The Football World as a Social Network

Matt Jarrams 30061191 CPSC 572
Owen Hunter 30075173 CPSC 572

2022-12-07

1 Project Overview

In this paper we collect all footballers playing in Europe's top 5 leagues since 2010 and create a social network by using their playing histories from their entire career, pairing any players they have played with over their career. From this resulting network, we attempt to find relationships and metrics using social network techniques such as community detection, information and influence spread through the network, and machine learning algorithms to predict nationality and teams based on certain edge (teammate) and node (player) attributes.

From our findings, we were able to create models to accurately predict the team that two players played for given their nationalities, ages and years of playing. As well as a model to predict the country a player came from based on how many years they spent in a particular league as well as social network metric values for a player.

We were also able to look at the structure of the football network by creating a model to recreate our network by simulating football transfers, player retirements, and promotion and relegation. These techniques allowed us to understand and explain our network structure as well as add context to our degree distribution.

Finally, we dive into how influence and information could possibly spread throughout the football world by creating probability models. We analyze the results and the positive and negative consequences that come along with the spread.

We believe we have only scratched the surface with this data set and analysis. with more computing power and fewer time constraints, we believe more in-depth models could be built for prediction. Furthermore, more time could be spent on what factors drive this network in an attempt to build a model that recreates this network.

We hope this initial analysis can be taken further by adding more data to compare across different eras of the game and with a more in-depth look at the nodes and edges further discoveries can be made about the highly connected footballing world.

2 Introduction

As lifelong sports fans, we have always thought football has a very unique structure to it. Being one of the richest sports in the world and the most popular according to WorldAtlas [1], it has a vast amount of data to explore. With multiple high-ranking professional leagues around the world, it is a sport that is played at an elite level internationally.

This popularity and international aspect make it very different from other North American sports which typically have only one or two countries that play it at the highest level (American Football and Baseball as examples). Football has many different countries, teams and players from all over the world, creating this unique environment of multiple different cultures, ages and backgrounds coming together to try and win games and trophies.

Some other separating factor from other sports football has, is that players are able to have some input into the teams they play for. Unlike North American sports like Ice Hockey, basketball, etc... players are not traded, they are bought by one team that will then pay a transfer fee to the parent club. That player will then have to agree on personal terms (wage, contract length, etc...) before the transfer can be complete. If the player and the new club do not reach an agreement then the player will not leave. This is helpful for us in making sense of some of the links we have in our network. In addition to this choice of the team aspect, football also has no "entry draft" from universities/colleges, this means players will typically be brought through youth academies and integrated into the first team depending on how well they progress. As opposed to North American sports where young players are mandated to play in university leagues or dedicated youth leagues before being drafted by professional franchises.

Theoretically, this could mean if a player aged 15 was seen as good enough for the first team, they could be playing with the first team with someone more than twice their age. A famous instance of this was a 16-year-old phenomenal Wayne Rooney who took the league by storm back in 2002 and Jude Bellingham being signed by Borussia Dortmund in 2020 at age 17 for 35 million Canadian Dollars [2]. However, it is much more common for players to start playing regularly in the 18-21 year range.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
			1 Soccer Training am Gym pm	2 Match Day minus 1 Light Training	3 **League Fixture**	4 Recovery Session am
5 **FA Cup Fixture**	6 Recovery Session am	7 *Off*	8 Soccer Training am Gym pm	9 Match Day minus 1 Light Training	10 **League Fixture**	11 Recovery Session am
12 Soccer Training am Gym pm	13 Match Day minus 1 Light Training	14 **Champions League Fixture**	15 Recovery Session am	16 Match Day minus 1 Light Training	17 **League Fixture**	18 Recovery Session am
19 *Off*	20 Match Day minus 1 Light Training	21 **EFL Cup Fixture**	22 Recovery Session am	23 Match Day minus 1 Light Training	24 **League Fixture**	25 Recovery Session am
26 *Off*	27 Match Day minus 1 Light Training	28 **Champions League Fixture**	29 Recovery Session am	30 Match Day minus 1 Light Training	31 **League Fixture**	

Figure 1: A Typical monthly schedule for a top professional soccer club in the Premier League [3]

We find this aspect particularly interesting that players at very different stages in life and maturity will spend a large amount of time training, travelling and competing together. As seen above in figure 1, professional footballers spend almost all their time together when they are working. We explore this further in our analysis, looking at how behaviour can spread throughout our network.

In this paper, we look to take on this unique player network and showcase how modelling players playing histories as a social network can uncover some interesting metrics and relationships.

Recently statistics and the importance of more complex data have made their way into the world of sports, especially football. In addition, there have been network analyses done in regard to football results and transfers. In "Is a social network approach relevant to football results?"[4] they use a social network approach for players passing to help in the performance of a team in a soccer match. Furthermore, there has been research into player transfers using social network methodologies. In "Football Transfers looked from a social network analysis perspective"[5], the author analyzed football transfers using network metrics to see how connected clubs were, predict success and investigate in how the change of ownership impacts transfer activity. In our paper, we plan to investigate more into the interactions between players as teammates than match results and to use transfers as a possible explanation for how our network forms.

We collected players who have played in the top 5 European leagues (English Premier League, Italian

Serie A, German Bundesliga, Spanish La Liga and French Ligue 1) dating back to 2010. We then looked back at all of these players playing history and constructed our network by creating links between players who played together in the same season on the same team.

This network approach allowed us to analyze attributes of our communities and clusters, how much of an impact players have on each other, how predictable player movements are, and what role nation, league, age and season have on predicting the nationalities of players and clubs players have played for.

3 Dataset

To collect our data we used Fbref.com. Fbref is a popular football encyclopedia which contains data from over 100,000 players and over 100 different leagues. For our data, we used the "Player Standard Stats 2022-2023 Big 5 European Leagues" tables from the 2010-2011 season all the way to the 2022-2023 season.

To collect our data, we first web-scraped all of these players' URLs, names, Fbref ID, birth year, and nation from the standard statistics table from each of the 13 seasons using the Beautiful Soup python library.

Player Standard Stats 2022-2023 Big 5 European Leagues

Share & Export ▾ | When table is sorted, hide non-qualifiers for rate stats | Glossary | Toggle Per90 Stats

								Playing Time				Performance						
Rk	Player	Nation	Pos	Squad	Comp	Age	Born	MP	Starts	Min	90s	Gls	Ast	G-PK	PK	PKatt	CrdY	CrdR
1	Brenden Aaronson	USA	MF,FW	Leeds United	Premier League	22-040	2000	14	14	1,189	13.2	1	2	1	0	0	2	0
2	Yunis Abdelhamid	MAR	DF	Reims	Ligue 1	35-064	1987	15	15	1,350	15.0	0	0	0	0	0	1	0
3	Himad Abdelli	FRA	MF,FW	Angers	Ligue 1	23-014	1999	7	2	231	2.6	0	0	0	0	0	0	0
4	Salis Abdul Samed	GHA	MF	Lens	Ligue 1	22-250	2000	15	15	1,349	15.0	1	0	1	0	0	2	0
5	Laurent Abergel	FRA	MF	Lorient	Ligue 1	29-303	1993	10	10	807	9.0	0	1	0	0	0	2	0
6	Matthis Abline	FRA	FW,MF	Rennes	Ligue 1	19-248	2003	10	0	105	1.2	1	0	1	0	0	0	0
7	Zakaria Aboukhlal	MAR	FW,MF	Toulouse	Ligue 1	22-286	2000	15	14	1,170	13.0	3	3	3	0	0	2	0
8	Tammy Abraham	ENG	FW	Roma	Serie A	25-060	1997	15	12	1,041	11.6	3	1	3	0	0	0	0
9	Francesco Acerbi	ITA	DF	Inter	Serie A	34-294	1988	9	7	660	7.3	0	1	0	0	0	1	0
10	Mohamed Achi	FRA	FW	Nantes	Ligue 1	20-319	2002	2	0	36	0.4	0	0	0	0	0	0	0
11	Marcos Acuña	ARG	DF	Sevilla	La Liga	31-034	1991	10	7	551	6.1	0	0	0	0	0	4	1
12	Che Adams	SCO	FW	Southampton	Premier League	26-141	1996	14	12	1,066	11.8	4	1	4	0	0	0	0
13	Tyler Adams	USA	MF	Leeds United	Premier League	23-290	1999	13	13	1,166	13.0	0	0	0	0	0	4	1
14	Sargis Adamyan	ARM	FW,MF	Köln	Bundesliga	29-192	1993	14	3	434	4.8	1	1	1	0	0	1	0
15	Tosin Adarabioyo	ENG	DF	Fulham	Premier League	25-068	1997	10	10	900	10.0	1	0	1	0	0	0	0
16	Martin Adeline	FRA	MF,FW	Reims	Ligue 1	18-364	2003	4	0	93	1.0	0	0	0	0	0	1	0
17	Karim Adeyemi	GER	FW,MF	Dortmund	Bundesliga	20-317	2002	11	7	489	5.4	0	0	0	0	0	3	0
18	Amine Adli	FRA	FW	Leverkusen	Bundesliga	22-205	2000	8	3	355	3.9	0	1	0	0	0	0	0
19	Yacine Adli	FRA	MF,FW	Milan	Serie A	22-125	2000	4	1	116	1.3	0	0	0	0	0	1	0
20	Michel Aebischer	SUI	FW,MF	Bologna	Serie A	25-329	1997	13	7	625	6.9	1	0	1	0	0	2	0

Show hidden rows 21 to 2383 ▾

Figure 2: Example of standard statistics table, highlighted columns were scraped

This resulting table of 9700 players formed the basis of our network's nodes.

