-18 season.

- **Team Scouts.** Scouts can leverage the insights from big data analytics to improve their decision-making. By analyzing advanced performance statistics of players, they can identify young talent and predict player potential. A primary example of how big data analytics has revolutionized player scouting is detailed in the article published by LatentView (2022). This article explains the shift from traditional methods of player scouting to more data-driven approaches such as predictive analysis and machine learning with the main examples being Barcelona FC, Brentford FC, and Liverpool FC who used the philosophy in their recruitment process defined as 'Quest for Rationality'. Gaining insights from these complex methods helped scouts identify undervalued talents that could not be identified by just analyzing traditional statistics (Goals, Assists) and overall helped them optimize the scouting process.

- **Researchers and Sports Scientists.** Big data analytics offers significant advantages to researchers who are interested in the application of data science in the field of sports and particularly in football. By leveraging vast datasets with data such as psychological metrics, physical attributes, and performance metrics, researchers and scientists can uncover hidden trends, patterns, and correlations that reveal how different factors influence a player's performance, long-term development, and overall well-being. Having access to these insights can revolutionize the fields of sports medicine and exercise science by enabling the development of personalized training programs based on the needs of each athlete, improving their performance efficiency and preventing them from getting major injuries. Overall, big data analytics enables sports scientists and researchers to make more data-driven decisions, combining their theoretical knowledge with the practical insights from the data analysis (Dergaa and Chamari, 2024).

1.5 Workplan and Structure of the Report

1.5.1 Workplan

In general, the workplan does not vary much from the figure shown in the proposal (Appendix A). The order of the tasks stayed the same, with only a few deviations. We began by focusing

on the event-level analysis of passing interactions between players and their positional tendencies by creating the necessary network graphs and heatmaps. Factor analysis was then applied on the passing events to associate each player with the underlying factors to have a better understanding of their role in the team's passing strategy. Then, we delved into the possession-based analysis of the passing events using unsupervised learning algorithms such as clustering and topic modeling to uncover underlying patterns and themes amongst the passing sequences. Due to the nature of the datasets, the collection of data and the preprocessing steps took a longer than expected but remained one of the very first steps alongside the necessary literature review. Also, some player information was missing so these cells needed to be modified manually. The interactions of players were captured for the construction of passing networks, while the events were grouped into possession episodes for sequential analysis. Finally, the evaluation of results and the writing of the report were given a significant amount of time to ensure the project was correctly completed and presented.

1.5.2 Structure of the Report

The report consists of six main chapters and extra supplementary material.

Chapter 1 introduces the background of the problem and the evolution of football analytics and modern challenges of data science in the field of sports. It also mentions the purpose of the project and why the analysis of passing dynamics is crucial for tactical understanding as well as the professional who can benefit from this project.

Chapter 2 provides a clear literature review and provides context about the use of data science today in sports and particularly in football. It also describes the key theory and practical use of algorithms used in this paper.

Chapter 3 provides a more clear review of the methods and algorithms used to complete the objectives. This includes the main preprocessing steps such as grouping passing events into possession episodes, encode the pass types and metrics in possessions, extracting spatial and temporal information about the passing events, constructing and evaluating algorithms for cluster analysis and topic modeling of possessions, while also constructing and evaluating the factor analysis algorithm to extract common passing factors and associate each player to the factors.

Chapter 4 explains the outputs produced by using the methods and algorithms. It highlights the most significant findings that will help us gain insights into the passing dynamics of Arsenal during the 2017-18 season.

Chapter 5 examines how the results we obtained align with our objectives.

Chapter 6 evaluates the project as a whole, highlighting the main conclusions and lessons learned, and also recommends areas of improvement for further work.

Supplementary Material includes the glossary, references, appendices and any additional files.

2. Context

2.1 Modern Applications of Data in Sports

2.1.1 Importance and Intricacies of Big Data Analytics in Football

Team sports are driven by an ongoing competition where strategic superiority over the opponent can lead to a victory. Big data analytics have a transformative role in football with a huge impact on decision-making, player performance evaluation, and strategic development. The evolution of football driven by data is highlighted by the transition from intuition-based approaches such as observational insights, subjective judgments and coach experience to more data-driven methodologies which are more precise and comprehensive. Initially, teams record simple statistics manually, but with the advancements of computer technology, more comprehensive data collection became possible (SoccerTAKE, 2023).

Big Data is defined by their Volume, Veracity and Velocity. In football, volume refers to the large amount of data being generated. Variety highlights the potential of analyzing diverse types of data, including structured data like positional logs, semi-structured data like XML files and unstructured data from video footage. Velocity refers to the rapid generation and analysis of real-time event, psychological, and positional data in comparison to delayed notational analysis methods where data collection and analysis is conducted after the event. These aspects are crucial in tactical analysis as they provide a more comprehensive understanding of team dynamics and performance. Big data technologies offer specific solutions for managing and analyzing these large, varied, and rapidly produced datasets, facilitating the process of extracting deeper insights into tactical behaviours. Big data analytics enables the detailed examination and modeling of complex tactical decisions, which were previously more challenging due to the insufficient and structured data (Rein and Memmert, 2016).

The sports industry, as well as others, have undergone significant transformations, driven by the growth of data generation and analytics. Lolli et al. (2024), explore the growing role of data analytics in professional football. The study surveys the perception of the value of data analytics across different members of staff in football. The paper compares the role of data analytics in sports to other industries. Similarly to fields like finance and healthcare, sports analytics can optimize performance emphasizing its critical role in informed decision making and competitive advantage.

The key findings of this survey highlight the low offer of professionals with expertise in data analytics and data engineering especially in national federations. In terms of the perceptions of

data utility, there was a strong belief that data analytics is a valuable asset for decision-making in areas such as player performance, recruitment and medical management. Just over half of the respondents either agreed or strongly agreed that data-driven information is the main guide to their decision-making. The most common method to present information to the coaching staff seems to be presentations from data scientists or match analysts. These findings highlight the growing appreciation for the role of data analytics in football (Lolli et al., 2024).

Respondents were also questioned about the usefulness of certain metrics for tactical analysis and match outcome predictions. The most relevant metrics were considered to be Ball Recovery Time, Possession Control, Line Breaks, Team Shape During Possession, Forward Pass Ratio, Pass Completion Rate, and Final Third Entries rather than simpler metrics like Expected Goals (xG) and Expected Assists (xA) of players and Team Threat. This helped us understand that analyzing sequential events that define possessions as a whole and mainly relying on metrics that indicate the team's passing patterns and dynamics are more effective in uncovering the tactics of a team, rather than just relying solely on individual skills. This is why we decided that this project should mainly focus on exploring the passing dynamics and patterns in a possession-based analysis of the game.

2.1.2 Episode-Based Analysis of Spatio-Temporal Events in Football

Due to the fluid nature of the game and the dynamic interactions and tactical changes over short periods, Andrienko N., Andrienko G. and Shirato (2023) present a methodology for analyzing episode-based data to understand the distribution of the multi-attribute dynamic characteristics of the game. An episode is defined as *«a time interval during which a dynamic process or behaviour occurs»* (Andrienko N., Andrienko G., and Shirato, 2023). In football terms, this means analyzing sets of episodes that include the sequence of multi-attribute variables that describe events that happen from the moment a team gains possession of the ball until the moment they lose it. This approach facilitates the process of understanding complex behaviours by identifying and analyzing patterns within episodes, providing insights into when, where, and under what conditions these patterns occur.

An example of the episode-based data analysis is the classification of transitions and counterattacks. Eusebio, Prieto-González and Marcelino (2024), offer insights of 88 performance indicators focusing on build-up play and transitions in the game and also identify 17 key performance indicators related to the defensive style of play, emphasizing the need to

adjust defensive strategies for optimized performance. While simple statistics can provide important insights, analyzing transitions and counterattacks offers a more holistic and detailed understanding of the game's sequential dynamics. The study notes that counterattacking transitions are more effective in creating scoring opportunities than any other type of attack, particularly when the opposing team is imbalanced. These insights can help coaches understand how their opponents handle transitions, allowing teams to anticipate and counter these strategies effectively.

2.1.3 The Ongoing Debate in Premier League Football: Long-Ball Game and Direct Play vs Possession-Based Football

Charles Reep and Bernard Benjamin, pioneers of football analytics, significantly influenced the tactical approaches that were used to win football games during the 1960s. Their work was primarily focused on examining the patterns in passing sequences of multiple matches from 1958 to 1967. Looking at the passing movements, they concluded that shorter passing sequences tend to be more effective and lead to more goal-scoring opportunities. They noted that on average 1 goal gets scored for every 10 shots (Reep and Benjamin, 1968). This indicates that the more chances a team can create, the more likely the team is to score, also known as the 'Chance Effect'. The findings of this study continue to be influential in the modern game. Many teams now aim to overcome the chance effect and employ direct play strategies like the 'Long-Ball' approach. This style of play focuses on either long crosses from defensive to offensive zones, attempting to break the lines and disrupt the opposition's defence or rapid, direct plays to advance the ball quickly up the field, minimizing lateral and backward passing (Hughes and Franks, 2005).

On the other hand, some experts had different views of this researcher. Many had evidence that possession football is more effective and can lead to more clear-cut chances, making it a more successful passing strategy that can lead to more goals. Pep Guardiola, one of the most influential figures in modern football, particularly known for his possession-based tactics, mentioned in an interview with Sky Sports that possession football is the key to his team's defensive success. Guardiola stated that *«Football is not A, B and C. What works last season will not necessarily work this season. What did not work last season, could work now»* (Bate, 2021), suggesting football being as dynamic and adaptable as it is, requires a more possession-based approach. While certain patterns might be successful in certain games, they must evolve and adapt to the changing nature of the game to remain effective against various tactical

approaches. Teams across the globe now try to mimic this style of play and focus on intricate passing networks and taking control of the game with relentless ball movement, a philosophy often referred to as ‘’Tiki-Taka’’ (Zamani, 2024).

The debate over the effectiveness of possession-based football versus direct play has been an ongoing discussion, from the late 50s to the present day. The growing usage of data science in football is fundamentally rooted in this ongoing dispute. The findings from these studies have influenced and shaped the objectives of this project to resolve this debate and find patterns and dynamics in passing to analyze their efficacy.

2.2 Theory Surrounding Data Science Algorithms Used to Analyze Passing Patterns and Dynamics

2.2.1 Advanced Network Science

- **Understanding Network Science and Its Role in Football Analytics**

Network science is being continuously developed by many scientists and mathematicians, who aim to study the interactions between the elements of a network in the most reliable and efficient way. The passing patterns and dynamics of a sports team are an excellent data source that can provide valuable information on team performance, as interactions between the elements of a network graph occur frequently and provide a distinct behaviour during various football matches with different contexts. The passing interaction graph is very similar to how data is arranged in a computer network and therefore network science can be a very effective way to analyze how the ball is being passed between players and uncover patterns of ball distribution (Zhou et al., 2023). The passing network is defined as «*a graphic that aims to describe how the players on a team were actually positioned. Using event data (documentation of every pass, shot, defensive action, etc. that took place during a game), the location of each player on the field is found by looking at the average x- and y-coordinates of the passes that person played during the match. Then, lines are drawn between players, where the thickness - and sometimes color - of each line signifies various attributes about the passes that took place between those players*» (Bush, 2024).

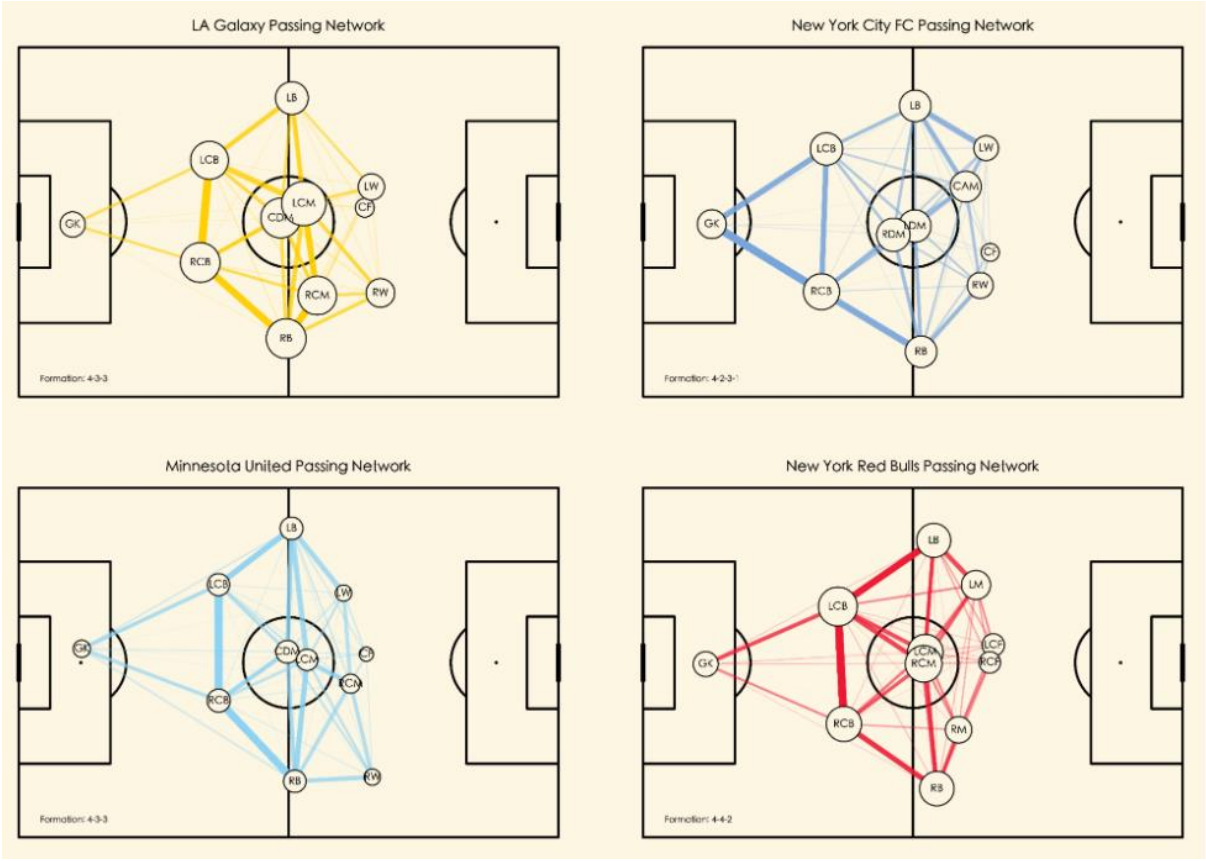


Image 2. Construction of Pitch-Player Passing Networks for Major League Soccer teams

Image Source: Bush, 2024

Image 2 provides some perfect examples of what Pitch-Player Passing Network graphs represent. The link widths are proportional to the number of passing interactions between players, whose vertical and horizontal positions (X and Y axis respectively) are calculated as the average coordinates of all passes made. This type of passing network is more holistic as it captures the strength of player-to-player interactions alongside their positions on the pitch.

- **Model Specification of Basic Passing Networks**

To analyze the two-way relationship of passing interactions between players, the original matrix of passes is transformed into a symmetric matrix. This symmetrix matrix also considers both the pass source and pass target, defined as the player who made the pass and the player who receives the pass. A simple passing network graph can be described by the following equation (Zhou et al., 2023):

$$\mathbf{G = (V , E)} \qquad \mathbf{[1]}$$

V represents players as nodes.

E represents the directed edges of the graph, indicating the direction and occurrence of a pass from one player to another.

- **Weighted Models of Directed Edges and Nodes**

This advanced model optimizes how the importance of passes between players is measured by factoring the quantity and relevance of the passes. To enhance the basic network graph model, weights are assigned to the edges representing the significance of passes between players. This type of model considers the number of passes and other relevant, external factors to assign a weight that reflects both the frequency and the strategic importance of passes between specific players (Zhou et al., 2023).

Node importance is also a crucial factor to consider while constructing an advanced network graph. Zhou et al. (2023), adapt the PageRank algorithm which is used by Google to rank the importance and connections of Web Pages to evaluate the players importance in the passing network while considering the macro-scale topology of the passing network, meaning that the node weight is adjusted not only by the number of interactions a player has, but also by the diversity of interactions (Buldu et al., 2018). In simple terms, a player who is the pass source or pass target across multiple connections will have a higher weight than a player who only has high pass interactions with certain players. To take these assumptions to consideration and improve our passing network, the PageRank value of player i is calculated by the following equation as mentioned by Zhou et al. (2023):

$$PR(i) = \alpha \sum_{j \in M} \frac{PR(j)}{W(j)} + (1 - \alpha)\frac{1}{N} \quad \textbf{[2]}$$

PR(i): Represents the PageRank index of player i, indicating their importance in the passing network.

α: This is the damping vector, which prevents the ranking from being overly influenced by isolated structures within the network.

$\sum_{j \in M} \frac{PR(j)}{W(j)}$: This part of the formula evaluates the importance of a player i based on the weighted contributions from other players who pass to them. A player who has many and different interactions with influential players j will have a higher PageRank (Zhou et al., 2023).

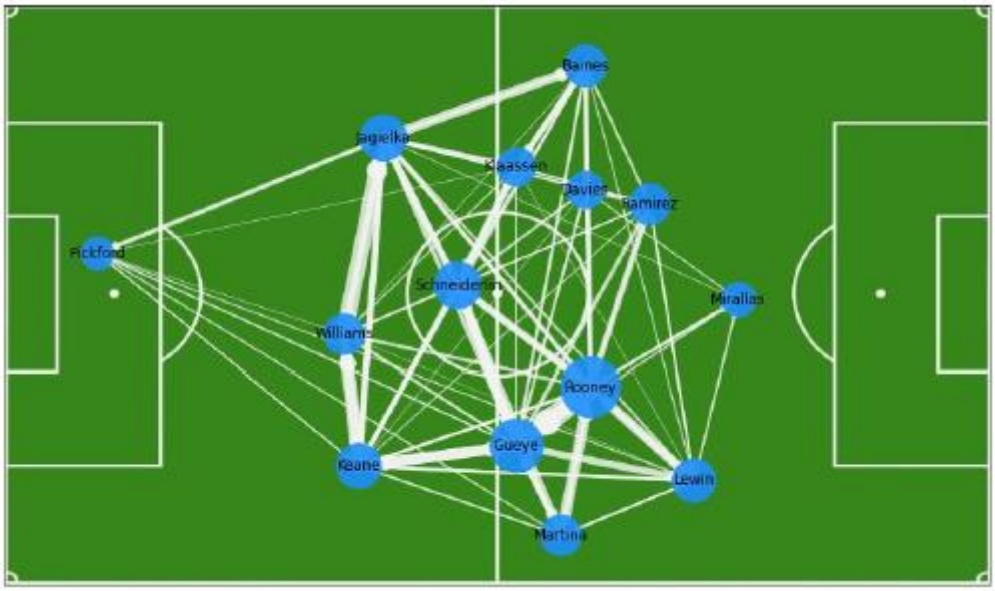


Image 3. An advanced passing network of Everton FC in their 1-0 win against Stoke City in the 2017-18 season

Image Source: Zhou et al., 2023

2.2.2 Clustering and Dimensionality Reduction Algorithms

- **Principal Component Analysis (PCA) for Dimensionality Reduction**

Principal Component Analysis is considered one of the most effective dimensionality reduction algorithms, used to reduce the number of features in a dataset while maintaining high variance to minimize the risk of losing valuable information (Nishida, 2018). To reduce the dimensionality of the data while retaining as much variability as possible, an orthogonal transformation of the data is needed where the set of possibly correlated variables needs to be into a set of linearly uncorrelated variables called principal components. The orthogonal transformation of the original data into the principal components is defined by the following equation (Li and Sang, 2018):

$$X_{j(i)}=N_{(i)}\times P_{(j)} \quad \textbf{[3]}$$

Xj(i) represent the new variables (principal components) that result from the above equation.

N(i) represents each observation, in our case each possession.

P(j) represents all the parameters or variables of the datasets, like types of passes, length of passes etc. that are more informative for the principal component X.

This equation rearranges the data into a new format that highlights the most important information while reducing the dimensionality and redundancy of the data.

To capture as much information as possible when generating the components, the following equation is used to pick out the aspects of the data that vary across different matches or possessions since these tend to be the most informative (Li and Sang, 2018):

$$P_{(j)} = \arg\max\left\{\frac{NP^T PN^T}{P^T P}\right\} \quad [4]$$

The goal of this equation is to find the principal component that maximizes the ratio of the transformed data's variance to the scale of the transformation matrix itself. Maximizing this ratio ensures that the principal components capture as much variance as possible (Li and Sang, 2018).

Image 4. Using PCA to group similar players based on their scores in various statistics

Image Source: Nishida, 2018

As we can see in Image 4, and on most occasions the first principal component captures the greatest variance in the data, with each subsequent component capturing as much variance as possible under the constraint of being orthogonal to the main principal component. This can help us associate the importance of each variable in the components and therefore, identify which factors are more significant in distinguishing possessions. For example, in the Image above, Finishing extends relatively far along the PC1 which explains 52% of the variance but not so much in the PC2 which explains only 12%). This indicates that finishing is a key factor in differentiating the data points, meaning that it's a reliable rating in differentiating player performance.

- **Finding the Optimal Number of Components in PCA**

One common approach to finding the number of components that reduce the dimensionality of the data while maintain high variance is by calculating the Cumulative Variance Explained by Number of Components. Nishida (2018), uses a line graph to see how much variance and information can be explained by adding an extra component. This can help identify after which number of components the variance explained does not significantly explain and therefore, find the optimal number of components to reduce the dimensionality of the data and keep as much relevant information as possible. Visually, the optimal number of components is indicated by the elbow point of the line graph.

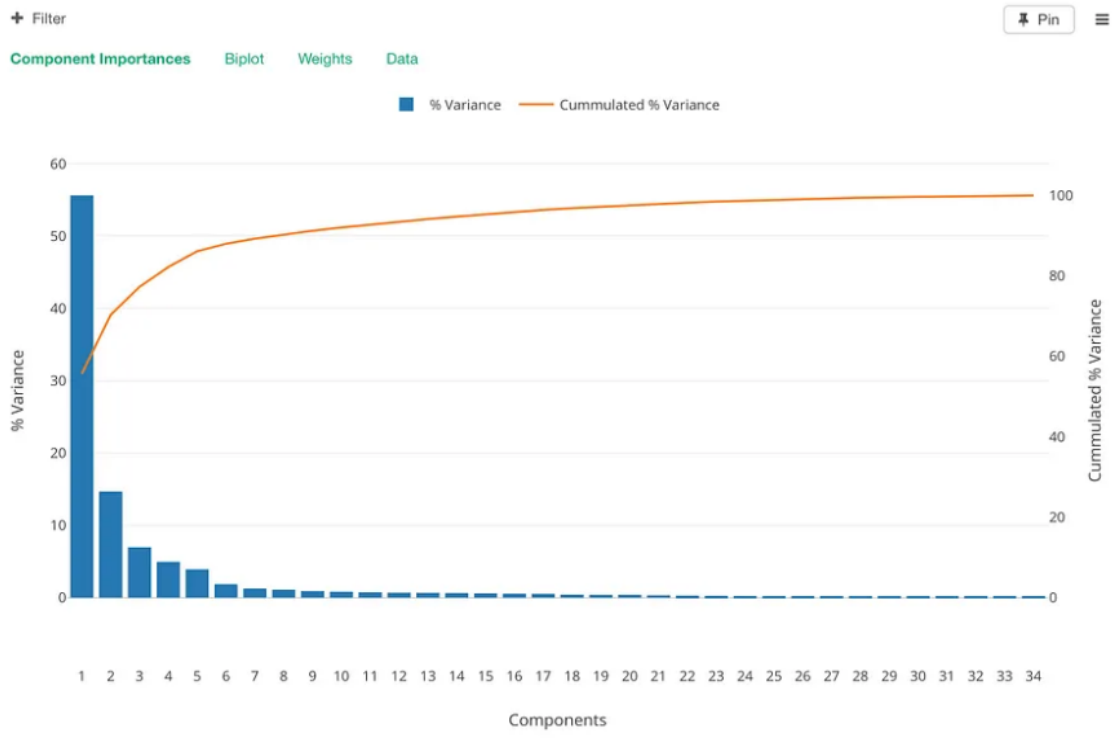


Image 5. Cumulative Variance explained by the number of components

Image Source: Nishida, 2018

Pasunuri, Venkaiah and Srivastava (2019), explore the integration of using Principal Component Analysis (PCA) with clustering algorithms to enhance clustering performance in high-dimensional datasets, in comparison to using clustering algorithms alone. The study demonstrates that by combining PCA before applying either K-Means clustering or Density-Based Clustering (DBSCAN), the accuracy and efficiency of clustering are significantly improved. This approach can effectively reduce the dimensionality and noise of the clusters, making them more distinct. The findings from this study significantly informed our approach to refining the methodology for grouping similar possession sequences to uncover patterns and dynamics, while also separating the occasional passing methods from the most frequent.

- **Factor Analysis in Football Player Evaluation**

Similarly to PCA, Factor Analysis is used for dimensionality reduction purposes. Quan, Li and Chen (2024), define FA as *«a method used to identify a few factors that can explain patterns observed in the data»*, while Tavakol and Wetzel (2020) give a more detailed explanation of FA as *«a method that allows us to simplify a set of complex variables or items using statistical procedures to explore the underlying dimensions that explain the relationships between the multiple variables/items».*

- **Mathematics Involved in Factor Analysis**

Factor Loadings are a crucial aspect of FA, as they help us understand the correlation of variables to each factor. Values closer to 0, indicate that there is no correlation between the variable and the factor. This can help us understand what each factor represents based on the correlated variables. Eigenvalues are used to determine the significance of each Factor extracted during FA. Factors with a greater value than 1 are considered relevant as to explaining a significant portion of the variance in the dataset. The eigenvalue is calculated by summing the squared factor loadings for all variables associated with a factor.

$$\lambda_j = \sum_{i=1}^{n} l_{ij}^2 \qquad \textbf{[5]}$$

A method for studying the magnitude of the factors is to calculate the variance explained by each factor as a percentage. A higher variance percentage means that there is a strong representation of the Factor as it captures the main idea that's being measured. The proportion of the variance explained is measured by dividing the sum of the squared factor loadings for

each factor (eigenvalue) to the total number of variables in the dataset (Tavakol and Wetzel, 2020).

$$\text{Proportion of Variance Explained} = \frac{\lambda_j}{p} \quad \textbf{[6]}$$

Where **λj** is the eigenvalue for the jth factor.

p is the total number of variables in the dataset.

Tavakol and Wetzel (2020) also discuss the misconception of using PCA instead of FA. The main difference between the two dimensionality-reduction algorithms relies on the purpose of their use. PCA focuses on reducing the dimensionality of the data by transforming the variables into smaller, variable-representative components that are uncorrelated between them. On the other hand, FA aims to uncover the underlying latent structure of the data by separating common variance from unique variance. This difference is crucial in understanding under what circumstances each method is more appropriate to use.

- **Clustering Algorithms**

With the rapid growth of data volume, data mining has become increasingly more difficult but remains an important role in extracting hidden, unknown, and potentially valuable information from the large amounts of data in databases. Data scientists use clustering as a data mining method for grouping similar data. In other words, clustering is a method applied to divide data points into subsets based on the similarity of their characteristics. Clustering methods can be divided into five main categories (Yin et al., 2023):

- **Partition Based Clustering**
- **Density-Based Clustering**
- **Hierarchical-Based Clustering**
- **Grid-Based Clustering**
- **Model-Based Clustering**

Partition-Based and Density-Based Clustering are the two most used methods but for different purposes. The main algorithms used on both occasions are K-Means and Density-Based Clustering. While K-Means partitions the data into a predefined number of spherical clusters, DBSCAN is used to identify clusters with arbitrary shapes, without requiring a predefined number of clusters to split the data on, making it a more appropriate algorithm for data with complex structures.

DBSCAN is notable for utilizing two main parameters (Nandi, 2022):

- The neighbourhood radius ε
- The minimum number of points required to form a cluster MinPts

By using these two parameters, the algorithm classifies data points as core points, border points and outliers based on local density, discovering dense clusters and noise. Data points that are outliers and are considered as points will not be assigned to a cluster, whereas in K-Means clustering even outliers will be assigned to a cluster (Nandi, 2022).

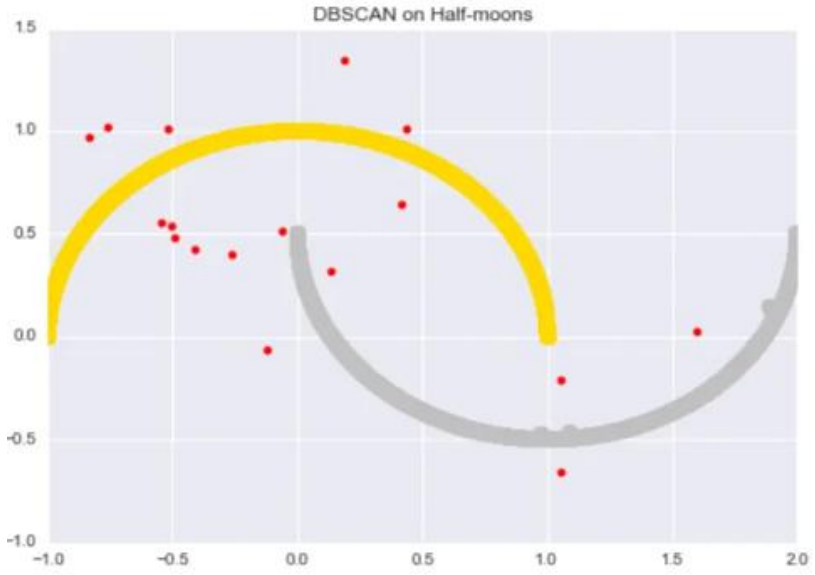


Image 6. DBSCAN of customer data

Image Source: Nandi, 2022

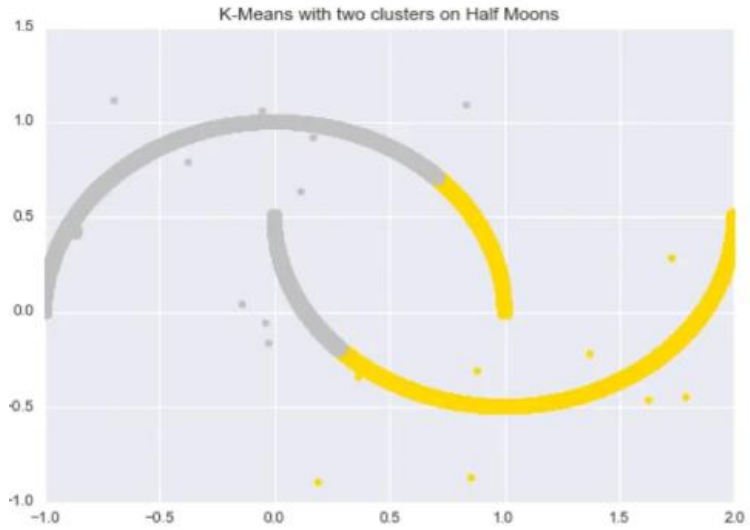


Image 7. K-Means Clustering of customer data

Image Source: Nandi, 2022

Images 6 and 7 provide a perfect example of the differences between DBSCAN and K-Means algorithms. K-Means clusters data based on the proximity of data points to the center of the clusters, aiming for compact, spherical groups, while DBSCAN groups data points based on closely they are packed together, making it appropriate for identyfing clusters of varying shapes and sizes based on density. This is why DBSCAN is able to discover the half-moon structure of the points in comparison to the K-Means Clustering. These differences should be considered when choosing an optimal clustering algorithm to group similar football passing sequences (possessions) together.

2.2.3 Visualization Methods of Multi-Dimensional Spatial and Temporal Data

Competitive sports data visualization is becoming a hot topic in the research field with many researchers proposing various useful tools that can help coaches and analysts find behavioural patterns. Various statistical information about how a footballer behaves on the field can help an analyst effectively identify the behavioural patterns including the overall contribution and degree of activity. However, competitive sports data contains multiple dimensions such as the space and time of the activities of each individual, making it more challenging for analysts to perceive the data intuitively since relying purely on numbers cannot fully represent the data analysis results, without taking into consideration the spatio-temporal dynamics (Du and Yuan, 2021).

In image 8, we can see that competitive sports data are classified as statistical information data and spatiotemporal information data. To effectively visualize and combine statistical information with their spatiotemporal occurrences, three key techniques are commonly used Du and Yan, 2020):

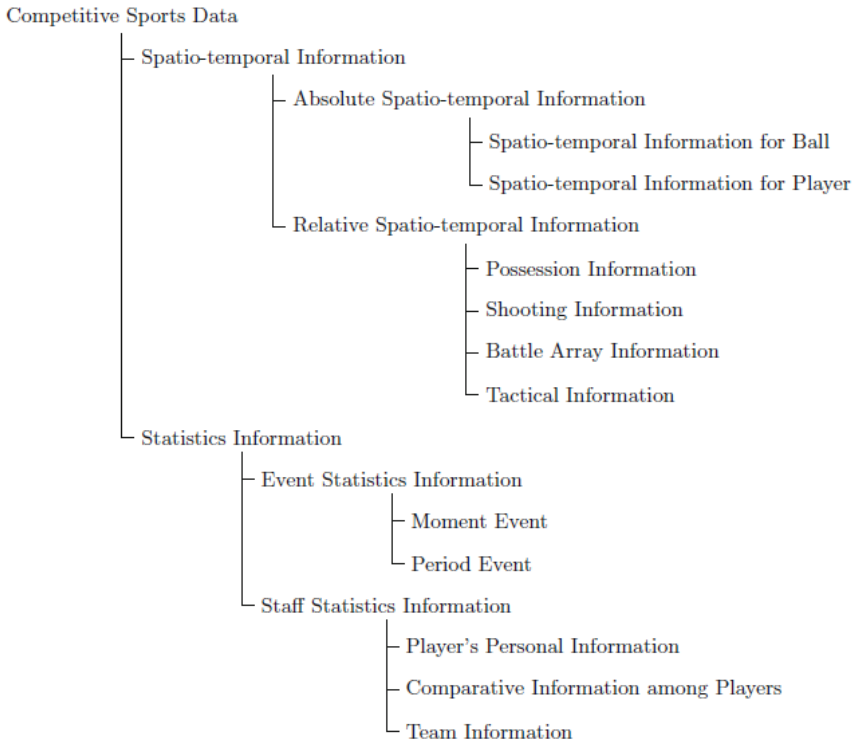


Image 8. Treemap that shows the structure and relationships of different pieces of information in competitive sports data

Image Source: Du and Yan, 2020.

A. High-dimensional data visualization techniques

These include techniques that explore the distribution patterns that indicate the relationships and influences in multiple dimensions. The main data visualization graphs include scatterplots, radar charts, parallel coordinate graphs and heatmaps. In our case where we aim to analyze the spatiotemporal intensity of passing activities by player, scatterplots show dimensional constraints and are not very effective when showing multiple dimensions simultaneously. Radar charts complement this by assessing player performance by comparing multiple attributes simultaneously and provide a very effective visualization graph for multi-dimensional data. Parallel coordinates are frequently used in large-scale data but are difficult to interpret due to dense lines overlapping. Therefore, Heatmaps were the most effective method for visualizing the distribution of passing events (Du and Yuan, 2021).

Choosing the appropriate heatmaps depends on the study objective and the eye movement measures that address these objectives. According to Bojko (2009), there are four main types of heatmaps, each offering a unique visualization perspective on the data.

- **Fixation Count Heatmap** which shows the accumulated number of fixations across participants.
- **Absolute Gaze Duration Heatmap** which also takes into perspective the accumulated time participants spent looking at the different areas of the stimulus.
- **Relative Gaze Duration Heatmap** which shows the accumulated time each participant spent fixating at the different areas of the stimulus relative to the total time the participant spent looking at the stimulus.
- **Participant Percentage Heatmap** which shows the percentage of participants who fixated on the different areas of the stimulus.

Since the goal of our study is to explore the frequency with which a player passes the ball in certain areas of the pitch, Fixation Count Heatmaps would be more effective in our case to highlight positions where passes are mostly concentrated. In contrast, Gaze Duration Heatmaps emphasize the total time a player focused on each area, while participant percentage heatmaps are more effective in showing how many different players made passes to each area. These findings are out of the scope of our goals since we are mainly interested in individual player movements and areas in which they pass the ball.

B. Time-series visualization

Time-series visualizations can display simultaneously the spatiotemporal information and various related statistical information. In our case, having the temporal information of the exact time that events occur during games, time-series visualization graphs such as line charts can be very useful in exploring the development of changing passing trends by identifying rising and falling curves in different time segments of games (Du and Yan, 2020).

C. Network Graph Visualization

As mentioned before, network graphs can't be only displayed statistically, but also dynamically by showing the spatial distribution of player and passing connections.

2.2.4 Non-negative Matrix Factorization Topic Modeling (NMF)

- **Defining Topic Modeling**

Topic Modeling is a clustering technique used in Natural Language Processing (NLP) to discover patterns in data by treating raw texts as tokens and then applying techniques such as

Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF). Andrienko N., Andrienko G., and Shirato (2023) refer to the use of topic modeling as a method to *«reveal patterns of co-occurrence of these tokens, which represent multi-attribute variations patterns in the context of episode-based data».*

Both Andrienko N., Andrienko G., and Shirato (2023) and Alfajri, Richasdy and Bijaksana (2022) mention that numerous evaluation studies have been made to compare the effectiveness of LDA and NMF in short-text data. Both highlight that NMF produces a more accurate rate when classifying text data into topics. These results influenced our decision-making as we decided to apply NMF to our short-text data describing football possession events.

- **Mathematics Involved in NMF Topic Modeling**

NMF is a method that decomposes a data matrix. The term Non-negative indicates the word representation matrix with the topics having non-negative values so the matrix is easier to interpret. This means that instead of feeding raw text to the topic modeling algorithm, the feature representations need to be transformed into a numerical scale. One very effective way of doing that, is to calculate the Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF provides a numerical way of taking into account the importance and relevance of different words within a collection of documents and can scale up words that appear in few documents but will help us more differentiate the topics. The calculation of TF-IDF is defined by the three equations below (Alfajri, Richasdy and Bijaksana 2022):

a. Calculate the value of TF by using equation (1) as follows:

$$TF(i,j) = \frac{freq(i,j)}{max_{ki,kj}}$$

b. Calculate the IDF value using equation (2) as follows:

$$IDF(i) = \log_2\left(\frac{N+1}{n_1+1}\right) + 1$$

c. Calculate the value of TF-IDF by using equation (3) as follows:

$$TF - IDF(i,j) = TF_{(i,j)} * IDF_{(i,j)}$$

Image 9. Equations of TF-IDF

Image Source: Alfajri, Richasdy and Bijaksana, 2022

NMF decomposition can be expressed in the following equation (Bala Priya, 2023):

$$V = W * H \quad \textbf{[7]}$$

Where Matrix **V** is a document word matrix that represents the document's text where each entry of the feature vector describes the number of words in the document. In other words, a Vij document-word matrix represents the number of times a word j appears in document i and is the product of W and H.

W is a weight matrix in which each row vector represents the vector of each word on the topic. In other words, a Wij word-topic matrix tells us how much each unique word i contributes to each topic j with a probability score.

H is a feature matrix in which each column vector represents the corpus of each document on the topic. In other words, a Hij topic-document matrix tells us how much each topic i contributes to each document j.

NMF topic modeling is a decomposition process that finds the matrices W and H so their product equals to V.

- **Evaluating NMF Topic Modeling Results**

A coherence score is a value that is used to measure the level of coherence of each topic. This can help us evaluate the uniqueness of each topic and how meaningful they are. Topic Coherence calculates the score of a single topic by measuring the level of semantic similarity between words, or in other words, how much these words appear in documents from the same topic and in the documents themselves. The score ranges from 0-1 with higher scores indicating better performance and more distinct topics. The most common measurement technique to evaluate NMF topic modeling is the c_v technique, which is defined by the below equation (Alfajri, Richasdy and Bijaksana, 2022):

$$NPMI(w_i w_j) = \sum_{j}^{N-1} \frac{\log \frac{p(wi,wj)}{p(wi).p(wj)}}{-\log p(wi,wj)} \quad \textbf{[8]}$$

Where **P(wi)** is the probability of a random occurrence of wi in the document.

P(wi, wj) is the probability of two words wi and wj appearing in the document at random.

N represents the total number of selected words w1, w2, … wn to calculate their occurrence.

3. Methods

3.1 Data Gathering and Preprocessing

3.1.1 Data Collection and Derivation Methods

To complete the objectives of this project, the data was obtained from the Figshare repository and specifically from the collection "Soccer match event dataset" (Version 5). Multiple datasets were extracted in JSON format and provide separate information about the detailed records of events from various football matches during the 2017-18 campaign. Specifically, 6 total datasets were obtained from Figshare (Pappalardo and Massucco, 2019).

- **Teams Dataset:** Includes information about each unique team. Arsenal is the team we are interested in.
- **Competitions Dataset:** Includes information about each unique competition that happened during the 2017-18 season. The competition we are interested in is the English first division.
- **Players Dataset:** Includes information about individual players.
- **PlayerRank Dataset:** Includes summary statistics of players and their role. This was merged with the players dataframe using the 'playerId' feature which was common in both dataframes.
- **Matches Dataset:** Includes information about the matches that happened during the 2017-18 season.
- **Events Dataset:** Includes detailed record of events in all football games during the 2017-18 season.

These tables were first filtered by keeping only records that consist of events in Arsenal games. After that, they were merged into a single dataset. To do that, we combined the tables one-by-one based on their common features and their relationships, according to Microsoft Support (2024). Matches, players and events were filtered out to match 'teamId' = 1609 which corresponds to Arsenal. The tables were merged using their one-to-many relationships, since the primary keys of the tables were foreign keys in the others. 'CompetitionId' was used to merge the competitions and matches tables, 'currentTeamId' was used to merge the players and teams tables, players and teams tables were merged by the 'currentTeamId' feature, 'matchId' was used to merge the events table with the already merged competitions-matches tables and finally 'player_Id' merged the already merged competitions-matches-events tables with the already merged players-teams tables.

This process helped us extract a single, combined dataset which included detailed information about the events, alongside the player information associated with each event, the matches and competition in which the events occurred, and the teams involved. A separate CSV file [1]provided information about the description of each 'tagId' which was used to map each tag to its description.

On-the-ball X and Y event coordinates were provided to associate each event with its origin and destination positions, each being a pair of coordinates (X, Y). Both X and Y coordinates are in the range of [0,100] and indicate the percentage of the field from the perspective of the attacking team. Since according to columns description of the dataset[2] the X coordinate indicates the event's nearness to the opponent's goal and Y indicates the nearness to the right side of the field, the X coordinates were mapped to a range of [0,120] and the Y coordinates to the range of [0,80] to represent the actual Matplotlib pitch coordinates according to Gozhulovski (2021[3]) who follows the STATSBOMB (X,Y) coordinate system.

New information on the events was extracted using derivation methods based on mathematical equations.

- **Event Length:** Defined as the Euclidean Distance [4] between the start and end points of an event, providing the direct length path.

$$\text{event\_length} = \sqrt{(end_x - start_x)^2 + (end_y - start_y)^2} \quad \textbf{[9]}$$

- **Event Angle:** Defined by the arctan2[5] function which calculates the angle between the positive X-axis and the vector to a point (X,Y), which is a very effective way to determine the direction of an event in Sports Analytics.

$$\text{event\_angle} = \text{degrees}(\arctan 2(end_x - start_x, end_y - start_y)) \quad \textbf{[10]}$$

- **Vertical Change:** Indicates the shift in position relative to the opponent's goal by calculating the distance between the end and start X coordinates of the event.

[1] https://figshare.com/articles/dataset/Mapping_of_tag_identifiers_to_tag_names/11743818?file=21385239
[2] https://figshare.com/articles/dataset/Events/7770599?file=14464685
[3] https://towardsdatascience.com/visualizing-football-game-data-6a124fab911b
[4] https://www.geeksforgeeks.org/euclidean-distance/
[5] https://www.slingacademy.com/article/numpy-understanding-tan-arctan-and-arctan2-functions-6-examples/

$$\text{vertical\_change} = end_x - start_x \quad \textbf{[11]}$$

- **Horizontal Change:** Indicates the shift in position of an event relative to the right side of the field by calculating the distance between the end and start Y coordinates of the event.

$$\text{horizontal\_change} = end_y - start_y \quad \textbf{[12]}$$

In image 10, we can see the X and Y coordinate values related to the areas on the pitch.

