A Network Tour of Football Transfer Market
NTDS 2019

Team 23
Saibo Geng, Olavo Gonçalves Bacelar, Xiao Zhou, Clarisse Poupon-Pourchot

I. Introduction

Football is a very popular sport in Europe that also involves very large ammounts of money. Player transfers play a central role in professional football, with an estimated worth of 25.5 billion. Twice a season, during the transfer windows, clubs try to build their squads by keeping their best players, transferring and/or signing others with the ultimate aim of building a better team and enhancing performances on the pitch - and sometimes they even specialize in this, finding unknow players, build them up and give them recognition, and then transfer them to bigger teams for profit. Under the assumption that the topological analysis of the network structure of the transfers could give us insight on this market, we will create a graph based on the transfer of soccer players, explore and visualize the data thoroughly, and exploit it with techniques learned in class, such as clustering.

II. Data Acquisition

The dataset that we chose to use is called **European Football Transfers Dataset**, which is available in Kaggle[1]. It packages data scraped from the website Transfermarkt[2], which publishes many information on the football transfer market. The original dataset deals with the transfers from the clubs in 9 leagues in Europe to and from other clubs all around the world, in the period 1992-2019. In our project, we will focus only on the top leagues in Europe in more recent years. Thus, we will filter the dataset to extract the transfers from 2010 to 2019 among **English Premier League, Spanish Primera Division, Italian Serie A, German Bundesliga, French Ligue 1, Dutch Eredivisie and Portugese Liga Nos**. After this filtering, we still arrived at a dataset consisting of more than 60420 transferences (including duplicated ones, when both clubs are part of these leagues). Each dataset include 11 features including: **club name, player, name, age, position, club involved name** (the name of the club to or from which the transfer was made), **fee transfer movement** (whether the transfer was to inside or to outside of the club), **fee, league name, year**.

In order to produce a signal which measures the **overall excellence** of a club, we scraped the **Soccer Power Index data** [3] from the election and football predictions FiveThirtyEight website and use it as an attribute for each club. This index contains a overall score (ranging from 0 to 100 for a total number), an average number of goal scored and an average number of goal conceded for a total number of 624 global clubs (not only European).

Another Kaggle European Soccer Database [4] is also downloaded as extra attributes of clubs. This data set contains more detailed measure of a club, such as defense pressure, chance created etc.

III. Data Preprocessing and cleaning

Unfortunately, our Player Transfer Data and Soccer Power Index data does not share the same schema for club names and league names. Mismatch is widespread, such as

- 'PSG' VS Paris 'Saint-Germain'
- 'Man. City' VS 'Manchester City'
- 'Primera Division' VS 'La Liga' (Spanish league)

To avoid ambiguity, we fix the name of the 7 selected European football leagues in all datasets as stated in the previous section.

The club names were more difficult to unify as we have more than 200 clubs and the club names can be in Spanish, French, German. We first remove all the prefixes and postfixes like FC, AS, CL etc. Then we hard-code the rest of clubs.

IV. Data Exploration and Network Structure

Our network is built upon the transfers between the considered leagues, i.e. both clubs involved in the transfer need to be in one of the 7 top European football leagues. This omits a considerable number of transfer between European clubs and South American's clubs for example, but it's necessary and quite natural since we don't have information on those clubs' transfers to other clubs. In the constructed graph, each node is one club and each vertex represents a transaction between two clubs from 2000 to 2019.

With the information provided, we can build different networks:

1) undirected network with the total number of transfers as weight of edge
2) undirected network with total number of transfer fee from as weight of edge
3) undirected network with a weight of edge designed as $w_{i,j} = \frac{2}{\frac{1}{s_{i,j}}+\frac{1}{b_{i,j}}}$ with $s_{i,j}$=transfer fee of selling from club-i to club-j,$b_{i,j}$ =transfer fee of buying from club-i to club-j.

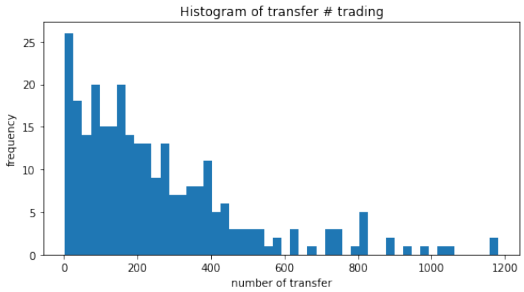


Figure 1: Histogram of the number of player transfers between the 7 European leagues, for each club, during the period from 2000 to 2019

The above figure shows the distribution of the number of player transfers between those 7 European leagues' clubs from 2000 to 2019.

