New Player Strategic Connections and Springboard Clubs: Football Network Analysis at a Player and Club Level

Abstract:
This paper examines the global football network at a club and player level in order to gather insights on how a new player can strategically position himself at the start of his professional football career. By examining measure of centrality and conducting network analysis, this paper identifies which connections with other players are the most important for a new player's long-term success and which clubs are most likely to lead a player to the top clubs in the long run. Through conducting this analysis, we also identified top clubs and top players. This analysis was conducted on two networks: one weighted undirected network that shows the relationships between players based on player weight and time on the same team and another directed network between clubs that transfer players.
Key Words: Network Analysis, Measures of centrality, PageRank, Football, Soccer

Katie Wallis, Philip Fang, Aman Jiddewar and Alex Su

Introduction:

Football (also known as "Soccer" in the United States) is a sport famous for its ability to bring people together. As the most watched and played sport worldwide, football has proven its ability to transcend both cultural and political boundaries. With so many different leagues, each consisting of numerous clubs, which are then made up by many different players themselves, the networks of players and clubs are large and complex. While one may initially believe that clubs that exist in various leagues across the globe are isolated, the incredible amount of trading of players that occurs ties clubs and players together into a tightly knit web.

With soccer's popularity ever growing, the standard of new talent joining the industry is ever increasing and a new player's network and training opportunities become more and more imperative. In this paper, we seek to analyze the football network on both a club and player level to see how a player can benefit his career by creating connections and by joining "springboard clubs", clubs whose connections and past behavior indicate the ability to train a new player and launch them upwards in the ranks towards the best clubs. Looking at both player connections and club ties, we built two separate networks. First, a directed unweighted network that shows ties between which clubs trade players with each other. Next, an undirected, weighted network that incorporates information about player value and relationship while accounting for inflation. Using these networks, we were able to identify top clubs and players and then determine a path to get to the top for a new player through connections with players and the use of springboard clubs. Ultimately, we determined that the best springboard clubs were clubs with highest levels of betweenness. These clubs tended to be clubs that ranked highly in less competitive leagues. Our analysis of the player network leads us to the conclusion that the best connections to have are with players that have the shortest "walk" or path to the top players.

Related Work:

The popularity of the sports network analytics has grown in past few years. With this increased popularity, there has been an increased availability of data and faster, more efficient computers which have made an analysis of big social networks possible. The key concepts of network analytics such as PageRank, Closeness, Betweenness, Exponential graph models, etc. have been in the literature for more than a few decades. Moreover, many large clubs have already began using these techniques to make important strategic decisions from players acquisition to game strategies. A very good example for this is "Moneyball: Tracking down how stats win Games!", which was one of its kind in early days of this exciting field of analytics.

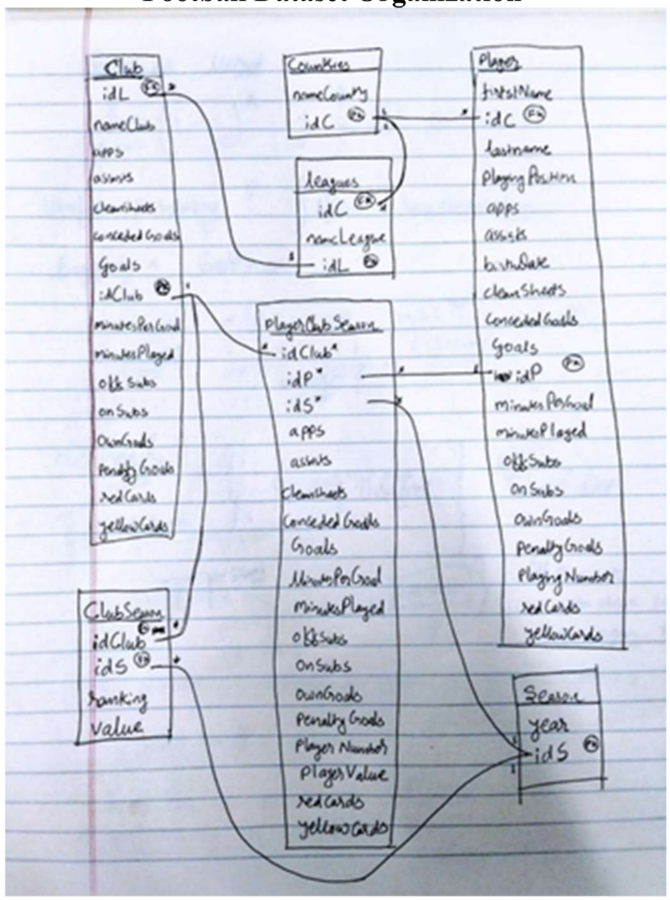
Reference [2], a related publication, discusses the ways in which different network centrality measures can be used in the club and players network. Additionally, reference [3] examines a large tennis player network across several decades to find the best player. Using these publications as guides, we defined connections between players based on the teams they have been together.

