Analysis of Pass Distribution Networks on Football Teams

André Diegues

July 15, 2016

DCC/FCUP

Abstract

This article studies the purpose and importance of the pass distribution in professional football teams. We'll analyse matches of the *UEFA Champions League* throughout the seasons from 2012/2013 to the present day. This article aims to discover things like: pass distribution patterns, the most influential player of a team, how is the pass distribution network related with the strategic approach of a team and so on. It also describes how every step of the project was approached and dealt with.

Contents

1 Introduction

A network is a set of vertices or nodes, with connections between them, called edges. Systems that take the form of networks are designated by graphs and almost everything can be represented in this form. Nowadays, networks fit to represent some of the most important things in the world, like social networks like *Facebook* or *Twitter*, biological networks such as neurological networks, information networks like the network of citations between academic papers and, lastly, technological networks which are all networks that are implemented by the human kind in the world such as water networks, transportation networks and even communication networks[5].

In this article we're going to represent through networks the pass distribution of football teams in the high-level of the sport looking specifically to the best European competition, the *UEFA Champions League*. We are hoping to find patterns in the networks that can reveal, for example, who is the most influential player of a team, how the absence of that player impacts the performance of his team and also if the team's strategy is perceivable from the amount of passes made per match.

2 Building the Networks

In order to represent all the teams in the *UEFA Champions League* we had to get the pass distribution of these teams. Fortunately all the data from the season 2012/2013 till the present season was available on the *UEFA* official website[1] in the total pass distribution .pdf files. We can see Figure 1, in the following page, as an example of these files.

Firstly we need to decide what data is relevant to collect. Throughout the process of collecting the data we conclude that we could derive more information than what we actually needed. So the process of choosing what is or what is not relevant was not easy. We decided that the kind of pass (short, medium or long pass) was not relevant because we could not identify to whom that pass was made or the percentage of success because we are not studying the accuracy or talent of a player to pass the ball. In other hand, we had to create some data from the collected one, for example, the average passes of a player to another player per match had to be calculated in order to make comparisons with teams that played less or more matches than others. We also kept track of the transfers made throughout the seasons in order to evaluate if the performance of a player (with respect to his passes made per match) is, somehow, connected to the change of team in a player's career.

[1] `http://www.uefa.org/mediaservices/presskits/uefachampionsleague/`

UEFA CHAMPIONS LEAGUE

Passing Distribution

Semi-finals 1st leg - Wednesday 6 May 2015

Camp Nou - Barcelona

FC Barcelona 3 - 0 FC Bayern München

From		TP	To: Marc-André ter Stegen 1	Gerard Piqué 3	Ivan Rakitić 4	Sergio Busquets 5	Andrés Iniesta 8	Luis Suárez 9	Lionel Messi 10	Neymar 11	Javier Mascherano 14	Jordi Alba 18	Daniel Alves 22	Xavi Hernández 6	Rafinha 12	Marc Bartra 15	Long PC	Long PA	Medium PC	Medium PA	Short PC	Short PA	Total PC	Total PA	Total %
ter Stegen	1	95'09"		3	-	2	1	1	1	-	5	11	2	-	-	-	11	17	12	12	3	3	26	32	81%
Gerard Piqué	3	95'09"	5		5	3	2	2	1	-	2	2	4	-	-	-	4	7	17	18	5	5	26	30	87%
Ivan Rakitić	4	81'58"	1	3		11	-	4	8	3	1	2	11	-	-	-	2	3	24	29	18	20	44	52	85%
Sergio Busquets	5	95'09"	1	4	7		9	1	2	2	4	8	7	-	-	-	6	6	24	26	15	19	45	51	88%
Andrés Iniesta	8	86'35"	-	1	4	6		2	4	7	4	10	3	-	-	-	2	2	21	24	18	20	41	46	89%
Luis Suárez	9	95'09"	-	-	3	1	-		6	2	-	1	-	-	1	-	0	1	9	13	5	9	14	23	61%
Lionel Messi	10	95'09"	-	-	9	3	3	6		6	-	2	2	-	-	-	4	4	16	17	11	16	31	37	84%
Neymar	11	95'09"	-	-	2	1	7	4	4		1	7	1	-	-	-	0	1	13	14	14	18	27	33	82%
Javier Mascherano	14	88'26"	7	4	2	4	5	1	1	-		4	3	-	-	-	6	9	21	22	4	5	31	36	86%
Jordi Alba	18	95'09"	4	-	1	3	8	5	4	16	6		-	-	-	1	3	4	29	34	16	20	48	58	83%
Daniel Alves	22	95'09"	-	4	14	2	3	1	12	1	2	1		-	-	-	2	2	22	26	16	20	40	48	83%
Xavi Hernández	6	13'11"	-	-	-	-	-	-	-	-	-	-	-		-	-	0	0	0	0	0	0	0	0	0%
Rafinha	12	8'34"	-	1	-	-	-	-	-	-	-	-	1	-		-	0	0	2	2	0	0	2	2	100%
Marc Bartra	15	6'43"	-	-	-	-	-	-	-	-	-	-	-	-	-		0	0	0	0	0	0	0	0	0%
Total passes received:			18	20	47	36	38	27	43	37	25	48	34	0	1	1	40	56	210	237	125	155	375	448	84%

