**Football Transfers Data Analyze**

**Team members:**

* **Roy Liberman –** [**roy.liberman@** **mail.huji.ac.il.com**](mailto:roy.liberman@%20mail.huji.ac.il.com)**, roy.liberman**
* **Eden Yerushalmy –** [**eden.yerushalmy@mail.huji. ac.il**](mailto:eden.yerushalmy@mail.huji.%20ac.il) **, eden107**
* **Moran Polack –** [**moran.polack@mail.huji.ac.il**](mailto:moran.polack@mail.huji.ac.il) **, polack1989**

\*\* GitHub address: https://github.com/polack1989/FootballTransfers

**European football transfer market introduction:**

The football industry in Europe is valuated in hundreds billions of dollars. One of the major business aspects of football is the Transfer Market. Every football season there are two period of times which called the transfer window. In those period of times, each team can buy and sell players according to the regulation of FIFA (International Federation of Association Football) and UEFA (Union of European Football Associations). There are three ways for a player to switch team –

1) Transfer – the destination team pays money to the origin team.

2) Loan – the player moves to the destination team for a limited time and returns to his team.

3) Free – the contract of the player in the origin team has ended, the player has the right to join any club.

FIFA and UEFA regulation supposed to prevent situation where teams has transfers agreements. The reason this regulation exists is to give more power to the players and keep the market open (without monopoles).

**Problem description:**

Analyze teams transfer history in order to study existence of financial connection between European football clubs in terms of the amount of player and money that was transferred between them in the last decade.

Our assumption was that exceptional connections as described above exist. We wanted to figure out if indeed the phenomenon occur and in which extent.

**Data**

Our data includes **all** **transfers** in Europe between the years 2007 - 2017.

Record is a tuple of the following values: *Player Name, Transfer Year, Original Team, Original Country, Destination Team, Destination Country, Transfer Type and Price.*

Where Country refers to a team (etc. Barcelona – Spain), Price scale is Millions of Euro.

The data contains 5,932 transfer records, 4,018 Players, 723 Teams and 79 Countries.

The size of the data is 1 MB.

We used a crawler to collect all the information. Site address - <http://www.soccernews.com/soccer-transfers>

Note – FIFA and UEFA **do not** publish this information, mainly because of pressure from the biggest teams in Europe, who believe that this kind of information will increase the price of players and regulation.

**Solution**

* **Data collection and organization –** we built a crawler using python package “Scrapy” to collect the date from the website. The website data was not fully associative. We addressed the following issues:
  + - Player names, Teams names – the names were not consistently (etc. case sensitive, sometimes written in Latin letters, spaces and special signs). Moreover, teams were written in several names (etc. Barcelona = “Barca”, “FC Barcelona”, “Barcelona”).

We changed all string to upper case, deleted spaces, and created a team mapping using dictionary to avoid multiple team names.

* + - Price – prices were written in several formats (etc. “15M”, “15Millions”, “15,000,000”, “15”, “500,000”, “NA”, “Loan”, “Free” and more).

We parsed it into two attributes – Transfer Type and Price. Transfer type is either Transfer, Loan or Free (see above introduction for more details). Price is number (Millions of Euros) in case it is transfer, null otherwise.

* + - Countries – the website data does not include the country of each team. We added the Countries manually while creating the team mapping dictionary.

We used python package named “XslxWritter” to create several excel tables of the data in a file named “**project\_data.xls**”.

* **Algorithm for finding cliques and communities -** We have used the clique percolation algorithm in order to extract communities of teams that has exceptional relations with one another.

We built a team graph which each node represents a team, there is an edge in the graph for each transfer between two teams. We also added weights on the edges that represent the amount of transfers between the teams and money that was transferred between them.

First we filtered out less relevant edges and nodes according to the amount of transfers between the teams, second we extracted every max-clique from the filtered graph, and filtered out the less significant cliques by calculating the cliques with the highest amount of transfers between the teams in the cliques.

At the end we have extracted the communities from the graph using only the cliques extracted in the previous stage.

Our results in the clique stage gave some interesting connections between the teams. The communities extracted from the cliques showed a clear image on the relations between the teams. We will extend the conversation on the details of the implementations and the analysis of the results in the experiment section.