There are many additional studies surrounding the subject of football data; however, they tend to focus more on game or club strategy while this report centers around examining the situation from a player perspective. Moreover, this report attempts to answer questions such as "what are springboard clubs?", "who should the player connect for long-term benefits?" and "how important is his country of origin, age, and, playing positions and affects the probability of the network to exist?".

Data:

We obtained the data for the networks from http://www.transfermarkt.co.uk after contacting the person in charge of the source. The data we retrieved contained information on all clubs and players around the world over a duration of 16 seasons. The database is constructed in SQL. We pulled data using MySQL, built connections in R and converted them into the following 7 tables: “Club”, “ClubSeason”, “Leagues”,” Countries”,” Players”, “PlayerClubSeason”, “Seasons”. A visual representing the layout of the data is drawn below.

Football Dataset Organization



We built networks for both the club level and player level. First, the club network is a directed one created for all the clubs in the world. The edges are calculated based on transactions in such a way that for every transaction where a player is transferred from club A to club B, the directed edge is created in the direction of club B. In total, there are 444 teams (vertices) in the network and 20,160 edges. The central coreness of the clubs is 51.8 among all 444 teams, showing a not very high cohesiveness.

On the player side, since soccer player network is a player collaboration network, we define players are connected if they ever played together at the same club. we have set up the network as an undirected weighted network consisting of 30,686 nodes and 1,509,767 edges.

Club Network Analysis:

Club Network Methods

One of our reports main objectives is to identify which clubs are “Springboard clubs”, clubs that best set up a new player for success in the future. These clubs must be clubs that a player can easily enter, allows the player to gain experience, and eventually lead a player to be transferred to the best clubs in the world.

Before we could properly identify “Springboard clubs”, we first needed to which teams are best in the world. The club ranks are determined by averaging the value across all 16 seasons. We used a coefficient of 3% to account for inflation over the 16-year period. When determining how to calculate these club ranks we also considered averaging club rank over the seasons as an alternative. However, this calculation was ultimately infeasible as clubs are only ranked within their leagues. Consequently, we were unable to rank clubs objectively using their rankings within the league and only used the method of averaging club value across the years to determine the rankings across all clubs. By using this method, we are assuming that clubs that have a higher market value are the superior clubs. This assumption may not be entirely correct, and the method’s validity could be further investigated in future research.

After identifying the best teams across the world, we focused on identifying the best “springboard club” using network analysis on the club network (created as described above). By applying different centrality methods, we were able to identify which club is most influential. However, when considering which club is the best “springboard” we also needed to take into consideration how easily a player could start there. To fulfill this condition, “betweenness” is the best criterion. In this case, the club will be involved in many transactions that have made it well-connected. Therefore, in deciding which club was the best club to start at as a “springboard club” we decided to look at which club had the highest level of betweenness.

Club Network Results

The top 20 clubs in the world are listed as shown below in Table 4.1. From the list, we can see that all the clubs are in Europe: 5 clubs are from Spain; 7 clubs are from Britain; 1 club is from Germany; 5 clubs are from Italy; and 2 clubs are from France. Comparing the results to the actual global club rank, we see that only 5 clubs are not in the list calculated by average value indicating that club rank has not changed much over the 16 seasons. It appears that the top clubs typically make large profits while attracting excellent players to keep them at the top.

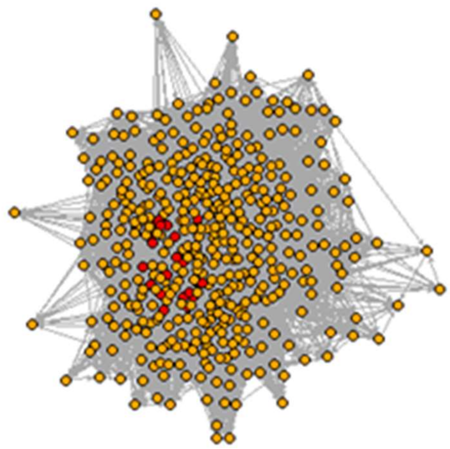
Table 4.1: Club rank by average value

idClub	avg_value	nameClub
131	352360272.7	FC Barcelona
418	342939545.5	Real Madrid
631	318404909.1	Chelsea FC
985	267822090.9	Manchester United
27	240341818.2	Bayern Munich
281	228774909.1	Manchester City
46	211474090.9	Inter Milan
506	208176909.1	Juventus FC
11	205819545.5	Arsenal FC
31	195006181.8	Liverpool FC
5	183805000	AC Milan
148	176983000	Tottenham Hotspur
583	142692090.9	Paris Saint-Germain
1049	138233636.4	Valencia CF
12	131811000	AS Roma
13	126433076.9	Atletico Madrid
430	121339454.5	ACF Fiorentina
1041	114877363.6	Olympique Lyon
29	114218000	Everton FC
368	107012272.7	Sevilla FC

