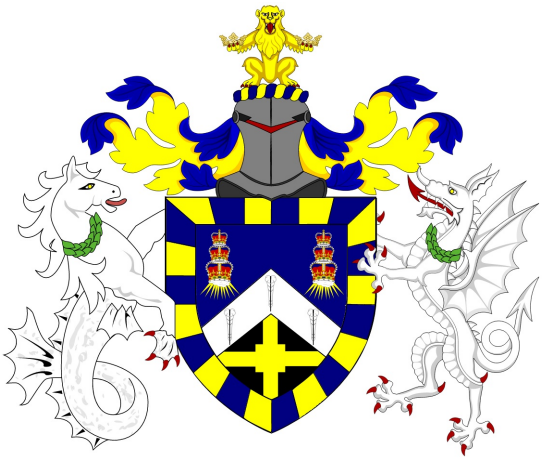
Data Analytics MSc Dissertation MTHM038, 2023/24

Network Analysis in Football:

A study of passing networks in the build up to a goal

Ellen Marie Leahy, ID 220865928

Supervisor: Dr. Hong Qi



A thesis presented for the degree of

Master of Science in *Data Analytics*

School of Mathematical Sciences

Queen Mary University of London

Declaration of original work

This declaration is made on September 2, 2024.

Student's Declaration: I Ellen Leahy hereby declare that the work in this thesis is my original work. I have not copied from any other students' work, work of mine submitted elsewhere, or from any other sources except where due reference or acknowledgement is made explicitly in the text. Furthermore, no part of this dissertation has been written for me by another person, by generative artificial intelligence (AI), or by AI-assisted technologies.

Referenced text has been flagged by:

1. Using italic fonts, **and**
2. using quotation marks "...", **and**
3. explicitly mentioning the source in the text.

Acknowledgements

I would like to thank my supervisor, Dr. Hong Qi, for her ongoing support of this project. I appreciate the time taken to review my work over the summer, and I hope I have written a report that she can be proud of.

I would also like to thank the many lecturers I had over the course of my studies, I learnt so much and I truly enjoyed combining two of my passions: mathematics and football.

Finally, I would like to thank my boyfriend. Thank you for always being there for me after long days in the library, I look forward to having more time to spend with you again.

Abstract

Data analysis has become more and more important in the world of football in recent decades. In particular, network analysis has allowed us to better understand how different teams play and score goals. In this report, I will analyse the goals scored by 98 teams across the top leagues in Europe in the 2017/18 season. For each team, I created two networks, a passing network between regions of the pitch and another between players, where I have included all passes in the phase of play directly before a goal was scored. I used several centrality metrics, including degree, betweenness and pagerank, to understand how the teams passed the ball. I then used UMAP to reduce the number of dimensions to two, such that classification analysis could be done. By analysing the centrality scores, I found a clear difference between the teams that finished top of the league and those that finished bottom. Furthermore, the results of classification analysis, using the SVM algorithm, validated that teams that finished in the top half of the league tend to have similar networks. Conversely, we see no correlation between league and play style.

Contents

1 Introduction **6**

1.1 Introduction to Football 6

1.2 Motivation for this work 7

1.3 Report Structure 9

2 Methods **10**

2.1 Procedures 10

2.2 Network Procedures 12

2.3 Network Analysis 13

2.3.1 Degree 14

2.3.2 Clustering Coefficient 14

2.3.3 Betweenness 15

2.3.4 Pagerank 15

2.3.5 Maximum Clique 15

2.4 Dimension Reduction 16

2.4.1 PCA 16

2.4.2 t-SNE 17

2.4.3 UMAP 18

2.5 Classification 19

2.5.1 K Nearest Neighbour 20

2.5.2 Support Vector Machines 20

3 Results **22**
3.1 Network Analysis 22
3.1.1 Network of Regions 22
3.1.2 Network of Players 24
3.2 Dimension Reduction 28
3.3 Classification 29

4 Discussion **33**

5 Conclusions **43**

Chapter 1

Introduction

1.1 Introduction to Football

Football is the world's most popular sport, with 3.5 billion global fans across all continents[1], and a combined total of 26 billion viewers watching the world cup in 2010 alone[2]. While football has been played in some form since before medieval times, the modern game as we know it originated in England in the 19^{th} century, when the official rules were first codified by the Football Association (FA), who, to this day, define the match rules in Britain[2].

The success of football may be routed in its simplicity; the only equipment needed is a ball, two goals and a pitch. Each team has 11 players, 10 of which are outfield players and are not allowed to touch the ball with their hands, while the goalkeeper can handle the ball when in the penalty box. The aim of the game is simply to get the ball into the opponents goal, and the team with the most goals after 90 minutes is the winner.

Football competitions can either be in league or cup format. Cup competitions are knock-out, where teams are randomly drawn against each other and only the winners progress to the next round. In a league competition, each team play every other team twice, home and away, and gain one point for a draw or three points for a win. The overall winner is the team with the most points at the end of the season. It is matches played in this latter format that I will be focused on in this report.

Each country has several leagues, with the winners of each being promoted into the one

above and the teams that come near the bottom being relegated down a tier. In this report, I will look at the top divisions across five European countries: England, France, Italy, Spain and Germany. Each league consists of 20 teams, except for the German league which has only 18, and thus I will be analysing a total of 98 teams.

1.2 Motivation for this work

Over the past few decades, data and analytics have become an intricate part of football[3][4]. Many companies have been established in this field, for example Rezzil (www.rezzil.com), who use data to generate drills and training games for footballers and other athletes, and PlaymakerAI (www.playmaker.ai), who use artificial intelligence to generate insights from football data to be shared with anyone from players to the media. Major football teams are also seeing the importance of statistics; in 2012, English giant Arsenal acquired football analytics company StatsDNA for over £2million, a company which uses statistics to scout new players[5].

In club football, players can be sold from one team to another. The price of footballers has risen dramatically in recent years; in 2011, the average price of a player in the English Premier League was £3.7m, in the summer of 2019 this number had increased to £14m[6]. With players in their peak costing clubs in excess of £100m, the importance of scouting players young cannot be ignored, and data can help us to do this.

Furthermore, it is important to understand the various playing styles in different clubs and leagues. The English Premier League has long since been heralded as the best league in the world, and players who flourish in the less competitive French League Un or German Bundesliga league are often believed to fall short when arriving in England. Data can be used, not just to measure how well a player is performing, but to understand how they contribute to a particular set up, and whether this performance level can be repeated in another team or league.

Betting is a huge part of football. From April 2022 to March 2023, the football betting industry alone saw a turnover of £1.58b[7]. In 2016, Leicester City Football Club topped the Premier League after almost being relegated the previous season, with only a 5000 to

1 chance of winning at the start of the season[8]. In the modern day, viewers can bet on anything, from the number of corners in a game to which players a club might buy, and as these viewers become more savvy to statistics, the betting industry must also become experts to ensure that they are still making a profit.

Network analysis is a key area of data analytics in football. Understanding the movement of the ball throughout the match, and the players positioning on and off the ball, can be a key indicator as to whether a goal is about to be scored or which team is more likely to win the match[9][10][11]. In the 2015/16 season, FIFA updated the rules to allow players to wear GPS trackers on the pitch, a tool which was previously only allowed on the training ground[12]. Since then, several papers have been published analysing the movement of players on the pitch.

While understandably, knowing the value of football in modern times, most of this data is kept within a club, in 2019, a dataset collected by Wyscout was published containing all temporal-spatial data across the top five leagues in Europe: the Premier League (England), La Liga (Spain), Bundesliga (Germany), Seria A (Italy) and Ligue Un (France)[13]. It is using this dataset that I will analyse the passing networks of teams in the build up to scoring a goal. My aim is to understand how these passing networks vary between teams competing in different leagues as well as comparing teams that finish top of their league versus those that finished bottom. I will investigate the centrality metrics as suggested by Clemente[14] and Lopez[15], and expand on their investigations to look at all teams across several leagues. I will also use dimension reduction techniques to consolidate the metrics into two dimensions, such that we can visualise the data and run classification analysis. This will allow us to understand which teams score similar goals.

I believe this analysis is useful for several reasons. Firstly, in European competitions, teams play those in other leagues, and having an understanding of how their opponents score goals could help teams to better defend against them. Secondly, teams can use this data to scout players who fit their style of play, or players who would complement those they already have. This analysis could also be used in the betting market to predict who will win the match or where the goals might come from. Finally, it could also be used as an indicator to understand how a team might perform in the future.

1.3 Report Structure

In chapter 2, I will discuss the methodology used in this report. In sections 2.1 and 2.2, I will discuss how I defined and extracted the passes in the lead up to a goal, and how I subsequently defined adjacency matrices from these passes for both players and positions on the pitch. In section 2.3, I will discuss the different centrality measures I used to analyse the passing networks for each team. In order to perform classification analysis, I used dimension reduction techniques to reduce the number of features down to two, in section 2.4 I will discuss the different algorithms used. Finally, for this chapter, in section 2.5 I will discuss the classifications algorithms used on the data.

In chapter 3 I will go through the results of each section of the analysis, and in chapter 4 I will discuss my findings and any conclusions that we can take from these results. Finally, in chapter 5 I will summarise my analysis and provide suggestions for further research which could be done in this area.

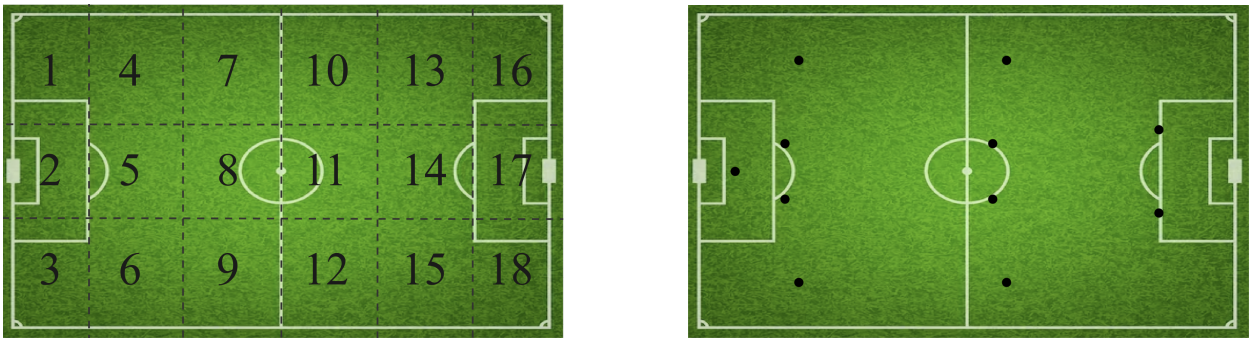
Chapter 2

Methods

2.1 Procedures

The dataset I will be analysing covers the 2017/18 season and was originally published in 2019[13]. The data is in JSON format and contains metadata on the teams, such as the league they play in, the players, for example what position they play, the matches, including which teams were playing, what were their starting line ups and who was substituted in, and the events. The events file includes every action that occurred during the game; such as passes, tackles, throw-ins, shots and saves, for this report I was only interested in the passes in the build up to a goal. To isolate these, I looped through each event until I found a goal, I then moved backwards through the events until I found an event which was not a pass or involved the team that did not score, for example a tackle or throw in. The goal events were then added to a new dictionary with each pass included in a sub-dictionary, along with all necessary metadata, such as the position on the pitch and which player was involved. As I was interested in the passing networks, own-goals and goals which were scored directly from a set-piece, such as a penalty or free-kick, were not included. I also removed any goal which had only one pass in the build up.

