In-Game Strategy Recommendations in Association Football:
A Study Based on Network Theory

Anton Heikinheimo, 38992
Master’s Thesis in Information Systems
Supervisor: Xiaolu Wang
Faculty of Social Sciences, Business and Economics
Åbo Akademi
2019
### Title: In-Game Strategy Recommendations in Association Football: A Study Based on Network Theory

#### Abstract:
Applying advanced data analytics to sports has become one of the cornerstones of a successful sports organization. There are direct business incentives for organizations to leverage these tools for better performance. Improved results increase the organization’s revenues by increasing the amount of competitions they partake in, increase the amount of merchandise sold and the number of spectators at their games. This study analyzes the existing body of literature surrounding network analysis applied to association football and extends it by suggesting new tools and techniques. Past research has failed to convert their findings into actionable insight. Therefore, one of the primary goals of this study is to suggest practical applications of network analysis in football. The dataset used in this study is the complete spatio-temporal data of the FIFA World Cup 2018. The World Cup data was chosen due to the fact that these teams do not have time to develop complex strategies. The problem with analyzing developed clubs and leagues is that they have built a unique strategy around their players and their strengths, and measuring success of said strategy will more likely only measure its execution. The data is transformed to a player passing network, where nodes represent players and connections represent passes between two players. Different network metrics are discussed and their application to football are explained. Using network analysis metrics coupled with feature engineering, this study identifies relevant features that significantly contribute to a team’s overall strategy. K-means clustering is applied to the features yielding three distinct strategies. The identification of strategies enabled an analysis of the strategic shifts in-game. When analyzing strategic shifts, it is possible to identify favorable actions based on a given scenario. A football manager can benefit from this information by practicing a certain playstyle before an upcoming game or by identifying weaknesses of the opponent’s strategy (pre-game analytics). The manager can also use this information to support strategic decision-making during a game (in-game analytics). This study is the first to suggests tools and techniques that can be utilized in-game. Additionally, this study suggests a new success metric: expected goals, which can be more effective than the outcome of the game for measuring team performance, especially when analyzing shorter timeframes of football games. Lastly, the foundation for feature selection in network analysis was set for future machine learning endeavors; this study was the first to make the distinction between player positions, separately considering subsystems of the whole team network. The study would benefit from additional data; however, the World Cup is only played every four years and therefore, the ability to draw conclusions based on a limited set of data is instrumental in the everchanging environment of sports analytics.

#### Keywords:
Network analysis, clustering, strategy, football

---

**Date:** 24.2.2019  
**Pages:** 84
# Table of Contents

1. **INTRODUCTION** .................................................................................................................. 1  
2. **OUTLINE** ............................................................................................................................. 2  
   2.1 **BACKGROUND AND MOTIVATION** ............................................................................... 3  
   2.2 **INVASION SPORTS** .......................................................................................................... 5  
   2.2.1 **Basketball** .................................................................................................................. 6  
   2.2.2 **Ice-Hockey** ............................................................................................................... 8  
   2.2.3 **Football** ................................................................................................................... 9  
   2.3 **MATCH ANALYSIS** ......................................................................................................... 10  
   2.3.1 **Pre-game** .................................................................................................................. 11  
   2.3.2 **In-game** .................................................................................................................... 12  
   2.3.3 **Post-game** ................................................................................................................ 13  
   2.4 **DATA TYPES AND APPLICATIONS** .............................................................................. 13  
   2.5 **NETWORK ANALYSIS** .................................................................................................. 15  
3. **RESEARCH QUESTION** ....................................................................................................... 16  
4. **METHODS** ........................................................................................................................... 17  
   4.1 **NETWORKS** .................................................................................................................... 17  
   4.2 **CLUSTER ANALYSIS** ...................................................................................................... 21  
   4.3 **EXPECTED GOALS** ........................................................................................................ 24  
5. **DATASET AND PRE-PROCESSING** ................................................................................... 25  
6. **LITERATURE REVIEW** ....................................................................................................... 26  
   6.1 **NETWORKS** ..................................................................................................................... 27  
   6.2 **CENTRALITY** ................................................................................................................... 29  
   6.2.1 **Degree Centrality** ....................................................................................................... 30  
   6.2.2 **Eigenvector centrality** .................................................................................................. 31  
   6.2.3 **Betweenness centrality** .............................................................................................. 33  
   6.2.4 **Pagerank** .................................................................................................................... 34  
   6.2.5 **Closeness centrality** .................................................................................................... 36  
   6.2.6 **Clustering Coefficient** .............................................................................................. 37  
   6.2.7 **Density** ...................................................................................................................... 39  
   6.3 **STYLES OF PLAY** .......................................................................................................... 40  
7. **ANALYSIS** ............................................................................................................................ 41  
   7.1 **DESCRIPTIVE STATISTICS** ............................................................................................ 41  
   7.2 **PASSING STATISTICS** ...................................................................................................... 42  
   7.3 **NETWORK ANALYSIS** .................................................................................................... 46  
   7.3.1 **Network eigenvector centrality** .................................................................................... 47  
   7.3.2 **Network betweenness centrality** ................................................................................ 47  
   7.3.3 **Network cluster coefficient** ....................................................................................... 50  
   7.3.4 **Network Pagerank** ..................................................................................................... 51  
   7.3.5 **K-means Clustering** .................................................................................................... 52  
   7.3.6 **Inspecting the clusters** .............................................................................................. 53  
   7.4 **CLOSER ANALYSIS OF EACH CLUSTER** ..................................................................... 57  
   7.5 **APPLICATIONS AND ANALYSIS OF STRATEGIES** .................................................... 60  
   7.6 **CONCLUSION** ............................................................................................................... 64  
   7.7 **FUTURE WORK** .............................................................................................................. 66  
8. **STRATEGIREKOMMENDATIONER FÖR PÅGÅENDE FOTBOLLSMATCHER. EN STUDIE MED UTGÅNGSPUNKT I NÄTVERKSTEORI.** .............................................................................. 66  
9. **REFERENCE LIST** ................................................................................................................. 69  
10. **APPENDICES** ...................................................................................................................... 77
1 Introduction

Analytical tools and techniques, which are used to transform large amounts of data to actionable insight, have seen explosive growth in the past decade. Data is gathered and harvested to increase an organization’s competitive capabilities. In the classic book *Moneyball*, Billy Beane is a pioneer in utilizing quantitative data to drive the success of the Major League Baseball team Oakland Athletics in 2002 (Lewis 2004). Since the publication of the book, sports analytics have seen large advancements. However, as early as in the 1950s, Charles Reep pioneered in the field of quantitative sports analytics. At a football game between Swindon Town and Bristol Rovers, he started collecting quantitative data by hand. He later became the first to publish an academic paper on quantitative analytics applied to association football (US terminology: soccer), where he, based on his observations, concluded that “the key to scoring goals and winning games was to transfer the ball as quickly as possible from back to front”. (Reep and Benjamin 1968; Reep et al. 1971; Pollard 2002)

The introduction of analytics was quickly accepted in some sports, while in others, it was met with opposition. For example, in baseball, sports analytics were quickly acknowledged, while in football, it met much resistance in the early years (McGarry and Franks 2003, 264-70). Football has always been considered a form of art, which is difficult to quantify (Friend and Mumcu 2016, 15-31). However, the situation has seen change, but even today, analytics are only utilized to a limited degree in certain sports. Fan engagement has played a large role in the growth of sports analytics as a result of the increasing wealth of data publicly available for anyone to leverage. Both the clubs and the spectators have an incentive to learn and apply analytics: clubs need to leverage data to gain a competitive advantage, whereas spectators mostly just want to understand the statistics presented to them by pundits. However, more invested spectators will want to perform analyses of their own, in hopes of earning money by betting, to beat their virtual league for prestige, or to simply gain a deeper understanding of the game at hand. (Gudmundsson, and Horton 2017; Cintia et al. 2015)
The purpose of this thesis is to investigate the existing literature on sports analytics as well as to suggest a new method for sports analytics. The research identifies a gap in the body of literature and suggests a unique approach to solving the problem. The suggested method is a recommendation framework to assist managers in decision-making throughout a football game.

2 Outline

The background section aims to introduce data analytics in sports as a topic. Various applications of data analytics in different sports will be presented and the motive behind data analytics in this field will be described: Why would someone want to analyze sports, and what makes it meaningful and valuable?

In the method section, the methods and tools applied in this study are discussed. Moreover, the choice of each method is motivated, and it is discussed why it is suitable given the problem and dataset. In the methods section, the network concept is presented and its past applications are outlined. Information about clustering and its past applications is also included, as is a section on why the K-means clustering algorithm was selected for this analysis. Lastly, the expected goals are described in terms of representativeness and meaningfulness.

In the literature review, the current body of literature is analyzed. First, the literature concerning networks is presented, and after that, each network metric is considered separately. The representation of a metric is explained, as are its past applications and findings in football analytics. In the theory section, also the literature concerning styles of play in football is covered. The purpose of the theory section is to convey an understanding of which network measures have been deemed meaningful so that this information can be utilized in the analysis. With the extensive literature review, the significance of this study will be proved, and so will the gap that it fulfills.

In the findings section, the actions needed to perform the analysis are presented step by step. The data and the findings will be described and visualizations enabling an
understanding of the problem and the solution will be provided. In the discussion section, the results provided in the analysis will be freely discussed. Moreover, limitations and possible extensions for further research will be outlined. The entire study will be brought to a close in a concluding section, where the thesis and its results are summarized.

2.1 Background and motivation

From a business standpoint, there are many incentives for sports clubs to leverage data. A good performance of a sports club directly improves the organization’s revenues, by increasing the amount and level of competitions they partake in, e.g. playoffs, Europa League and Champions League. These additional events provide the club with monetary rewards as well as with additional opportunities for tickets sales. Well-performing clubs also tend to sell more merchandise and have higher average attendance, which also allows them to charge a premium on their ads. In addition to increasing the on-pitch performance, data can be utilized to identify undervalued players, which can in its turn allow favorable player acquisitions or trades. For example, the ice-hockey team Vegas Golden Knights reached the Stanley cup in their inaugural year. This success is largely accredited to their data analytics. The GM of the Golden Knights was quoted stating:

We’re finally there (with analytics). I’d rather rely on the science now than just opinion. We understand that our instincts matter, our experience matters, but that data has become really reliable. (George McPhee, press conference at City National Arena, 7.6.2018, Las Vegas)

From an outside point-of-view, understanding the game of football has always been a lucrative proposition as the betting markets can be seen as a financial one. Understanding the game, and the factors that affect its outcome, can be used to beat the bookkeeper, and make a financial profit by betting on match outcomes. (Kumar, 2013)

A sports strategy can be defined as “the overall plan that is devised and adopted to achieve an aim or specific objective and is normally accomplished via the application of specific tactics” (Carling, Williams, and Reilly 2005, 11-17). Strategies and tactics
applied by the manager are important factors affecting the final outcome of a match (Yiannakos and Armatas 2006, 8-10). Therefore, sports teams adopt a combination of attacking and defensive styles of play to maximize their chances of winning. Fernandez-Navarro, Javier, et al. (2016, 1-2) define styles of play in football as: “the general behavior of the whole team to achieve the attacking and defensive objectives in the game”.

