Universiteit Leiden
The Netherlands

Opleiding Informatica & Economie

Identifying Successful Football Teams in the European Transfer Market: a Network Science Approach

Tristan Dieles

Supervisors:
Frank Takes and Carolina Mattsson

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

August 20, 2022

Abstract

Both the in-game performance and financial performance are descriptive of the landscape in which football clubs maneuver. Our goal is to explore if, and to what extent, a relation is present between a team's role in the transfer market and their performance within a particular football league. This research will be conducted by means of network analysis where the network is constructed over the eight best ranked European leagues over the past 28 years, and consists of teams as nodes and transactions between teams as edges. A team's role in the transfer market is expressed by a set of network position variables and a set of network engagement variables. A team's league performance is expressed on the basis of the standardized average position in the final standings over the past 28 years. The approach consists of multiple linear regression models that elucidate the relation between a team's role in the transfer market and their league performance. We found that the football transfer network follows characteristics that are similar to other real-world networks. Moreover, significant differences in competitive and trading profiles between leagues have been observed which corresponds to existing literature. Furthermore, we found that both the position and engagement of a team are correlated to league performance. Concretely, betweenness centrality and weighted in-degree are negatively correlated to a team's league performance when controlling for the league and degree of the team. Conversely, closeness centrality, clustering coefficient, weighted out-degree and weighted total degree are positively correlated when the team's league and degree are accounted for. This implies that a team's performance would benefit from a transfer policy that is directed towards obtaining a central role in the European network as well as in their league.

Contents

Chapter 1

Introduction

Football, or soccer in the United States, is the world's most popular sport by any objective metric. Football's popularity is especially apparent in Europe [Mat03]. In European countries, football teams compete against other teams in a national or regional league. At the end of a season, the best teams of a league promote to a higher league and the worst teams relegate to a lower league. Logically following from this dynamic, the best performing teams compete in the highest league on a national stage. This league often attracts the most interest of the fans which is accompanied by more financial resources for the teams.

Football research has attained increased relevance in academia over the last decades. One of the main topics within football research is the analysis of the in-game performance of football clubs. Clubs — and their teams — are often defined on the basis of their performance. This results from the fact that the access to higher leagues — which are also the most popular — is mainly determined by the team's performance [Fri07]. Along with financial stimuli, this gives teams an incentive to maximize match wins in the leagues they appear in. This research topic therefore focuses on determining what characteristics contribute to a team's ability to win matches. Much research has been devoted to what in-game characteristics benefit this goal ([GMB+20], [LZ20]). For this thesis, it is relevant to note that a season in football consists of a collection of matches that are played in one year between teams in the same league. A team's league performance is then defined as the collection of (in-game) performances over a season relative to its competitors.

More recently, due to the increased disclosure of financial information of football clubs [Fri07], another line of research approaches football clubs as businesses that, in general, rarely generate a profit [SS97], [HW10]. A major part of this financial aspect of football clubs is the circulation of players [DG12]. In Europe, a player can move clubs in two ways. A club can acquire a player from another club by exchanging money in return for the player; this is called a transfer. This gives the 'acquiring' club the rights to contract the player who gets paid a salary. Another option is to acquire the player for a certain period in which the 'acquiring' club pays the salary. However, the player remains under contract at his club; this is called a loan. Football's labour market is made up by these dynamics of teams buying and loaning a player from and to other teams, and by the contractual agreements between players and clubs [DG12]. The football transfer market as regarded in this thesis is defined solely by the former part.

Both research into the in-game performance and financial performance are descriptive of the landscape in which football clubs maneuver. In fact, the two are inter-related in the sense that they influence each other in a circular manner. In-game and league performance influence the revenue

that a club brings in. Consequently, this influences the financial aspect of the club. For example, if a team plays well, the club is likely to earn prize money, sell their players for a significantly higher price [Mou16], increase match attendance — although this also depends on the competitiveness of the league [Gyi20] — and attract more lucrative sponsorship deals. Part of this revenue allows the club to acquire new players. This dynamic is underlined by Mourao [Mou16], who describes the significant effect of league performance on obtaining a higher number of incoming players. On this basis, Pantuso and Hvattum [PH21] discuss tactical decisions to be applied to a club's transfer policy based on in-game data.

However, the inverse interaction between these two real-life intricacies has been explored less. In fact, Mourao expresses the yet unexplored potential of this research topic as a direction of further research. In other words, the influence of the circulation of players of a team on the league performance has not been a popular topic within football research. Clubs could take on a strategic role in the football transfer market. The market behaviour of a club that corresponds with their role influences their performance. For example, if a club's strategy is to develop players to profit on the transfer fee, their trading behaviour is partially signified by frequently buying and selling players. This could lead to less stability in the team and other teams acquiring the 'developed' players which, in turn, affects the team's league performance.

As strong method to uncover the intricate dynamics within the football transfer market, driven by each team's trading interactions, is network analysis. Network analysis encodes the interactions between a system's components in a structural manner, that are otherwise hard to capture. In this thesis, the system that we try to uncover is the football transfer market which is composed of teams and their trading interactions. The power of network analysis lies in the fact that all real-world networks are driven by a common set of underlying laws and principles. The universal properties of network characteristics allow for a robust and meaningful description of the network as a whole and of its individual components. In addition, the possibility to visually represent networks helps us to understand the interactions that occur within a system. The tool box of measures and metrics that network science offers to understand a system's components and their interaction is what makes network analysis a valuable method to tell the story of the football transfer network [Bar16], [New18].

This thesis attempts to construct an approach to assess if the performance of football teams can be expressed as a derivative of their position and engagement in the player transfer market. The player transfer market can be represented as a network — of which a more precise definition will be posed later. It is through this representation that a team's position and engagement will be defined. Hereafter, a measure to quantify a team's league performance is presented. The thesis proceeds by describing the characteristics of the network that is formed by player transactions between clubs (i.e., the market). Subsequently, in an attempt to capture the position and engagement of a team in the transfer market, this research proposes a model that relates a team's trading behaviour to their league performance.

This thesis works towards answering two distinct research questions; one descriptive question and one exploratory question. In order to gain a better overview of the characteristics of the European football transfer market, we pose the following descriptive question:

RQ1: *What are the characteristics of the European football transfer network?*

Based on the findings related to this question, we will explore the relation between a team's performance and their position in the network. The second research question is therefore:

RQ2: *How does domestic league performance of a football club depend on the position and engagement of this club in the European player transfer network?*

The remainder of this thesis is organized as follows: Chapter 2 discusses the data set obtained and how the data was processed before the analysis; Chapter 3 includes the related literature and definitions; Chapter 4 describes the layout of the statistical model and the setup of the experiments; Chapter 5 discusses the results; Chapter 6 concludes and provides suggestions for future work.

Chapter 2

Data

In this chapter, we will set forth the source of the data sets along with their properties. Subsequently, we will discuss the steps taken when preprocessing the data.

2.1 Data Source

This section details the data sets that are used in this research. This research is based upon two data sets. The transaction data set contains transfers between teams of the eight best European leagues over the last 29 years (https://github.com/ewenme/transfers). The transaction data set is identified by data points that consist of **a player** that is transferred from a **team** to another **team** at a certain **time** for a certain **price**. This data is used to construct the edges in the network.

The league result data set contains almost all domestic league results over the last 28 years for the leagues included in the transaction data set (https://www.football-data.co.uk/data.php). In comparison to the transaction data set, the league result data set missed 19 seasons for the Russian Premier Liga and one season for the Portugese Liga NOS. In Section 2.2 is detailed how we have processed this difference in available data. The league result data set consists of match records of a **team** that played against another **team** which resulted in a **final score**. This data will allow us to incorporate a league performance measure of the clubs.

2.2 Data Preprocessing

The main function of this section is to set forth the steps taken to guarantee a reliable, representative and reproducible configuration of the data. We have opted to use the past 28 seasons included in the transaction data set (instead of the past 29 seasons) to ensure that the time frames of both data sets are congruent. Furthermore, we have chosen to use the seasons that are missing from the league result data set to construct the transfer network as this does imply more accurate team characteristics in the context of network position and engagement for all teams analysed. A summary of the selected leagues and seasons of both data sets are displayed in Table 2.1, as well as the associated countries and country codes of these leagues.

The steps taken to work from the raw data to a reliable configuration for the research are indicated below:

1. The original data set contained transfers of clubs from the English Championship. These have been excluded as the analysis will be conducted over the top tiers of all countries in the data. This means that the clubs will be evenly distributed over the countries and their respective leagues. The data set then includes 136,339 transactions in the default configuration.

2. A club can be seen as the overarching association that unites a collection of teams. All youth teams and non-first teams of a certain club have been aggregated, such that the influence of affiliated teams on the position and performance of the first team can be interpreted collectively. This change of notation was applied in all 10,860 instances (which is 7.97% of the transactions), but did not affect the total number of transactions analysed.

3. Over the years, some clubs have changed names. Furthermore, some teams were noted under different names in different instances within the data set. The current or most recent name of a team was chosen as standard and the differing notations were adjusted. This did not lead to a change in the number of transactions.

4. Some transactions were present twice. This is the case when the player moved from a team present in the above competitions to another team present in the above competitions. Double notation of a transaction is also present when a player is loaned out. This transaction is taken into account twice: at the beginning of the loan and at the end. A loan deal between two teams present in the data sets could thus lead to these loan-transfers being documented up to four times. To account for the transactions that have been documented more than once, the first occurrence of every distinct transfer in the data set is used. 96,645 transactions are present after accounting for double documented transfers.

5. Moreover, some transfers existed where the player transferred to and from the same team. These transfers have probably not been documented properly and are therefore deleted from the configuration used in this research. This applied to 240 transactions.

6. Lastly, transactions containing players that retired, moved to or from an unknown club, or that became a free agent have been disregarded. This was a total of 4,954 instances.

In totality, this amounts to 91,451 transactions upon which the network will be based.

League in data set	Country	Country code	Transaction data	League result data
English Premier League	United Kingdom	GB	93/94 - 20/21	93/94 - 20/21
English Championship	United Kingdom	GB	None	None
French Ligue 1	France	FR	93/94 - 20/21	93/94 - 20/21
German 1.Bundesliga	Germany	DE	93/94 - 20/21	93/94 - 20/21
Italian Serie A	Italy	IT	93/94 - 20/21	93/94 - 20/21
Spanish La Liga	Spain	ES	93/94 - 20/21	93/94 - 20/21
Portugese Liga NOS[1]	Portugal	PT	93/94 - 20/21	94/95 - 20/21
Dutch Eredivisie	Netherlands	NL	93/94 - 20/21	93/94 - 20/21
Russian Premier Liga[1]	Russia	RU	93/94 - 20/21	12/13 - 20/21
Total			224	204

[1] We are less certain about these results as not all seasons included overlap across data sets

Table 2.1: An overview of the seasons corresponding to the leagues included the research.