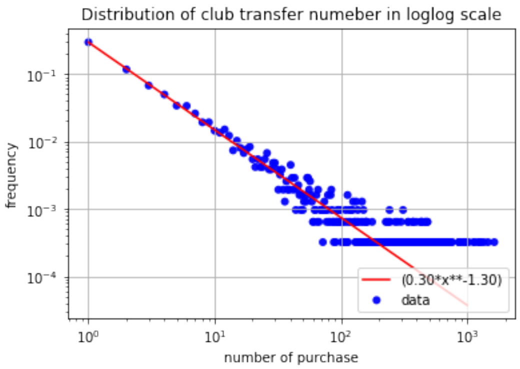


Figure 2: Distribution of club transfer number in loglog scale

The plot shows the distribution of club transfer number follows a power law $P(X) = 0.30e^{-1.30}$, thus we have a heavy-tailed distribution. In a heavy-tailed distribution, outliers are rare but not that rare, which means we can have some clubs with a huge number of transfers while most of the clubs have quite few transfers. But before going on, we need to point out **'transfers' here and in the following report** only refers to transfers within European top leagues. Thus a club with a tiny number of transfer in our data doesn't mean it didn't trade players, but rather it didn't trade players within top European leagues. In this case, several possibilities came up: 1. The club's transfers were done primarily with European's second or third division clubs 2. The club signed many young players from South America's clubs (which we don't take into account). 3. The club dealt with global clubs such as American or Asian clubs (which we don't take into account neither).

The clubs with the highest number of transfer are the following:

- Porto 1182
- Udinese 1153

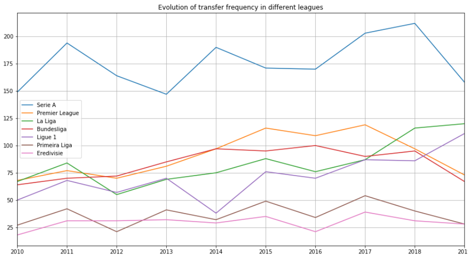


Figure 3: Evolution of transfer frequency in different leagues

- Roma 1063

Those are not necessarily the most famous clubs, which are typically in a middle position in terms of the number of transferences.

- Chelsea 887
- Liverpool 634
- Barcelona 578
- Real Madrid 592

V. Graph Properties Exploration

A. *Build graph*

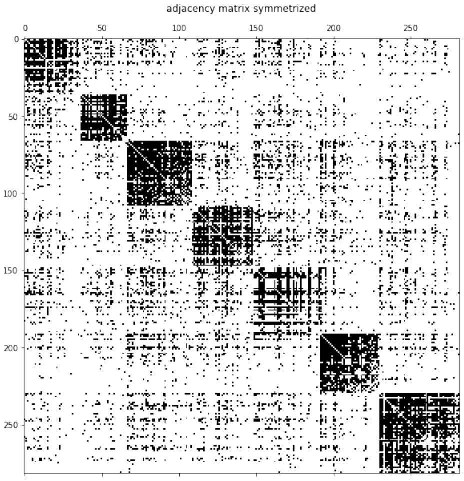


Figure 4: Adjacency matrix, where the the nodes(clubs) are sorted according to the league in the order: 'Bundesliga'(DEU), 'Eredivisie'(NLD), 'La Liga'(ESP), 'Ligue 1'(FRA), 'Premier League'(ENG), 'Primeira Liga'(PRT), 'Serie A'(ITA)

We first build the adjacency matrix. The transfer action is directed in reality, but here we first construct a symmetric adjacency matrix by neglecting the direction - A_sym=(A+A.T). We observe 7 blocks on the diagonal of the adjacency matrix, which corresponds to 7 leagues respectively. The size and the gray scale of the block presents qualitatively the number of links established inside the league. The block of Premier League is the least dense and this is coherent with the fact that Premier League is the most international league in the Europe. The block of Serie A is particularly large and this shows a extra-active transfer market inside Serie A.

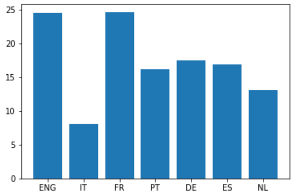


Figure 5: Percentage of the transferences that the clubs in a certain league made from and to their league, compared to all the transferences that involve their league

This can also be seen much more directly with the following graph which shows explicitly the total number of inter-league transfers for each league, as compared to the total number of transfers involving this league (see Figure **??**. Here, exceptionally, we do use the full dataset, which doesn't take out transfers involving leagues other than the 7 here considered.

