Predicting player transfers in the small world of football

Roland Kovacs, Laszlo Toka

MTA-BME Information Systems Research Group
Faculty of Electrical Engineering and Informatics
Budapest University of Technology and Economics, Hungary

Abstract. Player transfers form the squad of the football clubs and play an essential role in the success of the teams. A carefully selected player squad is a prerequisite for successful performance. Consequently, the main topic of the football world during summers is the transfer rumors. The aim of our research is to predict future player transfers using graph theory. In this paper, first, we examine the networks formed in the football world and whether if these networks have small-world property. To do this, we set up an acquaintance graph among professional footballers based on if they have ever been teammates. We make a similar graph for the managers, in which we consider two coaches connected if they have coached the same club. Moreover, we also analyze the network that has developed among the teams in the past 14 years, in which links illustrate player transfers. Using the graphs' metrics and the information about these transfers, we make a data mining model for predicting the future transfer of players. The model can be used to predict who will transfer into a selected league. Different leagues show different features as the most important ones when it comes to buying a player, but in every case that we studied, the features extracted from the graphs are among the most essential ones. These features improved the performance of the player transfer prediction model, giving sensible possibilities about the transfers that will happen. Network science has become widespread in recent years, allowing us to explore more and more networks. By examining complex networks, we can obtain information that would not otherwise be possible and that can have a massive effect on predictions. We show that by using this information we can create meaningful features that can improve the performance of the predictive models.

Keywords: sports analytics · European football · soccer · small-world · network · machine learning · player transfer.

1 Introduction

This article consists of two parts, first we perform network analysis using graph theory and then we use data science to predict football players' transfers using the information extracted from the networks that we build in the first part.

Networks appear in our everyday lives and affect numerous aspects like the spread of information, or the distribution of vaccines during a pandemic. Our

brain is essentially a network of connected neurons and our society likewise forms a network with various acquaintances.

Today we know plenty about complex networks and scale-free networks, thanks to numerous research results in this field [1–4]. These networks are usually small-world networks, for which the main characteristic is that the average shortest path between the vertices is small. This means that we can get from any vertex to another without having to go through a lot of other vertices. One of the best-known pieces of research on this topic is named after Milgram who proved that social networks are small-world networks, and Albert Barabasi proved the same about the Internet [1, 2]. In this paper, we analyze the networks formed in the world of football. We are looking for an answer to the question: do the graphs developed in that world have small-world properties? To do this we first examine the graph formed by the players' relationship of familiarity. We consider two players to be familiar if they have ever been teammates. Using a similar method, we create a graph among the managers in which they have a link if they have managed the same club. Not only relationships between people can be worth exploring, thus we also involve teams in our research and build a network based on the player transfers that emerged between those. With these networks we can analyze such complex problems that cannot be done in any other way since we can understand not only the single elements but also their relationships and interactions.

In football, player transfers play an essential role as those form the squad of the teams. The hottest topic of summers in the world of football is transfer rumors and guesses which team will purchase a given player. Football clubs want to strengthen their squads every year, and if they succeed, they could have great success next season. The opposite is also true, if the club fails to strengthen the squad, they may soon be at a disadvantage against the rival teams. Using the information of the graphs, we create a model that predicts future player transfers. The goal is not only to be able to predict transfers using past transfer data, but also to make more accurate forecast using information extracted from complex networks.

The paper is organized as follows. In Section 2 we present the results of the relevant research from network science and sports analytics. In Section 3 we present the basic properties of our constructed networks, with great emphasis on scale-free networks and the small-world property. Then, in Section 4, we specify the steps we took to develop the player transfer prediction model and describe the results of our research. Finally, we summarize the most important conclusions in Section 5.