The processed data set D described above is used for both research questions and is used to determine the network and its variables used in Chapter 5.

All teams that have been present in the leagues for at least one season have been used in the linear regression model (see Section 4.1). The teams that have been traded with, but that were not present in one of the leagues in at least one of the seasons mentioned in Table 2.1, have not been included. This means that a total of 334 teams have been used for the statistical analysis that is included for the second research question. All 334 teams possess all graph measures as discussed in Section 3.2.2 and a league performance measure as proposed in Section 3.2.4.

Chapter 3

Background and Related Work

This chapter provides the contextual framework and definitions that this research is embedded in. This is structured by first introducing the relevant existing literature. Secondly, based upon this contextual framework, the definitions of variables and attributes used in this thesis will be presented.

3.1 Related Work

This section lays out an overview of academic research in which this thesis is rooted.

3.1.1 League performance

In football, the predominant aspect upon which a team judged is their performance. Therefore, the study of a football team's performance has been a popular topic within football research. This interest is not only limited to football, but it applies to a multitude of sports. The unique setting where sports is accessible and relatable to the public makes sports an attractive field of research. Much effort has therefore been invested in analysing the performance of sports teams.

The league performance of a team is a derivative of its individual match performances. In football, to win a match, a team has to score. An extensive analysis of tactical in-game patterns that significantly contribute to a team's scoring ability is published by Goes et al [GMB+20]. The collection of match performances throughout a season constructs a team's seasonal performance (i.e., league performance). A match can therefore be seen as a sample of a team's seasonal performance. A team's seasonal performance regresses to the mean as the season progresses. This concept is illustrated by Beck and Meyer [BM11].

However, the exact relation between a match win and a team's league performance depends on the rating or ranking system the league applies. Different sports use different methods to rank its teams. An elaborate mathematical approach to these methods has been written by Langville and Meyer [LM12]. Differences between leagues can also be observed in the competitive profiles of leagues. This is a notion observed in this thesis and analysed in further detail by Vales et al [VLG+18]. Gymesi has found that a more competitive league increases match attendance [Gyi20]. Moreover, differences among leagues also appear on a smaller, in-game scale. Li and Zhao demonstrate this clearly by comparing goal scoring patterns across leagues [LZ20].

3.1.2 Financial performance

An aspect that has received increased attention within football research is the financial aspect. This increased prevalence within the academic spheres is a result of two shifting trends within European football. The first trend is the change in the regulatory framework regarding the disclosure of financial information by football teams. This has allowed for more insight in the financial aspects of football teams. Secondly, teams have earned increased revenue as a result of more lucrative sponsorship and broadcasting deals. These changes have affected the industry in which football teams compete and thus the financial and managerial structure of teams. This has been documented by Szymanski and Smith [SS97]. However, the increased revenue described in the latter trend has not led to teams becoming more profitable as they still tend to generate a loss [HW10].

The financial aspect of football entails a broad array of topics which ranges from forecasting matches to the determinants of managerial change within a team. This broad range of topics has been set forth substantially by Dobson and Goddard [DG12]. One major aspect within the financial realm that affects a team's financial performance is the labour market where players are contracted and traded. An extensive overview of the factors and actors that determine the dynamics of the football labour market has been reported by Frick [Fri07]. The most significant aspect of the labour market is the transfer of players. As teams have an objective related to in-game performance as well as maintaining financial sustainability, the interplay between managing a good performing squad while profiting on the transfers of the players is an insightful research direction. This has been explored in-depth by Mourao who revealed the significant effect of league performance on the attraction of players [Mou16]. Strategic decisions that result from this interplay have been researched and optimized by Pantuso and Hvattum through a chance-constrained model [PH21].

3.2 Definitions

In this section, we define the terminology and concepts used in the thesis for graphs and (its) measures as connected to related work.

3.2.1 Graphs

Consider data set D as introduced in Section 2.2. Data set D contains z records of transactions of players between two teams. Each transaction (i.e., instance) is represented as a triple $d_i = (u_i, v_i, p_i)$. Here, in the i-th transaction (where $1 \leq i \leq z$), a player p_i is transferred from team u_i to team v_i.

On this basis, directed graph G is defined as $G = (V, E)$. Here, V is a set of n nodes (i.e., teams). Moreover, E is a set of m edges (i.e., transfers of players between teams). An edge is an ordered pair of nodes u, v that is included in E when at least one instance of $d_i = (u, v, p_i)$ exists. The relation between u, v is signified by the attribute of weight w. This constitutes a notation of edges in the form (u, v, w). The weight associated to the pair of nodes u, v — denoted by $w(u, v)$ — corresponds to the number of instances that are present in D of players moving from team u to team v.

3.2.2 Node Measures

In order to express a team's position and engagement in the network, a range of node measures are used. Both global and local measures are used to reflect a team's position. The variables that are implemented to quantify a team's position are *betweenness centrality*; *closeness centrality* and *clustering coefficient*. The variables that are used to quantify a team's engagement are *degree*; *weighted in-degree*; *weighted out-degree*; and *weighted degree*. Moreover, the league of a team is used as a fixed node attribute to account for the differences among leagues ([VLG+18]) that influence the performance of the teams. These definitions are derived from [New18] and [Bar16].

Network position variables

Here, we define the variables used to quantify a team's position in the network. Given Graph G as defined in Section 3.2.1, we use betweenness centrality $B(u)$ as a global variable that is expressed as follows:

$$B(u) = \sum_{v,t \in V v \neq u \neq t} \frac{\sigma_{vt}(u)}{\sigma_{vt}} \tag{3.1}$$

In this thesis, we us the definition of a shortest path as proposed by Barabási [Bar16]. This definition is elaborated upon in Section 3.2.3. Here, σ_{vt} is the set of all shortest paths that connect nodes v and t. A subset of this set is $\sigma_{vt}(u)$. This denotes the shortest paths between nodes v and t that pass through node u. Nodes with a high betweenness centrality have relatively more shortest paths running through them. These teams tend to play an important role in connecting parts of the graph that are relatively unconnected as betweenness centrality is a measure that denotes the influence of a node on the flow of players.

Another global measure that describes the centrality of a node from a different angle is closeness centrality $C(u)$. The definition of closeness centrality is:

$$C(u) = (\frac{1}{n-1} \sum_{v \in V} d(u,v))^{-1} \tag{3.2}$$

Closeness centrality expresses the average distance $d(u,v)$ from node u to all other nodes v. Distance is defined as the length of the shortest path from node u to node v. Closeness centrality expresses the degree to which a node can efficiently spread information through the graph.

As a local measure to describe the degree to which a node u cluster together with its neighbours, the clustering coefficient $CC(u)$ is used. The clustering coefficient is defined as follows:

$$CC(u) = \frac{|\{(u,v,w) \vee (v,u,w) \in E : ((u,t,w) \vee (t,u,w) \in E) \wedge ((v,t,w) \vee (t,v,w) \in E)\}|}{D(u) \cdot (D(u)-1)} \tag{3.3}$$

In this definition $D(u) > 1$ signifies the degree of node u. The degree of a node indicates the number of trading partners team u has in the network and is defined in Equation 3.4. Clustering coefficient $CC(u)$ quantifies the presence of triangles between neighbouring nodes of u. The clustering coefficient therefore expresses the embeddedness of a certain node in its neighbourhood and serves as a measure to identify closely-knit groups.

Network engagement variables

This subsection is where the variables used to describe a team's engagement in the network are proposed. As a measure to express a team's engagement in the transfer network, we use degree $D(u)$. The degree of a node $D(u)$ is the sum of the in-degree $D_i(u)$ (all incoming edges) and the out-degree $D_o(u)$ (all outgoing edges). We therefore define $D(u)$ as:

$$D(u) = D_i(u) + D_o(u) \tag{3.4}$$

The degree $D(u)$ of a team is the number of trading partners team u has. The total number of trading partners consists of trading partners that have sent players ($D_i(u)$) to team u and trading partners that have received players ($D_o(u)$) from team u.

We furthermore use the measures of weighted in- and out-degree to quantify a team's engagement in the network. The sum of these variables forms the weighted (total) degree. Weighted in-degree, out-degree and total degree are expressed in the following manner:

$$W_i(u) = \sum_{v \in V} w(v, u) \tag{3.5a}$$

$$W_o(u) = \sum_{v \in V} w(u, v) \tag{3.5b}$$

$$W(u) = W_i(u) + W_o(u) \tag{3.5c}$$

Here, $W_i(u)$ and $W_o(u)$ are, respectively, the weighted in- and out-degree of team u. These measures express the number of players that have been sent from all teams v to team u (in-degree) and the number of players that are sent from team u to all teams v (out-degree). Consequently, $W(u)$ expresses the total number of transfers team u has engaged in. A higher value of $W_i(u)$, $W_o(u)$ and $W(u)$ indicates that the team has, respectively, bought more players, sold more players or both. These variables thus indicate the engagement of team u in the football transfer network in terms of trades made.

Node Attributes

To account for any structural differences between leagues, we incorporate the league L in which team u plays as a node attribute. This is denoted as L_{CO}. Here, league L is the first tier of country CO — indicated by the country codes as exhibited in Table 2.1. League L_{CO} is the set of teams that have featured in the first tier of country CO. So, for example, if team u has featured in the first division of Germany, then $u \in L_{DE}$.

3.2.3 Graph Measures

This section introduces a set of relevant graph measures. These measures will aid in interpreting and defining the network. In contrast with the variables introduced in the previous section, these measures consider the entire network instead of a single node. The measures that will be used for analysis are *average degree*; *average weighted degree*; *average path length*; *the diameter*; *graph density*; *weakly connected component*; *strongly connected component*; *modularity*; and *average*

clustering coefficient. These definitions are derived from [New18] and [Bar16].

Firstly, we define the average degree $\overline{D}$. This measure displays the average degree of all nodes u as defined in Equation 3.4. The definition of average degree is therefore:

$$\overline{D} = \frac{1}{n} \sum_{u \in V} D(u) \tag{3.6}$$

In this research, the average degree entails the average number of trading partners of each team in the network.

Secondly, we use the average weighted degree $\overline{W}$. The average weighted degree is the average of the weighted degrees $W(u)$ (see Equation 3.5c) of all nodes.

$$\overline{W} = \frac{1}{n} \sum_{u \in V} W(u) \tag{3.7}$$

In this context, the average weighted degree denotes the average number of transactions a team has made.

