Understanding of Soccer Player Transfers in 11 European Leagues

Gaoyuan Huang, Letian Xu, Xiaoxiao Deng

Introduction

For a soccer club, it is very important to consider the transfers of players during summers and winters. Player transfers will produce a great impact on both the loyalty of players in a team and the development of leagues and clubs.

The purpose of this analysis is to discover if there is any transfer preference among these clubs and leagues, which means a league or soccer club might be more willing to do business with some specific clubs or leagues. Social network analysis methods can offer some approaches, so this project is going to use them to examine the networks and to understand the hidden pattern behind it.

Data Preprocessing

Build all transfers dataset

The data we use is from European Soccer Database in Kaggle (https://www.kaggle.com/hugomathien/soccer).

In this part, we extracted useful player transfer information from the database. We refer to https://www.kaggle.com/nappon/the-fans-stay-loyal-the-players-move-on/notebook to build the TransferDf. It contains information about each player and each club he has ever been to during 2008-2016. The ClubFirst variable tells us when the player came to the current club. The ClubLast variable tells us when the player left the current club. Here is a screenshot of the beginning of the data:

player_name	Player	CurrentClub	FormerClub	ClubFirst	ClubLast	CurrentLeague	FormerLeague	CurrentCountry	FormerCountry
Aaron Galindo	Aaron Galindo	Eintracht Frankfurt	NA	2008	2009	Germany 1. Bundesliga	NA	Germany	NA
Aaron Hughes	Aaron Hughes	Fulham	NA	2008	2014	England Premier League	NA	England	NA
Aaron Hunt	Aaron Hunt	SV Werder Bremen	NA	2008	2014	Germany 1. Bundesliga	NA	Germany	NA
Aaron Lennon	Aaron Lennon	Tottenham Hotspur	NA	2008	2015	England Premier League	NA	England	NA
Aaron Meijers	Aaron Meijers	FC Volendam	NA	2008	2009	Netherlands Eredivisie	NA	Netherlands	NA
Aaron Mokoena	Aaron Mokoena	Blackburn Rovers	NA	2008	2009	England Premier League	NA	England	NA

For example, this is Zlatan Ibrahimović's transfer data. He was actually transferred from Juventus to Inter in 2006, but 2006 is not in our data, so the former club will be NA and we will just drop this row since it doesn't count as a transfer between year 2008-2016. For the rest of the transfer information, we can extract three edges from them. We will merge edges between same clubs later.

	player_name	Player	CurrentClub	FormerClub	ClubFirst	ClubLast	CurrentLeagu	FormerLeagu	CurrentCoun	FormerCountr
3837	Zlatan Ibrahimovic	Zlatan Ibrahimovic	Inter	NA	2008	2009	Italy Serie A	NA	Italy	NA
5813	Zlatan Ibrahimovic	Zlatan Ibrahimovic	FC Barcelona	Inter	2009	2010	Spain LIGA BI	Italy Serie A	Spain	Italy
7775	Zlatan Ibrahimovic	Zlatan Ibrahimovic	Milan	FC Barcelona	2010	2012	Italy Serie A	Spain LIGA BI	Italy	Spain
11960	Zlatan Ibrahimovic	Zlatan Ibrahimovic	Paris Saint-G	Milan	2012	2016	France Ligue	Italy Serie A	France	Italy

Build node list and edge list for transfers between clubs

We extracted all clubs as our node list from TransferDf data (Saved as team_node.csv, mode details in node_preprocess.R). It contains label as club's name, League as its league and a unique id we gave each club. The screenshot of the beginning of the data:

	label	League	id
1	Widzew ÅÃ	Poland Ekstra	1
2	KAA Gent	Belgium Jupil	2
3	MÃ¡laga CF	Spain LIGA BI	3
4	Sevilla FC	Spain LIGA BI	4
5	Hamburger S	Germany 1. E	5
6	Genoa	Italy Serie A	6

As the example shows, based on this TransferDf data, we were able to extract each transfer to create edge list. We set the club a player transfer from as V1 and the club he transfer to as V2. And we record the transfer time as Year. (Saved as soccer_edge.csv, more details in team2team.R)

V1	V2	Year
188	166	2012
188	210	2014
188	13	2011
188	252	2014
188	200	2014
188	48	2013
188	172	2015
188	93	2011
188	5	2011
188	127	2015

Add geographic attribute to team node list

We are interested in how the geographic location of each club influences its transfer, so we add longitude and latitude for each club in excel. We then save it as the final node list we will use named team_node_with_lnglat.csv. The structure of the data is here:

id	label	League	lat	lng
1	Widzew _¨d_	Poland Ekstraklasa	51.7592	19.456
2	KAA Gent	Belgium Jupiler League	51.0543	3.7174
3	M_laga CF	Spain LIGA BBVA	36.7213	-4.4214

4	Sevilla FC	Spain LIGA BBVA	37.384	-5.9705
5	Hamburger SV	Germany 1. Bundesliga	53.5511	9.9937

Merge multiple transfers between two clubs

Multiple transfers between two clubs can be merged as a new edge with weight indicating the frequency of their transfers.We first thought of this network as a undirected network, but later we realized that the in-degree and out-degree have a lot of meanings in transfer so we decided to make it a directed network. We also labeled all transfers according their types. By types we mean if the transfer happens between leagues, then the type is 0, otherwise it's a unique number of their league. So there are 12 types of edges. After doing all preprocessing above, we now have a network called dir_total_transfer_aggr_etype_geo.graphml which is the main network our analysis is based on. (More details of processing in merge_same_egdes.R)

We also did all the same things above to create a league transfer network. We can observe transfers from a different level and it can give us a whole picture of the data between leagues. But since this network only includes edges between different leagues, there are not sufficient information for us to explore. (More details of processing in build league node and edges.R)

All codes for creating graphml files can be found in Create_Graph.R.