Table 4.2: 2018 Actual Global Club Rank

Rank	Club
1	Juventus FC
2	Manchester City
3	FC Barcelona
4	Paris Saint-Germain
5	Atletico Madrid
6	Real Madrid
7	Liverpool FC
8	Bayern Munich
9	SSC Napoli
10	Tottenham Hotspur
11	Dortmund
12	Ajax
13	Chelsea FC
14	FC Porto
15	Arsenal FC
16	Manchester United
17	Sevilla FC
18	Inter Milan
19	AS Roma
20	FC Shakhtar Donetsk

Table 4.3 Plot of the network (Top Clubs in Red)



The result for centrality methods to find out springboard clubs are shown in the figures below. Eigenvector centrality is a measure of the influence of a node in a network. The 10 teams with highest eigenvector centrality are all from the list of top valuable clubs. From this, we conclude not only that teams of the highest value are usually also the most influential ones, but also that eigenvector centrality is less indicative of springboard clubs than other measures. In-degree and out-degree of clubs are better measures because they indicate the number of clubs a given club has transactions with. This shows different team operation patterns: some clubs tend to buy more players from other clubs while others prefer to cultivate their own young players and turn a profit from selling them. Additionally, we noticed the degree centrality of a given club is not great indicator of ranking, based on its low correlation.

Table4.4: Eigenvector

Id	nameClub	eigen
631	Chelsea FC	1
418	Real Madrid	0.882499
281	Manchester City	0.746823
131	FC Barcelona	0.695924
985	Manchester United	0.615364
46	Inter Milan	0.57025
506	Juventus FC	0.569342
31	Liverpool FC	0.514586
5	AC Milan	0.506432
148	Tottenham Hotspur	0.495994

Table4.5 Out-degree

Id	nameClub	deg_out
162	AS Monaco	106
720	FC Porto	101
31	Liverpool FC	100
336	Sporting CP	99
294	SL Benfica	99
46	Inter Milan	98
5	AC Milan	96
931	Fulham FC	95
82	VfL Wolfsburg	95
281	Manchester City	94

Table4.6 In-degree

id	nameClub	deg_in
931	Fulham FC	94
148	Tottenham Hotspur	86
294	SL Benfica	85
379	West Ham United	85
683	Olympiacos Piraeus	82
252	Genoa CFC	82
631	Chelsea FC	81
368	Sevilla FC	81
1084	MÃ¡laga CF	81
162	AS Monaco	80

Finally, we calculated betweenness to identify our springboard clubs because result gives us a more feasible place for a new player to begin. As we examined the results, we noticed that while most of the teams with high betweenness rank highly within their own league, the teams tended to reside in leagues that were less competitive. Because the league is less competitive, a strong albeit inexperienced player could still catch the eye of the team and propel his career. The team with highest betweenness is Standard Liege, which is the champion in Belgium. It has a betweenness of 1727, much larger than the last ten clubs. Having identified that team as a potential springboard club, we wanted to verify our assumptions by looking at the shortest number of walks it takes for a play to move from Standard Leige to a top team like Barcelona. Examining the results below, we see that it only takes between 1 and 2 walks to move from Standard Leige to all the top teams while it takes 2 walks to move from Bilbao Athletic, a famous club in Spain with very low betweenness to reach those same top teams. While the difference of 1 to 2 walks may at first seem insignificant when one considers the time and effort that is required to change teams, being that one step closer to the top teams could have a monumental impact on your career trajectory.

Table 4.7: Clubs with top betweenness

Id	nameClub	betw
3057	Standard LiÃ¨ge	1727.49
128	Skoda Xanthi	1676.093
7769	CFR Cluj	1654.653
197	AC Sparta Praha	1650.286
410	Udinese Calcio	1608.928
931	Fulham FC	1523.829
2441	AEK Athens	1519.166
312	Dinamo Bukarest	1449.289
162	AS Monaco	1407.499
336	Sporting CP	1333.098

Table 4.8: Clubs with least betweenness

Id	nameClub	betw
40058	New York City FC	0
45604	Orlando City SC	0
16247	Stal Dniprodzerzhynsk	0.486455
6688	Bilbao Athletic	3.404303
5818	Hobro IK	5.854912
40843	FC Voluntari	6.316418
2036	1.FC Heidenheim 1846	7.423638
4331	Trapani Calcio	10.64644
20519	Virtus Entella	10.94004
40812	CS U Craiova	15.65457