In order to validate that I had read in the data correctly, I calculated the total number of goals and goals per team in my dictionaries, prior to removing those with fewer than two passes. I then compared this to data on how many goals were scored by each team, which



(a) Regions. (b) Players.

Figure 2.1: Diagrams showing the regions and player mappings on the pitch.

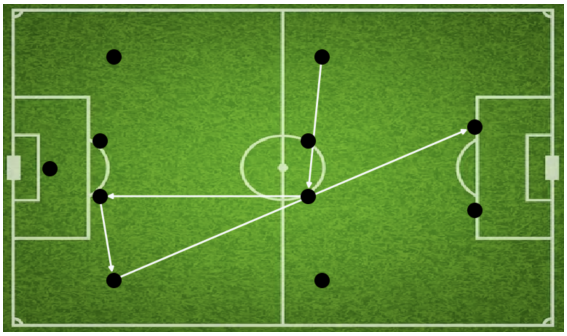
gave a breakdown of goals scored from open play, set-pieces and own-goals[16]. From this, I was able to validate that I had the correct number of goals for each team.

In this report, I will consider two different networks: the passing network between regions of the pitch and that between players. As the size of a football pitch can vary, the data contains the percentage distance the ball is from the opponents goal, with 0% along both the x and y axis being the top corner of the pitch furthest away from the goal that is being attacked, known as the opposition goal, and 100% along both axis being the opposite corner. Figure 2.1a shows the 18 sections of the pitch, as initially suggested by the Handbook of Soccer Match Analytics[17] and used in the paper by Clemente et al[14].

For the passing network between players, we only consider which player made the pass, and ignore which part of the pitch they were in at the time. For simplicity, I have chosen to visualise these networks using the classic 4-4-2 formation[18], where the team lines up with four defenders, four midfielders and two attackers in front of the goalkeeper, as shown in figure 2.1b. While this is a simplification and many teams will play different formations, such as 5-4-1 or 4-3-3, the exact position of the player is not important in this report, rather I will analyse whether the most important player for the team is further forward or in defense, and if the team has one player who is significantly more important than the others or if the distribution is more balanced. In order to assign the player a position index, I sorted the players by their position; goalkeeper, defenders, midfielders and then forwards, and numbered each player from 0 to 10, such that the goalkeeper will always be at index 0

0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

(a) Adjacency matrix of goal.



(b) Network of goal.

Figure 2.2: The adjacency matrix of a goal scored by Stoke City, and the corresponding network generated.

and the forward players will be at 9 and 10.

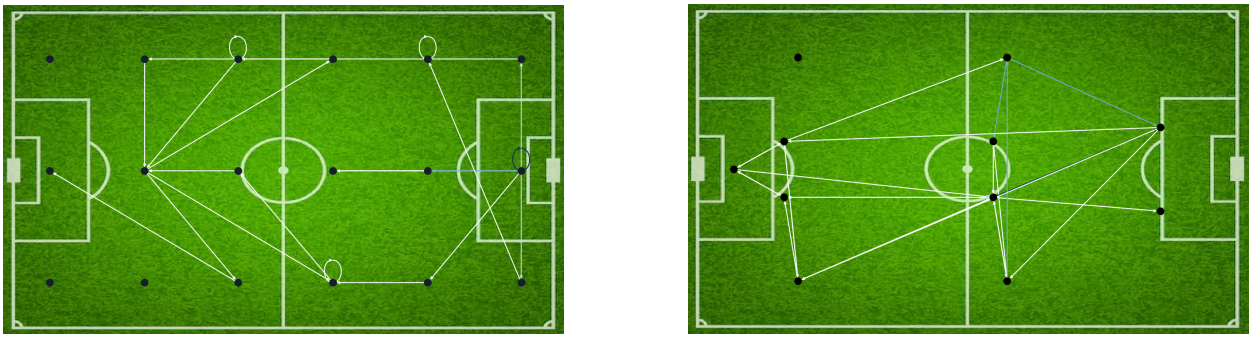
2.2 Network Procedures

For each attacking play, I defined an adjacency matrix, where each element is defined as follows:

$$A_{ij} = \begin{cases} X & \text{Number of passes from } j \text{ to } i \\ 0 & \text{If no passes were made from } j \text{ to } i. \end{cases} \tag{2.1}$$

This is a weighted and directed matrix, where the direction describes which player is making the pass and which is receiving, and the weight describes the number of passes between the two nodes[15]. As this is a directed network, we do not expect the adjacency matrices to be symmetrical. Figure 2.2 shows an example adjacency matrix for one of the goals and the corresponding network associated. A total adjacency matrix is then formed for each team by summing each element of the individual matrices together. The elements of this matrix are defined in the same way as for the individual attacking plays.

Therefore, each team we will have two final adjacency matrices; one for regions (18x18) and one for players (11x11). The regions matrix can have self loops, as a player can pass to another player within the same region, however, as a player cannot pass to themselves, the player matrix should not have any self loops. Figure 2.3 shows an example of the total



(a) Region network. (b) Player network.

Figure 2.3: The region and player networks for Stoke City Football Club.

region and player networks for Stoke City Football Club.

2.3 Network Analysis

In order to better understand each network and how they compare to those of other teams, I calculated a selection of centrality measures for each node of the network: degree, clustering coefficient, betweenness and pagerank, I also found the maximum clique of the network. I will discuss each in depth in sections 2.3.1 to 2.3.5. I then found the average value of each metric for each team.

For the players' passing network, as well as calculating the averages, I also found the the maximum centrality value in the team, in order to measure how much more central the most central player is compared to the average. I was also interested in understanding the distribution of centrality across the team. In order to calculate this, I defined a single metric for each player, the total centrality score, calculated by summing the normalised centrality measures. Normalisation is important here as some metrics, such as betweenness, are always between zero and one, while others, such as in and out-degree, are integers. We can then use this new metric to compare players within a team to find the overall most central player and to understand how their position compares to the position of central players in other teams. We can also look at the distribution of total centrality score across all players in the team, to find if teams have one or two very central players, or if all players are equally central.

2.3.1 Degree

The in-degree and out-degree of a node measures the number of times that region or player made or received a pass, and the total-degree is simply defined as the sum of these values[19]. This is a useful metric to understand how important a player or region was in the build up to a goal. The in-degree of the i^{th} node is defined as the sum of all elements of the i^{th} column of the adjacency matrix. Similarly, the out-degree is the sum of the i^{th} row of the adjacency matrix. The average degree values of the network are defined as the median of the node values, this is because the mean in-degree and out-degree will be equal and therefore do not give us any additional insight.

2.3.2 Clustering Coefficient

The clustering coefficient describes how well connected a node is to other nodes. This metric measures the transitivity of a network by finding the fraction of all triangles which contain node i that exist in the network[19]. For a weighted and directed graph, this can be defined as

$$c_i = \begin{cases} \frac{1}{u_i(u_i-1)} \sum_{j,k} \frac{(A_{ij}A_{kj}A_{ki})^{\frac{1}{3}}}{\max(A)} & \text{if } u_i > 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

where a_{ij} is the element in the i^{th} row and j^{th} column in adjacency matrix A[15]. The adjacency matrix must first be normalised such that all weights are between zero and one. The average clustering coefficient of the network is defined as

$$C =< c_i >= \frac{1}{N} \sum_{i=0}^{N} c_i \tag{2.3}$$

wherer N is the total number of nodes in the network. A player or region with a high clustering coefficient is well connected to other nodes. Similarly, a network with a high average clustering coefficient is one that has many nodes that are well connected. In the context of my analysis, this means that more players or regions of the pitch will have been involved in the lead up to a goal.

2.3.3 Betweenness

Betweenness centrality measures the extent to which node i is between any other nodes j and k. Specifically, a node is described as being more central if it is more involved in the movement between any other pair of nodes[19]. We define the normalised betweenness as

$$C_B(i) = \frac{1}{(N-1)(N-2)} \sum_{j \neq k \neq i} \frac{n^i_{jk}}{g_{jk}} \tag{2.4}$$

where g_{jk} is the number of geodesic paths between j and k and n^i_{jk} is the number of these paths that go through i[15]. The total graph betweenness is defined as the mean across all nodes. A player or region with a high betweenness value will be situated such that many of the plays go through them, and they are therefore of high importance to the team.

2.3.4 Pagerank

Pagerank, an algorithm designed by Google to rank search results, is a measure of popularity that assumes a node is more popular if it is connected to other popular nodes[19]. It is an approximate probability that a player or region of the pitch will have the ball after a reasonable number of passes. In this report, I will use NetworkX to calculate the pagerank of each node[20]. This algorithm takes hyperparameter α, the damping factor, this is the probability that the ball will continue moving along a path after taking a step. The value of the damping factor used for this analysis was the inverse of the largest eigenvalue of the network's adjacency matrix[21].

2.3.5 Maximum Clique

For an undirected graph, a clique is a group of nodes which create a complete graph, such that every vertex that could exist does exist[19]. In the case of a directed graph, we can update this definition such that each node in the clique must be connected by reciprocal edges. In order to calculate this, I created a copy of each network where only reciprocated edges exist, and then used the NetworkX function `find_cliques`[20] to find all the cliques

in the network. I then collected all cliques with size equal to the largest clique size.

2.4 Dimension Reduction

For the networks of regions, I calculated seven centrality metrics (mean total degree, median in-degree, median out-degree, mean clustering coefficient, mean betweenness, mean pagerank and maximum clique), and for the networks of players, I calculated 13 (the previous seven plus maximum in-degree, maximum out-degree, maximum total degree, maximum clustering coefficient, maximum betweenness and maximum pagerank). In order to run a classification algorithm and visually analyse the data, these can be reduced to only two dimensions using a dimension reduction algorithm.