2 Literature review

One of the biggest breakthroughs in the world of network science was brought by Pal Erdos and Alfred Renyi who discovered random networks [5]. The above-mentioned scientists both wanted to examine complex graphs, but at that time it was not that obvious how to model networks. They thought most of the

networks that occur in real life are unpredictable, asymmetric in structure, and rather appear random. Due to this assumption, the formation of graphs was characterized by the principle of randomness, which means that the best way to build a graph is to add the edges completely randomly between the nodes. Accordingly, each vertex has the same probability to collect edges, so most of the nodes have approximately the same degree. This means that if we want to draw a histogram of the degrees, we get a curve with a Poisson distribution. This was proven by Erdős's student, Béla Bollobás in 1982 [4]. The degree distribution, in this case, follows a bell curve, it has a maximum point, and the other vertices do not deviate much from this. We do not find vertices with very extreme degrees that differ from the average to a large extent. This suggests that, if we look at a social network, all people have nearly the same number of acquaintances. Or if we look at the World Wide Web and measure the connectivity of websites, pretty much each page points to the same number of other pages. Although most networks today are known to be non-random networks, their discovery has greatly contributed to the development of network science.

As stated above, numerous networks have been proven to be small-world networks. These networks have unique characteristics. According to Granovetter's studies, small-world networks have higher clustering coefficients thanks to the many complete subgraphs [6, 7]. Clustering coefficient is a metric in network science, which measures the probability if two neighbors of a vertex are also adjacent to each other. Real-world networks usually have a clustering coefficient between 0.1 and 0.5 [8]. Another vital feature of small-world networks is that the length of the average distance grows only logarithmically with the number of vertices [1]. Consequently, the average distance is relatively small. According to Amaral and his co-authors, the diameter, which is the largest shortest distance between nodes, is also small in the small-world networks [6]. Additionally, the degree distribution of such networks follows a power-law distribution. This means most of the nodes have a small degree while only a few have greater degrees. These nodes are responsible for the weak connections and we usually call those hubs. Such scale-free networks are all small-world networks [6] and inherently differ from random networks.

Research has already been done in the world of sports related to the analysis of networks. Yuji Yamamoto and Keiko Yokoyama examined the networks that emerged in a football game, representing the players and the passes between them. They concluded that the degree distribution follows a power law and that the exponent values are very similar to real world networks. They also managed to identify the key players who play a big role in the team's performance [9]. Javier López Pena and Hugo Touchette also used information about passes to create networks and to describe football strategy [10], just like Raffaele Trequattrini et al. did, who analyzed an UEFA Champions League match [11]. They visualized the line-up of the teams and determined the importance of the players. Pablo Medina et al. used social network analysis to determine match results. They not only developed and analyzed networks but also studied their relevance to the results [12]. Paolo Cintia et al. measured team performance with networks

as well. They not only used passes to determine the edges between players, but many other actions as well, like tackles, fouls, clearances, etc. They observed that the network indicators correlate with the success of the teams, and then used it to predict the outcome of the matches [13]. Filipe Manuel Clemente et al. suggested that defenders and midfielders have the most connectivity in the team [14]. Also Clemente et al. got similar results when they analyzed the Switzerland national football team in the 2014 FIFA World Cup [15]. E. Arriaza-Ardiles et al. used graph theory and complex networks to understand the play structure of the team. They used clustering and centrality metrics to describe the offensive play [16]. Opposed to the listed works that are all analyzing in-game relationships, our intention is to create a graph theoretical model on the club level in order to forecast player transfers between clubs.

3 Network research approach and findings

In this section we present the graphs we created for modeling the relationship among players, coaches and clubs, respectively.

3.1 Players' graph

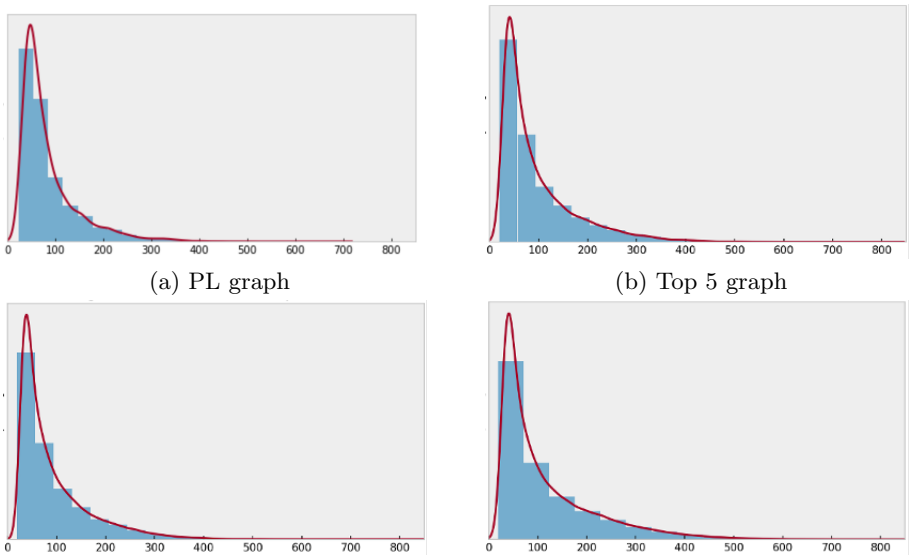
Table 1: The basic metrics of the player graphs.