Summary of network and Basic visualization

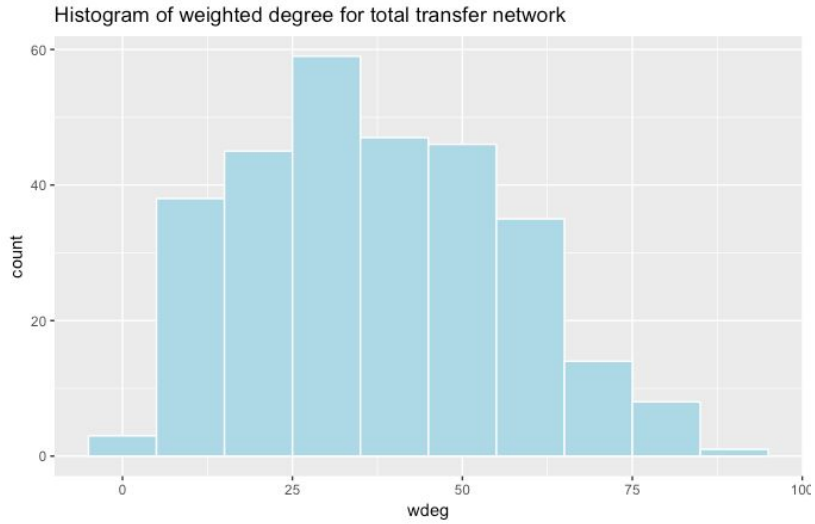
This part has been submitted as a R markdown file called basic.Rmd and there is a correspondent basic.html file for you to check.

Summary of network:

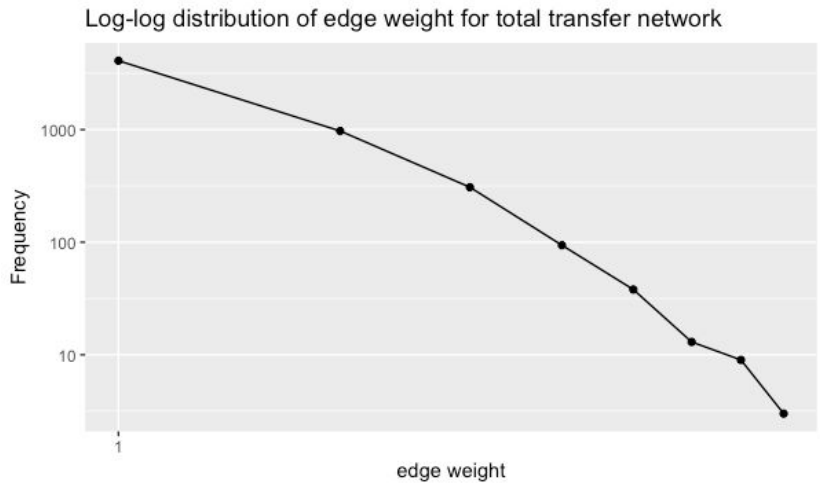
> summary(g)
IGRAPH ff39a7e DN-- 296 5534 --
+ attr: name (v/c), label (v/c), League (v/c), id (v/n), Weight (e/n), type (e/n)

variables:
name/id: id of the node(club)
label: club's name
League: the league the club belongs to
Weight: the weight of edge (transfer frequency)
type: the transfer's type

Here are some basic statistics of the network:



Most nodes have weighted degree in the middle, it's similar to normal distribution.



Before merging transfers, there are 7697 transfers and after merging there are 5534 edges. Most of the weight of edges are 1. The data is very skewed, so we use log-log plot to show the distribution.

For above plots please refer to basic_stat.R.