The emerging big data paradigm has successfully tackled the issue of data acquisition by automating or semi-automating data acquisition (Cintia et al. 2015, 1-3). At the 2018 FIFA World Cup, two optical tracking cameras recording the movement of each player were used at every game. Additionally, there are companies specializing in data acquisition and structuring, such as Opta (https://www.optasports.com/) and StatsBomb (https://statsbomb.com/), which base their business model on gathering and structuring data as a service. Data storing and computing is managed by advanced cloud-computing platforms available through the click of a button. Nevertheless, one of the biggest big data challenges remains the value extracted from it. In MIT Sloan’s management review of 2018, where 1900 business executives were surveyed, one of the main findings was that 59% of managers reported that data is a source of competitive advantage. (Ransbotham and Kiron 2018; Walker 2014; Manyika et al. 2011)

Sports managers and staff are presented with an abundance of data, which is to be augmented with their own qualitative analysis and intuition. A study performed by Franks and Miller (1988, 23-30) proved that on average, a person is only able to recall 42% of the critical match events during a full-length game. This has driven the advancements of quantitative data analyses in sports. Sport clubs apply different strategies and use a variety of tools. The key to a successful analysis is to create easily digestible data, as the overabundance of data has become an obstacle in itself (Rein and Memmert 2016, 2-5). There is no standardized set of tools, and therefore, it creates new ways to outperform rivals (Porter and Millar 1985, 14-20).
2.2 Invasion sports

“Invasion sports” is a group of sports where the objective of two teams is to compete for possession of an object such as a ball or puck within a fixed size pitch and timeframe. The primary goal is to try scoring by placing the object in the opposition’s goal, whilst simultaneously stopping the opposition from scoring. After the fixed timeframe has passed, the team with more scored goals wins, or in the case of equal amount of goals for both teams, it is decided as a draw (neither team wins, but both are generally rewarded 1/3 of the points of a win). Association football is the most well-known and popular form of invasion sport. The most popular sports competition in terms of TV viewership in all forms of sports is the FIFA World Cup, which is played every four years, and where thirty-two qualified national teams compete for prestige as well as for money (FIFA Homepage 2019) Invasion sports are an interesting topic for data analytics, as they are generally more complex and fast-paced, which makes them more difficult to interpret. In comparison, in sports such as tennis or baseball, the game is turn-based, and consists of only a few actors at a given time. (Gudmundsson and Horton 2017; Duch et al. 2010)

Baseball has historically been the most active community for sports analytics, due to the availability and accessibility of datasets. The most common application of analytics in sports is predicting the outcome of a game, which is of high interest to anyone. Betting markets can be compared to any other financial market, where assets are bought and sold, and the ability to outperform the market yields a return. Another motive to predict match outcomes is to unveil the variables affecting the outcome. If a club can understand the variables affecting the outcome of a game, it can effectively tweak them in its favor. If, for example, sprinting distance was a very strong predictor of match outcome, then a club could focus their practice sessions on speed and stamina. (Bunker and Thabtah 2017; Kumar 2013)

Machine learning is the most common set of tools used to predict outcomes. The task of predicting an outcome on previously unseen data is called classification (Abdelhamid et al. 2012, 1-2). In classification, the classifier is trained on historical data (training data) and once trained, used to predict a target variable (outcome). This type of a machine learning task is called supervised learning, since the target variable
is given (Mohammad et al. 2015, 4-6). The other type of machine learning tasks is unsupervised learning, where the target variables do not exist, but instead the algorithms group and interpret the data based only on the input data. (Witten and Frank 2011, Bunker and Thabtah 2017)

Figure 1: Supervised learning vs. unsupervised learning. (Bunker and Thabtah 2017)

2.2.1 Basketball

Sampaio et al. (2010) performed a factor analysis on box-scores (free-throws, 2-point field-goals, 3-point field-goals, passes, and errors) of well-performing teams and unimportant vs. important players. Well-performing teams were identified using two-step clustering, where the strongest teams were classified by a winning percentage of 69±8, intermediate teams by 43±5 and weak teams by 32±5. Important vs. unimportant players were identified through playing time, where important players had played 28±4 minutes per game, while less important players had 16±4 minutes. Furthermore, they also analyzed the seasonal variations from month to month, but no significant effect was observed. They concluded that the stronger teams were dominant in 2-point field-goals and passes, whereas weak teams were the weakest at defensive rebounding. Important players were on average characterized by an increase in all variables except errors made.

Franks et al. (2015) strive to quantify the defensive efforts of individual players. They argue that with current box-statistics, it is easy to determine the best offensive players, but that they cannot be used to quantify defensive efforts. They use spatio-temporal
game data in their analysis to define five metrics that can be used to identify defensive efforts: (Franks et al. 2015, 2)

**Volume Score:** The total magnitude of attempts which an individual defender faces.  
**Disruption Score:** The degree to which an individual defender is able to reduce the effectiveness of his assignment’s shots.  
**Defensive Shot Charts:** Like shot charts, but for defensive play. Visual depictions of an individual’s defensive prowess; we map both volume score and disruption score across the scoring area.  
**Shots Against:** A weighted average of the shots attempted against the defender per 100 possessions.  
**Counterpoints:** A weighted average of points scored against a particular defender per 100 possessions.

Additionally, one of the visualization tools used is a matchup matrix where we can see the points scored based on which player was marking them:

Hamdad et al. (2018) use a variety of data analysis techniques such as clustering, classification and regression to assist decision-makers in making data-driven decisions. In their analysis, they identify team weaknesses, players that can be used to overcome those weaknesses and predict the outcome of games, and the salaries of each player in the NBA. Cervone et al. (2014) use similar tools to determine the expected
possession value (EPV). EPV is a metric that gives the expected points of each possession, at any given time. EPV is dynamic in the sense that it increases or decreases as a result of a change in play, e.g. reduce in time, a pass or a dribble. In their analysis, Cervone et al. use player-tracking data of the NBA 2011-2012 season. They reason that the EPV is a powerful tool for coaches as it can be utilized to optimize the offense and the defense. Offensively, the team can use more high probability plays, and always try to maximize the EPV, whilst defensively, the team can identify and counteract the opponents EPV patterns. Here is an example of the EPV in a possession:

![Graph over the expected possession value. The ball carrier has probabilities assigned for each possible action as well as weighted average of the points those actions will yield. In the EPV the movement of the ball carrier is also considered but omitted from the graphical representation. (Cervone et al. 2014, 2)](image)

2.2.2 Ice-Hockey

In the context of ice-hockey, Bulmer (2018) applies a similar technique to the EPV of Cervone et al. (2018). Bulmer defines a Weighted Offensive Productivity Rating (wOPR). The wOPR is determined by a mixture of a player’s ability to pass, transition in zones, and shoot. The respective weights of these three variables were estimated by considering the contributions of shot assist, shot attempt, zone entry, and zone exit data. Bulmer hypothesizes that the wOPR is a better predictor of game success than
the player’s total points in the NCAA Division 1 hockey league. He concludes that the wOPR is statistically significant in predicting the outcome of a 5 on 5 match. It also outperformed as an indicator of play quality in cases where data availability was limited.

Kniffin et al. (2017) study the effects of strength and conditioning metrics on the in-game performance measurements of players in the NCAA Division 1 hockey league. They argue that the expected value of conditioning and strength programs is generally high but that there is little quantitative evidence supporting this, and therefore closer analytical inspection is needed. In their results, they find significant correlation between bench press performance and points scored. They present their most important finding as a positive relationship between strength and conditioning measures and playing time.

In hockey, the success of goalkeepers is largely determined by their ability to maintain a high save percentage (the percentage of shots saved vs. conceded). Naples et al. (2018) strive to create a more accurate metric of goalkeeper success. In their metric, they include the context of the shot, i.e. the shot type and distance. They suggest e.g. the following two approaches: save percentage, where all shots are considered, even if they do not hit the goal, and performance against the expected goals. Both of these methods yielded a more significant relationship with match outcome than the normal save percentage.

2.2.3 Football

In football, there have been multiple endeavors to predict the probabilities of game outcomes (Kumar 2013; Ulmer and Fernandez 2013; Timmaraju et al. 2013; Joseph et al. 2006). The successful ones are valuable intellectual property and therefore not distributed, so we cannot know exactly how e.g. betting companies derive their probabilities. Machine learning models can easily be tested in real world scenarios because fixed odds are given well in advance of upcoming games. Recently, also live betting has become more popular. Here, the odds are fluctuating and changing throughout the game, based on the events that are happening. This has also increased
the interest in real time predictions of in-game events and outcome probabilities. Google is one of the providers of real time probability scores.

In football, there are also many other applications of data analytics. Kumar (2013) seeks to analyze the variables that affect the player ratings given by pundits. He uses a large set of in-game statistics to analyze the variables affecting the ratings. In his study, he proves that up to 90% of the ratings are based on in-game events, such as shots, passes, tackles and pass accuracy. Hughes et al. (2012) took it upon themselves to identify key performance indicators for each player position. The performance indicators that they used can be divided into three categories: biomechanical indicators such as “ball release velocity”, technical performance indicators such as “tackles won and lost” and, lastly: tactical performance indicators such as “length of passes” (Kumar 2013, 3-20). From all of the different player positions that were analyzed, the key performance indicators were the most different for goalkeepers and outfield players.

Ruiz et al. (2017) use a variety of machine learning techniques to uncover the underlying factors that lead to Leicester City FC’s unexpected win of the 2015/2016 English Premier League Season. Their first finding was the defensive prowess of Leicester’s midfielders and goalkeeper. The midfielders successfully shut down many attempted shots, and those that did make it past the midfield, were saved by their goalkeeper. The second finding was Leicester’s disruptive strategy, as they effectively deterred the opponent’s attack through players such as N’golo Kante, Christian Fuchs and Danny Simpson. These were all listed among the top twenty players in terms of intercepting passes and Leicester was the team with the most players on this list (3 out of 20). Ruiz et al. conclude their analysis with evidence showing how Leicester has not been able to apply the same strategies in other seasons, and therefore has declined in performance.

2.3 Match analysis

The analytical lifecycle for a match has three stages: pre-game (preparations coming into the game), in-game (tactical shifts, substitutions etc.) and post-game. Out of these
three, the in-game stage is the most time-sensitive one, as the manager does not have much time to make decisions and hence needs to react and adapt to everything that is happening on the pitch. This also introduces the problem of effective data visualization, since the data needs to be formatted in an easily digestible way, whether that is graphical or qualitative. During half time, the manager has fifteen minutes to make tactical decisions for the second half and convey these to team. This is the period of the game when the manager has the largest impact on the trajectory and can turn a losing team around. (Carling et al. 2007)

2.3.1 Pre-game

In pre-game analytics, the purpose is to identify strengths, weaknesses, opportunities and threats of the upcoming fixture. This includes for example monitoring the fitness of your team and proactively avoiding injuries, but also analyzing the past fixtures of the upcoming opponent. The goal is to identify key success factors of the opposing team. Analysts will look at past games where the opponent has won to see what lead to that win, and contrary, look at lost games to understand the factors contributing to that loss. The manager will then develop a tactic with the purpose of winning. The Oxford dictionary definition of a tactic is: “An action or strategy carefully planned to achieve a specific end.” In a football context, end would refer to winning. Controversially, Gréhaigne and Godbout (1995, 3-10) make a distinction between “tactic” and “strategy”. They state that a strategy is the overall plan devised before the game, whilst tactic is an ongoing process as a result of interactions throughout a game. (Carling et al. 2007; Spinks et al. 2002; Reilly et al. 2003)

Tactics can further be divided into individual tactics, group tactics and team tactics. Individual tactics consists of player-specific instructions that are to be conducted during the game. Individual tactics can be e.g. a role in set pieces, opponents to mark and weaknesses to exploit. Group tactics is when a set of players, e.g. defenders, midfielders or attackers, play together in a specific way to achieve a common goal. The team tactics covers the overall plan for the whole team, and this can for instance include formation, movement and playing style. (Bisanz and Gerisch 1980; Carling et al. 2005).
For pre-game strategy, the manager has many variables that he can affect. For each match, the manager selects 11 starting players and 5 substitutes, from up to 30 players. Often, the managers simply select their best players. However, some teams have such a wide range of good players that they will specifically cater their lineup to counter the strategy of their opponent. The formation of the team is another factor that the manager needs to decide pre-game. The formation will be based on the general strategy of the game. A team hoping to play a draw will have more defenders and midfielders, while a team desperate for a win might invest in playing multiple forwards. Managers that are better informed of the strategy used by their opponent are more likely to counteract it and exploit their weaknesses. (Carling et al. 2007; Spinks et al. 2002)

The analyst will consider the opposing team’s passing networks and perform different factor analyses to identify key success factors of their gameplay. They will specifically be interested in games, where the said team has lost, and understand what was different in this game, compared to games where they have won. (Carling et al. 2007; Reilly et al. 2003)

2.3.2 In-game

In-game, the manager does not get many opportunities to affect the flow of the game. The manager will attempt to convey certain messages from the sidelines to the players on the pitch. Football stadiums can get loud during important games which hinders the manager’s ability to communicate with the players. The key event during the in-game analysis is the half-time, which is a 15-minute break after half of the game is played. During this time, the manager has the opportunity to communicate with the entire team without interruption. (Carling et al. 2007; Spinks et al. 2002; Reilly et al. 2003)

During half-time, the manager will communicate formational changes, strategic shifts and conditional statements. Such as, if the team is still losing at the 80th minute, then change formation, or if the game is still a draw at the 90th minute, then do something else. After the half-time, the second most influential changes the manager can perform are substitutions. These are usually made to substitute players suffering from fatigue or injury or to substitute players that will contribute to a new strategy. When a team is losing, the manager might substitute a defensive midfielder for an additional striker,
or when they are winning, a striker may be substituted for a defender. (Carling et al. 2007; Spinks et al. 2002; Reilly et al. 2003)

Player data is gathered throughout the duration of the game. Players wear different sensors that capture their heart rate, and a GPS. These sensors allow analysts to look at top speeds, distance covered, average speed, decrease or increase in performance etc. From these sensors, the analysts can identify players suffering from fatigue or predict players that will decline in performance. Additionally, cameras or semi-automated software is used to capture spatio-temporal event data. From this data, the analyst can identify strengths and weaknesses from both teams and try to counteract them. For example, this allows the analyst to look at dominant regions of the pitch. These are areas of the pitch, where the team has been most successful with possessing the ball. Spatio-temporal event data allow for complex network analysis that unveils characteristics of both teams and is a very powerful tool for the manager. (Carling et al. 2007; Reilly et al. 2003)

2.3.3 Post-game

The post-game analysis, the core of match analysis, is based on lessons learnt from the previous game. Teams look at past performances and try to understand the factors that contributed to wins and losses, and at this stage, detailed data is gathered and analyzed. The match analysis cycle both ends and begins with post-game analysis. The strategies that were developed during the pre-game stage and implemented in-game, are all based on the insight from the post-game analysis. (Carling et al. 2007; Spinks et al. 2002; Reilly et al. 2003)

2.4 Data types and applications

For match analysis, football clubs utilize two types of data: firstly, basic box-statistics such as shots, passes, tackles and pass completion rate and secondly, spatio-temporal data, which contains the sequence of samples, timestamp and location coordinates. In the football domain, spatio-temporal can be further divided into object trajectories, that capture the movement and location of players and/or the ball continuously and event logs that capture each event during the game, e.g. pass, shot, cross, tackle, and
its respective location, direction, outcome and timestamp. Spatio-temporal data is used to build heatmaps, intensity matrices, maps, passing networks etc., which can be useful in identifying weaknesses and untapped opportunities. (Gudmundsson and Horton 2017)

Figure 4: Intensity map over two players, playing from left to right. The marking shows their movement throughout the game. The intensity maps were developed from spatio-temporal data. (Gudmundsson and Horton 2017)

The use of spatio-temporal data analysis in sports, and especially in football, has been increasing. This is somewhat due to an increase of data availability, but also because of the promising results. Gudmundsson and Horton (2017) illustrate the rapid increase in count of academical papers on spatio-temporal analyses in different sports.