As a measure of connectivity throughout the network, we define the average path length $\overline{d}$. A path is an ordered list of edges (pairs of nodes) through which node u_0 and u_k are connected. Path P is defined as $P = \{(u_0, u_1), (u_1, u_2), ..., (u_{k-1}, u_k)\}$. The length of this path is then k. This coincides with the number of edges that are crossed. The shortest path between nodes u_0 and u_k is the path with the smallest length. Length k of the shortest path between nodes u and node v is also called the distance $d(u, v)$. The average path length takes the average distance d from all nodes u to all nodes v into account. Average path length $\overline{d}$ is expressed as:

$$\overline{d} = \sum_{u,v \in V} d(u, v) \tag{3.8}$$

The average distance, or path length, between nodes is indicative for the degree of connectivity within the network. A lower $\overline{d}$ indicates a more connected network.

On this basis, we define the diameter d_{max}. The diameter is the longest distance (shortest path) in the graph. The diameter also gives an indication of the connectivity of the network. It gives the distance of the two nodes (i.e., teams) that are furthest away from each other. A lower diameter thus means a more connected network.

Subsequently, we introduce the density ρ of a network. The density of a network forms a ratio between the total possible edges m_{max} of a, in this case, directed network and the edges actually present m.

$$m_{max} = n(n-1) \tag{3.9a}$$

$$\rho = \frac{m}{m_{max}} \tag{3.9b}$$

The density ρ implies how dense the network is connected as a whole as compared to what is maximally possible. We often see that real-world networks are sparse.

A weakly connected component is a subset of nodes that is maximal in size in which all nodes can be connected via a path where direction of the edge is disregarded. The weakly connected

component gives an indication whether a network is fully connected or the network is fragmented into non-connected sub-graphs (components) [EK10].

A strongly connected component is a subset of nodes which is also maximal in size in which all nodes can be connected via a path where the direction of the edges is taken into account. The largest strongly connected component (in terms of nodes included) is called the giant component.

A concept to distinguish the connectedness of parts of the graph is communities. A community is defined as a subset of nodes that are more strongly connected to each other relative to the rest of the network. Modularity maximization algorithms are a method for the detection of communities within a network. The modularity score Q is a tool with which the performance of a modularity maximization algorithm is measured. The modularity score is determined by establishing a subset of nodes for which the number of links is higher than expected. The calculation for the modularity score used in this research can be found in the article written by Blondel et al [BGLL08].

Lastly, we define the average clustering coefficient $\overline{CC}$. This is the average of the clustering coefficients $CC(u)$ of all nodes. The clustering coefficient of a single node is defined by Equation 3.3. The average clustering coefficient is then calculated as follows:

$$\overline{CC} = \frac{1}{n} \sum_{u \in V} CC(u) \tag{3.10}$$

This metric gives insight into the overall connectedness of nodes on a local level. It displays the average embeddedness of a node in the network.

3.2.4 League Performance Measure

This research will focus on relating a league performance variable to the network position of a team. The intended meaning of the league performance measure $P(u)$ is to display the dominance of team u in their respective league L_{CO} in a way that the measure is comparable among leagues. It should be noted that while the proposed measure is comparable among leagues in terms of dominance within a certain league, the measure does not account for cross-league comparisons of teams directly. An illustrative example derived from Table 3.2 is Real Madrid that has, on average, been as dominant in the Spanish La Liga as PSV Eindhoven has been in the Dutch Eredivisie.

Although viable methods for comparing the relative strengths of leagues exist [VLG+18], these methods are disregarded as we explore the effects on domestic league performance. These findings do pose as valuable comparative data for assessing the robustness of the constructed league performance measure in this research. A comparison as such is displayed in Table 3.1.

Furthermore, methods for rating and ranking teams, such as Colley's method or Elo's system — as set forth extensively by Langville and Meyer [LM12], are not suitable for this specific research.

To construct a robust league performance measure, the first step is to quantify a team's performance in any given year. The most applicable and relevant way to do so, is to use the rank in the final standings of a team. This score is then inversed relative to the number of clubs in the competition. For example, if Team A finishes first (1^{st}) out of twenty (20) clubs, Team A's score for that specific season will be 20. The score R of team u in season i is therefore calculated as:

$$R(u, i) = (X_{u_i})^{-1} \tag{3.11}$$

Here, X_{u_i} denotes the final rank X that team u obtained in year i. Note that a lower rank implies better performance. Therefore, we take the inverse of this number to obtain an increasing score for teams ending higher in the standings — which is denoted by a lower rank. Due to the dynamic nature of domestic club competitions (as a result of promotion and relegation), not all clubs will appear in one specific league for all seasons. Relegation to a lower league indicates worse performance. Therefore, $R(u, i) = 0$ when the team is playing in a lower league during season i.

Leagues within the data set have varying sizes. For this reason, R is multiplied with a constant c that accounts for this differing ratio. To normalise these differences, the assigned points will be standardized to a format of sixteen teams. The team that finished in last place will always score one point — regardless of the league. This means that teams within every league can gain at most sixteen (16) points and teams finishing in last place will score one (1) point. The distribution between these numbers differs per league size and corresponding factor. Constant c is the same for all teams in the same league and is expressed as follows:

$$c(L_{CO}) = \frac{16}{s_{L_{CO}}} \tag{3.12}$$

Constant c is dependent on league L_{CO} in which team u is included. The variable $s_{L_{CO}}$ expresses the size of league L in which team u plays.

The total score R of team u over the years that are available in the data set is simply the sum of the yearly acquired scores. This leads to the following notation:

$$R(u) = \sum_{i=1}^{m_{L_{CO}}} \max\{R(u, i) \cdot c(L_{CO}), 1\} \tag{3.13}$$

Variable $m_{L_{CO}}$ denotes the total number of years that have been taken into consideration for league L_{CO} in which team u plays. Finally, as discussed in Section 2.2, the leagues also differ in the number of years $m_{L_{CO}}$ included in the data set. To account for the differing years used to calculate $P(u)$, the score is normalized using $m_{L_{CO}}$. Combining the above, $P(u)$ is determined in the following way:

$$P(u) = \frac{R(u)}{m_{L_{CO}}} \tag{3.14}$$

In other words, a season is seen as a sample of 34 to 38 games collectively indicating a team's performance in that season. Total performance is determined by summing the individual seasons' performances while normalising for differences among leagues. This method takes the Regression-to-the-mean effects into account that are present when measuring a team's performance over a yearly basis. This builds on the notions about Regression-to-the-mean effects apparent within football as described and implemented by [BM11].

Furthermore, only teams present in the eight leagues in at least one of the included seasons will be used in the research. For reference, these seasons can be found in Table 2.1. The research will therefore only focus on the league performance of teams when — and only when — having appeared in the aforementioned eight leagues and, thus $P(u) > 0$.

As mentioned in the introduction of this section, research by Vales et al [VLG+18] poses as a good comparative instrument of the league performance measure. Table 3.1 shows that the findings of this thesis resemble the research closely. The distribution of P gives an insight in the

differing competitive profiles of the leagues which have also been observed by Vales et al. A lower standard deviation implies that teams within that sample (i.e., league) obtain a score that is, on average, closer to each other. A lower standard deviation within a league thus suggests a more competitive character as teams performances are relatively similar. The Dutch Eredivisie and the Russian Premier Liga were not incorporated in the research of Vales et al. Our findings suggest that these leagues have the highest deviations of league performance, thus implying they are the least competitive. Significant differences in the means are apparent between the Dutch Eredivisie and the Italian Serie A at the 5% level. Figure 3.1 exhibits these differences in the form of a box plot.

The distribution of the league performance indicator can be found in Figure 3.2. We see that it follows a right-tailed distribution indicating that, as expected, in-league dominance over the years is a feat reserved for only a small group of elite teams. The ten leading teams in this group of elite are displayed and ordered in Table 3.2.

League L_{CO}	Obs	Mean	Std. Dev.	Rank	Rank in [VLG+18]
English Premier League L_{GB}	49	3.48	4.10	3	3
French Ligue 1 L_{FR}	44	3.70	3.91	6	6
German 1.Bundesliga L_{DU}	42	3.62	4.06	4	5
Italian Serie A L_{IT}	51	3.06	4.01	5	4
Spanish La Liga L_{ES}	48	3.56	4.17	1	1
Portugese Liga NOS L_{PT}	43	3.32	4.10	2	2
Dutch Eredivisie L_{NL}	30	5.07	4.66	(1)	n/a
Russian Premier Liga L_{RU}	27	4.55	4.37	(2)	n/a
Total	334	3.68	4.14	-	-

Table 3.1: The competitive profiles of the leagues indicated by the distribution of P and a comparison of the results to research of Vales et al [VLG+18].

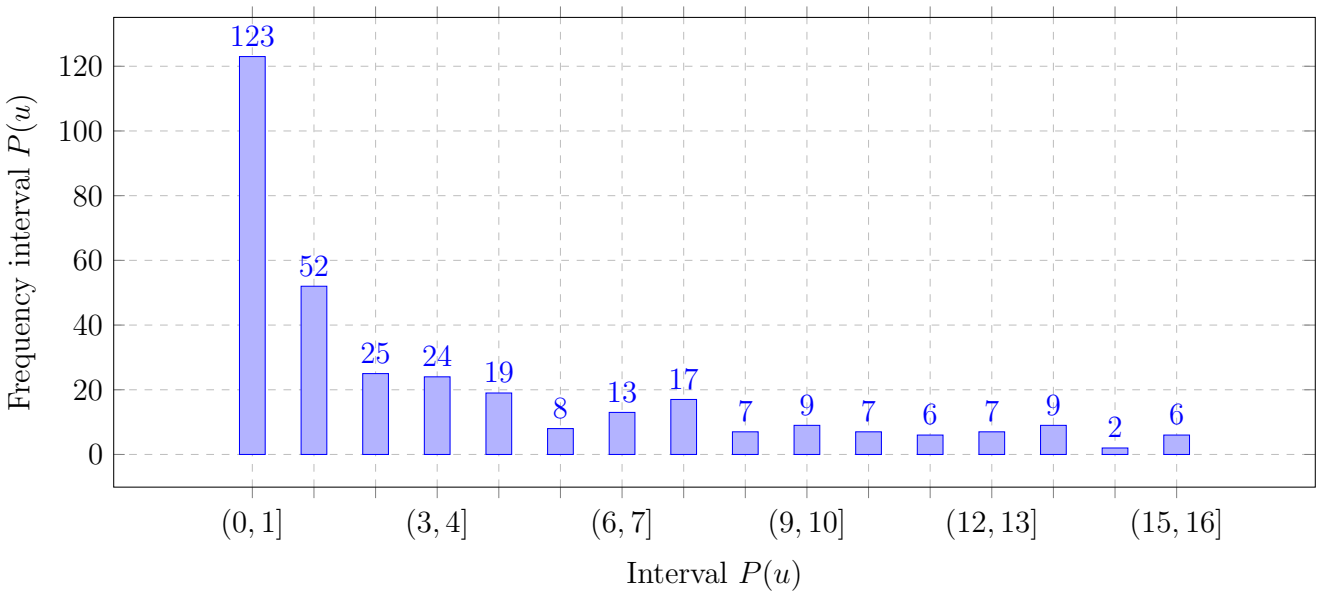


Figure 3.2: Distribution of the league performance measure $P(u)$.