Metrics	Premier League	Top 5	European	World
Nodes	6,407	22,509	242,827	92,969
Edges	271,083	1,070,595	1,964,482	5,299,404
Average degree	84.62	95.13	91.74	114.00
Max degree	486	570	583	801
Min degree	23	20	19	19

We created four different graphs with more and more leagues involved. In this way, we could examine how the network metrics have changed with the increasing number of nodes. We started the analysis with one of the most competitive leagues in the world, the English Premier League. For this graph, we used the information available on the official website of the league [18]. We studied the teams from the 1992/93 season to the 2020/21 one, so this part of our research covered the entire history of the Premier League. We created three more graphs: we named the first one as Top 5 as it includes the best 5 European football leagues; the second graph, named as European, contains 24 first division leagues from Europe; the third graph, named as World, contains 58 leagues from all over the world. Both first, second, third and in some cases even fourth divisions are included. These latter 3 graphs are based on FIFA computer game data, which is published every year with updated squads [19]. We used information about players since the 2007 release. We created familiarity graphs, in which

the players became the nodes and two players considered to be adjacent if they were teammates for at least a season. After defining the edge list, the developed graphs have the following metrics, summarized in Table 1.

First, we examined the degree distribution of the graphs, which can be decisive in answering our question. Small-world networks have power function distribution, which means that most players have only a few connections, but some players have a lot. Indeed, a power function is followed by the distribution of the degrees in the four graphs we created as Figure 1 shows.



(a) PL graph (b) Top 5 graph

(c) European graph (d) World graph

Fig. 1: Degree distribution of the graphs follow power law

The average shortest distance and the diameter are both relatively small in small-world networks. Quite precisely, these numbers are roughly equal to the logarithm of the number of nodes. The average distance is 2.63, while the diameter is 5 of the Premier League graph. As 6,407 base 10 logarithm is 3.8, the average distance is even smaller than what is required for a small-world graph. This means that in the Premier League, the distance between players is very short, averaging less than 3 players and even between the farthest players, the distance is only 5. As shown in Table 2, these metrics are also small for the other graphs and increase at a slower rate than the logarithm of the number of vertices. It is utterly amazing that in the examined leagues, which cover the entire world, players are just over three steps apart on average. What is more, it only takes a maximum of six steps to connect any two players, even if they play in the farthest parts of the world.

The third characteristic which we examined is the clustering coefficient metric. It seems logical that this should be comparatively big, since teammates form a complete subgraph within the graph. The whole network is made up of such subgraphs connected by players who have turned up in several teams. When examining the clustering coefficient, we are curious about the extent to which there are triangles in the graph, so this property is also commonly referred to as triadic closure [17]. We will use these two words as synonyms hereafter. This metric for the Premier League players' graph is 0.41 which is closer to the upper limit of the standard value described earlier. Table 2 suggests that with the increasing number of vertices the clustering coefficient becomes smaller. But as it is relatively still far from zero, the last condition is also met, so in the world of football, players make small-world networks.

Table 2: The distances and clustering coefficient of the graphs.