Figure 5: bar graph over the academical papers on spatio-temporal analyses in different sports. (Gudmundsson and Horton 2017)
Spatio-temporal data analytics allows us to analyze how the team builds up play and shifts formation when attacking or defending, winning or losing etc. It enables an analysis of the team’s interactions and of how the members work together (Gudmundsson and Horton 2017, 6-10). To understand teamwork is as crucial in sports as it is in business, art, science or any other form of human interaction. Understanding how a team works requires the analyst not only to analyze the individual members of the team, but also the interactions between the team members. In the context of teams and organizations, there is generally a consensus on the most important members of the team, even though it is not even clear what the variables affecting the final output of the team are. (Duch et al. 2010)

2.5 Network analysis

Network analysis is an emerging field in sports analytics. It is used to look beyond the individual player, and instead focus on the team as a single entity, a system built to accomplish a specific task. In football, the most common type of network is a passing network where nodes represent a player and the edges a pass between the two players. The passing network can be extended to represent a wealth of data, as we can see in this network graph by Arriaza-Ardiles et al. (2018, 15-19):
Figure 6: The radius of the nodes is proportional to the total number of passes and receptions associated to a player. The goalkeeper is in black, the defenders in blue, the mid-fielders in yellow, and the forwards in red. Each interaction is proportional to the number of passes sent by a player, where the link follows the colour of the kicker, reinforced by a little orange circle marking de origin, and a little white circle marking the destination. (Arriaza-Ardiles et al. 2018)

Despite the advancements in tools and techniques in network analysis, Cintia et al. (2015, 1-3) argue that these techniques have only been utilized to a limited degree so far, which they consider surprising given the overwhelming evidence of the complexity and dynamics of a football game. This, in combination with the recent availability of datasets, makes football network analysis an interesting and potent field for public research.

3 Research question

The aim of this thesis is to develop a system that can suggest a short-term shift in strategy, during half-time, in order to change the trajectory of the game, i.e. from losing to winning. The strategies will be based on network metrics, player positions and other in-game statistics. Unsupervised learning will be used to identify groupings of playing
styles and distribute them to strategies. The recommended strategic shift will be based on both teams’ play-style during the first half. In the results chapter, it will be examined how different strategies match up against each other, and what kind of a transition in strategy is best in a given scenario. The final outcome of the thesis will be an easily digestible and automated recommendation for the managers that they can easily convey to the team during a time-sensitive period such as half-time.

Network analysis, despite being a powerful tool for analyzing team behavior, is argued to have been utilized to a limited degree. This study aims to transform theoretical knowledge into practical applications that can be utilized by a sports organization for improved results.

4 Methods

4.1 Networks

A network is a collection of nodes (often also called a vertex, but node is the more common naming convention, so it will be used here) and edges. The node in a network is an entity, e.g. a person or a cell, and the edges are the connections between the nodes. Nodes in a network can have both a direction and a weight. For example, in the case of a passing network over a football game, players can be nodes and passes can be seen as edges between them. The network would be directed as there would be a pass from one node to another; weight could represent the frequency of passes between those nodes. The most common types of networks to analyze are the internet, a network with computers as nodes and data connections as edges, as well as human societies, where nodes can be people, families or communities, and an edge represents a relationship between those.
A network allows us to study the individual nodes, the interactions and patterns of interactions. Networks have applications in all disciplines. In the case of the internet, it allows us to analyze the flow of information, giving us the capability to identify bottlenecks and mitigate risk by identifying the most important computers. Networks can be applied to understand social structures, e.g. networks have been utilized to understand and predict the spread of diseases. Networks have also helped to analyze the spread of information and rumors on Twitter (Grabowicz et al. 2012). They have also been used to explain how different brain regions are coordinated to perform cognitive tasks (Papo et al. 2014). Social network analysis (SNA) is a subfield of network analysis, with the purpose of analyzing social structures through the use of network analysis. One of its first applications can be seen below. (Newman, 2010)
Networks are a simplified representation of something more complex. Nonetheless, they are a powerful tool to represent patterns in interactions and connections between entities, as they can be used to simplify and provide an understanding of how a complex system works. While networks are no recent idea, they have gained much momentum in multiple fields. Throughout the past decade, it has become one of the most prominent fields in both physics and mathematics due to the discovery of technological, social and biological systems that can be analyzed with their help. Network analysis is the combination of tools from four disciplines: graph theory, statistical physics, nonlinear dynamics and big data (Barabási 2016, 25).
Figure 9: A food web represented by a network. The nodes represent animals, fish or a group of a species. The directed edges represent flow of energy (calories). Using such networks, we can determine the most important actor in a given ecosystem (Newman 2010).

Understanding interactions between players in team sports is one of the most important and complex problems in sport sciences. This problem is only enforced when the number of players increases, since relatively, the impact of one player is decreased. In football, there are 11 players per team, whilst in other invasions sports such as basketball or hockey there are much less (5 and 6 respectively). Understanding player interactions in a team helps the analyst to generalize playing styles and to identify its most important components. Understanding the interactions of a team is the primary goal of social network analysis (Wasserman and Faust 1997, 12-18). Newman (2010, 70-73) emphasizes the visibility that networks give to the patterns of interactions in an organization. (Gudmundsson and Horton 2017)

Clemente et al. (2015, 9) conclude after their analysis that network metrics are a powerful tool for better understanding a team’s specific properties and support the
decision-making in order to improve the sports training process based on post-match-analysis. Networks allow us to capture elements of team performance and organization that would go unseen in classical analyses based on individual players (Buldu et al. 2018, 1-2). Network analysis describes the interactive behavior of a team, instead of using a reductionist approach to analyzing individuals in it (Mclean 2018, 3-7). Duch et al. (2010) proved the potency of network analysis as a tool for football analysis by correctly predicting the top-rated players of the 2008 Euro, with the help of network metrics.

Networks allow us to study the game at different levels of depth. At the microscale, we can analyze individual nodes, i.e. a single player and their roles in the network. At the mesoscale, we can analyze small groups of the network. The mesoscale allows us to, for example, analyze the existence or non-existence of passing triangles or play patterns between a set of players. Lastly is the macroscale, on which we analyze the network as a single entity. For each level of depth, we can analyze network measures, which help us understand the interactions and performances of individual players, a group of players, or the whole team. (Buldu et al. 2018)

4.2 Cluster analysis

Organizing complex data to distinct groupings is the first step to learning and understanding the data at hand (Anil 2010, 651). Cluster analysis is a form of pattern recognition, which is useful in discovering inter-relationships between heterogeneous data and in organizing it in homogenous clusters. Cluster analysis is a part of unsupervised learning because the data is unstructured, and without labels. Unsupervised learning is used to apply labels to the data. Inter-connectivity is a measure of how distinct the clusters are. The term learning is used in scenarios where, given a set of training data, we want to predict the behavior of unseen data (Anil 2010, 651). The Merriam-Webster international scientific vocabulary defines cluster analysis as:

a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics.
In summary, clustering is the process of grouping objects together based on their intrinsic characteristics and similarity. (Reddy and Ussenaiah 2012)

Data analysis can be divided into two types. The first is exploratory or descriptive data analysis, where the analyst is trying to understand the general structure and characteristics of the data. The second type is confirmatory or inferential, where the analysts have a predefined hypothesis which they try to confirm with the help of data analysis (Tukey 1977, 10-12). Tabachnick and Fidell (2007, 7-23) propose analysis of variance, linear regression, discriminant analysis, canonical correlation analysis, multi-dimensional scaling, factor analysis, principal component analysis and cluster as examples of tools and techniques available for analyzing data. The majority of available data is unstructured, which makes clustering a very potent tool in learning tasks.

Anil (2010, 652) offers an operational definition of clustering as:

```
Given a representation of n objects, find K groups based on a measure of similarity such that the similarities between objects in the same group are high while the similarities between objects in different groups are low
```

Clusters can differ in terms of form, shape, size and density. Hence, the measure of similarity will differ depending on the data at hand. Anil (2010, 652) argues that a cluster is a subjective entity of which significance and interpretation requires domain knowledge. Humans can visually discover clusters in a two- or three-dimensional space but require computer software when working with more complex data. Depending on the applied field, clustering has also been referred to as Q-analysis, typology, clumping, and taxonomy (Jain and Dubes 1988, 11-12). (Anil 2010)

Cluster analysis is used as a general tool in any analysis that consists of multivariate data. The list of potential applications is very extensive, but a few examples will be listed here. Clustering is an important tool used in computer vision tasks, such as image segmentation, where similar images are grouped together (Shi and Malik 2000, 903-904). An example of this is in Google image search results, where the system automatically detects sub groups of the search results:
Cluster analysis has also been applied to document search results, where document hierarchies are generated (Sahami 1998). Marketing firms utilize clustering as a form of customer segmentation in their marketing campaigns. In customer segmentation, customers are merged into groups, whose interests are the most similar, so e.g. e-mail campaigns can be tailored to the receiver (Arabie and Hubert 1994).

Cluster analysis can be further divided into two groups: hierarchical and partitional. In hierarchical clustering, the algorithm recursively finds nested clusters by either starting with every observation as a cluster of its own, and then merging the most similar clusters (called ‘agglomerative mode’) or by starting with all observations as the same cluster, and then recursively dividing the most dissimilar observations to separate clusters (called ‘divisive’ or ‘top-down approach’). Partitional clustering is not recursive or hierarchical, but instead finds all clusters simultaneously as a partition of the data.

The most popular partitioning clustering algorithm? is K-means clustering, which was introduced in 1955 (Anil 2010, 653-654). For hierarchical clustering, S-link and complete-link are most common. K-means is a method, which was independently discovered in multiple fields (MacQueen 1967; Ball and Hall 1965; Llyod 1982; Steinhaus 1956). Partitioning algorithms are generally preferred in pattern recognition. Anil (2010, 651) argues that ease of implementation, simplicity, efficiency, and empirical success are the reasons behind the popularity of the K-means clustering. Those, in combination with K-means’ potency in pattern recognition, is the reason why it will be used in this study.
4.3 Expected goals

Duch et al. (2015, 1) observe that one of the reasons football is a difficult sport to study is that the scores are generally low. This would pose problems to this study in particular, as the strategic shifts are based on the first half, where the score will often be as low as 0 – 0. To counteract this issue and to suggest a novel method for further research, expected goals (xG) will be used as an indicator of match outcome and team success. Using expected goals enables the analysis of shorter time periods. The team with higher expected goals will be considered the “winner” of the time period, if the two teams score are close to the same integer, it will be considered a draw. The strength of each win or loss will also be considered in the analysis.

Expected goals is a measurement developed as an extension to studies proving that shots are the ultimate measurement of performance in football (Armatas et al. 2009; Armatas and Yiannaskos 2010; Swarc 2004; Swarc 2007). Expected goals captures each chance of scoring a goal. It is a probability measurement based on past averages. Each chance of scoring is given a score between 0-1 (0-100%) based on historical averages of scoring from said location, angle of shot, body part used (header, weak foot, strong foot, back heel etc.), the type of pass that created the chance, whether or not a dribble was done and much more. Football statisticians have aimed at developing the most accurate expected goal measurement. (StatsBomb 2019; Rathke 2017)

The expected goal metric, which will be used in this study, was developed by StatsBomb and included in the World Cup dataset. StatsBomb’s expected goals captures a wide array of different metrics when valuing the chance of each goal scoring opportunity. In addition to the previously mentioned ones, this measurement captures the spatio-temporal data of all players included in the chance, as well as shot velocity. An example of an expected goal scenario can be seen in the appendices (Appendix A).

Kuper (2014) proves in his book “Soccernomics: Why England Loses, Why Spain, Germany, and Brazil Win, and Why the US, Japan, Australia and Even Iraq Are Destined to Become the Kings of the World’s Most Popular Sport”, that two thirds of a football game are random. This means that often the better team does not win the
game. On such occasions, the winning team will have less possession, less shots, less threatening attacks and less expected goals. When answering the question “who played better?”, we will often face contradicting statements due to this issue. This makes the game more interesting for spectators but problematic for analysts. In the FIFA World Cup 2018 dataset, in 66.6% of the games, the expected goal was an indicator of the outcome. In 33.3% of the games, the expected goal was not an indicator of the outcome. These findings corroborate those of Kuper (2014).