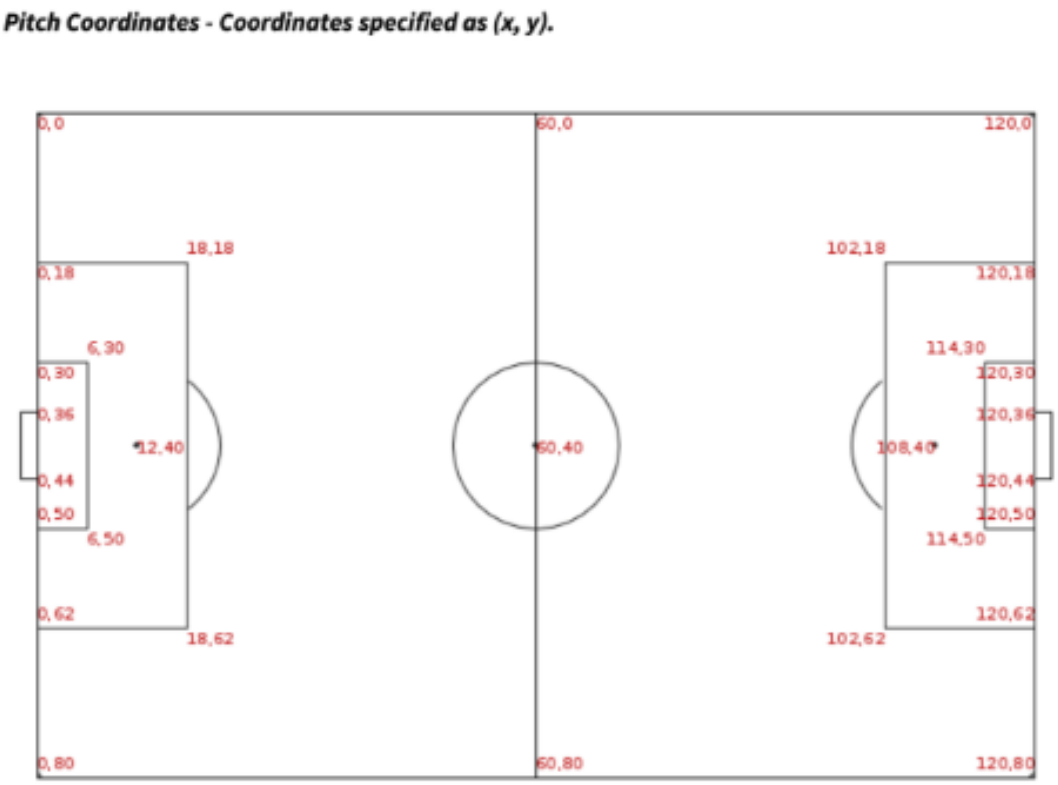


Image 10. STATSBOMB X and Y Pitch Coordinates

Image Source: https://medium.com/analytics-vidhya/how-to-create-football-pitches-goals-as-backgrounds-in-tableau-7b1a7800ae1c

3.1.2 Extracting Contextual Information of Events

Since the goals of this paper align with the identification and analysis of various patterns and dynamics in passing across multiple game scenarios, significant information was derived concerning the game context in which events occurred.

- **Home vs Away Games:** Events that occurred in home games were labeled as 1, while events that occurred in away games were labeled as 0.

- **Opponent Strength:** The opposition teams were grouped into 2 categories. Stronger teams that finished in the top 10 of the 2017-18 Premier League Season and weaker teams that finished in the bottom 10 of that season, according to Sky Sports[6].

- **Time Segment (mins**) Using the 'matchPeriod' and 'eventSec' features in the Events table, the events that occurred during the second half were transformed to show their timing from the start of the game by adding 2700 seconds (45 minutes), rather than their timing from the start of the second half. This helped us segment the timing of events into three categories. Events that occurred during the first 30 minutes, second 30 minutes, and last 30 minutes of games.

- **Current Score Difference:** Using the 'tags_description' feature of the events dataset, we were able to track down the current score of the game as the events occur.

- **Result:** Using the current score difference we were able to track down the result of the game which is defined by the current score of the last event. The current score values were encoded as 'Winning' for a Current Score Difference > 0, 'Drawing' for Current Score Difference = 0 and 'Losing' for Current Score Difference < 0.

3.1.3 Extracting Possession Episodes

Transforming event sequences into possessions provides us with a more structured view of Arsenal's passing dynamics. This approach highlights how Arsenal maintains control and orchestrates plays, offering insights into the effectiveness of different tactics and player contributions. The conditions under which we specified a new possession beginning were influenced by Stats Perform[7] and Andrienko N., Andrienko G., and Shirato (2023) who helped us define what a possession episode is and how the sequence of events should be split. If one of these conditions were met, this indicates that a new episode is starting.

[6] https://www.skysports.com/premier-league-table/2017
[7] https://www.statsperform.com/resource/introducing-a-possessions-framework/

- If the team of the following event changes using the 'teamId' column. This indicates that possession of the ball has changed even if temporarily. Even if the opposing team shortly interrupts the ball, we still want to extract a new episode to see how the team responds to the ball being intercepted. We are also interested in the outcome of each episode (last event) so even if an interruption is not successful, we still want to consider it as the outcome of the episode for our further analysis.
- If the match period changes. Using the 'matchPeriod' column indicating events that occur in different halves or in different games. 1^{st} --> 2^{nd} half indicates a different half, while 2^{nd} --> 1^{st} half indicating a different match.
- According to Andrienko N., Andrienko G., and Shirato (2023), a common problem when extracting episodes is the process of dividing the time span of a complex dynamic phenomenon described by multiple multivariate time series (MVTS) of event sequences into meaningful time periods that can enclose different relatively stable states or development states. To address this issue, using the 'time_segment (min)' column we created as explained above, to split possessions into 30-minute time segments.

The variables selected to describe the possessions include a comprehensive range match-related and player-related data. The match and team identifiers, along with the match period anchor each possession with a specific context. Temporal features such as the start, end and duration time of each possession provide a time-bound framework, while event-related variables that store the event names, sub events, and tags of all events associated with each possession into lists, give detailed insights into the actions during the possessions. Positional data of the events, including the starting and ending coordinates of the possessions, their horizontal and vertical amplitudes, the vertical and horizontal changes of all events associated with each possession, event lengths and angles were stored in lists and provide insights into the spatial dynamics of these possessions. Additional information such as the players involved in each possession, their positions and contextual details like home/away status, opponent team, time segment, current score when the possessions happen, and final score of the game that the possession occurs, enrich the description, making it a robust dataset for analysis. Finally, only possessions related to Arsenal with more than one event were extracted in the final dataset for analysis to make sure we don't mix episodes with ball possession of both teams. A total of 5948 unique Arsenal possessions from their games were extracted.

Summary Statistics of Possession Episodes		
Feature	**Statistic**	**Value**
Duration of Episodes	Total Possessions	5055
	AVG Duration (sec)	12.47
	MIN Duration (sec)	0
	MAX Duration (sec)	154.79
	MEDIAN Duration (sec)	7.51
Number of Events in Episodes	AVG Number of Events	5.49
	MIN Number of Events	2
	MAX Number of Events	51
	MEDIAN Number of Events	4
Vertical Amplitude of Episodes	AVG Vertical Amplitude	47.95
	MEDIAN Vertical Amplitude	42
Horizontal Amplitude of Episodes	AVG Horizontal Amplitude	40.42
	MEDIAN Horizontal Amplitude	39.2
Starting & Ending Coordinates of Episodes	AVG Starting Position X	55.26
	AVG Ending Position X	77.32
	AVG Starting Position Y	39.47
	AVG Ending Position Y	39.26

Table 1. Summary Statistics of Possession Episodes

Feature Name	Feature Description	Values	First Possession Example (PossessionId = 1)
matchId	Unique identifier for the match	Integer	2499719
teamId	Unique identifier for the team	Integer	1609
matchPeriod	Period of the match (e.g., first half, second half)	String (1H=First Half, 2H=Second Half)	1H
start_time	Time at which the possession starts	Float	2.758649
end_time	Time at which the possession ends	Float	12.548934
duration	Total duration of the possession	Float	9.790285
num_events	Number of events within the possession	Integer	6
events	Names of the events during the possession	List of Strings	['Pass', 'Pass', 'Pass', 'Pass', 'Pass', 'Pass']

sub_events	Specific sub-types of the events	List of Strings	['Simple pass', 'High pass', 'Head pass', 'Head pass', 'Simple pass', 'Simple pass']
tags_description	Descriptions of the tags associated with events	List of Strings	['Accurate', 'Accurate', 'Accurate', 'Accurate', 'Accurate', 'Not accurate']
eventsId	Unique identifiers for each event	List of Integers (Categorical Data)	[8, 8, 8, 8, 8, 8]
sub_eventsId	Unique identifiers for each sub-event	List of Floats (Categorical Data)	[85.0, 83.0, 82.0, 82.0, 85.0, 85.0]
players	Identifiers for the players involved in the possession	List of Integers	[25413, 370224, 3319, 120339, 167145, 3319]
player_names	Names of the players involved	List of Strings	['A. Lacazette', 'R. Holding', 'M. Ozil', 'M. Elsayed Elneny', 'H. Bellerin', 'M. Ozil']
position	Positions of the players involved (e.g., Forward, Midfielder)	List of Strings	['Forward', 'Defender', 'Midfielder', 'Midfielder', 'Defender', 'Midfielder']
opponent_strength	The strength rating of the opposing team	String	Top 10 Team
start_coords_x	X-coordinates where the possession starts	List of Floats	[58.8, 37.2, 61.2, 42.0, 49.2, 86.39999999999999]
start_coords_y	Y-coordinates where the possession starts	List of Floats	[39.2, 62.400000000000006, 60.0, 56.8, 76.0, 70.4]
end_coords_x	X-coordinates where the possession ends	List of Floats	[37.2, 61.2, 42.0, 49.2, 86.39999999999999, 92.4]
end_coords_y	Y-coordinates where the possession ends	List of Floats	[62.400000000000006, 60.0, 56.8, 76.0, 70.4, 60.0]
event_length	The length of each event during the possession	List of Floats	[31.699, 24.12, 19.465, 20.506, 37.619, 12.007]
event_angle	The angle of each event during the possession	List of Floats	[-42.955, 95.711, -99.462, 20.556, 98.561, 150.018]
vertical_change	Vertical movement of the ball during the possession	List of Floats	[-21.6, 24.0, -19.2, 7.199999999999996, 37.2, 6.0]
horizontal_change	Horizontal movement of the ball during the possession	List of Floats	[23.200000000000003, -2.4000000000000057, -3.199999999999996, 19.2, -5.599999999999994, -10.400000000000006]
time_segment (min)	The time segment in which the possession occurs (e.g., 0-30 minutes)	String (Categorical) (0-30=First 30 Minutes, 30-60=Second 30 Minutes, 60-90 Third 30 Minutes)	0-30
outcome	The result of the possession (e.g.,	String	Pass

	successful or unsuccessful)		
month	Month in which the match took place	Integer (Categorical)	8
home_game	Indicates if the match was a home game (1) or away game (0)	Binary	1
result	Outcome of the match (e.g., win, loss)	Integer (Categorical)	2
current_score_difference	Score difference at the time of the possession	Integer (Categorical)	0
vertical_amplitude	Vertical distance covered during the possession	Float	55.2
horizontal_amplitude	Horizontal distance covered during the possession	Float	36.8
starting_position_x	Starting X-coordinate of the possession	Float	58.8
starting_position_y	Starting Y-coordinate of the possession	Float	39.2
ending_position_x	Ending X-coordinate of the possession	Float	92.4
ending_position_y	Ending Y-coordinate of the possession	Float	60

Table 2. Feature Description of Possession-Episode Dataset

3.2 Identifying the Roles of Key Players That Influence a Team's Playing Style and Their Interactions With Other Players.

3.2.1 Studying the Roles of Individuals and the Connections Between Them in Distinct Game Scenarios

- **Calculating Passing Interactions to Study the Roles of Individuals and the Connections Between Them**

Network graphs can give a clear picture of Arsenal's passing patterns and dynamics and how the team operates on the field. Each node in the graph represents a player and each edge that connects two nodes represents a pass between two players. The thickness or color of the edge can indicate the strength of the relationships between the two nodes based on their total interactions. For our project, we extracted all passing interactions between players alongside the contextual information of the game which will be integrated into the passing networks to provide a more in-depth analysis (passing networks at home vs away games etc.). This symmetric dataframe considers the pass source and pass target, capturing the direction of the

passes. To estimate players' positions when they complete passing interactions, we calculated their average X, Y coordinates on the football pitch.

source	target	source_role	target_role	home_game	opponent_strength	time_segment	score_difference	source_position	target_position	current_score_grouped	normalized_weight_interaction	normalized_weight_interaction_by_role
A. Lacazette	R. Holding	ST	CB	1	Top 10 Team	0-30	0	Forward	Defender	Tied	0.331	0.12
R. Holding	M. Ozil	CB	LAM	1	Top 10 Team	0-30	0	Defender	Midfielder	Tied	0.335	0.14
M. Ozil	M. Elsayed	LAM	RCM	1	Top 10 Team	0-30	0	Midfielder	Midfielder	Tied	0.33	0.147
M. Elsayed Elneny	H. Bellerin	RCM	RWB	1	Top 10 Team	0-30	0	Midfielder	Defender	Tied	0.255	0.162
H. Bellerin	M. Ozil	RWB	LAM	1	Top 10 Team	0-30	0	Defender	Midfielder	Tied	0.192	0.149
H. Bellerin	G. Xhaka	RWB	LCM	1	Top 10 Team	0-30	0	Defender	Midfielder	Tied	0.158	0.119
G. Xhaka	S. Kolasinac	LCM	LWB	1	Top 10 Team	0-30	0	Midfielder	Defender	Tied	0.185	0.133
I. Monreal Eraso	P. Cech	LCB	GK	1	Top 10 Team	0-30	0	Defender	Goalkeeper	Tied	0.189	0.171
G. Xhaka	A. Oxlade-Chamberlain	LCM	RCM	1	Top 10 Team	0-30	0	Midfielder	Midfielder	Tied	0.286	0.118
A. Oxlade-Chamberlain	G. Xhaka	RCM	LCM	1	Top 10 Team	0-30	0	Midfielder	Midfielder	Tied	0.286	0.118
G. Xhaka	H. Bellerin	LCM	RWB	1	Top 10 Team	0-30	0	Midfielder	Defender	Tied	0.158	0.119
H. Bellerin	M. Elsayed Elneny	RWB	RCM	1	Top 10 Team	0-30	0	Defender	Midfielder	Tied	0.255	0.162
M. Elsayed Elneny	A. Lacazette	RCM	ST	1	Top 10 Team	0-30	0	Midfielder	Forward	Tied	0.326	0.126
M. Elsayed Elneny	G. Xhaka	RCM	LCM	1	Top 10 Team	0-30	1	Midfielder	Midfielder	Winning	0.242	0.118
H. Bellerin	M. Ozil	RWB	LAM	1	Top 10 Team	0-30	1	Defender	Midfielder	Winning	0.192	0.149

Table 3. First 15 Rows of Passing Interactions Dataframe

- **The Positional Frequency Bias Problem**

The analysis of player-passing interactions in football needs a nuanced approach that accounts for the variability in playing time among players. Simply counting interactions can lead to misleading conclusions since players with more game time naturally have more opportunities to pass the ball and therefore more passing interactions. Also, the versatility of player positions is another important factor when analyzing football interactions, especially in passing. For example, positions like right-back are typically less versatile because usually only one player occupies this role at a time, but positions like Left Center Midfield (LCM) can feature multiple players simultaneously, leading them to potentially be involved in more interactions. To address these issues, we are going to study the passing interactions of the 11 players with the most minutes played between them by position during the 2017-18 according to BDFutbol[8]. Also, we will be normalizing the data by calculating the interaction weight using the formula below:

$$\text{Interaction Weight} = \frac{1000}{(\text{minutes played by player 1} + \text{minutes played by player 2})} \quad \textbf{[13]}$$

This equation divides 1000 by the sum of the minutes played by both involved players for each interaction. This method ensures that the weight of each pass interaction reflects the actual involvement of the players in the team's passing strategy relative to their time on the field. This

[8] https://www.bdfutbol.com/en/t/t2017-182016.html

normalization allows us to evaluate the strength of the significance of player interactions more accurately, highlighting those who contribute effectively to the passing dynamics, regardless of their total minutes played. An example of the normalized interaction values is shown in Table 2 in the column 'normalized_weight_interaction'. The value of 1000 ensures larger and more distinguishable weights, making the passing network graphs more interpretable.

- **Adjusting the Node Sizes**

Zhou et al. (2023), mention in their paper the role of node significance as a key element when building an advanced network graph. Influenced by their adaptation of the PageRank algorithm to their methodology, we aimed to adjust the node sizes of the network graphs based on not only the quantity of their interactions, but also by their diversity and quality. To calculate the combined degree metric which takes into consideration all factors, for each node we calculated three unique metrics:

- Each node's **Interaction Quantity** by counting the total number of passing interactions in which the node was involved, either as the source or target of a pass.
- Each node's **Interaction Diversity** by identifying the number of unique players with whom a node interacts.
- Each node's **Strength of Interactions** is calculated as the sum of the weights of interactions with each unique player. This means that players who have many interactions with more influential players will be ranked higher.

These three metrics are combined to calculate the Combined Degree of each Node using the equation below:

Combined Degree=Interaction Quantity×log(Interaction Diversity+1)×Strength of Interactions **[14]**

This following equation that calculates the combined degree of each node is influenced by the PageRank indexing [2] since the core idea of both is that the importance of a node/player in a network graph is not just determined by the number of connections they have, but the quality and diversity of those connections. The PageRank equation factors a node's influence based on the overall weighted contributions of other influential nodes, which is similar to how the Combined Degree equation uses interaction quantity, diversity and strength to capture a node's significance to the network.

The reason to why we choose logarithmic scaling of Interaction Diversity is to dampen large variations so players with many diverse interactions would have a disproportionate influence on the combined degree. This creates a balance in the network, making it fairer by not letting players with a high diversity of weak interactions overpower those with a few, strong interactions (Robbins, 2012). Although in football, players in all positions will have interactions between them, certain player roles due to their positioning on the field will naturally have more diverse interactions with others. For example, an LCM will have more interactions than an RWB because they play more centrally and in closer proximity to other positions but might have weaker interactions than the RWB, indicating an LCM that does not pass the ball as much but an RWB that does but to fewer players. Logarithmic scaling minimizes the risk of these imbalances distorting the graph by fairly comparing the involvement of these positions.

The combined degrees were adjusted using a power law scaling to underscore differences between the node significances, using the formula of equations below:

$$\text{Normalized Degree} = \frac{\text{Adjusted Combined Degree} - \text{Min Adjusted Combined Degree}}{\text{Max Adjusted Combined Degree} - \text{Min Adjusted Combined Degree}} + \epsilon$$

$$\text{Scaled Degree} = (\text{Normalized Degree})^{1.5}$$

$$\text{Node Size} = 2000 \times \text{Scaled Degree} + 50$$

Image 11. Equations of Node size

We normalized the degree of the nodes because the differences between the values were small and it was difficult to differentiate the importance of players when first constructing the networks. So, to distinguish key players from others, we adjusted their sizes using the MinMaxScaler[9] theory to transform the range sizes between 0 and 1 with the highest node size being 1 and the lowest being 0. We also add a small constant of ε = 1*10^-9 to ensure non-zero values and make all nodes visible. We then applied Power Law Scaling[10] of 1.5 to exaggerate more the differences since it was still difficult to differentiate larger from smaller node sizes. Finally, after visualizing the network graphs, we noticed that the node sizes were too small and were not very visible, so we multiplied the scaled degree to 2000 and added a constant of 50 to increase the node sizes and make our graphs more interpretable.

[9] https://medium.com/@poojaviveksingh/all-about-min-max-scaling-c7da4e0044c5
[10] https://stackoverflow.blog/2011/07/21/power-laws/

- **Calculating Passing Interactions to Construct Passing Networks**

These steps underscore the careful consideration given when quantifying the contributions of each player in the network, not only by the volume of interactions, but also their strategic significance and diversity of connections. This methodology enables the visualization to reflect more nuanced insights into the dynamics of the team's passing game, aligning with the advanced network graph methodology discussed by Zhou et al. (2023). The same process was followed when calculating the edge weights and node sizes by each player position to construct a network graph in a 3-4-2-1 formation which was one of the most used formation during the 2017-18 season. Instead of showing the top 11 players by minutes, players were grouped based on their main role using the 'roleCluster' column. Multiple passing networks were constructed to show the differences in passing interactions in different game contexts.

3.2.2 Studying the Weights of Links Between Positions in Distinct Game Scenarios

Constructing difference graphs in the context of football analytics is important to understand better how teams alter their playing style, depending on the context, such as against top 10 vs bottom 10 opponents. Difference graphs highlight which specific passing interactions are more or less frequent and how the significance of different player roles changes in different contexts. This helps analysts to understand the strategic adjustments a team makes based on various conditions. In our case, we want to calculate the differences in proportions of each pass interaction. Because the count of pass interactions is different across multiple contexts, our aim is not to see the differences in counts, but how the distribution of pass interactions changes over different game contexts based on the total number of interactions in that context. For example, an interaction between two positions might occur more in home than away games because more passes are being made in general in home games but in away games the interaction is more frequent as a proportion to the total passes in away games. To address this issue an extract the proportional differences of interactions we used the below formula to normalize the interaction weights:

$$\text{Interaction Weight} = \left(\frac{\text{Normalized Weight Interaction by Role}}{\text{Total Pass Interactions in Context 1}} \right) \times 10,000 \qquad \textbf{[15]}$$

Where **Normalized Weight Interaction by Role** (Table 3) represents the normalized value of the pass intercations between two specific roles based on minutes played similarly to formula

[13] by dividing 1000 to the sum of the total minutes played between players who have either the source or target role.

Total Pass Interactions in Context 1 represents the total number of passes for a particular context (e.g. home game)

The multiplication by **10000** scales the result to give a larger, more interpretable value

To create the passing network difference graphs for two contexts (e.g. home vs away games), we needed to calculate the corresponding edge weight and node size differences of the two passing network graphs G1 and G2.

- **Edge Difference Calculation**

$$\text{difference\_edges}[(u,v)] = G1[u][v]['weight'] - G2[u][v]['weight'] \quad \textbf{[16]}$$

Where **u**, **v** represent two player roles that have more than one passing interaction (e.g. LB-LCM)

G1 and **G2**: These are directed passing network graphs that represent two unique game contexts (e.g. home vs away)

G1[u][v]['weight']: This is the total weight value of the edge from node u to node v in the graph G1 representing a unique passing interaction

G2[u][v]['weight']: Similarly, This is the total weight value of the edge from node u to node v in the graph G2 representing a unique passing interaction

Difference_edges[(u,v)]: This represents the difference in the weight value for a specific interaction u, v in the two Graphs G1 and G2.

The difference graph is plotted on a football pitch. The edge colours and widths represent the direction and magnitude of weight differences. This methodology helped us understand which passing lanes are significant when comparing the two contexts and how the involvement of each player's role in the passing network changes between the two contexts (home vs away etc.).

3.2.3 Factor Analysis of Passing Events

- **Preparing the Data for Factor Analysis**

Factor Analysis is an unsupervised learning method that allows us to transform complex passing statistics into underlying factors that represent core dimensions of a team's passing behaviour. This will help us associate each player with each factor to understand their role, how they contribute to the team's strategy and their overall strengths and weaknesses. The data that we are going to use for Factor Analysis can be segmented into three categories:

Types of Passes: Grouping the Event dataset by each player and keeping only events associated with passing, we counted the occurrences of each unique pass type in the 'subEventName' column which gives a detailed description of the pass type.

Description of Passes: Grouping the Event dataset by each player and keeping only the passing events, we counted the occurrences of each unique tag description in the 'tags_description' column.

Average Statistics of Passes: Grouping the Event dataset by each player and calculating the average length, angle, vertical change and horizontal change of their passes.

- **The Frequency Bias: Addressing Unequal Playing Time in Football Analytics**

To make a more accurate and equitable comparison of passing types and descriptions across players, regardless of their playing time or total passes made, we used percentages rather than the counts of each pass type. Since players have different amounts of time on the pitch their total pass counts can differ significantly making our Factor Analysis model biased towards players with more playing time. Our main interest is associating players with underlying factors based on the distribution rather than the volume of their passes. Using percentages is a way of normalizing the data, which means we can compare players with different amounts of playing time. This approach enables us to understand the specific tendencies and strengths of each player in the context of the team's overall passing strategy.

Further preparation of the data for our model includes checking the Collinearity and Communality scores alongside applying KMO and Bartlett's test to verify the suitability of the data and whether there is insufficient common variance to conduct a reliable factor analysis.

For the data to be suitable for Factor Analysis the below conditions need to be met (Quan, Li and Chen, 2024):

- ✓ Correlation Between two Variables < 0.7
- ✓ KMO Score > 0.5
- ✓ Bartlett's Test p-value Score < 0.05
- ✓ Communality Score of Each Variable > 0.5

Table 4 shows the input dataset of the factor analysis model.

Player	Cross %	Head Pass %	High Pass %	Launch %	Simple Pass %	Smart Pass %	% Pass Accuracy	Total Assists	% Total Counter Attack Passes	% Total Key Passes	% Total Dangerous Balls	% Total Final Third Passes	% Total Interceptions	% Total Through Balls	Avg Length of Passes	Avg Angle of Passes	Avg Vertical Change of Passes	Avg Horizontal Change of Passes
A. Iwobi	3	1	2	0	90	4	87	2	37	5	3	17	10	29	19.1	14.4	2.4	-1.8
A. Lacazette	4	3	2	0	86	4	80	4	28	22	0	15	6	29	16.5	1.2	-0.3	-0.4
A. Maitland-Niles	4	3	3	2	87	1	88	0	11	9	9	27	41	2	19.5	20.6	2.6	6.7
A. Oxlade-Chamberlai	14	1	10	1	75	0	78	0	5	10	0	60	20	5	27.8	19.3	5.1	-7.3
A. Ramsey	3	3	4	0	86	5	86	6	21	7	3	12	23	34	20.1	19.1	3.4	-0.1
A. Sanchez	4	2	12	0	71	11	76	3	17	16	1	17	1	47	20.6	18.1	5.4	8.9
C. Chambers	1	7	7	2	81	1	88	0	15	0	2	8	62	12	22.8	41.3	7.7	-4.4
D. Tackie Mensah Welbeck	1	6	0	0	89	3	86	2	35	13	0	3	23	26	16.5	-11.5	-1.3	2
F. Coquelin	0	6	9	1	84	0	87	0	14	0	29	0	57	0	21.1	17.6	6.1	1.7
G. Xhaka	0	5	8	1	85	1	89	3	26	2	3	5	34	31	22.5	35.2	6.3	-0.5
H. Bellerin	5	5	3	1	84	2	86	3	20	9	3	21	36	10	19.4	-7.2	-0.1	-9.3
H. Mkhitaryan	6	2	2	0	85	5	82	4	42	7	0	15	7	29	18.7	17.1	2.2	-3.2
I. Monreal Eraso	2	7	4	1	85	0	91	2	10	2	1	15	64	7	21.4	43.1	6.9	6.9
J. Wilshere	1	2	7	0	87	3	89	3	42	6	1	2	16	33	19.2	31.8	5.5	-1.5
K. Mavropanos	0	7	8	0	84	1	88	0	0	0	12	0	88	0	22	45.4	7.9	4.6
L. Koscielny	1	6	6	1	86	1	91	0	13	4	2	6	65	10	24.5	50.6	10.1	-3.9
M. Elsayed Elneny	1	2	4	0	92	0	94	1	18	3	3	5	59	13	19	30.1	3.9	-1.8
M. Ozil	3	1	3	0	89	4	88	6	29	13	2	17	9	29	19.1	24.7	3.4	-0.7
O. Giroud	3	16	3	0	73	4	77	0	0	10	0	10	40	40	14.1	-7.2	-1.2	0.6
P. Aubameyan	8	7	3	1	78	4	72	4	22	8	6	28	20	16	16.9	-7.8	-0.2	1.6
P. Mertesacke	0	8	0	3	89	0	94	0	12	0	6	0	75	6	24.5	69.4	10.9	1.3
R. Holding	0	6	7	1	85	1	87	0	14	0	7	0	74	5	24.6	49	9.6	-0.5
S. Kolasinac	5	6	4	2	83	1	85	4	25	7	6	17	30	15	20.4	11.5	1.5	9.8
S. Mustafi	1	8	7	1	82	1	88	0	13	3	3	8	58	15	25.1	53.5	10.9	-0.7
T. Walcott	19	0	0	0	76	5	78	0	0	0	0	50	0	50	21.7	4	-5.4	-9.6

Table 4. Factor Analysis Input Table

- **Applying Factor Analysis and Evaluating the Model**

According to (Quan, Li, and Chen, 2024), Scree Plots are effective in determining the optimal number of Factors. The optimal number is found when the eigenvalue drops below 1, indicating that additional factors are not contributing to explaining the variations in the data. In our case, 4 Factors seemed to be the appropriate number to split the data as seen in Figure 1.

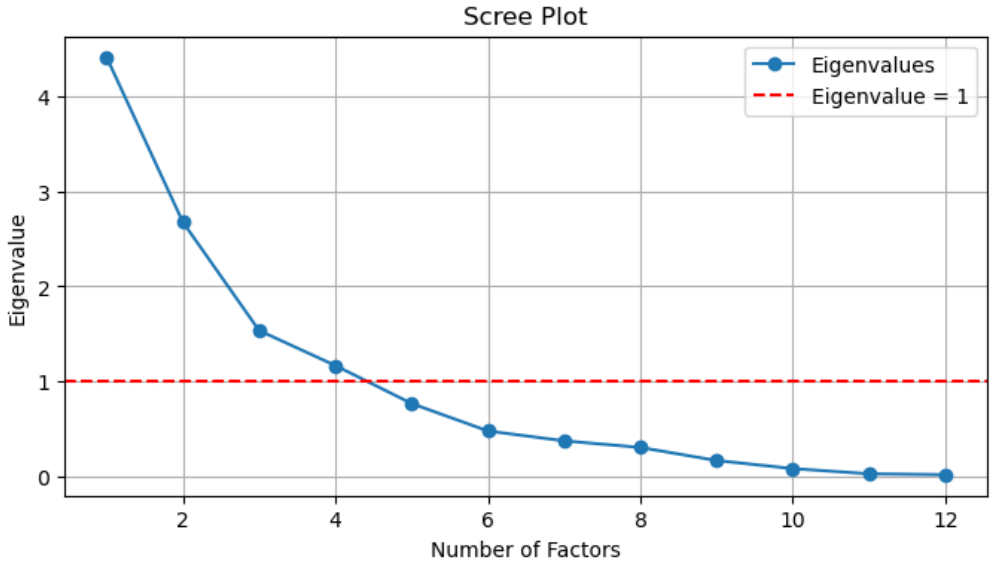


Figure 1. Scree Plot of eigenvalue across different number of factors

The further evaluation process involves calculating each player's total factor score by calculating the factor loadings to understand how the variables relate to the underlying factors. Next, we associate individual players by extracting their factor scores. Influenced by the formula used by Tavakol and Wetzel (2020) to study the magnitude of the factors, we calculated the variance contributions of each factor and derived their weights to prioritize and emphasize the factors that explain more of the data variability as specified in the equation [17]. Finally, we calculated each player's total factor score using a weighted sum formula [18], which combines their scores across all factors according to their respective weights.

$$\text{weights}_i = \frac{\text{variance}_i}{\text{total\_variance\_explained}} \quad \textbf{[17]}$$

Where **weights(i):** Represent the weight of the ith factor

Variance(i) is the variance contribution of the ith factor

Total_variance_explained: is the sum of the variance contributions of all considered factors.

$$\text{Total Factor Score} = W_1 \times F_1 + W_2 \times F_2 + W_3 \times F_3 + W_4 \times F_4 \quad \textbf{[18]}$$

W1, W2, W3, W4: These are the weights corresponding to each factor, derived from equation 17.

F1, F2, F3, F4: These are the individual scores of the player on each factor.

Players with higher total factor scores are generally considered more versatile in the team's passing strategy because they tend to have high scores across all factors, but this sometimes can lead to misjudgements as very high and very low factor scores can balance out. For this reason,

radar charts were created to simultaneously compare player performance across all factors and distinguish versatile from less versatile players. This process was influenced by the methodology outlined by Zhou et al. (2023) who used factor analysis to create a pentagonal evaluation model based on football player stats.

3.3 Analyzing the Positional Tendencies of Players Across Different Game Situations.

Even though network graphs can offer valuable insights into player positioning, to gain a deeper understanding of how an individual player's role and positioning may vary we illustrated player's average positions on the pitch. This method highlights the differences in positional tendencies under various match scenarios. Using data from Table 5 we calculated the average X and Y coordinates of passing events by player for different game contexts. Our goal is to identify any significant position changes in different game contexts and how they differ from their average positions on the pitch ('avg_start_x_base', 'avg_start_y_base').

fullName	avg_start_x _base	avg_start_y _base	avg_start_x _home	avg_start_y _home	avg_start_x _away	avg_start_y_ away	avg_start_x_ top10	avg_start_y _top10	avg_start_x_ bottom10	avg_start_y_ bottom1	avg_start_x _0_30	avg_start_y _0_30	avg_start_x _30_60	avg_start_y _30_60	avg_start_x _60_90	avg_start_y _60_90	avg_start_x_ winning	avg_start_y_ winning	avg_start_ x_tied	avg_start_ y_tied	avg_start_x _losing	avg_start_y _losing
A. Iwobi	73	44	74	40	72	47	70	45	76	43	74	46	73	44	70	40	70	36	75	51	73	40
A. Lacazette	81	43	83	44	79	42	80	42	81	43	80	42	81	43	82	43	81	44	80	43	83	41
A. Maitland-	62	28	63	27	61	28	62	29	62	27	62	29	64	28	60	26	60	33	62	21	64	36
A. Oxlade-Chamberlain	71	54	70	32	71	68	69	42	73	68	71	51	71	48	70	64	56	73	71	52	72	54
A. Ramsey	68	40	70	41	67	39	67	41	70	40	66	41	70	40	70	40	70	42	67	39	69	41
A. Sanchez	81	25	82	24	80	26	81	25	82	26	78	25	83	27	83	24	80	26	81	25	82	25
C. Chambers	48	53	46	41	50	63	48	59	49	49	49	52	45	52	51	55	44	50	48	52	56	56
D. Ospina	14	41	14	42	14	40	14	43	14	40	12	41	16	39	15	42	15	42	13	40		
D. Tackie Mensah	79	33	78	30	80	35	78	38	80	29	78	26	80	24	79	41	75	30	76	28	86	41
F. Coquelin	56	40	63	51	54	37	53	39	66	43	42	40	56	39	59	40	61	43	41	29	53	38
G. Xhaka	62	38	63	37	61	39	59	37	64	39	61	38	63	38	61	38	62	38	61	38	63	38
H. Bellerin	67	66	67	67	68	64	64	66	70	66	67	67	67	66	68	65	66	67	68	67	67	62
H. Mkhitaryan	73	50	72	55	74	45	72	49	75	50	73	51	74	49	73	49	72	52	71	47	75	49
I. Monreal	54	18	54	19	54	17	51	19	57	17	54	17	54	19	54	19	52	18	55	17	54	20
J. Wilshere	67	40	68	39	67	42	67	39	68	41	67	42	70	40	66	39	69	37	67	42	66	41
K. Mavropanos	41	27	43	24	38	30	41	27			41	27	42	27	38	27	43	24	39	28	40	31
L. Koscielny	47	50	46	52	48	48	45	49	48	51	46	51	48	50	47	49	45	54	48	50	48	46
M. Elsayed	62	44	62	44	63	43	62	45	63	43	63	41	62	44	63	48	61	44	63	43	66	46
M. Ozil	76	38	76	39	76	37	74	38	77	39	74	40	76	40	77	35	74	40	76	40	77	35
O. Giroud	78	36	75	36	81	35	79	34	78	36	71	32	75	42	80	35	73	36	79	37	84	32
P. Aubameyang	75	34	80	35	71	32	73	34	79	33	75	35	77	32	75	34	77	33	73	32	78	38
P. Cech	14	39	14	39	14	39	14	38	14	40	14	39	14	39	15	38	14	38	13	39	15	39
P. Mertesacker	42	40	45	41	40	38	35	38	45	40	42	37	47	43	38	40	37	42	40	37	47	41
R. Holding	46	39	50	47	43	33	44	40	48	39	44	39	47	40	47	39	49	43	45	34	44	43
S. Kolasinac	66	11	69	11	64	11	62	12	70	11	65	11	69	11	65	12	68	11	64	11	69	12
S. Mustafi	44	45	44	46	46	45	41	48	46	44	43	45	44	46	47	45	43	42	44	45	49	51
T. Walcott	80	70	86	69	72	72	90	67	77	71					80	70	83	70	97	68	76	70

Table 5. Average X, Y coordinates of passing events by player in different game contexts

Individual player passing heatmaps were also created to provide additional insights by illustrating the intensity and distribution of passes made by each player across different areas of the pitch. Unlike the positional scatterplots which focus on the average positions of passing, these heatmaps reveal the frequency and concentration of their passing events, highlighting the zones where each player is more active during ball distribution.

3.4 Identifying Distinct Passing Patterns and Dynamics Across Various Game Scenarios.

3.4.1 Density-Based Clustering to Group Similar Possessions: Separating Frequent From Occasional Possessions

When thinking about which method would more effectively group possessions, based on the comparison of DBSCAN and K-Means algorithms by Nandi (2022), we concluded that DBSCAN is more suited for football possessions due to the inherent nature of the game. DBSCAN effectively manages diverse shapes and densities of data points, meaning it can effectively separate frequent from occasional possession episodes that don't fit into a typical pattern, thus providing a more meaningful and accurate analysis of football possession dynamics. Football can sometimes be a random game so assigning every possession to a cluster can lead to misleading insights about possession types.

3.4.2 Preparing the Data

The 'First Possession Example (PossessionId = 1)' column of Table 1 can help us understand how the possession-based dataset is structured. Since Dimensionailty Reduction and Clustering algorithms require numerical input, we can't feed a list of string or integer values to our model. This means we first need to transform these lists into a structured format. So, we decided to standardize the data by counting occurrences of each unique value, for variables with lists of categorical data, and calculating the averages for columns with lists of numerical data, making the dataset suitable for algorithmic processing and analysis. For example, looking at the 'sub_events' of possessionId=1 (['Simple pass', 'High pass', 'Head pass', 'Head pass', 'Simple pass', 'Simple pass']), 'Simple_pass' occurs 3 times, 'High pass' 2 times and 'Head pass' only once. This means that possessionId=1 will have the value 3 for column 'Simple pass', 2 for 'High pass', 1 for 'Head pass' and 0 for any other unique value found in the 'sub_events' column. On the other hand, 'event_length' has a list of floats ([31.699, 24.12, 19.465, 20.506, 37.619, 12.007]), so we just calculating the average value of all items on the list. After this process was completed for the necessary columns, we standardized the data using the StandardScaler method, since the range of values across the columns was different and could lead to biased clustering. Standardization is a scaling technique where it converts the statistical distribution of the data into the below format (Mulani, 2022[11]):

$$z = \frac{x - \mu}{\sigma} \quad \textbf{[19]}$$

[11] https://www.digitalocean.com/community/tutorials/standardscaler-function-in-python

x is the original value of the feature

μ is the mean value of the feature

σ is the standard deviation of the feature

z is the standardized value of the feature

Now the entire dataset is scaled with a zero mean and unit variance (standard deviation = 1).

The problem with the methods we used above to transform the data into frequency counts is that we have increased the dimensionality of the dataset. Having many variables in our dataset makes clustering less accurate meaning we need to apply dimensionality-reduction. Nishida (2018) mentions in her paper that PCA can reduce the dimensionality of the data while maintain high variance and is a useful tool when classifying data with unsupervised learning. This is why the next step of our process was to reduce the number of features by at least 30% while maintaining around 90-95% variance. As a result, we were able to reduce the number of features from 23 to just 15 components while extracting around 95% of the cumulative variance.

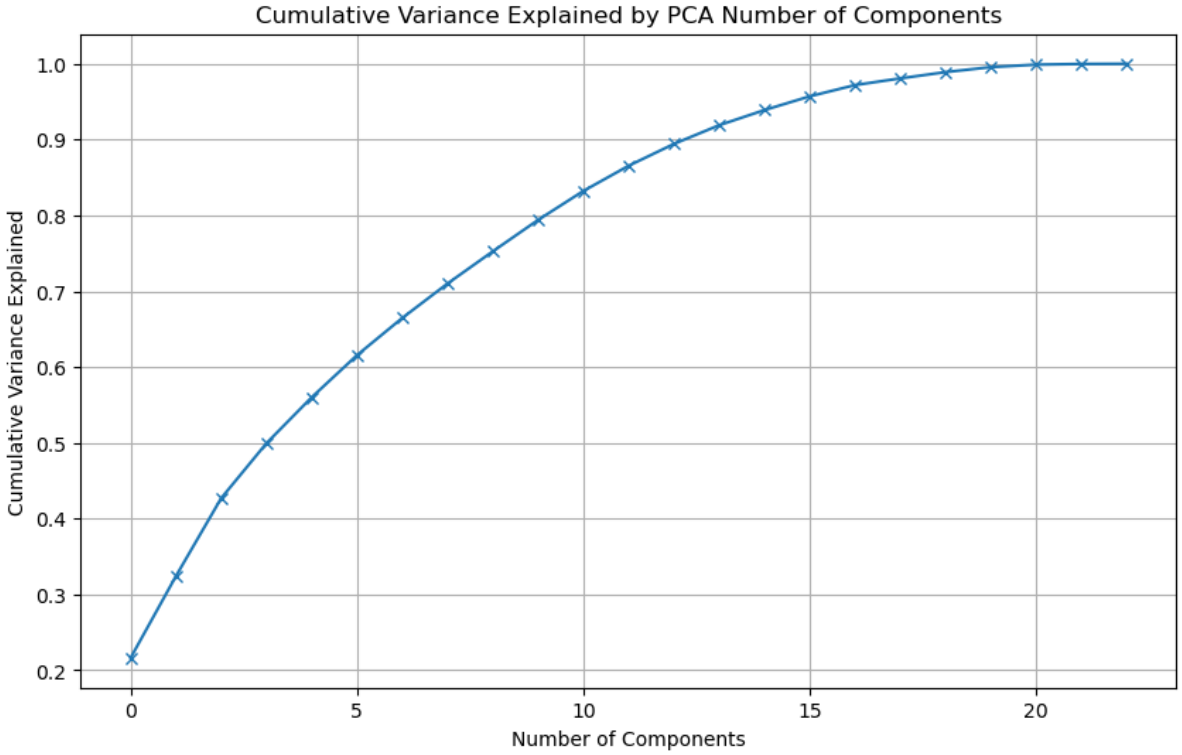


Figure 2. PCA Cumulative Variance Explained by Total Number of Components

3.4.3 Evaluating DBSCAN Algorithm

Nandi (2022) mentions that parameter sensitivity analysis is a crucial step when evaluating DBSCAN as the algorithm heavily depends on tuning the Epsilon (ε) and MinPts parameters. Mullin (2020)[12] demonstrates a very effective way of tuning these hyperparameters to separate noise from the rest of the data points using 'NearestNeighbors' library from 'sklearn' to plot the K-distance graph. Choosing the optimal 'MinPts' parameter mainly depends on domain knowledge and our familiarity with the dataset. Since our dataset is quite large and possessions tend to be rather different with diverse statistics, we chose a low number of MinPts=5 meaning that each cluster formed should have at least 5 similar possessions.