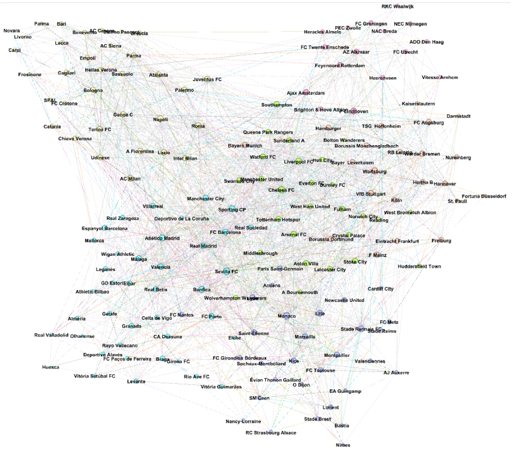


Figure 6: Visualization of the Built Network. For interactive visualization, refer the interactive network published on: https :// zx−joe.github . io / Soccer_Transfer_Network/

B. Graph Analysis

1) Full Graph: Our full graph is connected with diameter = 3 and average path = 1.9.

From the degree distribution, we recognize the type of our network to be scale-free. A scale-free network has many small degree nodes, and a few high-degree nodes (hubs), and that is indeed the case: in our graph most nodes are small clubs which typically achieve very few transfer with the 7 major leagues (in most cases, these transfers are of outgoing type for them), and a few others attract or sell a lot of players.

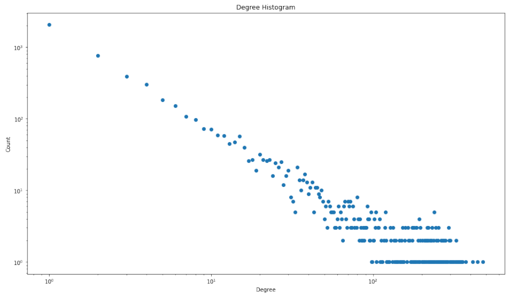


Figure 7: Degree Distribution of the graph in log-log scale

2) Reduced Graph:
Type: Undirected Graph
Number of nodes: 282
Number of edges: 6708
Average degree: 47.5745
Avg. Shortest Path Length: 1.9172
Diameter: 3
Sparsity: 0.1693

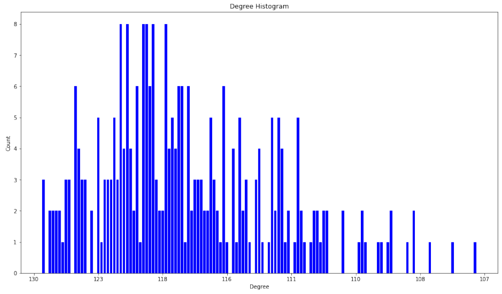


Figure 8: Degree Distribution of the reduced graph

The degree distribution of our second graph, which only considers transactions between clubs in 7 major leagues, seems to follow the Poisson distribution. The degrees of the majority of nodes are close to the average degree ¡k¿=47. We have a so-called homogeneous network. This argument is reasonable because our network is formed by the clubs of major 7 leagues, thus the clubs' transfer strategy and budget are indeed far more homogeneous than in the previous graph.

The network can be simulated by a Erdős–Rényi graph as follows:
Name: Simulated Erdős–Rényi graph
Type: Un directed Graph
Number of nodes: 282
Number of edges: 6781
Average degree: 48.0922
Avg. Shortest Path Length: 1.8292
Diameter: 3
Sparsity: 0.1711

Our interpretation towards this is: 1. Club's transfer intention is not random. Every club has a clear plan for their transfer in each season, either to increase their competitiveness or to make revenue by selling players. However, their objective club of transfer is quasi-random (or in other words: unpredictable) because every club only has 20 players (40 if we count team B) at a time and to find a player which has a reasonable market value and compatible with the club's tactics is hard. However, there may be some regional preference like domestic transfer or with South American clubs.

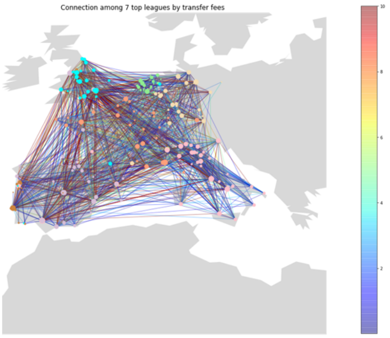


Figure 9: Visualization of Reduced Network on Europe Map, where the edge color means the amount of transfer fee between 2 club nodes. For more details refer to the appendices

3) *Clustering Coefficient:*

The clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. The global clustering coefficient of the reduced graph is 0.519. The most reputable clubs tend to have a clustering coefficient below average, ranging from 0.35 to 0.48. Small clubs tend to have a higher clustering coefficient. Our interpretation of this: top clubs and small clubs have different transaction models. Top club's specialize on finding the top talented players, thus they have less fixed trading partners. On the other hand, smaller clubs' transactions are based on budget, and they may establish a local transaction group network that trades between each other to economize negotiation time, and maybe even obtain a promotion price.