Anderson & Sally (2013, 5-7) argue that while the game has random elements, our methods are not sophisticated enough to fully grasp all facets of the game. Buchdahl (2016) proves through quantitative analysis how betting agencies (“bookies”) have not been able to improve the accuracy of their predictions throughout the past ten years, despite all the advancements in data analysis tools. The “winner” will more often be the team that played better, even in the those 33.3% of the cases where the outcome did not meet the result of expected goals.

5 Dataset and pre-processing

The data used in this study will be the complete spatio-temporal data of each game played in the FIFA World Cup 2018, provided by StatsBomb (StatsBomb 2019). The dataset consists of each event and the respective coordinates of the players in action. The specifics of the dataset can be found in the appendices (Appendix B).

World Cup data was specifically selected for this study as national teams have a limited time to prepare for games and tournaments. This leads to teams not having time to develop a distinct playing style but instead play football in its purest form. The problem with analyzing developed clubs and leagues is that they have built a unique strategy around their players, and their strengths, and measuring success of said strategy will more likely only measure the execution of their strategy. When analyzing World Cup data, we can make easily implementable shifts in strategy that will generally work regardless of the team they are facing. In summary, the strategies used in developed leagues are more long-term based, where teams will play in a certain
way, despite short-term losses, if that means that their playing style will eventually yield results. On the other hand, in the World Cup, everything is short-term: player selection, team formation and game lineups due to the volatile nature of the single elimination tournament.

In the World Cup 2018, we saw an unusual amount of upsets, e.g. Germany (the reigning champion and ranked one by FIFA (FIFA Coca Cola ranking 2018) finished last in their group, despite an easy group (Korean Republic rank 57, Sweden rank 24, Mexico rank 15), or Spain’s (rank 10) loss in the first round of the knockout stages against Russia (rank 70). A high number of upsets can be partially accredited to the random nature of the game and the single elimination structure of the World Cup, but also to managers playing to their teams’ strengths, and successfully implementing short-term strategies. Some of the best examples was Mexico vs. Germany in the group stages, where Mexico understood their position as the underdog. Mexico gave the possessional advantage to Germany, and only played fast counter attacks, and a very deep defensive line. Another example is Brazil vs. Belgium in the second round of knockout stages, where Belgium overcame the fan favorite and won despite having clearly less goal scoring opportunities and less possession.

6 Literature review

When analyzing matches, the most common approach is to codify events, actions and behaviors on the pitch (Franks & McGarry 1996, 4-7). Such an analysis simplifies the game to a number of dimensions on which inter-relationships can be identified. Lago-Peñas et al. (2011) illustrate in their study the football match factors that impact the outcome of the game positively or negatively. In their analysis, they investigate the Champions’ League group stages of seasons 2007/2008, 2008/2009, 2009/2010 and consider the following variables total shots, shots on goal, effectiveness, passes, successful passes, crosses, offsides committed and received, corners, ball possession, crosses against, fouls committed and received, corners against, yellow and red cards, venue, and quality of opposition. They conduct a one-way analysis as well as a discriminant analysis to differentiate between variables that impact the outcome of the
game. In their results, they conclude that winning teams had higher averages of total shots, shots on goal, effectiveness, passes, successful passes and ball possession, while losing teams were characterized by higher values in yellow cards and red cards. The problem with these findings is that they do not provide any value to the team in terms of strategy.

6.1 Networks

A more recent and prominent type of match analysis is network analysis. Here, the focus is not on individual players but on the whole team as a single organization. Networks can give insight on the strategies of a team, or on the importance of individual players in the network. This field has been fueled by the increasing amount of passing data made available. The idea of applying network theory to football games was first introduced in 1977 by Gould and Gatrell, who manually recorded every pass in the 1977 FA cup final between Manchester United and Liverpool. The main findings were that Liverpool had two disconnected subsystems and that Kevin Keegan was “the linchpin of Liverpool”, suggesting a high betweenness score. The study did not receive the recognition it deserved at the time but has since been seen as well ahead of its time. (Gudmundsson and Horton 2017)

The two most common types of networks are the passing networks and transition networks. Passing networks were first introduced by Passos et al. (2011, 170-176), where each player is represented by a node, and each pass between two players is represented by a directed edge, with a weight. The difference between a passing network and a transition network is that in a transition network, additional nodes are present, which represent the outcome (Gudmundsson and Horton, 2017). For example, in Skinner & Stephen (2015) study on basketball networks, a transition network was used to assess the effectiveness of the attack, with the transition network outcomes being either 0, 2 or 3 points. Additionally, Buldú et al. (2018) further identify three different subgroups of passing networks in football:

(i) player passing networks, where nodes are the players of a team (ii) pitch passing networks, where nodes are specific regions of the field connected through passes made by players occupying them or (iii) pitch-player passing networks, where nodes are a combination of a player and its position at the moment of the pass.
The work within football passing networks had remained quite silent and insignificant until the publication of Duch et al. (2010), which marked the beginning of a new era. Duch et al. proved the significance of networks and its application in quantifying the performance of individual players in a team sport. Their hypothesis was that while individual stats (passes, tackles, shots etc.) in certain sports, such as golf, baseball or track-events, are perfectly suitable for measuring the performance of individual players, this was not applicable to team sports, especially not to football. In football, it has always been difficult to measure the contributions of each player, due to the complexity and pace of the game. Additionally, in football, there is generally a low score, which means that measuring passes and scored goals rarely provides a reliable measurement. Duch et al. describe the contributions of a player as “hidden”, since the real contribution of a goal is not necessarily the final touch or the assist, which leads to a goal, but the entire build-up, whether it started from a dispossession or a goal kick. In their study, they develop a transition network, with shots on and off goal as separate nodes. Then, they quantify each player’s contribution in the buildup play towards those nodes. In their results, they find that out of the top twenty players that their model yielded, sixteen were also present in the official top-twenty, which was selected by football analysts and spectators. Thus, they concluded that networks provide a good measurement for player performance in football.
Cintia et al. (2015, 1-10) prove the predictive power of the network measures by predicting the outcomes of games in the Italian Serie A 2013/2014 season. In their conference paper, the performance of a team is described using three measurements: the mean degree of the network’s nodes, the variance of the degree of a network’s nodes, and a combination of the two (an index). A predictive accuracy of 53% is achieved, which defeats the null hypothesis of 48% as well as the other comparator used, i.e. the FIFA rankings. Achieving a higher accuracy than the simple ranking model suggests that there are some types of strategies (distinct combination of the three measurements) which are more efficient in some cases, meaning that a weaker team can beat a stronger team by playing in a certain way.

6.2 Centrality

Centrality measures were developed to answer the question regarding which node is the most important or central in the network. Centrality measures have been applied to understanding e.g. the influence, prestige or power of people or companies in a network. Centrality measures have been used to identify super-spreaders of diseases, and the most popular people in social networks. In sports, centrality measures have been utilized to identify the most influential players on the team. The first applications of centrality measures on human interactions were made by Bavelas (1948, 725-730), who was specifically intrigued by small-group communication and hypothesized a connection between centrality and influence. The early studies on the impact of centrality on group work reported that centrality was related with perception of leadership, personal satisfaction (of group members) and group efficiency (Leavitt 1949; Smith 1950; Bavelas 1950). (Newman 2018, 159-162)

The importance of a node can be defined in multiple ways, and therefore there is a variety of centrality measures. Borgatti (2005) proves that correctly applying the right centrality measure to a problem is critical as different centrality measures will yield different results. Therefore, understanding the context, and then applying the concept, is instrumental in a successful centrality analysis (Block et al. 2017). The most researched network measure is the degree centrality. (Newman 2018, 159-162)
Centrality measures have been used to identify the most influential twitter users in social networks (“influencers”) (Weng et al. 2010) and the most influential bloggers in specific communities (Akritidis et al. 2011). It has also been applied to forensic investigation in trying to identify suspects (Alzaabi et al. 2015). Zhang and Lin (2009) use a variety of centrality measures to identify the most essential proteins. Feeley et al. (2008) predicted the tenure and resignations of employees through the use of centrality metrics. In their study, they surveyed employees on their friendship. The results suggest that high out-degree was characteristic of employees that stayed at the company for longer. They also concluded that the strength of said friendships did not affect the results. Clemente et al. (2017) used four different centrality measures to identify the most connected and influential playing positions of the 2014 FIFA World Cup. In their results, they concluded that central midfielders were the players that most significantly contributed in building up the attack.

6.2.1 Degree Centrality

The simplest centrality measure is the degree centrality, which is a measurement of the number of edges in a node, sometimes simply called “degree”. In a directed graph, the degree centrality is extended to an in-degree and out-degree. In the case of a football passing network, the in-degree would be the count of received passes, and out-degree the amount of delivered passes. (Newman 2018, 159-60; Gudmundsson and Horton 2017, 17-18)

Degree centrality has a wide range of applications due to its simplicity. Additionally, it has a low cost of computation, compared to e.g. betweenness centrality, closeness centrality and eigenvector centrality.

Mclean et al. (2018) studied the games played throughout the 2016 EURO World Championship and analyzed the respective passing networks. In their analysis, they developed a pitch passing network on which they studied the prominent pitch zones on the field, based on the network measures. They concluded that there were no differences in team networks structures between knock out and group stages, and that high values of connectivity metrics did not differentiate between successful and unsuccessful teams. Lastly, their results suggest that the in-degree, out-degree and
within-degree of a pitch passing network can be analyzed to determine important and unimportant zones throughout the evolution of a game.

Clemente et al. analyzed Germany’s national team’s player-passing network of the 2014 FIFA World Cup. In the study, they prove how centrality can be used to identify collaboration patterns between players, both how they collaborate and the respective strengths and weaknesses of each team. In their results, they also conclude that the degree centrality is an effective discriminator between players. The central midfielders and central defenders have high degree centrality, and the goalkeeper and forwards have low degree centrality.

6.2.2 Eigenvector centrality

Wassily W. Leontief (1942) and John R. Seeley (1949) are considered as the eigenvector centrality pioneers. In their studies, they both realized the importance of the in-degree. Furthermore, Bonacich (1972) has been considered the person that developed its most modern form. Eigenvector centrality is a more precise measurement of degree centrality, since it does not only consider the degree of the node at hand, but also the neighbors of that node, i.e. its adjacent nodes. For a node to have a high eigenvector centrality, it does not only require a high degree centrality itself, but also the neighboring nodes need one. Degree centrality can be seen as a system, where each interaction is rewarded equally, while the eigenvector centrality system rewards interactions with more prominent (higher-centrality) nodes higher. The eigenvector centrality can be applied to both undirected and directed networks. However, complications can occur in directed networks, since the adjacency matrix will be asymmetrical. (Sargent and Bedford 2013; Newman 2018)

Eigenvector centrality has its applications in all fields. Lohmann (2010) applies the eigenvector centrality on fMRI data to draw conclusions on connectivity patterns in different spectral bands in the human brain. The eigenvector centrality has also been applied to ranking (American) college football teams nation-wide. In college football, teams rarely play out of their conference, making it difficult to determine the best team. Keener (1993) tackles this issue by ranking the college football teams based on network metrics such as the eigenvector centrality. Lastly, eigenvector centrality has
been applied to predicting the rise of lek-mating wire-tailed manakins in their social structures, where nodes represented male manakins, and the weight an index of the frequency of interactions (Blake et al. 2008). They were successful in predicting the male’s territorial tenure during this 4-year study.

![Weighted social network graph of wire-tailed manakins, which was used to predict the rise of males in their social structures. (Blake et al. 2008, 11-13)](image)

The eigenvector centrality also applies well to football passing networks. Grund (2012) studied just shy of 300 thousand passes, from 23 different teams, in the Premier league i.e. elite football. In his study, he analyzed the degree centrality as well as the Freeman centrality, which is a slight variation of the eigenvector centrality. In the results, he concludes that a low centralization of a network is associated with better team performance. He also confirms previous findings with panel data, considering unobserved characteristics.

Controversially, Linder (2017, 34-40) concludes that there is no significant advantage to playing decentralized. In her analysis, she looked at three teams of the season 2016/2017 premier league and was able to conclude that there was no advantage in playing a certain style of play as far as the outcome was concerned.
6.2.3 Betweenness centrality

The betweenness centrality, commonly just referred to as betweenness, is a measurement of the extent to which a node lies between the shortest path of two other nodes in the network. Suppose the case of football, where we want to move the ball from the goalkeeper to the forwards. Most often, this will require a sequence of passes from defender, to midfielder to forward. The players we use to move the ball to the forwards will have high betweenness, as they lie on the path between the goalkeeper and the forward. Nodes with the highest betweenness can be considered the most important, as other nodes in the network are dependent on the flow through them, in order to reach their destination. (Newman 2010, 173-176)

Betweenness has been used to explain the rise of the Medici family in Italy (Padgett and Ansell 1993; Jackson 2008). The Medici family had high betweenness in the network of influential families in Italy, which meant that they were the most important intermediary that could introduce and connect families.