From		TP	To: Manuel Neuer 1	Xabi Alonso 3	Medhi Benatia 5	Thiago Alcântara 6	Robert Lewandowski 9	Rafinha 13	Jérôme Boateng 17	Juan Bernat 18	Philipp Lahm 21	Thomas Müller 25	Bastian Schweinsteiger 31	Mario Götze 19	Long PC	Long PA	Medium PC	Medium PA	Short PC	Short PA	Total PC	Total PA	Total %
Manuel Neuer	1	95'09"		4	5	1	1	7	4	4	2	1	2	-	7	12	22	22	2	2	31	36	86%
Xabi Alonso	3	95'09"	3		5	17	1	10	7	5	11	3	13	1	6	7	48	54	22	24	76	85	89%
Medhi Benatia	5	95'09"	9	8		3	2	8	7	-	6	3	1	1	5	6	30	33	13	15	48	54	89%
Thiago Alcântara	6	95'09"	-	8	7		4	5	6	10	4	7	8	3	4	5	41	43	17	18	62	66	94%
Robert Lewandowski	9	95'09"	-	2	-	1		1	-	4	3	1	2	-	0	0	6	9	8	11	14	20	70%
Rafinha	13	95'09"	7	15	8	5	-		3	2	4	3	2	1	2	4	36	38	12	14	50	56	89%
Jérôme Boateng	17	95'09"	7	3	5	1	3	2		5	4	-	5	-	8	12	25	28	2	2	35	42	83%
Juan Bernat	18	95'09"	-	6	-	14	2	1	4		1	-	1	3	0	0	17	18	15	20	32	38	84%
Philipp Lahm	21	95'09"	2	13	6	7	1	5	1	3		8	8	1	1	1	33	36	21	24	55	61	90%
Thomas Müller	25	78'11"	-	1	1	-	8	2	-	1	6		5	-	1	1	13	13	10	16	24	30	80%
Schweinsteiger	31	95'09"	-	6	2	9	5	3	2	6	10	4		3	3	3	27	29	20	24	50	56	89%
Mario Götze	19	16'58"	-	2	1	2	-	1	1	3	-	-	1		0	0	7	7	4	5	11	12	92%
Total passes received:			28	66	40	60	27	45	35	43	51	30	46	13	37	51	305	330	146	175	488	556	88%

TP: Time played PA: Passes attempted PC: Passes completed %: Passing success percentage

00:22:28CET
07 May 2015

UEFA Media Information

Figure 1: Example of a pass distribution file from the official *UEFA* website.

2.1 Extracting the Data

Since the files are .pdf files, the data isn't editable, in that case we implemented a JAVA program that would download the file and right after the download would convert the .pdf file into a .txt file. This program would automatically invoke, to all the pass distribution files available in the website, the `wget` command following from the `pdftotext -raw` command (the `raw` option guarantees the integrity of the data) from the Linux Terminal. Also we saved the games played in each phase of the tournament in a file called "gamesXY.txt" where X is the year of the season and Y is the phase of the tournament.

2.2 Collecting the Data into Data Structures

Like in the section 2.1, we coded everything in JAVA. It is an object-oriented high level programming language with a very good API, a lot of documentation and community support, a very satisfying performance and exception handling that can provide flexibility when it comes to be creative. Creating a XML or SQL database wouldn't be necessary because we are not looking to be doing queries and JAVA has the particularity of freeing the allocated memory when the program terminates. The objective of this project is to study the pass distribution networks and all the queries can be easily done in JAVA if necessary.

The first impressions of the converted .txt files was that the data was all spread out and not organized at all. Due to this fact we decided to implement a JAVA program to parse the data in order to organize and, most importantly, to make it reliable and consistent. After that, the primary focus was to create data structures that could represent all the data found in one file.