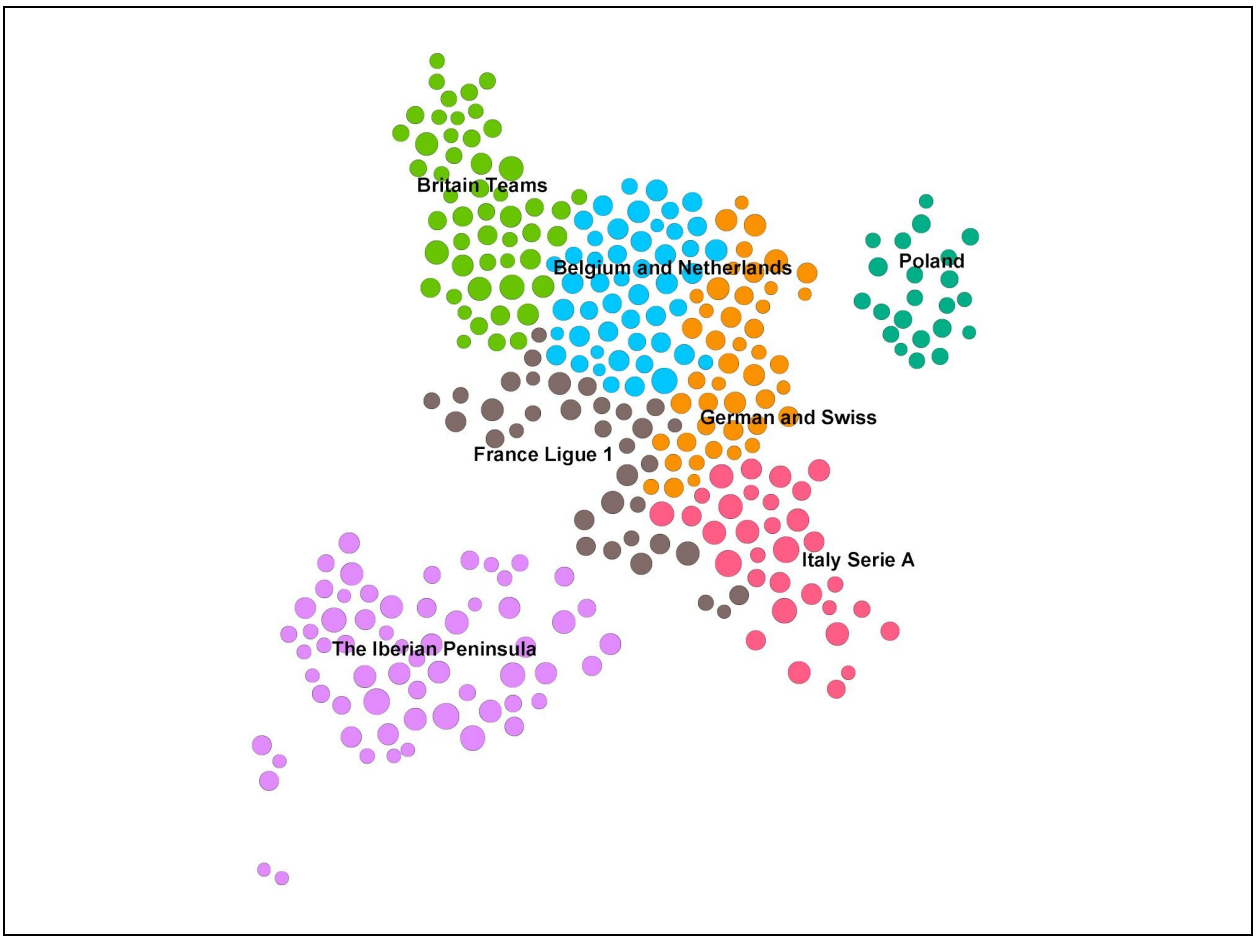
Analysis

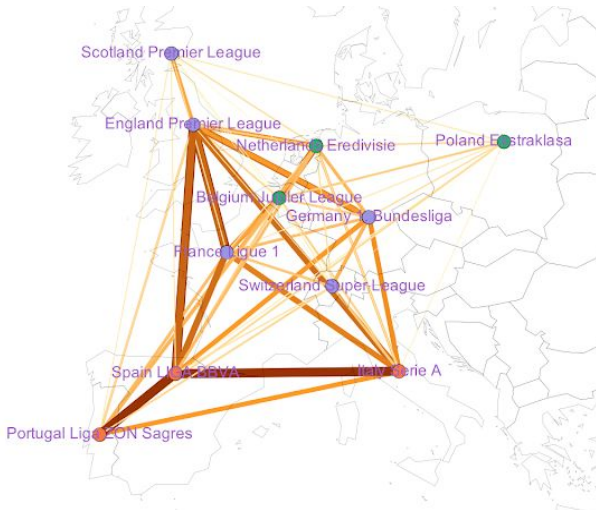
Our analysis focuses on identifying and comparing communities of clubs who frequently conduct transactions. Since some clubs prefer to trade with clubs in the same league, while some clubs prefer other leagues, the pattern of transaction might create some communities more than just leagues.

Using Gephi's modularity algorithm, 7 communities were identified in the network, whose size range from 22 to 62 clubs. Using Walktrap 10, we will stay with 11 communities and all clubs are still in the community consisting of clubs from their original leagues. Gephi's modularity communities were selected for subsequent analysis.

Comparing to the original data, new communities combine clubs from different leagues together, which indicates there exists certain transfer communities. For example, one of the giant components consists of clubs come from La Liga and Portugal Liga, which are very close to each other geographically. Also there are some communities including only one league's team, such as France Ligue 1 and Italy Serie A.

European Leagues' Transaction Structure By Modularity Community





Deeper and thicker edge color means more transfers.

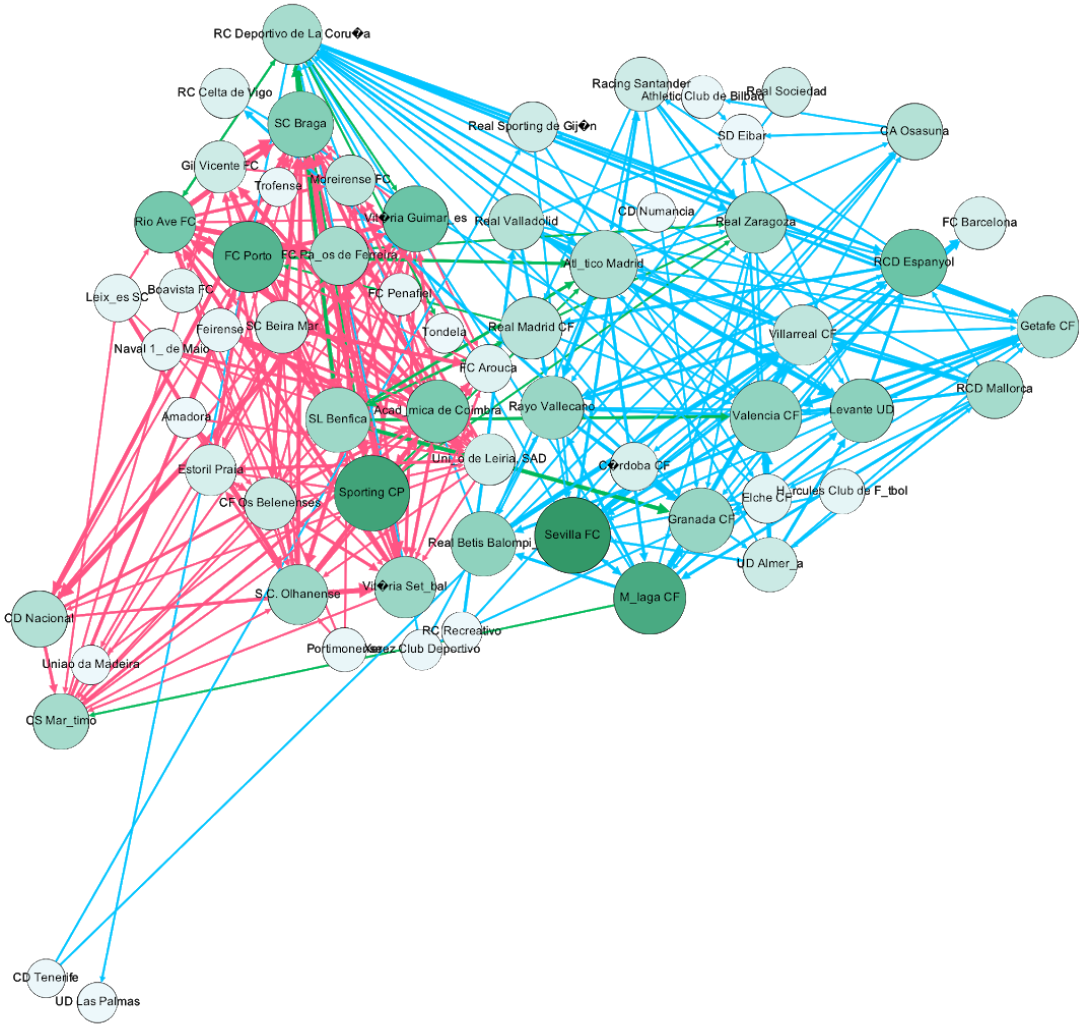
Before we go deeper into communities, we compare the communities we computed for league transfer network.