Figure 13: An unweighted and undirected social network graph over the influential Italian families. The figure depicts the importance of the Medici family, as the intermediary. (Borgatti 2005)
Pena and Touchette (2012, 525-528) hypothesize that a player’s betweenness in a player passing network is an indicator of the impact of removing that player. Strategically, this can lead to the opposing side benefitting from this information by focusing on shutting down this player (Cotta et al. 2013, 40-43). This can also help to assess the strength of said team if that particular player is injured or sent off. Due to these facts, Pena and Touchette argue that a team should strive to have a balanced betweenness, in order to mitigate risk and maximize team performance.

Gonçalves et al. (2017) sought to identify how passing networks and positioning variables affect the game outcome, specifically in youth elite association football. They studied two games, one in the under15 and one in the under17 age group. In their results, they concluded that lower betweenness network centrality scores, a proxy for the dependency on each player was a good performance indicator.

Duch et al. (2010) used flow centrality to perform ranking simulations of the 2008 Euro Cup. In their player ranking simulation, they used flow centrality, a variant of betweenness centrality. They developed a transition network with shots on and off goal as considered outcomes. The play sequences from acquiring the ball all the way to a shot were analyzed and the players were ranked based on their flow centrality to the shot on and off goal nodes. The final results were used to quantify the efforts of each player.

6.2.4 Pagerank

A variation of the eigenvector centrality measurement is the Pagerank, which is a measurement of tendency to land at a specific node after x number of steps. Google was one of the pioneers in the field of Pagerank, applying it to the rank web-pages in their search results (Sergey and Lawrence 1998). Sergey and Lawrence (1998) were pioneers in understanding the applicability of network theory to the world wide web. The success of their search engine algorithm was behind the idea that a link from one web page to another could be seen as an endorsement. This led them to think of the web as a network. In this network, webpages were nodes, and hyperlinks pointing to other websites were directed edges. The Pagerank algorithm does not only count the frequency of hyperlinks but also their status. Hyperlinks from websites that are popular
are scored higher, as are hyperlinks from websites that have a low count of hyperlinks. (Langville and Meyer 2011)

Since Pagerank’s introduction in web search engine algorithms, it has been used in multiple fields. Gleich (2015) mentions examples of applications in bibliometrics, social and information network analysis, link prediction and recommendation, networks, biology, chemistry, neuroscience, and physics. Gleich (2015, 35-38) shows how Pagerank can be applied to road and urban space design. With its help, designers can predict the traffic flow and human movement. Jiang et al. (2008) prove in their study that Pagerank is the best network measure for predicting traffic. Gleich (2015, 30-35), in his turn, illustrates how sports tournaments or leagues can be simulated using it; teams are nodes, and a win is a directed edge, from the losing to the winning team, and the weight is the score by which the winning team won.

In football, Pagerank can be seen as the proneness of the ball to land to a specific player after an either fixed or random number of passes. In addition to Pena & Touchette’s (2012, 520-522) findings in centrality, they also observed two different types of patterns in Pagerank within well-performing teams. They noted that the key to Uruguay’s and the Netherlands’ success was the well-balanced Pagerank of players, whilst the Spanish and German teams were successful due to a few high performing players: Lahm, Schweinsteiger and Xavi, with high values in Pagerank. These results show that there is no ideal composition of Pagerank, but a team such as the Spanish or the German is more dependent on specific players.
6.2.5 Closeness centrality

The closeness centrality is a measurement of the mean distance between each node and all other nodes in a network (Newman 2010, 171). The closeness centrality was first defined by Bavelas (1950), who sees it as the inverse of the farness of the node, which is the sum of its distance to all other nodes in the network. The standard distance metric in networks is the shortest path between any given pair of nodes. The closeness centrality is the inverse of that distance metric for every possible node combination.

Borgatti (1995) uses four centrality metrics to study the distribution of AIDS in a social group. In this context, he considers the closeness centrality as an index of the expected time for a disease to arrive at a given node, depending on which node it entered from. Ni et al. (2011) study the centrality metrics of a network composed of scientific papers as nodes, and citations as links. They aim to study the impact of each field of research, and their change over time. They conclude that the closeness centrality increases between 1960–1985 to 1985–2010, indicating that sciences have become more interconnected. They also find that social sciences are on average more interconnected than natural science disciplines. Lastly, they describe the emergence of computer
science and mathematics through an increase of closeness centrality. In this context, this means that mathematics and computer science have become a more important part of other fields of research, too.

Gonçalves et al. (2017, 10-11) defined the closeness centrality measurement in the football context as an indicator of high intra-team, well-connected passing relations and found that higher scores in closeness were indicative of a successful football team. Peña and Touchette (2012, 517-520) argue that the closeness centrality is an indicator for how easy it is to move the ball between any given players of a team. In Peña and Touchette’s (2012) study, the 2010 FIFA World Cup data was analyzed. Their main finding was that closeness centrality, Pagerank and clustering all correlated strongly. These findings were in line with the general perception of the players’ performance reported by the media.

6.2.6 Clustering Coefficient

The clustering coefficient measures the tendency for networks to cluster together (Clemente and Grasse 2017). In the case of a social network, the idea is that a friend of your friend is more likely to also be your friend, than a randomly picked person in the network. Simply put, the clustering coefficient is a measurement of cohesion.

Tabak et al. (2014) use the directed clustering coefficient to measure systematic risk in a Brazilian inter-bank network. In their results, they display a negative correlation between directed clustering coefficient and domestic interest rates. The clustering coefficient has also been used to explain the tendency of actors to play with the same partners in multiple movies (Watts and Strogatz 1998). Another application of the clustering coefficient has been the distribution of cell towers to optimize the coverage. With the help of networks and the clustering coefficient, Hsu & Helmy (2005, 2006) were able to identify the ideal placement of cell towers. In their network, cell towers were nodes and each overlapping connection with a cellphone was marked as an edge.

In a sports context, the clustering coefficient is indicative of a team with high cooperation between players. Fewell et al. (2012, 13-14), in their study on basketball
strategies during the 2010 NBA playoffs, concluded that a high clustering coefficient lead to larger team entropy, allowing a higher possibility of passing avenues or routes from player to player to reach the outcome of scoring in a transition network.

Cotta et al. (2013) analyzed the network of the Spanish national team, which won the 2010 FIFA World Cup. In their analysis, they studied variables such as number of passes, length of the chain of passes, and network measures such as player centrality and clustering coefficient. They concluded that a high clustering coefficient was indicative of the high passing and possessive game that was emblematic of the Spanish national team’s playstyle. They also found that the opposing team could effectively impact Spain’s clustering coefficient, passing length and passing speed negatively, as well as force key players from their natural position on the football pitch. Pena and Touchette (2012) also observed similar patterns when analyzing the FIFA World cup 2010.

Yamamoto & Yakamoto (2011) compared in their study the ability of a team to successfully create passing triangles to its ability to create scoring chances. Their results found a positive correlation between the two, meaning that the team that was more successful than their opponent in creating passing triangles, was also the more threatening side in terms of goal scoring opportunities.
6.2.7 Density

Network density is calculated as the number of edges divided by the total number of possible edges. The network density metric was modernized by Bott (1957). Balkundi & Harrison (2006, 49-68) show in their study how a team with high network density is more effective. They conducted their study on thirty-seven teams in their natural context to determine the hindering or helping factors on team performance from a managerial perspective. They concluded that high density, i.e. leaders who are central in teams, as well as interpersonal ties within the team are characteristic of an effective team, something they refer to as the “density-performance hypothesis”.

Fisher and Shavit (1995) apply networks to describe the differences in social networks between countries. In their study, they analyze the difference between Israel and the US. In their results, they find a disparity in network density, where the Israeli network has a higher density. They account this to the differences in social cultures, with the US individualism and Israeli group orientation. Grossetti (2007) performs a similar study where he compares the social networks of the French with those in the US. In this study, the network density was similar for both countries. But other differences were identified, such as the higher connectedness with kin in France, and high levels of isolation in the US (10% of respondents with four or less connections in the US vs. 1% in France).

In a football context, the density-performance hypothesis would suggest that a team where players have all successfully passed the ball between each other, would yield better results. In addition to the findings that Grund (2012) made in centrality, he also observed that an increase in network density (sometimes referred to as intensity) lead to increased performance. These results indicate that a player network, which is capable of effectively utilizing multiple passing routes and avenues, will be more successful. Clemente et al. (2015, 10-13) show in their results that in the games they analyzed, the average density is usually higher in the first half (0,48) than in the second (0,32). This suggests that, on average, teams are better connected in the first half, and that the impact of individual players increases in the second. Lastly, these findings go hand in hand with the findings of Fewell et al. (2012), who concluded that an NBA team with high entropy is more successful.
6.3 Styles of play

The purpose of this section is to examine the existing research within football styles of play. In the analysis chapter, this topic will be revisited, and a comparison will be made between the identified styles of play and the styles presented here. An attempt will also be made to investigate whether the styles of play identified in networks can exhibit these strategies.

There is overwhelming evidence suggesting that accounting the styles of play of football teams is an important step in all facets of match analysis (Buchheit and Laursen 2013; Reilly 2005; Bradley et al. 2011; Duarte et al. 2012; James et al 2002; Pollard and Reep 1997; Pollard, Reep and Hartley 1988). Originally, Pollard et al. (1988) analyzed the matches played between six national teams at the 1982 World Cup. In their analysis, they considered six variables as features of playing styles which led to the discovery of three distinct styles of play: firstly, direct and possessive (elaborate) style; secondly, the use of crosses, and thirdly, regaining possession close to the opponent’s or own goal. Direct play is a style of play where the team attempts to move the ball forward as fast as possible, through the use of fast and long passes (Fernandez-Navarro et al. 2016). A possessive style of play, on the other hand, is one where the team attempts to control the ball for the majority of the game, by avoiding dispossessions and risky passes.

Already in 1968, Reep & Benjamin identified the significance of direct play. They also discovered that one goal is scored in every ten shots taken, and that 80% of goals are scored after a sequence of three or less passes. However, the primary discovery Reep and Benjamin made was that a successful style of play can be based upon maximizing the chances of scoring (Tenga 2010).

Possessive style of play has been considered the most effective style (Hughes et al.1988; Hughes and Churchill 2004; Hughes and Franks 2005; Hughes and Snook 2006). Controversially, many studies have also proven the potency of direct play (Bate 1988; Hughes1990; Olsen and Larsen 1988; Reep and Benjamin 1968). However, research has proven that possessive play is only effective when the team is able to enter the final third of the pitch (Bate 1988; Hughes 1990; Hughes and Snook 2006).
To complicate the discussion even further, Collet (2012) proved that the possessive play style is only effective in elite football. The findings of this research have been applied by managers in British football (Hughes and Franks 2004).

In more recent years, there have been multiple endeavors to redefine styles of play, with the help of state-of-the-art datasets, and the emergence of new styles of play. One of the primary findings have been the distinctions made between possessive and direct play. Lago-Peñas et al. (2017) analyzed 240 games of the Chinese super league 2016 season. They perform linear regression and factor analysis, which yields five factors:

“Factor 1 (“possession” style of play, correlated with the ball possession, ball possession in opponent half and in the final third of the field, positional attacks, passes, accurate passes, passes forward and back), Factor 2 (set pieces attack, correlated positively with the number of set pieces attacks, and attacks), Factor 3 (counterattacking play, correlated with interceptions, interceptions in opponents half, recovered balls, and number of counterattacks) and Factor 4 and 5 (transitional play, correlated with lost balls, and picking up free balls).”

7 Analysis

The analysis chapter will consist of separate sections. First, descriptive statistics and visualizations of the data is presented and analyzed. The data is also used to describe the rationale behind the chosen analysis method. Then, the steps taken to perform the analysis are covered. The analysis is followed by an exploratory and descriptive analysis, a presentation of the results and real-world applications. Each step is supported by visualizations.

7.1 Descriptive statistics

The data used in this study is the spatio-temporal data covering each event throughout every game in the FIFA World Cup 2018. A single event can be a shot, pass, tackle, aerial duel etc. A single row of data and all of its variables can be seen in the appendices, but the most important content can be found in Table 1.
Table 1: The most important variables in a single row of data

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Example value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event type</td>
<td>Pass</td>
</tr>
<tr>
<td>Player</td>
<td>Paul Pogba</td>
</tr>
<tr>
<td>Recipient</td>
<td>Olivier Giroud</td>
</tr>
<tr>
<td>Start Location (x,y)</td>
<td>[50,50]</td>
</tr>
<tr>
<td>End Location (x,y)</td>
<td>[80,90]</td>
</tr>
<tr>
<td>Time stamp</td>
<td>34:16:00</td>
</tr>
<tr>
<td>Outcome</td>
<td>Successful</td>
</tr>
</tbody>
</table>

This data is then transformed to a network, by creating a directed connection between the players (Paul Pogba and Olivier Giroud) (Figure 16), with the weight of 1, which increases by 1 with each new pass.