We then visited all 9,700 of these player URLs and web-scraped the playing history from Fbref's "Standard Stats: Domestic Leagues" table. This gave us 104,556 rows of players who played in a particular season, with match stats like goals and number of appearances along with the team, league, and year they played on that team. Grouping this table on player names gave us our node attribute data.

Standard Stats: Domestic Leagues

Goal Logs Share & Export ▾ Glossary Toggle Per90 Stats

							Playing Time				Performance						
Season	Age	Squad	Country	Comp	LgRank	MP	Starts	Min	90s	Gls	Ast	G-PK	PK	PKatt	CrdY	CrdR
2017	16	Molde	NOR	1. Eliteserien	2nd	14	3	393	4.4	2	1	2	0	0	2	0
2018	17	Molde	NOR	1. Eliteserien	2nd	25	17	1,596	17.7	12	4	9	3	3	1	0
2018-2019	18	RB Salzburg	AUT	1. Bundesliga	1st	2	1	83	0.9	1	0	1	0	0	0	0
2019-2020	19	RB Salzburg	AUT	1. Bundesliga	1st	14	11	980	10.9	16	4	15	1	1	3	0
2019-2020	19	Dortmund	GER	1. Bundesliga	2nd	15	11	1,063	11.8	13	2	13	0	0	0	0
2020-2021	20	Dortmund	GER	1. Bundesliga	3rd	28	27	2,407	26.7	27	6	25	2	4	2	0
2021-2022	21	Dortmund	GER	1. Bundesliga	2nd	24	21	1,911	21.2	22	8	16	6	6	3	0
2022-2023	22	Manchester City	ENG	1. Premier League	2nd	10	10	841	9.3	15	3	14	1	1	0	0
6 Seasons		4 Clubs		4 Leagues		132	101	9,274	103.0	108	28	95	13	15	11	0
			Country	Comp	LgRank	MP	Starts	Min	90s	Gls	Ast	G-PK	PK	PKatt	CrdY	CrdR
Dortmund (3 Seasons)				1 League		67	59	5,381	59.8	62	16	54	8	10	5	0
Molde (2 Seasons)				1 League		39	20	1,989	22.1	14	5	11	3	3	3	0
RB Salzburg (2 Seasons)				1 League		16	12	1,063	11.8	17	4	16	1	1	3	0
Manchester City (1 Season)				1 League		10	10	841	9.3	15	3	14	1	1	0	0
Bundesliga (3 Seasons)						67	59	5,381	59.8	62	16	54	8	10	5	0
Eliteserien (2 Seasons)						39	20	1,989	22.1	14	5	11	3	3	3	0
Bundesliga (2 Seasons)						16	12	1,063	11.8	17	4	16	1	1	3	0
Premier League (1 Season)						10	10	841	9.3	15	3	14	1	1	0	0

Figure 3: Example of standard statistics table. Highlighted Columns were scraped

Once we had the players playing histories, we could build our social network. This was done by taking one row (corresponding to a particular player and season) and iterating through all other rows in the playing history and creating an edge for any player who played in the same team in the same year.

The resulting table had 901,971 rows, corresponding to links where two players had played together in the same year, grouping these on two player names that were the same and making the number of years they played together their edge weight we got our final edge dataset of 560,416 rows.

We use both the weighted version (number of seasons two players played together) and the multi-link version, where each row is a season they played together in the same team which has the possibility for multi-links. In some cases for our analyses, it makes sense to use the multi-link edge list.

Our network is undirected as when two players play with each other they interact in the same way and there is no clear definition for an in and out-degree.

For the most part, we had very few issues with data quality, apart from a couple of players where we needed to fill in the birth year and nationality of the player. As well we had to match the three-letter country code FBref used with a country list. However, apart from this, our time was mostly spent on data transformation than cleaning. We are very grateful for Fbref's work in keeping their data and statistics detailed and well-maintained.

4 Basic Statistics

In this section, we will outline the characteristics of our network by using some basic network metrics like the degree, clustering coefficient, average path length, etc. To give context to these metrics, we will compare them to a null model network created by double edge swaps to preserve degree distribution.

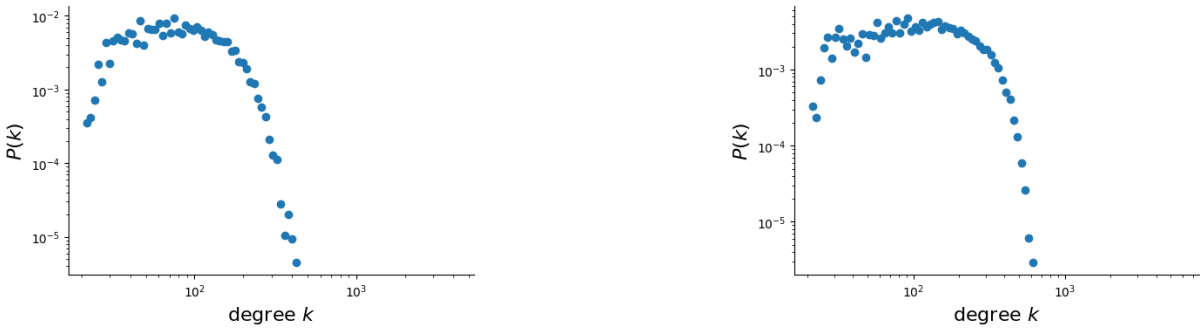
Table 1: Basic Metric Table

Basic Metrics		
Statistic	Our Network	Degree Preserving Null Model (100)
Average Shortest Path	2.53983	2.27801 +/- 0.0002
Clustering Coefficient	0.34821	0.01871 +/- 3.469 x 10^{-5}
Number of Communities	5	8.32 +/- 1.2400
Number of Nodes	9700	9700
Number of Edges	560416	560416
Connected Component	1	1 +/- 0

Two interesting metrics from this table are clustering and average shortest path length. Clustering measures the tendency of nodes to form neighbourhoods in the graph. Average path length takes the average of all the shortest paths from one node to every other node in the network. From our results, our graph has a much higher clustering coefficient than the random networks, meaning the edges in our graph tend to connect similar neighbours. This makes sense as we are looking at teammates and it is very likely for players to have played with other players in the league on other teams at some point in their careers. On average players in our network will play for roughly 5 (4.69) different teams in their career.

Our path length is slightly longer than the null model's average length, this is most likely due to players being a part of a certain team's clique. A bottom-of-the-table English team and a German team are not likely to interact in our network thus increasing the average path length. Whereas with the degree-preserving null network, random links across the network structure are far more likely thus reducing the average path of many of low degree nodes. With that being said, 2.53983 is a fascinatingly low number for a football network and it is very interesting to see that our players can reach any other player in our network with around 2-3 hops. Another interpretation of this metric is that players who are from other countries that are brought into these leagues have likely played for one of the top teams in their country (Ajax (Netherlands), Porto, Benfica (both Portugal) as some examples). Therefore, these feeder teams and leagues can explain why players can reach each other in a few number of hops. It is difficult for a null model to capture these intricacies as links are created randomly as opposed to ours which is made through a player's decision on what is best for their career and if the relative success they have had season to season This makes the relative similarity of the numbers quite interesting.

Degree Distribution :



(a) Unweighted Degree Distribution

(b) Multi-link Degree Distribution

Figure 4: Comparison of Degree Distribution

For both networks, the degree distribution follows a Poisson distribution which generally means a network is random. When investigating why it is random, it appears the inclusion of players' entire history from their early careers contributes highly to make the graph look more random. Using player history adds a lot more random edges connecting different team cliques. In addition, when taking the entire history a drawback is that if a player joins a team in one of the top 5 leagues from a smaller club in a non-major league in a different country, it is unlikely we will have a large number of players from

this small club in our network. This will mean their degree is likely to be lower than a player who for example started their career in one of our top divisions, like England. To add another reason for this a large number of players in our network begin their careers in one of the lower divisions from our top 5 leagues, like the English Championship (2nd division in English football). Since this league feeds into the Premier League a lot of young players will begin their careers in these lower leagues (often as a loaned player from their parent club in a top division), this creates a number of links with players in our network that a player coming from outside our main leagues will not have.

The same issue can occur for players who were older in our network in one of the earlier years we scraped (2010 or 2011). Especially if they retired, this meant they were less likely to be linked to future players in our network and will have fewer edges as most of their teammates will have not been included in our initial data scraping.

When player history is removed and we only focus on edges based in the top 5 leagues, our degree distribution now follows a truncated power law (figure 5) which is more what would be expected out of a sports social network. With that being said, we feel the complete player history is an important part of our network because it represents how players from the same foreign country begin their careers together and then disperse throughout the football world.