Now that we have selected our 'MinPts' value, we can move on and tune the optimal ε value. Following the methodology used by Rahmah and Sitanggang (2016), we plotted the K-distance graph to calculate the average Euclidean distance between each possession and its 4th nearest possession in ascending order using the formula of equations 20 and 21.

$$\text{distance}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \textbf{[20]}$$

Where **distance(x,y)** is the Euclidean Distances of two data points x and y.

$$K\_dist(p_i) = \text{distance}(p_i, \text{k-th nearest neighbor of } p_i) \quad \textbf{[21]}$$

Where **K_dist(pi)** is the K-distance of possession pi to the k-th nearest neighbor

The 'elbow point' shows the optimal value for the ε parameter (Rahmah and Sitanggang, 2016). In our case, Figure 3 indicates that the elbow point starts around 2-2.5. Since football possessions tend to have diverse statistics, we chose a low value of 2 to make sure that enough noise is detected and the clusters are meaningful. After finding the optimal parameters, DBSCAN was applied using the DBSCAN library from 'sklearn'.

[12] https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd

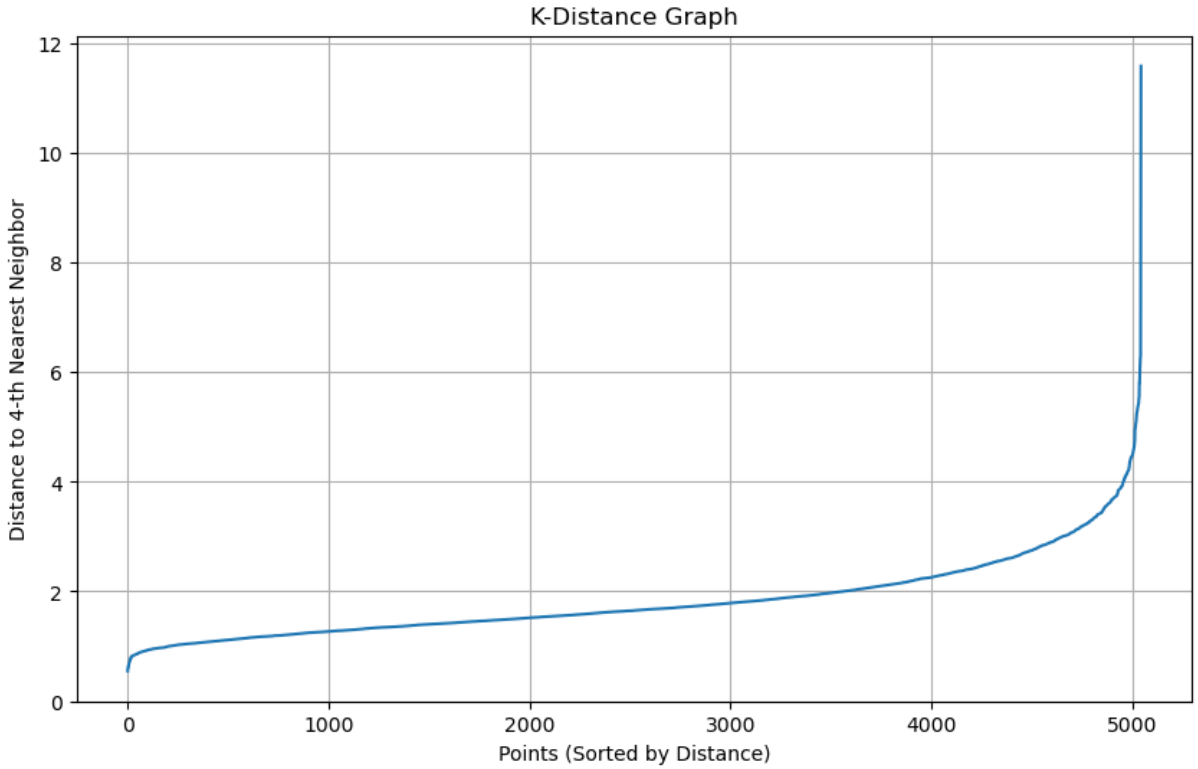


Figure 3. K-Distance Graph of Football Possessions

Around 30-35% of possessions were treated as noise which indicates that the parameters were not too strict but also could detect rare possessions and separate them from the clusters.

Looking at Figure 4, the clusters seem to be well separated across the Multi-Dimensional Scaling graph, indicating good performance. There is no overlapping between the clusters, which suggests that the ε parameter value of 2 was not too large.

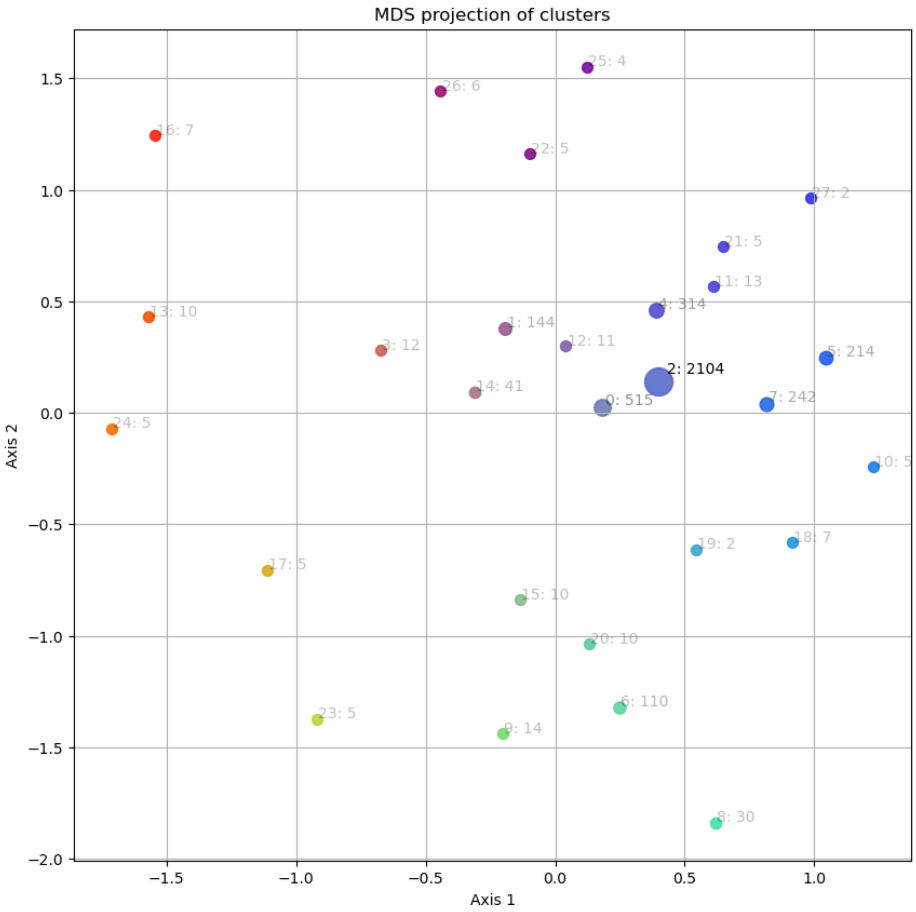


Figure 4. MDS 2D Projection of Possession Clusters

3.4.4 Comparing the Frequency of Passing Patterns Across Various Game Contexts

Our initial goal is to compare how often the most common passing patterns occur across various game scenarios. To achieve this, now that we have assigned possessions to clusters, we will be merging into a single dataset the cluster label of each possession alongside the contextual information we are interested in (Table 6). We will be comparing possession cluster occurrence at:

- **Home vs Away Games**
- **First 30 Minutes vs Second 30 Minutes vs Last 30 Minutes of Games**
- **Against Top 10 Opposition vs Against Bottom 10 Opposition**
- **When Winning During Games vs When Drawing During Games vs When Losing During Games**
- **Games Won vs Games Drew vs Games Lost**
- **Different Outcomes of Possessions (leading to shots, interceptions etc.)**

PossessionId	result	home_game	opponent_strength	time_segment (min)	current_score_grouped	outcome	Cluster
2	2	1	1	1	Tied	1	0
4	2	1	1	1	Tied	1	1
5	2	1	1	1	Tied	2	2
7	2	1	1	1	Winning	1	2
9	2	1	1	1	Winning	1	0
10	2	1	1	1	Winning	2	2
12	2	1	1	1	Tied	1	3
14	2	1	1	1	Tied	1	2
17	2	1	1	1	Tied	1	4
18	2	1	1	1	Tied	1	4
19	2	1	1	1	Tied	1	2
21	2	1	1	1	Tied	2	2
25	2	1	1	1	Tied	1	2
26	2	1	1	1	Tied	1	5
27	2	1	1	1	Tied	1	2

Table 6. Top 15 rows of possessions belonging to the Top 12 clusters by size alongside contextual information of games they occur

3.5 Uncovering Underlying Themes in Football Passing Styles That Reveal Distinctive Possession Strategies Across Various Game Situations.

3.5.1 Preparing the Data for NMF Topic Modeling

Both Andrienko N., Andrienko G., and Shirato (2023) and Alfajri, Richasdy and Bijaksana (2022) mention in their paper that Non-Negative Matrix Factorization (NMF) is a well-suited topic modeling algorithm for classifying short-text data into topics as it tends to be more accurate than other Topic Modeling algorithm. In our case, where we aim to uncover underlying themes in football possessions, we decided to apply NMF to our possessions dataset in which we have either short texts or numerical representations that describe passing attributes and statistics that characterize each possession (Table 2).

The first step was to encode the numerical representations of some data into categorical using basic Python functions[13]. Some examples to better understand this process can be seen in images 12 and 13.

[13] https://www.w3schools.com/python/python_functions.asp

```python
def encode_event_length(event_length):
    if float(event_length) <= 10:
        return 'Short_Pass'
    elif float(event_length) <= 40:
        return 'Medium_Pass'
    else:
        return 'Long_Pass'

df['event_length'] = df['event_length'].apply(lambda event_lengths: [encode_event_length(event_length) for event_length in event_
#We encode the possessions based on the length of the events

def encode_event_angle(event_angle):
    if -45 <= float(event_angle) <= 45:
        return 'Forward_Pass'
    elif 45 < float(event_angle) < 135 or -135 < float(event_angle) < -45:
        return 'Lateral_Pass'
    else:
        return 'Backward_Pass'

df['event_angle'] = df['event_angle'].apply(lambda event_angles: [encode_event_angle(event_angle) for event_angle in event_angle
#We encode the possessions based on the angle of the events

def encode_horizontal_amplitude(horizontal_amplitude):
    if horizontal_amplitude <= 42.0:
        return 'Compact_Play'
    elif horizontal_amplitude <= 60.0:
        return 'Balanced_Horizontal_Play'
    else:
        return 'Expansive_Width_Play'

df['horizontal_amplitude'] = df['horizontal_amplitude'].apply(encode_horizontal_amplitude)
#We encode the possessions based on the horizontal amplitude of the events

def encode_vertical_amplitude(vertical_amplitude):
    if vertical_amplitude <= 20.0:
        return 'Central_Play'
    elif vertical_amplitude <= 40.0:
        return 'Balanced_Vertical_Play'
    else:
        return 'Expansive_Vertical_Play'

df['vertical_amplitude'] = df['vertical_amplitude'].apply(encode_vertical_amplitude)
#We encode the possessions based on the vertical amplitude of the events
```

Image 12. Transforming Numerical Data into Categorical for NMF Topic Modeling

```python
def encode_start_x(start_x):
    if start_x <= 45:
        return 'Defensive_Third_Start'
    elif start_x <= 75:
        return 'Midfield_Third_Start'
    else:
        return 'Attacking_Third_Start'

df['starting_position_x'] = df['starting_position_x'].apply(encode_start_x)
#We encode the possessions based on the starting position of the X-axis

def encode_start_y(start_y):
    if start_y <= 15:
        return 'Left_Flank_Start'
    elif start_y <= 65:
        return 'Central_Zone_Start'
    else:
        return 'Right_Flank_Start'

df['starting_position_y'] = df['starting_position_y'].apply(encode_start_y)
#We encode the possessions based on the starting position of the Y-axis

def encode_end_x(end_x):
    if end_x <= 45:
        return 'Defensive_Third_End'
    elif end_x <= 75:
        return 'Midfield_Third_End'
    else:
        return 'Attacking_Third_End'

df['ending_position_x'] = df['ending_position_x'].apply(encode_end_x)
#We encode the possessions based on the ending position of the X-axis

def encode_end_y(end_y):
    if end_y <= 15:
        return 'Left_Flank_End'
    elif end_y <= 65:
        return 'Central_Zone_End'
    else:
        return 'Right_Flank_End'

df['ending_position_y'] = df['ending_position_y'].apply(encode_end_y)
#We encode the possessions based on the ending position of the Y-axis
```

Image 13. Transforming Numerical Data into Categorical for NMF Topic Modeling. X, Y Coordinates

The next step was to concatenate the texts into a single text corpus while maintaining their sequential format as seen in table 7.

Possession	Text Corpus
PossessionId = 1	Medium_Duration Medium_Activity Simple_pass High_pass Head_pass Head_pass Simple_pass Simple_pass Accurate Accurate Accurate Accurate Accurate Not_accurate Forward Defender Midfielder Midfielder Defender Midfielder Medium_Pass Medium_Pass Medium_Pass Medium_Pass Medium_Pass Medium_Pass Forward_Pass Lateral_Pass Lateral_Pass Forward_Pass Lateral_Pass Backward_Pass Tied Compact_Play Expansive_Vertical_Play Midfield_Third_Start Central_Zone_Start Attacking_Third_End Central_Zone_End Outcome_Pass
PossessionId = 2	Short_Duration Low_Activity Air_duel Head_pass Head_pass High_pass Won Accurate Interception Accurate Accurate Accurate nan Defender Midfielder Defender Short_Pass Medium_Pass Medium_Pass Medium_Pass Lateral_Pass Backward_Pass Lateral_Pass Lateral_Pass Tied Compact_Play Balanced_Vertical_Play Midfield_Third_Start Right_Flank_Start Attacking_Third_End Central_Zone_End Outcome_Pass
PossessionId = 3	Short_Duration Low_Activity Air_duel High_pass Lost Not_accurate Not_accurate Forward Defender Medium_Pass Medium_Pass Lateral_Pass Lateral_Pass Tied Compact_Play Balanced_Vertical_Play Attacking_Third_Start Central_Zone_Start Attacking_Third_End Central_Zone_End Outcome_Pass
PossessionId = 4	Short_Duration Low_Activity Simple_pass Launch Accurate Accurate Defender Goalkeeper Medium_Pass Long_Pass Lateral_Pass Lateral_Pass Tied Compact_Play Expansive_Vertical_Play Defensive_Third_Start Central_Zone_Start Midfield_Third_End Central_Zone_End Outcome_Pass
PossessionId = 5	Medium_Duration Low_Activity Free_Kick Simple_pass Ground_attacking_duel Accurate Accurate Free_space_left Won Accurate nan Midfielder Midfielder Medium_Pass Medium_Pass Short_Pass Backward_Pass Lateral_Pass Lateral_Pass Tied Compact_Play Balanced_Vertical_Play Attacking_Third_Start Central_Zone_Start Attacking_Third_End Left_Flank_End Outcome_Duel

Table 7. Text Corpus of First 5 Possessions

The final step of data pre-processing was to vectorize the data using the Inverse Document Frequency method (TF-IDF) as specified by Alfajri, Richasdy and Bijaksana (2022) and put into practice by Bala Priya (2023), to measure how often a passing term occurs in a possession and across all possessions. Passing terms that appear frequently across multiple possessions will have lower TF-IDF scores since they can help us distinguish the possession topics, while other words that are more unique to certain possessions and do not appear that often will have higher scores since they are more important in characterizing these possessions. This way, we can weigh down very frequent words that don't provide any uniqueness to the documents while scaling up rare ones that can be more effective in differentiating the topic of possessions. We also removed words that appear in less than 5% of possessions (min_df=0.05) and words that appear in more than 90% of documents (max_df=0.9) since these terms either don't really provide any information at all or are just too common amongst the possession-documents. Table 8 provides an example of the TF-IDF scores of four different terms for the first 5 possessions.

PossessionId	air_duel	attacking_third_end	attacking_third_start	backward_pass
0	0	0.130221195	0	0.123643145
1	0.333183609	0.129963904	0	0.123398851
2	0.426166912	0.166233615	0.252212524	0
3	0	0	0	0
4	0	0.159202314	0.241544512	0.151160299

Table 8. TF-IDF Scores of terms 'air_duel', 'attacking_third_end', 'attacking_third_start' and 'backward_pass' for the first 4 possessions

3.5.2 Evaluating NMF Topic Modeling Algorithm

Before applying NMF Topic Modeling we need to find the optimal number of components (topics) to classify the possessions. We first tried to calculate the reconstruction error for each number of topics using the Sparse[14] library in Python using the Forbenius Form of the reconstruction error but the results we got were not helpful in determining the optimal number of components. To address this issue, we used the 'c_v' topic coherence score [8] for various number of topics to tune the parameters and find the number of topics that extract the highest average coherence score across all their components. The c_v measure computes the pairwise word similarity score between the terms that belong to a topic (Alfajri, Richasdy and Bijaksana, 2022). In our case, we calculated the average coherence score for the top 10 words in each component. The number of components tested is in the range of 2-20 components. Table 9 provides the c_v scores for every number of topics. It seems that 8 topics have the highest average c_v coherence score indicating that splitting the possessions into 8 topics would extract the most meaningful topics.

[14] https://stackoverflow.com/questions/34123044/error-in-building-sparse-matrix-python-scipy-sparse

Average Coherence Score by Number of Components	
Number of Components (Topics)	c_v Score
2 Components	0.509
3 Components	0.514
4 Components	0.511
5 Components	0.557
6 Components	0.593
7 Components	0.585
8 Components	**0.603**
9 Components	0.574
10 Components	0.572
11 Components	0.561
12 Components	0.558
13 Components	0.549
14 Components	0.54
15 Components	0.535
16 Components	0.541
17 Components	0.557
18 Components	0.569
19 Components	0.537
20 Components	0.554

Table 9. Average 'c_v' Coherence Score of topics across different number of components

3.5.3 Visualizing the Topics and Comparing Their Frequency Across Various Game Contexts

Influenced by the methodology used by Bala Priya (2023) to visualize the topics, we used wordclouds to display the top 15 words in terms of their relative significance, with the most important terms having higher font size. After visualizing the topics and highlighting the most crucial terms that characterize them, we were interested in exploring under what circumstances these topics occur. This can reveal patterns and strategies that teams employ differently depending on the venue, opposition, or during different time segments. We will be comparing topic occurrences at:

- **Home vs Away Games**
- **First 30 Minutes vs Second 30 Minutes vs Last 30 Minutes of Games**
- **Against Top 10 Opposition vs Against Bottom 10 Opposition**

- **Games Won vs Games Drew vs Games Lost**
- **Frequency of Topics by Month**
- **Frequency of Topics by Gameweek**

3.5.4 Normalizing Possession Counts

When comparing the occurrences of each topic across different contexts, we need to take into consideration the difference in the total number of possessions between the two contexts. For example, since there are generally more Arsenal possessions in home games than in away games, simply comparing raw counts could lead to misleading results. To address this, we normalize the count of topics in each context using the below formula of equations:

$$\text{Normalized Count in Home Games} = \frac{\text{Actual Count of Possessions in Home Games for a Topic}}{\text{Total Number of Possessions in Home Games}} \quad \textbf{[22]}$$

$$\text{Normalized Count in Away Games} = \frac{\text{Actual Count of Possessions in Away Games for a Topic}}{\text{Total Number of Possessions in Away Games}} \quad \textbf{[23]}$$

This method gives us a proportion that represents how often a particular topic occurs relative to the total possessions in that context and allows for a fairer comparison of how the topics are distributed across the contexts, regardless of the differing total number of possessions. The same approach was utilized when also comparing the occurrences of each cluster across different contexts.

3.6 Finding the On-ball Passing Patterns and Themes That are Most Relevant to Success.

To assess the success rate of the underlying passing patterns and themes, measurements were based on two factors:

1. The frequency of each cluster and topic in **wins**, **draws** and **losses**.
2. Measuring the percentage of possessions leading to different **outcomes**, using the 'outcome' feature in the possession's dataset (Table 2). To be precise, we measured the distribution of each cluster leading to the following outcomes:
   - Non Accurate Passes
   - Duels
   - Others on the Ball
   - Shots

- Interruptions (Corner kicks)
- Offside

Outcomes such as ‘Non Accurate Passes’ (referenced as ‘pass’ but ‘not accurate’ in the ‘tags_description’ column) and ‘Duels’ indicate clusters that fail to create chances while ‘Shots’ can indicate successful clusters that lead to chances created.

3.7 Project Methodology Timeline

The timeline for applying the methodology was carefully structured to ensure a comprehensive understanding of both individual and collective behaviours on the pitch. We first focused on a more event-based analysis, where we examined the passing interactions between players, their positional tendencies and their playing styles through factor analysis. By first analyzing these steps at the micro level, we were able to gain a good understanding of how players operate. The whole process of our event-based lasted around 2 weeks (1st July – 14th July).

Once these individual and pairwise interactions were well understood, we transitioned our analysis to a more macro perspective, by grouping event sequences into possessions to explore the underlying patterns, dynamics and themes during ball control. The initial focus on event-level interactions and individual playing styles was crucial as it enabled us to better characterize the clusters and topics by understanding how specific players are involved based on what we already know of them. This approach helped us enrich the overall analysis of passing strategies and their implications on team performance. The macro-level analysis of the passing sequences required additional time (around 3 weeks, 15th July- 5th August), due to the extra preprocessing steps detailed in Chapters 3.4.2 and 3.5.1. Overall, the timing of each step in our methodology was essential to achieve our goals with a thorough analysis.

4. Results

4.1 The Roles of Key Players That Influence Arsenal's Playing Style and their Interactions with Other Players.

4.1.1 Comparison of Arsenal's Passing Interaction Networks Across Various Game Situations

The objective of this section is to compare the passing interactions between players in different game contexts to visualize how players are connected on the field and to associate each player with key factors that indicate their roles in Arsenal's passing strategy.

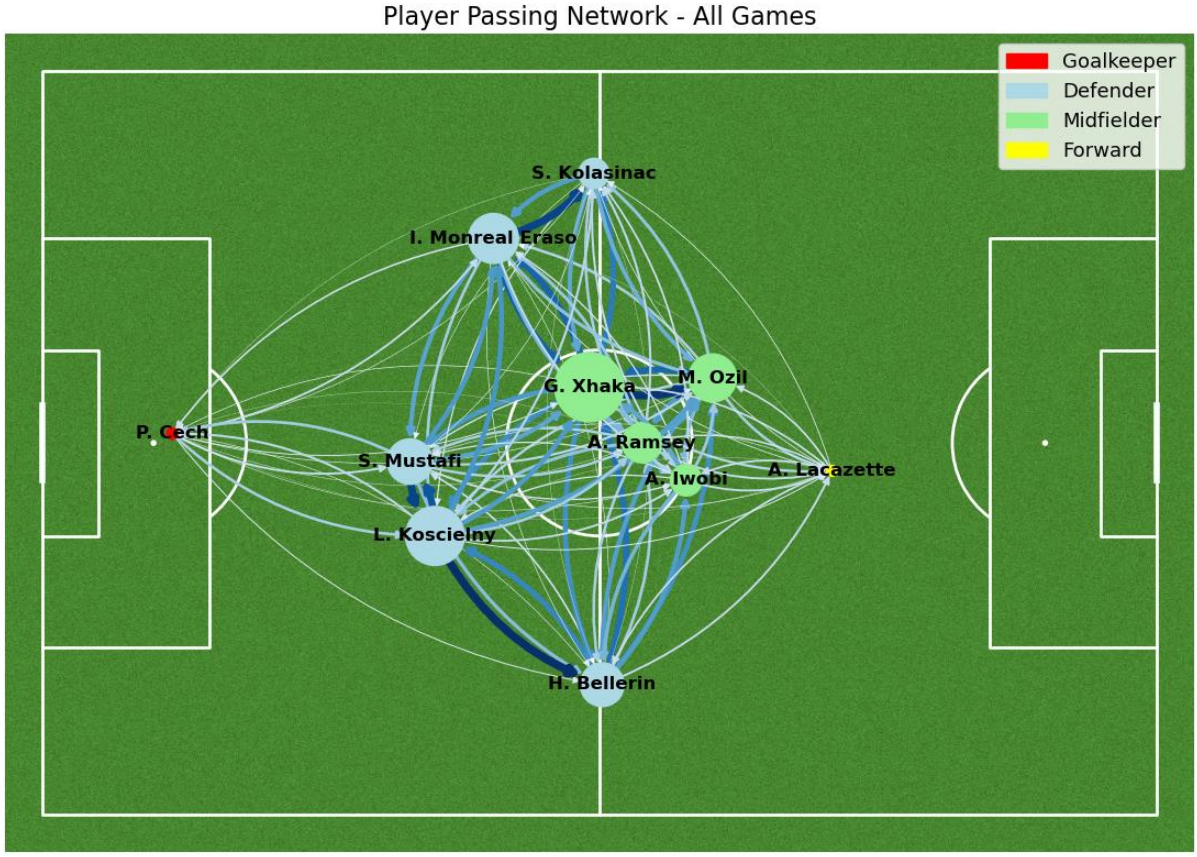


Figure 5. Player Passing Interactions Network – All Games

Figure 5 illustrates the Player Passing Network of the 11 players with the most minutes during the 2017-18 season. Arsenal's main formation involves the use of three center-backs in order for the wing-backs (Bellerin and Kolasinac) to be heavily involved in the attacking areas to provide width and support in creating chances. The centrality of Granit Xhaka in the network is immediately apparent, serving as a pivotal link between the defence and the attack. The presence of strong passing connections between the Wing-Backs and the central players indicates that once the ball is transitioned through the midfield, it is often directed towards the wing-backs who are positioned high up the pitch to exploit spaces on the flanks. The overall

structure reflects a team that compacts the midfield with the defense maintaining solid links to the midfield, allowing for the wing-backs to get into dangerous attacking areas and exploit wide spaces.

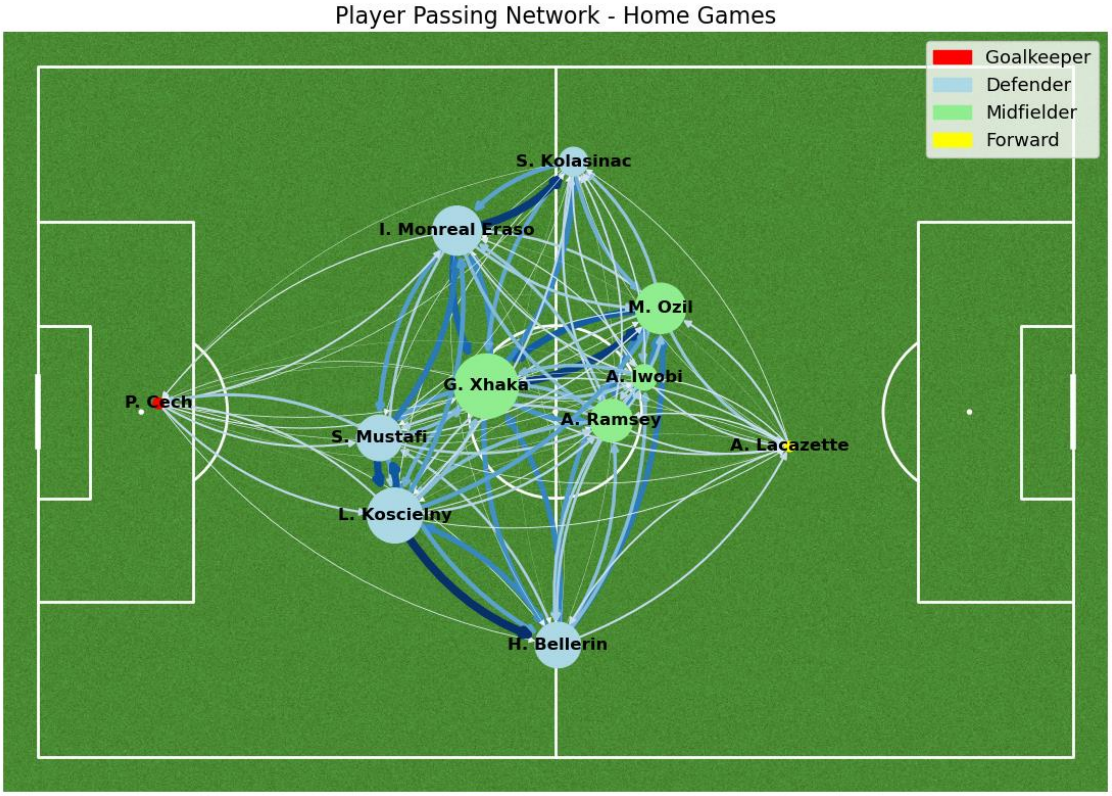


Figure 6. Player Passing Interactions Network – Home Games

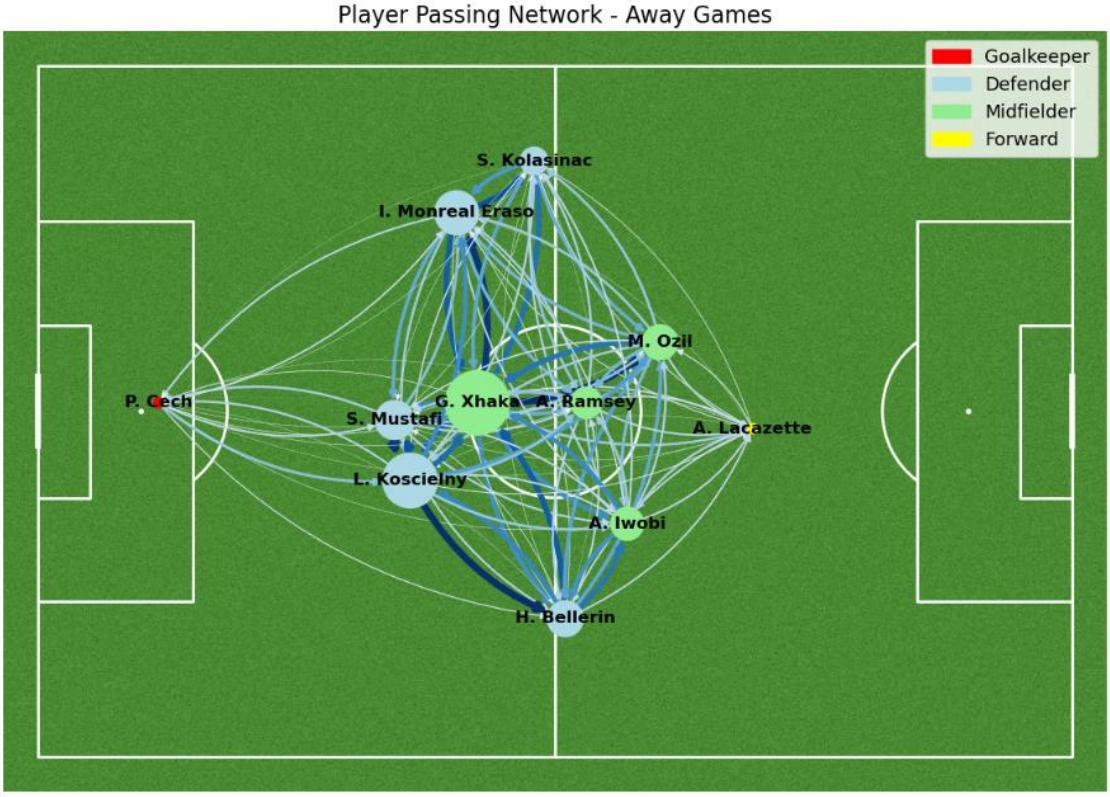


Figure 7. Player Passing Interactions Network – Away Games

When comparing the passing interactions between home and away games, it is evident that there are more frequent passing connections between players with players being generally more involved in passing as the nodes are bigger in home games suggesting that Arsenal is more confident and cohesive when playing at home and has more control of the game. In terms of the positioning of players, there do not seem to be major differences, but the positioning of A. Iwobi in away games which is more to the right side of the field and the stronger connections to H. Bellerin indicate that Arsenal tends to use more right-sided players to build up possessions in away games while the interactions are more equally distributed in Home Games.

When comparing the passing networks in games that Arsenal has played against the top 10 and against the bottom 10 teams, it is evident that the frequency of passing interactions is a lot higher against bottom 10 teams indicating that Arsenal does have a lot more possession in games with weaker opponents. In terms of the positioning of players, we can see that the wing-backs tend to get into more attacking positions against weaker opponents indicating their higher involvement in attacks.

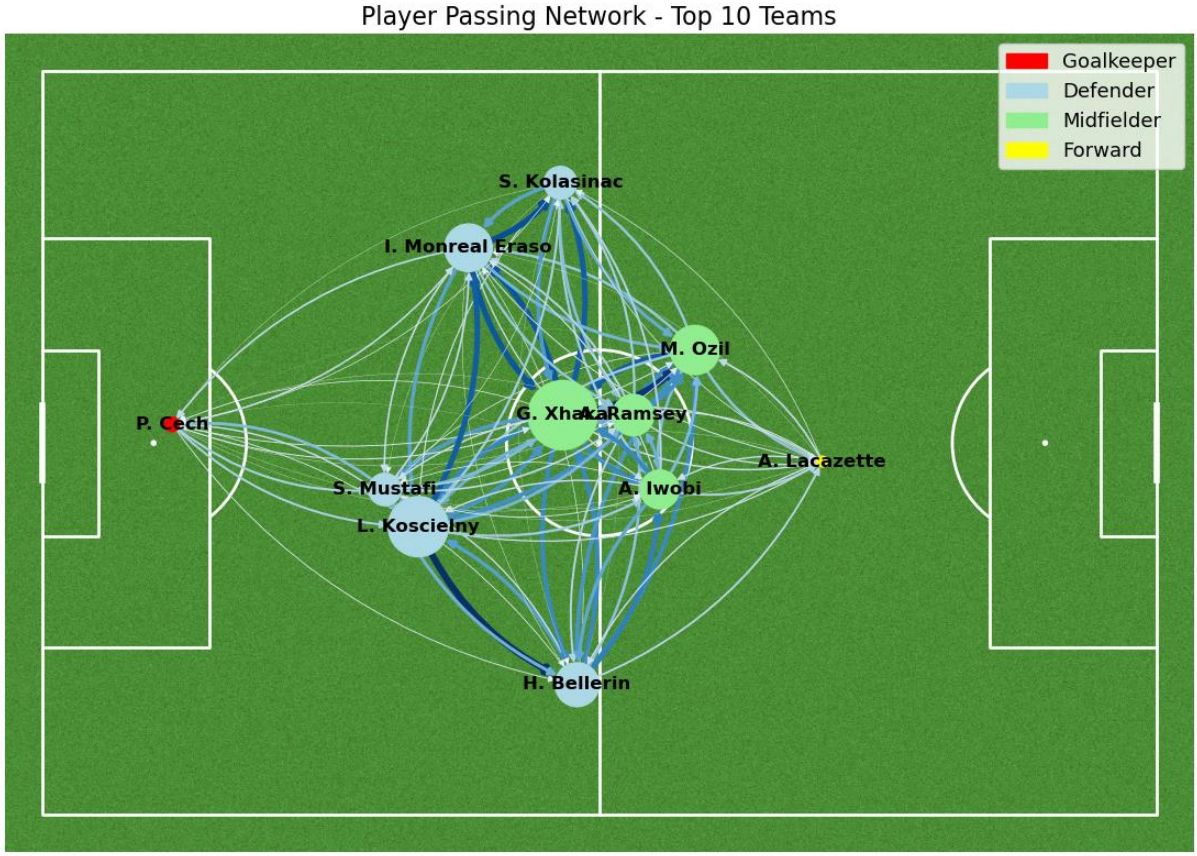


Figure 8. Player Passing Interactions Network – Against Top 10 Teams

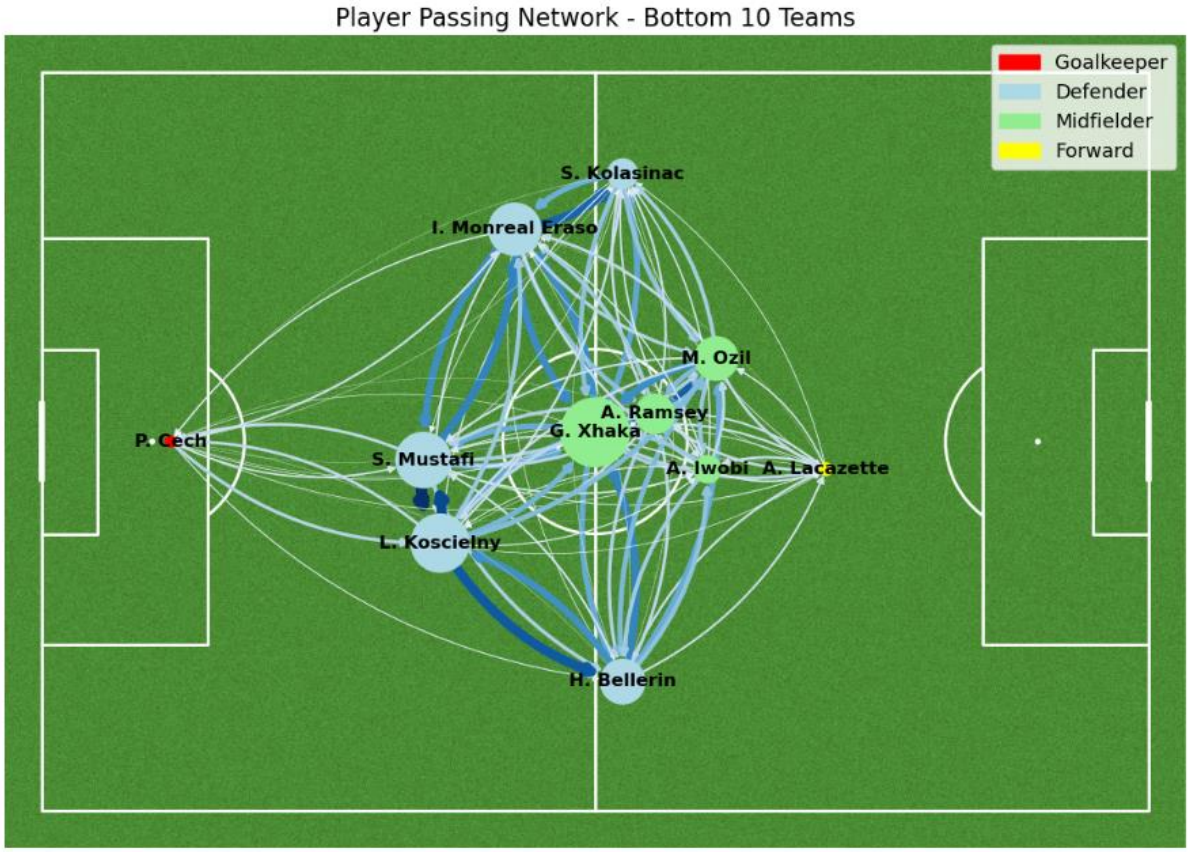


Figure 9. Player Passing Interactions Network – Against Bottom 10 Teams

One of our objectives was also to compare how the passing interactions and the positioning of players change in different time segments. Figures 10, 11 and 12, show Arsenal's passing network during the first, second and last 30 minutes of games. Overall, the passing networks seem to remain relatively consistent in terms of the positioning of players and the distribution of interactions. The network becomes denser during the 30-60 minutes of the game indicating more frequent passing interactions, especially between the central players during these periods of the game. A. Iwobi is more involved during the first 60 minutes of the game as he seems to be playing more as an attacking midfielder, drifting to the right to connect more with H. Bellerin. In the later stages of the game, Iwobi moves more centrally and is not as involved in passing, indicating a tactical change to his position to perhaps contribute more defensively in the central areas of the field.