Metrics	Premier League	Top 5	European	World
Nodes	6,407	22,509	242,827	92,969
Log of nodes	3.81	4.35	4.63	4.97
Average distance	2.63	3.00	3.17	3.24
Diameter	5	6	6	6
Clustering coeff.	0.41	0.31	0.31	0.25

As we can see, the world of the football players is small, no matter how many championships we take into account, since all the properties that apply to graphs with small-world properties are fulfilled in them.

3.2 Managers' graph

We did similar studies on the coaches of the Premier League. The used data is also collected since the 1992 season, available on the official Premier League website. Since two managers cannot lead the same team at once, we have defined the acquaintances in this network as two coaches knowing each other if they have managed the same team during their careers. Thus, we obtained a graph of 236 vertices and 1,843 edges, where the average degree is 15.62. This graph can be seen in Figure 2.

First, we examined the degree distribution, which shows a power function distribution in this case too. It is well observed that the degrees are much lower than among the players, which is understandable since there were far fewer coaches than active players. The smallest degree is 1 and the largest is 81, which is Sam Allerdyce's degree, who managed 7 different teams and holds the records with the most clubs coached in the history of Premier League.

Next, we examined the shortest path and diameter. Since the graph consisted of several components, we had to delete the smallest components to determine the average shortest path, as it can be determined only on one component. Two

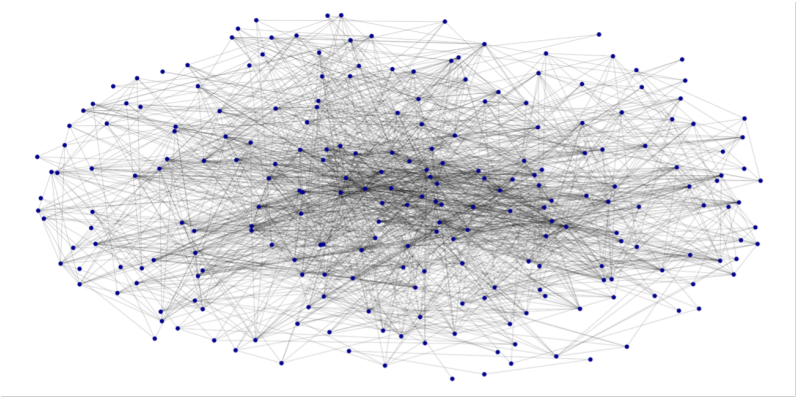


Fig. 2: The layout of the managers' graph.

components had to be removed with 3 vertices each. The average shortest path of the resulting graph is less than it is among players, 2.53, and the diameter is only 5. The former number is less than the logarithm of 1,837, which is 3.2, so the network also meets this criterion.

Finally, we examined the clustering coefficient metric, which also meets the requirements, as it is 0.55, that is particularly high. This means that a manager's two neighbors probably coached the same club.

The results show that not only do players make up a small world in the Premier League but coaches do as well.

3.3 Teams' graph

We have seen both players and coaches form a small world. We could also see from the example of the players that this is true not only to one league but to the whole world. But what about the teams? They are also in constant contact with each other, as they have a chance to purchase or borrow players from each other twice a year. For the study of the graph of player transfers, we used data available on Transfermarkt.com from 2007 to 2020, which includes all transfers between teams. Transfermarkt is the most well-known site that deals specifically with player evaluation.

In this case, the vertices of the graph are not players but teams, and two teams are adjacent if there was any transfer between them during the examined period. Thus, we obtained a network of 7,612 vertices with 49,805 edges. The average degree is 13.09, the highest is 330, and the lowest is 1.

The degree distribution follows a power curve. It is definitely true that most teams have only a low degree and there are only a few teams that have a high one. However, these teams are also extremely far from the average. The average shortest path is 3.6, which is less than 3.88, the logarithm of the number of nodes. Diameter is 7, which means that for some clubs it can take up to 7 steps to get to another, but on average 3-4 steps are enough.

The triadic closure is the smallest among the networks examined so far with the value of 0.18, but it is still sufficient to be a small-world network. It shows well that since now the nodes do not have teammates who are also adjacent to each other this value decreases. The relatively high value is likely given by the leagues, as within those, transfers are more common than usual.

In summary, this network is also a small-world network as the necessary conditions have been met. It also seems to be true in the world of football clubs that the world is small.