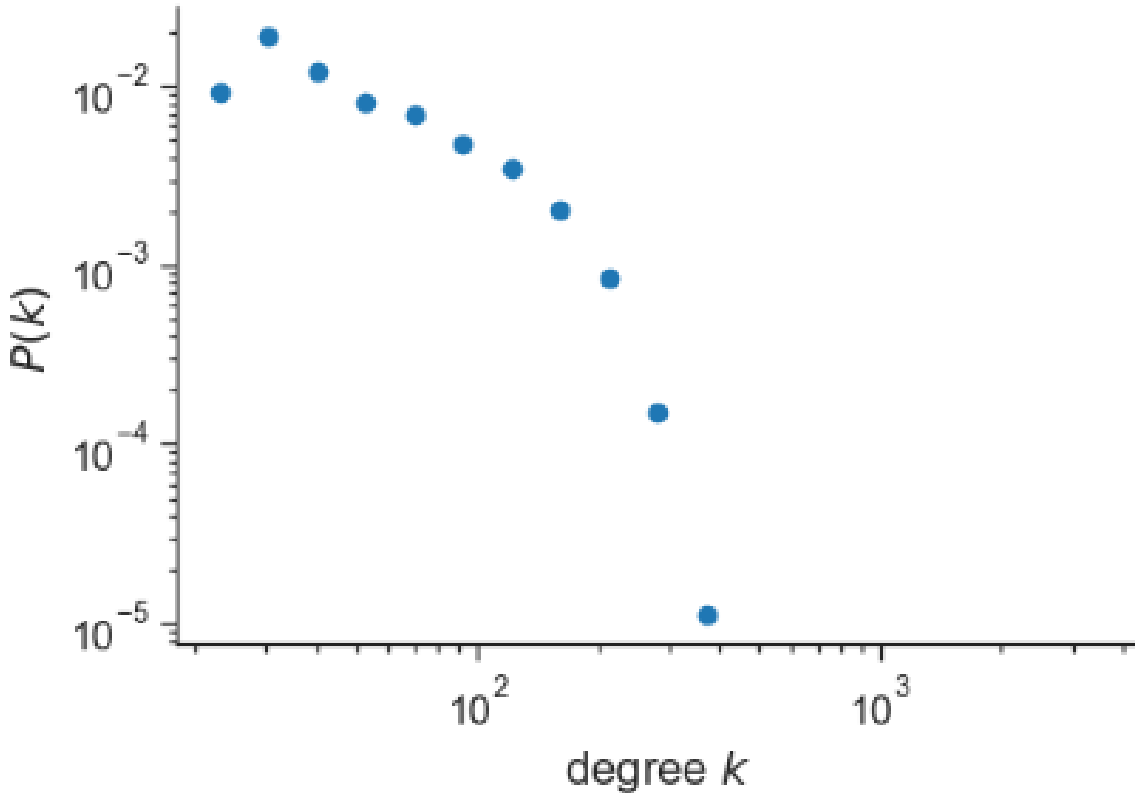


Figure 5: Degree distribution of network without player history

5 Network visualization

We visualized our network using the popular software Gephi. The network is arranged and coloured by the community. The five communities broadly correspond to the 5 main leagues for our network. We investigate this further in 6.2. Nodes are sized based on degree and higher-degree nodes tend to be drawn closer to the middle whereas low-degree nodes stay to the outside far away from other communities.

Figure 6: Entire Network - Coloured by Communities (5)

6 Results

Research Questions:

- How well can we predict what teams a link came from based on its characteristics? Is it possible to predict the nationality of a player based on their career path and personal information?
- What communities exist in our network? Are there patterns that emerge in the communities? Can we find out why these clusters form?
- How can information spread through our network? How much of an impact can a player's behaviour have on another? And what are the implications of these spreads?
- Can we build a model that recreates our network? What factors drive our network to the structure it has?

In this section, we will look at our research questions individually and broken up into subsections. Where we explain the motivation behind looking at the question, why we took the approach we did and lastly explain the importance of our analysis.

6.1 Country and Club Prediction

For our first result, we ran two classifier models. One on our unweighted multi-link graph to see how well we can predict the squad an edge came from. The second is on our nodes to predict the nationality of a player based on their career.

We first filter our edge list to only keep edges with a top 5 league squad to limit the number of labels the classifier needed to choose from, this left us with 180 squads. To run a decision tree on our edges we first needed to assign numbers to our non-number data which was each player's nation. Our other features were league position for that season and the age of both players. We then trained our decision tree model on 80% of our data and tested it on the remaining 20%.

Using the results, we paired our non-number data back to the original labels and found our accuracy to be around 0.56. We ran the classifier 100 times to protect against randomness; the results are in table 2.

Table 2: Edge Classifier Results

Edge Classifer	
Classifier	Decision Tree
Features	Age and Country of both players, League Rank
Accuracy	0.5648 +/- 0.0016
Recall	0.5673 +/- 0.0016
Precision	0.5647+/- 0.0016
F-1 Score	0.5648 +/- 0.0016

At first, we thought 56% seemed low but considering the classifier was choosing between 180 squads and correctly identified the squad over half the time, this was an impressive result. Adding in the season of the edge, the accuracy rate went up to approximately 95%. Additionally including the league improved it to 99%. However, these 2 classifiers felt a little meta and were more just using our dataset as a database lookup since with season, league rank and league there is only 1 possibility for what squad they played for.
To investigate what squad performed the best, we plotted precision vs recall for every squad on a scatter plot (Figure 7). The Serie A had both the best-performing teams and the worst. These ranged from 25% to 75%. To take it one step further, we grouped squads by league and plotted each as a box plot to compare league performance (Figure 8). Premier League here consistently performs the best while La Liga is the worst, although all 5 leagues are within similar ranges of each other.

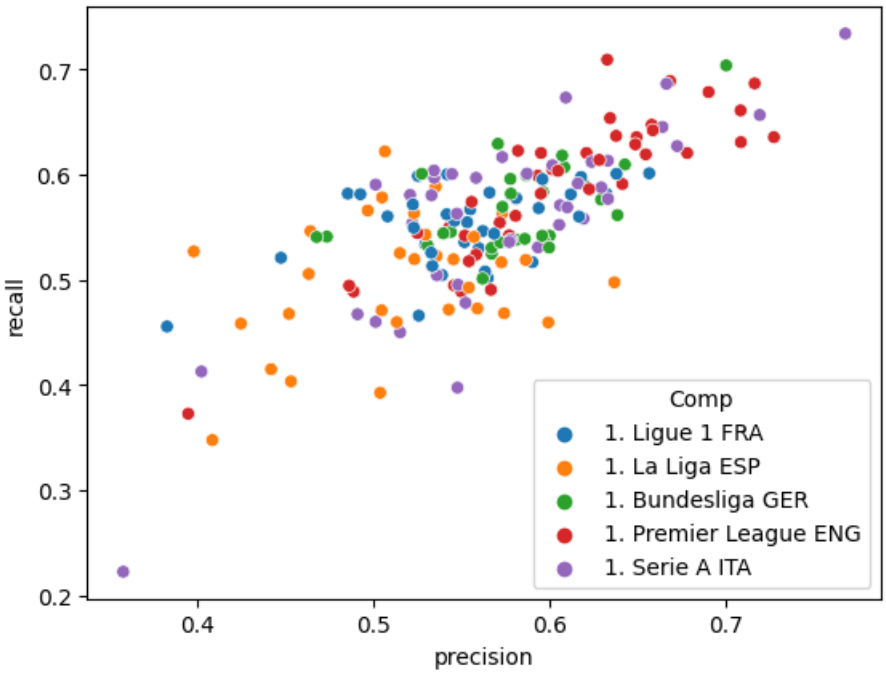


Figure 7: Scatter plot of Squads

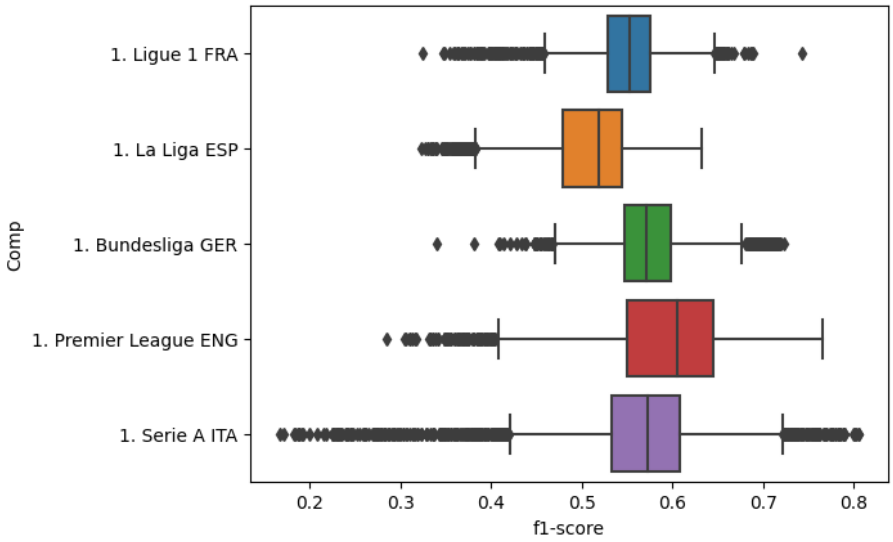


Figure 8: Box plot of Squads grouped by league

Our second classifier attempts to classify footballer's (nodes) country based on their career path of leagues played in along with the birth year, matches played, goals, network metrics, betweeness and closeness centrality. This model uses a random forest classifier to predict the nations. A random forest classifier is an ensemble which contains 100 decision trees each which classify based on a set of attributes. All the trees come together and the majority choice is picked as the label. Random tree classifiers tend to perform better than a single decision tree, especially on large data sets with a high number of features. We filtered our nodes on just the top 20 countries in our dataset so that the classifier would have a more concise set of labels all with a sizable amount of data points. We ran this classifier 100 times and the results are shown in table 3 and in figure 9 in form of a confusion matrix to show where exactly the classifier struggled.