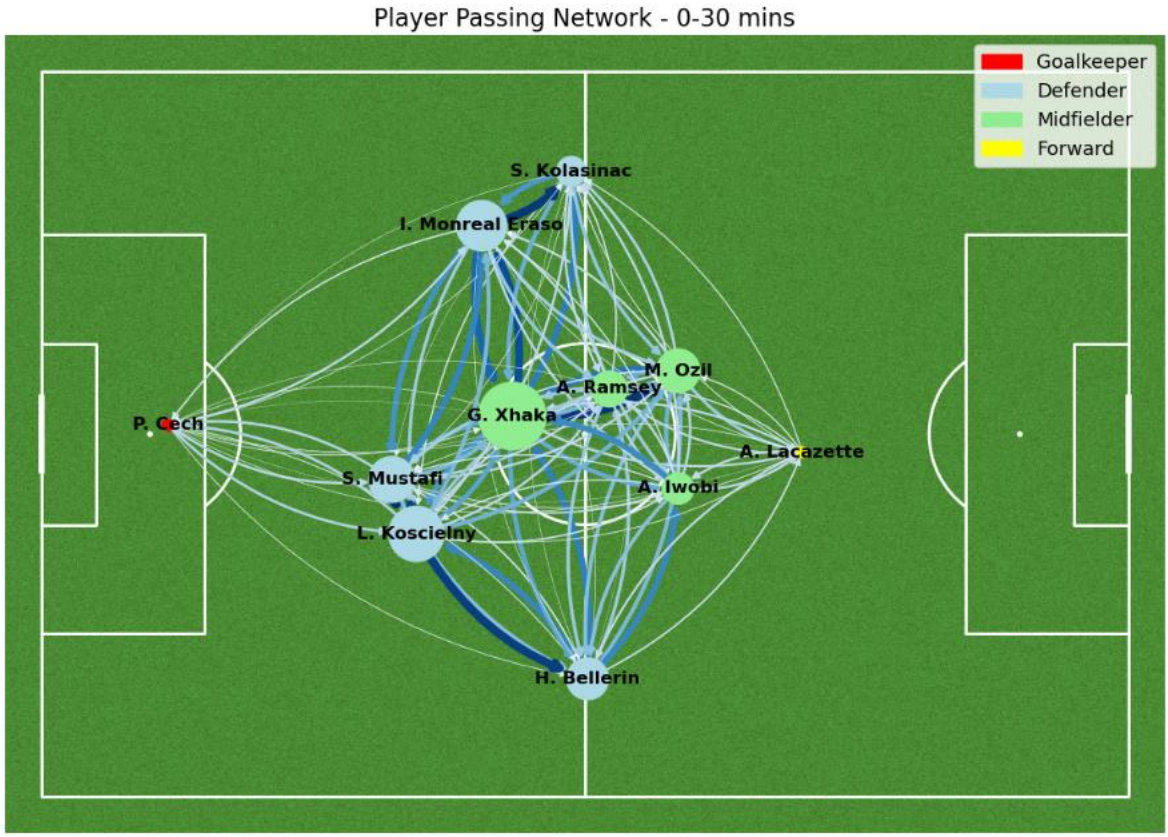


Figure 10. Player Passing Interactions Network – During 0-30 Mins

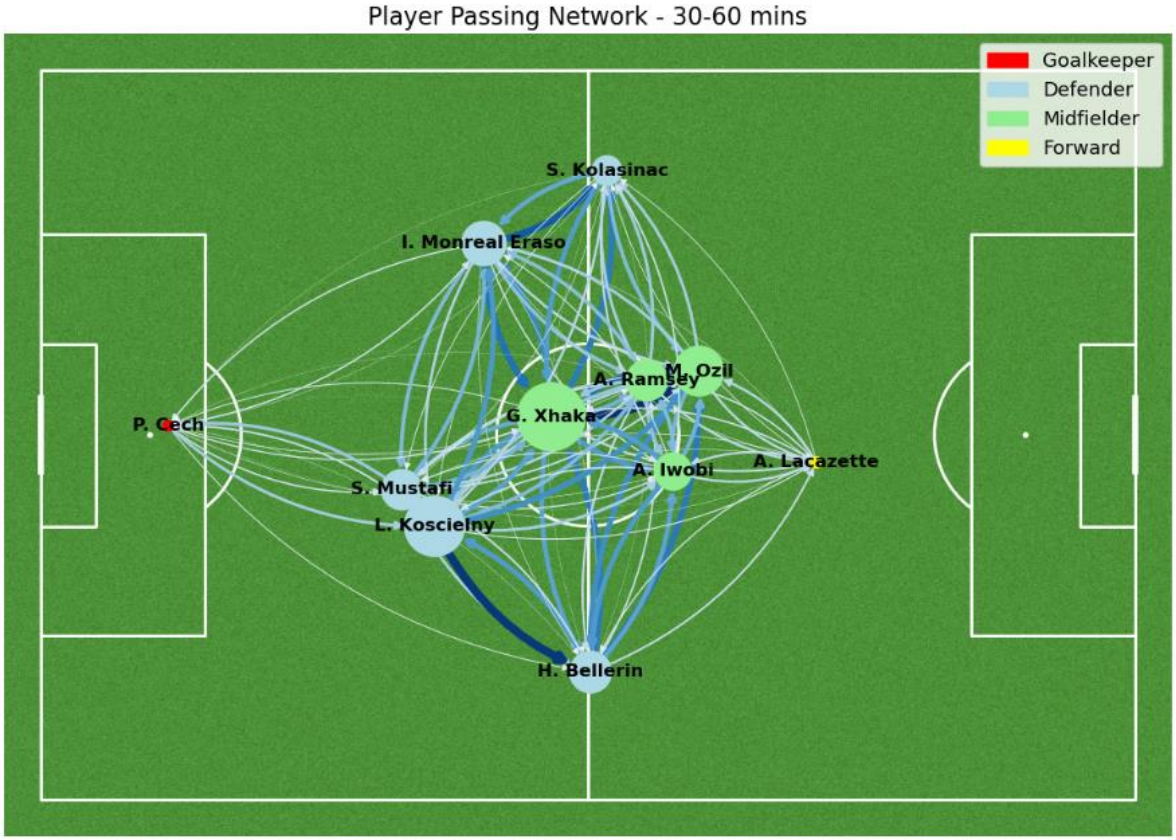


Figure 11. Player Passing Interactions Network – During 30-60 Mins

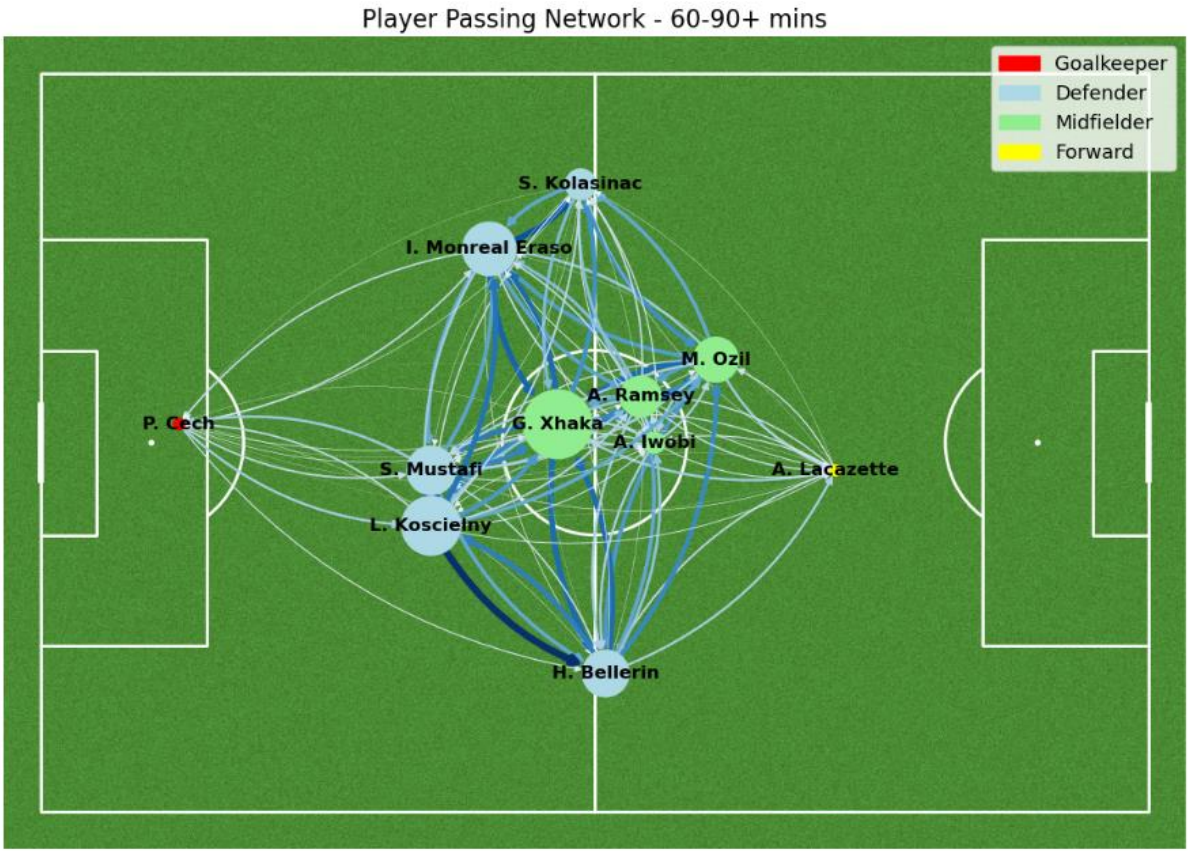


Figure 12. Player Passing Interactions Network – During 60-90+ Mins

Since Arsenal seems to maintain a similar structure and implement the same passing strategy across the whole game, comparing the passing interactions of situations where Arsenal is Winning during games, to when they are drawing or losing could provide a more insightful analysis.

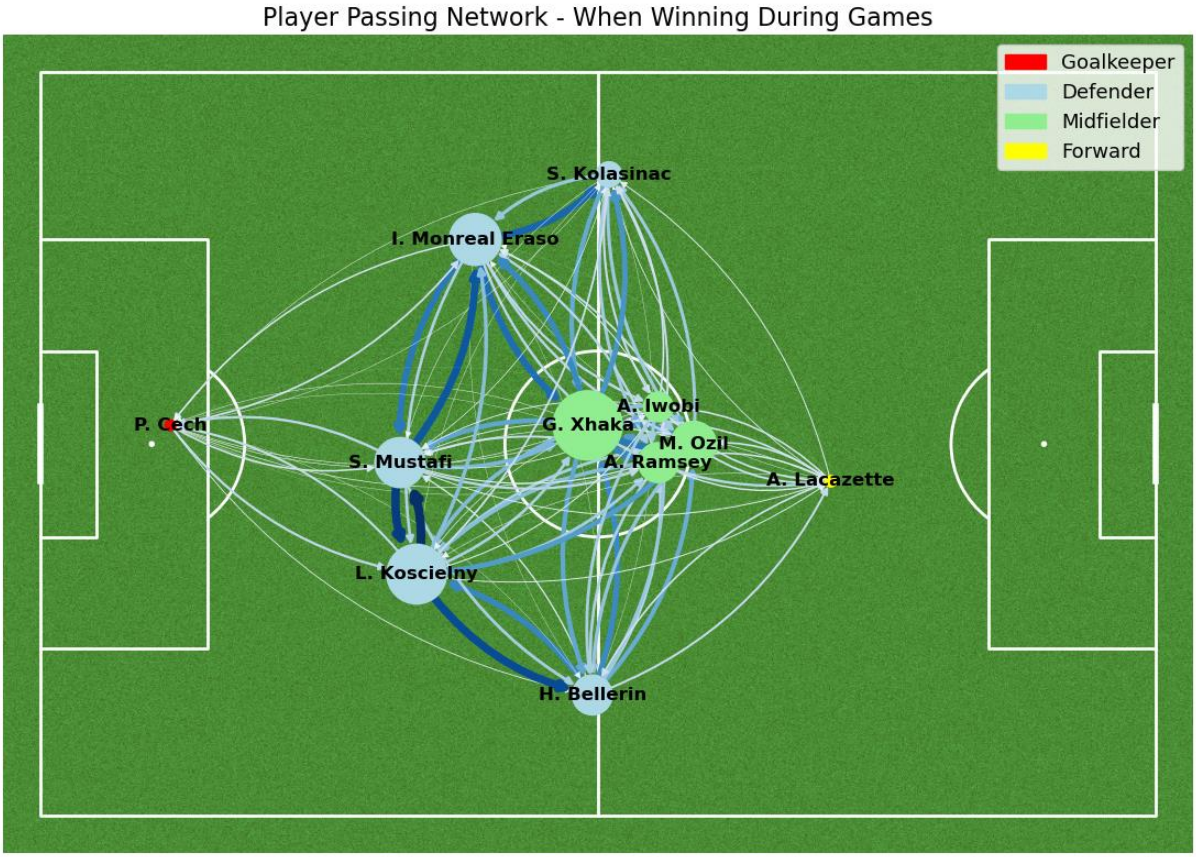


Figure 13. Player Passing Interactions Network – While Winning

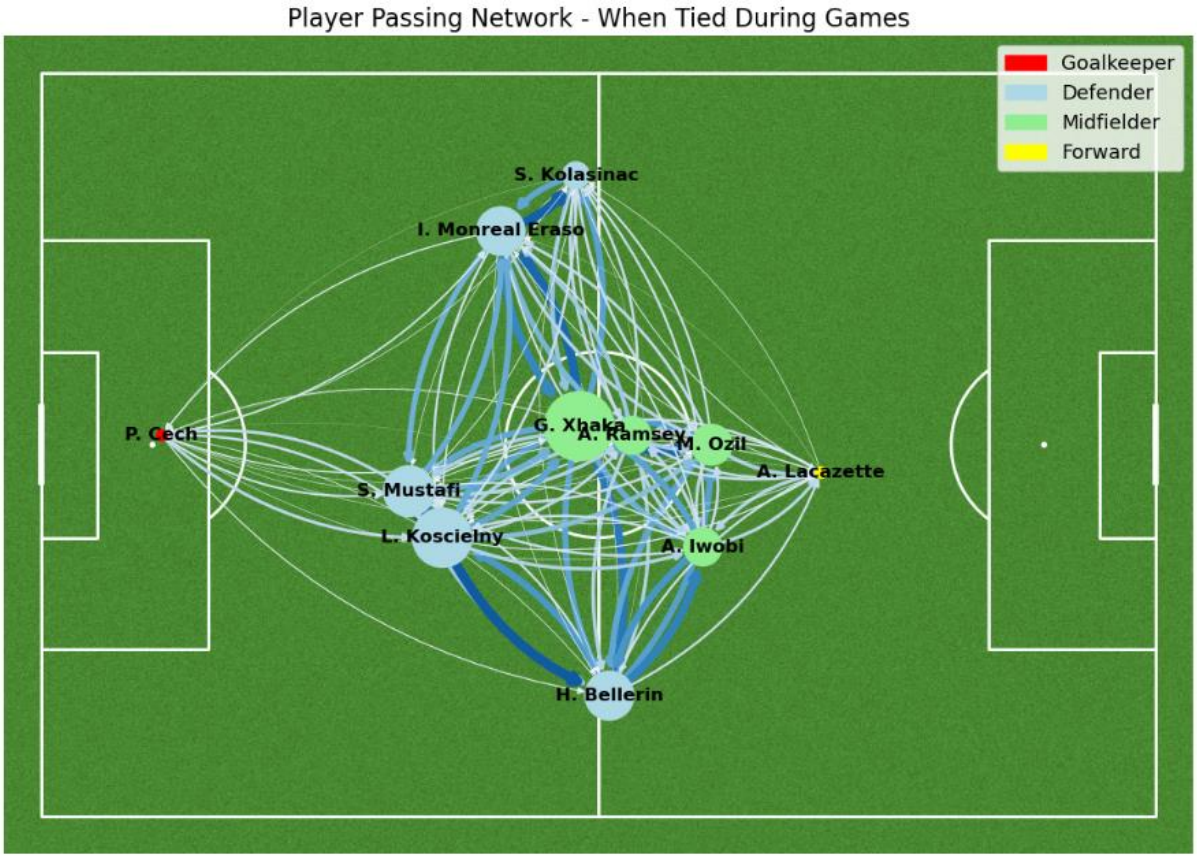


Figure 14. Player Passing Interactions Network – While Being Tied

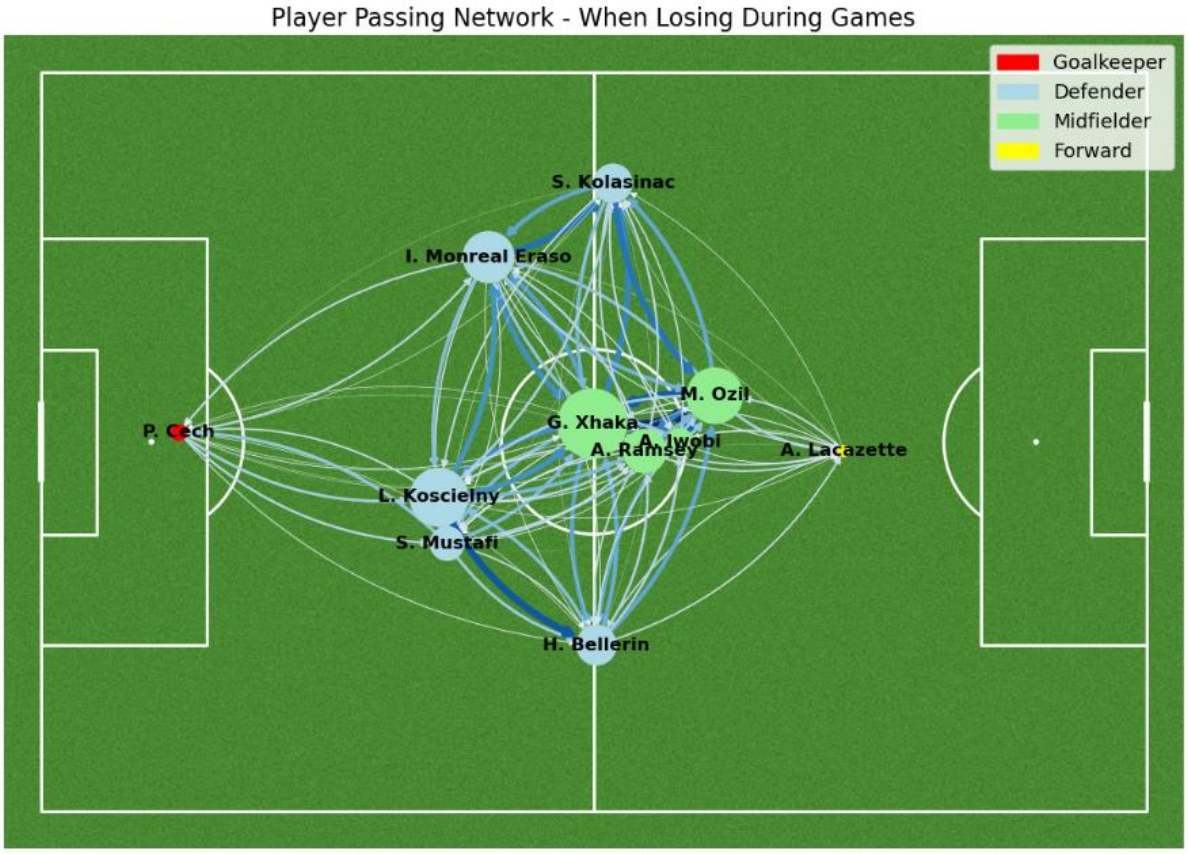
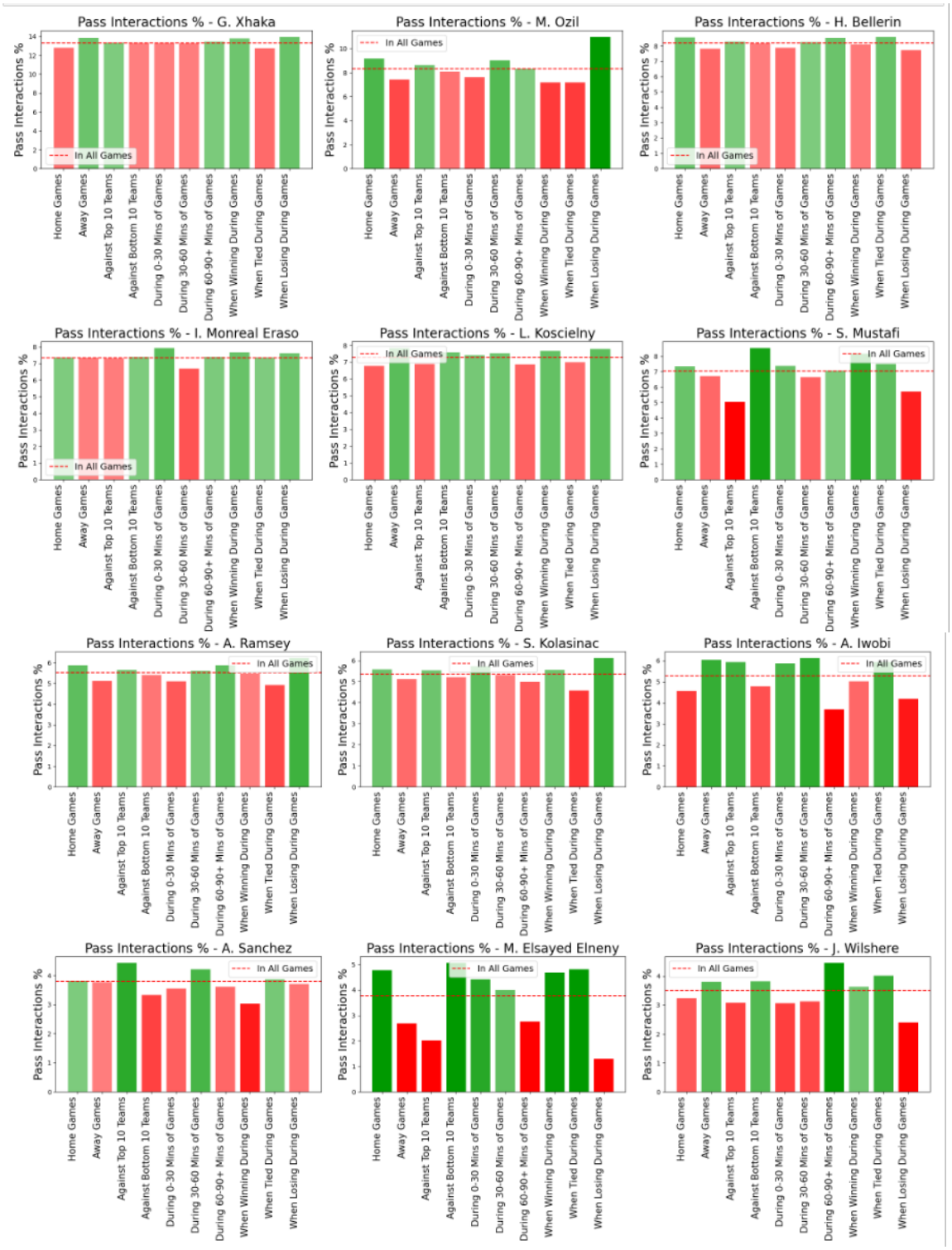


Figure 15. Player Passing Network Interactions – While Losing

Looking at figures 13, 14 and 15 we can see some significant differences in Arsenal's passing strategy. When winning, Arsenal seems to implement a more conservative play with a tendency towards more controlled and safer passes between the defenders and G. Xhaka who operates as a holding midfielder. There also seems to be less forward movement with connections to attacking midfielders and forwards being weaker. This indicates that when Arsenal takes the lead, the primary focus is to control the tempo of the game and distribute possession amongst defenders and midfielders. On the other hand, when Arsenal is tied, the networks become denser in the midfield and forward areas with players like Ozil and Iwobi getting more involved and moving into more attacking areas. There is also a slight shift towards the right side of the pitch with Bellerin having more interactions, indicating a strategy where Bellerin becomes more involved and attacking threats develop from the right flank. When Arsenal is losing, players like Kolasinac and Ozil get more involved and have stronger interactions with players near them. What is also interesting is that the four midfielders compact the central areas when Arsenal is winning but in the other two scenarios and especially when the game is tied, players like M. Ozil and A. Iwobi seem to move into wider areas, enhancing their interactions with the wing-backs. This indicates a strategy where the midfielders compact the central areas, either to

maintain possession more easily as midfielders are closer to each other or to protect these areas from opposition's attacks, while in situations where Arsenal needs to score a goal, attacking from the flanks is a priority.

To illustrate how the involvement of certain individuals is affected by the match conditions, a collection of bar graphs in Figure 16 shows the percentage of pass interactions for various Arsenal players under different game contexts.



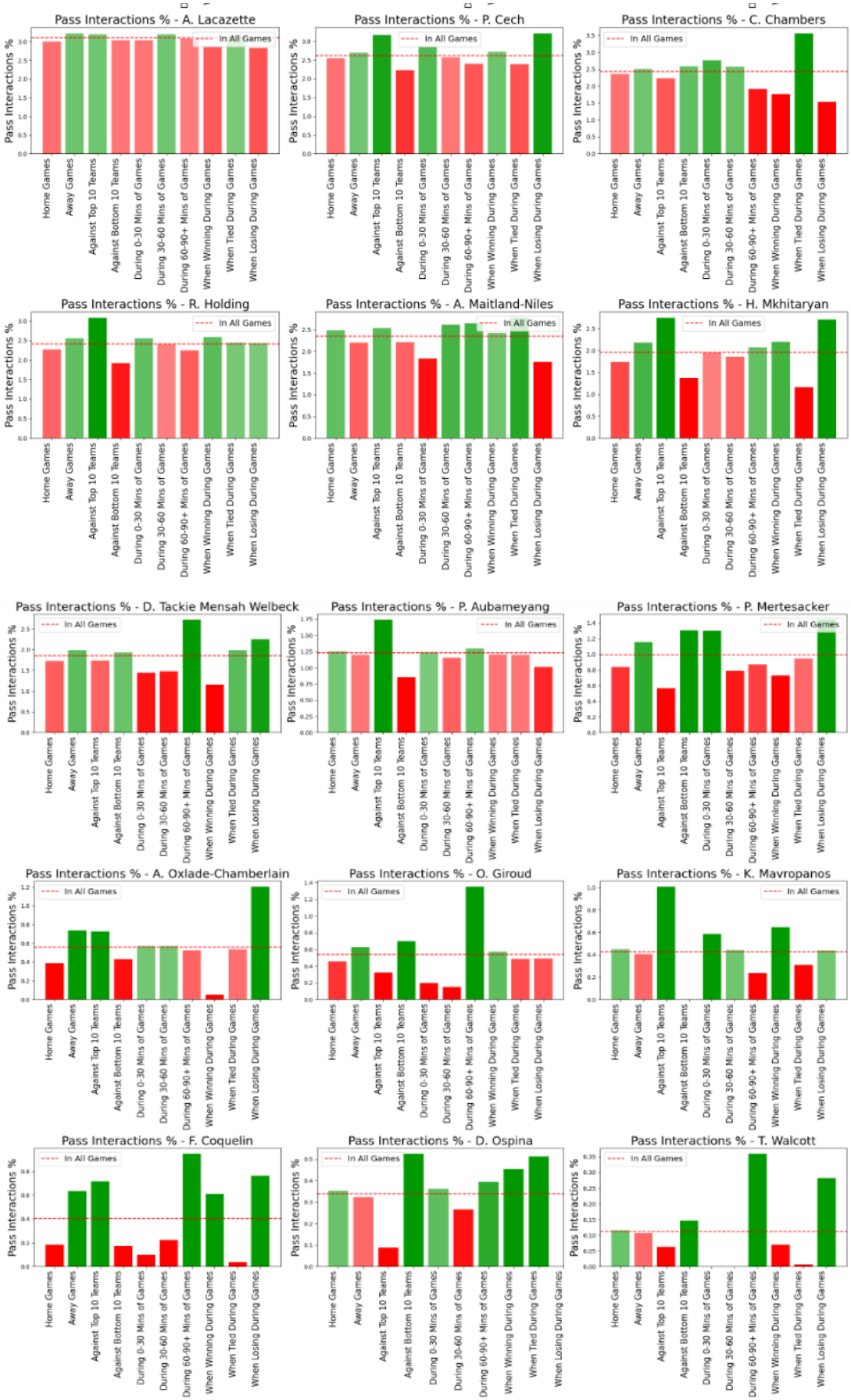


Figure 16. Pass % of players by game context

Notably, players like G. Xhaka and M. Ozil maintain a consistently high level of involvement across most conditions, reflecting their pivotal roles in Arsenal's playmaking, regardless of the game situation. Ozil's pass interactions significantly increases when the team is losing, indicating his pivotal role in driving the attack during crucial moments. On the defensive side, players like L. Koscielny and S. Mustafi show steady involvement across various scenarios but with more interactions against the bottom 10 teams and in situations where Arsenal is winning. This underlines that Arsenal's main goal is to distribute the ball across the center backs in these situation for more ball control and possession maintenance. The variability in participation for attacking midfielders like A. Iwobi and A. Ramsey, particularly in different time segments of games and when the team is winning, drawing or losing highlights their adaptive roles, switching between supporting defence and pushing forward. Players with lower overall interactions, such as O. Giroud, P. Mertesacker and F. Coquelin show greater spikes in specific conditions, reflecting their situational usage, likely in response to tactical changes or substitutions.

To add depth to the previous analysis, we created weight difference graphs to highlight the contextual differences in Arsenal's playstyle and the role differences by position in a 3-4-2-1 formation.

From Figure 17 we can notice that defenders have more passing interactions with various players to maintain possession and control the tempo of the game. This indicates that in away games, Arsenal tends to adopt a more cautious passing strategy, perhaps with forward players dropping deep and connecting more with the defenders. What is interesting is that a triangle of interactions between the RCB-LCM-RAM occurs more in away games while the RWB seems to receive the ball frequently from defenders and then pass to midfielders in home games. We can also see that many long passes from defenders to midfielders and forwards are being made in away games while in home games defenders tend to play more short passes between them and the wing-backs. This could indicate that more long-ball strategies are being used in away games while in home games Arsenal uses more build-up possessions.

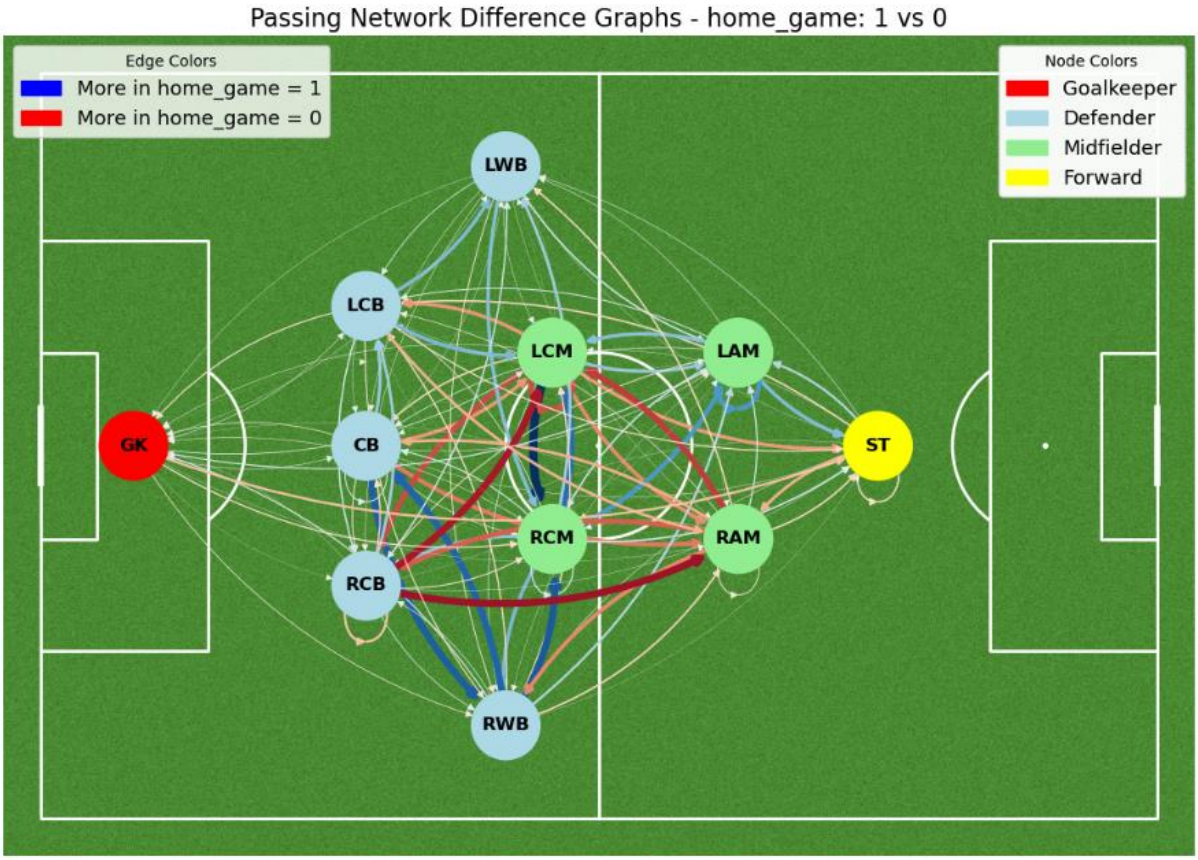


Figure 17. Weight Difference Graph – Home vs Away Games

Figure 18 indicates that against top 10 teams, there is a greater focus on distributing the ball through the central areas with more long, direct passes from the defenders to the attacking midfielders indicating more long-ball strategies being used against stronger opponents to minimize the risk of losing the ball in dangerous areas. On the other hand, against bottom 10 teams there seem to be more interactions between players closer to each other, especially between the defenders and the midfielders, indicating a different strategy with more short passes to build-up possessions more effectively. Attacking players seem to be more involved against the stronger teams due to them receiving the ball more often from defenders who operate as long-range playmakers.

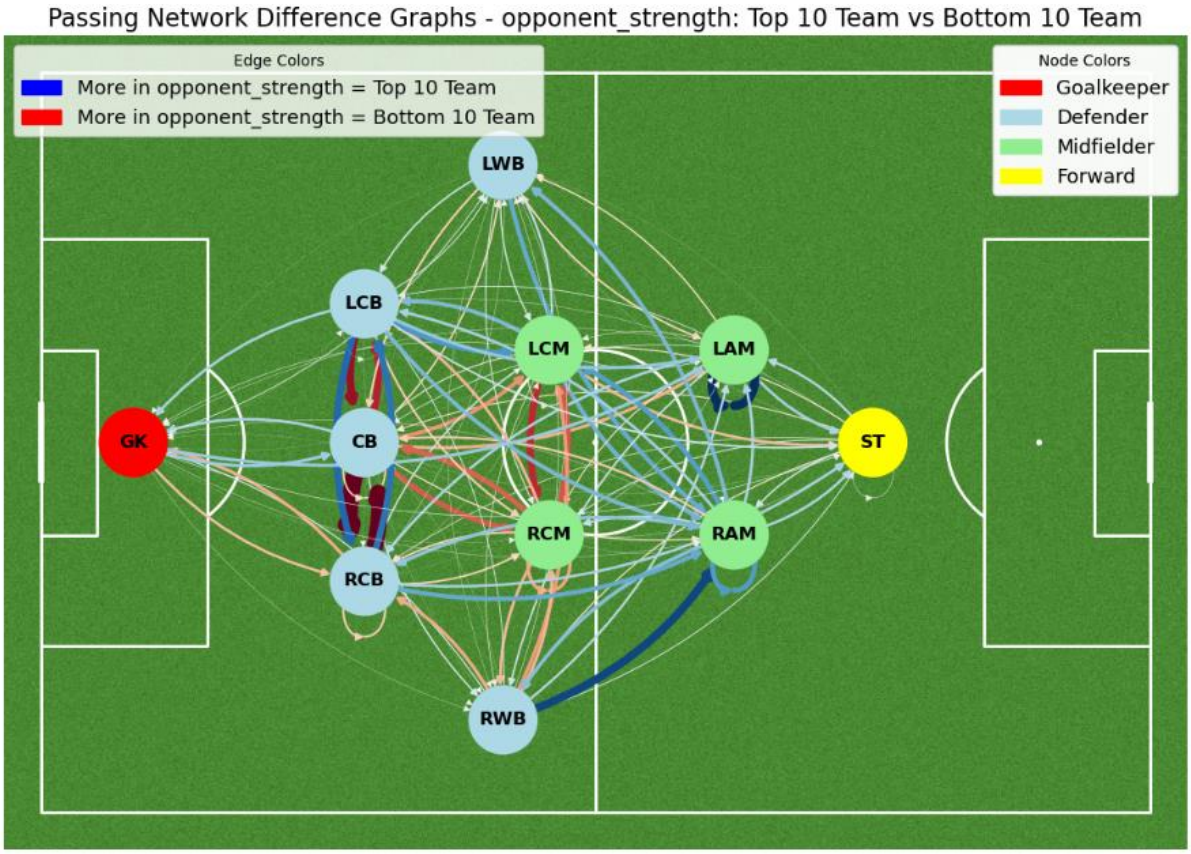


Figure 18. Weight Difference Graph – Against Top 10 vs Bottom 10 Teams

Figures 19 and 20 show the positional interactions over different time segments of the game. While the weight differences seem to be less, both diagrams show that different players become focal points of play as the game progresses. These shifts might reflect tactical changes, substitutions, or adaptations to the opposing team's tactics. This suggests that Arsenal tends to be a team not afraid to change tactics depending on the ongoing situation of games, rather than following a pre-set game plan, having flexible players who can adapt to different responsibilities.

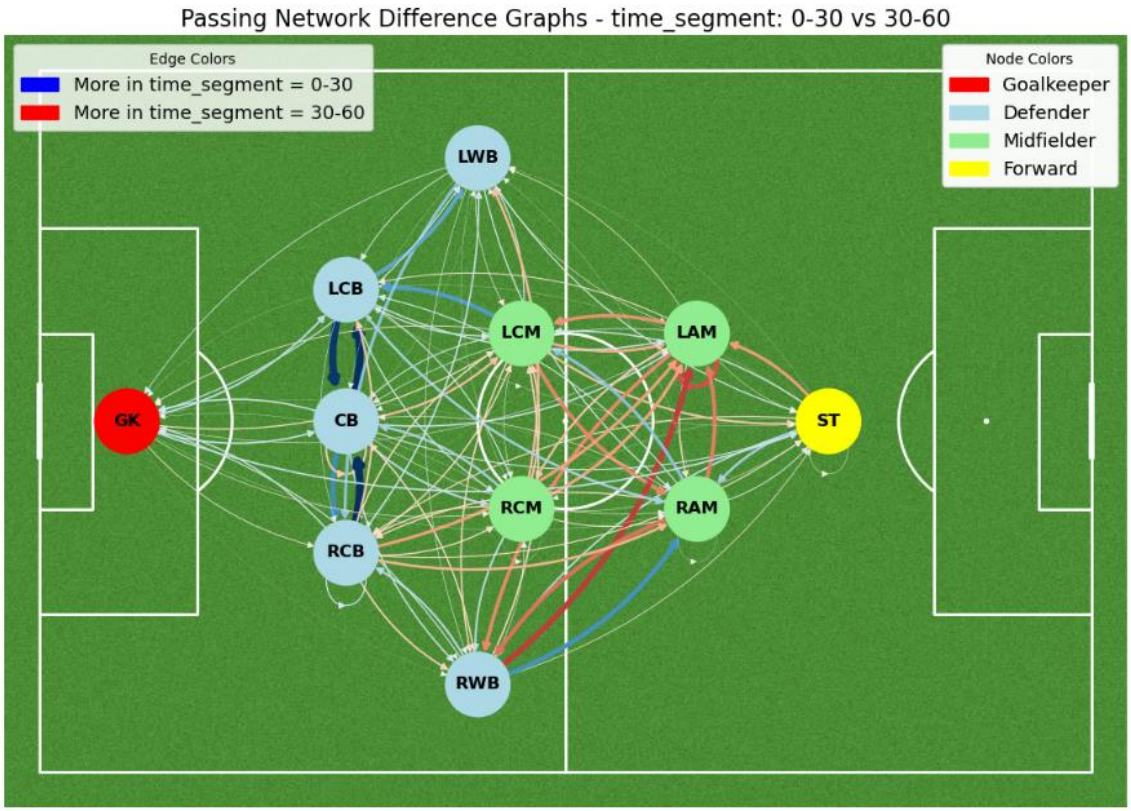


Figure 19. Weight Difference Graph – During the First vs Second 30 Minutes of Games

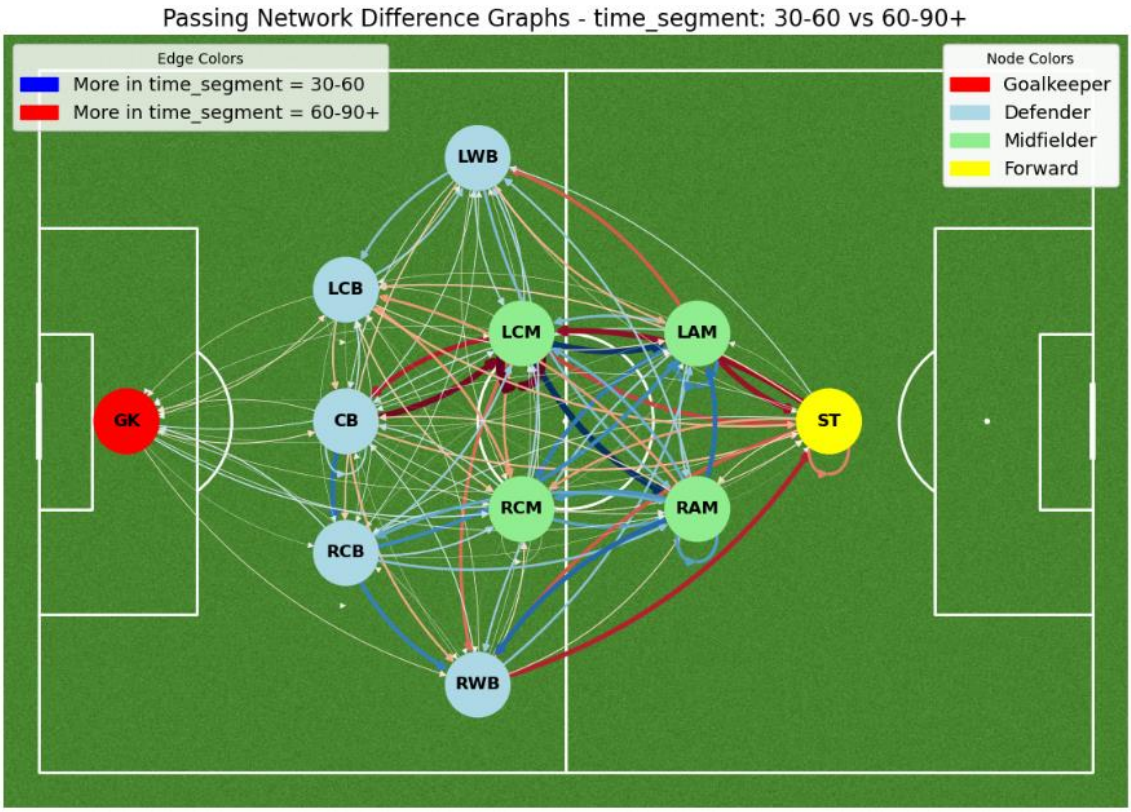


Figure 20. Weight Difference Graph – During the Second vs Last 30 Minutes of Games

Comparing the two diagrams and visualizing how the passing patterns change over time during games, what is noticeable is that the striker gets increasingly involved as the game goes by with him receiving more passes during the last 30 minutes of the game. Defenders seem to distribute the ball more between them during the early stages of the game. This likely indicates that Arsenal tends to start games slowly as they focus on maintaining possession with more sideways and backward passing between defenders, wing-backs and midfielders. The increased focus on the attacking phase and forward passing as the game goes by shows that Arsenal becomes more aggressive during the later stages of the game. This strategy allows Arsenal to exploit opponents's fatigue as the game progresses. A very common sequence of passing interactions between the CB-LCM-LAM-ST seems to occur during the last 30 minutes of the game, indicating a successful route to distribute the ball forward during this time segment.

Finally, to address how players adapt to different roles based on the current score of the game, we constructed difference graphs that compare passing interactions and overall player involvement when winning, drawing, or losing a game.