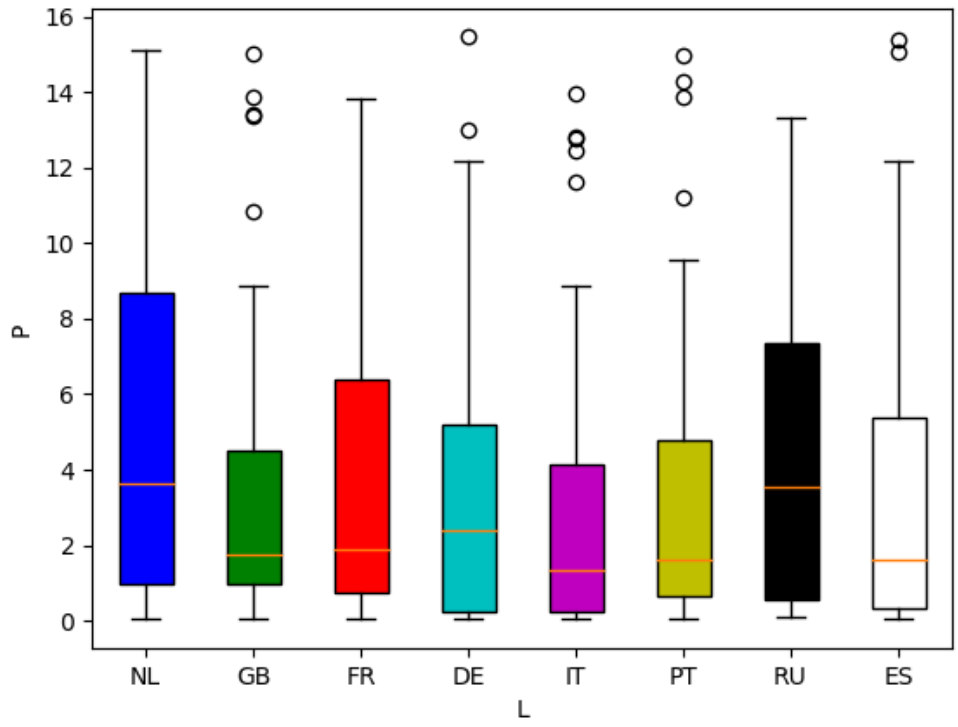


Figure 3.1: Box plots of the relation between leagues L_{CO} and the league performance measure P.

Rank	Team u	$P(u)$	League L_{CO}
1	Bayern Munich	15.46	German 1.Bundesliga
2	FC Barcelona	15.37	Spanish La Liga
3	Ajax Amsterdam	15.11	Dutch Eredivisie
4	Real Madrid	15.06	Spanish La Liga
5	PSV Eindhoven	15.05	Dutch Eredivisie
6	Manchester United	15.03	English Premier League
7	FC Porto	14.98	Portugese Liga NOS
8	SL Benfica	14.29	Portugese Liga NOS
9	Juventus FC	13.94	Italian Serie A
10	Sporting CP	13.89	Portugese Liga NOS

Table 3.2: The top ten most dominant teams in their respective leagues according to league performance measure $P(u)$ which is rounded for interpretability.

Chapter 4

Approach

This chapter will explain the approach taken to conduct the research. Firstly, by means of a definition of the linear regression model. Secondly, by introducing the setup of the experiment.

4.1 Linear Regression Model

This section proposes the linear regression model that is used for answering the second research question.

Note that we try to describe league performance of a team $P(u)$ on the basis of its position and engagement in the football transfer network. The position of a team is expressed by a combination of $B(u)$, $C(u)$ and $CC(u)$. The engagement of a team is expressed by $W_o(u)$ and $W_i(u)$ (see Section 3.2.2). Thus, we can say that response variable $Y = P(u)$. The explanatory variables $X_1, X_2, ..., X_5$ are respectively $B(u)$, $C(u)$, $CC(u)$, $W_i(u)$ and $W_o(u)$. A complete overview of the variables in the model and the associated node measures can be found in Table 4.1. This table also includes the signs that are expected based on existing literature.

This results in the following multiple regression equation:

$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \tag{4.1}$$

In this notation, α is the intercept and β_i is the coefficient belonging to the explanatory variable X_i. As the explanatory variables do not share a comparable scale, the coefficients are standardized. This enables us to compare the coefficients of variables on an equivalent scale.

Table 4.2 exhibits the variables and relevant descriptive statistics that are used for the regression models.

Variable in model	Node measure	Expected sign
X_1	Betweenness centrality B	$-$
X_2	Closeness centrality C	$+$
X_3	Clustering coefficient CC	$+$
X_4	Weighted out-degree W_o	$+$
X_5	Weighted in-degree W_i	$-$

Table 4.1: An overview of the explanatory variables.

Variable	Obs	Mean	Std. Dev.	Min.	Max.
League performance measure P	334	3.68	4.14	0.04	15.46
Betweenness centrality B	334	0.00348	0.00334	0.00001	0.02218
Closeness centrality C	334	0.378	0.0294	0.287	0.455
Clustering coefficient CC	334	0.176	0.113	0.030	0.569
Weighted in-degree W_i	334	191.04	102.49	17	638
Weigted out-degree W_o	334	198.25	177.63	13	1,067
Weighted total degree W	334	389.29	273.33	30	1,705

Table 4.2: Descriptive statistics over the teams used for the linear regression models.

4.1.1 Controlling for League

As demonstrated in Table 3.1 and documented by Vales et al [VLG+18], there exist differences among leagues in terms of competitive profiling. The competitive profile of a league denotes the distribution of relative strength within a league. If a league contains more teams that are evenly matched, the league is described to be more competitive. The composition of a league is therefore an important factor that determines the ability of a team to dominate (i.e., perform) in its league. We account for this by including the league L_{CO} of team u as a control variable. By controlling this variable, we compare teams within the same leagues. To use this categorical variable, we have opted to use dummy coding (see Table 4.3). The reference level is the Italian Serie A. This league was chosen as this league included the most observations.

League L_{CO} / Dummy variable	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	Obs.
Dutch Eredivisie L_{NL}	1	0	0	0	0	0	0	30
English Premier League L_{GB}	0	1	0	0	0	0	0	49
French Ligue 1 L_{FR}	0	0	1	0	0	0	0	44
German 1.Bundesliga L_{DU}	0	0	0	1	0	0	0	42
Portugese Liga NOS L_{PT}	0	0	0	0	1	0	0	43
Russian Premier Liga L_{RU}	0	0	0	0	0	1	0	27
Spanish La Liga L_{ES}	0	0	0	0	0	0	1	48
Italian Serie A L_{IT}	0	0	0	0	0	0	0	51

Table 4.3: Dummy coding for league L_{CO} as categorical variable.

4.1.2 Controlling for Degree

Mourao [Mou16] describes the significant impact a team's league performance has on the number of trading partners (degree). A successful team (in terms of league performance) is likely to experience an increase in financial resources as a result of the received prize money and more lucrative sponsorship deals. Besides, a better performance leads to the team's players becoming more attractive for other team that often are willing to pay more [Fri07]. Overall, this substantiates

the likelihood of a strong positive correlation between degree and league performance — which is also implied by the analysis in Section 5.2. Therefore, the number of trading partners of a team (degree) is controlled in the linear regression models.

4.2 Experimental Setup

This section introduces the setup of the experiment. This includes the steps taken after preprocessing the data such that the experiment could be conducted.

In Section 2.1 is detailed how the data was acquired and how the data was processed before analysis. After preprocessing the data, we used the software Gephi (https://gephi.org/) for its visualisation tools as used for Figure 5.1, the metrics as displayed introduced in Section 3.2.3 and the calculation of the node attributes set forth in Section 3.2.2. The latter were imported after calculation. Subsequently, the league performance measures were calculated for all teams through the method described in Section 3.2.4. The code used is available upon request.

Once all variables were calculated and stored in two separate files — one containing graph measures and one containing the league performance measure — these files were merged. The data set with graph measures was combined with the data set that contained the league performance variable. This enabled us to create and test the model as proposed in Section 4.1 by using the *pandas*-, *statsmodels*- and *patsy*-packages of Python (https://pandas.pydata.org/, https://www.statsmodels.org/stable/index.html, https://patsy.readthedocs.io/en/latest/). The code used is available upon request. This script is also used to generate the scatter plots that are displayed in Figure 5.4.

Chapter 5

Results

This chapter provides the results of the data analysis that is at the base of describing the characteristics to answer the first research question. Furthermore, this chapter contains the results of the linear regression model as proposed in Section 4.1 that is composed to answer the second research question.

5.1 Network Descriptives

In this section, we interpret the relevant network-variables (Table 5.1); a visual representation of the network (Figure 5.1); and relevant distributions (Figures 5.2 and 5.3). By combining and engaging with these, we provide an interpretation of the characteristics of the European football transfer market. This section builds on the definition of the network as described in Section 3.2.1. This means that, in the network, the teams are represented by nodes and the transactions between teams are represented by edges. The metrics that are presented in Table 5.1 and used in the text below are defined in Section 3.2.3. Furthermore, it uses node-variables as set forth in Section 3.2.2.

Metric	Value
Nodes n	4,876
Edges m	46,079
Average degree $\overline{D}$	18.9
Average weighted degree $\overline{W}$	37.5
Average path length $\overline{d}$	3.31
Diameter d_{max}	6
Graph density ρ	0.002
Weakly connected component(s)	1
Strongly connected component(s)	2,490
Nodes in giant component n_g	2,386
Modularity Q with Resolution of 1.0	0.596
Number of communities	8
Average clustering coefficient $\overline{CC}$	0.308

Table 5.1: An overview of network metrics of the European football transfer network.