Table 3: Node Classifier Table

Node Classifer	
Classifier	Random Forest Classifier
Features	# of Seasons, Squads, Comps Distribution of Comps, Matches played, Goals, Betweenness and Closeness Centrality
Accuracy	0.8001 +/- 0.0089
Recall	0.8001 +/- 0.0089
Precision	0.7708 +/- 0.0119
F-1 Score	0.7711 +/- 0.0105

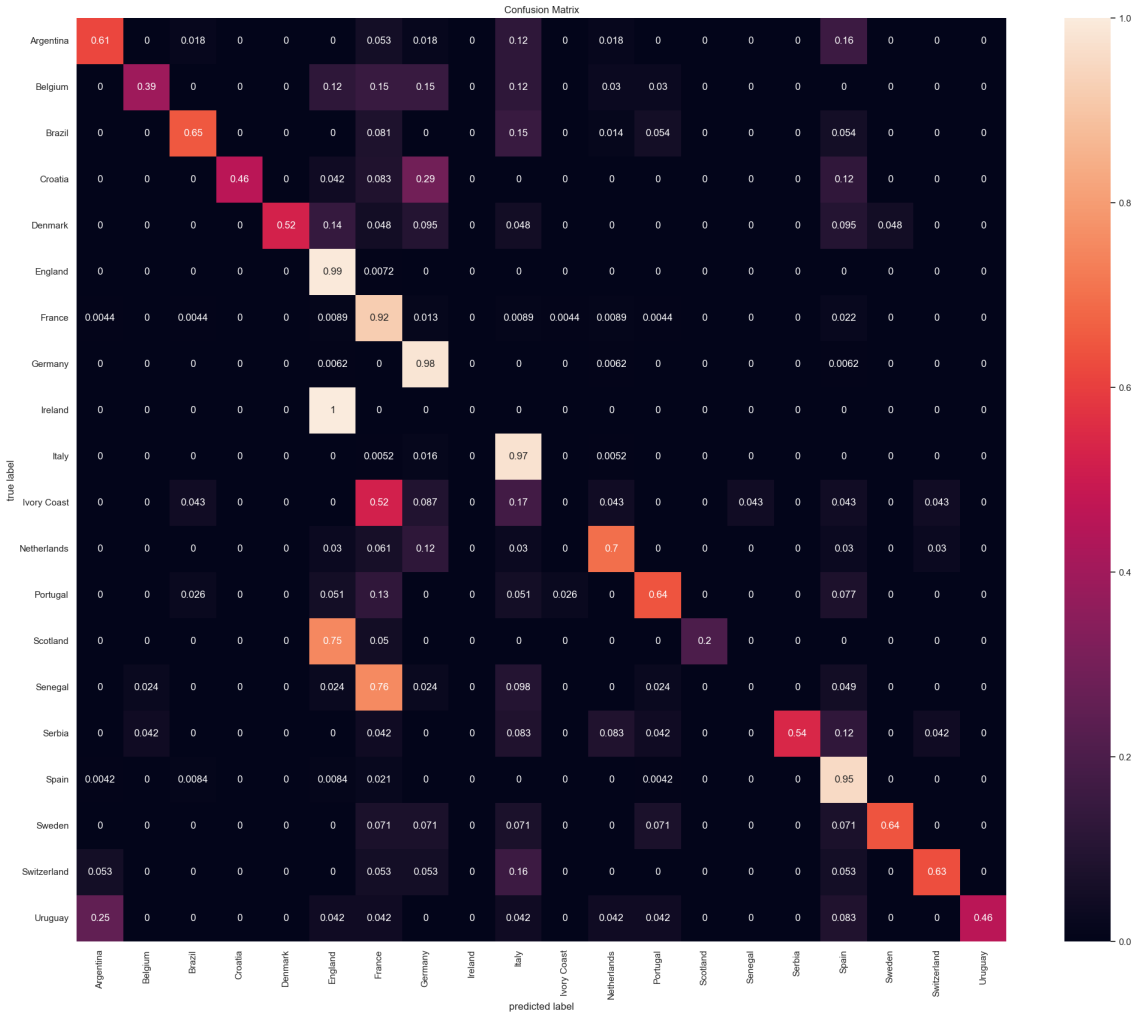


Figure 9: Confusion Matrix for node classifier

This classifier hovered around 80% on accuracy. When we take a look at the confusion matrix, overall the classifier performs well but does struggle in 2 main areas. The first being is incorrectly labelling Scotland and Ireland as England. This is likely due to Ireland and Scotland's close proximity to England, leading many of their nation's players to start out in the England football system. In general, our network has a high number of English players and the high number of English Players greatly outnumbers the Scottish and Irish therefore, the classifier tends to choose English when it is unsure about these 3 nations.

The second struggle point is African Countries, Senegal and Ivory Coast getting misrepresented as European countries, mainly France. There are a couple of different reasons for this. First of all, Senegal and Ivory Coast both lack a prominent football league of their own so many of their players move to Europe at a young age to continue their promising football careers. The second potential reason is many Senegalese players are actually born in France and other European countries but retain their citizenship of their parents in their home country. In this case, they are immediately integrated into the European soccer system therefore, our classifier has no way to distinguish them from a more dominant country like France in our data set.

This can open the discussion of the need to grow African countries' domestic leagues. African countries growing success in the football world has been on display at the 2022 world cup with great showings from Senegal, Morroco, Ghana, and Cameroon. With these teams outperforming the likes of world powerhouses like Belgium, Brazil and Spain the sky is the limit for African countries in football. However, the next step will be in building their own domestic leagues.

6.2 Community Structure and Link Prediction

In our next question, we look at the communities (shown in figure 6) of our network in greater detail and what drives them into these particular clusters. For community detection, we used the Louvain method described in section 8. This was very helpful for run-time due to the optimization aspects this function uses.

From running this function, our result had 5 different partitions. Furthermore, we ran the same algorithm on our 100 null model ensemble resulting in on average 8 communities with a standard deviation of 1.240. From this, we can conclude the 5 communities is significant to our specific network structure. We hypothesized from the fact we had 5 leagues of data and from using our own knowledge in investigating a selection of the nodes in these groups, that these clusters could correspond to each league.

To test this we ran the same metrics on our communities and our league sub-graphs. To strengthen our claim we view the country construction of both the sub-graphs and the community graphs to back up our community and league pairings.

To construct the 5 league sub-graphs we took only the edges between players in one of those top 5 leagues and created the 5 subsequent graphs. This meant we lost edges between players who had played together but not in that particular league, resulting in a subset of edges from our original network. In addition, because the edges are only for players in that league, players who had played in more than one of the top 5 leagues, were included in each of the sub-graphs they had edges in. However, their degree from the original network was split between these sub-graphs. This meant the most connected players in our sub-graphs were players who have played most of their careers in one league. The results are shown below.

Table 4: League sub-graph information

League	Max Degree	Min Degree	Mean Degree	Nodes	Edges
Serie A	363	16	81.011717	2475	100252
Premier League	222	19	59.652506	2354	70211
Bundesliga	249	18	60.441278	2035	61499
La Liga	263	17	64.066215	2507	80307
Ligue 1	269	17	62.079511	2616	81200

Table 5: Metric calculations on leauge sub-graphs

League	Country Similarity	Clustering Coefficient	AVG Shortest Path	Density	Domestic Ratio
Serie A	0.026123	0.634170	2.287405	0.032745	0.364035
Premier League	0.015250	0.689980	2.505569	0.025352	0.274207
Bundesliga	0.005721	0.684402	2.409154	0.029715	0.371359
La Liga	0.025144	0.672973	2.431889	0.025565	0.496296
Ligue 1	0.007823	0.679911	2.501902	0.023740	0.399101

From the above tables, it can be seen that there are no large differences in values between our leagues, except for a large number of edges in the Italian League and a low number in the German Bundesliga. We believe this is due to the relatively high number of transfers that occur in the Serie A with less player loyalty to clubs and Bundesliga being low due to having 2 fewer teams in the league compared to its four counterparts. This is supported by the table shown in figure 10 which shows the number of arrivals and departures that occur in the leagues between 2010 and 2022, where the Serie A has roughly 4 times as many.

Compact | Detailed

#	Competition	Expenditure	Arrivals	Income	Departures	Balance
1	Premier League	€18.27bn	5042	€8.72bn	5374	€-9,549.56m
2	Serie A	€10.69bn	11205	€9.18bn	11381	€-1,517.84m
3	LaLiga	€7.80bn	4096	€7.11bn	4418	€-685.35m
4	Bundesliga	€6.12bn	3087	€5.32bn	3442	€-790.53m
5	Ligue 1	€5.75bn	3538	€5.93bn	4084	€181.96m

Figure 10: Transfer Activity by League 2010/11-2022/23 [6]

We included two other metrics regarding country similarity and domestic ratio to inspect the nationalities of our leagues. The domestic ratio looks at the percentage of players whose nation matches the country of that league. La Liga is by far the largest at almost 50% meaning these teams are less likely to import players from international countries and tend to rely on homegrown talent. Premier League is on the other side of the spectrum as just 27% of their players are actually English. This speaks volumes about the international aspect of the Premier League and gives context on why it is considered the top league in the world as it is filled with international talent. 'Country Similarity' looks at the number of edges where both players share a nationality. This looks more at how teams are constructed and if they tend to go for players of the same nationality. A higher number means more same-country links.