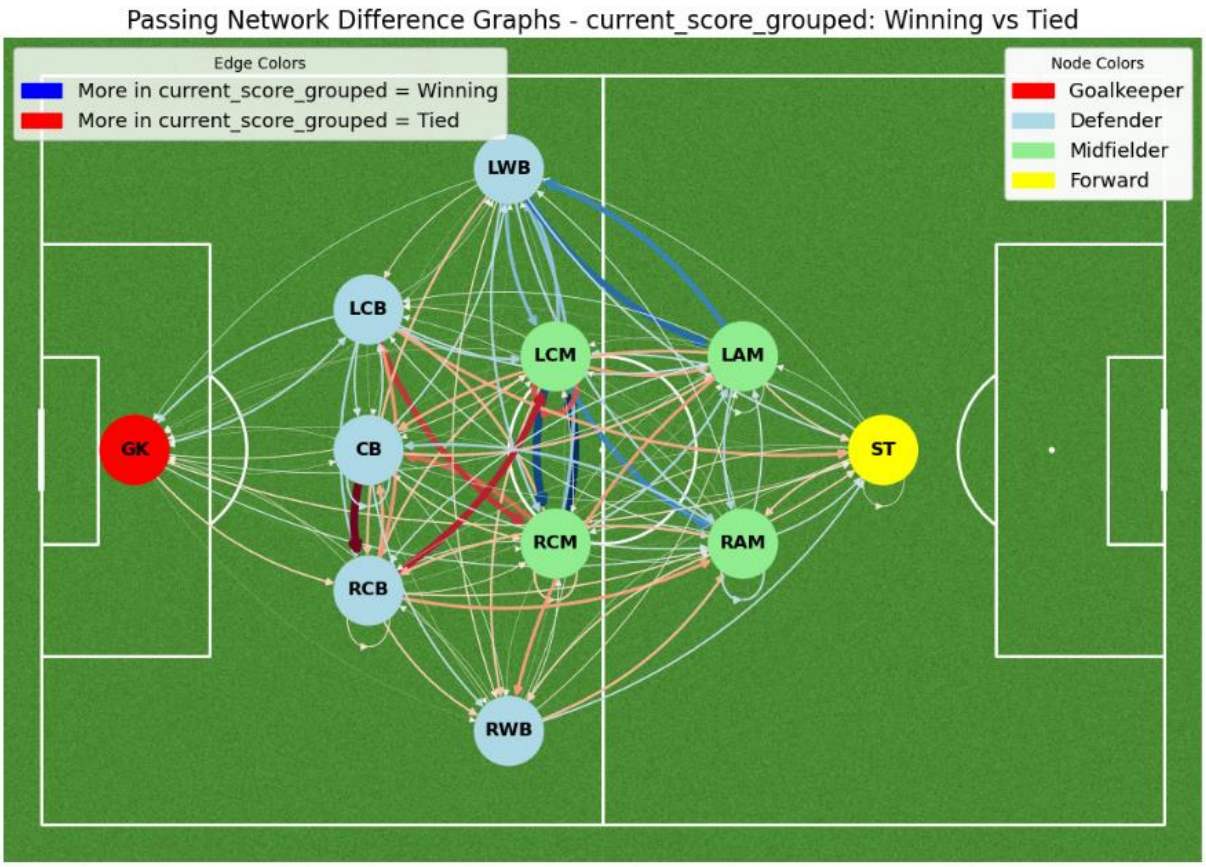


Figure 21. Weight Difference Graph – When winning vs when tied during games

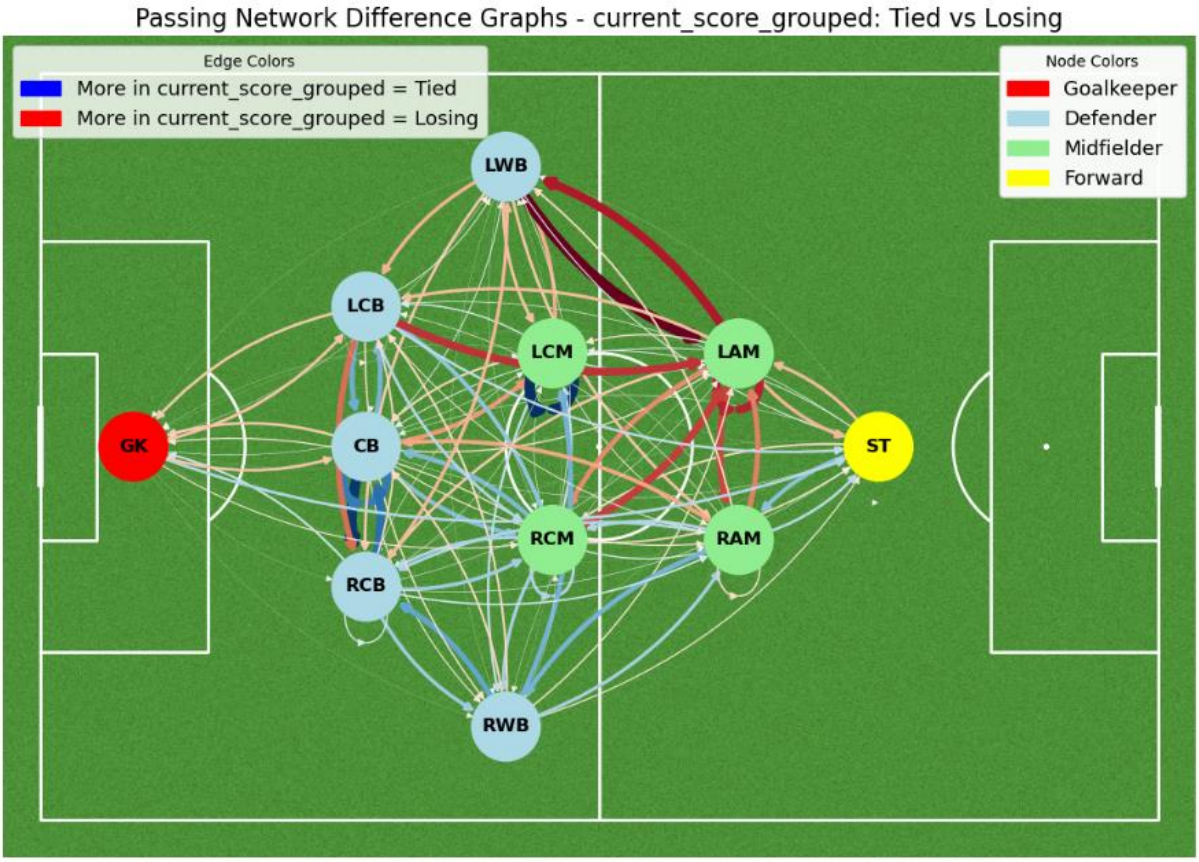


Figure 22. Weight Difference Graphs – When tied vs when losing during games

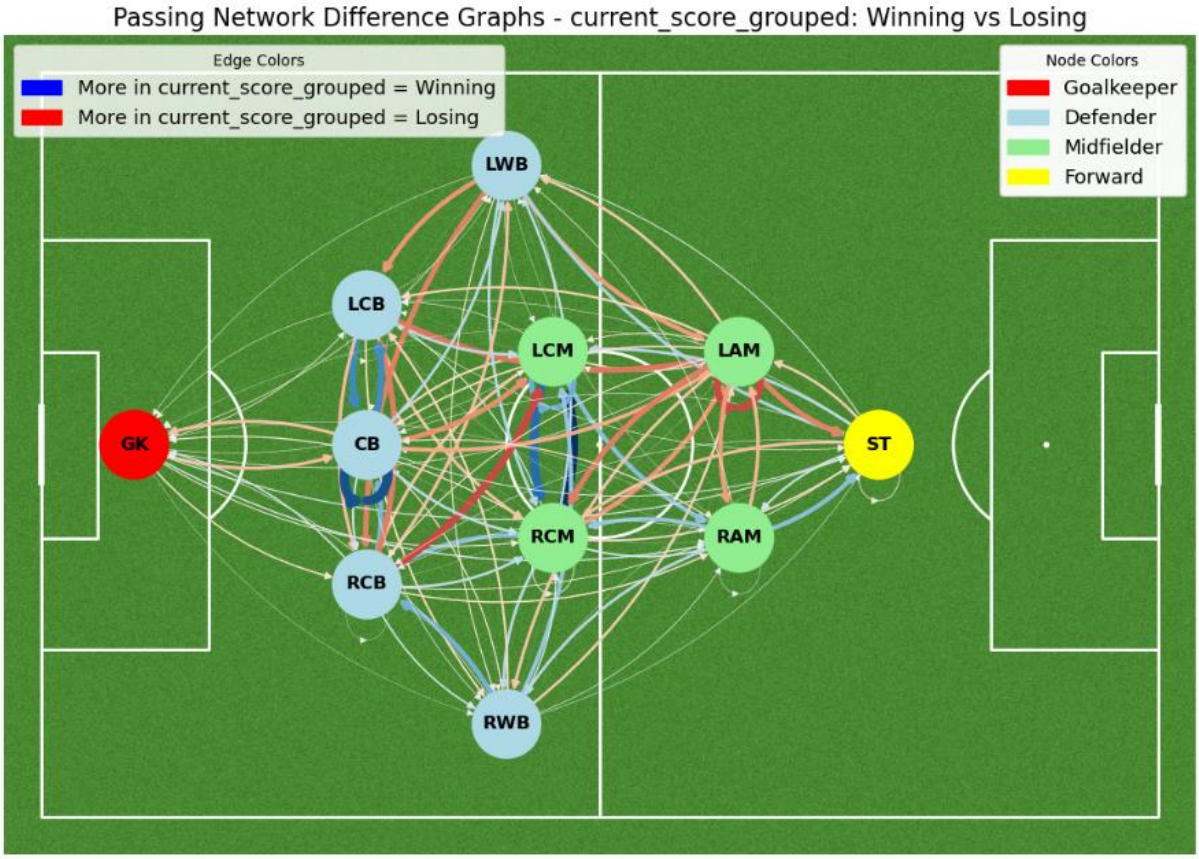


Figure 23. Weight Difference Graphs – When winning vs when losing during games

While difficult to interpret all three graphs, what is evident is that player involvement in Arsenal's passing strategy varies depending on the current score indicating that Arsenal tends to change tactics based on the ongoing situation of the game. What is very interesting is that Arsenal seems to heavily rely on left-attacking midfielders like M. Ozil in situations where the game is tied or when losing in comparison to when winning. More interactions between central players tend to happen when the game is tied in comparison to when winning, while in situations where Arsenal needs to score a goal, more interactions occur between left-sided players.

4.1.2 Passing Factor Analysis of Arsenal's Players

After analyzing passing networks and visualizing player connections and their positions, conducting factor analysis on passing statistics to associate players with underlying factors can further deepen our understanding of team dynamics and player roles. Table 10 represents the factor loadings for each feature that has a relative communality score.

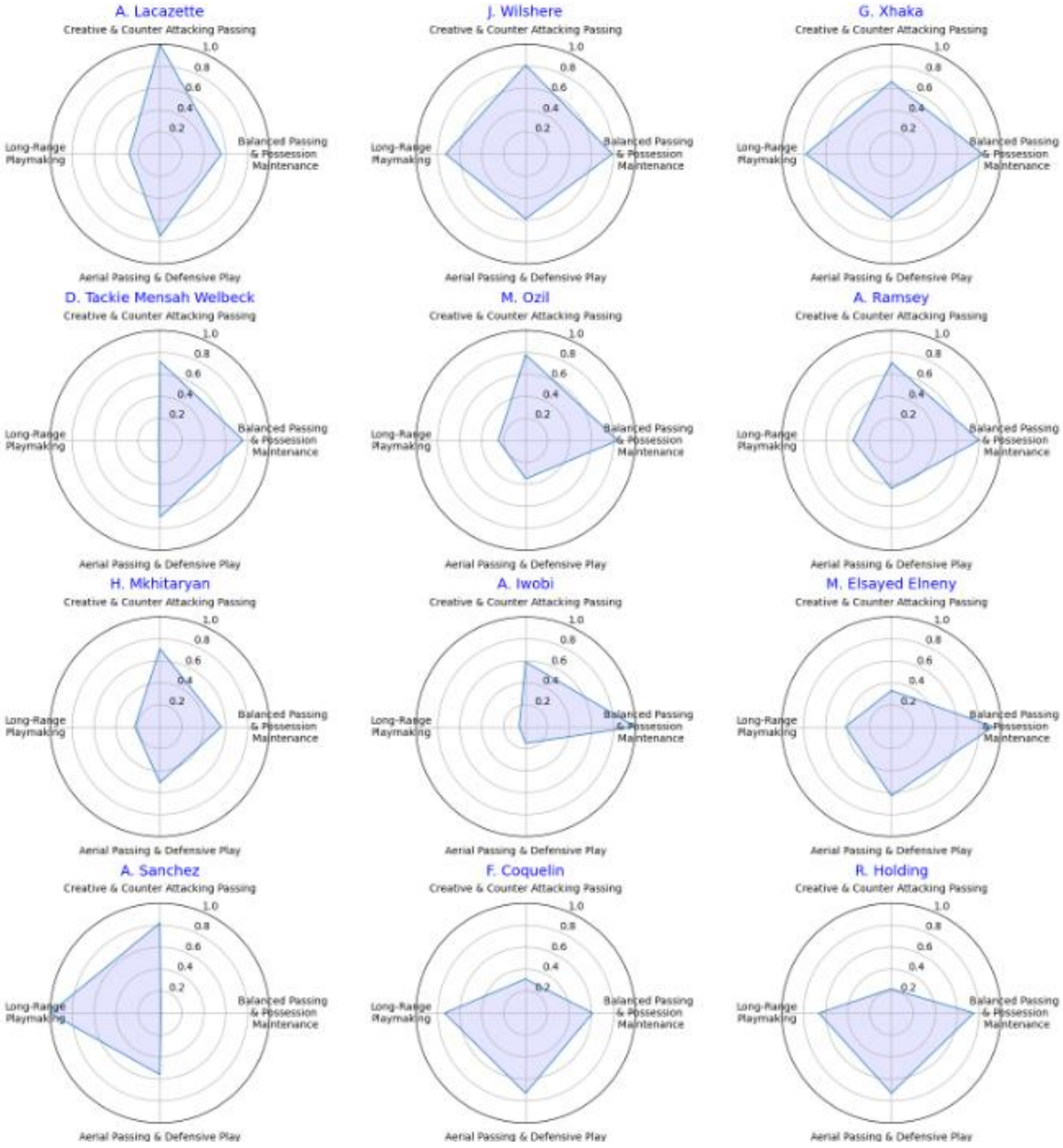
Features	Factor_1	Factor_2	Factor_3	Factor_4
Head Pass %	-0.3155	-0.210	0.780	-0.134
High Pass %	-0.1302	-0.180	0.015	0.914
Simple Pass %	0.0127	0.972	-0.036	-0.244
Smart Pass %	0.7089	-0.405	-0.080	0.041
% Pass Accuracy	-0.3928	0.831	0.083	0.050
Total Assists	0.7599	0.147	-0.141	-0.081
% Total Counter Attack Passes	0.7029	0.491	-0.123	-0.046
% Total Key Passes	0.6762	-0.252	-0.092	-0.145
% Total Dangerous Balls	-0.3093	0.153	0.227	0.234
% Total Final Third Passes	-0.0600	-0.607	-0.759	-0.221
% Total Interceptions	-0.7894	0.298	0.484	0.154
Avg Length of Passes	-0.6384	0.065	-0.347	0.452

Table 10. Factor Loadings

Looking at the factor scores of each variable, the four factors could be described as followed:

- **Factor 1: Creative & Counter-Attacking Passing**
- **Factor 2: Balanced Passing & Possession Maintenance**
- **Factor 3: Aerial Passing & Defensive Play**
- **Factor 4: Long-Range Playmaking**

The next step is to examine how each player is associated with each factor to understand better their roles. Figure 24 demonstrates radar charts that show the distribution of each player across the four key factors.



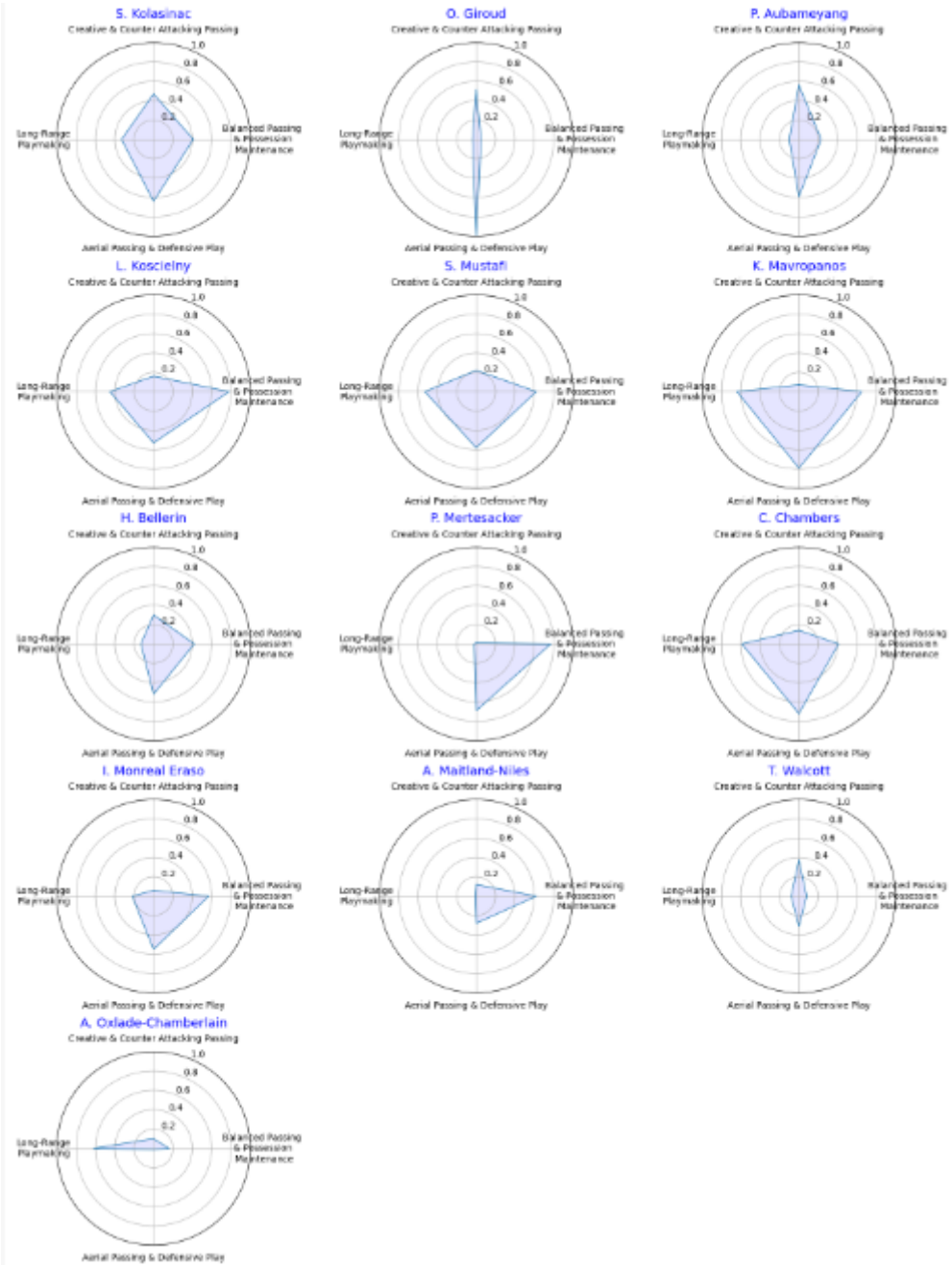


Figure 24. Radar Charts – Factor Scores by Player

Players like M. Ozil seem to be associated as creative and counter attacking players, while defensive players like L. Koscielny and S. Mustafi show a larger emphasis on aerial passes and their defensive responsibilities while other defenders such as C. Chambers and K. Mavropanos seem to also have active roles as long-range playmakers indicating their ability to contribute in possession. Players like G. Xhaka and J. Wilshere seem to be versatile as they show a more balanced radar chart, indicating their contribution across all factors.

Players with higher total factor scores are generally considered to be more versatile and effective across multiple dimensions of passing. Figure 25 shows that midfielders such as J.

Wilshere, G. Xhaka and M. Ozil are top performers and highly versatile in their roles. A. Lacazette, as a forward, shows that he is not only involved in goal-scoring but also participates in playmaking and maintaining possession. On the other hand, forwards such as P. Aubameyang and O. Giroud have mid-range scores, indicating their primary focus on goal-scoring rather than on passing metrics.

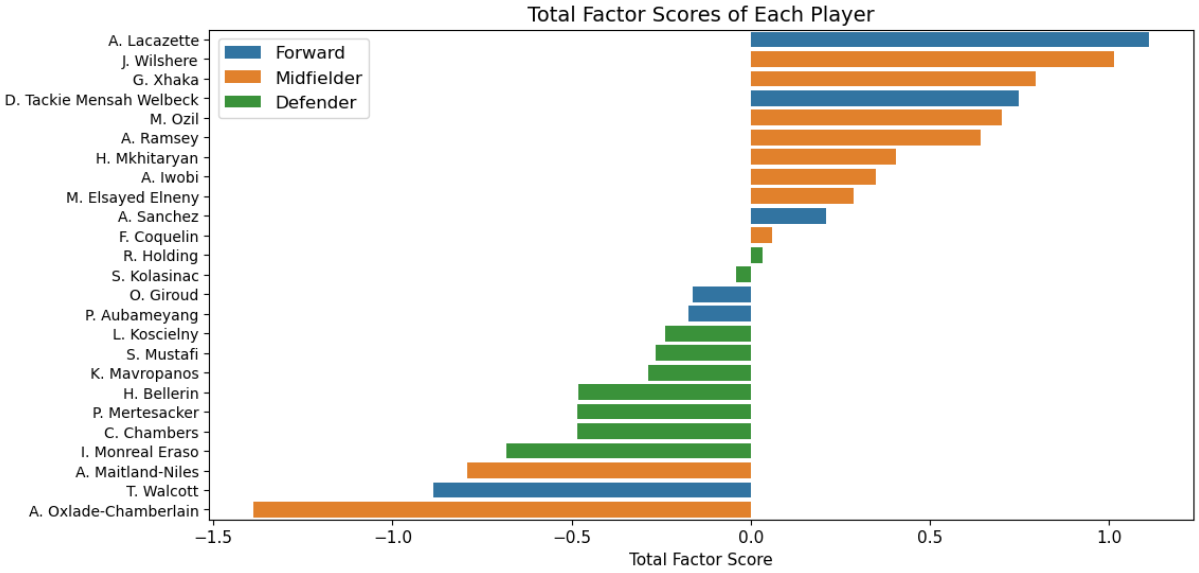


Figure 25. Total Factor Scores by Player

Finally, we created an MDS visualization scatter plot (figure 26) to illustrate the similarities and differences among players based on their factor scores. This revealed how closely players align in their passing roles and helped us identify clusters of players with similar passing profiles across different positions. We can see that the positions of players don't tend to overlap with a few exceptions which means that most players with similar positions tend to have similar factor scores and playing styles.

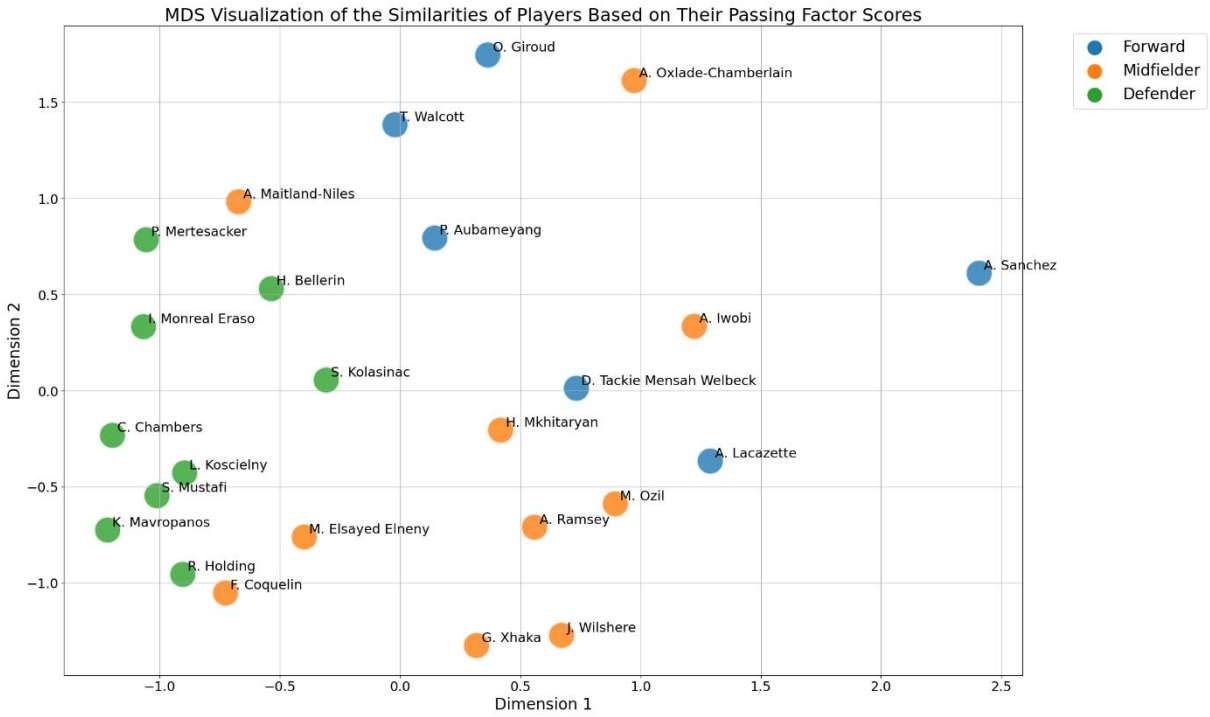


Figure 26. MDS Visualization of the similarities of players based on their passing factor scores

One thing we noticed when also comparing this graph to the radar charts of players and the total factor score plot is that players more to the bottom left are players that have low factor scores in creative and counter attacking passing but do show decent contribution in the other 3 factors. This area of the graph seems to be mainly occupied by defenders.

Players on the far top of the graph are players that have good scores in creative and counter attacking passing as well as in aerial passing and defensive play but are not associated with long-range playmaking or possession maintenance. These players show great imbalance between the factors and are players that do not have a major effect in possession. This area of the graph is occupied mainly by forwards. Players on the middle and lower part of the graph, seem to be the most versatile players on the team when it comes to the team's passing strategy. We can understand this since most of these players seem to have high overall scores on these factors. This area of the graph is mainly occupied by central midfielders. Players that cover the far-left area of the graph seem to have high scores in aerial passing and defensive play as well as possession maintenance but are not used for starting attacks. Looking at the passing networks of these players, they are mainly defenders who play as wing-backs.

A player who seems to have a unique playing style is A. Sanchez who is in the far-right of the graph. Looking at the positional changes of the player, he tends to cover the left flank of the attack with some tendencies to move centrally mainly in the final-third of the pitch. He seems

to be mainly associated with creative and counter-attacking passing, long-range playmaking and aerial passing and defensive play but does not seem to get really involved in the possession maintenance. His role is crucial in both attacking transitions and defensive situations, making him a valuable asset in fast-paced and high-risk scenarios.

4.2 The Positional Tendencies of Arsenal's Players Across Different Game Situations

4.2.1 Comparing the Positioning of Arsenal's Players Across Various Game Situations

Figures 24, 25 and 26 show the positional tendencies of players across various game scenarios and offer an insightful analysis of the team's tactical adjustments and player roles. Players like H. Bellerin, S. Kolasinac and L. Koscielny maintain relatively stable average positions across different scenarios, reflecting their consistent roles in the defensive line. On the other hand, midfielders and forwards show more noticeable variations in their positions. Players like G. Xhaka, M. Ozil and A. Lacazette who play more minutes and are more influential, show consistent positioning across various game scenarios, while other players like F. Coquelin, A. Oxlade-Chamberlain ,A. Iwobi and O. Giroud seem to be able to adjust in different positions, likely in response to tactical changes or substitutions.

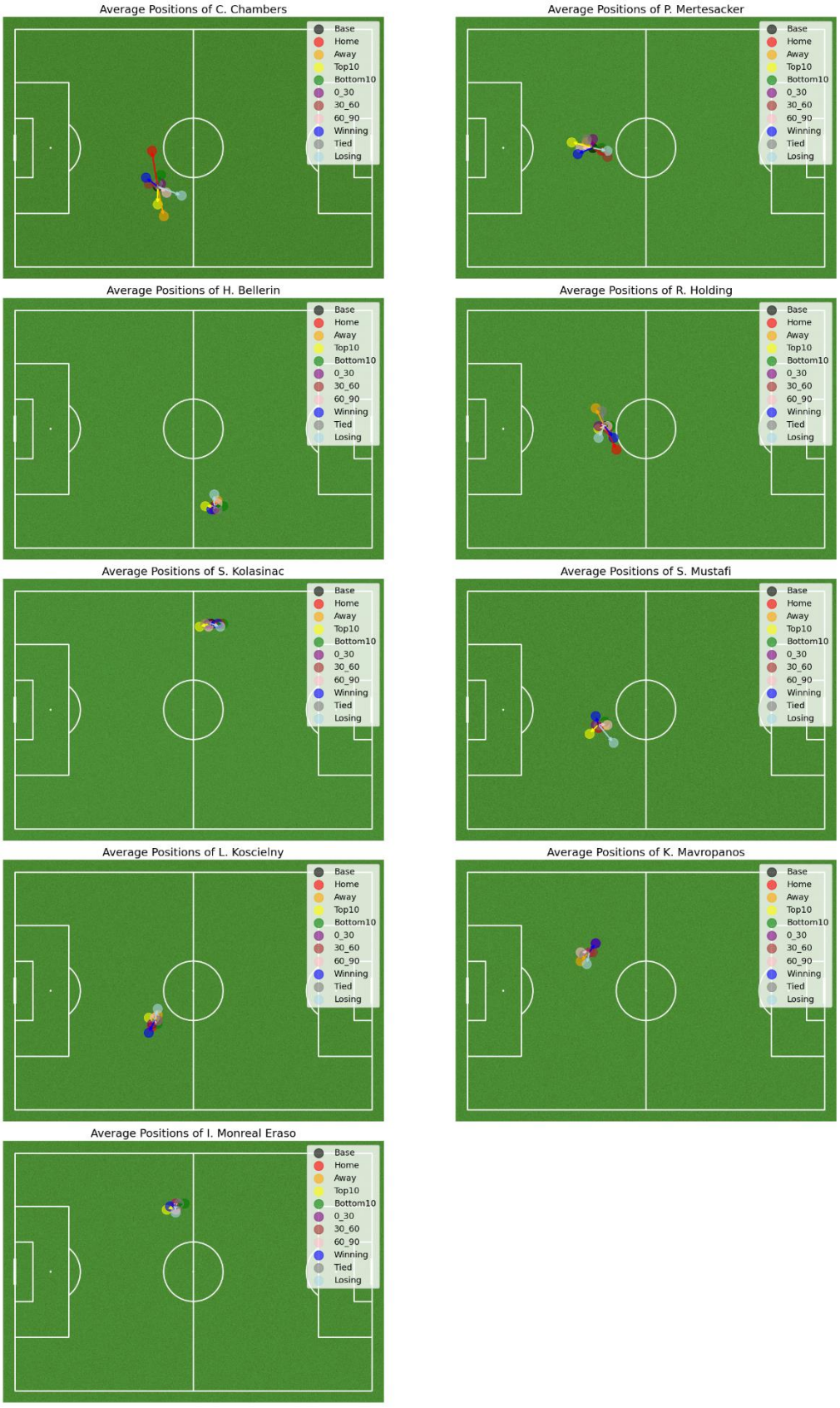


Figure 27. Average Positions of players by game context - Defenders

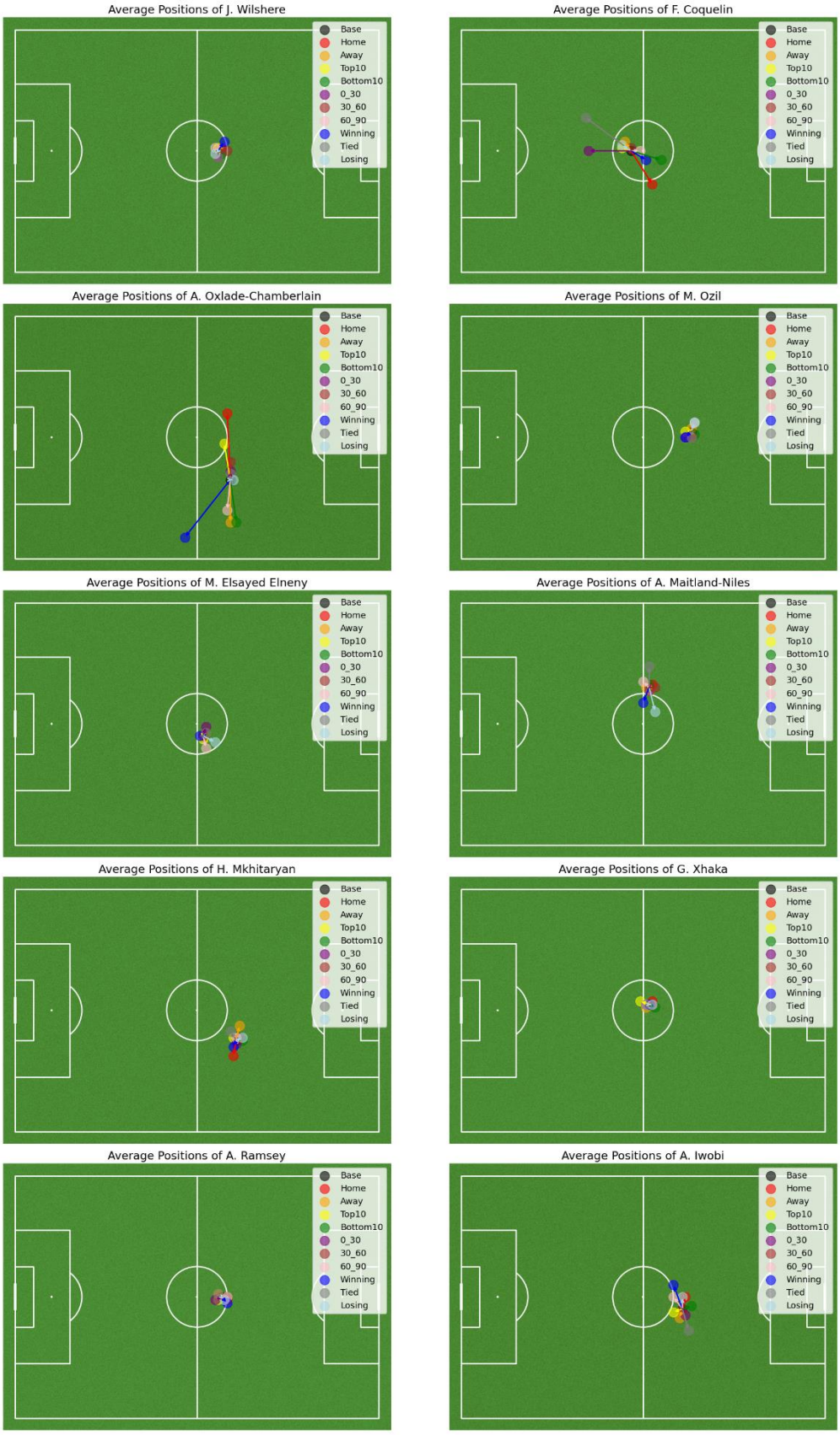


Figure 28. Average positions of players by game context - Midfielders

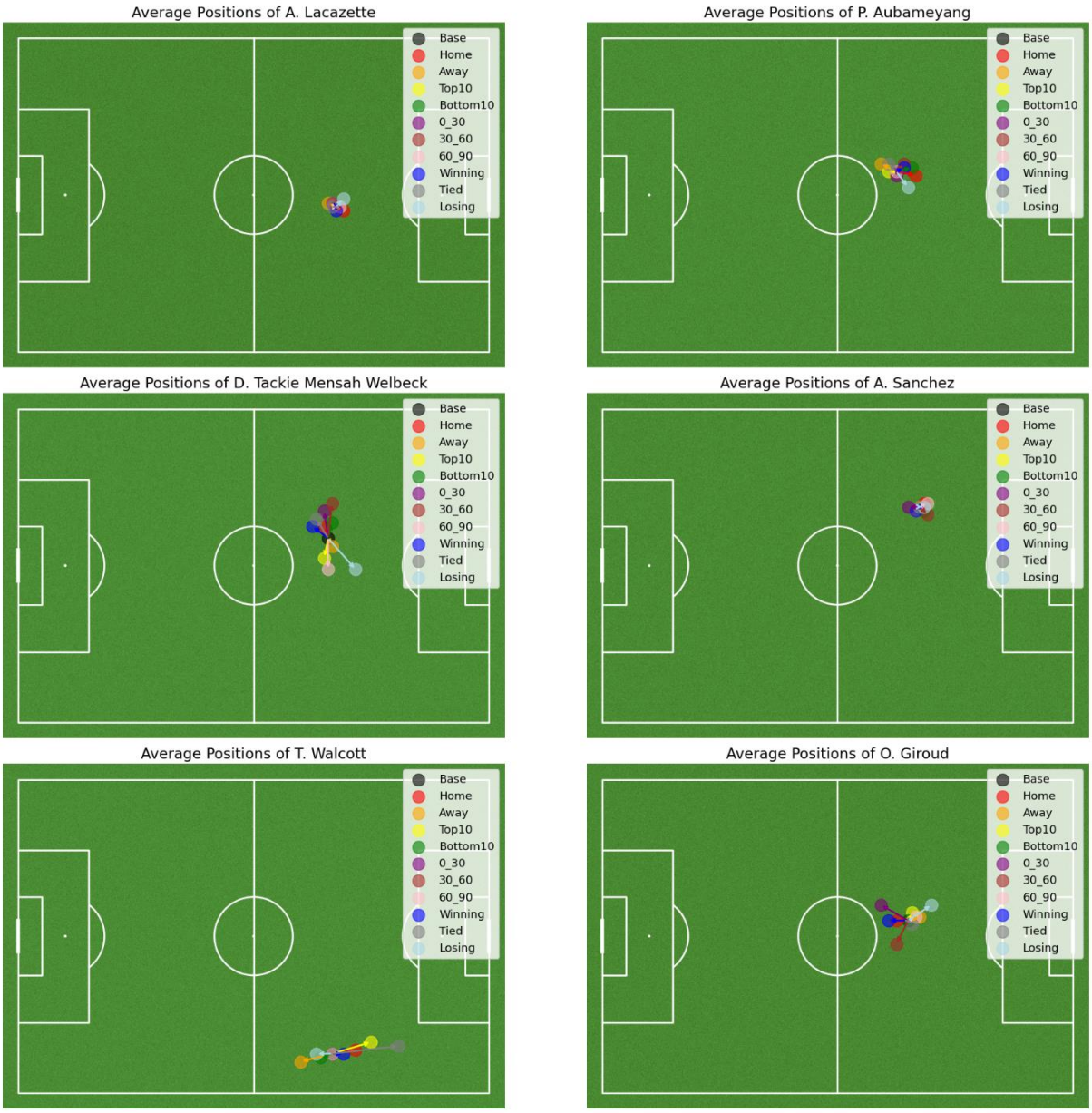


Figure 29. Average Positions of Players by game context - Forwards

To provide a visual summary of each player's role, positional tendencies and their contribution to Arsenal's overall passing strategy, we generated individual heatmaps of players that show where players contribute most on the field in terms of getting involved in the team's passing dynamics.

4.2.2 The Passing Heatmaps of Arsenal's Players

Looking at the heatmaps of central defenders like L. Koscielny, S. Mustafi, R. Holding, we understand that their passing distribution will be concentrated in their defensive flanks, indicating their contribution in build-up play or clearing the ball from danger zones. Wing-backs like H. Bellerin and S. Kolasinac seem to be more active near the sidelines and have the

highest concentration of passes in the attacking third areas. This indicates that the wing-backs have a high involvement in the attacking play rather than being traditional full backs. Their positioning indicates that they are responsible for providing width to the team's attack by overlapping to stretch the opponent's defense horizontally. I. Monreal Eraso seems to have a unique passing heatmap that indicates his versatility in the team's passing strategy. Although he is primarily a left center-back, his heatmap does suggest that he might sometimes be used as a left-back, providing a versatile role who can contribute both defensively and offensively.

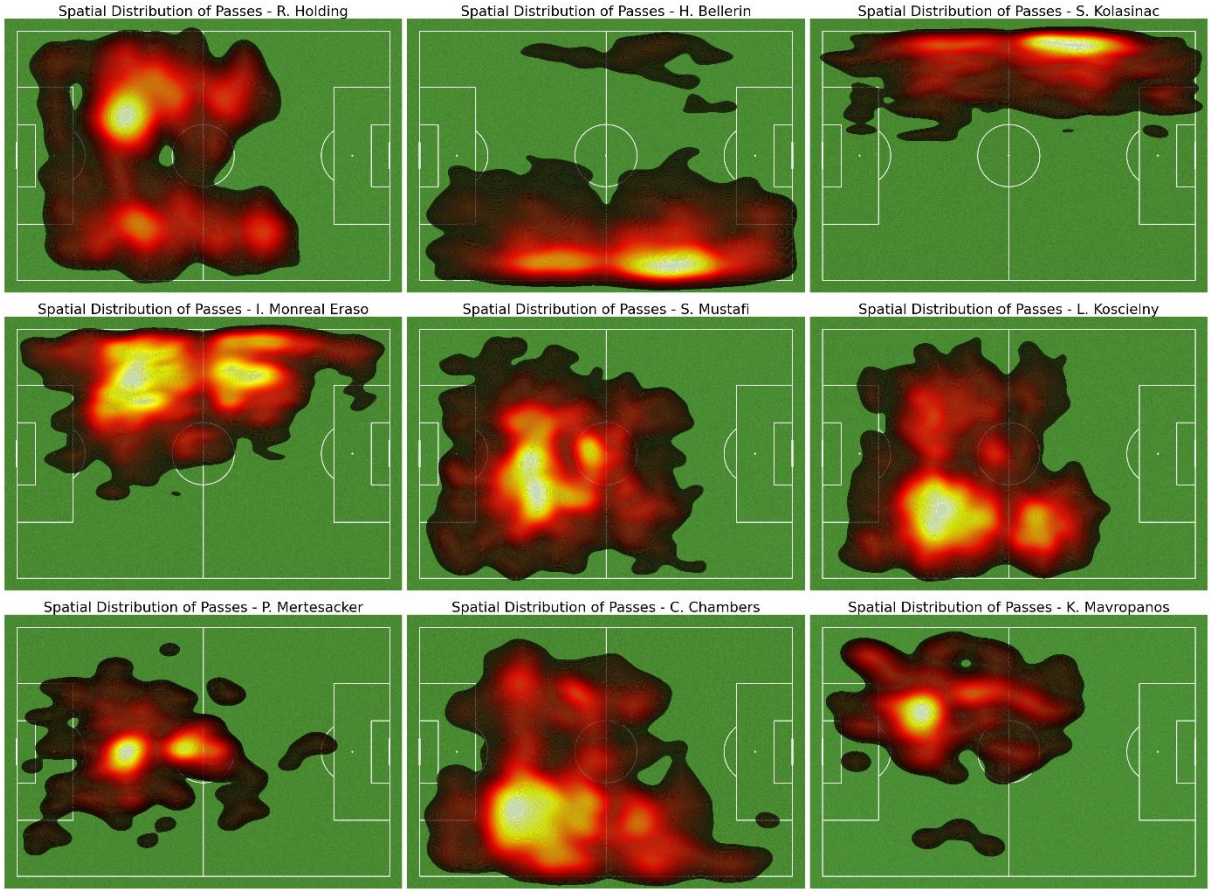


Figure 30. Player Passing Heatmaps - Defenders

Midfielders seem to show higher variability in their heatmaps and give us deeper role-specific insights. Players like M. Ozil, H. Mkhitaryan and A. Iwobi show significantly higher activity in the attacking areas which indicates that their main role is to create chances and distribute the ball forward. Looking also at Figure 26, they are relatively close in the MDS scatter plot which indicates they have similar roles. Ozil seems to have a freer role where he can move horizontally across the pitch operating as a traditional number 10, while Iwobi and Mkhitaryan are more active on the right flank. On the other hand, players like G. Xhaka, A. Ramsey and J. Wilshere seem to cover both defensive and offensive zones, suggesting their dynamic role in both supporting the defense and joining attacks as box-to-box midfielders. A. Maitland-Niles has a high concentration of passes on the left flanks indicating that he is a versatile player who can play either in midfield or as a wing-back. Looking at figure 26, his passing statistics are similar to other wing-backs like H. Bellerin and S. Kolasinac indicating his ability to play in both positions.

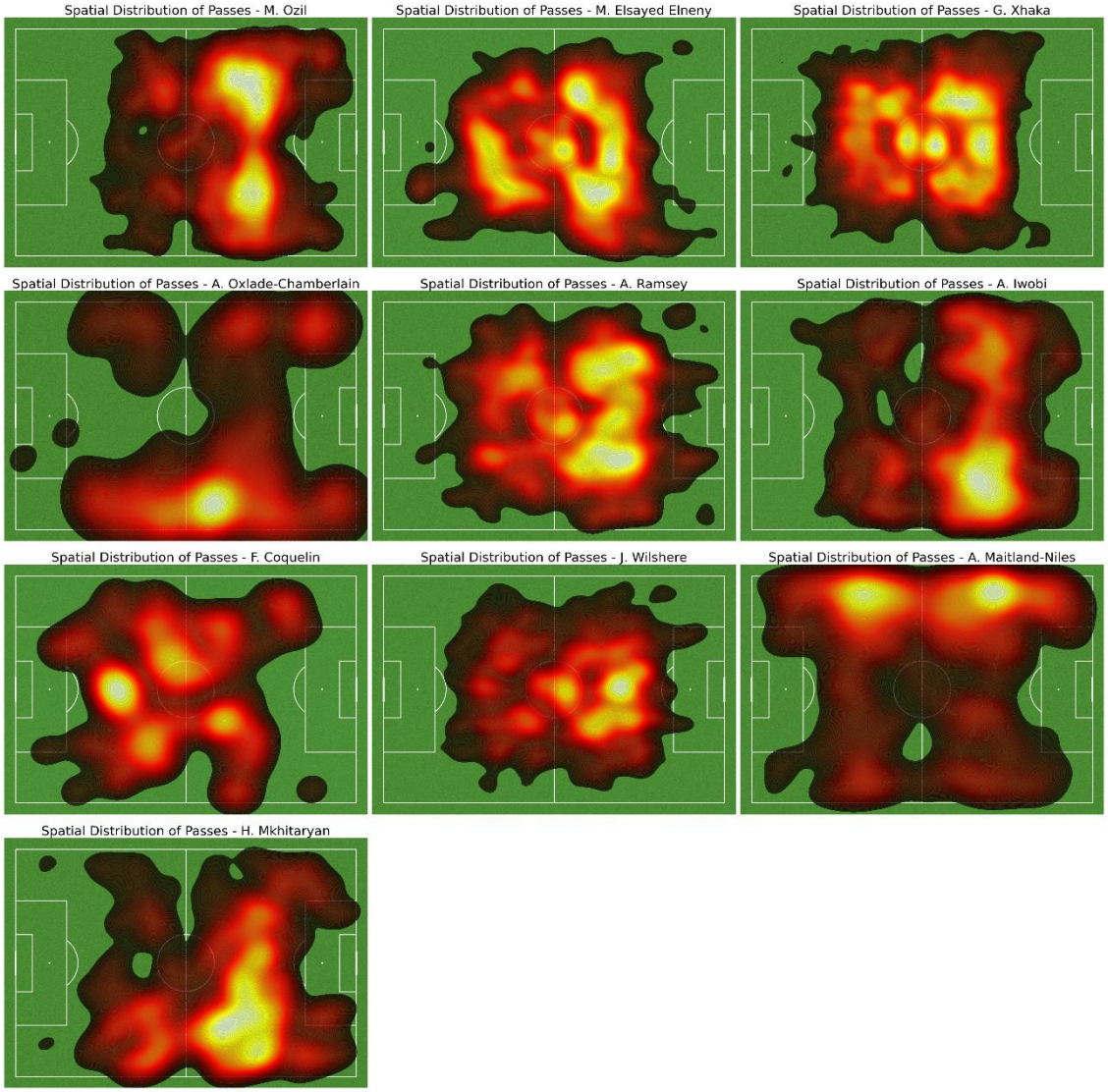


Figure 31. Player Passing Heatmaps - Midfielders

When analyzing the heatmaps of forwards based on their passing contributions, we can segment them into different categories based on their roles and positioning. Players like A. Lacazette and O. Giroud show a high concentration of passing in central attacking areas. Giroud's heatmap suggests he is a traditional target man who holds up the ball and distributes to teammates around him. A. Lacazette seems to move more into the half-spaces and tends to drop deep sometimes to link play. His nearness to creative players like Ozil and Mkhitaryan (Figure 26) and his high total factor score (Figure 25) indicate that he is a player who can also create chances. P. Aubameyang's heatmap shows significant activity both in central areas and towards the left side of the pitch. This suggests he often utilizes his pace to drift wide or cut inside from the left flank. Players like T. Walcott, D. Welbeck and A. Sanchez seem to operate as wingers moving along the touchlines in the opponent's half, reflecting their role in stretching the play to deliver crosses or cut inside and shoot.