4) *Centrality:*

In order to measure the importance of a club in European transfer market, we decide to combine 4 different algorithms available on networkx: degree centrality, page rank, betweenness centrality and closeness centrality. We use the mean of the 4 algorithms, and the 5 most central nodes are 'Sporting CP', 'Benfica', 'Monaco', 'Chelsea' and 'Porto': We find Portuguese Big Three (Porto, Sporting CP, Benfica) at the top. Indeed, there are many brilliant football players that have developed greatly while at these clubs and subsequently have been sold off at greater prices to other giant European clubs like Manchester United etc. For these reason, they are also called ”feeder clubs”.

To take into account the transfer fee factor, we construct a similar graph using transfer fee as weight of edge. By using the same algorithm on this weighted graph, we obtain the 5 most central nodes as 'Porto', 'Liverpool', 'Roma', 'Benfica' and 'Sevilla'.

To only use the transfer fee as weight may cause that a club only buys or only sells players can also occupy a high centrality. But this is against the UEFA's requirement: Income and expenditure from transfers of a club should be balanced.

Thus we use the weight designed as $w_{i,j} = \frac{2}{\frac{1}{s_{i,j}}+\frac{1}{b_{i,j}}}$ with $s_{i,j}$=transfer fee of selling from club-i to club-j,$b_{i,j}$ =transfer fee of buying from club-i to club-j to build the graph. The clubs with high centrality in this graph are those having achieved transactions with various clubs and keeps a balanced income / outcome. The top 5 clubs for this include 'Sporting CP', 'Benfica', 'Monaco', 'Chelsea' and 'Porto'.

VI. Graph Signal Processing of transfer network

A. *Graph Signal*

We wonder if the transfer network is related to the strength of a club. It is known that there is a substantial redistribution of players from big clubs to smaller clubs to reach a competitive balance. We use the **Soccer Power Index**[3] as attributes of nodes: a score from 0 to 100 measuring the overall excellence of club. Some nodes (clubs) don't figure in this index. These are typically small clubs which participated the 7 major leagues but not anymore. We fill up their score by the lowest score in their league minus 10.

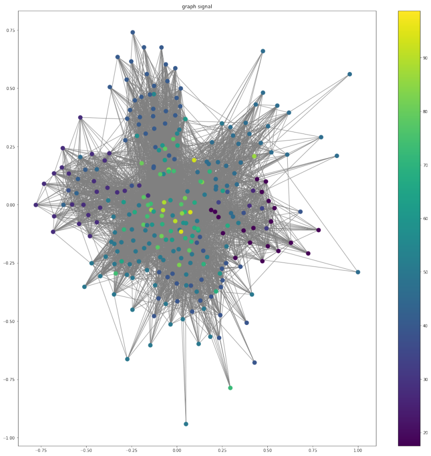


Figure 10: graph signal:Overall Score of Club

VII. Community detection

In recent years, many community detection algorithms have been proposed to unveil the structural properties of networks. In our case, we wonder if the network of transfer could give us enough information to detect communities of clubs. We first implement **K-means clustering**:

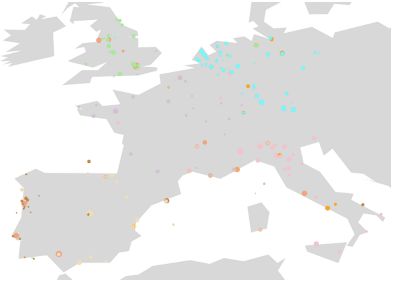


Figure 11: k-means clustering result

Given hyper-parameter K=7, K-means clustering is able to regroup nodes which are in the same league. This result is not surprising since graph clustering is based on the principle that pairs of nodes are more likely to be connected if they are both members of the same community. And as figure 4 suggests, there are significantly more links inside a league.

Another algorithm which shows us underlying community structure in our network is **modularity maximization**. It regroups nodes into 3 communities. A visualization is in the appendices.

VIII. Neural Network with Graph Properties

In this section we will use the properties of constructed graph and features in the dataset to predict the transfer fee category. Here we will merge the player data acquired from European Soccer Dataset[4]. Besides, properties obtained from our previous graph will be considered to find whether these graph properties will influence the prediction, including betweenness, centrality of club nodes, etc. The categories of transfer fee are manually defined according to distribution so that the labels will not be biased.