Given some consideration we decided that we wanted each team to have a graph for each different season. In order to do that we implemented the following data structures:

1. **Match**

   This class will save the basic data of a match, since we are focusing the performance of a team in each season the only things we need to know is the date of the match, the teams, their respective players and passes and the score.

   (a) **MatchScore**

   This class just saves the amount of goals that each team scored.

   (b) **MatchDate**

   This class composes the date into: weekday, day month year.

2. **Player**

   This class saves the identifier, the team, the shirt number, the name, the total time played so far, the season and the passes that the player made throughout the season.

3. **Team**

   Similarly to Player, this class will save an identifier, a name, a season, a list of matches and players that played in the season.

4. **Season**

   This class saves the year of the season, p.e., if the season is 2012/2013 the year is 2013, the teams and players that played in that year.

5. **Pass**

   This class saves the source player, the target player and the number of passes made between these two players in this order.

2.3 Converting the Data into Networks

Now that the data is collected we need a tool to reproduce the data into networks. In section 2.3.1 we'll introduce you to Gephi: An Open Source Software for Exploring and Manipulating Networks[1].

2.3.1 Gephi

Gephi is a powerful tool to visualize and analyse networks. It has a friendly and easy-to-use framework and the task of importing data is as easy as it can be. Anyone can build networks from spreadsheets or another file extensions, such as GraphML, GML, CSV and so on [2]. The Figure 2 is an example of the capabilities of Gephi with respect to network visualization and analysis.