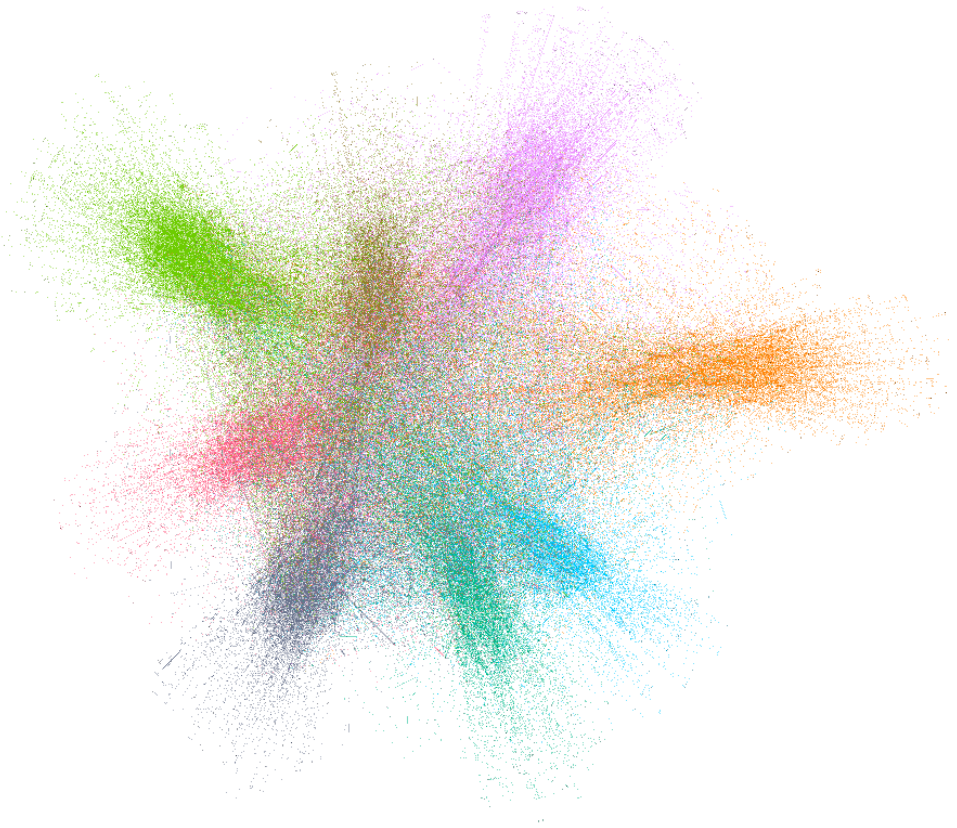


Figure 5.1: A visualisation of the transfer network coloured by community.

Legend:

Purple	: Portugese Liga NOS, 18.91%	Light green	: Italian Serie A, 15.91%
Orange	: Russian Premier Liga, 14.95%	Blue	: Dutch Eredivisie, 11.77%
Dark green	: German 1.Bundesliga, 10.6%	Red	: French Ligue 1, 10.15%
Grey	: English Premier League, 9.89%	Gold	: Spanish La Liga, 7.81%

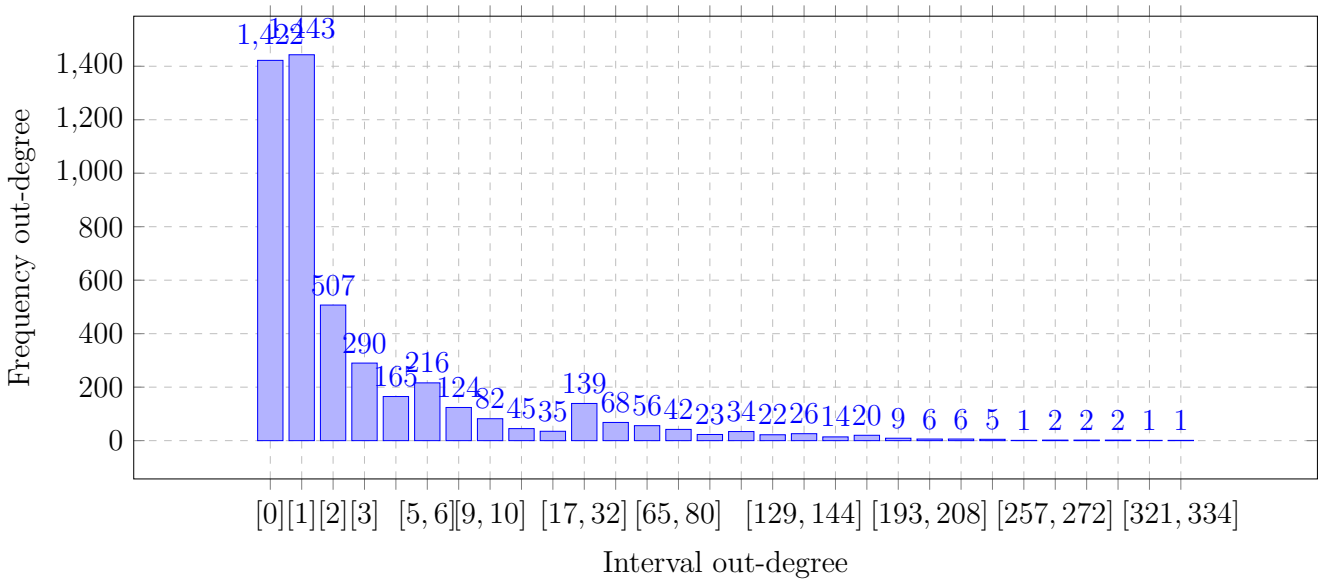


Figure 5.2: Distribution of out-degree.

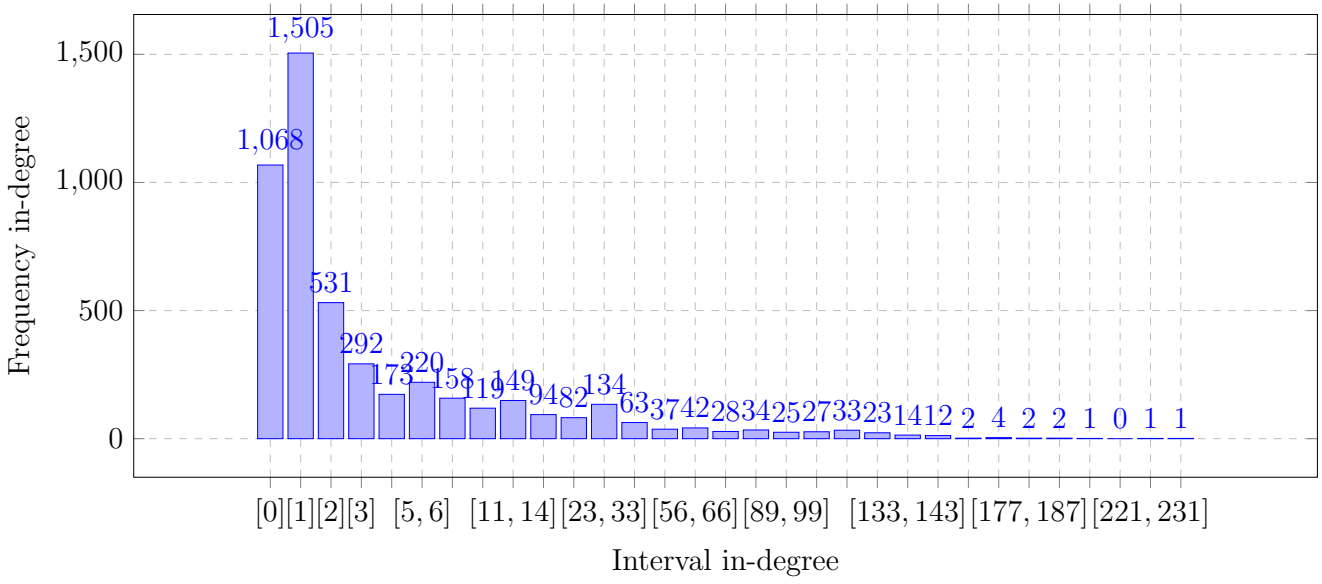


Figure 5.3: Distribution of in-degree.

From the above figures and table, we can deduce that the European football transfer network possesses the following characteristics that mark other real-world social networks:

- It is a sparse graph as it has a very low connectedness in terms of individual links between nodes compared to the possible links in the network — as expressed by the density. This reinforces the notion that trading players happens on a non-incidental and non-random basis.

- From Figures 5.2 and 5.3 follows that for both out-degree and in-degree, and thus total degree, the distribution is right-tailed. This indicates that the majority of nodes are lowly linked and there exists only a small group of 'hubs'. Hubs are defined by "nodes with a huge number of links" [Bar16]. In this thesis, hubs are defined as the ten nodes with the highest degree. In Table 5.2, these teams are displayed. The presence of hubs in random-networks is an exception rather than the rule that it is within social networks.

  Intriguingly, hubs only appear in either the Italian Serie A or in the Portugese Liga NOS. The presence of hubs in these leagues is expected to relate to the communities that are associated with both leagues being the largest. However, an explanation for this finding is unclear.

- All nodes belong to one weakly connected component. This shows the connected nature of the transfer network. The connected nature of the network is underlined by the fact that the giant component consists of 2,386 teams which amounts to almost 50% of all nodes.

  Interestingly, we see that teams form a community with teams that appear in the same league. Teams that do not appear in one of the eight leagues participate in the community they trade most often with. Communities are subgraphs that are locally dense connected [Bar16]. In other words, teams tend to trade more with teams that are in the same league. In fact, league as a categorical variable and community as a categorical variable match one-on-one. This is illustrated in Figure 5.1. The ranking of most frequent transfers between two teams serves as evidence for the apparent preference for trading with familiar teams. The top of this ranking is overwhelmingly dominated by transfers from an affiliated team to its first team.

- Another indication of the connected nature that is typical of real-world networks is the low pairwise node-to-node distance of the network. The average distance between nodes is 3.31 which means that all clubs are, on average, separated by 3.31 transfers.

  This corresponds to the "small-world" effect present in real-world networks. The small-world phenomenon builds on the six degrees of separation notion which states that in real-world networks, on average, any pair of nodes can be connected through a path of length six [Bar16]. In fact, in this context, *any* pair of teams can be connected through at most six transactions underlined by the diameter of six.