4 Predicting player transfers

To build the transfers' graph, we used historical information about the transfers. With this information advantageous correlation could be found, that can help the prediction of future transfers. Therefore we built a model for this purpose. We limited the data set only to transfers, as loan transactions are different in nature. Thus, starting from 2007, we had a total of 17 431 transfers which could be used to train the model. The goal of this model is to predict who will be transferred into a selected league, as the data is too spare to predict the same for clubs. During modeling we focused on the Top 5 leagues, but the same method can be applied to any other leagues as well.

The model was created based on the following features: the transferred player's age, market value, nationality, position, the league he was playing before the transfer, the FIFA computer game's player statistics, and information extracted from the graphs.

As the players' market value has been increasing steadily since 2007, we used a correctional scaling by dividing every value with the given season's 75th percentile value. In 2019, a player with the same parameters costed four times as much as in 2007, which would have led to a deterioration in the performance of the model. First we decided to do the correction with the mean of the seasons, but just a few high values can easily distort it, so it would not give us a clear picture about the market of the players. Then, we tried the median of the players' market value and noticed the distribution of it is unequal. Most of the players have a lower market value compared to the high profile players who play in the top leagues. Therefore, we decided to use the 75th percentile of the market values that is a decent indicator for a footballer playing in on of the top leagues. The result of the scaling is presented in Figure 3.

The FIFA statistics include various abilities that rank players on a scale of 1 to 100. The higher this number, the better the given ability of the player is. There are abilities like dribbling, finishing, short passes, prefered foot, international reputation and many more. These abilities can be grouped into six main abilities, that FIFA also uses for the online game modes. Using these grouped main abilities also, like Defending, Passing, Shooting, Pace, Dribbling, and Physique slightly improved our model. We also used features extracted from the graphs to get information not only about the transfers, but also about their relationships. We used the players' graph to get information about the players' acquaintance.

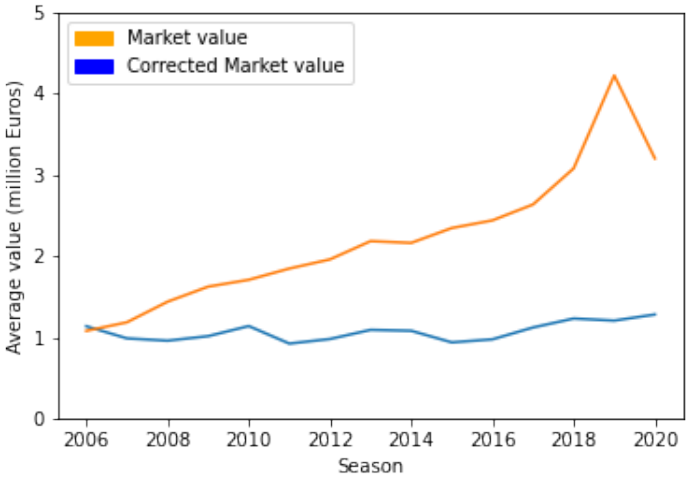


Fig. 3: Market value correctional scaling.

For this purpose, we built graphs for every year to avoid the scenario when a player gets an acquaintance with a player with whom he played later. For example, if a transfer happened in 2015, we used the graph that has the data from 2007 to 2014 to avoid a player's score being influenced by a later teammate . We extracted the degree number divided by the number of seasons the player has been playing; the eigenvector, which is a centrality metric to measure the node's importance in the graph; and the number of links each player has with the top 5 leagues. The number of years between two transfers of the same player has also been calculated, in order to get how many years the given player spent in a particular team.

We encoded the categorical variables with CatBoost Encoder to avoid creating hundreds of columns, which would have been caused by the Country and the League variables. To fill the empty values Iterative Imputer has been used, which uses modeling to predict the missing values in a column, as it resulted in the best performance. After standardizing the dataset and testing different models, XGBoost was found to be the most accurate one. We used cross validation to train the model, and grid search to find the best parameters. We chose the set of hyperparameters with the highest recall metric, that has a high accuracy as well, as we wanted to predict as many actual transfers as possible. The best set of parameters are 5000 esimators, 0.7 subsample, 6 maximum depth, and 0.1 learning rate.