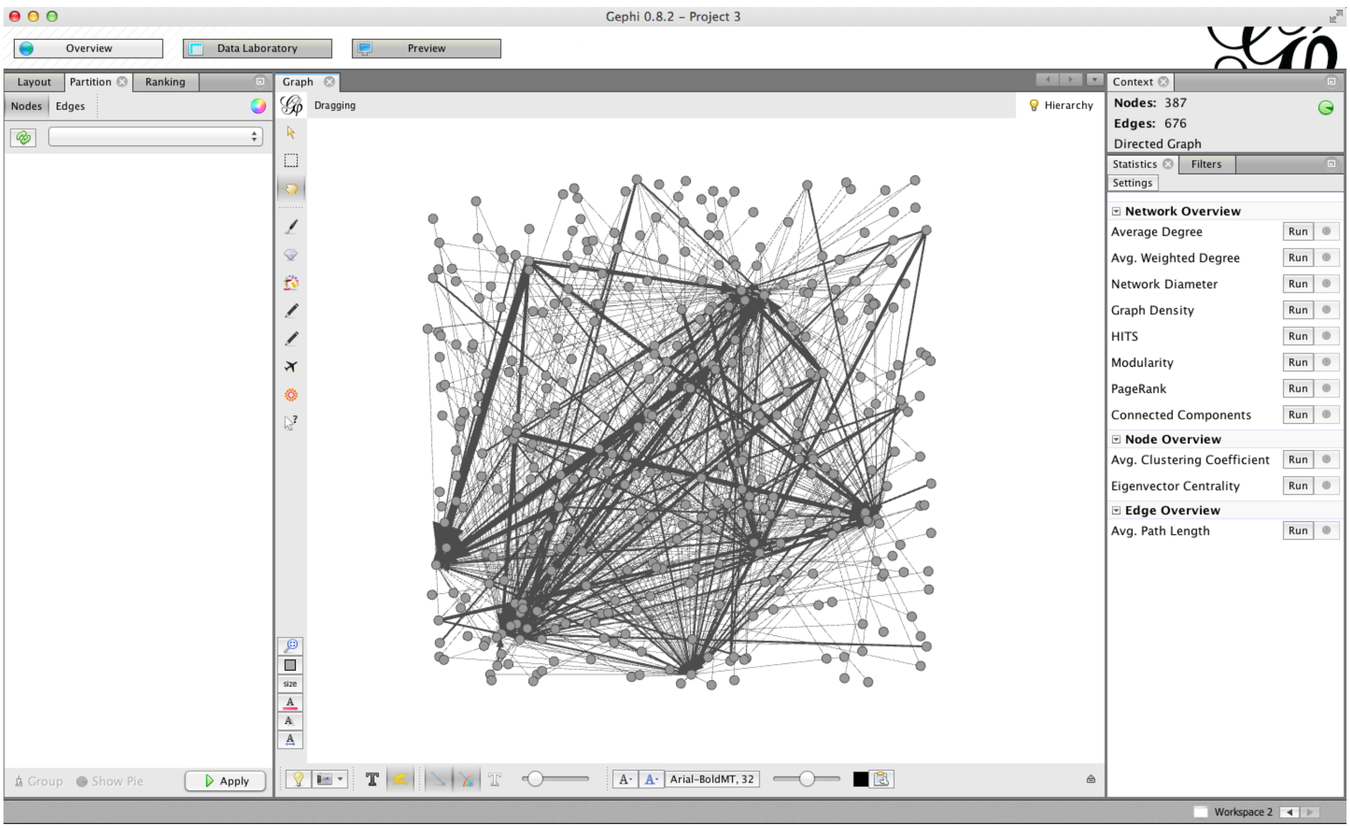

Figure 2: Example of a network built in Gephi.

2.3.2 Parsing the Data

Since we're using Gephi to analyse our pass distribution networks, we needed to rearrange the data so it would be easier to import it to Gephi in the future. In this way, we would not waste time importing data to the platform. We've decided to parse the data in two formats:

1. Spreadsheet Notation

   The Spreadsheet Notation is very easy to read and understand the data. For this project, any spreadsheet presented like Tables 1 and 2 can be easily imported into Gephi, we need only to specify which table represents the nodes and which table represents the edges and after that the network is built nice and easy.

Id	TeamNumber	Name
1682	20	Sergio Romero
908	7	Memphis Depay
464	8	Juan Mata
...	...	...

Table 1: Example of nodes.

Source	Target	Weight
1682	628	1
1682	1262	1
1682	745	2
...	...	...

Table 2: Example of edges.

2. GML Notation

   GML, Graph Modelling Language, consists of a hierarchical key-value lists and it features portability, simple syntax, extensibility and flexibility [3]. As we can see in Figure 3, firstly we must declare the nodes of the network before the edges. By doing this, we guarantee that the hierarchical key-value lists condition is preserved, avoiding information to be lost in the process of construction of the network.

```
graph [
	directed 1
	comment "Sample comment"
	node [
		id "A"
	]
	node [
		id "B"
	]
	node [
		id "C"
	]
	edge [
		source "A"
		target "B"
		label "Edge from node A to node B"
		weight "0.1"
	]
	edge [
		source "B"
		target "C"
		label "Edge from node B to node C"
		weight "0.5"
	]
	edge [
		source "C"
		target "A"
		label "Edge from node C to node A"
		weight "0.5"
	]
]
```

Figure 3: Example of a GML network.

3 Results

3.1 Dataset

In this section we're going to look over our dataset that was collected according with section 2.2. Table 3 represents all the dataset, per season, with the number of:

- Matches
- Teams
- Players
- Goals
- Passes

Season	Matches	Teams	Players	Goals	Passes
2012/2013	131	41	842	372	98120
2013/2014	130	41	843	379	111752
2014/2015	129	40	819	358	103129
2015/2016	123	39	786	340	100658
Total	**513**	**161**	**3280**	**1449**	**413659**

Table 3: Dataset.

In the perfect scenario, the number of matches and teams should be the same throughout the seasons. In this case, this does not occur due to corrupted files that we didn't want to include in the dataset in order to guarantee reliability and truthfulness.

3.2 General Statistics

In this section, we're going to analyse the dataset obtained in section 3.1. We're going to build hypothesis and verify with the dataset if they are true or false.

3.2.1 Hypothesis

1. *Teams that have the highest average Passes per Match ratio have a bigger percentage of wins in the tournament.*

2. *Teams that have the highest average Passes per Win ratio have a bigger percentage of wins and are likely to win the tournament.*

3. *Teams that have the lowest Passes per Goal ratio score more and, as a result, have a bigger percentage of wins.*

3.2.2 Proofs

Starting with Hypothesis 1, we can think that all the teams with a big average of passes per match would have a possessing play style. We can conclude that, given the fact that a team possesses the ball more time, it's unlikely to conceding goals or loosing matches. Let's take a look at Table 4:

Season	Team	Wins(%)	Average Passes/Match	Passes/Goal
2012/2013	FC Barcelona	41.7	649.4	222.7
	FC Kobenhavn	0.0	593.0	296.5
	FC Basel 1893	0.0	558.0	558.0
2013/2014	FC Barcelona	60.0	691.2	238.3
	FC Bayern Mnchen	58.3	662.4	214.8
	Paris Saint-Germain	62.5	599.5	177.6
2014/2015	FC Bayern Mnchen	66.7	600.9	156.8
	FC Barcelona	83.3	591.5	186.8
	Aalborg BK	0.0	510.5	170.2
2015/2016	FC Bayern Mnchen	66.7	674.9	197.5
	FC Barcelona	75.0	668.1	213.8
	FC Basel 1893	0.0	666.0	166.5

Table 4: The three teams with highest average passes made per match in each season, sorted in descendent order.