From this graph, we can see Spain and Portugal stay in one community but they have a new member, Italy Serie A. Same thing happened in the community of Scotland Premier League and England Premier League, they have France League 1, Switzerland Super League and Germany Bundesliga joining their community. The rest leagues formed another community. It's kind of surprise that Italy Serie A belongs to the same community with Spain and Portugal. Also French speaking countries like Belgium is not in the same community with France.

Comparing Communities

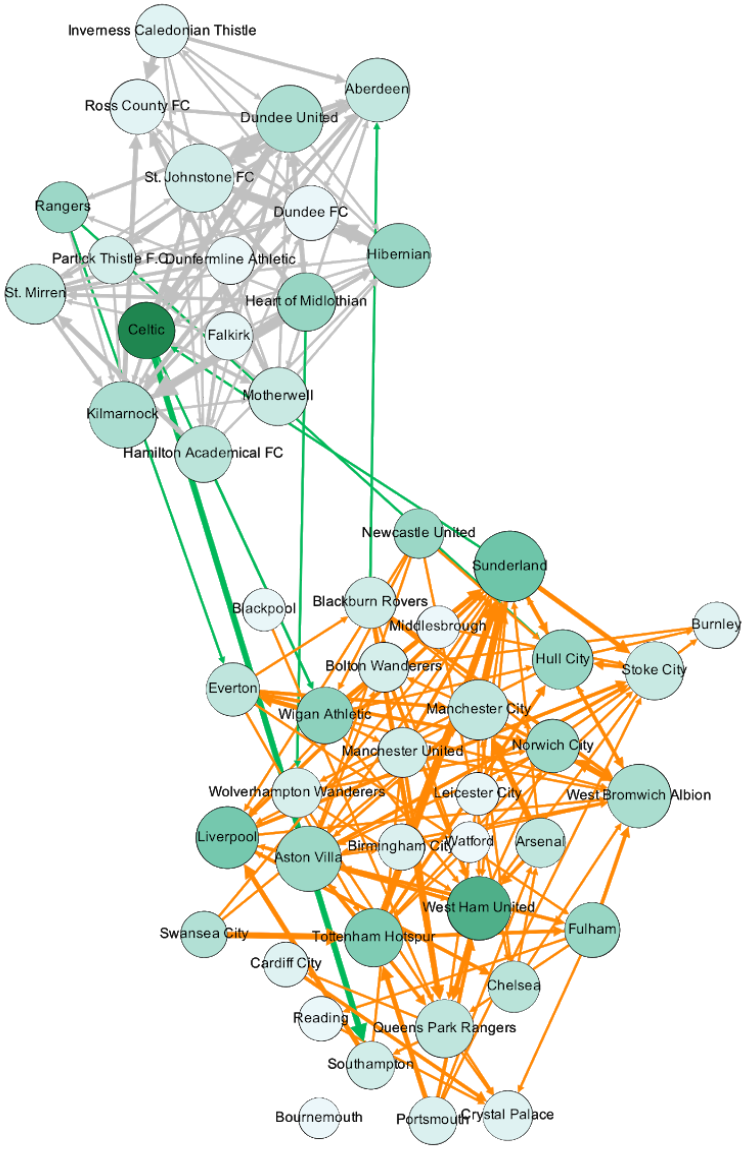
We used the two biggest communities for the following analysis:
The Iberian Peninsula, which combines Spain LIGA BBVA, the best soccer league in the world, and Portugal Liga ZON Sagres. Blue edge represents transfers between Spain LIGA BBVA, pink represents transfers between Portugal Liga ZON Sagres, and green represents transfers between them.

The Iberian Peninsula

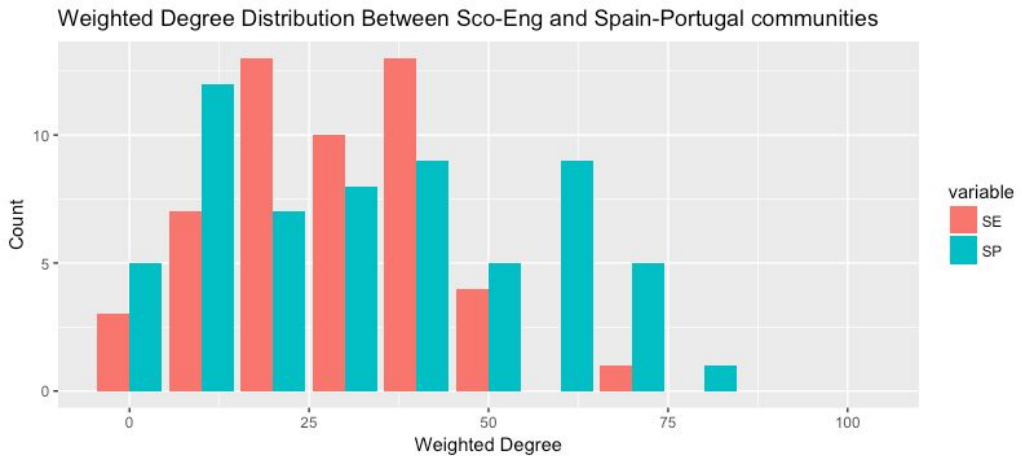


Great Britain, which combines England Premier League, the most successful league in business, and Scotland Premier League. For edge color, grey means transfers between Scotland, orange means transfers between England and green means transfers between them.