There are many different algorithms for dimension reduction, in this report I will investigate three: Principle Component Analysis (PCA), T-Distributed Stochastic Neighbor Embedding (t-SNE) and Universal Manifold Approximation and Projection (UMAP), which I will discuss in depth in the next sections. I will choose the technique which leads to the least amount of information loss, while also creating well clustered date. The trustworthiness function from the Scikit-learn library[22] can be used to measure how similar the new dataset is to the original, larger dataset. This function looks at the nearest neighbours of each node in the output space, and penalises any unexpected nearest neighbours based on which points are closest in the input space, in proportion to their rank in the input space. The value of trustworthiness is between zero and one, with one reflecting an output array that is identical in similarity to the input array.

2.4.1 PCA

PCA is a linear dimension reduction algorithm which uses eigenvectors to find the directions of maximum variation in the data and projects the data onto a new component system. There are five main steps of PCA[23]:

1. The data is standardized.

2. The covariance matrix is computed to quantify the relationship between each of the features.

3. The eigenvectors and eigenvalues of the covariance matrix are computed. The principle components are then found by sorting the eigenvalues in descending order, such that the first principle component is the eigenvector corresponding to the largest eigenvalue, the second corresponds to the second largest eigenvalue etc.

4. The feature vector is then defined such that each column is a principle component. The number of principle components included corresponds to the number of dimensions in our final dataset, such that if the goal is to have two dimensional data, the feature vector will have two columns.

5. The data is then recast along the principle component axis, this is done by simply multiplying the transpose of the feature vector by the transpose of the standardised data. If required, the data should then be unstandardized.

For this analysis, I used the PCA algorithm from the Scikit-learn library[22]. The main benefits of PCA are that it is easy to implement, easy to understand and effective at removing noise from the data[24]. The main drawback of PCA is that there is some information loss, and while this is true for all dimension reduction algorithms, PCA is particularly affected as it is a linear algorithm and this can lead to the data being over-simplified and nuance lost.

2.4.2 t-SNE

t-SNE is a non-linear dimension reduction algorithm, originally based on Stochastic Neighbor Embedding and updated to a t-distributed variant proposed by Laurens van der Maaten[25]. There are two main steps of t-SNE:

1. A probability distribution is constructed over pairs of high-dimensional objects, such that objects which are similar have a higher probability and those which are dissimilar have lower probabilities.

2. A similar probability distribution is then constructed over the points in the low-dimensional map, such that the Killback-Leibler divergence[26] (KL divergence) is minimised with respect to the position of the points between the two distributions.

The lower dimensional visualisation of the points is impacted by the parameters provided to the function. For this analysis, I used the KL divergence to finetune the perplexity, which defines the number of neighbours each point considers while running dimension reduction. I used the t-SNE algorithm in the Scikit-learn library[22].

An advantage of t-SNE over PCA is that it is non-linear, therefore if there are any polynomial relationships between features, t-SNE will out perform PCA. Furthermore, t-SNE is better at preserving relationships between data points. While PCA finds new dimensions to explain the variance of the data, t-SNE will investigate which points are close together in high-dimensional space and preserve these relationships in low-dimensional space.

The main drawback of t-SNE is that it is computationally expensive, as the algorithm analyses the pair-wise relationship for each data point. As this study has a relatively small amount of data, this is not an issue, but should be considered if this analysis were to be expanded in the future. Another drawback is that the hyper-parameter perplexity must be fine-tuned, adding a level of complexity. The algorithm is also non-deterministic, leading to a different result being returned following each run. Finally, patterns can be found in random noise, this is particularly relevant for this use-case as the data volume is low[27].

2.4.3 UMAP

UMAP is a nonlinear dimension reduction algorithm, introduced in 2018[28]. UMAP is based in manifold theory and topological data analysis and has three main steps:

1. A high-dimensional graph of the data is constructed where each node is a point and edges are drawn between similar points.

2. A fuzzy-simplicial set of the data is constructed, which quantifies the relationships between a subset of points, allowing for uncertainty. UMAP then attempts to find a lower dimensional representation of this fuzzy-simplicial set, preserving the relationships between points by minimising a cost function.

3. The low-dimensional representation is optimised using stochastic gradient descent (SGD); similar points in high-dimensional space are attracted to each other in low-dimensional space, while dissimilar points are repulsed.

To run this algorithm, I used the UMAP Python library[29]. As with t-SNE, UMAP has hyperparameters that need to be fine tuned: the number of nearest neighbours used and and the minimum effective distance between points, defined such that a smaller value will result in a more clustered embedding. To find the optimal value of these parameters, I ran the algorithm for a variety of values of nearest neighbours and minimum distance and chose the values which lead to the maximum trustworthiness score.

UMAP has many benefits; similar to t-SNE, it is non-linear and can therefore preserve polynomial relationships between features. Furthermore, compared to t-SNE, it is typically better at preserving global relationships between data points. It is also a much more scalable and efficient solution than t-SNE. UMAP has several drawbacks too however. Firstly, it is also a non-deterministic algorithm, leading to different results on each run and it is sensitive to the choice of hyper-parameters, emphasising the importance in choosing these. It also does not have an inverse function that can be used to map points back to the high-dimension space.

2.5 Classification

By plotting the results of dimension reduction, we see that we get data that is a good candidate for classification analysis, discussed futher in section 3.2. I tested two algorithms: k nearest neighbours (K-NN) and support vector machine (SVM), discussed in the next two sections. Both algorithms are supervised and require labelled data; they train on a sample of data and then use the resulting algorithm to classify a test dataset. We can use accuracy, precision and recall to measure how well the algorithm is performing. Accuracy is simply the fraction of test data points which were correctly classified. Precision is the fraction of total classified elements which were correctly classified, and recall is the fraction of elements that were in a particular group that were classified as such.

2.5.1 K Nearest Neighbour

K-NN is a supervised classification algorithm which uses adjacent data points to classify a new data point[30]. There are five main steps to K-NN:

1. Select the number of neighbours the algorithm should use, K.
2. For each data point, calculate the euclidean distance to all other data points.
3. Take the K points closest to our new point.
4. Count the number of data points in each category amongst the K neighbours.
5. Assign the new data points to the most popular category.
6. Repeat for each new data point.

To choose the value of K, we can run the algorithm for a range of values and select the option with the maximum accuracy score. If several values of K provide the best accuracy, we should use the smallest value, as this will ensure we are not using data points which are too far away. For this analysis, I used the k-NN algorithm in the Scikit-learn library[22].

The advantages of k-NN are that is is easy to understand and implement. It is also a lazy algorithm, which means that is does not build a model or make any generalisations, it stores the dataset and performs a new computation for any new data point that is added. However, this means that it is computationally heavy; for each data point, the distance to all other points needs to be calculated to find the nearest neighbours, this can cause issues when the data volume increases. Furthermore, the results depend on our choice of K and the distance metric used[31].

2.5.2 Support Vector Machines

SVM is a supervised machine learning algorithm which classifies datasets into groups by defining a line or hyperplane which separates the various categories[32]. There are seven steps in the SVM algorithm:

1. A training set is provided, including feature vectors, x_i, and corresponding class labels, y_i, which are either -1 or 1.

2. A hyperplane is defined and can be written as

$$y_i(w \cdot x_i + b) \begin{cases} = 1 & \text{where } x_i \text{ belongs to the closest support vector} \\ \geq 1 & \text{where } x_i \text{ belongs to another vector.} \end{cases} \tag{2.5}$$

3. The perpendicular distance between the hyperplane and the point closest to it is maximised, given by $\frac{2}{||w||}$.

4. The Lagrangian is formed, such that

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum \alpha_i [y_i(w \cdot x_i + b) - 1] \tag{2.6}$$

   where α_i are the Lagrange multipliers.

5. The saddle points of the Lagrangian are calculated by computing the gradients with respect to w and b and setting equal to 0.

6. The gradients are substituted back into the Lagrangian to obtain the problem:

$$\begin{aligned} \text{Maximize: } W(\alpha) = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\ \text{Subject to: } \sum (\alpha_i y_i) = 0 \text{ and } \alpha_i \geq 0. \end{aligned} \tag{2.7}$$

7. This quadratic problem can then be solved using optimization techniques, for example gradient ascent.

For this report, I used the more simple linear SVM, from the Scikit-learn library[22], as the data is well separated and a quadratic algorithm is not needed.

SVM works well for data that is well separated, which is the case for our dataset. It is also much more memory efficient than K-NN. SVM works better on smaller datasets, however, it does not do well if there is a greater number of features than training points[33].

Chapter 3

Results

3.1 Network Analysis

The entire results set for the network analysis can be found on my GitHub repository (https://github.com/ellenleahy95/dissertation/tree/main/Data/output/centralities/excel). In this section, I will focus on only the teams that finished top or bottom of the league.

3.1.1 Network of Regions

Table 3.1 shows the average centrality measures of regions of the pitch for the teams which finished first in the league and those that finished last. The average total degree and median in and out-degrees are both significantly larger for the teams finishing top of the league. The French team, Paris Saint-Germain (PSG), in particular have high degree scores, and interestingly, Metz, the team who finished bottom of the French league, have a higher value of total degree than many of the other teams that finished bottom of their league. This might suggest that the French league have more passing plays in the lead up to a goal, using more areas of the pitch. For those teams finishing bottom, the median in-degree and out-degree scores are either zero or one, suggesting that the majority of regions are not involved in the build up to a goal. This is likely due to their defenders becoming less involved in the

Team	League	Total Degree	In-Degree	Out-Degree	Clustering Coefficient	Betweenness	Pagerank	Max. Clique
Man. City	EPL	20.44	5.0	7.0	0.02	0.10	0.055	3
Barcelona	La Liga	16.44	4.5	7.0	0.02	0.19	0.055	3
B. Munich	Bundesliga	15.67	5.0	6.0	0.04	0.12	0.055	3
Juventus	Serie A	12.78	4.0	5.5	0.05	0.13	0.053	3
PSG	Ligue Un	23.56	8.5	10.5	0.04	0.17	0.055	3
WBA	EPL	2.22	0.0	0.0	0.03	0.01	0.052	1
Málaga	La Liga	0.89	0.0	1.0	0.02	0.01	0.052	1
Kóln	Bundesliga	4.11	0.5	1.0	0.02	0.06	0.051	2
Benevento	Serie A	1.78	0.0	0.0	0.02	0.01	0.047	1
Metz	Ligue Un	4.11	0.0	1.0	0.02	0.03	0.054	1

Table 3.1: Table showing the average centrality measure for the regions of the pitch for the teams that finished top and bottom of their respective leagues.

attacking play, I will discuss this in more detail in chapter 4.