As we can see, the winning percentage is not that high in some teams that have the best average passes per match which means that these teams didn't play in a possessive attacking style rather in a controlling possessive style. We conclude that having the highest average of passes per match doesn't implies bigger percentage of wins, so this hypothesis is false.

In Hypothesis 2, the highest average of passes per win is an important component to analyse because we can conclude if the amount of passes is actually relevant to win a match or if it's a winning condition.

Season	Team	Wins(%)	Losses(%)	Average Passes/Win	Passes/Goal
2012/2013	FC Porto	62.5	25.0	468.2	201.7
	FC Bayern Mnchen	76.9	15.4	430.8	136.1
	Paris Saint-Germain	60.0	10.0	415.8	131.8
2013/2014	FC Barcelona	60.0	20.0	712.8	238.3
	Paris Saint-Germain	62.5	25.0	651.6	177.6
	Arsenal FC	60.0	30.0	544.5	215.1
2014/2015	FC Bayern Mnchen	66.7	25.0	621.9	156.8
	FC Barcelona	83.3	16.7	605.3	186.8
	Real Madrid CF	66.7	16.7	534.3	185.4
2015/2016	FC Bayern Mnchen	66.7	16.7	683.8	197.5
	FC Barcelona	75.0	12.5	678.8	213.8
	Paris Saint-Germain	60.0	20.0	663.7	273.0

Table 5: The three teams with highest percentage of wins per season, sorted in descendent order of average of passes per win.

In Table 5, we can observe that this component is actually relevant due to the percentage of wins although we have already proven, in Hypothesis 1, that the teams with highest average of passes per match don't have the bigger percentage of wins in the competition and, also, not always the teams with highest average passes per win are the winner of the competition, as we can see in Table 7.

We agree that the amount of passes by itself is not enough to prove that this hypothesis is true or false but we can say that teams that have a winning percentage bigger than 60% play a possessive style.

In Season	Average Passes/Match
2012/2013	749.0
2013/2014	804.1
2014/2015	802.6
2015/2016	806.4

Table 6: Average of passes per match.

Season	Team
2012/2013	FC Bayern de Munique
2013/2014	Real Madrid CF
2014/2015	FC Barcelona
2015/2016	Real Madrid CF

Table 7: Winners of the tournament.

Analysing Hypothesis 3, we can say that for it to be true the teams must have possession of the ball most of the time. Observing Table 8, we see that that does not happen. The winning percentage of these teams are at most 50% which indicates that the play style of these teams is not possessive but rather counter-attack or defensive.

We conclude that this hypothesis is false but assuming it was true this would imply a huge amount of goals, according to Table 6, which give us the average amount of passes in a match since 2012/2013. We also can see the rise of passes per match between 2012/2013 and 2013/2014 showing that teams were trying to play in a possessing style.

Season	Team	Wins(%)	Losses(%)	Draws(%)	Passes/Goal
2012/2013	CFR 1907 Cluj	50.0	33.3	16.7	77.7
	Chelsea FC	50.0	33.3	16.7	88.0
	FC Nordsjaelland	0.0	83.3	16.7	89.3
2013/2014	FC Bayer 04 Leverkusen	0.0	100.0	0.0	49.5
	FC Paos de Ferreira	0.0	100.0	0.0	78.9
	PFC Ludogorets 1945	0.0	100.0	0.0	80.8
2014/2015	FC BATE Borisov	25.0	62.5	12.5	63.3
	Malmo FF	25.0	75.0	0.0	74.8
	RSC Anderlecht	0.0	50.0	50.0	80.2
2015/2016	Maccabi Tel-Aviv FC	0.0	0.0	100.0	44.5
	Club Brugge KV	0.0	100.0	0.0	52.6
	Malmo FF	25.0	75.0	0.0	66.4

Table 8: The three teams with the lowest passes per goal ratio per season, sorted in ascendant order.

After analysing these results, we still have to question if the pass amount is a winning condition just by itself. Let's take a look at Figure 4, which represents the last three winners of the tournament associated with the respective average of passes per match.