Great Britain



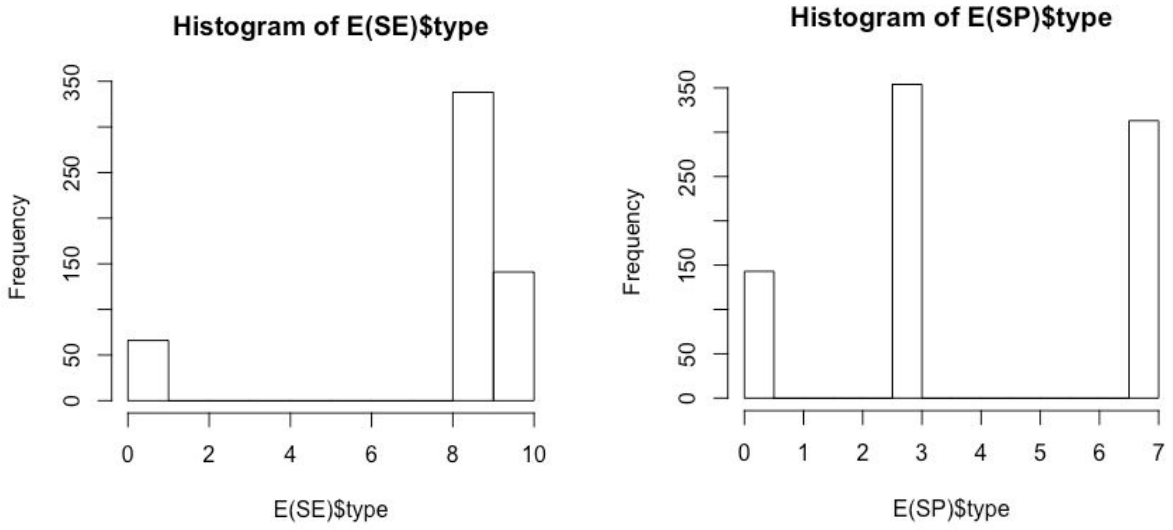
In both visualizations, node size is scaled by node degree and node color indicates their betweenness (darker green indicates higher betweenness centrality). We remove edges whose weights are less than 2, which means only 1 transaction happened between two clubs.



SE: Great Britain. SP:The Iberian Peninsula.

By observing this histogram, we can find some difference. The Iberian Peninsula community tends to have more higher weighted degree clubs in general, which means there are some clubs in Iberian Peninsula community not only transfers a lot but also have some frequent business partner.

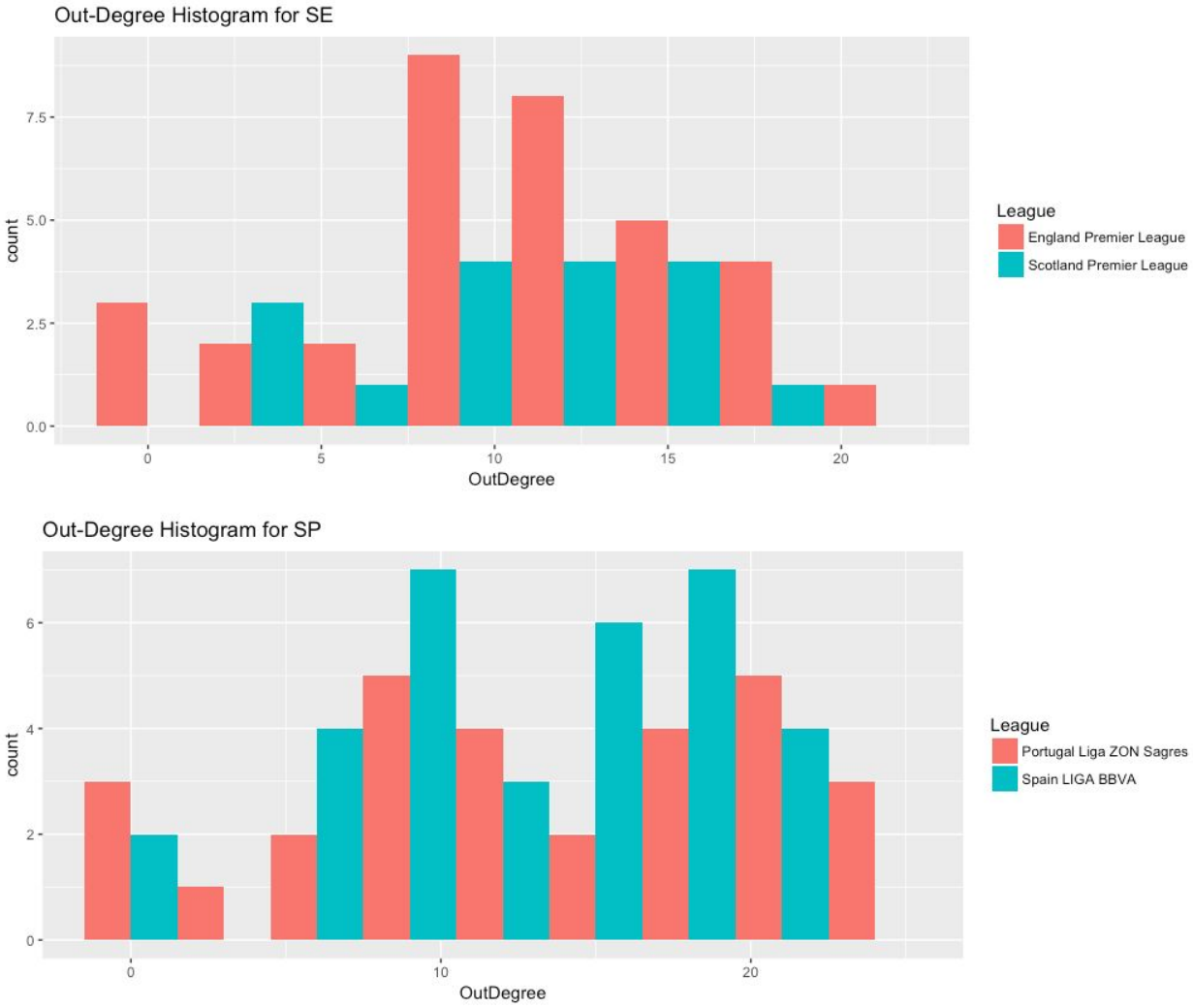
Compared with Great Britain, transactions between soccer clubs in Spain LIGA BBVA and Portugal Liga ZON Sagres are more often than Great Britain(type equal to zero meaning transactions between leagues, number 8 type means transactions happened in England Premier League, number 9 type means transactions happened in Scotland Premier League, number 3 type means transactions happened in Spain LIGA BBVA, number 7 type means happened in Portugal Liga ZON Sagres).