- Real-world networks tend to have more triangles than expected on a random basis. We can see that the average clustering coefficient $\overline{CC}$ is 0.308 which is about 159 times higher than expected ($\overline{CC_e} = \frac{\overline{D}}{n}$). These findings are in line with the literature written about the real-world difference between $\overline{CC}$ and $\overline{CC_e}$ [Bar16].

This collectively shows that the European football transfer market as represented as a network is centered around hubs. These hubs are characterized by a high degree of incoming and outgoing transactions.

Furthermore, the transfer market is divided into communities (i.e., leagues) and clusters. This displays the tendency of teams to trade with a familiar set of teams — often, linked to their respective league. On the contrary, teams that do not seem to share the same league are less likely to trade with each other and this results in a sparse network. We can therefore pose that transfers within communities are abundant, however, transfers between communities are relatively scarce.

The general exception to this are the hubs. The high degree of hubs aids in connecting the different communities. This is underlined by the apparent correlation between degree and the tendency of being located on shortest paths between nodes (degree D and betweenness centrality B share $r^2 = 0.779$). Hubs are therefore vital in minimizing the distance between nodes and thus, in connecting the teams in the transfer network.

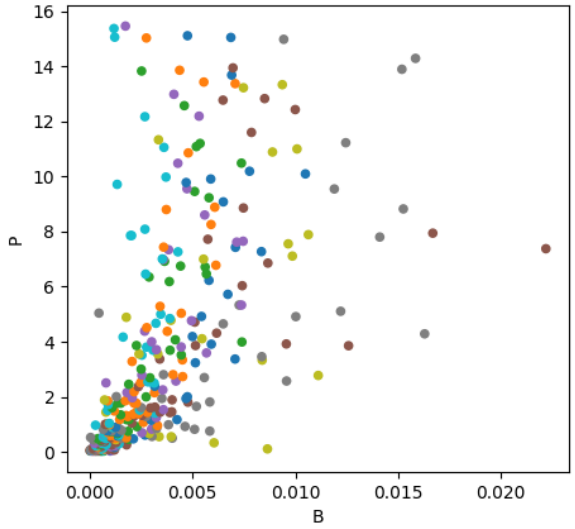
Rank	Team u	$D(u)$	League L_{CO}
1	Udinese Calcio	557	Italian Serie A L_{IT}
2	Parma Calcio 1913	539	Italian Serie A L_{IT}
3	SL Benfica	503	Portugese Liga NOS L_{PT}
4	Sporting CP	463	Portugese Liga NOS L_{PT}
5	Genoa CFC	442	Italian Serie A L_{IT}
6	Inter Milan	437	Italian Serie A L_{IT}
7	AS Roma	417	Italian Serie A L_{IT}
8	Vitória Setúbal FC	407	Portugese Liga NOS L_{PT}
9	Juventus FC	405	Italian Serie A L_{IT}
10	SC Braga	395	Portugese Liga NOS L_{PT}

Table 5.2: Hubs in the European transfer network.

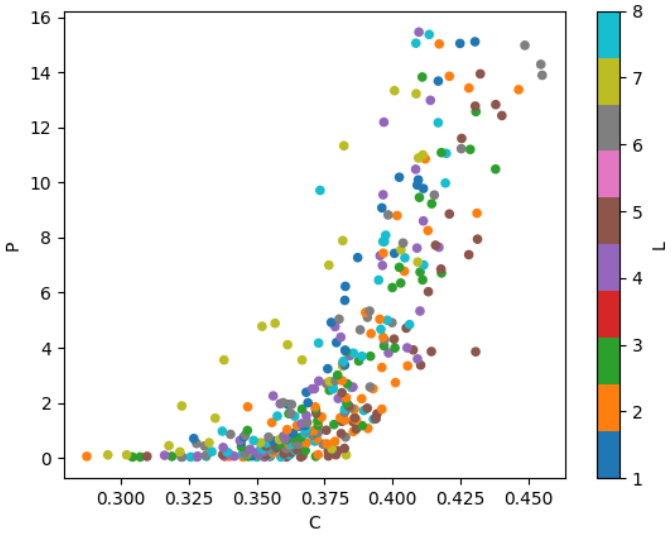
5.2 Network-driven Modelling of Team Performance

This section presents the results of the relations between the node measures and the league performance measures. Furthermore, this section provides various linear regression models that seek to answer the second research question. This is structured by first presenting the scatter plots in Figure 5.4 of the relations of the individual variables with league performance. Thereafter, the results of multiple models that predict league performance on the basis of the team's network position and engagement are exhibited in Table 5.3. The models that are displayed in this Table have standardized coefficients so that comparison of coefficients between attributes is simplified. We will then proceed to interpret these findings in the context of the second research question. The results are summarized in Table 5.4. The correlation-matrix of all variables is displayed in Table 5.5.

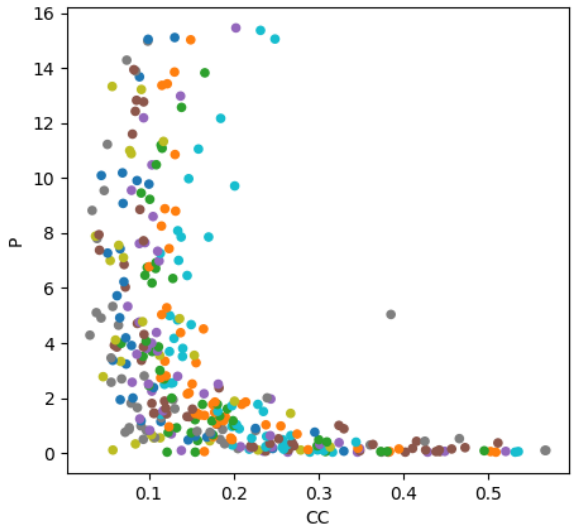
This section uses the variables (node attributes) as proposed in Section 3.2.2 and the definition of the linear regression model as introduced in Section 4.1. Recall that variables $X_{NL}, ..., X_{ES}$ are dummy variables for leagues of which an overview can be found in Table 4.3.



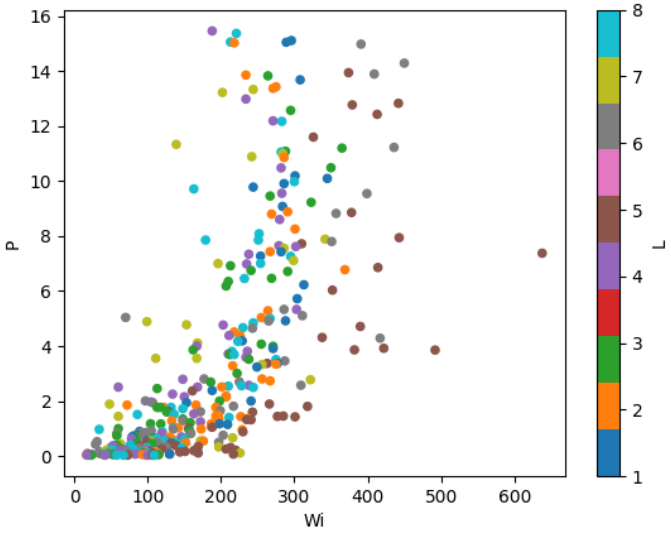
(a) Scatter plot of the relation between betweenness centrality B and league performance P.



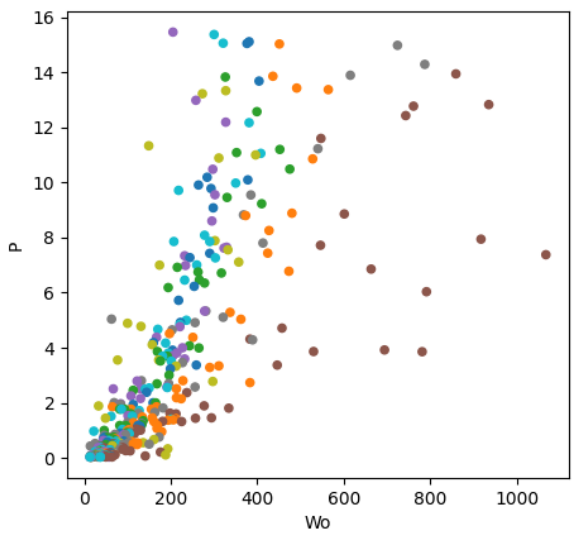
(b) Scatter plot of the relation between closeness centrality C and league performance P.



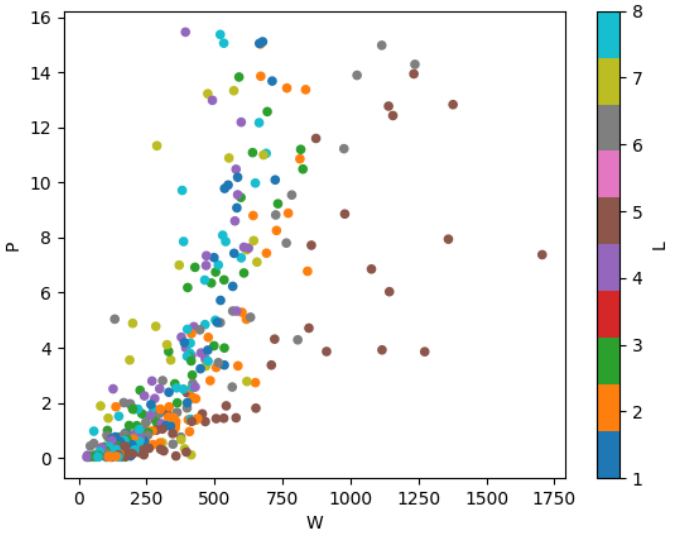
(c) Scatter plot of the relation between clustering coefficient CC and league performance P.



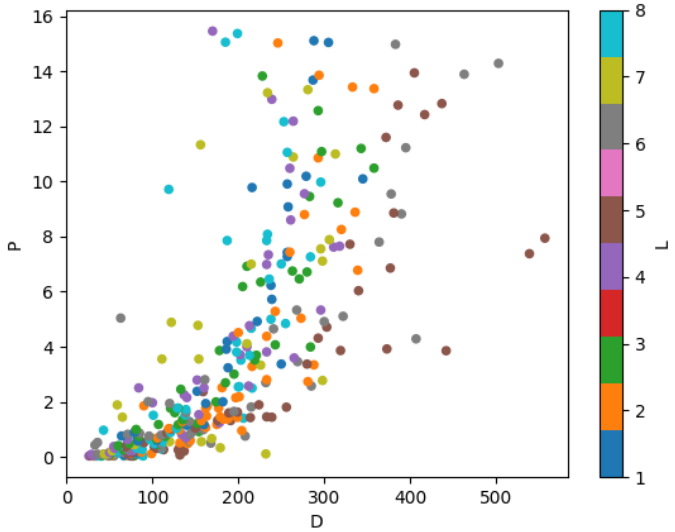
(d) Scatter plot of the relation between weighted in-degree W_i and league performance P.