All teams who finished top of the league had a maximum clique of three, meaning that they had three areas of the pitch that had passes in both directions in the lead up to a goal. No other team had a larger clique than these teams. For the teams who finished bottom of the league, only Köln had a maximum clique more than one, the minimum possible value. This suggests that these teams did not move the ball between regions in the lead up to a goal and, instead, were more direct in their movement.

Teams that finished top of the league also had significantly higher average betweenness compared to those that finished bottom. As betweenness measures how important a region of the pitch is when passing between other regions of the pitch, this aligns with other metrics suggesting that the weaker teams are more direct in their attacking play, whereas the top teams are able to pass the ball around more freely.

Conversely, there is not a significant difference between pagerank and clustering coefficients between these sets of teams. Pagerank measures the probability that a region will get the ball given the passes that have occurred. This suggests that neither set of teams play in a predictable way and that there is an element of randomness. The clustering coefficient measures how well connected the nodes in a network are by counting the number of triangles.

Team	League	Total Degree	In-Degree	Out-Degree	Clustering Coefficient	Betweenness	Pagerank	Max. Clique
Man. City	EPL	20.33	16	19	0.05	0.05	0.091	7
Barcelona	La Liga	16.44	12	14	0.06	0.07	0.091	5
B. Munich	Bundesliga	15.56	12	15	0.12	0.05	0.092	5
Juventus	Serie A	12.78	12	11	0.03	0.05	0.091	5
PSG	Ligue Un	23.44	24	20	0.03	0.05	0.091	7
WBA	EPL	2.22	1	2	0.09	0.04	0.090	2
Málaga	La Liga	0.89	0	0	0.03	0.01	0.054	1
Köln	Bundesliga	4.11	4	2	0.19	0.06	0.089	2
Benevento	Serie A	1.78	1	1	0.01	0.03	0.085	2
Metz	Ligue Un	4.11	3	3	0.15	0.05	0.090	2

Table 3.2: Table showing the average centrality measure for the networks of players for the teams that finished top and bottom of their respective leagues.

When we consider that the maximum clique is only three, and each network has a total of 18 nodes, it is not surprising that even the best teams are not playing in a well connected network. This is likely due to the fact that the ball will most likely be in the opposition half in the build up to a goal, regions 10 to 18, and regions 1 to 9 are likely to be quite disconnected. In section 4, I will analyse which nodes are in each clique to validate this theory.

3.1.2 Network of Players

As discussed previously, for the network of players, I analysed both the average centrality measures, shown in table 3.2, and the maximum values, shown in table 3.3. As for the regions, the average total degree and median in and out-degrees are both significantly larger for those teams which finish in the top spot. Metz also outperforms the other bottom placed teams in these metrics, suggesting they likely have more possession and passing than other teams that finish last.

The maximum clique is much larger for the teams that finished first compared to those that finished bottom. In contrast to the regions network, the cliques cover a much higher proportion of the nodes, with PSG and Manchester City both having cliques of seven out of a

Team	League	Total Degree	In-Degree	Out-Degree	Clustering Coefficient	Betweenness	Pagerank
Man. City	EPL	54	31	26	0.08	0.13	0.093
Barcelona	La Liga	55	28	27	0.09	0.24	0.097
B. Munich	Bundesliga	40	23	17	0.51	0.14	0.094
Juventus	Serie A	35	19	17	0.17	0.14	0.099
PSG	Ligue Un	54	27	28	0.05	0.13	0.093
WBA	EPL	8	6	4	0.50	0.14	0.13
Málaga	La Liga	3	3	3	0.31	0.02	0.21
Köln	Bundesliga	15	7	9	0.62	0.16	0.11
Benevento	Serie A	8	6	5	0.06	0.09	0.12
Metz	Ligue Un	14	9	9	0.99	0.15	0.12

Table 3.3: Table showing the maximum centrality measure across the players for the teams that finished top and bottom of their respective leagues.

possible 11 players, while the other top teams have maximum cliques of five. Conversely, the bottom teams still have very small maximum cliques, with only two players, again suggesting that they are not moving the ball around and that fewer players are getting involved in the attacking play.

Similar to the networks of regions, there is no significant difference in pagerank between the teams that finished top versus those that bottom of the league. While Málaga does have a significantly lower value than other teams, their maximum value of pagerank is significantly higher. This suggests that Málaga have a player that is significantly more central than the rest of their team, as the probability of them receiving the ball is quite high, whereas for other teams, the distribution across all players is more even.

The average betweenness values of the teams that finished top of the league are all quite similar at 0.05, with only Barcelona having a higher value of 0.07. Köln and Metz have a similar value to these top teams, while only Málaga has a significantly lower value. When we look at the maximum value of betweenness in the teams, we see a similar pattern; Barcelona have a much higher value than any other team, suggesting that they have a player who many of the passes go through, often referred to as a play-maker. Again, Málaga have a very low maximum betweenness compared to the other teams, suggesting that they do not pass the

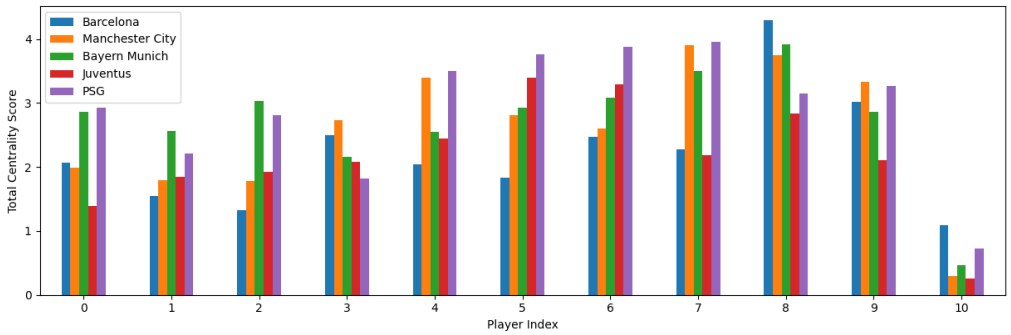
ball around in the same way and likely play much more direct football.

When analysing the average clustering coefficient, we do not see a significant difference between the two sets of teams, in fact the team with the highest average clustering coefficient is Köln, who finished bottom of the Bundesliga. Furthermore, when analysing the maximum value of clustering coefficient, the teams with the highest values are actually those that finished bottom of their league. This suggests that these lower ranked teams have a single player who is very well connected, whereas the top teams may have more players who are less well connected. This highlights the importance of individuals in poorer teams, and team-work in better teams.

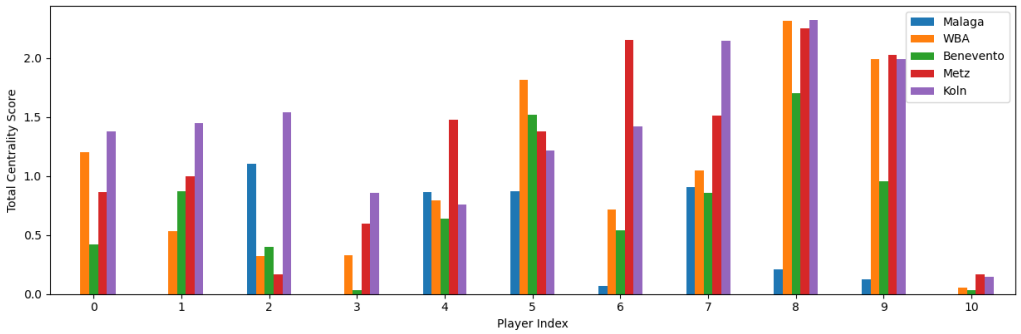
The graphs in figure 3.1 show the normalised total centrality scores for each player for the top (3.1a) and bottom (3.1b) teams. The total centrality scores for the teams that finished top of the league are significantly higher than those that finished bottom, with almost the entire team being quite central. Interestingly, the striker is the least central player for all teams. This is likely because the striker is more likely to shoot when they get the ball or be in a position such that they are not heavily involved in the build up play. It would be a simplification to say they are the least important player as they are likely the one to score the goals.

Furthermore, we can see that all teams, except for Málaga, have position 8, a midfielder, as their most central player. Málaga's most central player was in position 2, a defender, although this appears to be due to the majority of players on the team having very low centrality scores, rather than this defender being particularly involved. The majority of top teams have either position 7 or 8 as their most central figure, however for Juventas, their key player is in position 5.

Interestingly, we can see that Juventas not only have their most central player as being further back than the other teams, they also have a much more central midfield compared to the rest of the team, which we do not see in other clubs. It is likely that Juventas have more possession compared to other teams and their midfield players dictate much of these passing plays. PSG also have similarly important midfield players, although their forward players are more central than Juventas'.

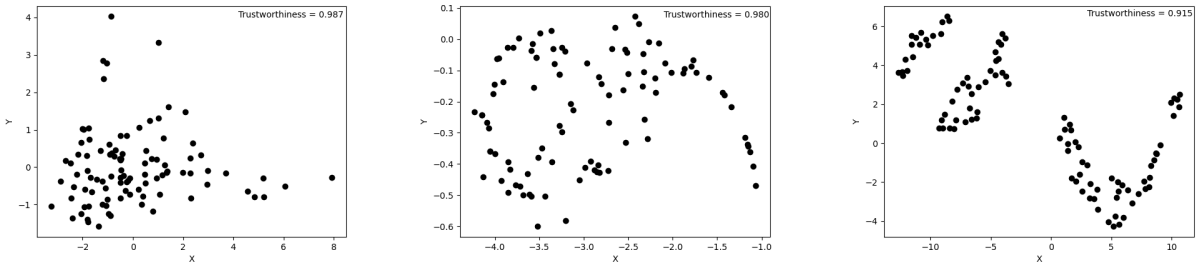


(a) Top teams.



(b) Bottom teams.

Figure 3.1: The overall centrality score for each player index for the top and bottom teams.



(a) PCA. (b) t-SNE. (c) UMAP.

Figure 3.2: The dimension reduction results for PCA, t-SNE and UMAP for the networks of regions.