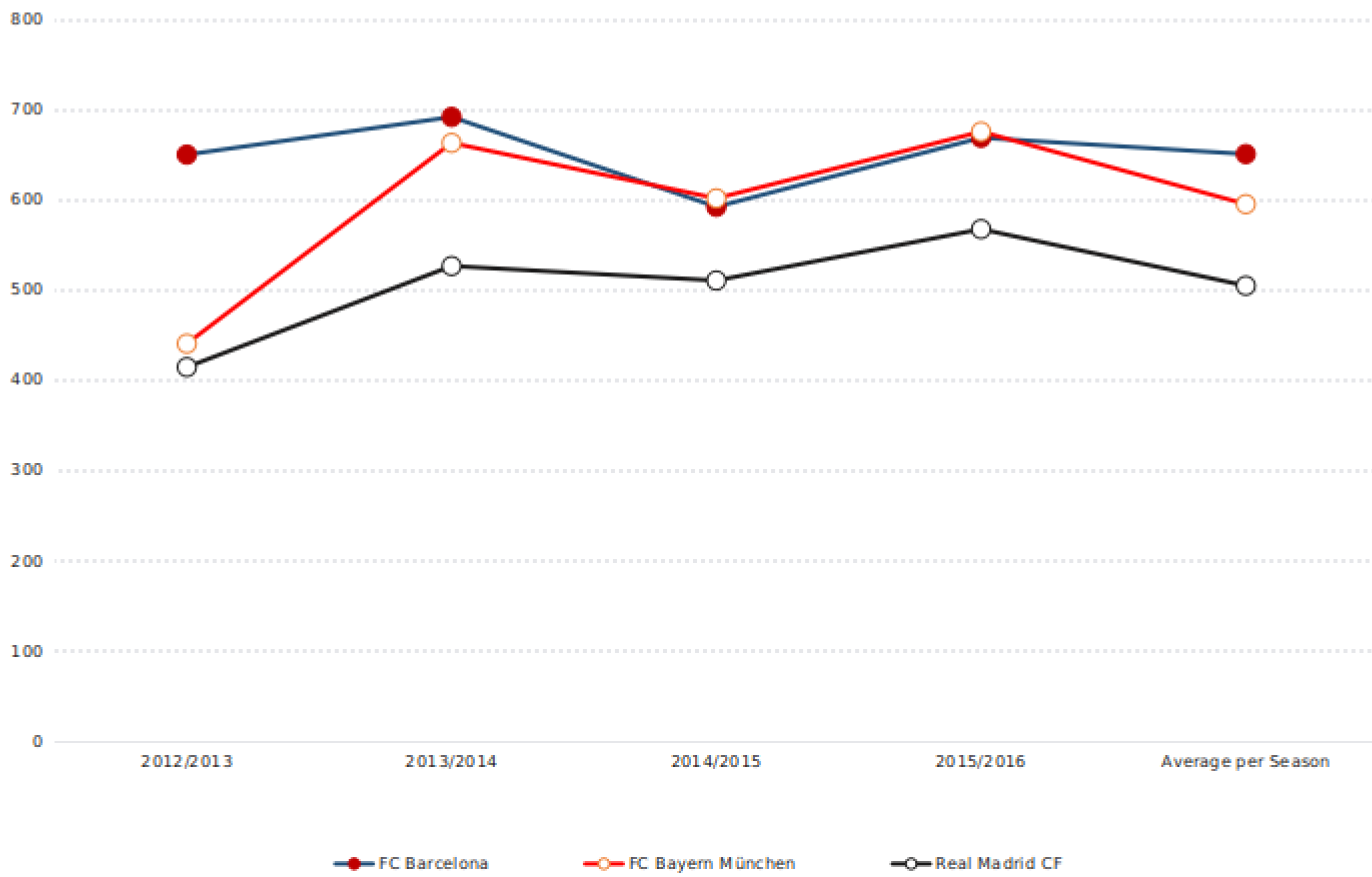

Figure 4: Average of passes made by the last four champions of the tournament and its average per season.

We can clearly see that Real Madrid CF, who won the tournament in 2013/2014 and 2015/2016, doesn't have the highest average of passes per match and FC Barcelona, who won the tournament in 2014/2015, had his lowest average passes per match so we can say the same that when analysing Hypothesis 2: The passes by itself doesn't guarantee winning the match but it can certainly lead to it.

3.3 A detailed analysis on a single Team

In this section we'll be taking a closer look into a single team: Arsenal FC. This team was chosen because not only did it played the four seasons we're analysing but the coach was always the same, which means that the formation used by the team was, with rare exceptions, the same: 4-2-3-1.