(e) Scatter plot of the relation between weighted out-degree W_o and league performance P.



(f) Scatter plot of the relation between weighted degree W and league performance P.



(g) Scatter plot of the relation between degree D and league performance P.

Figure 5.4: Scatter plots of the relations between the explanatory variables and the league performance measure.
Legend:
1: L_{NL} 2: L_{GB} 3: L_{FR} 4: L_{DU}
5: L_{IT} 6: L_{PT} 7: L_{RU} 8: L_{ES}

Variable								
Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
B	0.12**	0.07		0.11*	−0.19***	−0.38***	−0.31***	−0.71***
C	0.72***	0.89***		0.69***	0.66***	0.52***		
CC	0.02	0.16***		−0.04	0.14***	0.19***	0.09*	0.35***
W_i			0.04	−0.49***	−0.21**			
W_o			0.72***	0.50***	0.71***			
W						0.84***		
D							1.10***	1.67***
L_{NL}		0.85***			1.25***	1.22***		1.33***
L_{GB}		0.31***			0.43***	0.52***		0.26**
L_{FR}		0.47***			0.73***	0.92***		0.76***
L_{DE}		0.68***			0.96***	1.06***		0.94***
L_{PT}		0.44***			0.83***	0.98***		0.88***
L_{RU}		1.19***			1.49***	1.52***		1.47***
L_{ES}		0.56***			0.78***	0.87***		0.65***
α	0.00	−0.50	0.00	0.00	−0.79	−0.81	0.00	−0.70
r^2	0.626	0.703	0.564	0.674	0.779	0.763	0.595	0.717
Obs.	334	334	334	334	334	334	334	334
Note:	$^*p < 0.1$	$^{**}p < 0.05$	$^{***}p < 0.01$					

Table 5.3: Results of multiple linear regression models.

In Table 5.3, the columns denote different linear regression models. The rows introduce the variables. When a variable is used in a model, the (rounded) coefficient of said variable is displayed in the corresponding cell. The number of asterisks next to the coefficient denote at what level the result is significant. As an example, in Model 1, the node attribute betweenness centrality B is used and its coefficient $\beta_B = \beta_1 = 0.122$. This is statistically significant at the 5% level. An elaboration on what variables are included in what model can be found in the enumerated list below:

1. Model 1 uses all network position variables.
2. Model 2 uses all network position variables and accounts for the league in which the team plays.
3. Model 3 uses W_i and W_o as network engagement variables.
4. Model 4 uses all network position variables and W_i and W_o as network engagement variables.
5. Model 5 uses all network position variables and W_i and W_o as network engagement variables. Furthermore, this model accounts for the league in which a team plays.
6. Model 6 uses all network position variables and W as network engagement variable. Furthermore, this model accounts for the league in which a team plays.

7. Model 7 includes degree D as a control variable for B and CC.

8. Model 8 includes degree D as a control variable for B and CC. Furthermore, this model accounts for the league in which a team plays.

The relations between the individual variables and the league performance measure as displayed in Figure 5.4 suggest that a wider reach within the network seems to be positively correlated to better performance. A wider reach here is defined by three characteristics. Firstly, a wider reach in terms of a more central position in the network (higher B and C, Figures 5.4a and 5.4b). Secondly, a wider reach that expands beyond a team's neighbourhood (lower CC, Figure 5.4c). Lastly, a wider reach in terms of being more active in trading players (higher W_i, W_o and W, Figures 5.4d, 5.4e and 5.4f).

An analysis of the models that incorporate the set of variables that define a team's position (B, C and CC, Model 1) or a team's engagement (W_i and W_o, Model 3) seems to confirm this notion – with the exception of CC. However, the results of CC and W_i in these models are not significant at the 10% level.

When accounting for both the position and engagement of a team in the network (Model 4), we again find that a more central position within the network (a higher C) is correlated to a better league performance, but also in this model we find that the (seemingly negatively correlated) result of CC is not significant. Interestingly, we can observe in Model 4 that, if a team is similarly embedded in the network, weighted in-degree correlates negatively to league performance. This may be explained by a difference in the loaning behaviour of teams as documented by Pantuso and Hvattum [PH21]. Lower and mid-tier teams are often constrained by a budget which is why loaning a player is more alluring. Expressing engagement by W_i and W_o individually leads to a more accurate model, but combining them (W) shows the significant positive correlation that the overall engagement of a team has with league performance P. When comparing teams that are similarly positioned, the engagement of a team is positively correlated to league performance.

A shift in the apparent correlation between B and P can be found in Model 4. Furthermore, an apparent shift in correlation between CC and P is also observable in Models 1 and 4. These results are not significant, but might suggest a tainted representation in Figures 5.4a and 5.4c as there might exist another variable that causes these correlations to seem negative and positive respectively. A potential variable that might be at the root of this misrepresentation is degree D as degree has an influence on the construction of most. Degree correlates negatively with clustering coefficient, significant at the 1% level. This is sensible as a higher degree means that more neighbouring nodes of node u will have to share an edge (trade) with each other which, by means of chance, is less likely. Furthermore, degree correlates positively with betweenness centrality, also significant at the 1% level. As degree itself is significantly and strongly correlated to P (see Figure 5.4g), its association with B and CC influences the relation with P when not accounting for D. We find that, when accounting for D, such as in Model 7, the coefficients of CC and B flip signs. The high correlation of degree with closeness centrality and weighted (in-, out- and total) degree (see Table 5.5) may explain why the relations of these two variables seems to differ from the scatter plots in Models 1 an 4.

When controlling for league, we observe a similar result. We find that, when comparing teams within the same league — and thus, the same competitive profile — a central role in terms of closeness centrality C and an active approach within the transfer market in terms of weighted

out-degree W_o and weighted total degree W remain positively correlated to league performance. In Models 2, 5, 6 and 8, we find that, when controlling for league, embeddeness within a team's neighbourhood — clustering coefficient CC — is positively correlated to league performance. This suggests that significant differences of clustering coefficients among leagues exist. This is confirmed by Figure 5.5 and by further examination of the relation between league and clustering coefficient in Table 5.6. Significant differences at the 1% level are observable between the Dutch Eredivisie and the English Premier League, Italian Serie A and the Spanish La Liga, and between the Russian Premier Liga and the English Premier League, Italian Serie A and the Spanish La Liga. Here, teams from the Dutch Eredivisie and the Russian Premier League have significantly lower clustering coefficients. Differences in clustering coefficient between leagues become more apparent when controlling for the degree of teams. These findings suggest that some leagues are more inward-oriented (a higher CC) and some leagues are more outward-oriented (a lower CC) in terms of their trading behaviour.

Lastly, when including leagues in the analyses, we can see that the competitiveness of a league, and differences among those as observed in Section 3.2.4, have a significant effect on the ability of a team to dominate (e.g., the team's score on the league performance measure). The inclusion of leagues in Models 2, 5, 6 and 8 increases the accuracy of those models drastically in terms of ability to explain the variance of the league performance measure (as expressed by the r^2).

A model that combines the sets of variables that account for the position of a team and for the team's engagement explains the variance of league performance the best. Furthermore, taking the specific competitive profiles of leagues into account improves the accuracy of the model drastically. Most variables use degree in their computation which, when not accounted for, affects the correlation between the explanatory variables and league performance P as is displayed in Figure 5.4. Degree indirectly influences these variables through C and W. As these variables express a more specific element of a node than degree does, they are able to more accurately describe a team's league performance.

Outcome / Variable	B	C	CC	D	W_i	W_o	W
P	$-$	$+$	$+$	$+$	$-$	$+$	$+$

Table 5.4: Summary of the relations between variables and league performance.

Variable	B	C	CC	D	W_i	W_o	W	L	P
B		0.456	0.455	0.779	0.704	0.619	0.682	0.146	0.351
C	0.456		0.343	0.792	0.748	0.694	0.750	0.047	0.619
CC	0.455	0.343		0.514	0.435	0.319	0.377	0.084	0.232
D	0.779	0.792	0.514		0.916	0.868	0.930	0.040	0.567
W_i	0.704	0.748	0.435	0.916		0.805	0.917	0.089	0.464
W_o	0.619	0.694	0.319	0.868	0.805		0.973	0.096	0.564
W	0.682	0.750	0.377	0.930	0.917	0.973		0.095	0.552
L	0.146	0.047	0.084	0.040	0.089	0.096	0.095		0.019
P	0.351	0.619	0.232	0.567	0.464	0.564	0.552	0.019	

Table 5.5: Correlation matrix of the chosen variables. When the variables share a correlation where $r^2 > 0.7$, the corresponding cells is yellow. When the attributes share a correlation where $r^2 > 0.9$, the corresponding cells are red.

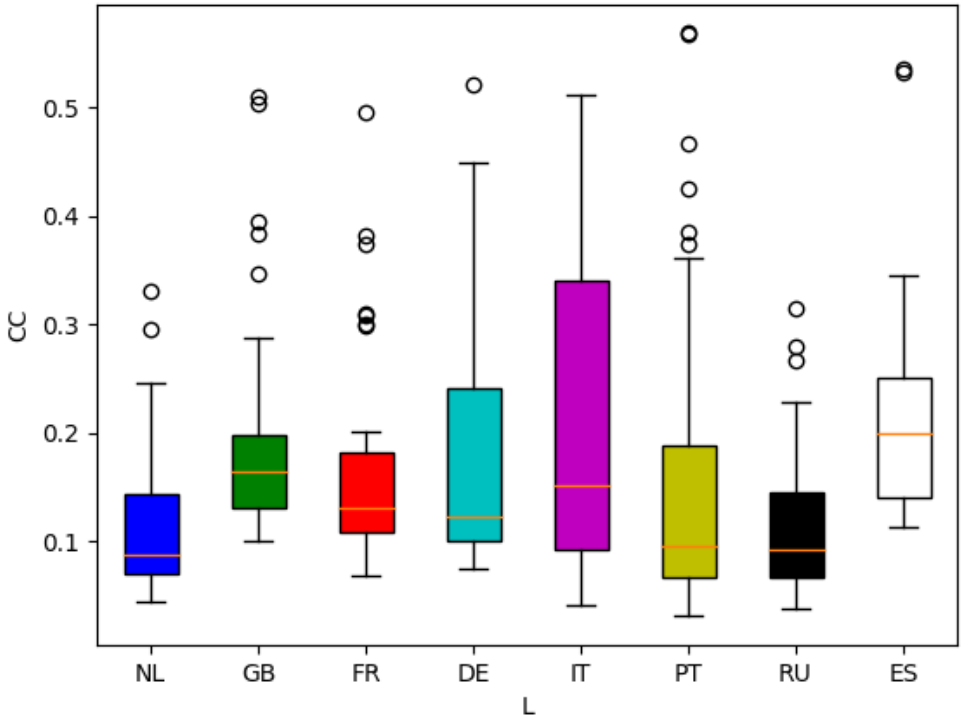


Figure 5.5: Box plots of the relation between leagues L_{CO} and closeness centrality CC.