We will use the basic metrics from our league sub-graphs shown in the tables above to compare to the communities the algorithm detected. If they are similar it will support our hypothesis that our communities are the 5 leagues. If they differ we will have the opportunity to look further into the structure of our community to find an explanation. Below are the results of running the same metrics on our leagues.

Table 6: Metric calculations on communities detected by algorithm

Community	Predicted League	Country Similarity	Clustering Coefficient	Avg. Shortest Path	Density
1	Serie A	0.043418	0.412141	2.071862	0.053713
2	Premier League	0.028718	0.380569	2.101809	0.052515
3	Bundesliga	0.032806	0.447319	2.167758	0.043466
4	La Liga	0.061011	0.432766	2.166448	0.041736
5	Ligue 1	0.014819	0.476477	2.253642	0.035277

Table 7: Community graph information

Community	Predicted League	Max Degree	Min Degree	Mean Degree	Nodes	Edges
1	Serie A	330	19	108.016899	2012	108665
2	Premier League	227	18	87.700778	1671	73274
3	Bundesliga	243	19	76.847937	1769	67972
4	La Liga	255	15	86.226415	2067	89115
5	Ligue 1	228	19	76.444649	2168	82866

When comparing the sets of two graphs it can be seen the patterns of the five leagues do tend to follow in the 5 communities in our graph. In particular, the number of edges is fairly similar and follows the same pattern of one community having a high number of edges compared to the rest as the Serie A has. When looking at the rest of the values calculated for both league sub-graphs and the community graphs we have attached a column for the predicted league we believe that community roughly corresponds to.

Compared to the league sub-graphs, generally, the community graphs have a higher mean degree and a lower number of nodes. The lower number of nodes comes from every node only being counted once as nodes can only be part of one community whereas a player could've played in all top 5 leagues thus being in all 5 league sub-graphs. A great example of this is the Premier League graph versus the community graph we hypothesize the Premier League represents. The league has far fewer nodes (1671 to 2354) and we believe this is because many international players come to the Premier League for a shorter time so even though they played in the league, they are not necessarily part of the community of players coming up through the English football system, hence the smaller home country percentage.

To further confirm the communities correspond to the predicted leagues, we took a look at the nation composition of every sub-league graph and community and the results are rather definitive (table 8).

Table 8: League/Community Nation Composistion

League or Community	Country 1	Country 2	Country 3	Country 4	Country 5
Serie A	Italy (913)	Brazil (168)	Argentina (148)	France (95)	Uruguay (69)
Community 1	Italy (909)	Brazil (131)	Argentina (111)	Uruguay (58)	Croatia (43)
Premier League	England (657)	Spain (144)	France (138)	Ireland (89)	Netherlands (78)
Community 2	England (657)	Ireland (90)	Scotland (71)	Wales (59)	Netherlands (57)
Bundesliga	Germany (765)	France (76)	Austria (74)	Netherlands (72)	Brazil (72)
Community 3	Germany (761)	Austria (72)	Netherlands (59)	Brazil (58)	Switzerland (57)
La Liga	Spain (1273)	Argentina (177)	Brazil (118)	France (100)	Portugal (85)
Community 4	Spain (1241)	Argentina (134)	Brazil (80)	Portugal (85)	Uruguay (50)
Ligue 1	France (1066)	Brazil (119)	Senegal (110)	Mali (83)	Ivory-Coast (73)
Community 5	France (985)	Senegal (96)	Brazil (91)	Mali (75)	Ivory-Coast (56)

For each league, there is a community which almost exactly matches the nation composition of each league. The only differences are other main countries are present in another league but not in the community. For example, the Premier League's second and third top nations are Spain and France however in the community, these nations are not in the top 5. As mentioned earlier, even though these players did play in the premier league, they most likely also grew up and played in their home country and therefore belong more to that community. This further explains the lower number of players in the Premier League community.

The high number of players that come to the Premier League from other leagues is shown below in the table where we analyze the inter-community links between our communities with the predicted league labels attached. From this, we see the Premier League is in 2 of the top 3, which can further explain the fewer number of nodes, the lower domestic ratio and the fact that Spain and France are missing from the top countries in the predicted Premier League community (community 2) country composition table.

Table 9: Inter Community Edge Combinations

Count	Community1	Community2
17264	La Liga	Seria A
16959	Premier Leauge	Ligue 1
16445	Premier Leauge	La Liga
14049	Seria A	Ligue 1
14046	La Liga	Ligue 1
14008	Bundesliga	Premier Leauge
13613	Seria A	Premier Leauge
11224	Ligue 1	Bundesliga
10926	Bundesliga	Seria A
9787	Bundesliga	La Liga

Taking all of these results together, we believe that the core players will typically tend to stay in and around the players in their league. This also could be due to the fact that all of our 5 major leagues also have lower divisions with teams that feed into the top division. These teams tend to not have the financial power that the top teams do to sign players from other countries and rely much more heavily

on domestic players through local scouting. (use citation for this)

This can be seen from our feeder table showing which leagues dominate our edges.

Table 10: Top 15 Feeder Leagues (Proportion of Players Played in League)

League	Number of Seasons	Unique Players	Country	Proportion
Premier League	11747	2708	England	27.92%
Ligue 1	9881	2664	France	27.46%
Serie A	10869	2517	Italy	25.95%
La Liga	9883	2508	Spain	25.86%
Bundesliga	8972	2125	Germany	21.91%
Championship	5889	1550	England	15.98%
Segunda División	5124	1485	Spain	15.31%
Ligue 2	3996	1359	France	14.01%
Serie B	5428	1357	Italy	13.99%
Bundesliga	3575	1131	Germany	11.66%
Super League	1604	731	Greece	7.54%
Süper Lig	1837	714	Turkey	7.36%
Primeira Liga	1992	616	Portugal	6.35%
Eredivisie	2097	608	Netherlands	6.29%
First Division A	1001	461	Belgium	4.75%

This table shows the top 15 leagues that players have played in at some point in their careers. It isn't until the 11^{th} row we find a league outside of our five main countries. This reinforces the point that a player's career path to one of the top 5 leagues is seen more through the lower division in one of Germany, Spain, England, France or Italy rather than another country.

However, when looking at our table above we see the 5 feeder leagues (2nd divisions) to our top leagues are the next most played in the league from all our player's career paths. Therefore, in our sub-graphs all of these links are removed reducing the degree of all nodes who have played in these divisions, thus resulting in a lower average degree for the whole league sub-graph. In the communities, these edges are left in but only if the two players are in the same community. We believe this is the cause for the higher average degree as we know from table 10 these feeder leagues are the next most common leagues for our players (nodes) to play in. This implies that clubs will tend to look at domestic players that have been tried and tested in their competition before looking abroad.

From this, we can conclude that yes, our generated network communities correspond to each of our top 5 leagues. Taking this further, they represent the 5 complete football league structures in each of the 5 top nations in our dataset as second and third-division leagues play a large part in defining communities and bringing players back to their home communities who have played in multiple leagues.

Following on from the previous question regarding predicting aspects of our network, we wanted to see if the similarity of neighbours in our network could also be used to predict future teammates. In an attempt to test the idea that players will tend to look to stay in their neighbourhood and will look to play with similar players their teammates know.

To test this we used a link prediction algorithm from the NetworkX python library called Jaccard Coefficient. Which is a coefficient assigned to two nodes (link) that is calculated by dividing the intersection of all the two nodes' common neighbours divided by the union of both of their total neighbours.

The higher the value (0-1) the more common neighbours they have, and the lower the number the fewer neighbours they share.

To apply this technique we first had to split our graph on a date, we choose 2015 since this roughly

split our network in half in both edge size and by the number of seasons we collected.

We then calculated the Jaccard Coefficient on all node pairs without a link in our pre-2015 graph (all players who played together before the 2015-16 season). Filtering any player with a degree less than 20 out, since we wanted players who had played on at least one team.

This returned a coefficient list of all pairs of nodes in our graph that were not already connected by a link (had not played together). This returned 22,676,467 rows of node pairs, with 7667313 having Jaccard values above 0. Meaning we had 7,667,313 node pairs that shared at least one neighbour, 53891 sharing at least 10%, 1820 sharing at least 25% and 41 with at least half. Our maximum pair sharing is 65%.

We then wanted to use this list of returned edges with higher Jaccard values to see if players who share a large number of common neighbours that haven't played together will eventually play on the same team in the future.

Our methodology for this was to filter our Jaccard values by certain benchmark values and use the filtered lists to compare with our post-2015 edges. The number of times we found a match was the number of correctly predicted future edges. Running this test was to see if any of these players who had high overlap in their neighbourhoods (high Jaccard values) without playing together would end up becoming teammates in the future.

Table 11: Table Showing How Well Jaccard Values Predict Future Teammates

Jaccard Value	Number of Predicted	Correct Predictions	Percentage
≥ 0.10	53891	1542	2.86%
≥ 0.25	1820	169	9.29%
≥ 0.40	210	23	10.95%
≥ 0.50	41	5	12.20%

From these results, we were not able to conclude this is a particularly accurate measure for future link prediction. We believe we would need to refine this model and take it further before making any claims regarding predicting future teammates based on similarities.