Figure 16: Network representation of one row of data

7.2 Passing statistics

Figure 17 shows the accumulated data of all events from all of the games played in the World Cup 2018. As the figure shows, the majority of the events during a football game is either passing or receiving the ball. In addition to passing and ball receipt, events such as pressure, i.e. applying pressure on the player with possession, ball recovery and duel were the third, fourth and fifth most common events in the tournament.
Figure 17: The 10 most frequent events in all games played

Figure 18 shows the ten least common events in all games. Own goal was the least frequent event, despite being relatively frequent in this year’s World Cup, where the own goal record from 1998 was more than doubled. Additionally, offsides, errors, bad behavior and referee ball-drops were very infrequent.

Figure 18: The 10 most infrequent events in all games played

Detailed information of passes is also captured. Passes are divided into a multitude of types; a pass can be categorized by its height, body part used and pass intention. In
Figure 19, the variety and count of different passes used in the World Cup 2018 is displayed:

![Pass types](image)

*Figure 19: The frequency of different pass types used*

A regular pass, which has no other intention than to move the ball from one player to another is not separately noted here. The most common intent with passes were switch, where a team moves the ball along the width of the pitch, from left to right or vice versa. The second most common was crossing, where a player passes the ball into the opposing team’s penalty area. The third most common was shot assist, which is a pass that lead to a shot.

Lastly, the pass height is captured (Figure 20). The most common pass type is a ground pass, where the ball never leaves the ground. Second are high passes where the ball, at its peak, goes over shoulder height. The least common pass height is the low pass, where the ball leaves the ground, but never over shoulder height.
In Table 2, the passing statistics from all games played is presented, as is the difference between winning and losing teams. In the Winning Team column are stats describing the average values of those teams that won the match, and in the Losing Team column are those that lost. Winning teams had approximately thirty more passes on average per game as well as a slightly higher pass success percentage. The surprising finding in Table 2 is that losing teams had higher average distance on successful passes, even if the difference is minimal. Losing teams also had a higher average of meters gained per pass.

Table 2: Statistics of differences in passing between winning and losing teams.

<table>
<thead>
<tr>
<th></th>
<th>Winning Team</th>
<th>Losing Team</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. count of passes per game</td>
<td>460</td>
<td>428</td>
</tr>
<tr>
<td>Avg. pass success %</td>
<td>82.1%</td>
<td>81.2%</td>
</tr>
<tr>
<td></td>
<td>19.79 m</td>
<td>19.96 m</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Avg. distance of a successful</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pass</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. distance of an unsuccessful pass</td>
<td>28.86 m</td>
<td>29.34 m</td>
</tr>
<tr>
<td>Avg. meters gained per pass</td>
<td>3.55 m</td>
<td>3.65 m</td>
</tr>
<tr>
<td>Avg. meters gained by passing per game</td>
<td>1633.62 m</td>
<td>1559.80 m</td>
</tr>
<tr>
<td>Avg. count of passes per possession</td>
<td>8.5</td>
<td>7.7</td>
</tr>
</tbody>
</table>

The longer distance of passes is due to a higher frequency of counter attacking and a direct style of play. Losing teams tended to take advantage of lost possession of the winning team and attempt to quickly move the ball forward. The higher meters gained per pass is a result of winning teams passing the ball more backwards and sideways. Winning teams were characterized by a more possessive style of play. The most telling statistic of the table is the average count of passes per possession. Winning teams were more successful at retaining the ball and dominating possession. The statistics were slightly skewed by the large amount of upsets in the World Cup. When analyzing a larger dataset, the differences in passing statistics of winning and losing teams would be more apparent.

The use of the pitch length and width was also analyzed but no significant differences between losing and winning teams were identified.

7.3 Network Analysis

In this section, the dataset is analyzed with the help of social network analysis package for Python, NetworkX. This section starts with visualizations and analysis of team networks where max or min values were present in the network metrics. Each network will focus on a specific network metric. Furthermore, each network graph will be analyzed, identifying how the network metric is visible in the visualization, and what kind of behavior is expected from a network with such values.
With the help of a player passing network, a variety of different variables can be communicated and visualized in an easily digestible way. In the network graph below, a multitude of factors can be seen. First, the node size represents the relative number of passes, which that node has received. The thickness and opacity (transparency) of the passing lines are based on the frequency of passes between the two nodes: a thick line indicates many passes. The size of the arrowhead is an indicator of directionality of the passing route. The color of the node is an indicator of how successfully that player has received passes, ranging from green (successfully) to yellow (unsuccessfully). Lastly, the players are placed in their average location captured throughout the game. The average location is calculated as the sum of x and y coordinates, divided by the count of events that the player has participated in. Lastly, if a player is substituted off, the new player’s actions are counted towards the node of the player that was substituted. The player numbers and network metrics can be seen in the table below the figure.

7.3.1 Network eigenvector centrality
Figure 21: England’s player passing network from their game against Sweden in the group stages, which ended 0 – 2 in favor of England. Just by glancing at the graph, one can make many conclusions about England’s overall game strategy. There are some passing routes that are very apparent, between #5 (John Stones) and #2 (Kyle Walker), as well as between #2 and #12 (Kieran Trippier). The pass success rate is very good at the defenders but starts to deteriorate higher up the pitch towards the strikers (#9 & #10). Of all the games played in the World Cup 2018, this English side was the highest in average player eigenvector centrality, and lowest in deviation of eigenvector centrality. The high centrality values can be seen by visually inspecting the network graph. Players are very well-connected and have utilized a wide range of different available passing routes. #10 (Raheem Sterling) has the highest centrality value. By inspecting the graph, his critical role in connecting and moving the ball up the pitch can be identified. Overall, England’s strategy revolves around playing the...
ball forward to wingers (#18 and #12), who in turn distribute the ball by crossing or passing the ball towards the center.

Figure 22 presents the network graph of the side with the lowest average eigenvector centrality, Uruguay, in their game against Portugal, which they won 2 – 1. By examining the centrality values, the dominance of certain Uruguayan players can be spotted, namely #9, #21 and #8. These three players have high centrality values, while others have low; this is an indication of a team that is highly dependent on a few players. The graph also clearly illustrates this strategy; there is very little possessive play around the defenders (compared to England’s network graph) and the aim of the team is to move the ball to #9 as fast as possible. The Uruguayan side lacks structure and is vulnerable to injury or man-marking of key players. In comparison to England’s player passing network, which had the highest average centrality, Uruguay’s network has some players that are barely utilized. This can be considered as a weakness since Uruguay’s threats are very apparent and vulnerable to counter-attacking.
7.3.2 Network betweenness centrality

The next network graph presents France’s passing network in their game against Denmark, which ended 0–0. In this network, France had the overall highest average betweenness values. A high average betweenness is indicative of a team where multiple passing routes are used when moving the ball from one point to another. This phenomenon is identifiable by visually inspecting the graph. France has an evenly distributed team both length- and widthwise. France has a short average pass length, moving the ball from defenders to midfielders, who in turn distribute the ball either to the forwards or back to the defenders. Very few passes are made directly from defender to forward. A team with a well-distributed betweenness is less vulnerable to countering since the dependency on players is evenly distributed across the team.
7.3.3 Network cluster coefficient

The clustering coefficient is characteristic of a team that creates passing triangles, and has player synergies, where certain players are often playing together. A high clustering coefficient is often emblematic of a team that prefers a possessive play style. In the next graph, Morocco’s passing network in their game against Portugal, which they lost 1 – 0, is presented. In this game, Morocco had the overall highest average clustering coefficient. Multiple passing triangles can be spotted in the graph, the most apparent one being between #2, #14 and #7. A high clustering coefficient is generally a positive indicator, since it indicates that the team has successfully possessed the ball and succeeded in passing it. Past research indicated that a high clustering coefficient was characteristic for winning teams.
Figure 24: Morocco’s player passing network

7.3.4 Network Pagerank

The overall highest Pagerank for a single player in all games was Sardar Azmoun (#20) from Iran in their game against Morocco, which Iran won 1 – 0. Sardar Azmoun also had the highest eigenvector centrality of all of the players. When inspecting the graph and the network metrics, Iran’s intended achievement in this game becomes very apparent: the players were trying to move the ball to Sardar Azmoun, with through passes, high passes and crosses. High value in Pagerank also signals that Sardar Azmoun was likely to lose possession, as a result of shot or dispossession. A high value of Pagerank for strikers is generally good, if the striker is able to retain the ball.
In this section, various examples of changes and differences in network metrics were presented with the help of network graphs. Different network metrics are clearly visible through the visualization of the network. Different playstyles are visible both in the network metrics and in the network visualization. Large amounts of passing data can be quickly understood with the help of a network graph. A team’s passing strategy can be summarized with the help of a network visualization. The network metrics provide more in-depth insight and can be used to look beyond the visualization. The next section will focus on identifying distinct strategies by the use of k-means clustering.

7.3.5 K-means Clustering

For clustering, each game is split to four parts: home vs. away 1st half; away vs. home 1st half; home vs. away 2nd half and finally, away vs. home 2nd half. Features are
separately extracted from all these four phases of a game. Through the use of feature engineering activities such as feature selection and feature extraction, the optimal features were found. Football domain expertise and past research was also used to support the feature engineering phase. The selected features and their description are found in Table 3. The features represent separate subsystems of a team network. The features highlight network metrics of specific player positions, making the distinction between strikers, midfielders, defenders and players on the right and left wing.

Table 3: The features selected for clustering

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Clustering</td>
<td>The average clustering coefficient of a team</td>
</tr>
<tr>
<td>Average Betweenness</td>
<td>The average betweenness centrality of a team</td>
</tr>
<tr>
<td>Deviation Betweenness</td>
<td>The standard deviation of the betweenness centrality of a team</td>
</tr>
<tr>
<td>Deviation Pagerank</td>
<td>The standard deviation of the Pagerank of a team</td>
</tr>
<tr>
<td>Average Degree</td>
<td>The average degree of a team</td>
</tr>
<tr>
<td>Defensive Betweenness</td>
<td>The average betweenness of a team’s defenders</td>
</tr>
<tr>
<td>Defensive Centrality</td>
<td>The average eigenvector centrality of a team’s defenders</td>
</tr>
<tr>
<td>Defensive Closeness</td>
<td>The average closeness centrality of a team’s defenders</td>
</tr>
<tr>
<td>Defensive Pagerank</td>
<td>The average Pagerank of a team’s defenders</td>
</tr>
<tr>
<td>Central Clustering</td>
<td>The average clustering coefficient of a team’s central (length wise) players, i.e. center backs, central midfielders and central strikers. Goalkeeper is excluded.</td>
</tr>
<tr>
<td>Midfield Clustering</td>
<td>The average clustering coefficient of a team’s midfielders</td>
</tr>
<tr>
<td>Right Clustering</td>
<td>The average clustering coefficient of a teams’ right-wing players. The players were decided based on their average position throughout the game, some teams had multiple players on the wing, while some even had 0.</td>
</tr>
<tr>
<td>Left Clustering</td>
<td>The average clustering coefficient of a team’s left-wing players. The players were decided based on their average position throughout the game, some teams had multiple players on the wing, while some even had 0.</td>
</tr>
<tr>
<td>Striker Centrality</td>
<td>The average eigenvector centrality of a teams’ strikers</td>
</tr>
<tr>
<td>Striker Pagerank</td>
<td>The average Pagerank of a team’s strikers</td>
</tr>
<tr>
<td>Striker Clustering</td>
<td>The average clustering coefficient of a team’s strikers</td>
</tr>
</tbody>
</table>
To identify the optimal number of clusters in the dataset, the elbow method was used and cross-referenced with the silhouette method. The silhouette and elbow method both yielded three as the optimal number of clusters. Three clusters indicate that in the World Cup, three distinct strategies were used.

Principal component analysis, a tool for visually inspecting clusters, was used to visualize the multifeatured data in a two-dimensional space. The cluster centers are visualized with black X marks in Figure 26.