League L_{CO}	Obs	Mean	Std. Dev.
Dutch Eredivisie L_{NL}	30	0.116	0.072
English Premier League L_{GB}	49	0.190	0.093
French Ligue 1 L_{FR}	44	0.167	0.091
German 1.Bundesliga L_{DU}	42	0.179	0.112
Italian Serie A L_{IT}	51	0.215	0.143
Portugese Liga NOS L_{PT}	43	0.158	0.143
Russian Premier Liga L_{RU}	27	0.120	0.075
Spanish La Liga L_{ES}	48	0.212	0.091
Total	334	0.176	0.113

Table 5.6: Distributions of clustering coefficient CC within the leagues analysed.

5.3 Discussion and Limitations

This section discusses the results from Section 5.1 and 5.2 and provides an overview of the wider context and limitations of this research.

The network presented in Section 5.1 features characteristics that coincide with related literature that is written on real-world social networks [Bar16]. The results suggest that the network is sparse with little interaction between leagues, but plenty of interaction within the distinct leagues. Teams that transcend this trend are the hubs as displayed in Table 5.2.

The findings in Section 5.2 imply that these hubs, teams with a high degree D, overall have a higher league performance P. Degree seems to be strongly and positively linked with a better league performance. This is also implied by research of Mourao [Mou16] and Frick [Fri07]. Furthermore, degree seems to affect the explanatory variables as degree is used in their computation. Accounting for the degree of teams, we see that the sign of betweenness centrality and clustering coefficient flip.

On this basis, we find that, although degree is significantly correlated to league performance, it is not the most accurate in explaining the variance of the league performance measure. More specific variables seem to benefit this. A model that uses both network position variables and network engagement variables seems to be most accurate. Model 5 includes both network position variables (betweenness centrality, closeness centrality, clustering coefficient) as well as network engagement variables (weighted in-degree, weighted out-degree). It furthermore accounts for the league. This model seems to describe the variance of P the best. The distinction between leagues seemed significant as differences in competitive profiles of leagues are observable. This is in line with related literature [VLG+18]. Furthermore, but yet unexplored, considerable differences in the network variables have been found between leagues. However, it is beyond this thesis' scope to map the leagues' trading profiles. Future research into this topic, would allow for a comparison of a league's competitive character and the recruitment of players on a league-wide level.

It can however not be stated that all network position variables contribute in the same direction or to a similar degree. We find that betweenness centrality is negatively correlated when accounting for degree, significant at the 1% level. On the other hand, closeness centrality and clustering coefficient are positively correlated with a better league performance measure. These findings suggest that teams that are closely embedded within the neighbourhood — their league — and possess a central role with regards to the whole network tend to perform better. However,

if a team's position is characterized by being a link between leagues, performance seems to be lower. The reason that is at the basis of the apparent negative correlation between being a 'broker' between groups in the network and league performance is unclear, and further research into this topic is encouraged.

The results of the network engagement variables also provide differing results. We can observe that the number of outgoing players W_o is positively correlated to a better league performance. Conversely, the number of incoming players W_i is negatively linked to P. The underlying reason that may explain the different effects of W_i and W_o is a distinction between transfers and loans. This would match related literature that states that top teams tend to buy the players they desire due to their larger budget. While, on the other hand, lower and mid-tier teams opt for a loan [PH21].

In the direction of network engagement variables, we furthermore see that the total number of transfers a team has engaged in (W) is strongly linked to a better performance. This might relate to the finding that a higher degree is correlated to better performance. Similarly to research into the relation between degree and league performance, further research related to the total number of transfers might benefit from taking a financial indicator, such as income, into account. Implementing a financial indicator to control for the results would allow for a robust interpretation of the effect of the abundant circulation of players as related literature argues in favour of maintaining a stable, non-changing squad to achieve better results [Mou16].

Chapter 6

Conclusion

This thesis has approached the European football transfer market from a network science perspective. In the football transfer network, teams were represented as nodes, and transfers of players between teams were represented as edges. This research first constructed and described the network using relevant network metrics. Subsequently, a football team's position and engagement in the football transfer network both were quantified using a set of node measures. On this basis, the relation between these measures and teams' domestic league performance was explored. The network was constructed by the analysis of 91,451 transfers of players between teams in 28 seasons. These team played in the eight most prominent European professional leagues. Furthermore, domestic league performance of teams was retrieved by analysis of the final standings of these eight leagues over 204 seasons in total. On this basis, the network analysis conducted in this thesis has aided in constructing a bridge between research into the in-game performance of football teams and research into the financial performance of football teams.

We found that the football transfer network shares characteristics that mark other real-world social networks. This is substantiated by the low density of the network which suggests that incidental links are not abundant. Furthermore, the left-skewed distribution of the network in terms of degree implies the presence of a small group of high-degree hubs and the majority of teams being lowly linked. We see that teams cluster with teams from the same league. Teams tend to be embedded and rooted in their direct neighbourhood. There are some teams that transcend this phenomenon and that connect the different communities within the network. Overall, this makes for a connected network on a local level that is bridged through hubs and a dense core. This is in line with the small-world effect observed in other real-world social networks.

Our research revealed that league performance correlates to both a team's position and engagement in the network. We found that, in terms of a team's position, betweenness centrality relates negatively to a team's league performance, but closeness centrality and clustering coefficient are positively correlated to league performance. In terms of engagement, weighted out-degree and total degree are positively linked with league performance while weighted in-degree relates negatively to a team's domestic league results. These results become more apparent when accounting for the number of trading partners and for the league in which the team plays. Moreover, significant differences in the composition and competitiveness of leagues have been observed which corresponds with existing literature. Furthermore, differences in network position and engagement measures between leagues also differ significantly.

These findings imply that teams that are part of a densely connected neighbourhood and

that have a central position in the network, on average, perform better in their domestic league — demonstrated through the closeness centrality and clustering coefficient. However, teams that connect different groups (i.e., leagues) of the network, tend to perform worse — characterized by a high betweenness centrality. In terms of engagement, teams that sell or loan out players more do better in their respective domestic league. Conversely, if a team buys or loans more players, this relation is reversed. This is in line with existing literature which states that lower or mid-tier teams are more likely to loan a player from another team. Based on these findings, a policy that is directed towards long-term financial health such that a team can actively mingle in the transfer market and obtain a central position is expected to increase a team's dominance in their domestic league over time.

This thesis provides a foundation for multiple directions of further research. Firstly, including a financial indicator as a control variable could contribute to our understanding of the relation between a team's position in the football transfer network and its league performance. Another direction of research is a more detailed analysis of the structure of and directionality throughout the network that explores the flow of players through the market (accounted for their age). Such an analysis would broaden our current understanding of the negative relation between betweenness centrality and league performance. Furthermore, this research is built upon an average base of variables over time. Future research into the effects and dynamics of time on these variables seems potent in discovering trends and suggesting long-term policy implications. In addition, an analysis over time could expose trends that have not been captured in this research as the data has been aggregated. Another fruitful direction of research is to increase the scope of this research to different regions and countries is a promising research direction. This would introduce comparative results to examine the robustness of these findings as well as illuminate the possibility of wider applications of the results shared in this thesis. Lastly, replicating these results in the context of a different sport is a direction that helps to increase the understanding of the economy of sports and the differences among them.

Bibliography

[Bar16] Albert-László Barabási. *Network Science*. Cambridge University Press, 2016.

[BGLL08] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):10008–10012, 2008.

[BM11] Nikolaus Beck and Mark Meyer. Modeling team performance: Theoretical and empirical annotations on the analysis of football data. *Empirical Economics*, 43:335–356, 2011.

[DG12] Stephen Dobson and John Goddard. *The Economics of Football*. Cambridge University Press, 2012.

[EK10] David Easley and Jon Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.

[Fri07] Bernd Frick. The football players' labor market: Empirical evidence from the major european leagues. *Scottish Journal of Political Economy*, 54(3):422–446, 2007.

[GMB+20] F. R. Goes, L. A. Meerhoff, M. J. O. Bueno, D. M. Rodrigues, F. A. Moura, M. S. Brink, M. T. Elferink-Gemser, A. J. Knobbe, S. A. Cunha, R. S. Torres, and K. A. P. M. Lemmink. Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science*, 21(4):481–496, 2020.

[Gyi20] András Gyimesi. League ranking mobility affects attendance: Evidence from european soccer leagues. *Journal of Sports Economics*, 21(8):808–828, 2020.

[HW10] Sean Hamil and Geoff Walters. Financial performance in english professional football: 'an inconvenient truth'. *Soccer & Society*, 11(4):354–372, 2010.

[LM12] Amy N. Langville and Carl D. Meyer. *The Science of Rating and Ranking*. Princeton University Press, 2012.

[LZ20] Chunhua Li and Yangqing Zhao. Comparison of goal scoring patterns in "the big five" european football leagues. *Frontiers in Psychology*, 11, 2020.

[Mat03] V.A. Matheson. European football: A survey of the literature. Technical report, Williams College, Department of Economics, 2003.

[Mou16] Paulo Reis Mourao. Soccer transfers, team efficiency and the sports cycle in the most valued european soccer leagues – have european soccer teams been efficient in trading players? *Applied Economics*, 48(56):5513–5524, 2016.

[New18] Mark Newman. *Networks*. Oxford University Press, 2018.

[PH21] G. Pantuso and L.M. Hvattum. Maximizing performance with an eye on the fnances: a chance-constrained model for football transfer market decisions. *TOP*, 29:583—-611, 2021.

[SS97] Stefan Szymanski and Ron Smith. The english football industry: profit, performance and industrial structure. *International Review of Applied Economics*, 11(1):135–153, 1997.

[VLG+18] Angel Vales, Carlos Casal López, Pedro Gómez, Hugo Blanco Pita, and Jaime Serra Olivares. Competitive profile differences between the best-ranked european football championships. *Human Movement*, 18(5):97–105, 2018.