With more time we would like to investigate these results further to see how many of these predicted teammates are from youth players to senior team players, where the young player was playing in a youth team prior to the date we split our data. As these common neighbours would be other young players playing in the senior squad at the same club. However, these young common neighbours might just be players who are one or two years further in their development than the predicted young player in the edge.

A possible way to improve this algorithm would be to include similar countries, national teams or youth teams. Providing this extra information could help in identifying players who know each other from their national teams or from previous youth academies. Adding this information could uncover possible friendships or interactions where a player could influence another decision on joining a new team not included in our network. This information could help in better predicting who is likely to play together in the future.

Despite the results not being as significant as we liked, we found it interesting to investigate the probability of a player sharing teammates with other players they haven't played with as a possible indicator for link prediction to be a worthwhile approach.

6.3 Information Cascade and Age Influence

In our next question, we tackle two important topics; how information can spread throughout our network, how much influence exists in our network along with how it can spread. Football is also a very unique occupation as all professional sports are. It is extremely team-oriented and at certain times of the season, teammates will spend more time together than they will with their families. This creates a very intense environment, where players spend the majority of their time training, travelling, playing or just generally interacting at the training ground with their teammates. With so much time spent together, information and rumours are very likely to spread. Younger players will also tend to look up to older players and be influenced by them.

To begin, we first look at information spread. To analyze this we modified an information cascade algorithm template that we were given in class by our professor Dr. Towlson. This algorithm is included in our notebook and works by taking in a graph and an integer greater than 0 specifying the number of rounds to run the information spread and a list of initial nodes that are said to be activated. Activated in our context is to say the information reached these nodes.

In our information cascading algorithm, we use certain criteria to increase a player's chance of spreading information to that player. We based this on common factors in how friendships are built. Time spent together, if they have played on multiple different teams together and if they are from the same country (especially important for language).

How we approached this question was to use 3 groups of 5 individuals, one of which is a group of known players for their sportsmanship and charity work. A group of negative individuals known for bad sportsmanship and in some cases criminal activity, a group of players from the Barcelona team around 2008-2018 who were incredibly successful and played for very few teams outside Barcelona. Lastly, we generated 1000 groups of 5 nodes to compare our chosen groups to.

Below we outline the individuals we have grouped with a reference to the justification for the choice. We would like to make it clear, although we provide references to justify our choice of which player is in which group. These remain our own opinions and should not be taken as fact.

For our group of positive individuals we used:

- Juan Mata: Charity [7]
- Didier Drogba: Diplomatic and Charity [8]
- Xavi: Grand Cross for Sporting Merit recipient [9]
- Marcus Rashford: Charity [10]
- Miroslav Klose: Fair play award [11]

For our group of negative individuals we used:

- John Terry: Infidelity [12]
- Luis Suarez: Biting and unsportsmanlike [13]
- Neymar: Excessive diving [14]
- Ryan Giggs: Assault [15]
- Benjamin Mendy: Sexual Assault Charges [16]

The Barcelona group we choose:

- Xavi (1998–2015)
- Andres Iniesta (2002–2018)

- Lionel Messi (2004–2021)
- Gerard Pique (2008–2022)
- Carles Puyol (1999–2014)

In addition to our chosen groups, to look at how information cascades in relation to degree throughout our network we took the last 10 smallest degree nodes and 10 highest degree nodes and used 1000 random degree nodes. In this approach, we run the algorithm for multiple rounds to evaluate how much propagation exists amongst the node groups. Below are our results:

Table 12: Information Cascade Algorithm (Nodes hand-picked)

Player Group	Individuals Reached	Direct Neighbors (Degree)	More Than 5 Seasons Together
Positive Individuals	112.99 +/- 8.324	637	42
Random Group (1000)	100.484 +/- 27.405	583.097 +/- 136.868	6.142 +/- 6.124
Negative Individuals	115.96 +/- 9.107	701	52
Successful Core Group	94.24 +/- 6.734	637	34

From the table of results detailing our chosen players, we see that there isn't too much variation in the number of individuals that spread to. However, there are still some interesting takeaways, specifically the degree impact. The group of players we choose are fairly similar in degree with the negative individuals having the highest and also the highest number of individuals reached. Interestingly the successful core group and the positive individuals we choose have the same degree, but the core group has reached a fewer number than the positive group. We believe this could be due to the fact that the core group we choose played together for such a long time and played with similar players since the Barcelona team at this time was very dominant and kept the player group together. This means they have fewer unique players to reach than the other group with a similar degree since they did not play together for the same length of time. This is why we see a fewer number in the '> 5 seasons' column since the links these players have outside of the Barcelona team will not be for very long since the majority of them played for Barcelona for their entire career. This also makes sense as these players will not have as large ego networks as other players who have moved around more. Since each time you transfer to a new team, you are adding roughly 20 new players to your network or strengthening links if you have played with them before. For these players who played for the same team for years, will only add players to their network when new players join their club, which is very unlikely to be more than 20.

This is why we believe the best information spreaders are groups of individuals who have a high degree (lots of teammates over their career) and played for more than 1 team to create a large ego network but also to choose players who played for teams for longer than 5 seasons so they have the opportunity to integrate themselves into the teams and build relationships.
Using the random node groups helped us come to this conclusion as the degree and the number of '> 5 season' was the lowest compared to the others, thus, giving the lowest reach value. In the next table we look at the network as a whole:

Table 13: Information Cascade Algorithm (Nodes chosen on degree)

	Top 10 Degree	Random Group(1000)	Bottom 10 Degree
Direct Neighbors (Degree)	3794	1153.65 +/- 186.115	224
1 Round	521.71 +/- 18.013	196.46 +/- 36.779	40.31 +/- 4.772
2 Rounds	5059.07 +/- 114.009	2999.904 +/- 459.504	508.39 +/- 82.176
3 Rounds	9532.03 +/- 24.67	9068.73 +/- 245.142	4811.99 +/- 556.925
4 Rounds	9690.74 +/- 3.279	9687.508 +/- 4.205	9556.81 +/- 76.135

In this table, we looked at the most highly connected players (10 highest-degree players) and the least connected (lowest 10 degree players). We also generated 10 random player groups in our network to compare the results. To analyze the impact reach of our network we ran the algorithm for 1-4 rounds for each

group. Running the algorithm on 1 round, only checks the direct neighbours of each player, running for 2 rounds, checks the neighbours of the set that were reached by the first run of the algorithm. Then for 3 rounds, they will check the neighbours of the individuals reached at round 2 and so on for multiple rounds.

We ran this to test how fast information propagates throughout the network and how many rounds it would take to reach the whole or a very large percentage of the entire network.

From the table, we see after the first round that the top 10 degrees perform the best with the random being average and the bottom 10 only reaching around 40. This ordering trend continues for the remaining rounds. However, what is interesting is that after 2 rounds the random groups are still around 2000 fewer than the top 10, yet, when running for the third round they both get very close to the whole network of 9700. The same thing occurs between the 3rd and 4th rounds for the bottom 10, the bottom 10 group reaches around half the network at round 3 and almost the entire network by the last.

This indicates our network needs around 3 rounds on average to spread information to the whole network for degrees with around average degrees (115) and 4 rounds for very low degrees (20). When looking further into our bottom 10 degrees to work out how they manage to reach a large percentage of players after 2-3 rounds we looked at the ages of these players and found them to be very young players majority born between 2001-2004.

This means they are players who are breaking through the academy and into the first team at a very young age. What this means for their links in our network is they have a small number of direct links with players who are closer to the average degree and thus, when running the algorithm for multiple rounds we start to reach more of the network through these more prominent players.

This algorithm also gives an application for our shortest path value, which we found to be around 2.5, which gives these rounds some context in that our nodes on average can reach each other in 2-3 hops.

Next, we look into how influence can spread in our network. As we mentioned in our introduction, football teams have large amounts of variation in the age of their players. Due to this variation, it is very possible young players can end up playing with their childhood hero. As players at the highest level, younger players will want to look at older more senior players as examples. This makes it vital for senior players in a team to conduct themselves in a professional manner to set a positive example and club culture for the rest of the team and especially the young players.

These examples include being on time, working hard for the team, respecting the coaching staff and staying out of trouble on and off the pitch. Unfortunately, not all footballers in the past have met this standard.

In more serious cases, some footballers have been involved in criminal activity. Including illegal betting, sexual assault, assault and driving under the influence just to name a few. Using our network we want to analyze a few sets of individuals and groups of individuals to see how much of a positive or negative impact they have on the players around them using some common influencing metrics that make a football player more senior. In addition, we want to see how fast information or rumours could spread through our network using similar metrics.