* **Data visualization –** Cytoscape software for graph (network)visualization and Excel for charts.

**Experiment**

* **Evaluation criteria**

In order to tell if the player transfer relations we found are exceptional in terms of the amount of transfers made between the teams, we calculated the ratio between the averages of transfers inside a clique to the average of transfers outside the clique.

The calculation is:

[(number Of Transfers Inside A clique) / (The Size of the Clique)] / [(number Of Transfers with Teams outside the Clique) / (the number of teams outside the clique that players were transferred to)]

We define success if the average ratio is significantly bigger than 1, I.e. if the amount of players that was transferred inside the cliques in comparison to the size of the cliques was significantly higher than the same average with teams outside the cliques.

* **Setup**

After Extracting and organizing the data, in order to examine the phenomenon of exceptional player transfer relations between European football clubs, we used the clique percolation algorithm. We assumed that a group of more than two teams having exceptional relations will be characterized by substantial amount of players or money that was transferred between the teams.

Graph representation:

The transfers data was represented as an undirected graph.

Nodes: node in the graph represents a football club.

Edge: two nodes such that there is a transfer between them i.e. player was transferred between the teams.

Weights: On each edge we assigned two kinds of weights the first is the total amount of transfers and the second one is the total amount of money that was transferred between the teams.

Data filtering:

*Preliminary Filtering* - After the first Processing of the algorithm we got un-informative results, for example two teams that has one transfer between them. After analyzing the data and the graph we figured out that the data had too much noise.

We have filtered out all edges that has less than two transfers between them **or** less than a total of 3 million euros that was transferred between the teams.

We chose these parameters and thresholds by analyzing the average transfer cost and the average amount of transfers between the teams.

The process of choosing the thresholds also included examining the results of the algorithm and the substantial data that it produced using the thresholds and parameters that was set and according to the evaluation criteria decided.

After analyzing the results we noticed that we are losing important data by filtering teams that has 10 or more transfers and we decided that teams that has less than five transfers in ten years are teams that are not relevant to our research. Finally we found a threshold that in our perspective filter out only the less relevant data as good as possible. The Change in the mean amount of transfers supports our claim.

|  |  |  |
| --- | --- | --- |
|  | Before edge Filtering | After edge filtering |
| Mean amount of player  transfers per team | 2 | 8 |

*Clique Filtering* - in order to find exceptional relations in the transfers graph, we looked for max cliques in the graph. We defined strong clique as a clique with many player transfers between the teams in the clique. We assumed that exceptional relation between teams will be characterized as a strong clique. In order to find the strongest cliques we extracted all the cliques from the transfer graph, and extracted the strongest cliques from each clique size, i.e. we calculated the strength of every clique by summing the amount of players that was transferred between the teams in the clique and extracted only the Strongest five cliques from each clique size.