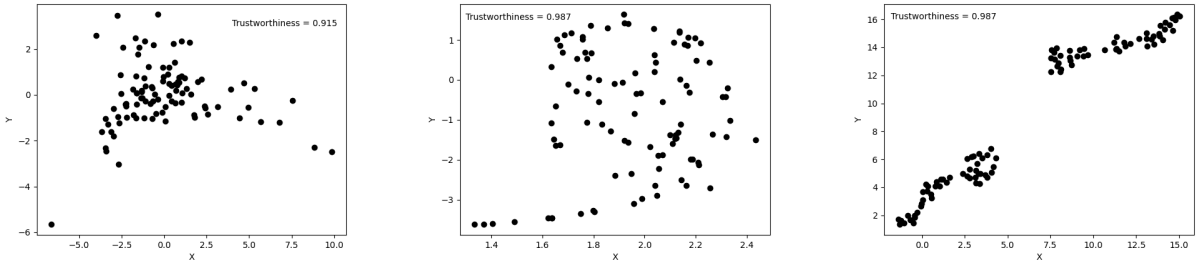
3.2 Dimension Reduction

Prior to running dimension reduction, I investigated which metrics to use and the best values for the hyperparameters. For the network of players, I investigated using only the average centrality measures, only the maximum centrality measures and all centrality measures. I found for all dimension reduction algorithms, using all measures resulted in more clearly defined clusters. For the networks of regions, I used the average metrics.

For both the networks of players and regions, the perplexity which gave the lowest value of KL divergence was 80. For the networks of regions, the values of nearest neighbours and minimum distance which gave the highest value of trustworthiness was 5 and 0.5 respectively, and for the network of players it was 10 and 0.25.

Figure 3.2, shows the results of dimension reduction for the networks of regions for each of the algorithms, along with the trustworthiness score. PCA (3.2a) has a significantly lower trustworthiness score than t-SNE and UMAP, this is likely due to the fact that this is a simple, linear algorithm and may not be able to include all of the complexities in the data. UMAP (3.2c) performs slightly better in terms of trustworthiness than t-SNE (3.2b) and shows much more defined clusters, it will therefore be used going forward.

Figure 3.3 shows the same results for the player networks. As before, PCA (3.3a) is the worst performing algorithm. To three decimal places, UMAP (3.3c) and t-SNE (3.3b) cannot be separated, in fact, to the fourth decimal place, t-SNE has a slightly higher value



(a) PCA. (b) t-SNE. (c) UMAP.

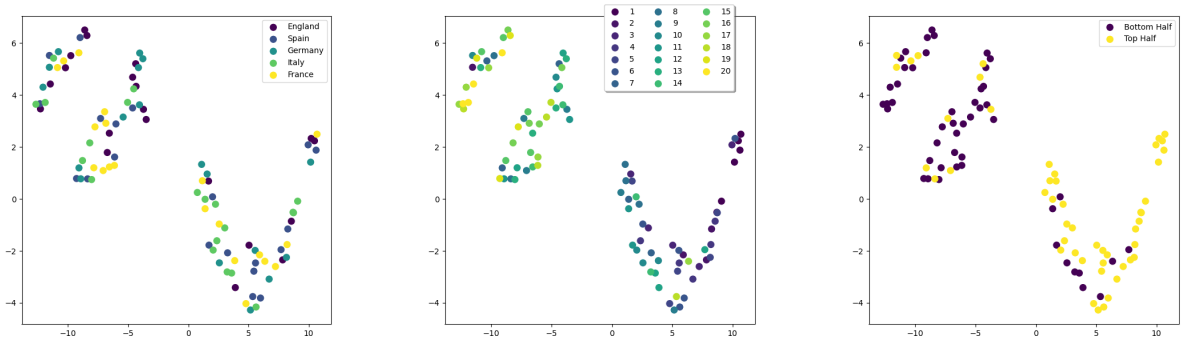
Figure 3.3: The dimension reduction results for the PCA, t-SNE and UMAP for centrality measures for the networks of players.

of trustworthiness. However, t-SNE does not show clear clusters, whereas UMAP does, thus we will move forward with UMAP for both network types.

3.3 Classification

As shown in figures 3.2c and 3.3c, the dimension reduction using UMAP returns data that is well clustered. Starting with the networks of regions on the pitch, we can see three well defined clusters, with two being closer together than the third. Prior to running a classification algorithm, I plot the data coloured by the potential labels: league, final league position and whether the team finished in the top half of the league or not. These charts can be seen in figure 3.4.

Figure 3.4a shows no clear relationship between the centrality measures and which league the team play in, whereas figure 3.4b does show some correlation between these metrics and final league position. The points in the cluster on the right tend to be darker, representing the teams which finished higher in the league. Figure 3.4c consolidates this, with the majority of points in the two clusters on the left belonging to teams that finished in the bottom half of the league, and the points in the cluster in the bottom right tend to be teams that finished in the top half. The graphs in figure 3.5 show similar results for the networks of players. There is no relationship between centrality metrics and league, however, we do see one cluster in the bottom left where the majority of teams finished in the top half of the league, and the



(a) Coloured by league. (b) Coloured by league position. (c) Coloured by finishing half.

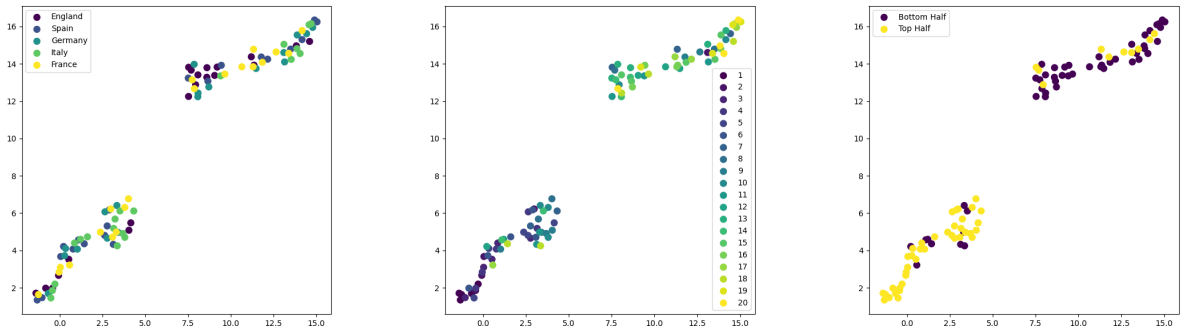
Figure 3.4: Graphs of networks of regions, following UMAP centrality measures, coloured by labels.

cluster in the top right contains mostly teams that finished in the bottom half of the league.

I ran K-NN and SVM on both datasets, using whether the team finished in the top or bottom half of the table as a label. I trained the model on 80% of the data, and used the final 20% for testing. I used accuracy, precision and recall to measure the performance of each algorithm.

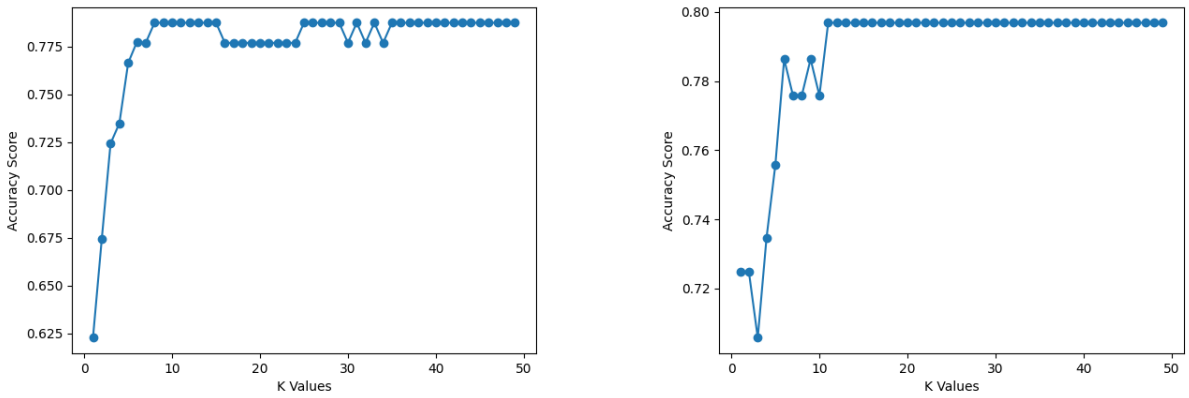
For K-NN, I first had to choose the value of K which maximises the accuracy. For both graphs, I ran the algorithm for values of K between 1 and 50 and plot the accuracy score calculated for each one, figure 3.6 shows the results of this. For the networks of regions, the smallest value of K which gives the best accuracy is eight, and for the networks of players it is 11. The accuracy score for the network of regions was 0.80, and the precision and recall were 0.83 and 0.76 respectively. For the network of players, the accuracy was 0.81, and the precision and recall were 0.83 and 0.81.

Figure 3.7 shows the results of SVM, including the accuracy, precision and recall. We can see that SVM outperforms K-NN, with precision scores of 0.95 and 0.90 for the networks of regions and players respectively. These results conclude that we can use the centrality metrics to determine whether a team will finish in the top or bottom half of a league.



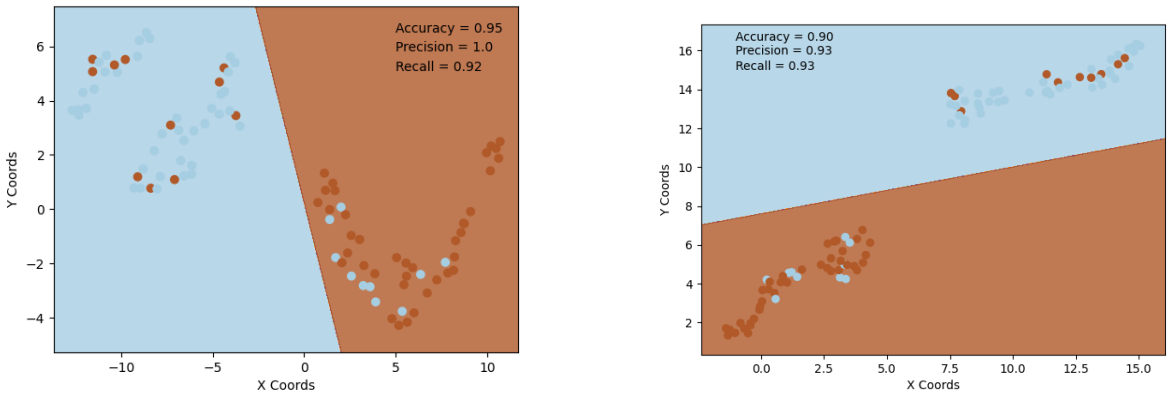
(a) Coloured by league. (b) Coloured by league position. (c) Coloured by finishing half.