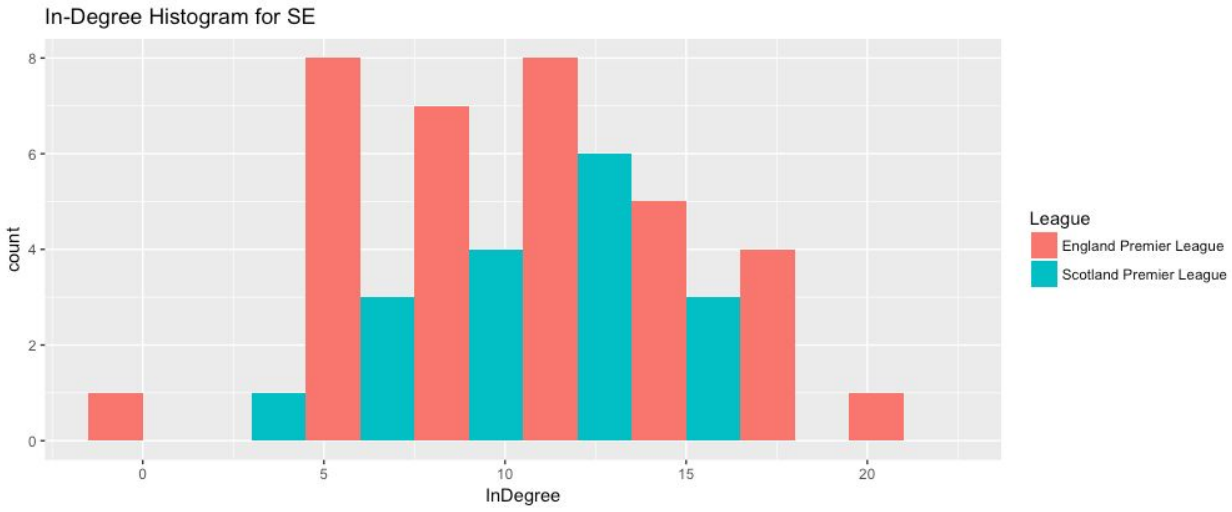


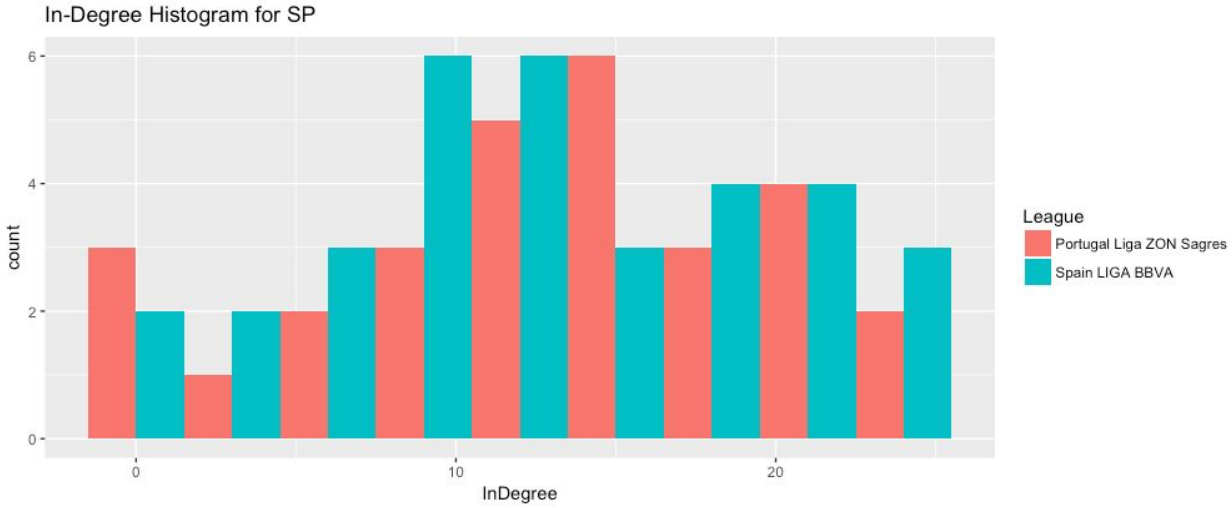
After visualizing in and out degree distributions in two communities, we find out that in the Iberian Peninsula community, clubs from Portugal Liga ZON Sagres have lower out-degree than clubs in Spain LIGA BBVA. First reason would be Spain LIGA BBVA has a successful youth training system. So it brings clubs in Spain LIGA BBVA the ability to supply their first team from

their youth team, and it works well. Now that they have the backup plan, clubs in Spain LIGA BBVA are inclined to let players out instead of using contract to force them to stay.

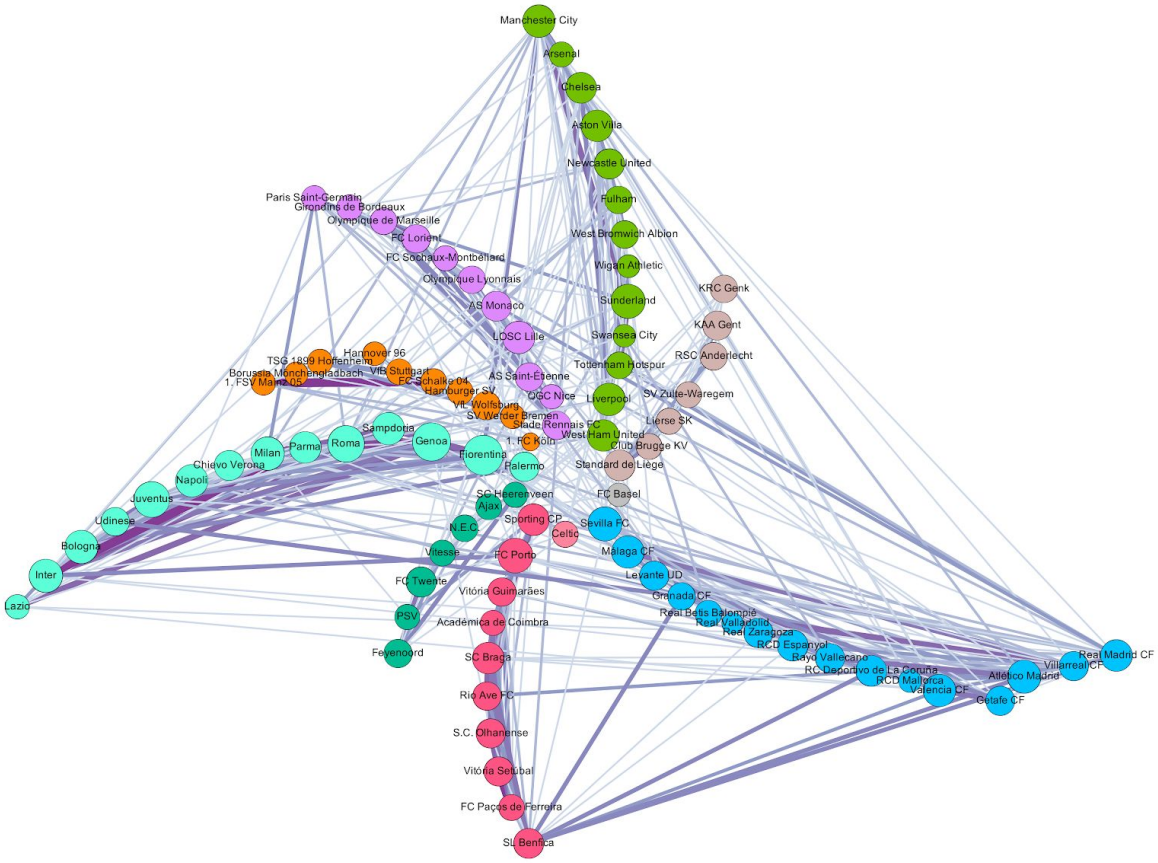
In the Great Britain, England Premier League has higher in and out degree than Scotland Premier League. That can reflect England Premier League is more competitive than Scotland Premier League, since the club needs to be maintain the competition by selling and buying players.