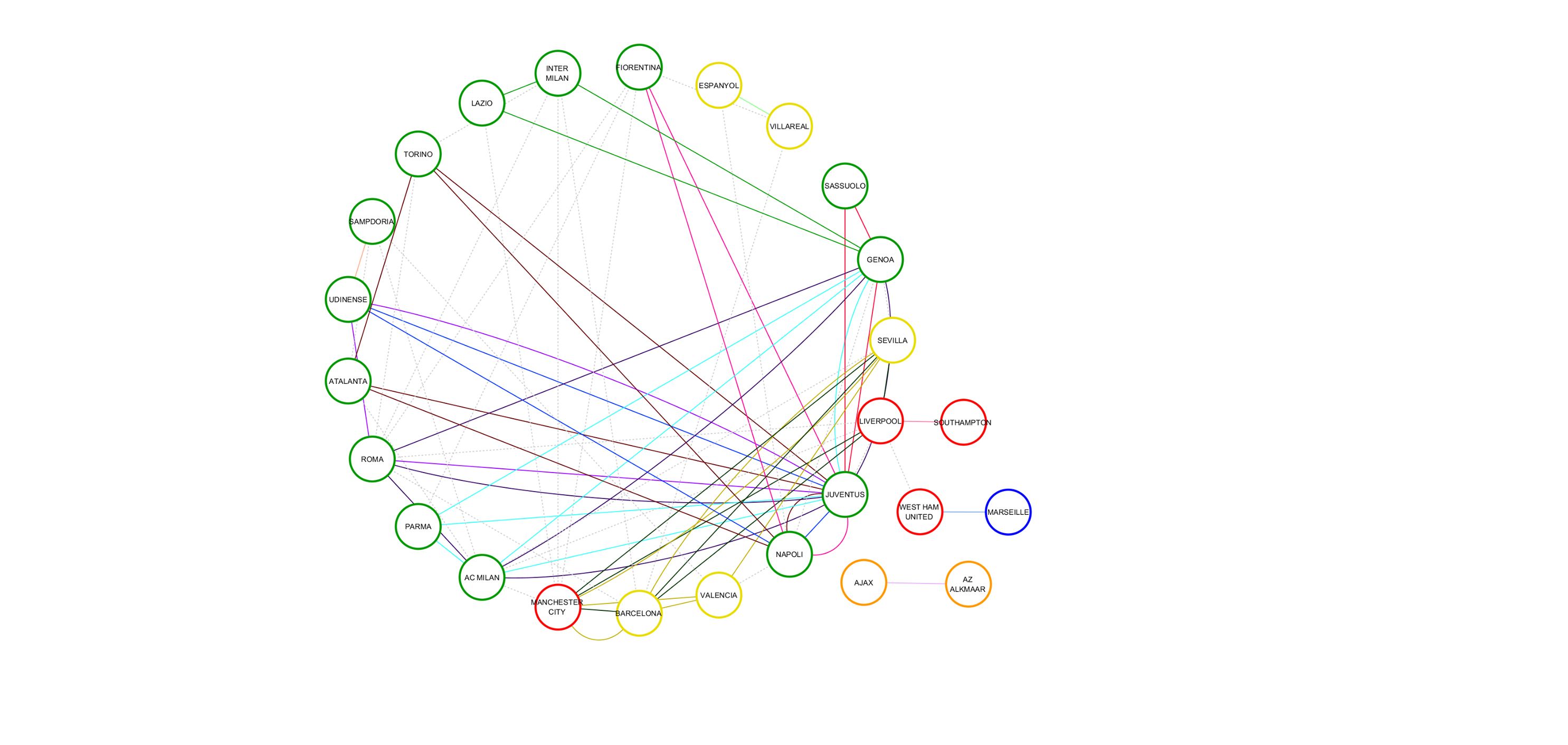
|  |  |  |  |
| --- | --- | --- | --- |
|  | Nodes | Edges | Cliques |
| Before Filtering | 627 | 2590 | 261 |
| After Filtering | 145 | 359 | 15 |

*Clique and community extraction -* We figured that there may be a connection between the strongest cliques we found, we wanted to see if by calculating the communities extracted from the strongest cliques will generate a hidden connection or give a different point of view on the connection between the cliques. We used the clique percolation algorithm on the filtered graph and extracted the communities only from the strongest cliques extracted all the max-clique from the graph. The implementation of the algorithm is the standard one as the one taught in class, the main difference is that we activated the algorithm on filtered graph and relevant cliques only.

* **Results and Visualization**
* Clique summarized

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | Clique Size | Clique Edge color | Teams in Clique | Amount of transfers | Total amount of money M Euro |
| 1 | 2 |  | 'SAMPDORIA', 'UDINENSE' | 6 | 25.4 |
| 2 |  | 'LIVERPOOL', 'SOUTHAMPTON' | 5 | 119.9 |
| 3 |  | 'MARSEILLE', 'WEST HAM UNITED' | 4 | 48 |
| 4 |  | 'AJAX', 'AZ ALKMAAR' | 4 | 10 |
| 5 |  | 'VILLAREAL', 'ESPANYOL' | 4 | 11 |
| 6 | 3 |  | 'NAPOLI', 'UDINENSE', 'JUVENTUS' | 18 | 198 |
| 7 |  | 'LAZIO', 'INTER MILAN', 'GENOA' | 15 | 62.5 |
| 8 |  | 'GENOA', 'JUVENTUS', 'SASSUOLO' | 15 | 69.5 |
| 9 |  | UDINENSE', 'ROMA', 'JUVENTUS | 15 | 122.35 |
| 10 |  | 'FIORENTINA', 'JUVENTUS', 'NAPOLI' | 14 | 132 |
| 11 | 4 |  | 'GENOA', 'JUVENTUS', 'AC MILAN', 'PARMA' | 27 | 69 |
| 12 |  | 'GENOA', 'JUVENTUS', 'AC MILAN', 'ROMA | 24 | 152 |
| 13 |  | 'VALENCIA', 'MANCHESTER CITY', 'SEVILLA', 'BARCELONA' | 23 | 453.4 |
| 14 |  | LIVERPOOL', 'MANCHESTER CITY', 'SEVILLA', 'BARCELONA' | 19 | 387 |
| 15 |  | TORINO', 'JUVENTUS', 'ATALANTA', 'NAPOLI' | 18 | 158.5 |