When looking at our age algorithm, which was very similar to our information cascading algorithm, the only difference was how we assigned our multiplier. For older players with a 5 or 10-year age gap between a teammate, we greatly increased the multiplier and then also applied the same metric of the number of seasons as this is still important for how players interact and more time together results in a higher probability for influence. It is important for good behaviour to spread across the team, older players need to quickly integrate younger players into the culture of their team. We wanted to look at this for a select number of nodes to see if there is potential for a positive or negative impact from possible troublemakers.
Below are our results from this investigation

Table 14: Age Influence Algorithm Results

Player Group	Influenced Individuals	Direct Neighbors (Degree)	Greater than 5 Year Age Gap
Positive Individuals	143.33 +/- 9.388	637	108
Random Group (1000)	96.789 +/- 28.978	577 +/- 128.47	39.86 +/- 31.867
Negative Individuals	161.73 +/- 10.966	701	154
Successful Core Group	115.75 +/- 6.704	637	114

From looking at the table we can see the negative individuals have played with the highest number of players they are 5 years older than or more and as well as having the highest degree it is not surprising they influence the highest number of teammates. However, when looking at the proportion of their neighbours they have influenced they both influenced roughly 22% of their teammates. Due to all players in during this time in both groups being quite well-known and successful players in football, it is both comforting and concerning. On one hand, it is encouraging to know players who are known for advocating in favour of good sportsmanship and charitable organizations have the ability to influence a great number of the players they play with, especially youth. However, when looking at the other side knowing that players whose morals on and off the field have been questioned are still seen as club legends and idols to many young players, it is troubling to know they too can impact players at the same rate.

When looking at the successful core from Barcelona we see fewer players influenced, this is again likely due to the fact these players have a much larger overlap in their teammates. With only 259 unique players were played compared to 637 (negative group), 507 (positive group) and an average of 576.17 for the random group. This shows the impact players playing together for so long can be. This group of players widely regarded as some of the best players to ever play the game were also luckily seen as very good role models. This is likely to have contributed to a large amount of success they had in winning all major trophies in club football. Barcelona during this time also relied on an extremely talented and world-renowned academy[17], which all of these players came through. This meant during this time a lot of the squad was made up of generations from this academy. Creating a squad of players who had not only played together for many seasons but also with age gaps.

This again reinforces our point that creating teams with players who have played together for a long time mixed with groups of young players can be extremely beneficial in this case as this was a very successful era for Barcelona and the Spanish national team. Although, this could have the reverse effect if there is a core group of players in a team that create a negative atmosphere and this is spread to newer younger players coming through that continue this negative club culture.

When comparing our results to the random group we do see lower numbers of individuals impacted with around only 16% of teammates influenced. Although with only approximately 40 cases of more than 5-year age gaps this does still show that age is not the only variable impacting the results, with the number of teammates and the number of seasons played together still an important factor. With also the highest proportion of unique players played it does highlight that a collection of random players were not as effective as the groups we choose, which is to be expected considering we chose high-profile players to investigate.

Taking this algorithm one step further we look at a specific ego network for one player we outlined in our negative group, Ryan Giggs. Ryan Giggs is being accused of using controlling and coercive behaviour as well as assault. He is currently facing trial and it should be noted that at the time of writing has not been proven to be guilty. We are just using him as an example of how a high-profile player who played at the elite level for such a long time could create concern for teams with younger players.

When running the algorithm just on Ryan Giggs, we find 48 individuals that he played with where there is a 10-year age gap. Our algorithm returns 29 individuals he influenced, of those 29, 17 had a 10-year or more age gap with Ryan Giggs. We have attached the table below of these individuals. A lot of these players are well into their 30's now and have degrees in our network greater than our average.

Table 15: Ryan Giggs Player Reach

Name	Country	Current Age	Degree
Nick-Powell	England	28	191
Phil-Bardsley	Scotland	37	226
Danny-Simpson	England	35	254
Jonathan-Spector	United States	36	146
Shinji-Kagawa	Japan	33	183
Anders-Lindegaard	Denmark	38	129
David-de-Gea	Spain	32	149
Fraizer-Campbell	England	35	264
Marouane-Fellaini	Belgium	35	146
Gerard-Pique	Spain	35	185
Chris-Smalling	England	33	175
Fabio	Brazil	32	235
Carlos-Tevez	Argentina	38	171
Ritchie-De-Laet	Belgium	34	233
Darron-Gibson	Ireland	35	164
Zeki-Fryers	England	30	166
Will-Keane	Ireland	29	198

Ryan Giggs was part of a very successful group of players that went on to win a great deal of trophies in the '90s and 2000s. He was a hero for a lot of young players and kids that are now senior players in a lot of teams today in 2022. Whether he is guilty or not of the allegations he is standing trial for, it shows how the culture senior players today in the modern game were exposed to in their youth careers could have been very negative. One hopes this sort of behaviour was not taken on board by these young players although, this can not be guaranteed with players today still being accused of such disturbing behaviour[18].

A potential use case for this sort our algorithm and data set could be for clubs or governing bodies of football looking to identify possible vulnerable players that might have been influenced to hopefully attempt to prevent this kind of behaviour in the future.

On the other hand, we hope this could be used for possible detection of players passing on positive habits. If a club is looking for a strong leader for its youthful squad, this could be used to find senior players who have played with successful youth players in the past.

The results from our influence algorithm have highlighted the impact a core group of players can have on a team, due to any new nodes joining this team having multiple chances to be influenced by the group of nodes. With a higher chance if they are a new younger player. In addition, we show that high-profile positive and negative individuals with longevity in our network have a high level of reach to be able to influence their teammates highlighting a professional athlete's responsibility to uphold good moral standing.

The results from this investigation raise a point from a much larger issue which is the importance elite professional athletes have to set an example for not only fans of their team or the sport, but also for their colleagues. The majority of professional athletes were once fans and will have had heroes growing up that they might one day play on the same team as. When a young player is starting to be integrated into the first team squad from their youth team, they will look up to the older senior players for how they carry themselves to try and one day reach their level. This means whether these talented players like it or not, they have a responsibility to act as role models on and off the pitch.

6.4 Generative Network

To conclude our results, towards the end of our research into this network, we wanted to find a way to conclude and bring all of these analysis results together to try and find what drives the football social network at the highest level.

To do this we attempted to create our own network, using defined functions for how our links are created. This was predominately done in 4 steps:

- Creating the initial network as it would look like in 2010
- Simulating the transfer marketing of players switching teams or retiring
- Simulating promotion and demotion by removing whole squads from the network and adding new ones
- Simulating new players entering the Top 5 leagues

Below we detail our steps in this investigation

1. We began with an initial set of 98 teams with 22 players that form a fully connected graph. We then estimated based on averages in our network how many players would be connected to the others in the rest of the network based on their history before 2010.
2. In step 2 we worked on simulating the transfer window to account for the movement of players to new teams or retiring. The transfer window occurs twice a year (once during the summer and for one month in January) in which a player can move to a different team upon agreement between the two clubs and between the new club and player. This transfer creates new links for this player to all of his new teammates.
3. We then simulated promotion and relegation by removing whole squads from the network and adding new ones. In football, at the end of each year, the bottom three teams in each league get relegated to the league below, and the top 3 teams from that league join the new league. This involves adding new cliques to our network while removing the old cliques so they stop gaining edges with players who have not played in the top 5 leagues.
4. The final step is adding new players who make their Top 5 debut in a season. Football has high turnover and there are always new young players joining top teams looking to make an impact. In our network, this involves adding a node and then adding edges to every teammate of the team he joined.

The first step gets run once then each of the next 3 steps gets simulated 13 times each, once for each season in our dataset. Each step is based on certain probabilities so to limit randomness in our result, we ran the model 100 times then took the averages of important metrics and compared them to our network.

Our results are shown in the table below

Generative Model Results		
Statistic	Our Network	Generative Model (100)
Average Shortest Path	2.53983	2.42027 +/- 0.007
Clustering Coefficient	0.34821	0.25496 +/- 0.001
Number of Nodes	9700	9700 +/- 0
Number of Edges	560416	550966.51 +/- 5493.91
Min Degree	21	22.54 +/- 2.26
Max Degree	423	483.18 +/-17.62
Average Degree	115.55	113.60 +/- 1.133