In-Degree Histogram for SE
count
InDegree
0
2
4
6
8
0
5
10
15
20
League
England Premier League
Scotland Premier League


In-Degree Histogram for SP
count
InDegree
0
2
4
6
0
10
20
League
Portugal Liga ZON Sagres
Spain LIGA BBVA

Stepping stone



In the real world, if a soccer player wants to join top clubs like Real Madrid, Barcelona or Manchester United, he needs to start from a bottom club like some amateur soccer clubs, and then join another club in the higher level step by step.

Some clubs are really good at finding talented players and they will buy those players at very low transaction fee when the players are still young and then transfer them to other top-level clubs to earn transaction fee after a few years. We named this kind clubs "stepping stone".
In the social network analysis, especially in this two communities, betweenness centrality will be the most appropriate measurement for us to find out the “stepping stone” node.

We choose Radial Axis layout to create our network. In this process, nodes are grouped by leagues and ordered by betweenness centrality. We also remove some nodes with total degree less than 40, which means this club had transaction with less than 40 other clubs, and remove some edges with weight less than 2, which just keep clubs who more than two transactions with each other stay in the network. And the remaining nodes will help us find “stepping stone”.

One thing interesting is that most good clubs in their leagues, which means top 3 or top 4 clubs, do not have a relative high betweenness centrality. Such as Real Madrid, Paris Saint-Germain

F.C., Chelsea and Manchester City, they do have a high degree, however it seems like, instead of seeking to trade with unfamiliar clubs, they prefer to find qualified promising players from clubs they are familiar with. This conservative transaction strategy let those top class clubs had few chance to meet up with new cooperated clubs, which lowers down their betweenness centrality.

As above visualization shows, nodes in the center are the “stepping stone”, which means the soccer players are being transferred into these clubs to gain more attention or reputation, such as West Ham United, Sporting CP (where Cristiano Ronaldo comes from), and Sevilla FC (where Sergio Ramos comes from). These clubs scouted well and developed young players before selling those players to bigger clubs in Europe. Obviously, they are our important nodes in this case.

CUG/QAP Tests

Conditional Uniform Graph(CUG) Test is used to find out if the assortativities arise by chance. Quadratic Assignment Procedure(QAP) test will also be used in this analysis for holding network structure fixed and randomly scramble the vertices.

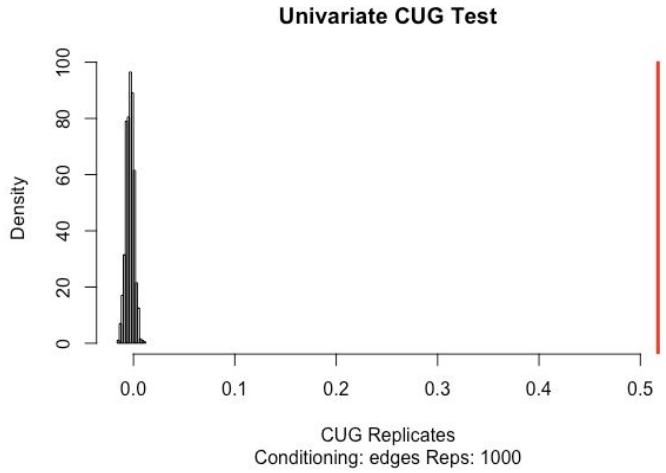
In this process, we used “league” as type parameter to test this directed network. And we also used “edges” as the cmode, which means when their distributions of edges are the same, we will examine if the assortativities are significantly different from random networks.

As the result shows, these three networks’ assortativities are significantly different from random networks. We can reject that the assortativities of these networks arise by chance, the factor of league does affect them. It is further proved that there is indeed transfer preference.

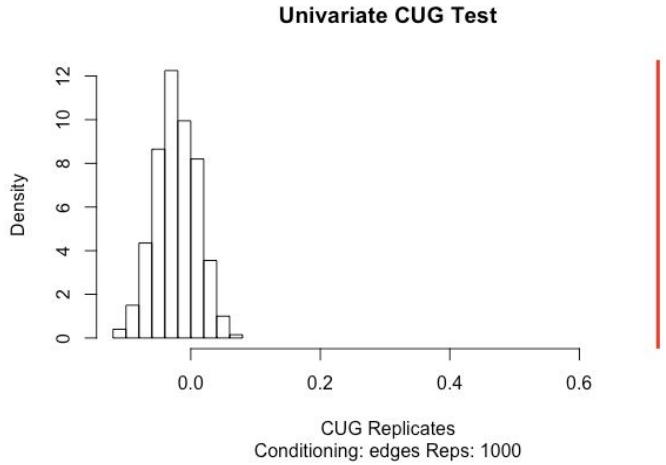
CUG test			
	Assortativity	Pr(X>=obs)	Pr(X<=obs)
Entire Network	0.52	0	1
Scotland Premier League and England Premier League	0.72	0	1
Spain LIGA BBVA and Portugal Liga ZON Sagres	0.65	0	1