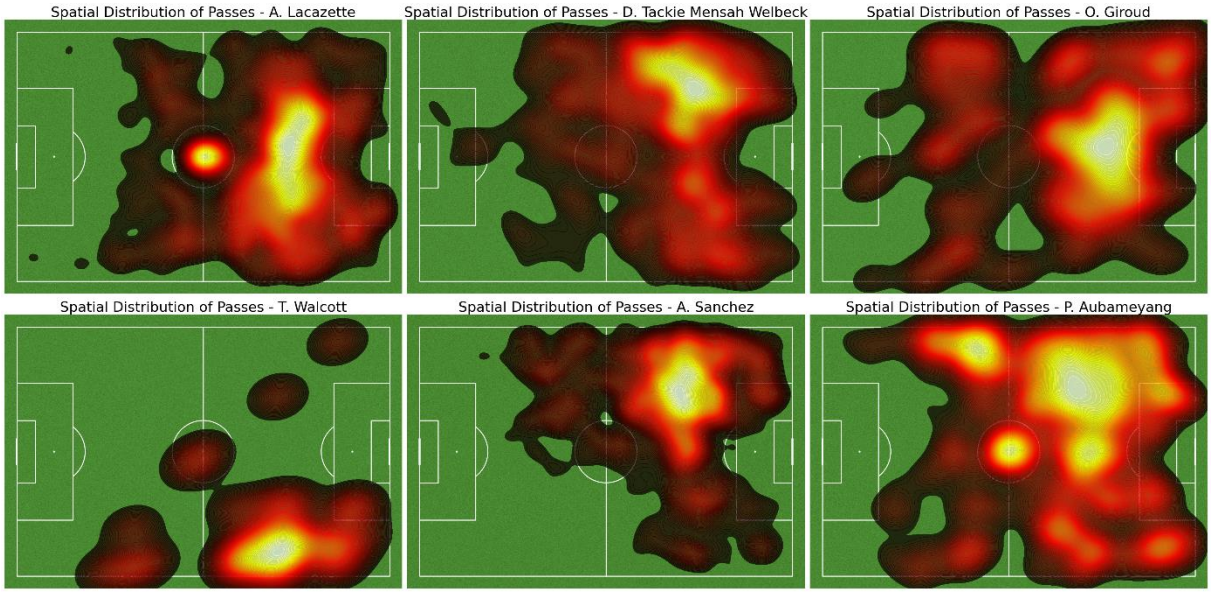


Figure 32. Player Passing Heatmaps - Forwards

After analyzing the passing networks across various game contexts, we observe that the key differences across different situations are more evident in the interaction between players and their roles, rather than their positioning. This suggests that the team's tactical adjustments are more dynamic and strategy-oriented depending on the game's context. The stability in player positioning despite these changes indicates a consistent formation, with the team relying on familiar spatial arrangements while varying the intensity and frequency of interactions. Overall, it seems that Arsenal's passing strategy appears to be adaptive, prioritizing the specific opponent and the game context, rather than applying a uniform approach, making their tactics more responsive to the strengths and weaknesses of different opponents.

After addressing the similarities of players and their contribution to the team's passing strategy, while also looking at their heatmaps, we notice that players who show similarities are also players that have similar positions and move in the same areas of the pitch. Defenders and Attackers seem to be way less versatile in the team's passing strategy than the midfielders and show more focus on certain aspects of possessions. This indicates that Arsenal plays position-based football rather than total football, where each player based on their position has certain duties when it comes to the team's passing strategy rather than being positionally flexible.

4.3 Arsenal's Distinct Passing Patterns and Dynamics Across Various Game Scenarios.

4.3.1 Arsenal's Possession Cluster Analysis

Transitioning our analysis to a more macro perspective, we grouped event sequences into possessions to explore the underlying passing patterns and dynamics. Table 11 shows that a total of 28 clusters were extracted and around 31% of possessions were occasional and treated as noise, indicating high variability in Arsenal's passing strategy. Looking at the cluster sizes, only 7 clusters have more than 100 possessions as most of them seem to indicate rare passing strategies.

Cluster	Cluster Size
2	2104
-1	1194
0	515
4	314
7	242
5	214
1	144
6	110
14	41
8	30
9	14
11	13
3	12
12	11
15	10
13	10
20	10
18	7
16	7
26	6
24	5
21	5
23	5
17	5
22	5
10	5
25	4
19	2
27	2

Table 11. Possession Cluster Sizes

Tables 12 and 13 show the possession dynamics of each cluster and statistics related to the pass type distribution and the positional roles.

Cluster Dynamics Summary Table													
Cluster	Total Possessions	horizontal_amplitude	vertical_amplitude	avg_event_length	avg_event_angle	avg_vertical_change	avg_horizontal_change	duration	num_events	starting_position_x	starting_position_y	ending_position_x	ending_position_y
2	2104	34.23	33.85	17.65	30.67	5.22	0.71	9.62	4.87	56.66	39.62	78.06	41.12
0	515	43.43	49.74	25.35	30.11	9.55	0.89	11.33	5.04	48.77	39.63	85.3	42.58
4	314	19.75	23.81	15.33	36.67	5.55	-0.46	4.3	3.06	47.03	40.37	63.24	38.12
7	242	34.84	42.5	19.36	44.71	8.64	0.98	9.23	4.64	65.7	38.19	101.2	42.3
5	214	39.68	27.87	22.85	29.06	5.61	3.22	5.61	3.46	90.03	39.5	109.77	47.62
1	144	29.05	45.52	28.29	-0.44	8.57	-0.92	6.59	3.23	33.8	36.66	56.73	35.87
6	110	51.96	32.7	33.02	4.64	29.1	11.87	6.32	3.97	87	37.32	120	40
14	41	37.64	50.46	21.15	47.32	10.94	1.34	10.92	4.59	13.32	39.06	60.32	41.35
8	30	47.49	25.2	18.2	-3.93	45.92	18.93	2.97	2.47	101.84	41.25	120	40
9	14	64	53.03	44.47	-10.4	16.07	13.43	8.65	4.5	60.09	51.09	120	40
11	13	15.08	22.34	14.06	42.21	7.78	-1.13	3.17	2.77	34.71	18.95	54.83	16.74
3	12	36.6	71.5	31.3	78.95	22.16	3.43	10.06	3.58	13.8	42.8	85.3	52.73
12	11	37.38	42.65	25.25	38.83	9.07	-3.83	7.28	3.82	34.36	46.91	70.91	31.56
20	10	57.92	105.48	33.32	10.25	-13.12	-6.95	10.34	6.3	74.52	40.32	0	0
13	10	67.28	84.24	55.87	58.92	34.22	23.09	15.71	2.3	0.84	0	85.08	48
15	10	53.28	103.44	41.5	19.76	-10.36	-5.18	12.11	5.2	53.04	25.44	0.84	1.2
18	7	59.66	81.77	22.94	19.47	7.8	-3.13	22.53	9.86	37.37	62.63	112.8	33.26
16	7	25.94	62.4	35.21	98.49	31.2	-11.83	5.52	2	13.71	41.37	76.11	17.71
26	6	15.73	17.4	13.58	80.45	6.7	0	3.1	2.5	25.6	31.33	41.4	32
17	5	47.04	108.24	43.22	16.46	-12.87	-2.11	8.17	4.2	120	80	57.84	61.76
10	5	72	76.08	30.83	13.43	7.93	11.39	14.65	5.6	49.92	15.2	93.36	77.6
21	5	28.8	26.64	14.48	12.21	2.06	8	5.43	4	24.96	40.96	37.44	68
22	5	16.48	25.2	17.26	97.12	11.96	1.39	1.68	2.2	29.52	56.48	54.72	59.52
23	5	75.04	107.76	51.16	15.39	-14.42	-12.47	50.06	4.4	120	80	88.08	17.92
24	5	73.12	52.32	37.07	64.11	19.78	21.57	9.43	3.2	0	0	51.12	45.6
25	4	13.4	28.5	18.47	86.06	14.25	-0.4	1.97	2	56.1	18.4	84.6	17.6
19	2	71.2	78	35.43	9.8	8.92	-1.66	18.4	6.5	45.6	14.8	104.4	4
27	2	16	12.6	10	1.88	1.44	3.2	5.38	4	23.4	61.6	29.4	74.4

Table 12. Cluster Summary Statistics – Pass Dynamics

Cluster Pass Distribution and Positional Role Statistics (top 12 clusters by size)											
Cluster	Total Possessions	Simple pass %	High pass %	Smart pass %	Launch %	Cross %	Hand pass %	Goalkeeper %	Defender %	Midfielder %	Forward %
2	2104	100	0	0	0	0	0	1.55	39.18	46.8	12.47
0	515	75.16	24.84	0	0	0	0	2.27	46.07	41.73	9.94
4	314	100	0	0	0	0	0	0.54	47.8	39.98	11.68
7	242	73.81	0	26.19	0	0	0	0.09	31.95	50.23	17.73
5	214	61.3	0	0	0	38.7	0	0	27.36	51.9	20.73
1	144	56.23	0	0	43.77	0	0	19.38	48.55	24.28	7.8
6	110	100	0	0	0	0	0	0	22.38	54.74	22.87
14	41	72.67	0	0	0	0	27.33	23.66	31.72	41.4	3.23
8	30	34.78	0	0	0	65.22	0	0	48.48	33.33	18.18
9	14	67.44	0	32.56	0	0	0	0	30.65	50	19.35
11	13	100	0	0	0	0	0	0	41.67	33.33	25
3	12	42.86	28.57	0	0	0	28.57	27.91	27.91	44.19	0
12	11	54.17	45.83	0	0	0	0	0	59.52	28.57	11.9
20	10	79.59	0	0	0	20.41	0	0	31.67	60	8.33
13	10	16.67	83.33	0	0	0	0	47.83	39.13	8.7	4.35
15	10	71.43	28.57	0	0	0	0	2.08	29.17	50	18.75
18	7	87.93	0	0	0	12.07	0	0	50	42.65	7.35
16	7	0	0	0	100	0	0	57.14	7.14	0	35.71
26	6	0	0	0	100	0	0	0	66.67	26.67	6.67
17	5	58.33	0	0	0	0	41.67	47.62	28.57	23.81	0
10	5	73.68	26.32	0	0	0	0	3.7	77.78	18.52	0
21	5	100	0	0	0	0	0	0	40	60	0
22	5	0	0	0	0	0	0	0	63.64	36.36	0
23	5	100	0	0	0	0	0	27.27	54.55	18.18	0
24	5	37.5	0	0	0	0	62.5	62.5	12.5	25	0
25	4	0	100	0	0	0	0	0	62.5	25	12.5
19	2	81.82	18.18	0	0	0	0	0	69.23	23.08	7.69
27	2	100	0	0	0	0	0	0	75	25	0

Table 13. Cluster Summary Statistics – Pass Distribution and Positional Roles

Cluster 2 seems to be the most dominant possession type indicating that Arsenal heavily relies on a core strategy centered on high-volume, low-risk-passing with forwards dropping deep aiming to control the game's tempo and maintain possession effectively. Highly frequent clusters tend to involve a high volume of simple and short passes, with a more balanced distribution across different positional roles and with an emphasis on defenders and midfielders. Frequent clusters generally start and end in the middle and forward areas of the pitch, indicating a strategy focused on counter-pressing the opponent to maintain possession in the opponent's half.

On the other hand, less frequent clusters show a significant percentage in specialized pass types such as high passes and launches, indicating that these clusters may be involved in specific tactical scenarios like counter-attacks or long-range playmaking. Many of these clusters show a high involvement of the goalkeeper and the defender and long-range passes, indicating their occasional involvement in starting and building up possessions from the back line or putting long balls to wide areas for attackers to get in behind the opponent's defensive line.

Looking at other frequent possession types, **cluster 0** involves lengthy build-up plays starting from the defensive third, where defenders such as C. Chambers and S. Mustafi who are highly associated with long-range playmaking launch long balls, utilizing the pace of full-backs and wide forwards to break defensive lines in situations where Arsenal could be pressed.

Cluster 4 involves rapid, compact possessions centralized around key players like G. Xhaka and M. Ozil focusing on quick build-up play in situations where Arsenal could intercept the ball and attempt to counter attack.

Cluster 7 indicates successful advancement into attacking areas through smart, diagonal passes by creative midfielders like M. Ozil and A. Ramsey, aiming to break down compact defences.

Cluster 5 represents attacks developing in left attacking areas, with significant contributions from full-backs like S. Kolasinac or forwards like A. Sanchez and P. Aubameyang emphasizing crosses into the box.

Clusters 1 and 3 seem quite similar (figure 4) and mainly involve possessions where the goalkeeper or the defenders use long balls to exploit spaces behind the opponents, with a particular focus on finding wide players. These possessions could occur in situations where the opponent is high up the pitch to pressure Arsenal.

Cluster 11 focuses on build-up play with forwards like A. Lacazette dropping deep to the midfield third to assist in possession maintenance. These possessions could occur in situations where the opponent's midfield is compact, with the forward dropping deep to outnumber the opposition's midfield to distribute the ball forward.

Cluster 14 initiates from the goalkeeper, involving short passes for controlled build-up, spreading the opposition and methodically progressing through the lines.

Clusters 6, 8 and 9 involve possession starting in the attacking areas indicating that the forwards and attacking midfielders press high in the attacking third to force turnovers and regain possession.

4.3.2 The Frequency of Arsenal's Possession Clusters Across Various Game Situations

Now that we have described the clusters and understood the situations in which they would occur and the main players involved, we are going to compare their frequency across different game contexts.

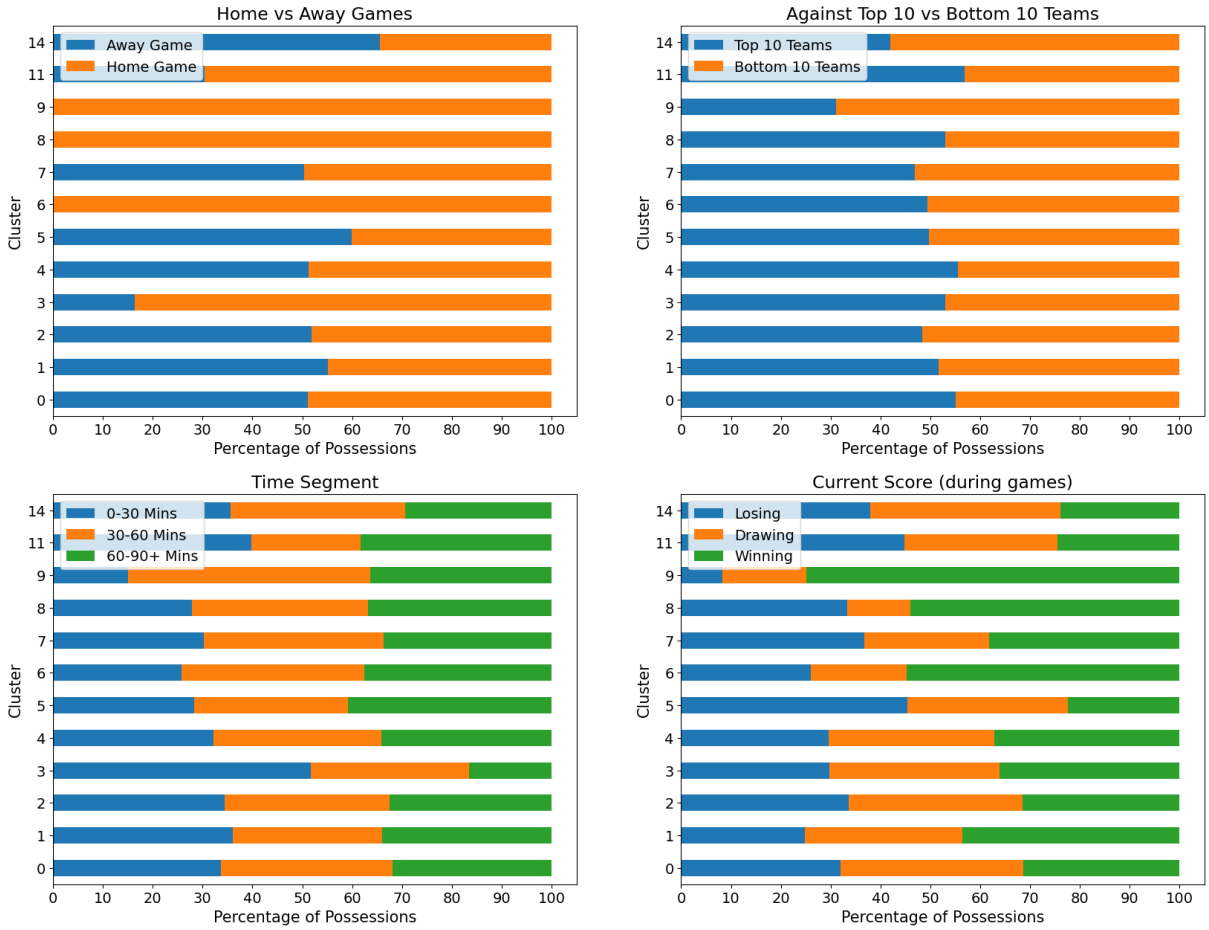


Figure 33. Distribution of clusters in different game contexts

Looking at Figure 33, some clusters seem to be equally distributed in home and away games but clusters such as 6, 8 and 9 only seem to occur in Home Games, indicating that Arsenal tends to get players higher up the pitch in home games to press the oppositions's defenders, influenced by the home crowd and pitch familiarity. Clusters 11 and 3 also feature more often in home games indicating frequent buildup possessions to control the tempo of the game with forwards dropping deep and long balls to wide spaces.

Opponent strength seems to be an important factor into how Arsenal approaches games and develops their possessions. As expected, cluster 7 appears more against bottom 10 teams, indicating that weaker opponents tend to adopt a more compact and defensive approach which finds arsenal in situations with more possession of the ball and is trying to stretch the opposition's defense with diagonal passes to wide areas. Clusters 6 and 9 are more frequent against the bottom 10 teams indicating that Arsenal tends to player higher up the pitch against weaker opponents to press the opposition's defense. Cluster 14 seems to appear more against weaker opponents while cluster 4 appears more against top 10 teams indicating that against weaker opponents Arsenal is more confident in building up possessions from the goalkeeper and controlling the game's tempo with longer possessions. On the other hand, against stronger opposition Arsenal favors quicker forward passing to minimize the risk of losing the ball in dangerous areas. Against stronger teams, possession is more contested and evenly distributed indicating a more defensive approach.

Looking at the distribution of cluster occurrence in different time segments of games, Cluster 3, which involves long-range playmaking from the goalkeeper or defenders targeting wide players with long balls, primarily occurs during the early stages of games. This tactic could be applied to counteract team's high-pressing strategies. By quickly transitioning the balls to wide areas, Arsenal capitalizes on the spaces left behind and leverages the high energy levels of players during the first few minutes.

It's also clear that there are distinct patterns that reveal how the current score of the game affects Arsenal's possession tactics. Clusters 6, 8 and 9 have similar principles and appear more in situations where Arsenal is winning the game, trying to pressure opposition's defenders high up the pitch to force errors and maintain control of the game. This shows that Arsenal maintains an offensive posture even after gaining the lead, seeking to extend their lead rather than adopting a purely defensive strategy. Clusters 2, 0 and 14 display similar patterns, characterized by longer possessions that initiate from defensive zones. These plays typically involve a

strategic build-up, either progressing through or by utilizing the full width of the pitch by either leveraging wing-backs or wingers to get in behind of the opposition's back line. These clusters appear more in scenarios where Arsenal is losing but especially when the score is tied, suggesting a tactical preference to retain control of the game while actively seeking to gain the lead. On the other hand, cluster 11 is more frequent when Arsenal is losing. This indicates that in such situations, forward players tend to drop deep to help in maintaining possession. This pattern suggests difficulty in advancing the ball into the final third with opposing teams likely adopting a more defensive posture when taking the lead against Arsenal. Finally, we can see that cluster 5 possessions, which involve attacks developing in the left flank and crosses in the box occur when Arsenal is losing, indicating that Arsenal relies on left-sided players to get in attacking positions and threaten the opposition when trailing.

Overall, the analysis shows that Arsenal's ball possession strategies are highly varying and adaptive based on factors like opponent strength, match location and current score. As expected, at home or against weaker teams, Arsenal adopts a more aggressive, possession-centric approach, while tougher away games may see a more conservative style. When being pressed, Arsenal utilizes the full width of the pitch, exploiting spaces in wide areas by getting the full-backs in attacking positions, whereas their build-up play is more centralized with wide players being less involved.

4.4 Arsenal's Underlying Passing Style Themes that Reveal their Distinctive Possession Strategies Across Various Game Situations.

4.4.1 Analyzing Arsenal's Possessions Using Word Cloud Insights

To gain insights into the key topics related to Arsenal's possessions, we constructed word clouds that highlight the 15 most frequently mentioned terms. Our analysis categorizes Arsenal's possession tactics into eight distinct topics.

Figure 34. Possession Topic Wordclouds

Topic 1 - Controlled Buildup: This strategy involves patient build-up play from the back, utilizing defenders in long possessions using the entire width of the field with both forward and backward passing, aiming to maintain possession through a stable defense.

Topic 2 - Compact Defense: This strategy focuses on a defensive and compact central setup, characterized by short, low-activity possessions mainly developing in the defensive areas and marked by frequent interceptions and difficulties in maintaining possession.

Topic 3 – Long-Range Playmaking: This topic involves possession where goalkeepers and defenders executing long passes in wide areas to exploit spaces behind the opposition's line.

Topic 4 – Offensive Flank Play: Emphasizes possessions in attacking areas, with plays often developing in wide areas, especially on the left, to deliver crosses or cut inside and shoot. These possessions mainly start at in attacking areas indicating that Arsenal plays with a high defensive line to pressure the opposition.

Topic 5 – Midfield Focus: This topic revolves around simple and short passes from midfielders in central areas, aiming to control the tempo of the game and transition the ball forward.

Topic 6 – Right-Side Attacks: Describes possessions occurring in situations where Arsenal intercepts the ball in dangerous areas and attempts quick attacks from the right flank to surprise disorganized defenses.

Topic 7 – Direct Forward Play: This topic represents fast transitioning with more direct passing. Forward players potentially drop deep to help out to either start counter-attacks or exploit spaces when the opposing team plays with a high defensive line and Arsenal intercepts the ball.

Topic 8 – Dynamic offensive play: Characterised by a fast, aggressive style of play with smart, risky passing, indicating a focus on rapid transitions and line-breaking movements by creative midfielders. These possessions tend to occur more frequently in counter-attacks.

4.4.2 Comparing the Frequency of Arsenal's Possession Topics Across Various Game Situations

Figure 35 shows that possession-oriented strategies such as controlled buildup (topic 1) and midfield focus (topic 5) are slightly more common in home games, suggesting a greater control and confidence in home games. On the other hand, defensive and long-ball strategies (topics 2 and 3) show a reverse trend, indicating a more conservative approach where Arsenal does not prioritize keeping possession when playing away. Offensive flank play (topic 4) also seems to occur more in home games, indicating that Arsenal is much more aggressive in exploiting wide areas in home games.

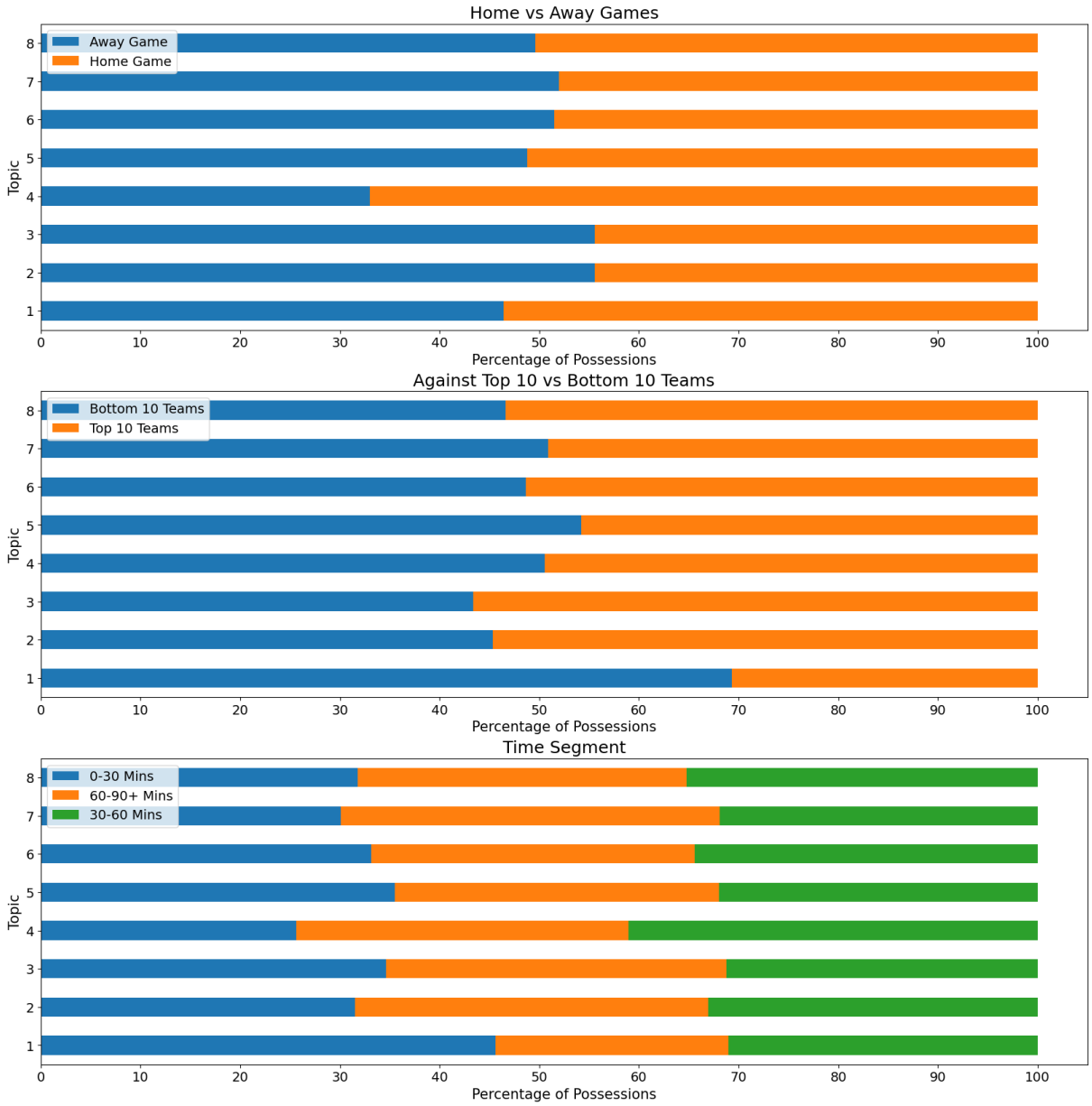


Figure 35. Distribution of topics in different game contexts

It's also clear that Arsenal adapts their passing strategies depending on their opponent. Strategies that prioritize possession maintenance such as controlled buildup (topic 1) and midfield focus (topic 5) occur a lot more frequently against weaker opponents while more defensive strategies prioritizing long-balls and counter-attack (topics 2, 3 and 8) show an opposite trend. This indicates that Arsenal aims to control the tempo of the game against weaker opponents, whereas against strong teams Arsenal implements a more strategic posture focusing on exploiting wide spaces after absorbing pressure or counter-attacking opportunities.

The distribution of possession topics at different time segments in the game shows how Arsenal adjusts its strategy based on the match phase, fitness levels and perhaps the scoreline. We can

see that controlled buildup (topic 1) is more prevalent during the earlier stages of the game, indicating that Arsenal attempts to maintain possession from the get-go aiming to set the tempo of the game without taking excessive risks to assess the opposition's setup. Offensive flank play (topic 4) increases during the second 30 minutes, indicating a shift to more aggressive tactics as the game opens up, finding more space on the flanks. Finally, direct forward play is dominant in the later stages of the game indicating a strategy which aims to exploit the opponent's fatigue as spaces might open up. As players tire, it's easier to catch the defenses off-guard with fast attacks becoming more efficient.

Finally, to provide a visual representation of Arsenal's tactical adaptations and the frequency of each strategic topic throughout the season, we have created Figure 36 which shows the frequency of each topic by gameweek. We can also see the match outcome and result, the team that Arsenal faced in each gameweek, whether the game was at home or away, and the final position of the opponent (FP) in the end of the 2017-18 season.

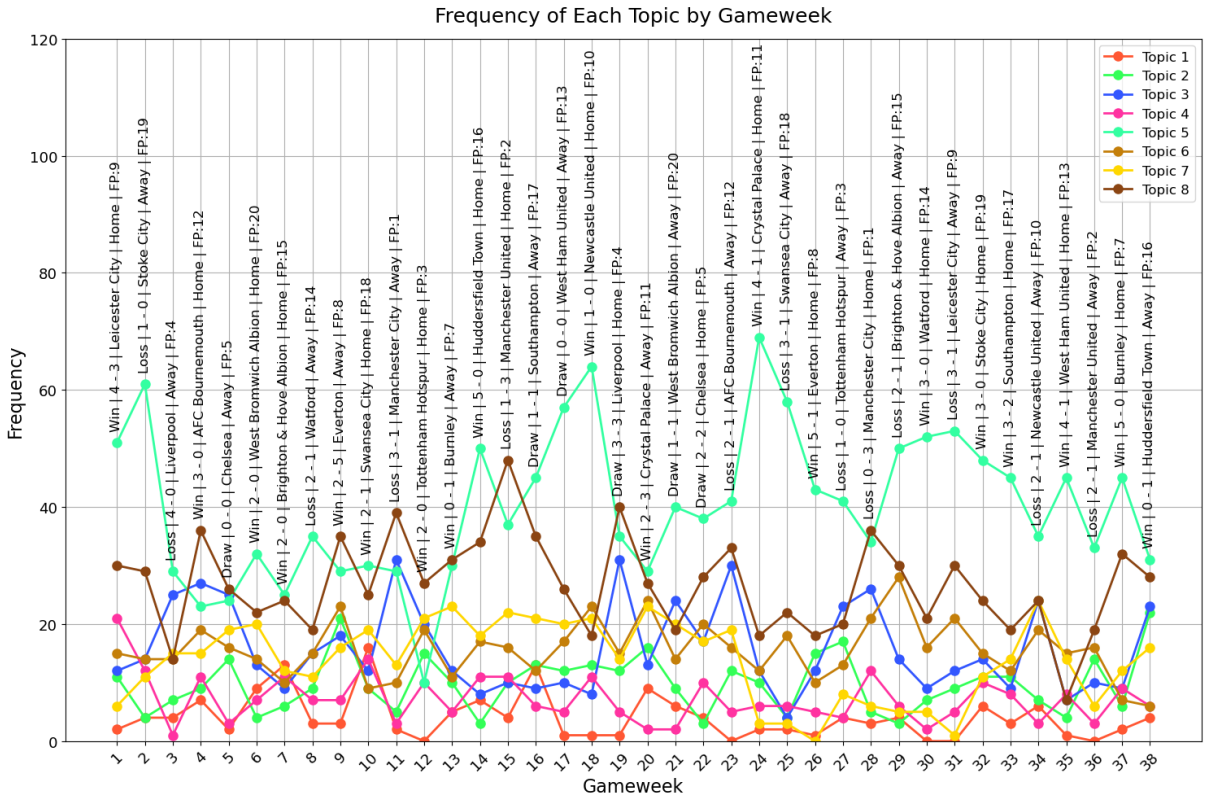


Figure 36. Frequency of topics by gameweek

Midfield focus (topic 5) seems to be the most frequently used passing strategy but with significant spikes. This indicates a high, adaptive use of midfield-driven play to control the game. Increases in frequency could suggest games where controlling the midfield was particularly crucial. Dynamic attacks (topic 8) also seem to be a frequently used strategy,

especially when facing top teams. It seems that whenever Arsenal emphasizes dynamic attacks, they reduce focus on midfield control especially against stronger opponents as we saw in Figure 35. Overall, these two possession topics seem to be the core strategies that Arsenal implements but in different game situations. The rest of the topics seem to be less frequent and more adaptive into certain situations. Long-range playmaking (topic 3) seems to be used inconsistently throughout the season but alongside dynamic attacks indicating their similarities. Topics 4 and 6, which represent possessions that start and develop in wide attacking areas when Arsenal successfully presses and intercepts the ball, are used less frequently and show spikes throughout the season, indicating their adaptive use in specific situations such as home games, as seen in Figure 35 for topic 4. Controlled buildup (topic 1), compact defense (topic 2) and direct forward play (topic 7) show similar spikes throughout the season and seem to be used more occasionally in certain situations.

Overall, the strong presence of strategies that focus on midfield dominance in central areas highlights the key roles that midfielders like G. Xhaka and M. Ozil have in Arsenal's passing strategy as a team that prefers to distribute the ball in central areas. The trends suggest that Arsenal is more confident in keeping possession at home games and against weaker opponents by adopting a more possession-focused and offensive approach, particularly with controlled buildup and wide play. In away games and against stronger opponents Arsenal tends to change their tactics and apply a more conservative approach, relying mainly on long-range playmaking and counter-attacks. Arsenal is a team that focuses on maintaining possession early in matches to dictate the pace, gradually then shifting to quicker attacks by using the full width of the pitch, showcasing their robust stamina and physical power.

4.5 Arsenal's On-ball Passing Patterns and Themes that are Most Relevant to their Success.

Clusters 6, 8 and 9, which are similar and define possessions that develop in situations where forward players press the opponent's defenders in the attacking third and force turnovers, seem to have the highest success rate based on how often they were used in games won. This indicates that Arsenal is very effective in these situations. Cluster 14 which is characterized by build-up possessions starting from the goalkeeper on the left flank shows the lowest win ratio and mainly occurs in games lost, indicating potential struggles attacking from that side of the field.

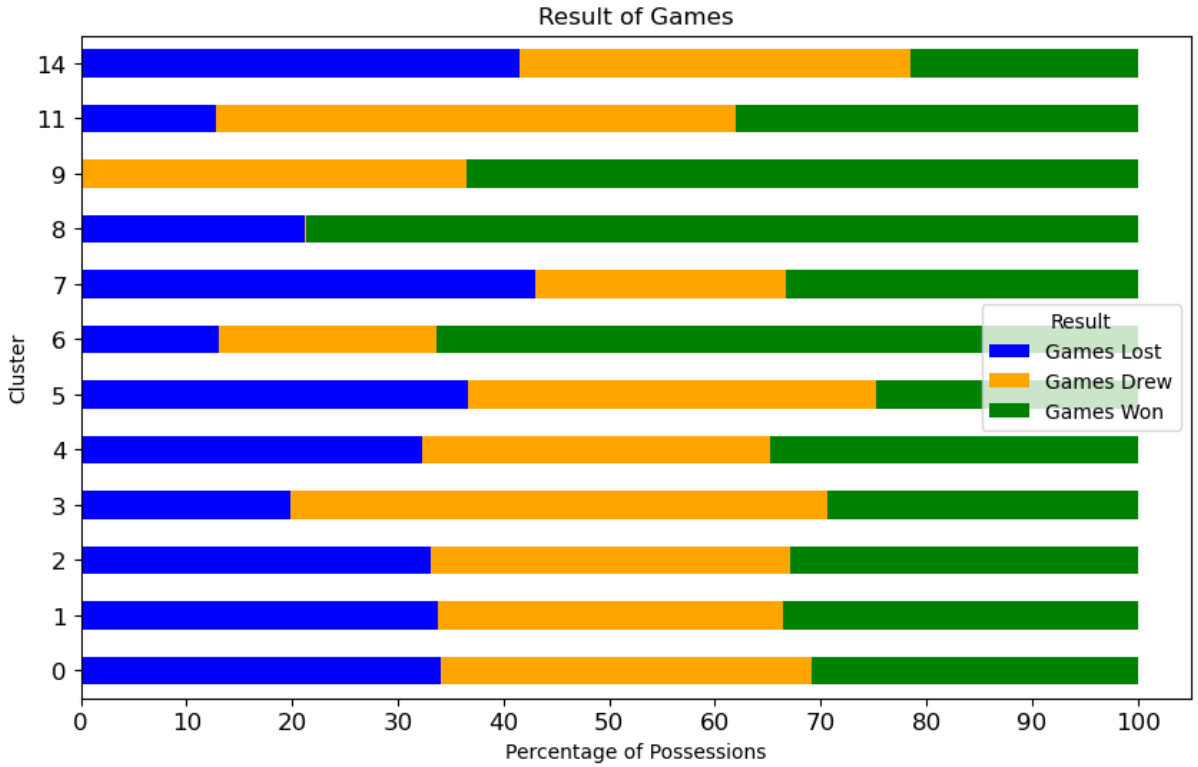


Figure 37. Distribution of clusters based on results of games

Figure 38 shows that clusters 6, 8 and 9 often lead to shots, indicating that Arsenal is a dangerous team when re-gaining possession in attacking areas but with the risk of attacks often leading to offsides. Clusters 0, 1 and 3 which represent long-ball strategies seem to be less effective with many attacks often leading to non-accurate passes, indicating a general struggle of completing long balls. Clusters 2, 14 and 11 which represent build-up possessions starting from defensive areas and moving centrally, show an often struggle of ball progression with many possessions leading to either non-accurate passes or duels, putting Arsenal at risk.

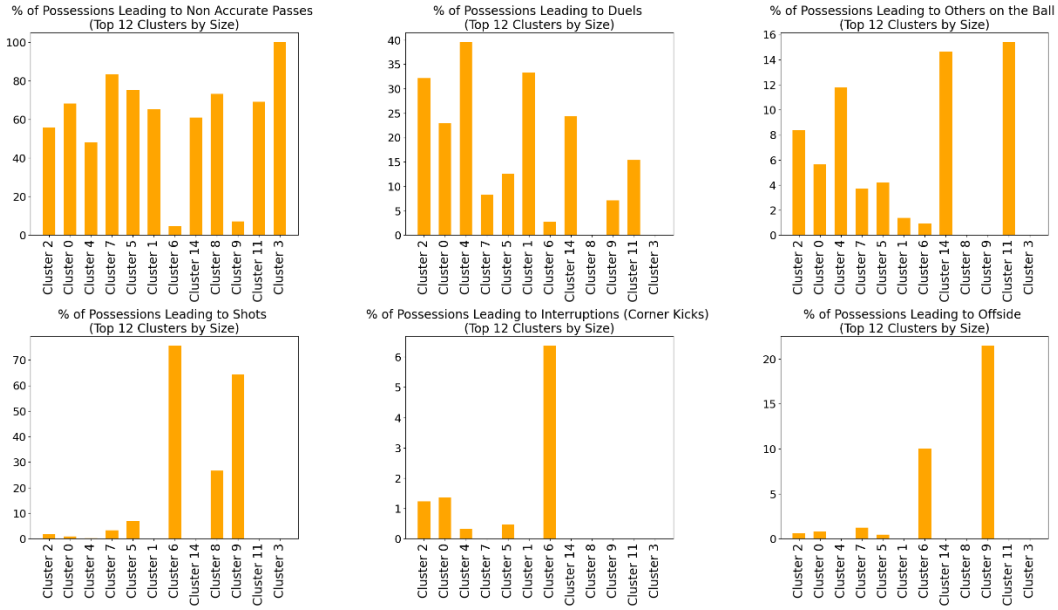


Figure 38. Distribution of cluster frequency across different possession outcomes

Figure 39 shows the win rate of each possession topic. Topics 4 and 1, which represent possessions developing from the flanks either starting from defensive or attacking areas, seem to have the highest win percentage, while topic 3 which is characterized as long-range playmaking from the goalkeeper and the defenders has the lowest. The frequency for the rest of the topics seems to be more equally distributed across the three possible match outcomes.

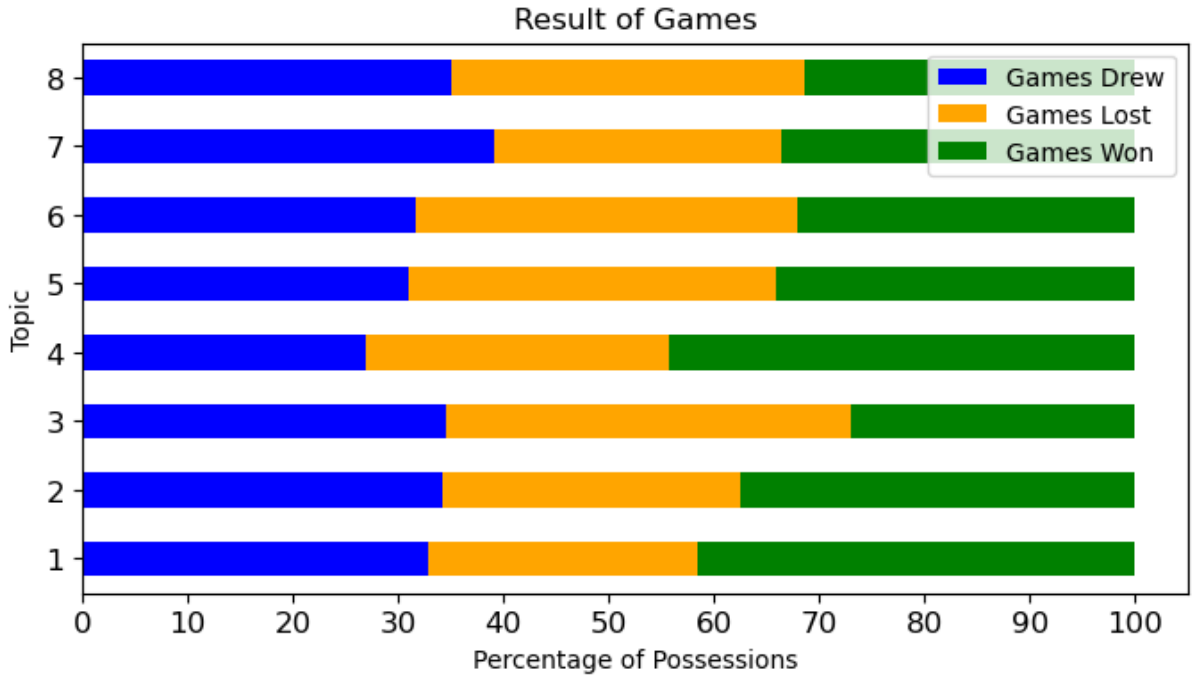


Figure 39. Occurrence of possession topics in different game outcomes

Overall, it is evident that Arsenal is more effective when re-gaining possession in attacking areas. This indicates that Arsenal is efficacious at counter-pressing high up the pitch, with the attacking midfielders being excellent at initiating quick transitions. On the other hand, long-ball passing from the defensive lines seems to be one of Arsenal's weaknesses with most possessions leading to non-accurate passes. This indicates that Arsenal's goalkeepers and defenders tend to struggle with the ball on their feet. During build-up play possessions, Arsenal is more effective when the ball is distributed across the full width of the pitch with wing-backs and wide forwards getting more involved, rather than when moving centrally. This highlights the ability of their wide players to exploit space and stretch the opponent's defense, while providing accurate cross into the box.