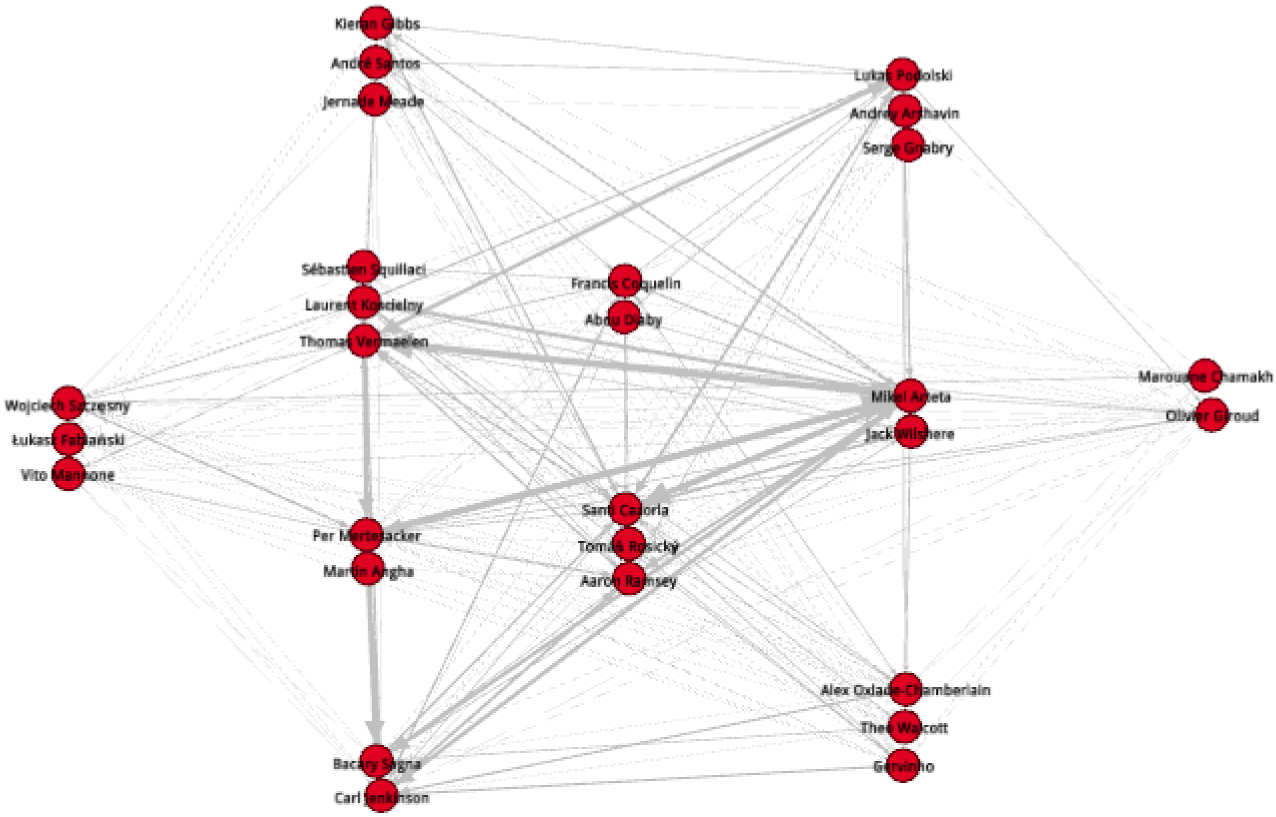

Figure 5: Passing Distribution Network of Arsenal FC in 2012/2013 on a 4-2-3-1 formation.

Speaking from personal experience on watching this team playing, I could say that Arsenal FC is a team that likes to construct an attack slowly and to possess the ball to look for a glitch in the opponent's defence to make the last pass to score a goal. Meanwhile, the team circulates the ball between the defenders and mid-fielders.

If we look to Figure 5 and analyse the network that's exactly what's happening (the thicker is the edge the more passes were made from one player to another): Per Mertesacker receives more passes from Mikel Arteta and sends more passes to Bacary Sagna, which sends more passes to Mikel Arteta, who sends the ball either to Santi Cazorla or Thomas Vermaelen and so on. In Table 9 we can verify that.

Season	**Goalkeepers**	**Defenders**	**Midfielders**	**Strikers**
2012/2013	107	1329	1694	138
2013/2014	140	1706	2533	353
2014/2015	77	1294	1943	189
2015/2016	63	1008	1768	142
Average	96.75	1334.25	1984.5	205.5

Table 9: Amount of passes made to the different positions of the field.