Figure 3.5: Graphs of networks of players centrality measures, following UMAP, coloured by various labels.



(a) Regions. (b) Players.

Figure 3.6: Plots showing the accuracy for a range of K values for the networks of regions and players for the k-NN algorithm.



(a) Regions. (b) Players.

Figure 3.7: The results of SVM for the networks of regions and players.

Chapter 4

Discussion

Comparing centrality measures for the top and bottom teams shows significant differences between the two groups. Firstly, looking at degree measures, we see that the median out-degree tends to be one to two points higher than the median in-degree for the teams finishing top of the league. Conversely, the numbers are more similar for teams finishing at the bottom of the league. For the networks of players, the majority of top teams show a similar pattern, however PSG has a median in-degree of 24 and median out-degree of only 20. Köln also has a higher in-degree than out-degree.

To understand the reason behind these differences, we can compare the average in and out-degree for each node. Figure 4.1 shows the mean in and out-degrees for each region and player index for the teams that finished top and bottom of their league. Immediately, we notice the stark difference between the two sets of teams, in that those that finished top of the league have significantly higher degree values across all regions and players. For all teams, region 17, in front of goal, has the highest in-degree, however, the difference is much more significant for teams finishing bottom of the league. For the teams that finished top of the league, we notice that region 14, just outside the penalty box in the centre of the pitch, has a high out-degree, while for teams that finished bottom, regions 13 and 15 have similar values. This suggests that this is a region that stronger teams take advantage of much more than weaker teams, who might be forced out onto the wing by a strong defence.

Furthermore, the teams at the bottom of the league have significantly more regions with

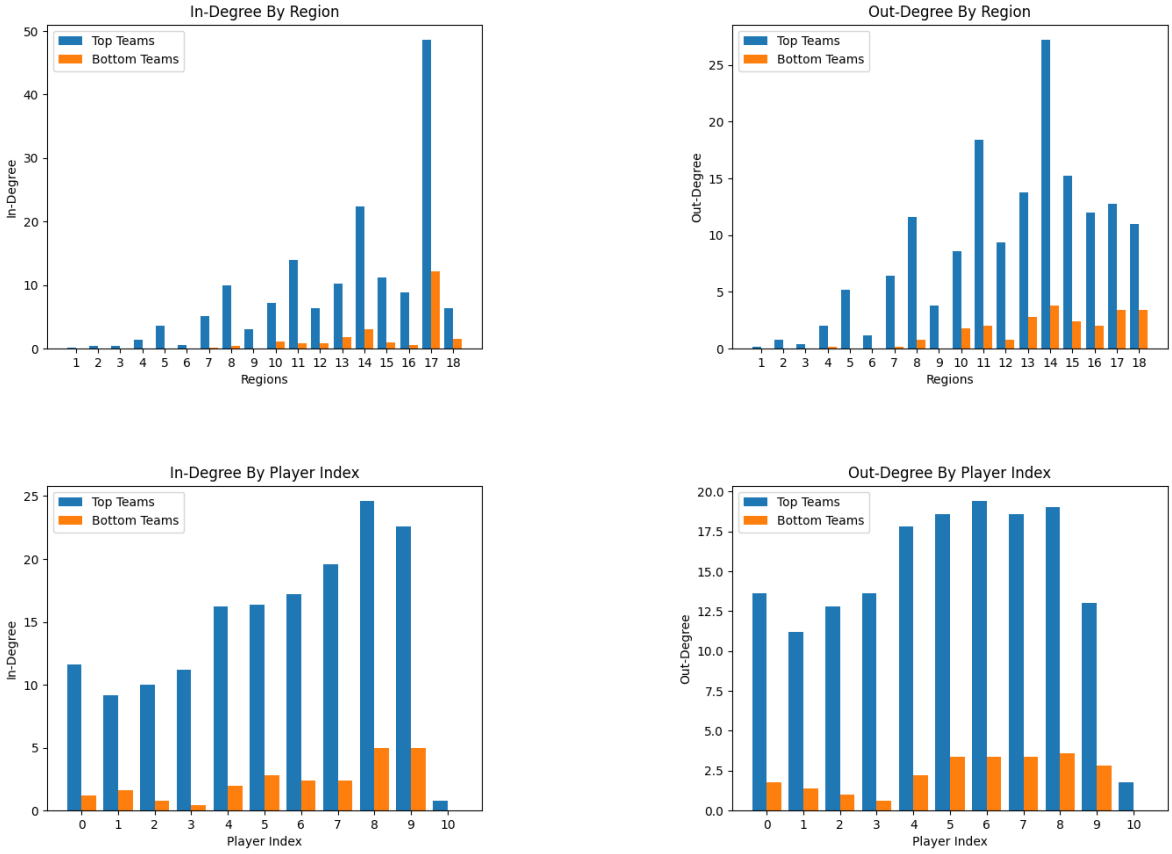


Figure 4.1: Average in and out-degree by region and player index for the top and bottom teams in each league.

zero in-degree and out-degree values. Málaga and Benevento have 13 and 12 regions respectively where no passes occurred, whereas the most of any top team is four, and PSG and Barcelona have only one and two respectively. This stresses the importance of moving the ball around the whole pitch for the top teams, whereas teams at the bottom of the league are much more direct. This is likely at least partially due to the fact that they are less capable of holding onto possession[34], and therefore must be more direct when attacking.

Conversely, when looking at the degree metrics for the networks of players, the distribution is much more similar for the two sets of teams. Forward players have the highest degree values for all teams, however, for the top teams, the defensive and midfield players also have high degree values, whereas those in the teams finishing bottom tend to have single digit degree scores. Although, unlike for the regions, we do see that most players are involved for every team. This reflects that, particularly for the teams finishing bottom, their defensive players are only getting involved higher up the pitch, around the halfway line. Whereas the top teams are able to stretch their team along the entire pitch, creating longer passes and forcing the defence to work harder.

Another interesting difference is the role of the goalkeeper. For the teams finishing bottom of the league, the goalkeeper has very low in-degree and out-degree scores, with Köln having the highest values at only four and three respectively. Whereas, the majority of the teams that are finishing at the top of the league have goalkeepers with in and out-degree values in the double figures. This reflects a new role that the goalkeeper is playing in the modern game, which involves them being able to pass the ball around the back line, known as playing out from the back[35]. This would also explain why these teams use more regions of the pitch, as their goalkeepers are more confident in their ability on the ball close to their own goal, stretching out the defence. The Juventus goalkeeper, however, has in and out-degree values of only five and six respectively, suggesting that they do not play this particular style of football, or, at least, it is not where their goals originate from.

The average clustering coefficient is the one centrality metric where the bottom teams have higher values than the top teams. Of these 10 teams, the teams with the highest average clustering coefficient for the network of regions are West Bromich Albion, Köln and Benevento, followed closely by Juventus. It is particularly interesting to see Juventus in the

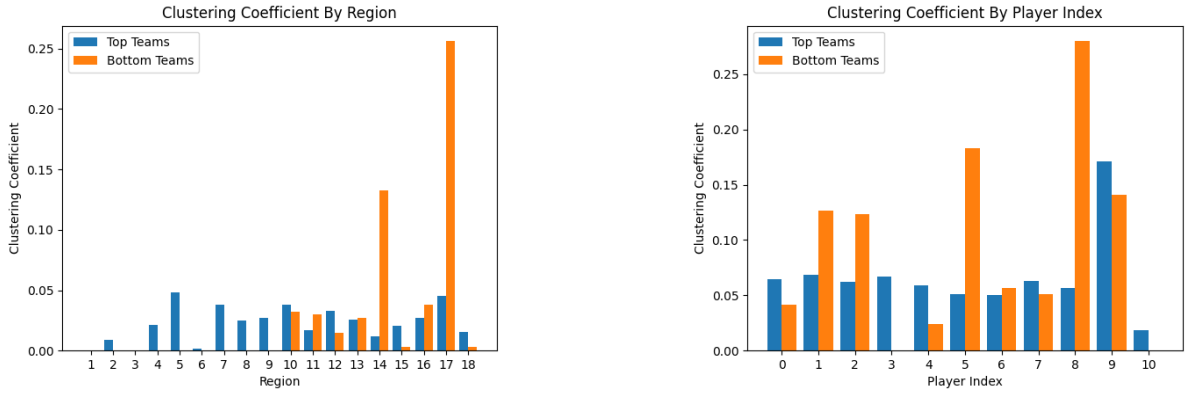


Figure 4.2: Average clustering coefficient by region and player index for the top and bottom teams in each league.

mix here as their degree metrics were more similar to the teams that finished bottom of the league, suggesting that this is another play style that links these teams.

We can understand this better by investigating the distribution of clustering coefficient values across nodes. Figure 4.2 shows the average clustering coefficient by region and player index for the teams that finished top and bottom of their league. The teams that finished bottom of the league, who typically had a higher average clustering coefficient, actually have more regions where the clustering coefficient is zero, with about half of their regions falling into this category. Whereas, the top teams, who have lower average clustering coefficient values, tend to have more non-zero regions. Juventus only has four regions with zero clustering coefficient, so in actual fact are more similar to the top teams in this regard. We see a similar pattern for both regions and players, with the top teams having a relatively even distribution of clustering coefficient and the teams that finished bottom having a few regions or players that are significantly more central with regards to this metric. Therefore, while these teams may have a high average clustering coefficient, this does not reflect a better connected network as a whole, and they likely play quite a predictable brand of football that is easier to defend against.

When analysing the centrality scores for the networks of players, we notice that Metz

performs better than the other lowest ranked teams. For all centrality measures, except pagerank, Metz also has a high maximum value when compared to other teams that finished bottom of the league. While the league finishing position is a good indicator of how good a team is, there is of course some randomness involved and therefore it is useful to check their finishing position across several seasons. By finishing bottom of the league, all of these teams would have been relegated to the lower division, however Metz were promoted again the following season. When we consider that it typically takes an average of six seasons to bounce back from relegation[36], this shows that Metz are playing good football. However, when in the top division, they do tend to finish in the bottom half, with tenth being their top finish of the last decade[37]. Metz play in Ligue Un, the French top division, and studies have previously shown that the strict financial rules in France have provided an even playing field amongst the French teams, but affected their performance against other European clubs[38][39]. Therefore, we would likely not expect to see as big of a gap between French teams as we see between teams in other leagues.

Málaga have a very low average value of pagerank, but a high maximum value, when compared to other centrality measures for the networks of players. As previously discussed, pagerank is a measure of the probability of a player getting the ball, so having a player with a very large value suggests that Málaga have one key player who is very important to their goal scoring. Figure 4.3 shows Málaga's pagerank scores for each player and region. The player with the highest pagerank is at index 2, a defender. This is initially surprising, however when we look at their squad that season, we see that their defenders are actually some of their most experienced and expensive players[37]. Players in position 4 and 7 also have relatively high pagerank scores, along with regions 11 and 14, the central regions in the opposition half. This suggests that they score many fast break goals, where possession is won back near their own goal and the ball is quickly moved down the pitch. These goals tend to be scored by weaker teams who are unable to hold possession. When we look at Málaga's other league results, we find that they never got back into the top division, and in 2021/22 were relegated again to the third division of Spanish football[37], suggesting that they are playing consistently worse football than other teams.

Analysing the maximum clique values for networks of regions and players across the top

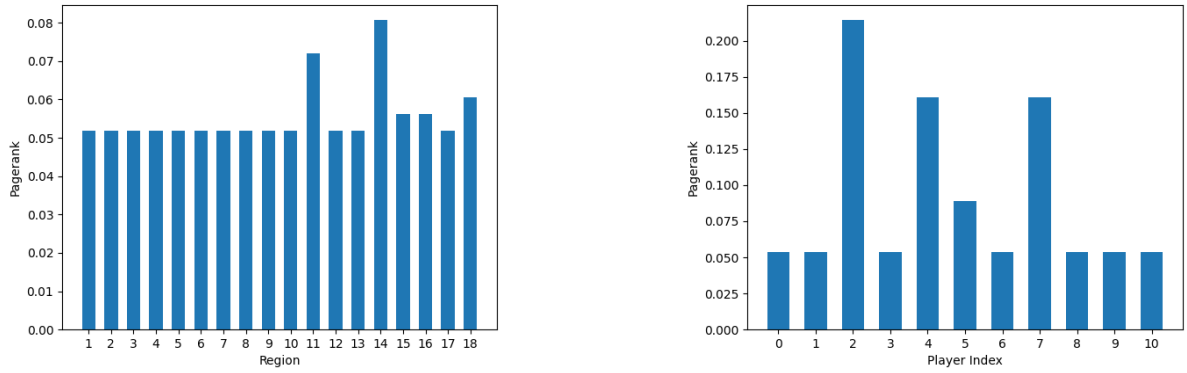


Figure 4.3: Pagerank by player and region for Málaga.

and bottom teams provides some valuable insights. Starting with the networks of regions on the pitch, even for the top teams, these are quite small, with only three nodes. However, digging in further, we find that each of the top teams have at least two cliques with three nodes, and PSG having four and Manchester City seven of this size. This aligns with the play style of having triangles that the ball is moved around[40]. The most popular regions in these cliques are between index 10 and 15, from the halfway line to the edge of the penalty box. Only Manchester City has regions directly in front of goal included in their cliques. This aligns with the particular style of football coached by Manchester City manager Pep Guardiola, who likes to have possession for a high proportion of the game and pass the ball safely between players waiting for the ideal opportunity to shoot, which is often very close to goal[41]. Of the bottom teams, only Köln has a clique with more than one region, and even that is only of size two, suggesting that the lower ranked teams pass the ball around much less in the build up to a goal.

The maximum cliques for the network of players cover a much higher fraction of the nodes, with three of the top teams having cliques containing five nodes, and Manchester City and PSG both having maximum cliques sizes of seven. PSG actually has two cliques of size seven, although both include six of the same nodes. These cliques stretch the length of the team, including both defensive and attacking players, which, when we consider the small

cliques in the networks of regions, suggests that these players are moving high up the pitch in attack. Málaga is the only team who finished bottom of the league that does not have a clique larger than one, all other bottom teams have several cliques of size two, but none larger. Further emphasising that these weaker teams are passing the ball around between players less than the top teams and, instead, being more direct on goal.

Looking at betweenness, we see that all of the top teams have similar median metric values, with only Barcelona's being slightly greater. However, when we look at the maximum values, we see that Barcelona's is much higher than the other teams. The player in this position for Barcelona is at index 8, which is likely either an attacking midfielder or a winger. At this time, Barcelona had €100m midfielder Coutinho[37], a Brazilian player known for his technical ability on the ball. They also had Iniesta, a Spanish player with great ball control[37]. Having these strong, technical players in the centre of the pitch would allow the team to pass the ball around in a manner that a less technical team could not achieve.

Comparing the average and the maximum values for the networks of centrality metrics across teams provides some interesting insights. For degree metrics, the teams finishing top of the league have maximum values approximately twice as high as their average, whereas the teams finishing bottom of the league have a maximum greater by a factor of three to six. PSG actually has median in and out-degree values that are quite close to the maximum values, suggesting that the players are all equally involved. Conversely, the discrepancy between these values for the bottom teams suggests that they have a single or small number of very important players in the build up to their goals. Clustering coefficient follows a similar behaviour, with all of the bottom teams, except for Benevento, having a large difference between the average and maximum. Interestingly, Bayern Munich also has a maximum value which is significantly higher than the average. Again, this reflects having central players in the team, rather than having a strong team of 11 players who can manage the game evenly between them.

Comparison between different football leagues is hotly debated both in the media and academically[42]. Many people believe that the English league is the best in the world, with many of the most high profile teams and players being based in England and the league being the most competitive (however, this may no longer be the case, with Manchester

City winning five of the six titles since 2017/18[37]). However, my analysis did not show a significant difference between teams in different leagues. The result of classification analysis found that there was no relationship between metric values and leagues, rather teams that performed well were more similar, regardless of the league they played in. This suggests that these leagues are all of a similar standard.

One simply has to look at the tables of metric centrality scores to see the clear delineation between the teams finishing top and those finishing bottom of the league. In general, the teams finishing top of their league tend to play a more fluid football, with more players and regions of the pitch involved, making and receiving passes prior to a goal being scored. Furthermore, they have a higher betweenness score, suggesting that more players are between others and providing passing channels. This result is emphasised when we look at the results of dimension reduction in figures 3.4 and 3.5. In figure 3.4b, which shows the results of dimension reduction for the networks of regions, coloured by league finishing position, we can clearly see a shift from lighter colours in the top left to darker colours in the bottom right of the chart. Similarly for figure 3.5b, which shows the same results for the networks of players, the teams in the bottom left of the chart finished higher on average than those in the top right. Furthermore, looking at the results of the Champions League, the competition played by the top teams across all leagues in Europe, we find that there were six different winners across three countries in the last 10 years, and if we include runners-up, this number increases to 12 teams across five countries[37]. This emphasises a competitiveness across leagues in Europe, such that no single league is dominating over the others.

While there is a clear distinction between the two clusters, some teams that finished in the top half are clustered with the teams that finished in the bottom half, and vice versa. For simplicity, I will refer to the cluster with majority teams that finished in the top half as cluster A and the cluster of teams that majority finished in the bottom half as cluster B. Looking into the English teams first, four teams finished in the top half but were in cluster B. Of these four teams, in the six years since this data was produced, two have been relegated; Burnley and Leicester, and another, Everton, have been close to relegation several times[37]. The other team, Newcastle United, finished in tenth position, just inside the top half, and finished in the bottom half of the league in the four seasons directly following this one. Their

results only changed when they were bought by Saudi Arabia in 2021[43]. This suggests that their finishing in the top half in 2017/18 was more luck than skill and that their play style was a hint at what was to come. Bournemouth did finish in the top half of the league, but were in cluster B. In this instance, Bournemouth do appear to be an anomaly as they have not finished in the top half of the league since and were relegated in 2019/20[37]. It would be interesting to analyse the goals scored by Bournemouth in subsequent seasons to see if this pattern remains consistent.

Similarly, of the La Liga teams that finished in the top half but were in cluster B, Girona and Eibar have both since been relegated and Getafe has finished in the bottom half for four out of the six seasons[37]. For the reverse situation, it is more nuanced, with Deportivo, Espanyol and Alavés all being relegated since then, but Real Sociedad staying consistently in the top half of the league[37]. Similarly, the only team that didn't finish top of the league in the Bundesliga but were in cluster A was Hertha Berlin, who have since been relegated[37]. In the reverse situation, Schalke and Stuttgart have both been relegated, although it is worth noting that Stuttgart have since got back into the Bundesliga and finished second in the 2019/20 season[37]. In Serie A, the Italian league, the only two teams who were clustered differently than their final position, Udinese and Genoa, both finished in the bottom half of the league, but were in cluster A. Both have consistently finished in the bottom half, and Genoa has since been relegated[37]. Finally, in the French Ligue Un, Dijon and Lille finished in the bottom half, but were in cluster A. Lille has consistently finished in the top half of the league, whereas Dijon has since been relagated[37]. Saint Étienne and Nantes finished in the top half of the league but where in cluster B, both have since consistently finished in the bottom, and Saint Étienne were relegated in 2021/22[37].

These results show that there is much more randomness for poorer teams compared to top teams. The teams that consistently finish in the top half of the league tend to score similar goals, while poorer teams can occasionally match this style of play. There could be a variety of reasons driving this. One cause might be that smaller teams tend to change managers more often[44]. In football, managers get fired if the team's owners do not believe they are performing as well as expected. A new manager might change the way the team plays, and can lead to a new manager bounce, a period where a team plays better than expected under

a new manager[45]. Another reason might be that these teams are trying to play a style more similar to that of the top teams, however, they do not have they squad to achieve this and thus, score fewer goals, leading to a lower league position. This is something that we could understand in more detail by analysing these teams across several seasons.

Chapter 5

Conclusions

In this study, I expanded on previous work by Lopez and Clemente. I created passing networks for the teams in the top five leagues across Europe to understand how the ball is passed prior to a goal being scored. Lopez's study looked at all passes by two national teams in the 2010 World Cup[15], while Clemente's looked at the passes before a goal was scored or conceded by one club team in Portugal across a single season[14]. My analysis looked at several of the same centrality metrics of these papers, however, by including all teams across several leagues, I was not only able to understand how individual teams scored goals, I was able to compare these networks for teams within a league, and across leagues.

I found that for both the network of players and the network of regions on the pitch, the teams that finished top of their league had significantly better centrality scores. Players and regions both had higher in and out-degree values, and the average total degree was significantly higher. Teams that finished top of the league also had more nodes involved in the network, with a maximum clique of three for the networks of regions and five or seven for the networks of players, compared to only one or two for either network for the teams finishing bottom of the table. Conversely, teams finishing bottom had similar clustering coefficients to those that finished top of the league, and some even had higher values for the network of players. I found that this was a result of skewness, due to these teams having one or two players which are significantly more central than any other, whereas the distribution is more even for top teams. This is reflected in the high value of maximum

clustering coefficient found for these weaker sides.

By normalising and summing each of the centrality measures, I was able to calculate a total centrality score for each of the players. Interestingly, the player in position 10, or the furthest forward player, had the lowest centrality score for all teams. This is likely due to to the fact that, while they are important in goal scoring, they are not a key part of the build up play. I also found that the teams finishing top of the league had an even distribution of total centrality score across the whole team, with the back line and midfield being of equal importance. For the teams that finished bottom, only Köln had a similar pattern, with the remaining four teams having defensive lines which were significantly less central than their attacks.

I tested three techniques to reduce the number of dimensions of the data to two, such that the results could be easily plot and classification analysis done. Of the three methods, UMAP performed the best, providing data which was most similar to the original whilst also creating well-defined clusters. For the network of regions, I used the average centrality scores for each team, and for the networks of players I used the average and maximum centrality scores. As the two-dimensional data was well clustered, I tested K-NN and SVM for the classification analysis, with SVM performing better.

Both network types gave similar results in the classification analysis, suggesting that teams that do well, or finish in the top half of the league, score similar goals. There was no relationship between the league a team played in and their passing network, suggesting that Europe as a whole play a similar brand of football. This may be due to the fact that managers tend to stay within one continent, for example, Pep Guardiola managed Barcelona, Bayern Munich and, now, Manchester City, who all play similar styles of football and have won the league several times[37]. The outliers, the teams that played as if they finished bottom of the league and finished top, tended to be teams that had a one off good season, and since then have consistently finished in the bottom or even been relegated to a lower league. This suggests that this analysis can predict how well a team will do, not just in the current season, but in future seasons as well.

There are many avenues for follow up studies in this area, for example, a longitudinal study across several seasons. While football is a technical game, there is an element of

randomness and teams can over-perform or under-perform in a given season. By analysing, say, five seasons, we could remove much of the randomness and increase the accuracy of our classification algorithm. Another interesting analysis would be to investigate if European teams play a different style of football to teams in other continents. It could be that the top teams in Europe score similar goals because they have a pool of managers who move between them, or because they play each other in the Champions League. Therefore it would be interesting to analyse the goals scored in other continents to see if the same trends appear. Finally, with the women's game emerging and more data being made available, it would be interesting to see how different the goals scored by the women are compared to those scored by the men, and if the same patterns can be seen in the women's game. We could also investigate if the women's team for a club play a similar style of football to their corresponding men's team.

Bibliography

[1] Veroutsos, E. The Most Popular Sports In The World. [online]. World Atlas: New Jersey; 2023. [Accessed on 26 August 2024]. Available from https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html.

[2] Weil, E. et al. Football. [online]. Encyclopedia Britannica: Edinburgh; 2024 [Accessed 28 August 2024]. Available from: https://www.britannica.com/sports/football-soccer

[3] Herberger T., Litke C. The Impact of Big Data and Sports Analytics on Professional Football: A Systematic Literature Review. In: Herberger T., and Dótsch J. Digitalization, Digital Transformation and Sustainability in the Global Economy. Cham:Springer, 2021. p. 147-171.

[4] Thakkar P., Shah M. An Assessment of Football Through the Lens of Data Science. *Ann. Data. Sci.* [online]. 2021;8, 823-836. Available from: https://doi.org/10.1007/s40745-021-00323-2.

[5] Smith R. How Arsenal and Arsène Wenger Bought Into Analytics. [online]. New York Times: New York; 2017 [Accessed August 6th 2024]. Available from: https://www.nytimes.com/2017/02/03/sports/soccer/arsenal-arsene-wenger-analytics.html

[6] Sherlock H. The average cost of a Premier League player and why we should all despair. [online]. Football Fancast: UK; 2019 [Accessed 6 August

2024]. Available from https://www.footballfancast.com/premier-league/the-average-cost-of-a-premier-league-player-and-why-we-should-all-despair/

[7] Statista, Statista Research Department. Off course betting turnover in the gambling industry in Great Britain from April 2009 to March 2023, by sector. [online]. Statista: Hamburg; 2024 [Accessed 6 August 2024]. Available from: https://www.statista.com/statistics/203411/revenue-of-off-course-betting-in-the-uk/

[8] Oxera. Odds on? What was the probability of Leicester City's 2016 success? [online]. Oxera: Oxford; 2018 [Accessed on 7 August 2024]. Available from: https://www.oxera.com/wp-content/uploads/2018/07/What-was-the-probability-of-Leicester-Citys-2016-success-1.pdf-1.pdf

[9] Ichinose G. et al. Robustness of football passing networks against continuous node and link removals. Chaos, Solitons and Fractals [online]. 2021;147:110973. [Accessed on August 7 2024]. Available from: https://arxiv.org/abs/2003.13465.

[10] Trequattrini R. et al. Network analysis and football team performance: a first application. Team Performance Management. 2915;21(1):85-110.

[11] Gama, J. et al. Network analysis and intra-team activity in attacking phases of professional football. International Journal of Performance Analysis in Sport. 2014;14(3):692–708.

[12] Rago V. et al. Methods to collect and interpret external training load using microtechnology incorporating GPS in professional football: a systematic review. Research in Sports Medicine. 2019;28(3):437-458.

[13] Pappalardo, L., Cintia, P., Rossi, A. et al. *A public data set of spatio-temporal match events in soccer competitions.* Sci Data **6**, 236 (2019). https://doi.org/10.1038/s41597-019-0247-7

[14] Clemente F, Martins F, Mendes R. (2016). Analysis of scored and conceded goals by a football team throughout a season: A network analysis. Kinesiology. 48. 103-114. 10.26582/k.48.1.5.

[15] Lopez Pena J, Touchette H. A network theory analysis of football strategies. eprint arXiv. 2012;arXiv:1206.6904. [Accessed July 27 2024] Available from: https://ui.adsabs.harvard.edu/abs/2012arXiv1206.6904L

[16] Who Scored. Premier League. [online]. ODI: London; 2024 [Accessed 26 July 2024]. Available from: https://www.whoscored.com/Regions/252/Tournaments/2/Seasons/6829/Stages/15151/TeamStatistics/England-Premier-League-2017-2018

[17] Carling C, Mark Williams A, Reilly T. Handbook of soccer match analytics: a systematic approach to improving performance. 1st ed. London & New York: Taylor & Francis Group.

[18] Bauer P, Anzer G, and Shaw L. Putting team formations in association football into context. IOS. 2023; 9(1):39-59.

[19] Menczer F, Fortunato S, Davis CA, A first course in network science. 2nd ed. Cambridge University Press.

[20] Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science Conference, 2008, Pasadena, CA. pp. 11-15.

[21] Scheie F. Damping factor analysis for pageRank [Internet] [Dissertation]. 2022. Available from: https://urn.kb.se/resolve?urn=urn:nbn:se:mdh:diva-59143

[22] Pedregosa F et al. Scikit-learn: Machine learning in python. Joirnal of machine learning research. 2011;12:2825-2830

[23] Mackiewicz A, RAtajczak W. Principal components analysis (PCA). Computers and Geosciences. 1993;19(3):302-342.

[24] Velliangiri S et al. A review of dimensionality reduction techniques for efficient computation. Procedia Computer Science. 2019;165:104-111.

[25] Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9(86):2579-2605.

[26] Van Erven T., Harremos P. Rényi divergence and Killback-Leibler divergence. IEEE Transactions on Information Theory. 2014;60(7):3797-3820.

[27] Linderman, G et al. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. Nature Methods. 2019;16(3):243-245.

[28] McInnes L et al. UMAP: uniform manifold approximation and projection. Journal of Open Source Software. 2018;3(29):861. Available from: https://doi.org/10.21105/joss.00861

[29] McInnes L et al. UMAP: Uniform manifold approximation and projection. The Journal of Open Source Software. 2018;3(29):861.

[30] Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 2016 Jun;4(11):218.

[31] K. Taunk, S. De, S. Verma and A. Swetapadma, A brief review of nearest neighbor algorithm for learning and classification. International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 1255-1260.

[32] Suthaharan, S. Support Vector Machine. In: Suthaharan, S. Machine Learning Models and Algorithms for Big Data Classification. Boston, MA: Springer; 2016. p. 207-235.

[33] Yue, S., Li, P. and Hao, P. SVM classification:Its contents and challenges. Appl. Math. Chin. Univ. 2003;18:332–342.

[34] Collet, C. The possession game? A comparative analysis of ball retention and team success in European and international football, 2007–2010. Journal of Sports Sciences. 2012;31(2),123–136.

[35] Obetko M. Technical–tactical profile of an elite soccer goalkeeper. Journal of Physical Education and Sport. 2022;22(1):38-46.

[36] Cerqua A. If you get knocked down, how long before you get up again? International Journal of Sport Finance. 2014;9(4):284.

[37] FBRef. [online]. Philadelphia;2024 [Accessed 1 August 2024]. Available from: https://fbref.com/en/about

[38] Andreff W. Financial and sporting performance in French football Ligue 1: Influence on the players' market. International Journal of Financial Studies. 2018;6(4):91.

[39] Andreff W. French professional football: how much different? In: Goddard J, Sloane P. Handbook on the economics of professional football. Cheltenham: Edward Elgar Publishing; 2014. p. 298-319.

[40] Elancheran S. Geometry in football: Understanding the triangle. Outside of the boot [online]. 2016 Nov 3. [Accessed on August 8]. Available from: https://outsideoftheboot.com/2016/11/23/geometry-in-football-understanding-the-triangle/#:~:text=If%20three%20players%20form%20a,passes%20with%20the%20third%20player

[41] Immler S et al. Guardiola, Klopp, and Pochettino: The purveyors of what? The use of passing network analysis to identify and compare coaching styles in professional football. Frontiers in Sports and Active Living. 2021;3.

[42] Dellal A et al. omparison of physical and technical performance in European soccer match-play: FA Premier League and La Liga. European Journal of Sport Science. 2011;11(1):51-59.

[43] Marshall S et al. The ethics behind the recent takeover of Newcastle United Football Club. Journal of Physical Fitness, Medicine and Treatment in Sports. 2022;9(4):555768.

[44] Hughes M et al. Short-term versus long-retm impact of managers: Evidence from the football industry. British Jounral of Management. 2010;21(2):571-589.

[45] Goddard J. The football manager. In: Goddard J, Sloane P. Handbook on the economics of professional football. Cheltenham: Edward Elgar Publishing; 2014. p. 298-319.