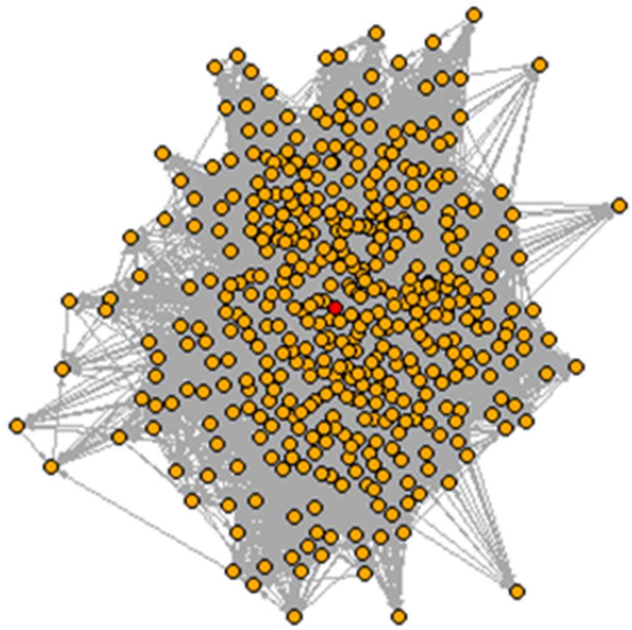
Table 4.9: Number of Shortest Paths From Standard Liege to Top Clubs

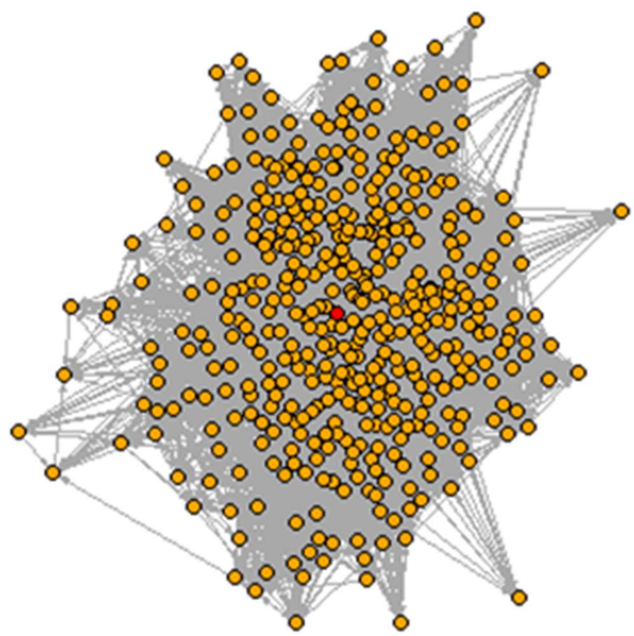
IdClub	nameClub	no. of paths
131	FC Barcelona	2
418	Real Madrid	2
631	Chelsea FC	2
985	Manchester United	2
27	Bayern Munich	2
281	Manchester City	1
46	Inter Milan	1
506	Juventus FC	1
11	Arsenal FC	2
31	Liverpool FC	1
5	AC Milan	1
148	Tottenham Hotspur	1
583	Paris Saint-Germain	2
1049	Valencia CF	1
12	AS Roma	1
13	Atletico Madrid	2
430	ACF Fiorentina	2
1041	Olympique Lyon	1
29	Everton FC	1
368	Sevilla FC	1

Table 4.10: Number of Shortest Paths From Bilbao Athletic to top clubs

IdClub	nameClub	no. of paths
131	FC Barcelona	2
418	Real Madrid	2
631	Chelsea FC	2
985	Manchester United	2
27	Bayern Munich	2
281	Manchester City	2
46	Inter Milan	2
506	Juventus FC	2
11	Arsenal FC	2
31	Liverpool FC	2
5	AC Milan	2
148	Tottenham Hotspur	2
583	Paris Saint-Germain	2
1049	Valencia CF	2
12	AS Roma	2
13	Atletico Madrid	1
430	ACF Fiorentina	2
1041	Olympique Lyon	2
29	Everton FC	2
368	Sevilla FC	2

Table 4.11: Plot of the network (Springboard Club in Red)





Player Network Analysis

Player Network Methods

On player side, we want to help a new player to improve the skills by approaching to top players. And according to the research, we found that the players that play with the best players are usually strong players themselves as measured by the price a club has paid for that player. This result matches our intuition as typically higher ranked teams frequently attract and afford several very highly ranked players. Conversely, we believe low value players may change a lot of clubs and change a lot of teammates in a couple of seasons.

We believe this comes about as a consequence of new players' ability to more quickly rise through the ranks when they start out as their skills improve and their connections grow.

Edge Weight:
In general, player market value is a good identifier of the quality of a player. Therefore, we choose to use the combined market value of the two adjacent players when calculating the edge weight. Additionally, since our data spans over fifteen seasons, we adjusted for inflation to accurately compare the dollar values of players over time. In so doing, we ensure that good players that played in the past are not penalized.

As shown below, the formula for calculating the weight of a specific edge is:

$$\text{Weight} = (\text{player Value x} + \text{player Value y}) * (1 + \text{inflation rate} * (15 - \text{idS})) / 100000.$$

In this equation, player Value x and player Value y correspond to the market value of the two adjacent players; idS represents the number of seasons in which players played together; and inflation rate represents average inflation rate. The equation is divided by 100000, to standardize the results along the same numeric scale.