We implemented neural networks for the classification problem. The major part of the model is a 3-layer MLP with 64 hidden units, and ADAM is chosen for the optimizer. The effect of regularization trick (mainly batch normalization) and graph properties will be studied according to comparison among the learning curves. The machine learning part is implemented in Tensorflow with Keras.

From Figure12, it can be observed that batch normalization can improve the prediction performance through comparisons of learning curves in loss and accuracy. The graph properties show an improvement on the accuracy of validation set (also test set in our case). However, the overall accuracy is not very high. We analyze it may be due to the fact that the transfer fee might be difficult to predict just from the considered features. For example, when a good player's contract is going to expire soon, other clubs may not pay so high a price. Also, it will be more efficient to consider the trend of transfer fee within some period, which requires models concerning time series. Also, if a club payed a lot of money for a player, then it's likely to demand a high fee when selling.

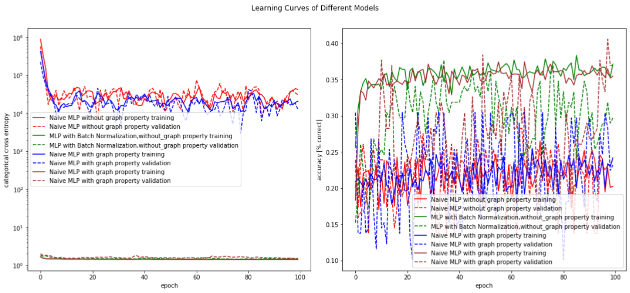


Figure 12: Learning Curves of different neural network models with and without graph properties

IX. Conclusion and Limitation

Due to our dataset size, we can only proceed Europe major leagues scale analysis. But our work is highly scalable and can easily be applied to a larger dataset, which would allow us not to throw away important graph data and better understand the patterns in this market. Also note that the inter-continental interactions are an important factor as transfers with Non-UEFA Countries represents 12% of all transactions according to UEFA's data. In the network structure analysis part, we explored different facets of the economic importance of a club based on transfer records, which goes beyond economic or competitiveness measures. The results in community detection suggest that the affiliation relation of club and league can be recognized to some extent from the network of transfer. For the machine learning part, we conclude that the graph property has a positive influence on the transfer fee prediction. However, for more accurate learning, we will consider time series model like LSTM in the future.

X. Appendices

A. *Size of top 7 European soccer leagues*

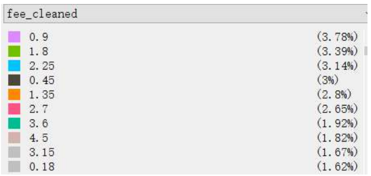
1) 'Bundesliga'(DEU), 18 clubs
2) 'Eredivisie'(NLD), 18 clubs
3) 'La Liga'(ESP), 20 clubs
4) 'Ligue 1'(FRA), 20 clubs
5) 'Premier League'(ENG), 20 clubs
6) 'Primeira Liga'(PRT), 18 clubs
7) 'Serie A'(ITA) 20 clubs

B. *Edge color-transfer fee table for figure 7*

C. *Modularity Maximization result*

References

[1] "European football transfers dataset," https://www.kaggle.com/giovannibeli/european-football-transfers-database.



fee_cleaned		
🟪	0. 9	(3. 78%)
🟩	1. 8	(3. 39%)
🟦	2. 25	(3. 14%)
⬛	0. 45	(3%)
🟧	1. 35	(2. 8%)
🟥	2. 7	(2. 65%)
🟩	3. 6	(1. 92%)
■	4. 5	(1. 82%)
■	3. 15	(1. 67%)
■	0. 18	(1. 62%)

Figure 13: edge color transfer fee table for interactive network publication

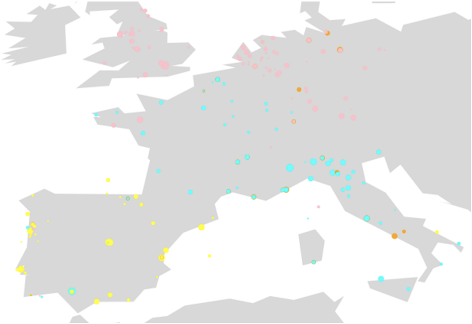


Figure 14: edge color transfer fee table

[2] "Soccer transfer market statistics," https://www.transfermarkt.co.uk/.

[3] "Global club soccer rankings," https://projects.fivethirtyeight.com/global-club-soccer-rankings/.

[4] "European soccer database," https://www.kaggle.com/hugomathien/soccer.