Using Gephi, the task of analysing which player is the most important or which one connects the most players of the network or the one who received and/or send more passes is an easy task. Things like *Degree, Closeness and Betweenness Centrality*[4] can be calculated with a simple click. Table 10 shows us that procedure.

Name	In-Degree	Out-Degree	Degree	Closeness	Betweenness
Aaron Ramsey	27	26	53	1.0	77.6
Thomas Vermaelen	23	23	46	0.9	30.3
Mikel Arteta	22	22	44	0.8	23.4
Carl Jenkinson	22	22	44	0.8	42.2
Per Mertesacker	20	21	41	0.8	16.4
Francis Coquelin	23	20	43	0.8	38.3
Santi Cazorla	22	19	41	0.8	19.2
Olivier Giroud	19	19	38	0.8	8.0
Laurent Koscielny	19	19	38	0.8	8.6
Vito Mannone	11	18	29	0.8	2.9
Wojciech Szczesny	11	18	29	0.8	8.5
Lukas Podolski	19	17	36	0.7	6.7
Gervinho	19	17	36	0.7	17.7
Alex Oxlade-Chamberlain	19	17	36	0.7	12.5
Kieran Gibbs	16	16	32	0.7	5.2
Theo Walcott	15	14	29	0.7	2.0
Bacary Sagna	14	14	28	0.7	2.2
Jack Wilshere	14	14	28	0.7	1.6
Tomas Rosicky	19	13	32	0.7	6.9
Marouane Chamakh	10	10	20	0.6	1.9
Andre Santos	9	10	19	0.6	0.3
Sebastien Squillaci	8	10	18	0.6	1.0
Abou Diaby	11	9	20	0.6	0.1
Lukasz Fabianski	6	9	15	0.6	0.0
Jernade Meade	7	8	15	0.6	1.2
Andrey Arshavin	8	7	15	0.6	1.2
Serge Gnabry	4	5	9	0.6	0.0
Martin Angha	3	3	6	0.5	0.0

Table 10: Results calculated with Gephi, sorted by Closeness Centrality in descendant order.

In this particular case, the most important player (Closeness Centrality), the player that received and/or sent more passes (Degree Centrality) and the player who connects the majority of the network (Betweenness Centrality) are all the same: **Aaron Ramsey**. But was he always the best?

Season	Degree	Closeness Centrality	Betweeness Centrality
2013/2014	Bacary Sagna	Bacary Sagna	Bacary Sagna
2014/2015	Alexis Sanchez/Santi Cazorla	Santi Cazorla/Per Mertesacker	Alexis Sanchez
2015/2016	Mesut Ozil	Mesut Ozil/Laurent Koscielny	Mesut Ozil

Table 11: Players that had the highest results per season.

From Table 11 we can conclude that, although the formation and the coach were the same, players tend to play different roles, new signings may change the play style of a team or opponent teams may try to counter the performance of a player in the match. Also, we can say that passing performances can lead to a transfer, for instance, Bacary Sagna was in good form in 2013/2014 and he got transferred to Manchester City FC in the following season.

4 Conclusions

With this project, we were able to conclude that teams that play in a possessive style are more likely to be victorious in comparison to teams that play either defensively or in counter-attack, the amount of passes just by itself cannot guarantee a victory in a match, which means there are other relevant factors, like the amount of shots on target, that can help guaranteeing a win in a match and the passing performance of a player can be a decisive factor for a player to transfer to another team.

Although the *UEFA Champions League* is the best European competition, I'd like, in the future, to investigate the *UEFA Euro 2016* competition and it'd be interesting to study not only the pass distribution but, for example, the shots on and off target. It would be also interesting to study how these competitions influence the transfer market.

Developing this project made me learn a lot of new concepts and new ways to work in a field of study that is of my interest and that made it a very pleasant experience.

References

[1] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. Gephi: an open source software for exploring and manipulating networks. 2009.

[2] K. Cherven. *Network Graph Analysis and Visualization with Gephi*. Community experience distilled. Packt Publishing, 2013.

[3] Michael Himsolt. Gml: A portable graph file format. *Html page under http://www. fmi. uni-passau. de/graphlet/gml/gml-tr. html, Universität Passau*, 1997.

[4] Loet Leydesdorff. Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9):1303–1319, 2007.

[5] M. E. J. Newman. The structure and function of complex networks. *SIAM REVIEW*, 45:167–256, 2003.