5. Discussion

5.1 Objective Assessment

The first objective of identifying the roles of key players that influence a team's playing style and how players interact with passes in different game situations was completed with the use of network graphs and factor analysis of player passing statistics. For this project, our main goal was to explore the tactical adjustments Arsenal made during the 2017-18 season based on the venue, the opponent strength, the current score during games and in different time segments to see whether Arsenal employs distinct passing strategies based on these contexts or if they maintain a consistent approach regardless of these factors. This led us to compare the passing networks across multiple games that describe a certain context (e.g. home vs away games).

The problem with this approach is that different players might be used game-by-game which could lead to misjudgments. Taking this into consideration, we included the 11 players with the highest number of minutes played between them and to construct the difference graphs we grouped players by position using the 'roleCluster' column. This ensured that we included as much data as possible and we always compared the same players. By identifying patterns across different game contexts such as changes in ball circulation, key passing routes or player involvement, coaches can adjust their strategies to optimize performance in specific scenarios, enhancing decision-making and their preparation for future matches.

To make fair comparisons of how Arsenal players are associated with the passing factors, goalkeepers were excluded as their role is unique and the distribution of their passes significantly differs from those made by players in other positions. Using statistics that indicate the types of passes made by players alongside their description and their spatial coordinates on the pitch provided us with an in-depth analysis of how each individual is being utilized and how they contribute in Arsenal's passing strategy. In our case where we only compare Arsenal players, the number of rows was limited to just 27 with a total of 23 variables. According to Quan, Li and Chen (2024) the ratio between the rows and the variables when applying factor analysis should be at least 2/1 for the comparisons to be accurate. To address this issue, we applied PCA and reduced the dimensionality of the dataset to 15 variables while maintaining 95% of the variance. Comparing our approach to similar studies that apply factor analysis, most studies use factor analysis to uncover underlying factors between players of different teams in similar positions while our approach is a bit different as our objective is to compare the distribution of passes between players in the same team, providing a focused insight into the unique strengths and roles within the team itself.

The overall objective results produced top-level findings that helped us understand how players are utilized in various game situations based on their factor scores and the playing styles observed from the network graphs.

- The network graphs in away games and against the top 10 teams suggest that Arsenal uses long-ball strategies in comparison to home games and against weaker opponents where defenders have more passing interactions between them indicating that Arsenal builds up possessions from the backline.
- Players who are highly associated with long-range playmaking (factor 3) are more useful in away games and against stronger opponents.

The second objective is designed to address the first as we aim to analyze the positional tendencies of players. We first illustrated how the average positions of players differ in various game scenarios and then constructed passing heatmaps of players to enrich the findings concerning players' skillsets by providing a spatial dimension to the passing factors identified between players. Completing this objective provided us with some interesting findings not only at an individual player level but also about the overall tactical adjustments. These findings highlight the tactical use of players and can guide coaching decisions, opposition analysis and player development.

- Although players interact differently across various game contexts, their positioning is rather consistent. This suggests that Arsenal's tactical adjustments are more dynamic and strategy-oriented while maintaining a consistent formation.
- Arsenal's two main full-backs, H. Bellerin and S. Kolasinac seem to have similar roles and are highly associated not only with their defensive responsibilities but also with possession maintenance and counter-attack football. Looking at their heatmaps, most of their passes are being made in the final third, which indicates that their main responsibility is to provide width in Arsenal's attacks by overlapping and stretching the opponent's defense.
- Although I. Monreal's main position is Left-Center-Back, his heatmap suggests that he is a versatile player who can also play as a Left-Back but seems to complete more passes in the defensive third indicating that he is being used more as a traditional left-back, prioritizing his defensive duties, in comparison to H. Bellerin.

The third and fourth objectives are designed for the macro-level analysis of Arsenal's passing dynamics by grouping event sequences into possessions and exploring the key patterns and

themes of Arsenal's passing strategy. The 'teamId' column helped us group consecutive passing events of the same team to ensure that the events of different teams were not mixed up. We also excluded possessions that had only one event since we aim to explore the patterns of consecutive passing sequences. Clusters and topics of possessions were formed based on the spatiotemporal dynamics of the passing sequences, the distribution of different pass types and the involvement of player positions. The number of clusters and topics extracted show high variability and prove that football is a random game.

- 28 possession clusters were formed
- 31% of possessions were not assigned to a cluster and were treated as noise
- 8 possession topics were formed

When interpreting each cluster and topic, our main goal was to think about the situations in which they would occur during a game. This approach aligns with the methodology and theoretical framework outlined by Andrienko N., Andrienko G., and Shirato (2023) who aim to analyze the sequence of multi-attribute variables that describe events from the moment a team wins possession until they lose it by providing insights into where, when and under what conditions the patterns occur. This approach was very effective with some of the most significant insights being the following:

- Many possessions tend to start in the attacking third, which could indicate that Arsenal tends to counter-press the opposition's defense to force mistakes.
- Some possessions show high contribution of goalkeepers, forwards and long passes, which indicate and often use of long-ball strategies.

The next step was to compare how frequently the main clusters and topics occur in home compared to away games, when playing against weaker and stronger teams, in different time segments during the games and based on the current score of the game. We notice that possessions occur inconsistently between the contexts indicating their situational use.

- Arsenal's clusters and themes that focus on long-ball strategies occur way more in away games and against stronger opponents.
- On the other hand, in home games and against weaker opponents Arsenal prioritizes maintaining possession and building possessions from the back.

These insights can be assessed by professionals and coaches to fine-tune their tactics and overall strategy. By understanding the patterns and themes in Arsenal's possession episodes, they can

identify key trends such as when Arsenal is more likely to counter-press or utilize long-ball strategies. This can help coaches adjust their training sessions and prepare for different game situations to either exploit Arsenal's weaknesses or counter their strengths.

Our final objective was to assess the success rate of the most significant patterns and themes in Arsenal's possession episodes by analyzing their outcomes and how frequently they occur in games won, lost, or drawn. This objective aligns with one of the main goals of this study which is to address and provide insights into the ongoing debate of whether possession-based football or direct football is more effective in the modern game, with many studies having different views on the subject. To identify the outcome of each episode, we extracted the last event of the sequence before Arsenal lost possession using the 'eventName' column while also looking at the 'tags_description' column which provides us with more details about the event and helped us interpret the outcome of the possessions. Some of the most interesting findings from this part of the study are the following:

- Arsenal seems to be more effective using a hybrid of possession-based football and direct football, with a great emphasis on counter-pressing and quick transitions than a purely possession-based approach like 'Tiki-Taka'.
- Arsenal is most effective and creates more scoring opportunities, when regaining possession in attacking areas by exploiting disorganized defenses through few, quick and forward passes.
- The above insight suggests that Arsenal is suited to a more direct, fast-paced style with their attacking midfielders excelling in rapid transitions.
- Possession-based football seems to be effective for Arsenal in certain situations especially when facing weaker opponents who tend to sit deep to maintain a more compact defensive structure.
- In situations such as the above, Arsenal is more effective in building possessions using many lateral and backward passes, with wide players getting highly involved to stretch the opposition's defense.
- Long-ball strategies from the defensive third seem to be a weakness with Arsenal struggling to create chances using this style of play.

Overall, Arsenal seems to thrive applying a balanced approach that mixes high pressing, quick transitions and controlled buildup from the wide areas, while appearing less suited to traditional 'direct football' that focuses on bypassing the midfield with long balls. On the other hand, it's

also evident from our findings that there are external factors that determine the success of both strategies. This proves that there is no passing strategy that is universally effective or guaranteed to work for every team in all situations. The key is to use each passing strategy appropriately, taking into consideration the context of the game and its specific dynamics. These insights are valuable not only for opposing teams' coaches and analysts to identify Arsenal's weaknesses and strengths, but also for Arsenal's coaching staff to refine their tactics by maximizing their strengths and improving on their weaknesses.

5.2 Answering the Research Question

Can Data Science Techniques Uncover Hidden Patterns and Dynamics in Football, and Provide Insights into the Skillsets of Players in Passing?

This paper aimed to help understand whether data science techniques can be used to explore the passing dynamics and strategies of a team and assess the skillsets of players in passing to help coaches and analysts make more informed, data-driven decisions on tactical adjustments, match preparation and player development. This approach seeks to minimize subjectivity in football decision-making by promoting data-driven insights to enhance team performance and produce better results. It is clear from the outputs discussed that the use of data science techniques can definitely help coaches and analysts make informed decisions that can give them an edge over their opponents. Overall, data science can be a valuable tool for explorative research in the area of football, especially for studying the passing dynamics.

6. Evaluation, Reflections and Conclusions

6.1 Overall Evaluation of the Project

Overall, the results of this study have been positive in leveraging data science techniques to explore team passing dynamics and evaluate players' skillsets. By analyzing passing sequences and identifying patterns and themes, we have been able to quantify the similarities in possessions and uncover strategic tendencies and relationships between players. Furthermore, the methods applied provided us with valuable insights into individual player performance by assessing the passing metrics of players and the distribution of their passes.

When comparing player performance in sports, many researchers often overlook the impact of playing time as players with more time on the pitch will naturally complete more passes which can skew comparisons and make it difficult to assess true performance levels. Similarly, when comparing the occurrence of different passing patterns and themes across different contexts, it is essential to consider the total number of passes in each context, so the analysis does not become biased, as teams tend to have more possessions in certain contexts. This paper was able to address these factors and ensure that the comparisons made are accurate and effective.

The graphs chosen to present the findings of this paper were premediated to ensure they could be easily interpreted not only by academics but also by sports enthusiasts who may not have specialized knowledge. We prioritized clarity and simplicity, using mainly visualizations such as tables, bar charts, scatterplots and line charts that can clearly highlight key patterns and insights from our analysis. By avoiding overly complex representations, we made the data accessible to a wider audience, ensuring that both experts and casual viewers can understand the results and apply them practically.

Choosing the appropriate datasets and processing them into a suitable structure took a significant amount of time. This meant carefully understanding table relationships to combine them into a single, merged dataset that captures information about the events, alongside player information who are relevant to the events, the teams involved and the fixtures and competitions in which the events occurred. Having seen the complexity of the datasets, time was adjusted to suit the project plan.

On an event level, the main methods used in this paper were network science to model the passing interactions between players and heatmaps to visually represent passing data intensity. On a possession-episode level, unsupervised learning techniques were used such as cluster analysis and topic modeling to identify hidden patterns or topics within the episodes and factor

analysis to uncover underlying factors that explain the variance in the data. The choice of these methods was predefined not only based on the structure and characteristics of the data but also informed by a thorough review of the literature in Chapter 2, where studies with similar goals have employed these techniques effectively. This ensured that the selected methods were well-suited for addressing the studies' objectives and providing meaningful insights. These methods did not significantly differ from those discussed in the proposal but during the process, for methods like cluster analysis, we needed to analyze literature that compares K-Means to Density-Based Clustering. Nandi (2022) explores the theoretical comparison of both algorithms helping us acknowledge that DBSCAN is more well-suited for this study due to the inherent nature of the game.

Labeling the findings of this paper as successful or not was a difficult thing to do. While the insights are data-driven and the methodology used was rigorous, interpreting them accurately was challenging as it requires specialized knowledge of football tactics and strategies. This is why the team selected was from a league with which there was already a strong familiarity. The events selected for analysis were all from a single season (2017-18 season) to simplify the evaluation process, as there are numerous studies and analyses available for that season, most notably FBREF[15] who provides us with detailed statistics of that season and Football London[16] who review Arsenal's season that year highlighting key players by providing a tactical analysis. While these studies helped us assess whether our findings were coherent, we ensured that we did not heavily rely on them, allowing the interpretations and conclusions of our study to remain original and based on our own analysis and point of view.

6.2 Lessons Learned

- In terms of the overall tactical insights we gained from this paper, one key lesson learned is the importance of comparing passing dynamics across different contexts, such as home vs away games or against stronger and weaker teams, due to the situational nature of the game. Tactics and player behaviours often adapt to certain situations during the game such as defensive pressure, the crowd influence or the state of the game. By analyzing these contextual differences, we gained a more detailed understanding of how Arsenal adjusts

[15] https://fbref.com/en/squads/18bb7c10/2017-2018/Arsenal-Stats

[16] https://www.football.london/arsenal-fc/news/arsenal-latest-news-ramsey-arteta-14654202

their passing strategies to suit different challenges and match situations, revealing their tactical flexibility.

- Another lesson learned is based on the use of factor analysis to identify the key variables that define the underlying factors in Arsenal's passing dynamics. The results show that Arsenal's passing strategy was built around structured, accurate and often simple passing, with a key focus on maintaining possession and controlling the attacking third of the field. Direct and risky passes played a smaller role indicating that Arsenal emphasized possession-based football rather than high-risk tactics.

- One of the technical hurdles we encountered during this study was defining the possession episodes. Choosing the parameters to define our episodes like touches, turnovers or shots played a crucial role and could significantly affect the clusters and topics formed. This indicates how domain-specific knowledge can influence the methodological choices in data science-related projects.

- Another lesson learned was that density-based clustering (DBSCAN) is much more suitable for football possession data than any other clustering algorithm due to its ability to treat irregular possessions as noise. In football, many possessions are either too short or too long, making them unique, with DBSCAN effectively separating these from typical patterns.

- When applying DBSCAN, fine-tuning the parameters like Nneighbors and the minimum distance for data points to be grouped in the same cluster is a crucial step to avoid over-clustering. This ensures that we extract meaningful clusters while considering football's variation and randomness.

- Deciding which topic modeling algorithm to use for extracting possession topics was important due to the structured nature of football event data. Topic modeling algorithms are commonly used in text data to uncover hidden themes within a large collection of documents. Football possessions lack a natural 'vocabulary', making feature engineering an essential step for transforming event data into a suitable structure for topic modeling.

- When comparing the two main topic modeling algorithms (Latent Dirichlet Allocation and Non-Negative Matrix Factorization), we found that NMF is more suited for extracting

themes from short texts such as football event data, due to its ability to handle compressed texts more effectively.

6.3 Proposals for Further Work

- Instead of focusing solely on Arsenal, future work could expand to compare multiple team's passing strategies during the same season. This would provide a broader context to evaluate how Arsenal's passing networks and patterns are compared to other top teams in the league and how successful they are. We could also compare Arsenal's passing networks and dynamics with their opponents' passing strategies, positional tendencies and defensive structure when playing against Arsenal. This will help us analyze how opponents set up defensively to block passing lanes, identify the most frequent passing lanes or perhaps understand how they adapt their pressing to counter Arsenal's strategies.

- Looking back to our approach for grouping similar possessions, one issue with our models is that passing patterns and themes are extracted based on solely passing statistics without considering external factors such as player movement or defensive actions. Further studies could expand the scope of the analysis to integrate player movement data to capture off-the-ball events, team shape during different phases of possessions and defensive actions such as tackles or pressure events alongside the passing data. Combining tracking data, defensive actions and passing data can offer a holistic view of the key patterns and themes that highlight how defensive positioning and player movement affect Arsenal's passing dynamics.

- Since the 'Events' dataset that we used for this paper includes events from both Arsenal and their opponents, one way that we could analyze Arsenal's passing networks simultaneously with their opponent's defensive actions is by expanding the possession episodes to include both teams' actions by perhaps considering a single episode to be a sequence of events that leads to a shot, a corner kick or a foul using the 'tags_description' column. By expanding the possession episodes to include both teams' actions, we could gain a deeper view on the tactical battle that unfolds during different game contexts. At the same time though, this would increase the complexity of the episodes requiring more advanced feature engineering methods to represent both attacking and defensive actions meaningfully to then create the passing networks and to cluster the episodes effectively.

- Further research could examine how Arsenal's passing networks and patterns differ when using various formations (e.g. 3-4-2-1 vs 4-2-3-1). This could reveal how the positional structure impacts passing networks and how player roles change while also assessing each formation's success.

- This paper looks at one season's worth of English Football League data and a key proposal could be to expand the datasets and include data of multiple, consecutive seasons. Knowing that the 2017-18 season was Arsene Wenger's last year as the manager of Arsenal, it would be interesting to see how Arsenal's passing dynamics changed over time, following the appointment of a new manager. This would reveal how the external factors of a new managerial philosophy and changes in the squad impact the core dynamics of Arsenal's passing game.

7. Glossary

7.1 Technical Terms

Word/Phrase	Definition
Big Data	Datasets that are too large or complex to be dealt with by traditional data-processing application software
Data Mining	The process of extracting and discovering patterns in large datasets
Machine Learning	A field of study in artificial intelligence related to the development of statistical algorithms that can learn from data and generalize to new, unseen data to perform tasks without any instructions
Unsupervised Learning	A framework in Machine Learning where algorithms learn patterns exclusively from unlabeled data
Cluster Analysis	The process of grouping similar data points with similar characteristics
Dimensionality Reduction	The transformation of the data from a high-dimensional space into a low-dimensional space while retaining meaningful properties from the original data
Principal Component Analysis (PCA)	A linear dimensionality technique for reducing the number of features in a dataset while capturing the largest variation in the data
Natural Language Processing (NLP)	The application of computational techniques for analyzing or synthesizing natural language or speech
Network Science	The study of complex networks represented by nodes and the connections between them
Topic Modeling	A field of study in Natural Language Processing for discovering hidden semantic structures in a text body or a collection of documents
Factor Analysis	A statistical method used to describe the variability among correlated features in a dataset with a potentially lower number of unobserved features called factors
Collinearity	When the predictors in a regression model are linearly dependent

Table 14. Glossary – Technical Terms

7.2 Football Terms

Word/Phrase	Definition
Expected Goals (xG)	A metric used to represent the probability of a scoring opportunity that may result to a goal
Expected Assists (xA)	A metric used to represent the probability that a completed pass will become a goal assist
Top 10 Teams	Teams that finished in the top 10 positions (1-10) of the 2017-18 English Premier League table
Bottom 10 Teams	Teams that finished in the bottom 10 positions (11-20) of the 2017-18 English Premier League table
Build-Up	Possessions that involve patient ball progression from the defensive to the attacking third, aiming to create scoring opportunities
Counter-Attack	An attack made in response to or in defence against an attack made by the opponent
Line Break	A pass that goes through a line of the opponent's team formation
High Pass	A pass in which the ball goes above shoulder height
Launch	A goal kick from the defending team after the ball goes out of bounds directly into their goal line, having last been touched from the attacking team
Others on the Ball	A turnover from the attacking team
Offside	An attacker is in an offside position if any of the body parts that can score a goal (excluding the hands and arms) are in the opponents' half of the pitch and closer to the opponents' goal line than the ball and the last outfield player
Holding Midfielder	The purest form of a defensive midfielder with a role to hold his position and stay close to the defenders
Box-to-Box Midfielder	A midfielder with a hybrid profile, neither strictly defensive or strictly attacking meaning he needs to equally contribute defensively and offensively
Full-Backs	The wide defenders in a back line of four or five defenders. These positions are occupied by the Left and Right Backs in a backline of four players and the Left and Right Wing-Backs in a backline of five players

Table 15. Glossary – Football Terms

8. References

8.1 Academic References

Alfajri, A., Richasdy, D. and Bijaksana, M.A. (2022) 'Topic Modelling Using Non-Negative Matrix Factorization (NMF) for Telkom University Entry Selection from Instagram Comments', *Journal of Computer System and Informatics (JoSYC)*, 3(4), pp. 485–492. Available at: https://doi.org/10.47065/josyc.v3i4.2212

Andrienko, N., Andrienko, G. and Shirato, G. (2023) 'Episodes and Topics in Multivariate Temporal Data', *Computer Graphics Forum*, 42(6). Available at: https://doi.org/10.1111/cgf.14926

Buldú, J.M. *et al.* (2018) 'Using Network Science to Analyse Football Passing Networks: Dynamics, Space, Time, and the Multilayer Nature of the Game', *Frontiers in Psychology*, 9. Available at: https://doi.org/10.3389/fpsyg.2018.01900

'Clustering High-Dimensional Data: A Reduction-Level Fusion of PCA and Random Projection' (2019) in Pasunuri, R., Venkaiah, V. C., and Srivastava, A., *Advances in Intelligent Systems and Computing*. Singapore: Springer Singapore, pp. 479–487. Available at: https://doi.org/10.1007/978-981-13-1280-9_44

Dergaa, I. and Chamari, K. (2024) 'Big Data in Sports Medicine and Exercise Science: Integrating Theory and Practice for Future Innovations', *Tunisian Journal of Sports Science and Medicine*, 2(1), pp. 1–13. Available at: https://doi.org/10.61838/kman.tjssm.2.1.1

Du, M. and Yuan, X. (2021) 'A survey of competitive sports data visualization and visual analysis', *Journal of Visualization*, 24(1), pp. 47–67. Available at: https://doi.org/10.1007/s12650-020-00687-2

Eusebio, P., Prieto-González, P. and Marcelino, R. (2024) 'Decoding the complexities of transitions in football: a comprehensive narrative review', *German Journal of Exercise and*

Sport Research [Preprint]. Available at: https://doi.org/10.1007/s12662-024-00951-9

Hughes, M. and Franks, I. (2005) 'Analysis of passing sequences, shots and goals in soccer', *Journal of Sports Sciences*, 25(5), pp. 509-514. Available at: https://www.tandfonline.com/doi/abs/10.1080/02640410410001716779

'Informative or Misleading? Heatmaps Deconstructed' (2009) in Bojko, A., *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 30–39. Available at: https://doi.org/10.1007/978-3-642-02574-7_4

Li, W. and Sang, G. (2018) 'Principal Component Analysis on Football Competitions via Linear Model in China Football Association Super League Tournament', in *Proceedings of the 2018 International Conference on Management, Economics, Education and Social Sciences (MEESS 2018)*. *2018 International Conference on Management, Economics, Education and Social Sciences (MEESS 2018)*, Shanghai, China: Atlantis Press. Available at: https://doi.org/10.2991/meess-18.2018.12

Lolli, L. *et al.* (2024) 'Data analytics in the football industry: a survey investigating operational frameworks and practices in professional clubs and national federations from around the world', *Science and Medicine in Football*, pp. 1–10. Available at: https://doi.org/10.1080/24733938.2024.2341837

Quan, T., Li, X. and Chen, S. (2024) 'Exploratory Factor Analysis: A Pentagonal Evaluation Model Based on Football Player Stats', *Applied Mathematics and Nonlinear Sciences*, 9(1). Available at: https://doi.org/10.2478/amns.2023.2.01450

Rahmah, N. and Sitanggang, I.S. (2016) 'Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra', *IOP Conference Series: Earth and Environmental Science*, 31, p. 012012. Available at: https://doi.org/10.1088/1755-1315/31/1/012012

Reep, C. and Benjamin, B. (1968) 'Skill and Chance in Association Football', *Journal of the Royal Statistical Society. Series A (General)*, 131(4), p. 581. Available at: https://doi.org/10.2307/2343726

Rein, R. and Memmert, D. (2016) 'Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science', *SpringerPlus*, 5(1). Available at: https://doi.org/10.1186/s40064-016-3108-2

Tavakol, M. and Wetzel, A. (2020) 'Factor Analysis: a means for theory and instrument development in support of construct validity', *International Journal of Medical Education*, 11, pp. 245–247. Available at: https://doi.org/10.5116/ijme.5f96.0f4a

Vermeulen, E. (2018) 'Big data in sport analytics: applications and risks', *South Africa* [Preprint]

Yin, L. *et al.* (2023) 'Improvement of DBSCAN Algorithm Based on K-Dist Graph for Adaptive Determining Parameters', *Electronics*, 12(15), p. 3213. Available at: https://doi.org/10.3390/electronics12153213

Zhou, W. *et al.* (2023) 'An Improved Passing Network for Evaluating Football Team Performance', *Applied Sciences*, 13(2), p. 845. Available at: https://doi.org/10.3390/app13020845

8.2 Online References

Bala Priya, C. (2023) *Topic Modeling Tutorial – How to Use SVD and NMF in Python*. freeCodeCamp [Online]. Available at: https://www.freecodecamp.org/news/advanced-topic-modeling-how-to-use-svd-nmf-in-python [Accessed: 15/08/2024]

Bate, A. (2021) Pep Guardiola exclusive interview: *Possession football is the secret of Man City's defensive success.* Sky sports [Online]. Available at: https://www.skysports.com/football/news/11662/12202414/pep-guardiola-exclusive-

interview-possession-football-is-the-secret-of-man-citys-defensive-success [Accessed: 22/07/2024]

Bush, S. (2024) *HOW TO USE PASSING NETWORKS IN SOCCER*. American Soccer Analysis [Online]. Available at: https://www.americansocceranalysis.com/home/2024/6/20/how-to-use-passing-networks-in-soccer [Accessed: 01/08/2024]

Dougramaji, M. (2023) *Data Analytics in football: How LFC Used Data to Gain the Edge*. Rockborne [Online]. Available at: https://rockborne.com/graduates/blog/data-analytics-in-football-lfc/ [Accessed: 25/07/2024]

LatentView (2022) *Redefining Football Player Scouting with Predictive Analytics*. [Online]. Available at: https://www.latentview.com/blog/redefining-football-player-scouting-with-predictive-analytics/ [Accessed: 22/07/2024]

Microsoft Support (2024) *Guide to table relationships*. [Online]. Available at: https://support.microsoft.com/en-us/office/guide-to-table-relationships-30446197-4fbe-457b-b992-2f6fb812b58f [Accessed: 18/08/2024]

Nandi, M. (2022) *Density-based clustering*. Domino [Online]. Available at: https://domino.ai/blog/topology-and-density-based-clustering [Accessed: 09/08/2024]

Nassoori, J. (2022) *AC Milan's 'Mind Room': The story behind an innovative psychology lab.* BBC Sport [Online]. Available at: https://www.bbc.com/sport/football/61245521 [Accessed: 21/07/2024]

Nishida, K. (2018) *An Introduction to Principal Component Analysis (PCA) with 2018 World Soccer Players Data*. Medium [Online]. Available at: https://blog.exploratory.io/an-introduction-to-principal-component-analysis-pca-with-2018-world-soccer-players-data-

810d84a14eab [Accessed: 06/08/2024]

Pappalardo, L. and Massucco, E. (2019) *Soccer match event dataset* (Version 5). Figshare, Collection [Online]. Available at: https://figshare.com/collections/Soccer_match_event_dataset/4415000/5 [Accessed: 01/07/2024]

Ritchie, D. (2020) *The History of Sports Analysis: The Man Who Ruined English Football.* Nacsport [Online]. Available at: https://www.nacsport.com/blog/en-gb/News/the-history-of-sports-analysis-the-man-who-ruined-english-football [Accessed: 20/07/2024]

Robbins, N. (2012) *When Should I Use Logarithmic Scales in My Charts and Graphs?* Forbes [Online]. Available at: https://www.forbes.com/sites/naomirobbins/2012/01/19/when-should-i-use-logarithmic-scales-in-my-charts-and-graphs/ [Accessed: 21/08/2024]

Sekan, F. (2023) *A short history of data analysis in football.* Medium [Online]. Available at: https://medium.com/@filip.sekan/short-history-of-data-analysis-in-football-ce1963e428ae [Accessed: 20/07/2024]

SoccerTAKE (2023) *Soccer Analytics: How Data is Changing the Game*. [Online]. Available at: https://www.soccertake.com/performance/soccer-analytics-how-data-is-changing-the-game [Accessed: 19/07/2024]

Zamani, A. (2024) *The Guardiola Paradox; Possession Perfection or Stifling the Beautiful Game?*. Medium [Online]. Available at: https://medium.com/@amoszamani6/the-guardiola-paradox-possession-perfection-or-stifling-the-beautiful-game-5e58bc6edd82 [Accessed: 23/07/2024]

Appendix A. Project Proposal

Dissertation Proposal

Visual Analytics of Events in Football Games – Analysis of Passing Strategy

Author: Nicholas Tsioras

Supervisor: Dr Gennady Andrienko

1. Introduction

1.1 Problem

Today, football is one of the most popular sports and has become very interesting from a scientific point of view. Thanks to the availability and ability of sensing technologies, football teams have easier access to precise records of spatio-temporal events and statistics such as passes, shots or tackles in a game. In the modern era, football has become a complex interplay of strategies, tactics and skills. This has made football analytics an indispensable tool for clubs, coaches, and players. As technology continues to evolve, data collection methods have become more sophisticated and therefore the role of analytics in football is constantly growing.

In football, understanding the dynamics of passing is crucial for coaches and players to understand the overall tactics of their opponents. With football constantly changing, watching games and videos is not enough to understand the passing strategies of teams and it needs a structured approach which combines data analytics with football knowledge. Coaches and analysts are constantly seeking ways to gain competitive advantage, and one very effective way of doing that is to analyze and understand passing patterns and networks. With the volume and variety of data increasingly rapidly on football events, it has become very difficult to effectively interpret and capture the complexities of passing interaction using only traditional methods, particularly in dynamic game situations where strategies and tactics evolve in real-time.

1.2 Purpose of the Project

The purpose of this project is to explore and visualize the passing dynamics and patterns of a single football team across various football matches. For this project, we will be using the publicly available dataset from Figshare which includes statistics from games from a single season for the top 5 European leagues and we will employ various analytical and modelling techniques, for feature extraction, dimensionality reduction and pattern recognition. Analyzing football quantitatively has been challenging due to the nature of the data and the limitations of it. However, the evolution of data science presents a crucial opportunity to analyze comprehensive time-series data at coordinate level for football analytics. The link to access the datasets is shown below:
Link for Datasets: https://figshare.com/collections/Soccer_match_event_dataset/4415000/2 .

1.3 Objectives

- Aggregate and preprocess raw data from football events, with a primary focus on passing events

- Process raw data and construct algorithms to extract relevant attributes and common themes from passing events, such as the passing type, timing, length, and outcome alongside the passer and receiver.
- Construct passing networks based on the main passing topics and unravel their structural underpinnings.
- Implement interactive visualization tools to analyze the passing network and any additional contextual information.
- Validate the efficacy of the proposed framework by selecting a particular team and applying the developed framework to extract passing events, construct passing networks, and analyze relevant attributes to uncover insights into passing patterns, team strategies, player roles and game dynamics.

1.4 Research Questions

Through this project, we will be looking at answering the following questions:

- How do passing strategies change across distinct game scenarios, such as whether a team is playing at home or away, and against different levels and types of opponents?
- What is the typical passing style of a team and how does that change over the course of a whole season?
- Which are considered to be the key players that influence a team's playing style?
- Which are the main factors that change the passing strategy and tactics of a team?
- How can the coaching staff utilize visual tools and exploit data analytics to understand passing data better and make effective decisions on their team?

1.5 Products of the Work

The main products of this study will include tools and insights that will help us understand the passing strategies in football games.

- We will create visual graphs that demonstrate the interactions of players for a certain team showing who passes to whom during matches, and when the passes happen by reconstructing the passing networks for a selected team.
- We will apply advanced modelling techniques such as topic modelling to identify the main 'topics' in the passing structure of a team and gain contextual insights into how various contexts influence the passing dynamics of a team across multiple games.
- Other modelling techniques such as principal component analysis (PCA), multidimensional scaling (MDS) and clustering will be used to gain a deeper understanding of passing patterns and tactical changes over different game situations. These techniques will simplify complex data to reveal hidden patterns and group similar passing events together.
- We will create interactive visualization dashboards to explore the passing networks across multiple games and how they change dynamically.

1.6 Beneficiaries

Three main types of people can massively benefit from this project:

- Coaches and analysts seeking deeper insights into the passing dynamics and strategies of the team we will be analyzing. This will help them refine their tactics, optimize player positioning, and create strategies that can exploit the team's weaknesses. Overall, this will help them bolster their teams' performance.

- Researchers in Sports Science who are interested in the application of data science and data analytics in the field of sports. The insights from this project can be used by researchers for future studies and advancements in sports analytics.
- Football enthusiasts and analysts who want to deepen their understanding of the game's strategic nuances. Through our project, football enthusiasts can dive into the tactical complexity employed by teams and players and appreciate the game more. The deep insights we aim to uncover can foster informed discussion and debates within the football community.

1.7 Project Scope

In this project, we aim to provide a comprehensive analysis of passing events in football matches. The analysis will cover all games of a single team from the top 5 European leagues and will focus on the main passing events and patterns while considering contextual factors such as home vs away games, the strength of the opponent and the timing of the passing events. While the primary focus of this project is to understand passing patterns and game context, we will use advanced modelling techniques to dig deeper into the data and make more sense of passing behaviours.

2. Critical Context / Literature Review

2.1 Literature Review

- **The importance of big data and visual analytics in football**

Rein and Memmert (2016) talk about the growing importance of big data in football analytics in the modern era and highlight advanced data collection and processing technologies have revolutionized the way football performance is being analyzed. They talk about how the modern game has changed with complex interplay of strategies and tactics, indicating the need for sophisticated analytical techniques that can extract meaningful insights from large datasets. This underscores the need to incorporate data science techniques into football analytics to improve the understanding of the game dynamics.

- **Understanding Tactical Actions and Advanced Network Science**

Clemente et al. (2016) present a comprehensive study that investigates the network structure and centralization tendencies from different professional teams in Spanish La Liga and the English Premier League. The methodological framework of this project aligns closely to our project in which we aim to construct and analyse passing networks. This paper offers a deep understanding of how the centralization tendencies can help us understand how different players and positions influence the passing dynamics. This is crucial for the project's objective the typical passing networks of a team and the main factors that influence tactical changes.
Buldu et al. (2018) provide an in-depth examination of using network science to analyse passing networks in football. This study provides a unique approach and identifies three types of passing networks based on the spatial positioning of players and the temporal evolution of events. These networks show not only who passes to whom, but also where on the field each pass happened providing a more comprehensive analysis of the passing dynamics of teams. Buldu et al. (2018) also explains how passing networks can be analysed at different levels by using topological scales such as Microscale, where each player is analysed individually focusing on their roles in the team, Mesoscale, which examines the interactions between a small group of players, and

Macroscale which look at the entire team as a whole. The paper evaluates the passing networks at different levels with different metrics that can help us understand how to identify key players and their roles in a team, how they contribute to the overall strategy and how the passing strategy of a team changes during a game.

Caicedo-Parada et al. (2020) is another paper that focuses on how passing networks can be used to understand tactical actions within the game. This study utilizes advanced network science methods, graph theory, statistics, topic modelling and big data analytics to provide a full analysis of passing networks in football. This paper offers a detailed and methodological review of the analysis of passing networks in football. The study's emphasis on using advanced performance indicators and network metrics aligns with the goals of our project reconstruct and analyze passing networks in different context. It also uses advanced modelling techniques to group similar passing events together, which aligns with the approach we will be using for our project. Similarly, Goncalves et al. (2017) examine how passing networks and positioning variables relate to match outcomes in youth football. This study focuses on how to use network centrality measures such as the closeness and betweenness of players along with spatio-temporal data to provide insights into how the spatio-temporal relationships of players influence team performance. This paper shows how network centrality measures can be used to identify key players and their roles in passing networks.

- **Visualization Techniques in Football Analytics**

Li and Sang (2018) while also Quan et al. (2024) explore the application of Principal Component Analysis (PCA), factor analysis and clustering to evaluate tactical strategies in football. In these papers, PCA is utilized to transform various performance parameters into a set of linearly uncorrelated variables known as principal components. This transformation is used to reduce the complexity of the data by reducing the dimensionality and highlighting only the most important factors that influence a match outcome. The methodologies and findings from this paper can be directly applied to improve our analysis. By using PCA and factor analysis, we can reduce the complexity of our dataset which includes various passing attributes such as pass length, changes in ball position, and other contextual information and it can help us identify the most significant factors influencing passing strategies.

Clarence Thunberg's Kalt's thesis (2024) explores the use of cluster analysis in football teams of the top 5 European leagues of the 2022-23 season. This study employs hierarchical clustering using Euclidean distance and Ward's algorithm to group teams based on different performance metrics. While the scope of this project is a bit different to ours, we can still use the methodologies and findings from this paper to enhance our analysis by using clustering techniques to categorize and interpret different passing strategies. In our case, clustering can be used to group similar passing sequences together, by analyzing different attributes such as passing length and timing. This will help us identify clusters of similar passing patterns and understand common passing strategies used by teams in different game scenarios.

- **Contextual Influence on Passing & Sequential Pass Analysis**

Andrienko et al. (2023) introduce the concept of 'episodes' to segment continuous multivariate time-series data into manageable intervals for detailed analysis. Their approach focuses on analyzing episode-based data to understand the multivariate dynamic characteristics across a set of episodes using topic modelling and interactive visualization techniques. By applying

similar techniques, in this project we aim to uncover and visualize patterns in football passing strategies by transforming episodes into ‘texts’ to extract interpretable patterns to reveal tactical insights from multivariate data. Following a similar approach to this project will enable us to understand context-specific dynamics of passing networks. Zhou et al. (2023), investigate the influence of contextual factors, specifically match location (home vs away games) and the strength of the opponent, on team tactics and passing dynamics. This study highlights the importance of taking event—specific factors into consideration when analyzing passing networks.

2.2 Conclusion

The critical context provided from these papers, sets the stage for our study on analysing passing networks. By examining them, we gained a clear understanding of why passing analysis is important in football, how contextual factors affect it, and what advanced techniques and methods should be used for an in-depth analysis. This literature review helped us shape the methods we will be using in our project and by highlighting the importance of our research for enhancing football tactics and player performance analysis.

3. Approaches: Methods & Tools for Design, Analysis & Evaluation

3.1 Data Collection and Preprocessing

The data for this project will be obtained from the Figshare repository and specifically from the collection ‘’Soccer Match Events Dataset’’. The datasets are provided in JSON format and includes detailed records of events from football matches, including the timing of player movements and actions captioned at a coordinate level. The datasets we will be using are the Events dataset, Matches dataset, Teams dataset, Players dataset and Competitions dataset, which all have at least one common column that can be used to merge them. To parse the JSON files and convert them into structured dataframes we will be using the Python programming language. We will be then cleaning the data to handle missing values by either removing any events that are missing important information or have any inconsistencies. To filter out the data for the selected team, we will be using the unique identifier ‘teamId’ variable which exists in both the Teams and Events dataset, ‘matchId’ which exists in both the Events and Matches dataset, ‘competitionId’ which exists in both the matches and competitions datasets and ‘playerId’ which exists in both events and players dataset which will allow us to assign the player names that are associated with each event. These variables will be used to merge the datasets either through Python or SQL to then make the process of filtering out events from matches concerning only the single team of interest. In this way, our dataset will be structured to include information about different events focusing on passes, together with information about the matches, competitions and players. Finally, after we have combined the datasets together, we will be removing any duplicates and inconsistencies. Additional features will be extracted from the existing ones, such as the pass length by calculating the Euclidean distance between the start and the end coordinates of each pass, the vertical and horizontal change which will be calculated from the difference in vertical and horizontal position respectively of the ball from the start to the end of the pass, and the pass angles, to gain insights into the directionality of the pass.

3.2 Analytical Methods and Visualizations

- **Passing Network Construction & Contextual Segmentation of Passes:** We will be constructing passing networks for a single team across multiple games where nodes will

represent players and edges will represent the passes between them. Also, the data will be segmented in different game contexts to compare the passing networks across these scenarios. By visualizing passing network graphs with nodes and edges, we can identify key players and significant passing routes. The contextual information that will be integrated into the passing networks to provide a more in-depth analysis includes:

  - **Match Location:** Home vs Away Games
  - **Opponent Strength:** Higher vs Lower Ranked Teams
  - **Timing of the Game:** Analyzing how passing strategies change every quarter of the game
  - **Player Roles & Passes Between Players:** Differentiate passes by defenders, midfielders and attackers to identify key players and roles
  - **In-Game Events:** Analyze the differences in the passing dynamics before and after a crucial event happens in a game (e.g. red card)

The networks will be analyzed in three main topological scales: The microscale, where each player will be analysed individually focusing on their roles in the team, the mesoscale, which examines the interactions between a small group of players, and the macroscale which look at the entire team (Buldu et al., 2018).

- **Analyzing the Similarities of Passing Sequences Through Clustering & Dimensionality Reduction:** To assess the similarity of passing sequences, we will first extract the relevant features from the Events dataset such as the pass coordinates, lengths, directions and timestamps. Then we will be using Principal Component analysis (PCA) or Multidimensional Scaling (MDS) to reduce the dimensionality of the data while also retaining the key patterns. Finally, we will use clustering algorithms such as k-means or hierarchical clustering to group similar passing sequences together while maintaining the contextual information of the events. This approach will help us identify common passing episodes and tactical similarities across different game contexts.
- **Defining Possession Episodes with Topic Modelling for Sequential Pass Analysis:** From the events dataset, we will extract the passing events and convert them into feature representations while including details such as the type, direction, outcome, and the coordinates of the pitch. Then we will identify possessions by grouping consecutive passes made by the same team, starting a new possession when the ball changes possession. Each pass will be converted into a token that includes relevant features such as the start and end positions, pass type and outcome. The tokens representing pass sequences will then be combined to represent a whole pass episode which will represent an entire possession alongside additional contextual information (e.g. home vs away games). These documents will then be used as inputs for applying Non-negative Matrix Factorization (NNMF) topic modelling to identify the main topics of ball possession where each topic represents a common passing pattern or strategy. Finally, we will explore how frequently each topic occurs in different contexts such as home vs away games or against stronger or weaker opponents. To visualize the topics identified through NNMF topic modelling we will be creating word clouds, heatmaps and bar charts to identify dominant topics and their common terms in different match segments.
- **Identifying Underlying Factors That Explain the Variations in Passing Behaviours:** To identify underlying factors that explain variations in passing behaviours, factor analysis will be applied to various technical statistics of football players to identify the main factors that underscore common passing styles such as aggressive passing, long-ball tactics, or

direct attack. Algorithms that can be used for factor analysis include Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA) to identify the relationships between the variables and the latent factors. These factors will be evaluated with the use of the pentagonal factor weighting method to categorize players' performance in each factor and analyze how they contribute in the team's playing style (Quan et al. 2024).

- **Player and Team Heatmaps:** We will visualize the spatial distribution of passes for individual players with interactive heatmaps to identify the main areas of influence.

3.3 Evaluation

The success of this dissertation will be assessed by data accuracy, the validity of our analysis, and practical insights. Data accuracy will be ensured by cross-checking with referencing sources and validating metrics. The effectiveness of the analytical methods that will be implemented will be evaluated through statistical tests and comparisons with other existing studies in sports analytics. Potential limitations will be recognised, and the assumptions made during the analysis will be clearly stated to provide context for the findings. By incorporating these evaluation methods, we aim to ensure the credibility, robustness and practical applicability of this dissertation.

3.4 Ethical, Legal and Professional Considerations

We must ensure that no personal information is disclosed and the data is anonymous. Also, we must ensure that the research will be completed without any chance of legal, professional or any ethical issues occurring. By reviewing the City University Ethics Review Questionnaire, we need to ensure that none of the questions require further action. Due to the nature of the data which is provided by Figshare, we can ensure that the datasets that are being used are compliant with the relevant regulations and there are no legal issues to be concerned about.

4. Work Plan

Figure 1 demonstrates the work plan process on a weekly basis with the use of a Gantt chart. The project will begin in late June 2024 and finish in late September 2024. If any unexpected deviations occur from this current work plan, it will need to be adapted to a new one. We aim to complete the writing of the report in around 2 months with a few weeks to spare in case any of the risks included in the risk assessment table occur.