* Clique Graph



* Communities summarized and Graph

|  |  |  |
| --- | --- | --- |
| Community size | Teams in community | Total amount of money M Euro |
| 11 | GENOA', 'NAPOLI', 'TORINO', 'SASSUOLO', 'PARMA', 'ATALANTA', 'JUVENTUS', 'ROMA', 'UDINENSE', 'FIORENTINA', 'AC MILAN' | 573.9 |
| 3 | 'LAZIO', 'INTER MILAN', 'GENOA' | 62.5 |
| 5 | 'LIVERPOOL', 'SEVILLA', 'MANCHESTER CITY', 'BARCELONA', 'VALENCIA' | 652.2 |

C:\Users\eyerusha\Documents\University\מחט\commGraph_corrected.tif

**Clique result analysis –**

We found 15 significant cliques, we can see that most of the teams in the two sized cliques are from different countries and include mediocre sized teams.

All the teams in the three sized cliques are all from the same country Italy and include teams from different sizes.

In the four sized cliques we can see that most of the cliques include teams from the same countries Italy and Spain.

The four sized cliques that include teams from different countries include teams with management from the same country, like the clique LIVERPOOL, MANCHESTER CITY, SEVILLA, BARCELONA all two of them are Spanish with Spanish management and two of them are English with Spanish management.

The four sized cliques group include most of the biggest teams in the premier leagues of Europe.

After calculating the evaluation criteria for all the teams in each clique we found out that in average the teams we found transfer 1.6 times more players per teams in clique than players per teams outside their clique (all other teams). For example the figure below, show that inside Liverpool and Southampton clique Liverpool and Southampton transfer in average 2.28 more players to each other than to all other teams.

**Community result analysis –**

In the community graph we can see strong connections between teams in Italy, and some teams in Spain.

The community that include teams from different countries, England and Spain, are managed by Spanish directors that believes in Spanish football.

Two communities from the three contains teams from Italy.

There is no community of teams from England and Germany which are in the top 5 premier leagues in Europe.

* **Impediments**

The main issue we encountered during the experiment was how to extract the important data and filter out all the irrelevance data. The issue was characterized as finding the parameters and thresholds that will achieve this goal.

**Future Work**

Investigate the correlation between the amount of money paid for a player and the time period that player stayed in the team. Moreover, check if there is a correlation between the amount of money team invests and the achievements of the team and player specifically.

To do so, more data about the players (etc. personal achievements, amount of goal scored and more) and about the teams (achievements and awards)

Another interesting aspect is to add to our data information about the team’s budget and understand the influence of transfer market on it. Team budget constructed by few major components:

* Transfer Market
* Broadcasting Rights and competition money (winning awards)
* Brands and Sponsors

It will be interesting to discover and try to understand if there a relation between the amounts of money team invest in transfer market and the bank balance of the teams.

FIFA and UEFA do not publish information about budgets of the teams.

By the way – it seems like the amount of money does not related to the bank balance of a team, but to the amount of money owner have...).