Having now created the undirected weighted network, we next wanted to analyze which player would be the best connection for a new player to have in the network, taking into account the difficulty of forming a connection with that player. First, we needed to discover the most "important" player in the network. We defined node importance here by the following four methods: (1) Degree centrality can show the direct connections, (2) Closeness centrality can show the reachability of a specific player, (3) Betweenness centrality can tell whether the player is in a breakage position in the network, (4) PageRank can show the importance of the player in the network by using eigenvalue methodology and also taking weight into consideration.

After identifying the most important player in the network, we can suggest that a new player to engage with those players to get close to the best player and calculate the best and reasonable path for newbie to become the top player he always wants to be.

Player Network Results

The top 20 players based only on their market value are shown in Table 4.12. However, we expect the top 20 players that we reevaluate base on the four methods will be different.

Table 4.12. Top 20 player based on the market value

	idP	lastName	firstName	count_year	playervalue
1:	28003	Messi	Lionel	11	84000000
2:	8198	Ronaldo	Cristiano	11	77000000
3:	68290		Neymar	8	70000000
4:	44352	SuÃ¡rez	Luis	10	63000000
5:	39381	Bale	Gareth	9	56000000
6:	88103	RodrÃguez	James	8	56000000
7:	58358	MÃ¼ller	Thomas	7	52500000
8:	3373	GaÃºcho	Ronaldinho	12	49000000
9:	38253	Lewandowski	Robert	6	49000000
10:	50202	Hazard	Eden	8	49000000
11:	3332	Rooney	Wayne	11	45500000
12:	7600	Iniesta	AndrÃ©s	11	45500000
13:	7607		Xavi	10	45500000
14:	26399	AgÃ¼ero	Sergio	11	42000000
15:	39152		Falcao	13	42000000
16:	48280	Cavani	Edinson	9	42000000
17:	88755	De Bruyne	Kevin	8	42000000
18:	3366		KakÃ¡	13	38500000
19:	8806	FÃ bregas	Cesc	11	38500000
20:	40433	SÃ¡nchez	Alexis	8	38500000

In the following results, we are identified that player who has the highest direct connections in the player network in Table 4.13. This will allow the new player to increase his network once he connects to, for instance, Benjamin Finke who has 523 direct connections.

Table 4.13 Top 10 player with the highest degree centrality across 15 years

	player_id	firstName	lastName	degree
1	1324	Benjamin	Finke	523
2	3290	Graham	Barrett	478
3	3654	Ifeanyi	Udeze	475
4	5217	Geri	Cipi	466
5	2804	Torsten	Mattuschka	458
6	3182	Ashley	Cole	450
7	692	Tim	Borowski	448
8	518	Mike	Barten	442
9	2918	Ladislav	Maier	442
10	2215	Petar	Kushev	440

(1) Closeness centrality

With identify the top players base on closeness centrality, the new player has a higher chance to approach to any players he wants to connect with.

Table 4.14 Top 10 player with the highest closeness centrality across 15 years

	player_id	firstName	lastName	degree	closeness
1	5460	Christophe	Le Roux	171	8.085362e-07
2	18396	Lars	Wallaeys	67	8.077795e-07
3	15964	Istv<e1>n	Ferenczi	144	8.076858e-07
4	18388	Marlin	Piana	195	8.065106e-07
5	19347	Christoph	Stahl	143	8.056810e-07
6	15976	Vasili	Khomutovski	80	8.052108e-07
7	15266	Omonigho	Temile	246	8.046039e-07
8	15974	Esteban		96	8.039310e-07
9	16353	Micka<eb>l	Chr<e9>tien Basser	161	8.038693e-07
10	15368	Vyacheslav	Malafeev	156	8.027947e-07

(2) Betweenness centrality

Since the player has connections only if they have played in the same team, so by connecting with the high betweenness player will allow newbie to connect with player in another team easily.

Table 4.15 Top 10 player with the highest betweenness centrality across 15 years

	player_id	firstName	lastName	degree	closeness	betweenness
1	19347	Christoph	Stahl	143	8.056810e-07	17185443
2	15964	Istv<e1>n	Ferenczi	144	8.076858e-07	13380824
3	13007	Anders	Nielsen	95	7.917243e-07	8983659
4	18388	Marlin	Piana	195	8.065106e-07	8024365
5	5460	Christophe	Le Roux	171	8.085362e-07	7678546
6	23269	Juan	Luis Mora	134	7.857711e-07	7074253
7	19061	Frank	van der Struijk	171	7.880280e-07	6986939
8	19370	Frederik	D'Hollander	92	7.906264e-07	6915039
9	3463	Michael	Chopra	272	7.737285e-07	6817662
10	15368	Vyacheslav	Malafeev	156	8.027947e-07	6530526

(4) PageRank

By applying PageRank on player network, we can reveal the best players in Table 5.3. Since we wanted to identify the best players in the last fifteen seasons, we expected the well-known and valued names of soccer to be at the top of the list in this stage of analysis.