QAP test			
	Assortativity	Pr(X>=obs)	Pr(X<=obs)
Entire Network	0.52	0	1
Scotland Premier League and England Premier League	0.72	0	1
Spain LIGA BBVA and Portugal Liga ZON Sagres	0.65	0	1

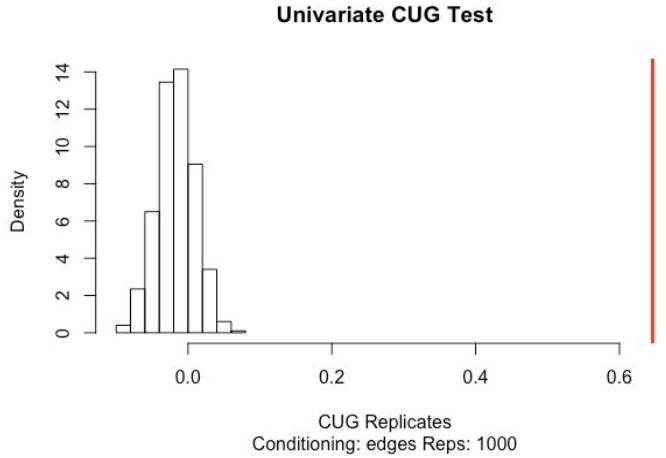
CUG test for entire network



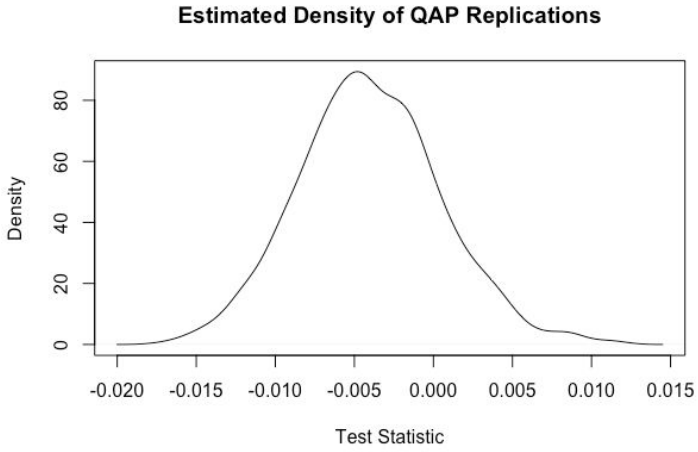
CUG test for Great Britain



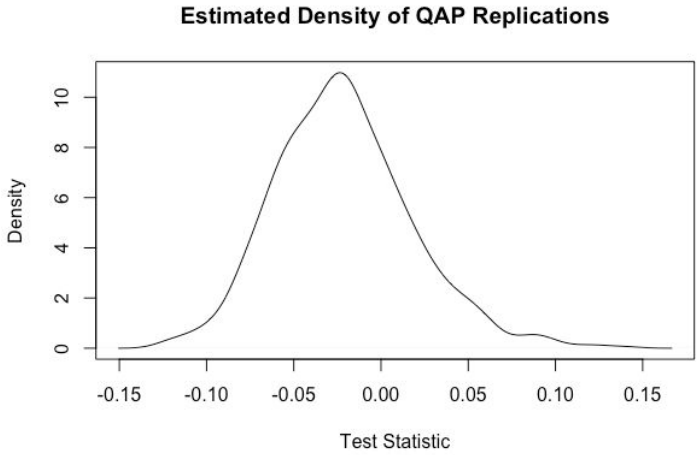
CUG test for The Iberian Peninsula



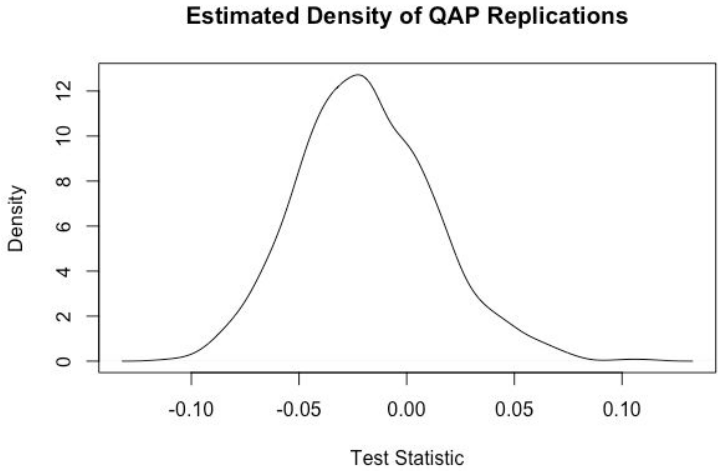
QAP test for entire network



QAP test for Great Britain



QAP test for The Iberian Peninsula



Conclusion

After analyzing the european soccer database by social network analysis methods used above, we can clearly recognize the transfer communities. In this process, we can also find that not each league becomes a community, some communities are made up of two european leagues because of the intensive player transfers between them. For example, Spain LIGA BBVA and Portugal Liga ZON Sagres build up a community,The Iberian Peninsula. Spain and Portugal are geographically close and it reflects there might be a proximity principle for soccer player transfers.

What's more, there is one or two "stepping stone" in most leagues, such as West Ham United from England Premier League and Sevilla FC from Spain LIGA BBVA. Although this kind of clubs might not be the strongest clubs in their leagues, they can act as a middle point of many clubs and has its importance in the European soccer. A soccer player is willing to join these clubs to gain more attention or reputation and that makes them more likely to join some top clubs in the future. This is the reason for the clubs' high betweenness.

According to the result from CUG test and QAP test, the assortativities of our networks differ from random networks. It proves that the factor of leagues does affect the transactions of the soccer players.

For the limits of our analysis, we don't have time to explore how the transactions relate with the ranking of each club, how the transactions change over time, how the transfer communities change over time and what if we take transaction fee into account how the network structure will change. There are so many attributes we haven't used. We hope we can get a chance to make it better in the future.