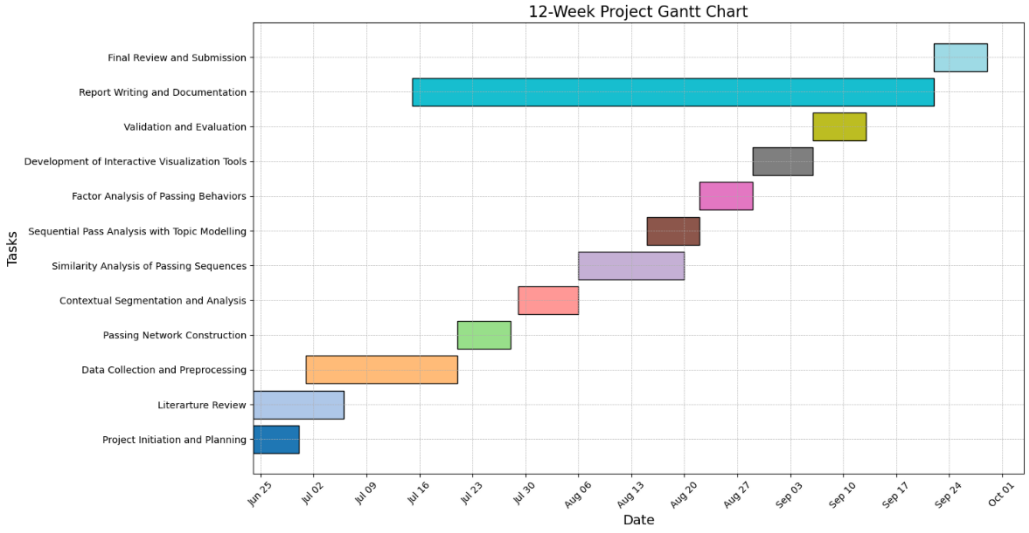


Figure 30. Gantt Chart of 12-Week Project Plan

5. Risks

Risk Description	Likelihood (1-3)	Consequences (1-5)	Impact (L x C)	Mitigation Strategy
Data Quality Issues	2	5	10	Use advanced data cleaning and preprocessing techniques
Insufficient Player Tracking Data	2	3	6	Ensure comprehensive data collection mechanisms and verify data completeness
Algorithm Performance Issues in Network Analysis	2	4	8	Use well-documented algorithms and test them thoroughly in various scenarios
High Dimensionality of Data	2	3	6	Apply dimensionality reduction techniques like PCA or MDS to manage and analyze the data
Clairty of Visualizations	1	4	4	Develop clear visualizations and provide comprehensive explanations
Misunderstanding With Supervisor	1	4	4	Consistent communication with supervisor to define researh goals and methods
Software or Tool Failure	1	5	5	Have alternative software and tools ready to use
Data Loss	1	5	5	Implement robust data backup and version control systems
Unclear Project Objectives	1	4	4	Regularly review and clarify project goals with the supervisor
Time Management	2	4	8	Develop a work plan and follow it
Computational Resource Limitations	2	5	10	Ensure access to sufficient computational resources and optimize code efficiently
Insufficient Contextual Analysis of Passes	2	3	6	Ensure thorough analysis of passes in different contexts (game scenarios)

Table 16. Risks Assessment

References

Andrineko, N., Andrienko, G., and Shirato, G., (2023). Episodes and Topics in Multivariate Temporal Data. *Computer Graphics Forum*, 42(6), el4926. Available at: http://dx.doi.org/10.1111/cgf.14926 . [Accessed: 17-05-2024].

Buldu, J.M., Busquets, J., Martinez, 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, Article 1900. doi: 10.3389/fpsyg.2018.01900 . [Accessed: 13-05-2024].

Caicedo-Parada, S.A., Lago- Peñas, C., & Ortega-Toro, E. (2020). Passing Networks and Tactical Action in Football: A Systematic Review. *International Journal of Environmental Research and Public Health*, 17(18), 6649. doi:10.3390/ijerph17186649. Available at: https://www.mdpi.com/1660-4601/17/18/6649 . [Accessed: 12-05-2024].

Clemente, F.M., José, F., Oliveira, N., Martins, F.M.L, Sousa Mendes, R., Figueiredo, A.J., Wog, D.P., & Kalamaras, D. (2016). Network structure and centralization tendencies in professional football teams from Spanish La Liga and English Premier Leagues. *Journal of Human Sport and Exercise*, 11(3), pp.376-389. Available at: http://www.redalyc.org/articulo.oa?id=301050351005 . [Accessed 15-05-2024].

Gonçalves, B., Coutinho, D., Santos, S., Lago- Peñas C., Jiménez, S., & Samplaio, J. (2017). Exploring Team Passing Networks and Player Movement Dynamics in Youth Association Football.

PLoS ONE, 12(1), e0171156. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0171156 .[Accessed: 12-05-2024].

Li, W., & Sang, G. (2018). Principal component analysis on football competitions via linear model in China Football Association Super League Tournament. *Proceedings of the International Conference on Management, Economics, Education and Social Sciences (MEESS 2018).* Advances in Social Science, Education and Humanities Research, volume 236. Atlantis Press, pp. 57-61. Available at: https://www.researchgate.net/publication/327191791_Principal_Component_Analysis_on_Football_Competitions_via_Linear_Model_in_China_Football_Association_Super_League_Tournament . [Accessed: 16-05-2024].

Quan, T., Li, X., and Chen, S., (2024). Exploratory Factor Analysis: A Pentagonal Evaluation Model Based on Football Player Stats. *Applied Mathematics and Nonlinear Sciences*, 9(1),pp.1-13. Available at: https://www.researchgate.net/publication/377548226 . [Accessed: 17-05-2024].

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. *SplringerPlus*, 5, Article 1410. Available at: https://springerplus.springeropen.com/articles/10.1186/s40064-016-3108-2 . [Accessed: 16-05-2024].

Thunberg Kalt, C. (2024). *Cluster Analysis on Football Teams Performance*. Bachelor Thesis in Statistics, supervised by Rauf Ahmad, January 8, 2024. Available at: https://uu.diva-portal.org/smash/get/diva2:1833447/FULLTEXT01.pdf . [Accessed 14-05-2024].

Zhou, W., Yu, G., You, S. and Wang, Z., (2023). An Improved Passing Network for Evaluating Football Team Performance. Applied Sciences, 13(2), p.845. Available at: https://www.mdpi.com/2076-3417/13/2/845 . [Accessed: 17-05-2024].

Research Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

http://www.city.ac.uk/department-computer-science/research-ethics

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:
PART A: Ethics Checklist. All students must complete this part.
The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.
PART B: Ethics Proportionate Review Form. Students who have answered "no" to all questions in A1, A2 and A3 and "yes" to question 4 in A4 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in such cases that are considered to involve MINIMAL risk. The approval may be ***provisional*** – *identifying the planned research as* likely to involve MINIMAL RISK. In such cases you must additionally seek ***full approval*** from the supervisor as the project progresses and details are established. ***Full approval*** must be acquired in writing, before beginning the planned research.

1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.* *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
3.1	Does your research involve participants who are under the age of 18?	**NO**

3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.* *For example, students studying on a particular course or module.* *If yes, then approval is also required from the Head of Department or Programme Director.*	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**

Appendix B. Supplementary Material

B1. Correlation Heatmap of Arsenal Players' Passing Statistics

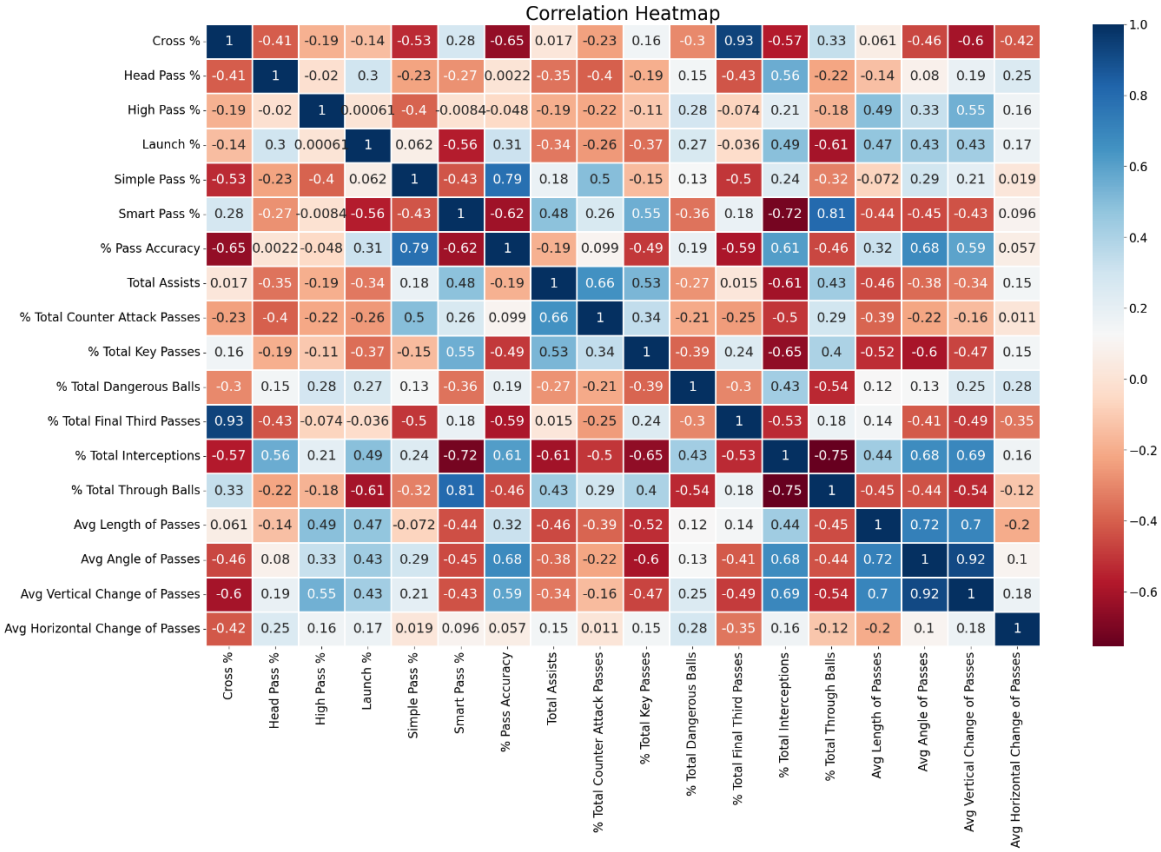


Figure 41. Correlation heatmap of passing statistics for Factor Analysis

B2. Occurrence of Possession Topics by Month

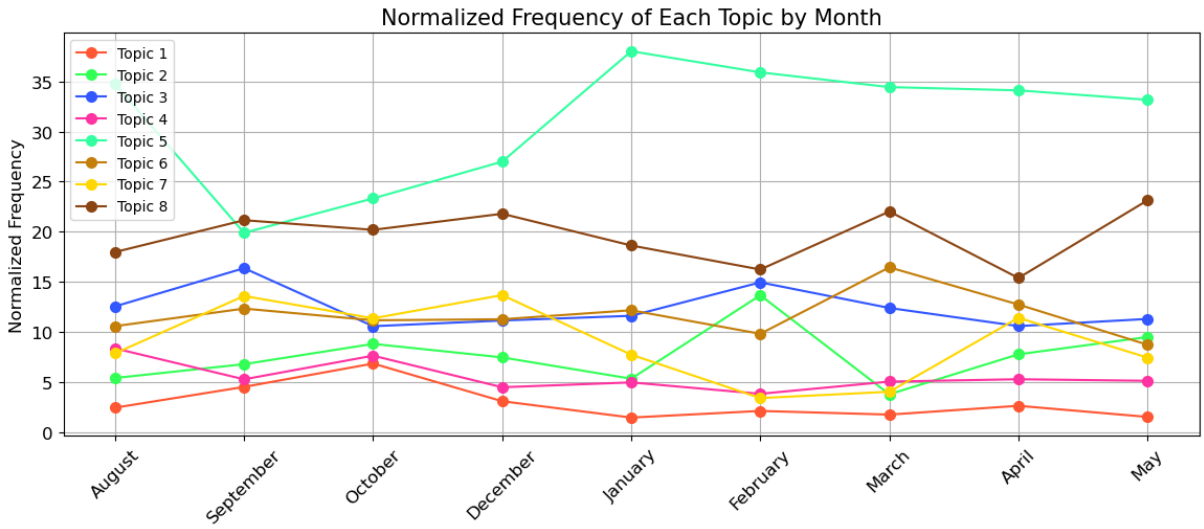


Figure 42. Normalized frequency of each topic by month

B3. Passing Networks by Player Roles in Different Game Contexts

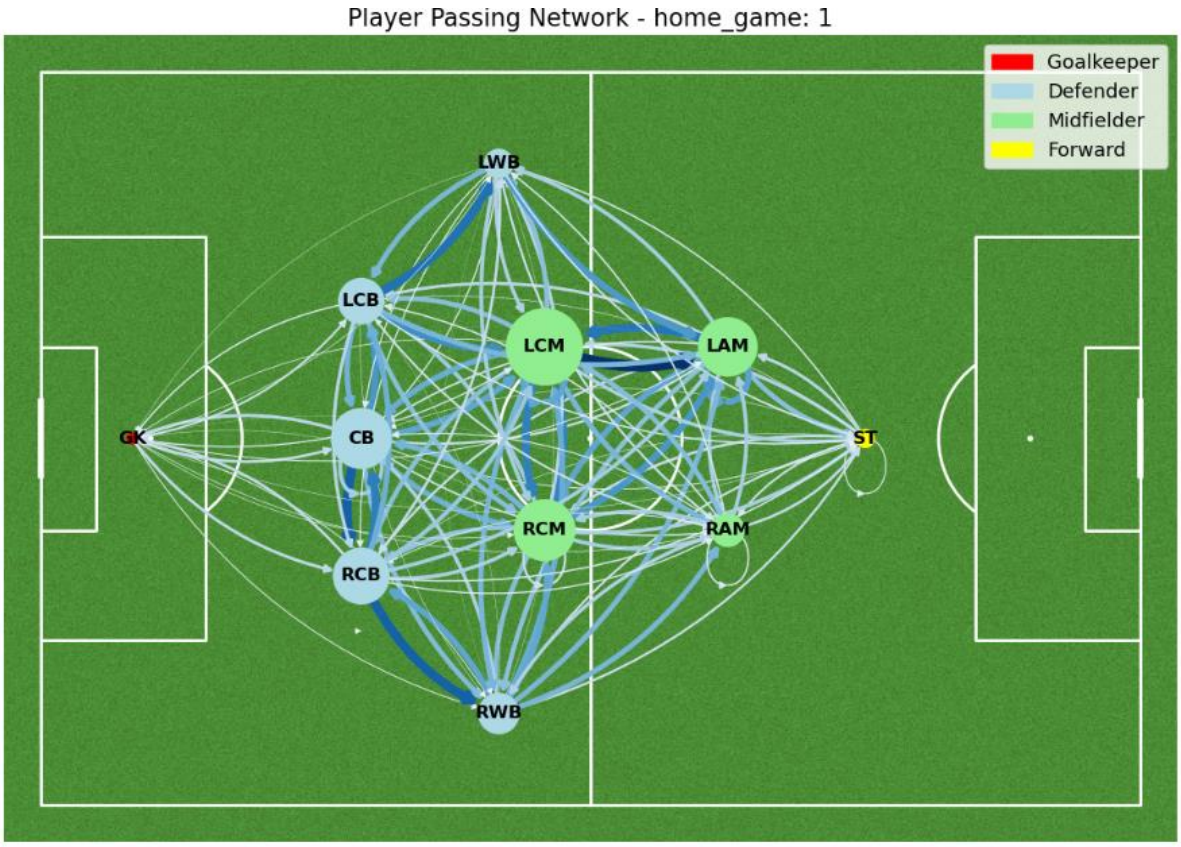


Figure 43. Passing Networks by Player Roles – Home Games

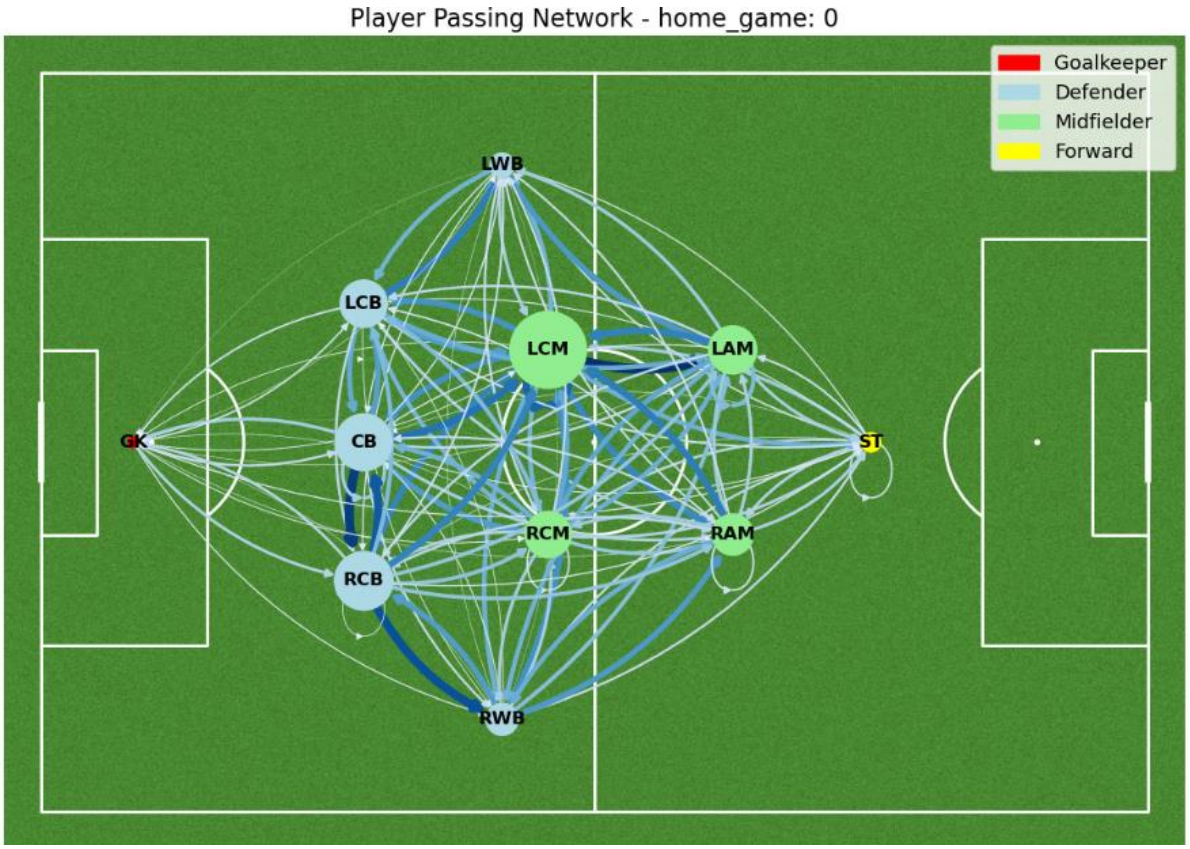


Figure 44. Passing Networks by Player Roles – Away Games

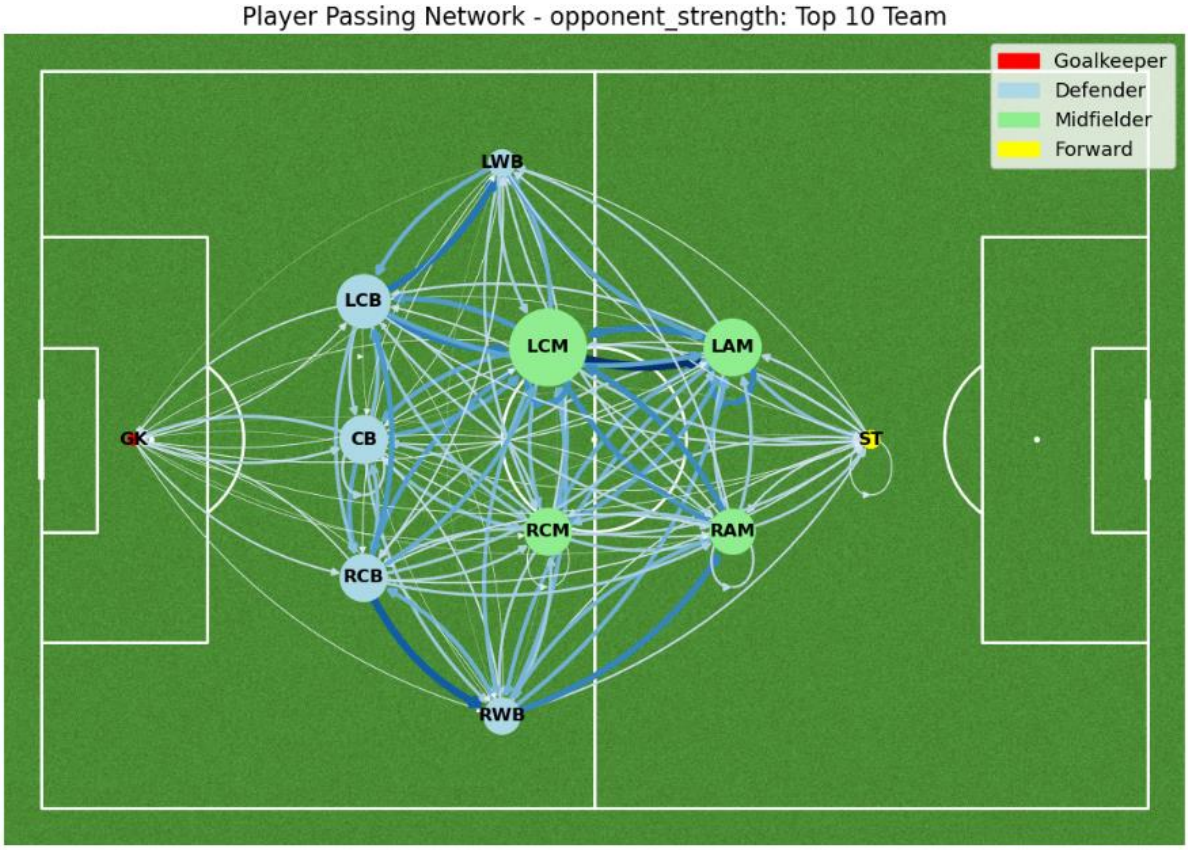


Figure 45. Passing Networks by Player Roles – Against Top 10 Teams

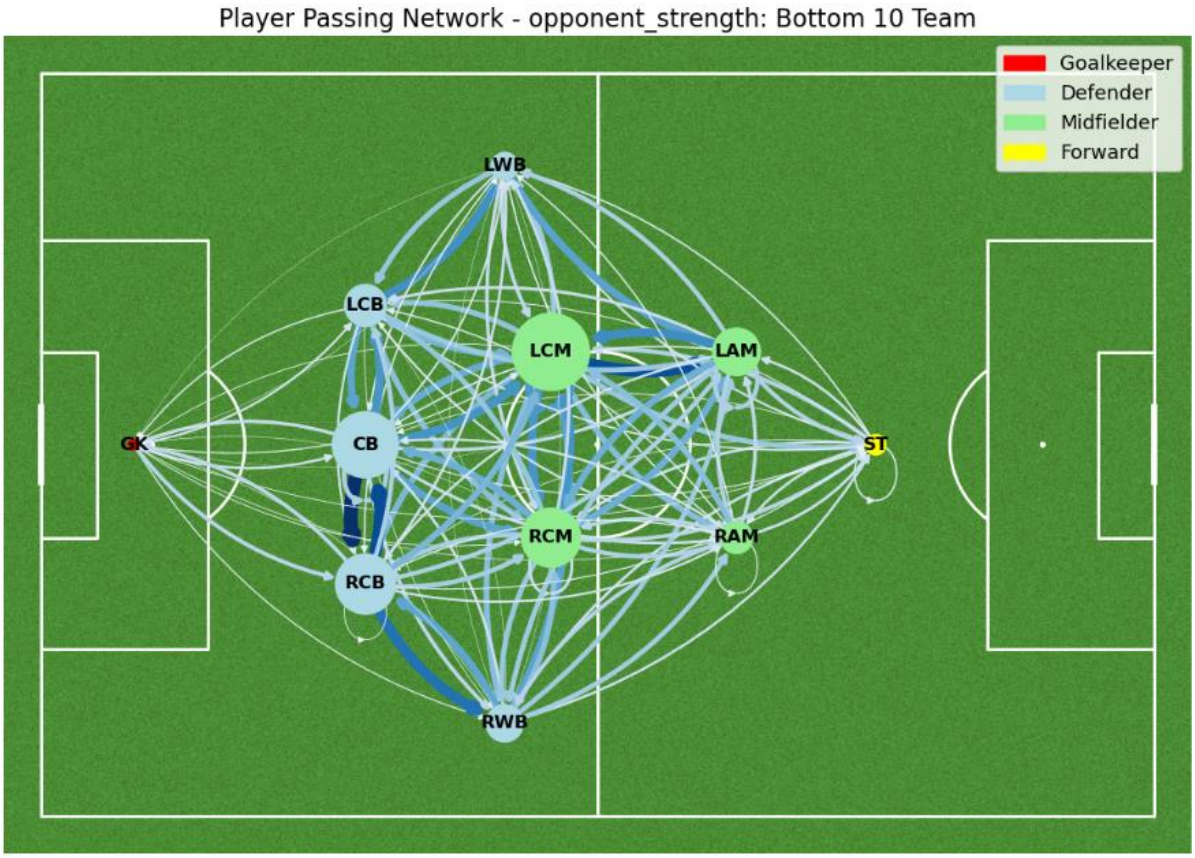


Figure 46. Passing Networks by Player Roles – Against Bottom 10 Teams

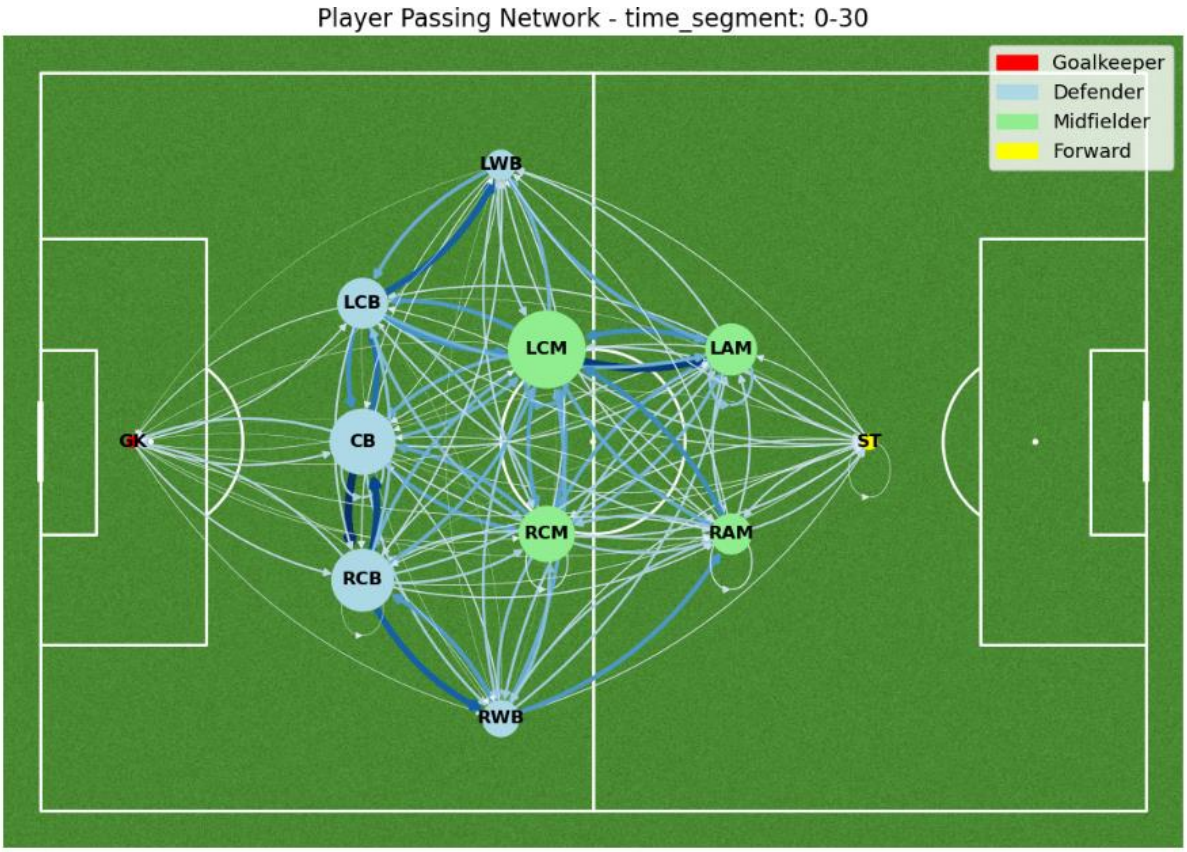


Figure 47. Passing Networks by Player Roles – During 0-30 Mins of Games

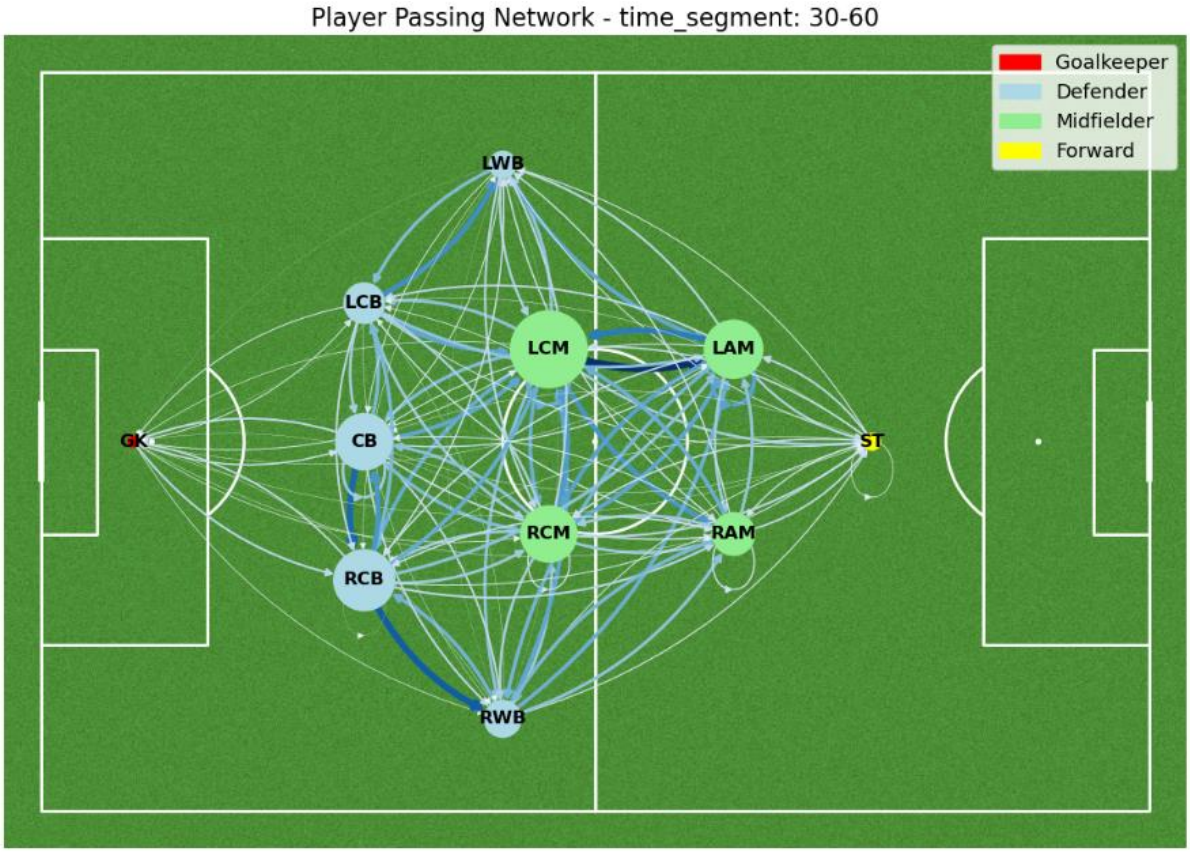


Figure 48. Passing Networks by Player Roles – During 30-60 Mins of Games

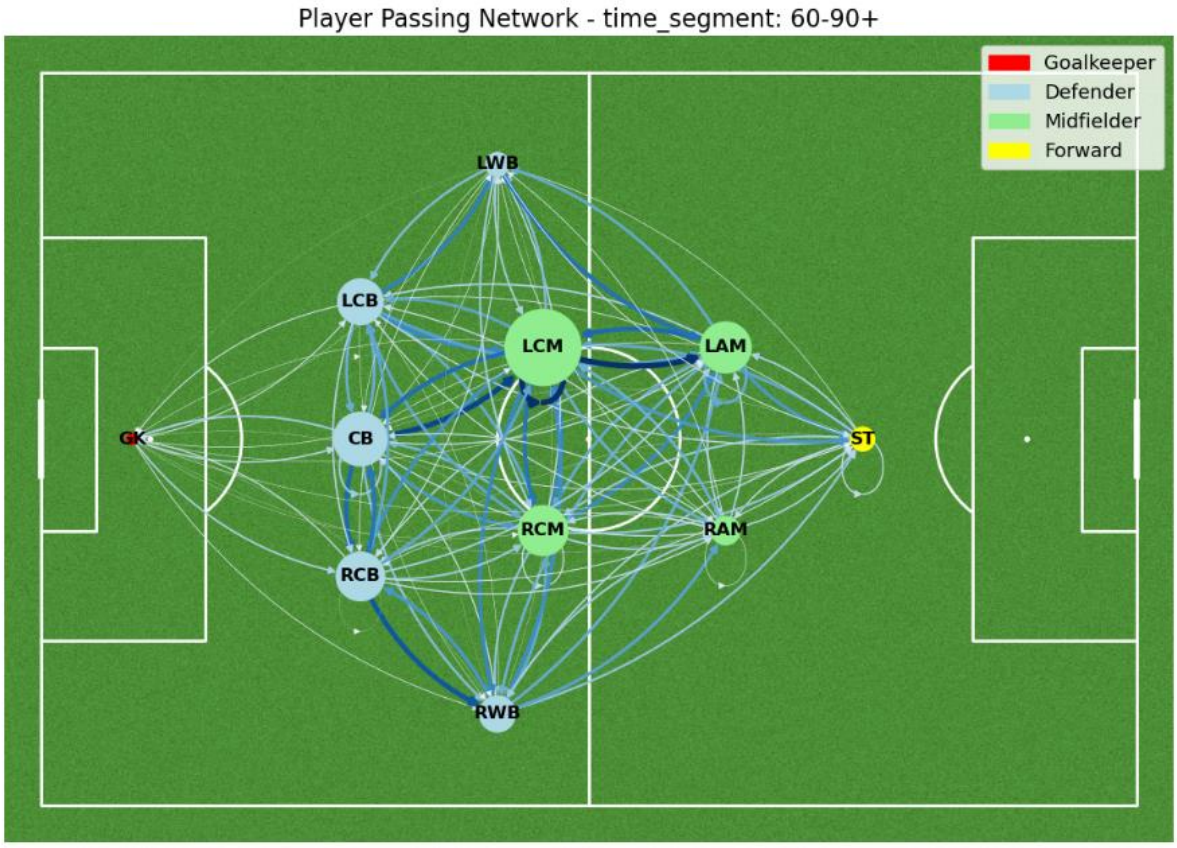


Figure 49. Passing Networks by Player Roles – During 60-90+ Mins of Games

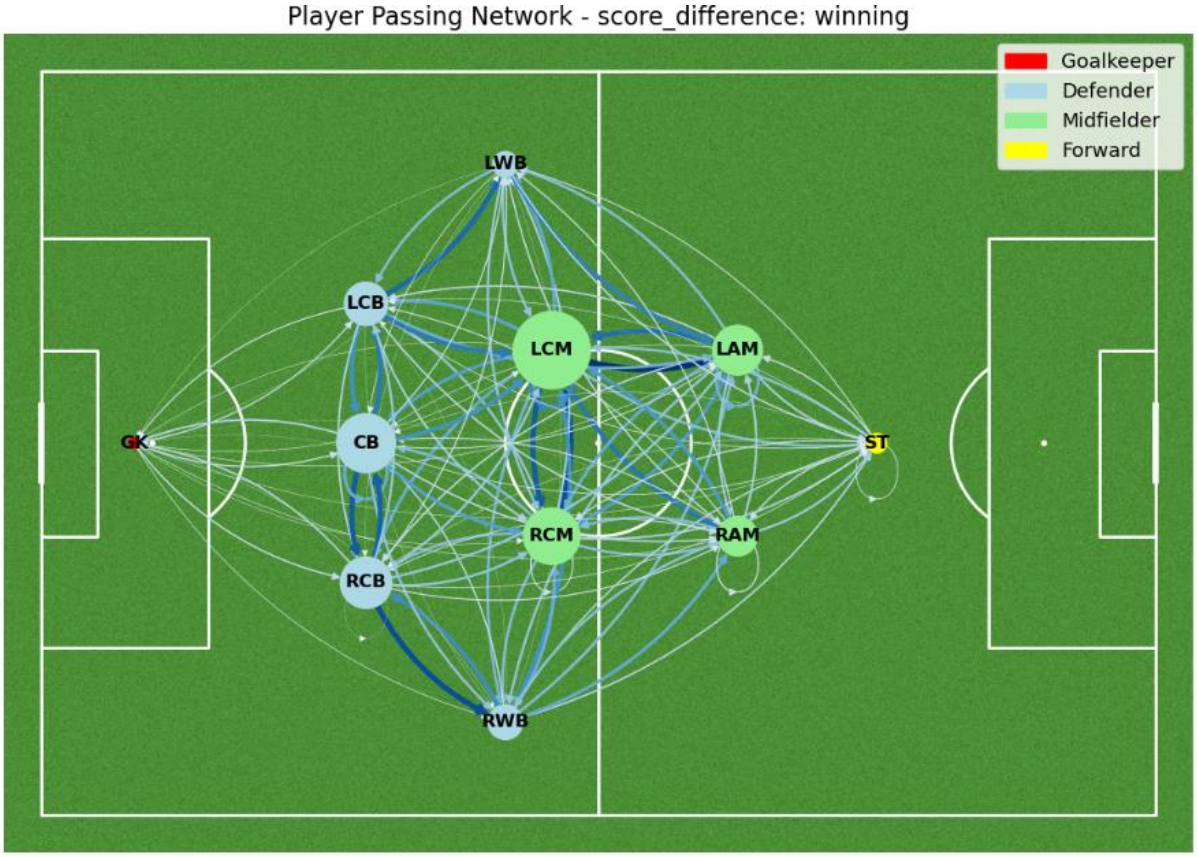


Figure 50. Passing Networks by Player Roles – When Winning During Games

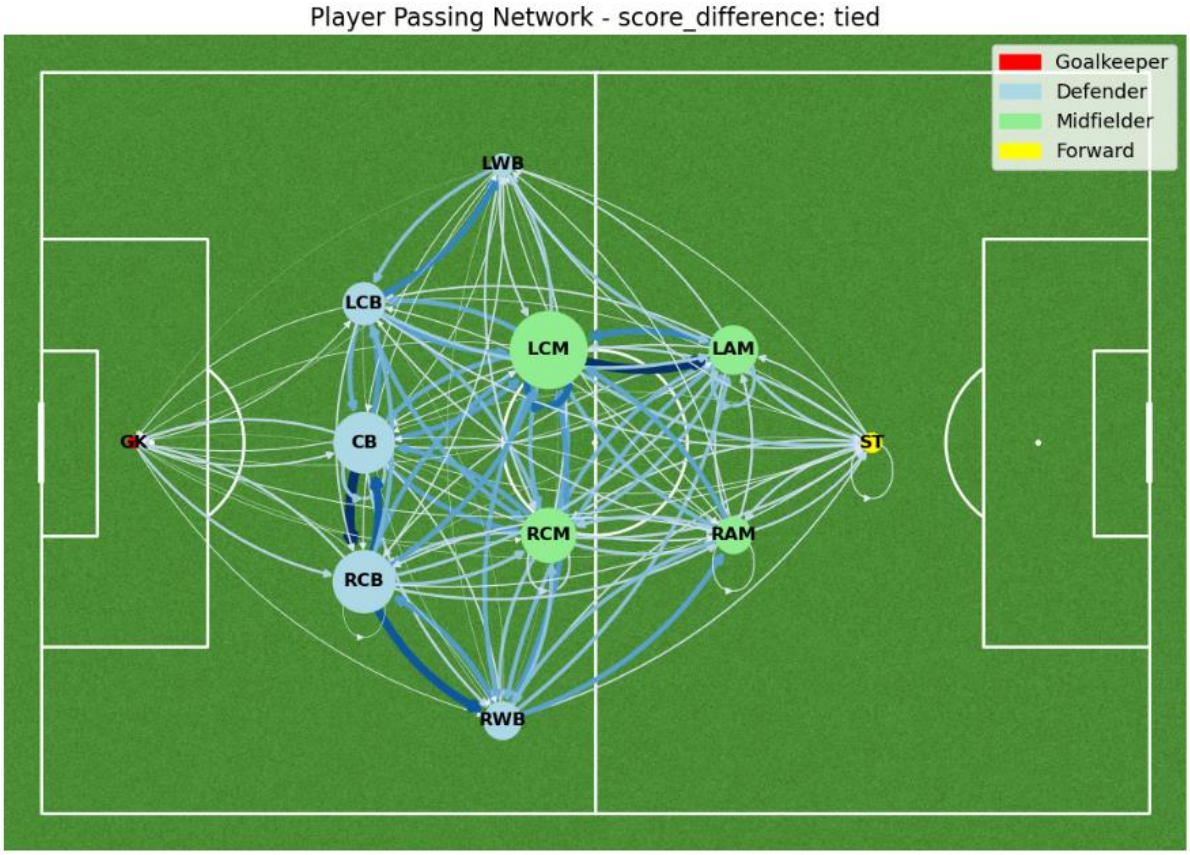


Figure 51. Passing Networks by Player Roles – When Tied During Games

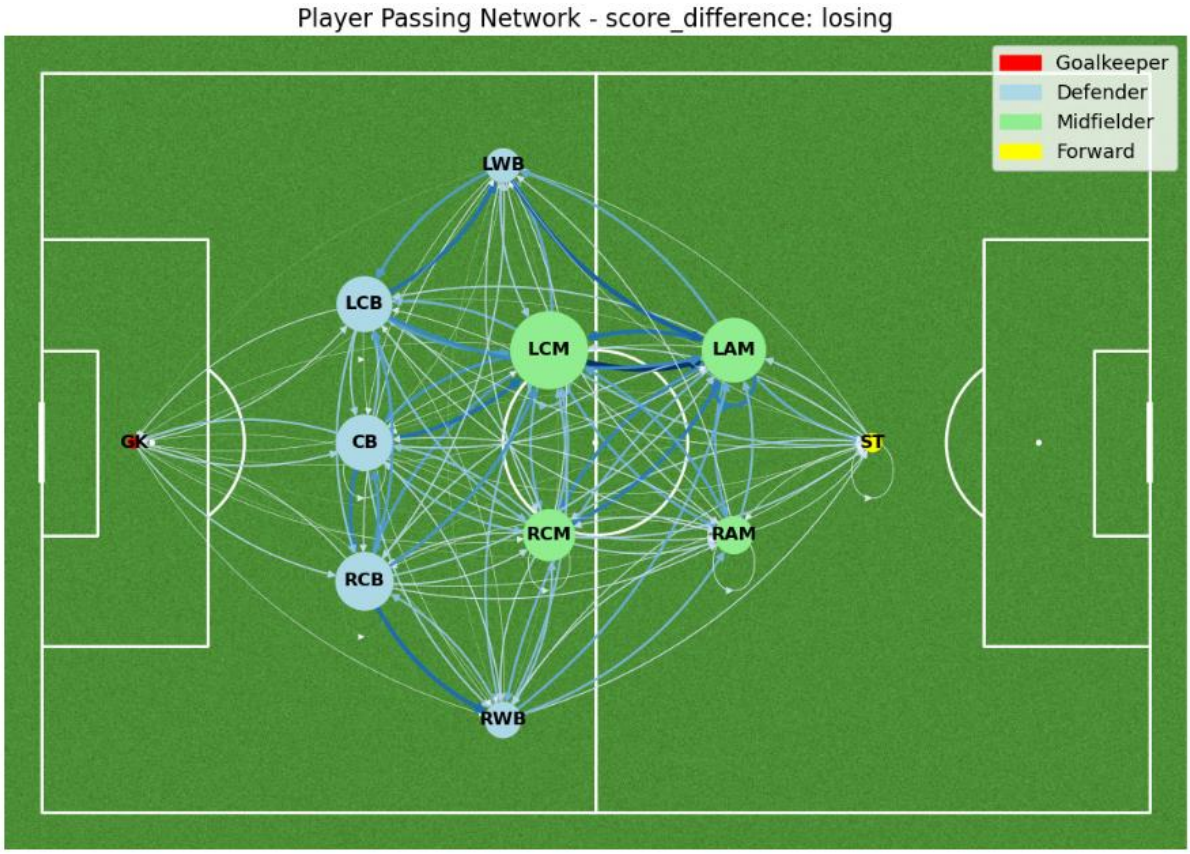


Figure 52. Passing Networks by Player Roles – When Losing During Games