Table 4.16 Top 10 player with the highest PageRank across 15 year

	player_id	firstName	lastName	degree	closeness	betweenness	pagerank
1	3522	Andriy	Shevchenko	296	1.301845e-07	0	0.0007557969
2	3504	Alpay	<d6>zalan	327	1.004520e-07	0	0.0005874447
3	2692	Mounir	Zitouni	318	1.991898e-07	0	0.0005593546
4	155	Michael	Sternkopf	322	2.907401e-07	0	0.0005410277
5	1262	Miroslav	Karhan	280	2.173439e-07	0	0.0005218284
6	1253	Pavel	David	371	5.653052e-07	0	0.0005118156
7	89	Z<e9>	Roberto	331	4.245759e-07	0	0.0005079129
8	2905	Mamadou	Kant<e9>	371	2.437665e-07	0	0.0005014846
9	512	Hans	van de Haar	306	5.806844e-07	0	0.0004839355
10	1324	Benjamin	Finke	523	6.404253e-07	0	0.0004498464

Next, in order to identify the best players and to find the players that have similar features to the new player, we chose age as the most vital feature because it can be an indicator of the experience a player has in professional soccer. We analyzed each age group individually to identify the most valuable players in that age group. The player has the highest PageRank score is the best player in their age groups.

Age around 30:

Table 4.17 Top 10 player in age around 30

	player_id	firstName	lastName	pagerank	birthDate	age_group
1	2905	Mamadou	Kant<e9>	5.014846e-04	1984	30
2	12359	Niki	M<e4>enp<e4><e4>	3.925389e-04	1985	30
3	4233	Adam	Federici	3.467119e-04	1985	30
4	4763	Bo	Storm	2.710271e-04	1987	30
5	9917	Henri	Siqueira-Barras	2.651627e-04	1985	30
6	3432	Florian	Mohr	2.583155e-04	1984	30
7	1671	Mirnel	Sadovic	2.571297e-04	1984	30
8	3762	Peter	Whittingham	2.564765e-04	1984	30
9	7781	Alexis		2.550619e-04	1985	30
10	7335	Justin	Hoyte	2.352095e-04	1984	30

Age around 40:

Table 4.18 Top 10 player in age around 40

	player_id	firstName	lastName	pagerank	birthDate	age_group
1	3522	Andriy	Shevchenko	0.0007557969	1976	40
2	3504	Alpay	<d6>zalan	0.0005874447	1973	40
3	1262	Miroslav	Karhan	0.0005218284	1976	40
4	1253	Pavel	David	0.0005118156	1978	40
5	89	Z<e9>	Roberto	0.0005079129	1974	40
6	512	Hans	van de Haar	0.0004839355	1975	40
7	1324	Benjamin	Finke	0.0004498464	1982	40
8	2980	J<fc>rgen	Friedl	0.0004405301	1981	40
9	1574	Mohamadou	Idrissou	0.0004402619	1980	40
10	4589	Patrick	Pothuizen	0.0004387913	1972	40

Age around 50:

Table 4.19 Top 10 player in age around

	player_id	firstName	lastName	pagerank	birthDate	age_group
1	2692	Mounir	Zitouni	0.0005593546	1970	50
2	155	Michael	Sternkopf	0.0005410277	1970	50
3	530	Andreas	Herzog	0.0004090537	1968	50
4	5865	Stefano	Torrisi	0.0003804608	1971	50
5	151	Zdenko	Miletic	0.0003783630	1968	50
6	179	Vasile	Miriuta	0.0003498872	1968	50
7	3524	Marco	Simone	0.0003334036	1969	50
8	140	J<f6>rg	Bode	0.0003224853	1969	50
9	560	Martin	Max	0.0003195850	1968	50
10	3429	Kennet	Andersson	0.0003026737	1967	50

To help a new player better acquire different skills, we also examined positions as different categories. By identifying the skillset required for different positions at age group 30, The new player must focus on a certain sub group of players to approach to learn the specific skills.

Position: ATT

Table 4.20 The best 10 player playing as an Attacker in age around 30

	player_id	firstName	lastName	pagerank	birthDate	age_group	playingPosition
1	2905	Mamadou	Kant<e9>	5.014846e-04	1984	30	ATT
2	4763	Bo	Storm	2.710271e-04	1987	30	ATT
3	1671	Mirnel	Sadovic	2.571297e-04	1984	30	ATT
4	7416	Roman	Kienast	2.068371e-04	1984	30	ATT
5	12551	<c9>ric	Mouloungui	1.963677e-04	1984	30	ATT
6	3169	Daniele	Fiorentino	1.853549e-04	1984	30	ATT
7	6322	Patrick	Helmes	1.542759e-04	1984	30	ATT
8	3975	Domi	Kumbela	1.479630e-04	1984	30	ATT
9	2087	Dustin	Heun	1.468281e-04	1984	30	ATT
10	8883	Emmanuel	Adebayor	1.432303e-04	1984	30	ATT

Position: DEF

Table 4.21 The best 10 player playing as a Defender in age around 30

	player_id	firstName	lastName	pagerank	birthDate	age_group	playingPosition
1	9917	Henri	Siqueira-Barras	2.651627e-04	1985	30	DEF
2	3432	Florian	Mohr	2.583155e-04	1984	30	DEF
3	7781	Alexis		2.550619e-04	1985	30	DEF
4	7335	Justin	Hoyte	2.352095e-04	1984	30	DEF
5	11695	Philipp	J<e4>schke	2.055680e-04	1984	30	DEF
6	7445	Can	Arat	1.680928e-04	1984	30	DEF
7	16638	Edson	Henrique	1.574408e-04	1987	30	DEF
8	9915	Reto	Ziegler	1.365331e-04	1986	30	DEF
9	8628	Matthias	Langkamp	1.342427e-04	1984	30	DEF
10	15588	Artem	Madilov	1.250576e-04	1985	30	DEF

Position: MID

Table 4.22 The best 10 player playing as a Midfielder in age around 30

	player_id	firstName	lastName	pagerank	birthDate	age_group	playingPosition
1	3762	Peter	Whittingham	2.564765e-04	1984	30	MID
2	12392	Emerse	Fa<e9>	2.196452e-04	1984	30	MID
3	12285	Kim	Tandrup	2.165859e-04	1987	30	MID
4	9161	Sammy	Clingan	2.147603e-04	1984	30	MID
5	12386	Gr<e9>gory	Bourillon	1.934115e-04	1984	30	MID
6	4218	Floribert	Ngalula	1.874510e-04	1987	30	MID
7	12917	Thibaut	Detal	1.818162e-04	1985	30	MID
8	13270	Matt	Bloomfield	1.805321e-04	1984	30	MID
9	9870	Christos	Tsevas	1.751956e-04	1985	30	MID
10	12866	Alexandros	Karachalios	1.553421e-04	1985	30	MID

Position: GK

Table 4.23 The best 10 player playing as a Goalkeeper in age around 30

	player_id	firstName	lastName	pagerank	birthDate	age_group	playingPosition
1	12359	Niki	M<e4>enp<e4><e4>	3.925389e-04	1985	30	GK
2	4233	Adam	Federici	3.467119e-04	1985	30	GK
3	3181	Lenny	Pidgeley	2.298628e-04	1984	30	GK
4	15570	Igor	Akinfeev	1.963052e-04	1986	30	GK
5	16408	Orestis	Karnezis	1.333080e-04	1985	30	GK
6	15607	Sergey	Panov	1.263890e-04	1989	30	GK
7	13326	Johnny	Leoni	1.196744e-04	1984	30	GK
8	17734	Polat	Keser	1.138407e-04	1985	30	GK
9	14004	Lance	Cronin	1.111276e-04	1985	30	GK
10	7914	Carlos	Kameni	1.055880e-04	1984	30	GK

According to Table 5.8, If a new player wanted to become the best defender, Florian Mohr would be the best player to approach to improve his defense skills. Ultimately by applying shortest path methods, we can provide a given new player a precise number of steps for him to reach to Florian Mohr.

For instance, since the player is new to professional soccer, it is possible that the only connection that player has is through a player like Manuel Hartmann, who is the second lowest value player in DEF position in age around 30. After calculating the shortest path, we see that the new player would be another 10.3482(~11) steps away from a top player like Florian Mohr.

Network Model:

To specify the probability distribution of the players' network from the set of random networks, we used the *ergm* package in R. Using maximum likelihood, this method estimates for the factors age group, country, and playing positions is as shown below.

Age Groups:

- Age group 1: Less than or equal to 20
- Age group 2: Age greater than 20 and less than or equal to 28.
- Age group 3: Age greater than 28 and less than or equal to 38.
- Age group 4: Age greater than 38.

Playing position:

- ATT: Attacker.
- DEF: Defender.
- GK: Goalkeeper
- MID: Midfielder

```
==========================
Summary of model fit
==========================

Formula:   players_n ~ edges + triangle + nodematch("countryID") +
    nodematch("age", diff = T) + nodematch("playingPosition",
    diff = T)

Iterations:  20 out of 20

Monte Carlo MLE Results:
                                Estimate Std. Error MCMC %  z value Pr(>|z|)
edges                         -1.576e+01  3.987e-02     43 -395.179   <1e-04 ***
triangle                       3.053e+00  7.001e-04     18 4360.533   <1e-04 ***
nodematch.countryID            1.919e+00  7.339e-02     97   26.145   <1e-04 ***
nodematch.age.2                2.936e+00  3.267e-01    100    8.987   <1e-04 ***
nodematch.age.3                7.483e-01  4.611e-02     83   16.229   <1e-04 ***
nodematch.age.4               -1.915e+00  5.908e-02     98  -32.418   <1e-04 ***
nodematch.playingPosition.ATT -2.618e-01  1.109e-01     99   -2.361   0.0182 *
nodematch.playingPosition.DEF -1.634e+00  6.834e-02     97  -23.916   <1e-04 ***
nodematch.playingPosition.GK   6.205e+00  8.769e-02     97   70.768   <1e-04 ***
nodematch.playingPosition.MID -6.696e-03  5.266e-02     94   -0.127   0.8988
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 15847245  on 11431371  degrees of freedom
 Residual Deviance:   659784  on 11431361  degrees of freedom

AIC: 659804    BIC: 659947    (Smaller is better.)
```

The above results suggest that players in age group 2 and 3, are more likely to be connected. This in simple terms means all teams have more players from age group 2 and 3. When combined age attribute with the

country, we get to see that players from same countries have higher tendency to form an edge than players from different countries. These results give conditional log-odds of a tie between two edges.
This modeling procedure is reasonably important to concretize our belief of how the players tend to create an edge. The results suggest homophily in the network based on countries and age groups.

Conclusion:

In conclusion, the player network we analyzed has over 30 thousand nodes, with over 1 million edges in the network. The club network has 444 nodes and more than 20 thousand edges. The network centrality algorithms were somewhat computationally expensive; however, the results are consistent with real-world expectations. A very important factor in the players' value evaluation is the inflation rate, which ensures the top players' list should have the best players from previous seasons.
Using the same network of players, PageRank is used to find top players in different age groups as well as at different playing position. The top players from age category 30 are Mamadou Kant, Niki Menp, Adam Federici, Bo Storm. Overall top players are Andriy Shevchenko, Alpay Zalan, Mounir Zitouni, and Michael Sternkoph.
The results from club network are consistent with our initial expectations. The clubs from less valued leagues are good springboard clubs as they give more opportunity for young players to gain experience. These are the clubs on which the big clubs have their eyes on to find potential good players. The top four springboard clubs are Standard Liege, AC Sparta Praha, Skoda Xanthi, and CFR Cluj from ProLeague, SynotLiga, SuperLeagueGreece, and Liga1 leagues, respectively.
Combining two results is critical here, we will take top four springboard clubs and rank players in these springboard clubs in latest season according to their PageRank, this would give us the names of players who should be targeted to create connection or get closer. Below table suggests that among the springboard club who are the top players according to their PageRank. These results can be interpreted as suggestion to the new player who are the players once he gets into a club to have more interactions in order to have higher chances of getting more connected in the global social network of all the payers.

firstName	lastName	pagerank	playingPosition	nameClub
Markus	Steinhöfer	3.201963e-05	DEF	AC Sparta Praha
Christos	Lisgaras	2.830529e-05	DEF	Skoda Xanthi
Ondrej	Mazuch	1.167102e-05	DEF	AC Sparta Praha
Jonathan	Legear	4.489673e-05	MID	Standard Liège
Ovidiu	Herea	6.823519e-06	MID	Skoda Xanthi

The ERGM model which was tried for the players' network could not be run due to the *ergm* library limitations, hence we could only run it on season 5, which has close to 5000 edges.

Future Work:

In future we would like to see the network behavior to see how this social network has evolved in years. Include factor of tie strength reduction using a good approximation function of edge decay. The coreness of the network would be affected because of the decaying edges, which can give useful insights.
The algorithm can be modified to use rank by giving weighted levels for each league, with the help of domain experts. We can incorporate variables such as club managers and how club policies change over time for more accurate results. The data about the skill level can be utilized to predict the position of a player in the whole network after a period. For instance, we could use a skill as behavior variable and see how the edges change with waves defining the connections at each season or multiple seasons.
We can also extend this application on other sports competition that also have the set up as team and compare to other team, for instance, basketball and baseball. To further utilize our result, we can analyze deeper into the connection players have not only within a team, but also the connection that player made

within a match. So, we are able to come up with a strategy that helps team to win the match based on social network analytics.

Reference:

1. World Football / Soccer Clubs Ranking: https://footballdatabase.com/ranking/world/1
2. *Identifying top football players and springboard clubs from a football player collaboration and club transfer networks* by Matic Tribuˇson, Matevˇz Leniˇc University of Ljubljana, Faculty of Computer and Information Science, Veˇcna pot 113, SI-1000 Ljubljana, Slovenia
3. F. Radicchi, M. Perc, Who is the best player ever? a complex network analysis of the history of professional tennis, PloS one 6 (2) (2011) e17249.
4. https://statnet.org/trac/raw-attachment/wiki/Sunbelt2016/ergm_tutorial.html
5. www.transfermarkt.co.uk. website that has players transfer data.