![Figure 26: Principal component analysis of clusters. Cluster centers can be seen by black x on the graph.](image)
7.3.6 Inspecting the clusters

To understand the clusters and their characteristics, a heatmap was developed, in which the averages of all features for each cluster can be seen. All features are standardized separately and are placed on a scale from -1 (min) to 1 (max). In the heatmap, dark blue represented high values, while lighter colors represent lower values. The features are presented on the Y-axis and the clusters can be seen on the bottom X-axis (Figure 27).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Clustering</td>
<td>0.93</td>
<td>-0.52</td>
<td>-0.73</td>
</tr>
<tr>
<td>Average Betweenness</td>
<td>-0.13</td>
<td>-0.063</td>
<td>0.21</td>
</tr>
<tr>
<td>Deviation Betweenness</td>
<td>-0.13</td>
<td>0.16</td>
<td>0.042</td>
</tr>
<tr>
<td>Deviation Page Rank</td>
<td>-0.21</td>
<td>0.9</td>
<td>-0.43</td>
</tr>
<tr>
<td>Average Degree</td>
<td>0.096</td>
<td>-0.72</td>
<td>0.43</td>
</tr>
<tr>
<td>Defensive Betweenness</td>
<td>-0.094</td>
<td>-0.89</td>
<td>0.79</td>
</tr>
<tr>
<td>Defensive Centrality</td>
<td>0.1</td>
<td>-0.61</td>
<td>0.34</td>
</tr>
<tr>
<td>Defensive Closeness</td>
<td>0.75</td>
<td>-0.77</td>
<td>-0.33</td>
</tr>
<tr>
<td>Defensive Page Rank</td>
<td>-0.078</td>
<td>-0.94</td>
<td>0.8</td>
</tr>
<tr>
<td>Central Clustering</td>
<td>0.85</td>
<td>-0.45</td>
<td>-0.69</td>
</tr>
<tr>
<td>Midfield Clustering</td>
<td>0.92</td>
<td>-0.51</td>
<td>-0.73</td>
</tr>
<tr>
<td>Right Clustering</td>
<td>0.78</td>
<td>-0.41</td>
<td>-0.64</td>
</tr>
<tr>
<td>Left Clustering</td>
<td>0.79</td>
<td>-0.5</td>
<td>-0.58</td>
</tr>
<tr>
<td>Striker Centrality</td>
<td>0.12</td>
<td>0.85</td>
<td>-0.78</td>
</tr>
<tr>
<td>Striker Page Rank</td>
<td>0.044</td>
<td>0.85</td>
<td>-0.69</td>
</tr>
<tr>
<td>Striker Clustering</td>
<td>0.79</td>
<td>-0.067</td>
<td>-0.91</td>
</tr>
</tbody>
</table>

Figure 27: Heatmap over the feature averages of each cluster

The heatmap serves as a powerful tool to understand the characteristics of the clusters. Cluster 1 is characterized by a high average clustering coefficient, and their play revolves around the midfield. In cluster 1, strikers have low centrality measures but a high clustering coefficient, indicating that the strikers are involved in building up the play from the midfield. Cluster 2 is the opposite of cluster 1. Cluster 2 strategy revolves around striker centrality and Pagerank. Cluster 2 has low values in all network metrics.
in the defense and midfield. A high deviation in Pagerank signals that the strategy involves playing the ball to selected players (in this case strikers) while other players are less involved. Lastly, cluster 3 has low values in average clustering and in all network metric values of the strikers and midfielders. Cluster 3 is characteristic of a defensive team where the defenders build up play and serve as the mediators on the pitch. The defender’s role as mediators can be seen in their high values of betweenness centrality.

7.4 Closer analysis of each cluster

Another excellent tool to aid in understanding the attributes of each cluster is to analyze each cluster center. In this section, the network graph that is closest to each cluster center is analyzed by inspecting the network metrics and visually analyzing a network graph. The networks that are closest to the cluster centers resemble the average of the observations in that cluster, as it is the most central one. Therefore, the cluster centers are a good overall representation of the observations in each cluster.

7.4.1.1 Cluster 1

Cluster 1 was characterized by play revolving heavily around their midfield. The cluster 1 cluster center is Argentina’s player passing network in the 2nd period in their game against Nigeria, in the group stages (Figure 28). The focus around the midfield can be instantly spotted in the network graph. The play revolves around possessing and passing the ball in the midfield. Players such as #7 and #14 are key players in the midfield area. #7 in particular has received the most passes, with a high success rate, he has also distributed the ball around the pitch well. The high striker clustering value can be seen in in their forward #9 (Lionel Messi). Messi is also known for dropping low and building play with the midfielders, which was also the case in this game. Argentina was unsuccessful in creating scoring chances in this game and period, as they only have an expected goal of 0.189. This might be a result of their inability to transfer the ball to their strikers.
Cluster 2 was characterized by a strategy revolving around their strikers, and their ability to receive passes and create scoring chances. The cluster 2 cluster center is Russia’s player passing network in period 2 against Egypt, in the group stages (Figure 29). All roads lead to Rome, or in the case of Russia’s network, all passes lead to #22 (Artem Dzyuba). #22 is the biggest node in this network, meaning that he has received the most passes; the color of the node also indicates that he has unsuccessfully received many of those passes. #22 also has the highest centrality and Pagerank values of anyone in the network. All forward players have high centrality values, while defenders and midfielders have low. To move the ball to the forward placed players is
clearly the strategy of cluster 2. High-risk and high-reward forward passing, and crossing is indicative of cluster 2.

![Figure 29: Network of cluster center of cluster 2](image)

<table>
<thead>
<tr>
<th>Number</th>
<th>Betweenness</th>
<th>Centrality</th>
<th>Clustering</th>
<th>Page Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Igor Akinfeev</td>
<td>#1</td>
<td>0.036</td>
<td>0.044</td>
<td>0.149</td>
</tr>
<tr>
<td>Mario Figueres Fernandez</td>
<td>#2</td>
<td>0.002</td>
<td>0.299</td>
<td>0.174</td>
</tr>
<tr>
<td>Ilya Kutepov</td>
<td>#3</td>
<td>0.006</td>
<td>0.163</td>
<td>0.158</td>
</tr>
<tr>
<td>Sergei Ignashevich</td>
<td>#4</td>
<td>0.162</td>
<td>0.19</td>
<td>0.117</td>
</tr>
<tr>
<td>Denis Cheryshev</td>
<td>#6</td>
<td>0.195</td>
<td>0.293</td>
<td>0.131</td>
</tr>
<tr>
<td>Yuri Gazinsky</td>
<td>#6</td>
<td>0.008</td>
<td>0.286</td>
<td>0.16</td>
</tr>
<tr>
<td>Roman Zobnin</td>
<td>#11</td>
<td>0.134</td>
<td>0.186</td>
<td>0.141</td>
</tr>
<tr>
<td>Aleksandr Golovin</td>
<td>#17</td>
<td>0.081</td>
<td>0.402</td>
<td>0.161</td>
</tr>
<tr>
<td>Yuri Zhirkov</td>
<td>#16</td>
<td>0.202</td>
<td>0.23</td>
<td>0.14</td>
</tr>
<tr>
<td>Aleksandr Samoedov</td>
<td>#19</td>
<td>0.006</td>
<td>0.469</td>
<td>0.171</td>
</tr>
<tr>
<td>Artem Dzyuba</td>
<td>#22</td>
<td>0.03</td>
<td>0.457</td>
<td>0.156</td>
</tr>
</tbody>
</table>

### 7.4.1.3 Cluster 3

Defenders and their role in the team is in the essence of cluster 3. The cluster 3 cluster center is Denmark’s player passing network of the 1st period in their game against Peru, in the group stages (Figure 30). The network graph clearly telegraphs the common passing routes of the Danish team. #6, #4 and #8 were all heavily involved in the passing game. The centrality and Pagerank values are low at forward positions, which is indicative of a team that has failed to translate high possession and many
passes to actual scoring chances. Denmark had an expected goal of 0.348 in this period. The clustering coefficient is evenly distributed throughout the team.

![Denmark's Player Passing Network](image)

Figure 30: Network of cluster center of cluster 3

<table>
<thead>
<tr>
<th>Number</th>
<th>Betweenness</th>
<th>Centrality</th>
<th>Clustering</th>
<th>Page Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kasper Schmeichel</td>
<td>#1</td>
<td>0.136</td>
<td>0.108</td>
<td>0.111</td>
</tr>
<tr>
<td>Simon Thorup Kjær</td>
<td>#4</td>
<td>0.144</td>
<td>0.393</td>
<td>0.14</td>
</tr>
<tr>
<td>Andreas Christensen</td>
<td>#6</td>
<td>0.034</td>
<td>0.468</td>
<td>0.171</td>
</tr>
<tr>
<td>William Kvist Jørgensen</td>
<td>#7</td>
<td>0.026</td>
<td>0.292</td>
<td>0.142</td>
</tr>
<tr>
<td>Thomas Delaney</td>
<td>#8</td>
<td>0.032</td>
<td>0.398</td>
<td>0.172</td>
</tr>
<tr>
<td>Nicolai Jørgensen</td>
<td>#9</td>
<td>0.156</td>
<td>0.167</td>
<td>0.127</td>
</tr>
<tr>
<td>Christian Eriksen</td>
<td>#10</td>
<td>0.08</td>
<td>0.306</td>
<td>0.131</td>
</tr>
<tr>
<td>Henrik Dalsgaard</td>
<td>#14</td>
<td>0.247</td>
<td>0.186</td>
<td>0.112</td>
</tr>
<tr>
<td>Jens Stryger Larsen</td>
<td>#17</td>
<td>0.056</td>
<td>0.266</td>
<td>0.133</td>
</tr>
<tr>
<td>Yussuf Yurary Poulsen</td>
<td>#20</td>
<td>0.044</td>
<td>0.21</td>
<td>0.120</td>
</tr>
<tr>
<td>Pione Polo Ememija</td>
<td>#23</td>
<td>0.003</td>
<td>0.315</td>
<td>0.166</td>
</tr>
</tbody>
</table>

7.5 Applications and analysis of strategies

In this chapter, the real-world applications and results of strategic matchups are presented. The expected goal difference, the difference between the home and away teams’ expected goals, is one of the key metrics in analyzing a team’s performance. In Table 4, an example case is introduced, where the expected goal difference is +1 for the home team and -1 for the away team, for the first half of the game.
Table 4: An example case showcasing the “expected goal difference” metric

<table>
<thead>
<tr>
<th>Team</th>
<th>Cluster</th>
<th>Period</th>
<th>Expected goal</th>
<th>Expected goal difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>1</td>
<td>1</td>
<td>2.35</td>
<td>+1.00</td>
</tr>
<tr>
<td>Away</td>
<td>3</td>
<td>1</td>
<td>1.35</td>
<td>-1.00</td>
</tr>
</tbody>
</table>

Each game has been divided into two halves, the first and the second. In the table below, the count of each cluster is demonstrated, as is their expected goal difference mean, standard deviation, min, 25% percentile, 50% percentile, 75% percentile and max during the time of each half (45 minutes). Table 5 shows that cluster 1 was the most common strategy with 103 appearances, cluster 3 had 85 and cluster 2 had 64. On average, they all performed equally well. The highest standard deviation was in cluster 2. This is not surprising, since the high-risk, high-reward playstyle is likely to yield mixed results. Cluster 2 also showed the highest min value of -2.647. Cluster 3 had overall the most consistent results, having the lowest deviation and highest min and lowest max values.

Table 5: Table showing the expected goal difference for the three clusters.

<table>
<thead>
<tr>
<th></th>
<th>count</th>
<th>count</th>
<th>mean</th>
<th>Std</th>
<th>Min</th>
<th>25 %</th>
<th>50 %</th>
<th>75 %</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Expected Goal Difference</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>103</td>
<td>0.005</td>
<td>0.769</td>
<td>-2.190</td>
<td>0.499</td>
<td>0.035</td>
<td>0.455</td>
<td>2.647</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>64</td>
<td>-0.009</td>
<td>0.899</td>
<td>-2.647</td>
<td>-0.461</td>
<td>-0.047</td>
<td>0.519</td>
<td>2.190</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>85</td>
<td>0.001</td>
<td>0.745</td>
<td>-1.777</td>
<td>-0.495</td>
<td>0.003</td>
<td>0.488</td>
<td>2.403</td>
<td></td>
</tr>
</tbody>
</table>

Next, the expected goal effect of a strategic shift is analyzed. The expected goal effect is the difference between the second period and first period expected goal difference of a team. The expected goal effect is illustrated in Table 6. The away team goes from a -1 expected goal difference in the first half to a +3 expected goal difference in the second half, resulting in a total expected goal effect of +4 as a result of their strategic shift.
Once the expected goal effect is calculated, the strategic shifts in all of the games played in the World Cup can be analyzed. In Table 7, the expected goal effect of different strategy shift combinations is presented. In the first column is the first half’s strategy cluster, and in the second column the strategy cluster used in the second half. Here again, the success metric is the expected goal effect. The most telling column is the mean column, which shows the average expected goal effect of the strategic shift.
Cluster 1 was the only one to yield a positive expected goal effect when not shifting strategy. Cluster 2 and 3 both present a negative expected goal effect when not shifting strategy. This is an indication of a strategy that is more susceptible to countering. The most interesting finding in both of the tables presented so far is that no cluster is performing better than the other. Instead, the clusters are scenario-based, meaning that they yield different results based on the situation. The expected goal effect matrix can prove to be a powerful tool due to that reason. The min, max and std columns can be used to determine the volatility of the strategic shift, which allows the team to identify their situation and plan accordingly. Lastly, none of the strategies are overwhelmingly weighted towards the first or second period. This means that none of the identified strategies are a result of fatigue.

The same data can also be aggregated in a way where different cluster strategies can be compared on how they match each other. In Table 8, a strategy matchup matrix is presented. The two first columns represent the matchup of clusters, and the key metric here is the expected goal difference. Matchups of the same cluster strategy will have a mean expected goal difference of 0. The matchup matrix can be used to identify favorable matchups, which can be used for a strategic benefit by changing the strategy based on the opponent’s play style. These results can be utilized both pre-game and in-game.

Table 8: the matchup matrix shows the expected goal difference between the different cluster matchups.