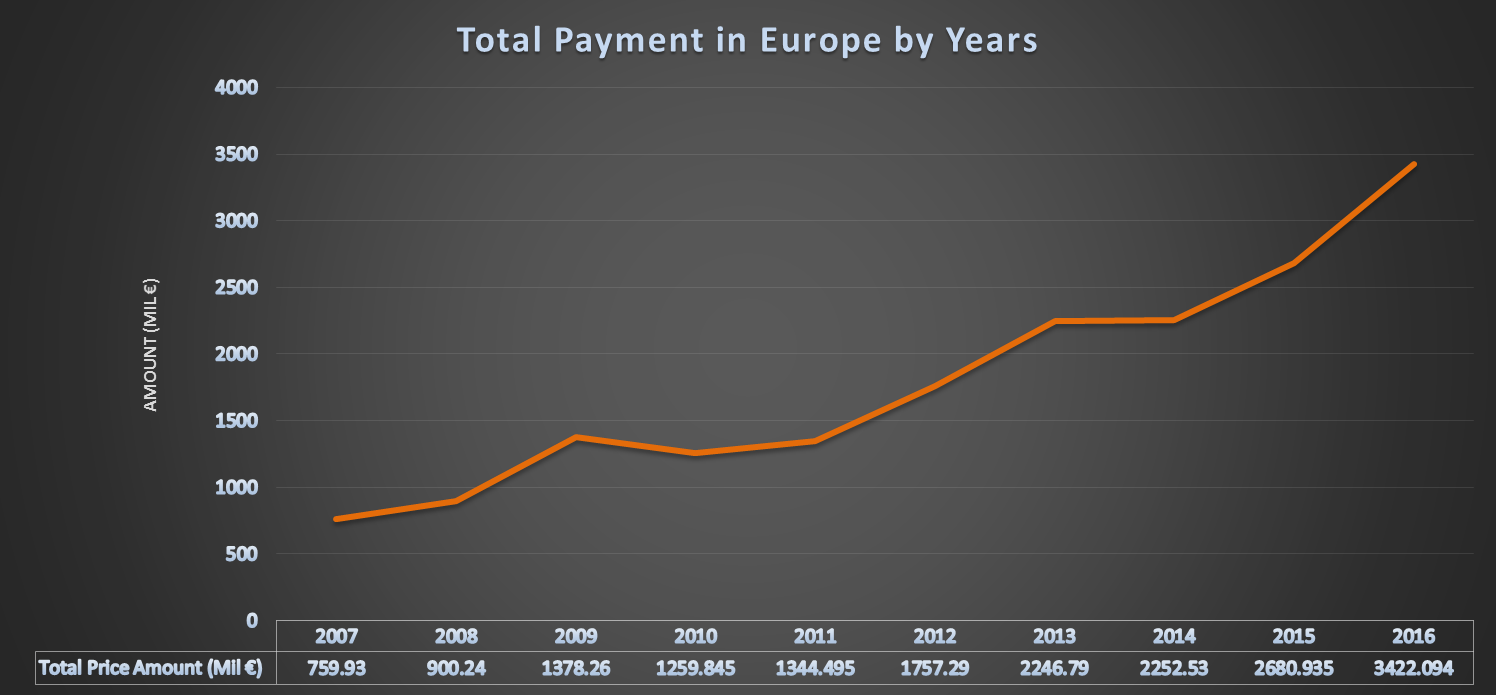
**Brief conclusion**

According to the data presented above there are groups of teams that transfer more players within the group than outside the group, in average the teams presented above transfer 1.6 times more players per team in the group than players per teams to teams outside their group, according to the evaluation criteria.

It is hard to say that the data presented is actually implies on un-ethical relations between the teams in the different cliques. There may be other parameters that influenced the data like ethnical, financial or geographical parameters.

However, if there is an un-ethical relation between teams, according to our data most probably they will be characterized by a significantly bigger amount of players transferred per team in the clique than other teams and both teams will probably be from the same country

**Appendix charts**

****

|  |
| --- |
| The amounts clubs spending on buying players increased significantly over the years.  There are 3 main reasons :  1) Sponsors and Media increased the amount of money that has been given to the clubs (mainly because of popularity). 2) New owners of a big number of teams – the new owners invested big amount of money in their teams after they bought it.  3) Reasons 1-2 increased the competition between clubs about each and each player. |
|
|
|

The graph shows the subtraction between team’s income and outcome from players transfer deals over the last decade. The 9 teams marked with red has the biggest negative balance, all of those teams ranked as top clubs in the world. The other 9 marked with green are all quality teams but not top teams, all of them have the highest profit from players transfers deals (mid-level clubs).

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |