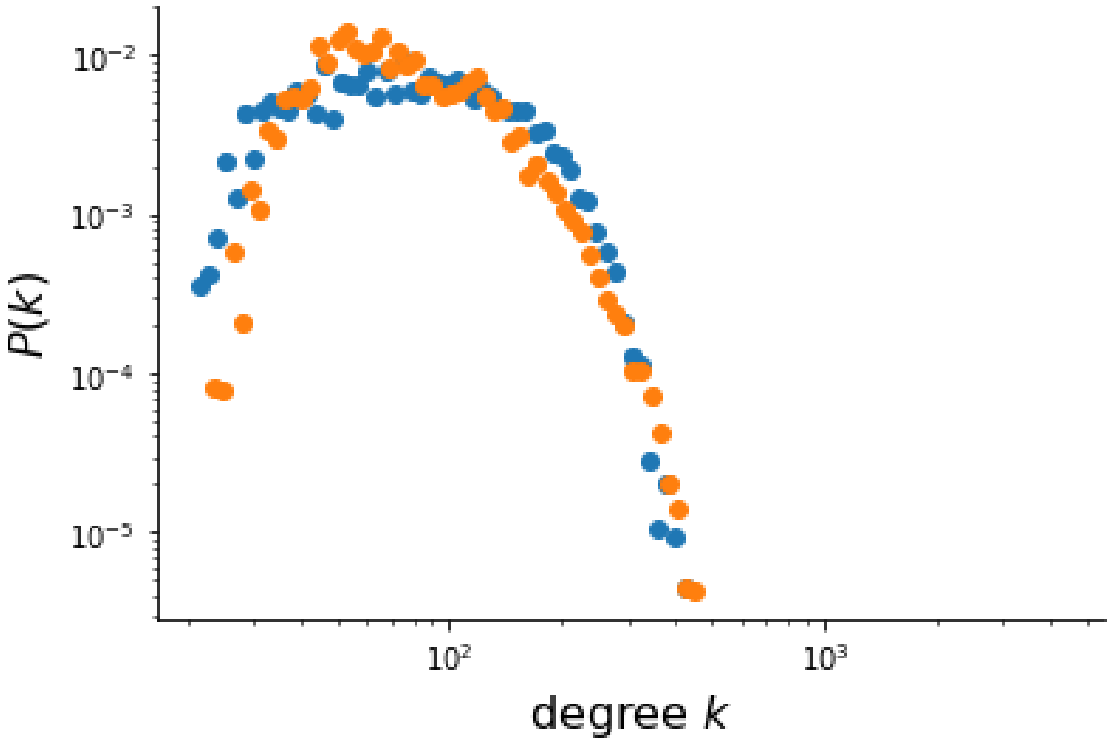


Figure 11: Orange is Generative Model and Blue is the Original Network

The biggest strength of this model was the fairly accurate reproduction of the degree distribution. It follows the same Poisson distribution as our network and the min and max degrees are similar. The model could be improved by further analyzing the transfer market and recreating trends as opposed to basing it on probability. This would greatly improve the clustering coefficient which is the main weakness of the model. Having said that, this is still a very high clustering coefficient compared to our original null model (0.0187). The shortest path is also a little shorter than our network but reducing the randomness of the transfer network would most likely improve this, as well as nodes when then be more likely to stay in one part of the network. Adding age and nation would also help the model with the shortest path and cluster coefficient as well. Since by adding ages to our nodes we could better predict the number of previous players they have played with in their career history, the likeliness of retirement and the chances of a transfer.

Taking this generative network further, it would be interesting to look at the different community structures that exist and how similar to our network, this would be. We leave this as a possible future investigation for ourselves or anyone taking our research further.

7 Discussion

From our results, it can be seen that the footballing world is quite complex, especially as each player more often than not does not choose who they will be interacting with. Although, despite an originally quite misleading degree distribution and a messy blob for a network, we believe we have shown there is a structure behind the social football network that can be explained. We did this through community detection, centrality measure analysis, information and influence algorithms on nodes, link prediction and prediction of node and edge attributes based on data. All of this was combined to try and recreate our network to find the attributes that most impact the structure.

During our analysis, one of the biggest things we noticed was how connected some players are and the very few hops that are needed to find the other players. Despite being a sparse network, we found this an interesting property. In addition, we found it quite remarkable how well a team and country could be predicted based on certain personal and career data. It was interesting to try and see if some teams had

certain profiles they look for when building a style of play.

We were very surprised not to see more clusters, we had expected to find denser communities in relation to certain successful teams. This was not the case as we found this was more based on the league. We believe this was due to the high number of average teams a player will play for in their career.

When looking at the overall success of our research, we are satisfied with our results and the groundwork we have put in for future work for ourselves or others. We are especially happy with the dataset we created from scratch and how we built this network without much vision of how the final project would turn out. The project has taken a very different path from what we had both imagined at the start, with our limited knowledge of Network Science. However, as we learned new techniques in the Network Science field we think we were able to refine some of the research questions and add some completely new ones we had not thought of at the beginning. These new questions allowed us to dig deeper into the overall structure of our network as opposed to focusing on subsections of our network and relying on other data analytical approaches not specific to Network Science.

Some limitation of our dataset is that we do rely on the quality of data from Fbref. Due to the large amount of data we scrapped, we did not check every single player's history to make sure it was in fact correct. Another possible quality issue could come from our selection criteria, we do not take into account the number of games a certain player has played for a team in all cases. This was largely due to the fact a young player will likely not play many games for a team but will still be seeing his teammates every day. Thus, by filtering for matches played we would lose these key youth player interactions.

In terms of our approach, some of our techniques and algorithms use metrics that we have personally decided are the best based on our own knowledge and things we found out from our network. A change in some of our values or features used to find relationships could create different results.

In addition to some of these issues, a possible limitation of our research is the consideration of only men's football data. In the coming years as Women's football becomes more popular we believe that a similar style project could be done and some interesting comparisons could be made between the two different independent networks.

We hope our work on this project will be taken further by other passionate sports fans and if nothing else we hope the dataset scraped can be used. In the future, we would like to further improve our model to recreate our network along with digging deeper into finding some smaller communities and neighbourhoods to better predict and analyze relationships.

Other future work we are interested in, is looking at other sports to see how unique the football world is compared to for example the basketball network or the ice hockey network. We would like to test if similar techniques or results would reappear in another sporting network.

8 Methods

In this section, we detail and credit the software libraries we used, with an explanation of the specific methods we used. Along with the library, we explain what basic statistics and calculations the library was used for.

Table 16: Software Used for Methods

Method	Software Used	Explanation
Scraping	Beautiful Soup	beautiful soup is a python package made for parsing HTML and XML documents and extracting data from them. When scraping Fbref.com we followed a tutorial by Michael O'Donnell.
Data Cleaning	Pandas and Jupytar Notebook	Pandas is a powerful data analysis tool built on top of the python language. It is best used in a Jupytar notebook which allows for an interactive python experience
Network Visualization	Gephi	Gephi is an open-source network visualization software that also has functionality for network analysis. We strictly used Gephi for visualizing our network.
Network Analysis	Networkx (NX)	Networkx is a python package for network creation manipulation and analysis. All of our work related to networks (besides visualization) was done in Networkx. This includes the shortest path, cluster coefficient, and centrality measures.
Community Detection	NX Louvain Method	The Louvain method was created by Vincent Blondel. It is a greedy algorithm designed to extract communities from large networks. The method has been implemented in networkx
Edge Prediction	NX Jaccard Coefficient	The Jaccard coefficient is a statistic for evaluating the similarity between sets created by Grove Karl Gilbert in 1884. In the network environment, the method is implemented in Networkx and can be used to give a percentage of shared neighbours for two nodes.
Classifiers	Scikit Learn	Sci-kit learn is a machine language python library which includes methods for implementing decision trees and random forest classifiers.
Box and Scatter Plots	Seaborn	Seaborn is a data visualization python library which is great for creating plots right inside of a Jupytar notebook.

All of our code can be found at on GitHub `https://github.com/OwenH/The-Football-World-as-a-Social-Network`. We have structured GitHub to have 1 interactive notebook for each part of our research paper. The notebooks are listed below. The GitHub also includes .csv files and raw notebooks used in the data wrangling process.

- Data
  - Raw .csvs collected from scraping
  - Combined .csv of all players
  - .csv of all scraped player histories
  - Node and edge .csv files
  - 100 randomly generated graphs stored as edge .csv files
  - A collection of .csv files containing metrics such as Centrality measures, Jaccard values, etc...
- Notebook used for scraping data
- Notebook used to construct our network
- Basic Statistics Notebook

- Classification Notebook
- Community Analysis Notebook
- Information Cascade and Influence Notebook
- Network Simulation Notebook

9 References

1. https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html
2. https://www.skysports.com/football/news/11899/12029767/jude-bellingham-signs-for-borussia-dortmund-from-birmingham
3. Ranchordas, Mayur & Dawson, Joel & Russell, Mark. (2017). Practical nutritional recovery strategies for elite soccer players when limited time separates repeated matches. Journal of the International Society of Sports Nutrition. 14. 10.1186/s12970-017-0193-8.
4. Pablo Medina, Sebastián Carrasco, José Rogan, Felipe Montes, Jose D. Meisel, Pablo Lemoine, Carlos Lago Peñas, Juan Alejandro Valdivia, Is a social network approach relevant to football results?,Chaos, Solitons & Fractals, Volume 142, 2021, 110369, ISSN 0960-0779, https://doi.org/10.1016/j.chaos.2020.110369 (https://www.sciencedirect.com/science/article/pii/S0960077920307633)
5. Kapanova, K. (2014, June 8). Football Transfers looked from a social network analysis perspective. https://www.academia.edu/2046327/Football_Transfers_looked_from_a_social_network_analysis_perspective
6. https://www.transfermarkt.us/transfers/transfersalden/statistik/plus/1sa=&saison_id=2010&saison_id_bis=2022&land_id=&nat=&kontinent_id=&pos=&w_s=&plus=1
7. https://juanmata8.com/en/common-goal/#
8. https://www.cnn.com/2022/11/16/football/drogba-globe-soccer-off-the-pitch-ctw-spt-intl/index.html
9. https://www.marca.com/en/2015/07/17/en/football/spanish_football/1437144869.html
10. https://fareshare.org.uk/marcus-rashford/#:~:text=FareShare's%20work%20fighting%20hunger%20in,and%20at%20times%2C%20food%20banks
11. https://sports.ndtv.com/football/germanys-miroslav-klose-wins-fair-play-prize-1546540
12. https://www.wsj.com/articles/SB10001424052748704259304575043212033975040
13. https://www.insider.com/world-cup-2022-luis-suarez-bites-players-2022-11
14. https://www.sportingnews.com/us/soccer/news/fake-injuries-neymar-world-cup-2022-players-dive-flop-simulate/bpd34zcdycdgj6reckd0dlhx
15. https://www.bbc.com/news/uk-wales-57280487
16. https://www.bbc.com/news/uk-england-manchester-63666378
17. https://www.espn.com/soccer/league-name/story/2113963/headline
18. https://www.bbc.com/news/uk-england-manchester-63699399