<table>
<thead>
<tr>
<th>Clust1</th>
<th>Clust2</th>
<th>Count</th>
<th>Mean</th>
<th>Std</th>
<th>Min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>44</td>
<td>0</td>
<td>0.655</td>
<td>-1.244</td>
<td>1.244</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>28</td>
<td>-0.097</td>
<td>0.979</td>
<td>-2.190</td>
<td>2.647</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>31</td>
<td>0.104</td>
<td>0.717</td>
<td>-1.061</td>
<td>1.777</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>28</td>
<td>0.097</td>
<td>0.979</td>
<td>-2.647</td>
<td>2.190</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>18</td>
<td>0</td>
<td>0.801</td>
<td>-1.714</td>
<td>1.714</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>18</td>
<td>-0.184</td>
<td>0.884</td>
<td>-2.403</td>
<td>1.641</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>31</td>
<td>-0.104</td>
<td>0.717</td>
<td>-1.777</td>
<td>1.061</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>18</td>
<td>0.184</td>
<td>0.884</td>
<td>-1.641</td>
<td>2.403</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>36</td>
<td>0</td>
<td>0.696</td>
<td>-1.472</td>
<td>1.472</td>
</tr>
</tbody>
</table>

The matchup matrix (Table 8) presents slight advantages in certain matchups. For example, cluster 3 has an expected goal difference of 0.184 against cluster 2. This is
an interesting finding considering the uniqueness and differences in the two strategies. Additionally, some matchups are more volatile in their results, and can therefore be used to avoid upsets.

7.6 Conclusion

Applying advanced data analytics to sports has become one of the cornerstones of a successful sports organization. Understanding what impacts the performance of a sports team and how is unique for each sport. In team sports, understanding and analyzing the cooperation and interaction between individuals is extremely significant and the importance is only emphasized in sports with large teams, such as football. One of the most prominent tools used to analyze complex systems is social network analysis.

This study was conducted on all of the matches played in the World Cup 2018 with data provided by StatsBomb. This material was selected as it contained accurate data from the events of the games played in the World Cup. These matches were specifically well-suited for the problem, as the World Cup national teams do not have a significant amount of time to practice and perfect strategies, which could create biased results.

The purpose of this study was to cover the body of literature surrounding social network analysis applied to sports and to extend it by suggesting new methods and results in the realm of football. Past research has proven that social network analysis is an effective tool for sports analytics, and that it can be used to measure the performance of a team. Past research has failed to convert these findings into actionable insight and to suggest methods how football managers can benefit from them. One of the contributions of this study was the discovery of new tools and techniques, which support managers in strategic decision-making. This study is the first to suggest tools and techniques that can be utilized in-game. Additionally, this study suggests a new success metric: expected goals, which can be more effective than the outcome of the game for measuring team performance. Lastly, the foundation for feature selection in network analysis was set for future machine learning endeavors;
this study was the first to make the distinction between player positions, separately considering subsystems of the whole team network.

In the analysis and results chapter, tools and techniques that can be directly utilized by a football manager to support strategic decision making were suggested. One of these tools is the identification of different football strategies by using $k$-means clustering. By studying past research, it was possible to identify network metrics that differentiate between successful and unsuccessful teams. The network metrics and how they translate to the football pitch were also explained. This knowledge was utilized in the selection of features for the clustering. The technique yielded three distinct strategies that were thoroughly analyzed, identifying both weaknesses, strengths and characteristics of each strategy. The identification of distinct strategies enabled an analysis of the strategic shifts in-game. When analyzing strategic shifts, it was possible to identify favorable actions based on a given scenario. A football manager can benefit from this information by practicing a certain playstyle before an upcoming game (pre-game analytics) or by identifying weaknesses of the opponent’s strategy. The manager can also use this information to support strategic decision making during a game (in-game analytics).

Another contribution of this study was the proposal of a new metric for measuring success in football. One of the problems in studying football is the random nature of the sport; the better team does not always win. This results in much variance, which makes the sport difficult to study. My proposed method was to analyze the expected goals instead of the game outcome. The expected goal is a metric, which places a value on each shot taken, that is based on past averages. Here, shot location, direction, placement of defenders, shot type any much more is considered. The expected goals differed to the actual outcome in 33% of the cases, corroborating previous findings of football randomness. In this study, the difference in the two teams’ expected goals as a metric for assessing strategies was utilized. This also allowed for splitting the game into smaller parts without having to fear a lack of results.
7.7 Future work

In this study, the focus was on football due to its complexity and the availability of datasets. However, the approach of this study can be extended and applied to the analysis of any system that is the result of interactions between entities. The features selected for clustering will need to be customized individually based on the activity. The success metric will also need to be decided based on the activity; even the use of multiple success metrics could be beneficial in some scenarios.

In the realm of football, network analysis is growing in popularity. One of the promising areas of network analysis, for which there is a research gap, is the analysis of how networks impact and interact with each other. How can the network of one team disrupt or unsettle the opponent’s network? Such an analysis would allow the analyst to identify passing networks that are firstly, static, meaning that the network remains constant regardless of opponent, and secondly, dynamic, i.e. those that dynamically shift, and change based on the opposing network.

When using network analysis as a tool to discover, identify and label strategies, it would be beneficial to work with larger datasets. Additionally, performing a similar study on a different league and comparing the results would yield interesting results, in other words, identifying whether different leagues have different playstyles, and how they differ. Also, an interesting area for further research would be to predict changes in networks to preemptively accommodate to those changes.

8 Strategirekommendationer för pågående fotbollsmatcher. En studie med utgångspunkt i nätverksteori.

Avancerad tillämpad dataanalys har av flera orsaker blivit en av hörnstenarna i en framgångsrik, modern idrottsorganisation. Tack vare förbättrade resultat ökar organisationens intäkter i och med att antalet tävlingar som lagen deltar i stiger, samtidigt som mängden sålda produkter och åskådaraantalet på matcherna troligen också ökar. (Ransbotham och Kiron 2018)


I början av avhandlingen diskuteras och förklaras olika nätverksmått och deras tillämpning på fotboll. Ett nätverks centralitetsmått mäter den relativa vikten hos varje nod i nätverket. Egenvektorcenralitet är ett mått på hur central noden är, Pagerank mäter bollens förmåga att landa vid en viss nod efter n antal passningar, och

I och med att de olika strategierna identifierades, blev det möjligt att analysera de strategiska förändringarna i spelet och därmed att identifiera gynnsamma åtgärder baserade på ett givet scenario (Tabell 7 och 8). Fotbollstränare eller -managers kan dra nytta av denna information genom att träna en viss spelstil inför en kommande match eller genom att identifiera svagheter i motståndarens strategi. De kan också använda informationen som stöd för strategiskt beslutsfattande under en match.

Denna studie är den första som föreslår verktyg och metoder som kan användas under matchens lopp. Ett annat nytt resultat är att jag föreslår ett nytt mått för framgång i fotboll. Ett av problemen med att studera fotboll är nämligen idrottsgrenens slumpmässiga karaktär; det bättre laget vinner inte alltid och p.g.a. den stora variationen blir spelet svårt att analysera. För att kringgå detta problem föreslår jag i min avhandling att man analyserar de förväntade målen i stället för det reella spelutfallet. "Förväntade mål" (Bilaga A) är ett mått som på varje skjutet skott sätter

Slutligen fastställdes grunden för val av attribut i nätverksanalys för framtida maskininlärningsansatser (Tabell 3). Denna studie är den första som skiljer mellan spelarpositioner genom att skapa delsystem i hela lagnätverket. Attributen fångar separata nätverksvärden för olika spelarpositioner, identifierar spelstilar och strategier, där spelare utnyttjas till olika grad.

Studien skulle gynnas av ytterligare data. VM spelas dock bara vart fjärde år och därför kan förmågan att dra slutsatser utifrån en begränsad mängd data spela en stor roll inom sports analytics ständigt växande universum.

9 Reference List


Swarc, A. (2004). Effectiveness of Brazilian and German Teams and the teams defeated by them during the 17th FIFA World Cup. Kinesiology, 36(1), 83-89.


scoring in Norwegian professional soccer. Journal of Sports Sciences, 28(3), 237-
244.
Timmaraju, A. S., Palnitkar, A., & Khanna, V. (2013). Game ON! Predicting English
Premier League Match Outcomes.
Publishing Company.
Ulmer, B., Fernandez, M., & Peterson, M. (2013). Predicting Soccer Match Results in the
English Premier League (Doctoral dissertation, Doctoral dissertation, Ph. D.
dissertation, Stanford).
world’networks. nature, 393(6684), 440.
Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010, February). Twitterrank: finding topic-
sensitive influential twitterers. In Proceedings of the third ACM international
conference on Web search and data mining (pp. 261-270). ACM.
techniques. United State: Morgan Kauffman.
games. PloS one, 6(12), e29638.
in Sport, 6(1), 178–188.

10 Appendices

Appendix A.

Description: Expected goal visualized, in this scenario Kelechi Iheanacho takes a
shot at goal with an expected goal of 0.29. The expected goal is low due to defender
placement, being in close proximity to the Kelechi Iheanacho.
Leicester City (home)

Kelechi Iheanacho

Match Date: 2018-03-03, Shot Time: 72:37

Bournemouth (away)

Asmir Begovic

Leicester City

Kelechi Iheanacho
3 Ben Chilwell
5 Wes Morgan
9 Jamie Vardy
13 Harry Maguire
22 Matty James
25 Wilfred Ndidi
26 Riyad Mahrez

Bournemouth

Asmir Begovic 27
Simon Francis 2
Steve Cook 3
Dan Gosling 4
Nathan Ake 5
Charlie Daniels 11
Callum Wilson 13
Adam Smith 15
Lewis Cook 16
Joshua King 17
Junior Stanislas 19

Shot Result | Penalty? | Body Part | Buildup | Assisted? | Play Type | xG
--- | --- | --- | --- | --- | --- | ---
Miss | | Right Foot | | From Corner (Left) | | 0.29

Goal Keeper Action | Body Part | Technique | Position | Outcome
--- | --- | --- | --- | ---
Shot Faced | Other | | Moving |
Appendix B.

Description: the columns and the description of each field can be seen in its entirety here. (StatsBomb, homepage, 2019)

<table>
<thead>
<tr>
<th>Column</th>
<th>Type</th>
<th>Child Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>Uuid</td>
<td></td>
<td>The unique identifier for each event</td>
</tr>
<tr>
<td>Index</td>
<td>Integer</td>
<td></td>
<td>Sequence notation for the ordering of events</td>
</tr>
<tr>
<td>Period</td>
<td>Integer</td>
<td></td>
<td>The part of the match the timestamp relates to</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1 = first half, 2 = second half)</td>
</tr>
<tr>
<td>Timestamp</td>
<td>Timestamp</td>
<td></td>
<td>The point in the match the event takes place.</td>
</tr>
<tr>
<td>Minute</td>
<td>Integer</td>
<td></td>
<td>The minute part of the timestamp</td>
</tr>
<tr>
<td>Second</td>
<td>Integer</td>
<td></td>
<td>The second part of the timestamp</td>
</tr>
<tr>
<td>Name</td>
<td>Text</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Possession</td>
<td>Integer</td>
<td>Each possession is given a unique integer within the scope of the match. Events in the same possession have the same identifier.</td>
<td></td>
</tr>
<tr>
<td>Possession_team</td>
<td>Object</td>
<td>Id</td>
<td>Integer</td>
</tr>
<tr>
<td></td>
<td>Name</td>
<td>Text</td>
<td></td>
</tr>
<tr>
<td>Play_pattern</td>
<td>Object</td>
<td>Id</td>
<td>Integer</td>
</tr>
<tr>
<td></td>
<td>Name</td>
<td>Text</td>
<td></td>
</tr>
<tr>
<td>Team</td>
<td>Object</td>
<td>Id</td>
<td>Integer</td>
</tr>
<tr>
<td>Name</td>
<td>Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>The name of the team this event relates to</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Player</td>
<td>Object</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Id</td>
<td>Integer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>The id of the player this event relates to. Player object will only display if the event is tied to a specific player</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>Array[x,y]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>Array containing two integer values. These are the x and y coordinates of the event. This only displays if the event has pitch coordinates.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration</td>
<td>Decimal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration</td>
<td>If relevant, the length in seconds the event lasted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under_pressure</td>
<td>Boolean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under_pressure</td>
<td>If an event was done whilst pressure was</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Related_events</td>
<td>Array[uuid,uuid,uuid …]</td>
<td>A comma separated list of the Ids of related events. For example, a shot might be related to the Goalkeeper event, and a Block Event. The corresponding events will have the Id of the shot in their related_events column.</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>---------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>[event_type_name]</td>
<td>Object</td>
<td>For some event types, additional details are added with additional details specific to that event type. e.g. for shot events a shot object is added, containing details about the shot (shot_type, body_part used etc.) The contents of these types are detailed below.</td>
<td></td>
</tr>
<tr>
<td>Tactics</td>
<td>Object</td>
<td>Formatio n</td>
<td>Text</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>------------</td>
<td>------</td>
</tr>
<tr>
<td>For events with a set of match positions relevant (starting XI, tactical shift), the “tactics” object is added. The formation item describes the formation being used.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lineup</td>
<td>array</td>
<td>Collection of Player Position Objects (detailed later).</td>
<td></td>
</tr>
</tbody>
</table>