As the data set contains many leagues, the number of transfers into the selected league is dwarfed relative to the entire data set. This degrades the quality of the model, so we used the SMOTE (Synthetic Minority Oversampling Technique) to reproduce the favorable data if there were too few positive cases in

the selected league and it led to performance degradation. The SMOTE technique uses oversampling with the k-nearest algorithm to smooth the data set. After trying several models, oversampling with SVM (Support Vector Machine) worked the best. Using this method the recall metric increased greatly, but creating too many records resulted in over-fitting, so only a few percent has been added to the positive cases.

After finding the best parameters of the model, we used it on the FIFA 21 computer game's data, that includes almost 20,000 players. We ran the model for the Top 5 European leagues, and compared the most important features according to the XGBoost model. These are listed in Table 3.

Table 3: The most important features of the Top 5 leagues.

Premier League	La Liga	Ligue 1	Serie A	Bundesliga
League	Country	League	Serie A	League
MarketValue	League	Ligue 1	League	Market Value
Eigenvector	Market Value	Country	Country	Age
Country	Foot	IntReputation	MarketValue	Reactions
Age	Eigenvector	Age	Degree	IntReputation
IntReputation	GK kicking	Premier League	Foot	Country
Foot	Defending	GK diving	La Liga	Short passing
GK kicking	Passing	Eigenvector	Eigenvector	GK diving
Physical	IntReputation	Market Value	Age	Defending
Shooting	Age	La Liga	Sliding Tackle	Dribbling

In some leagues the most meaningful variables are the graph variables, most notably the number of links with the league for which the transfer prediction is made, like the Ligue 1 and Serie A. These leagues are particularly characterized by transfers predominantly within the league. We denote each of these features by the name of the respective league. The eigenvector also turns up among the most vital features as one of the graph variables. The country, market value, current league, and age also play great part in the prediction. The international reputation feature is only missing from the Serie A's top 10, while the preferred foot is missing from the Ligue 1 and Bundesliga. The main attributes from FIFA also appear. In the Premier League the most essential ones are the shooting and the physical attribute, while in the La Liga the defending and passing.

The accuracy of the models are 93-95 depending on the selected league. The F1 score differs among the leagues. The highest one is the Ligue 1 with 86, and the lowest one is for the La Liga with 46. There are numerous false positive predictions, as in the result there are 1500-3000 players predicted to transfer into the selected league. This is mainly because we do not take into account some vital features, like performance features or when the players' contracts expire. Involving these features could further improve our model. Also worth mentioning that due to the loss of revenue caused by the coronavirus pandemic, clubs spend

much less this season, and purchase much fewer players, then in those years on which the model was trained.

The La Liga and the Bundesliga have the fewest number of transfers in the training set, and the model works the worst for these two leagues. However, for the Premier League and for the Serie A it predicted 25 out of 47, and 24 out of 64 transfers well, respectively, that have recently happened. For the Ligue 1, 15 of 38 transfers were predicted correctly. Running the models without the graph metrics resulted in worse result, except for the Bundesliga. Without those features the model produces significantly more false positives and the accuracy is 2 percent higher, when those metrics are included.

Table 4 shows some of the transfers predicted correctly by the model for the Top 5 leagues.

Table 4: Some of the correctly predicted transfers

Premier League	La Liga	Serie A	Bundesliga	Ligue 1
B. White	R. De Paul	M. Darmian	J. Gvardiol	J. Lucas
E. Buendía	M. Depay	F. Tomori	K. Boateng	D. Da Silva
A. Townsend	S. Agüero	M. Maignan	R. Hack	L. Balerdi
J. Grealish	D. Alaba	E. Hysaj	M. Uth	L. Badé
B. Soumaré	E. Lamela	R. Patrício	G. Haraguchi	A. Bassi

Overall, the model works well for the purpose of showing who could be considered if a team wants to buy a player for the above mentioned three leagues based on the results so far. The information extracted from the graphs improved the performance of the model, as the accuracy is increased by reducing the false positive predictions.

5 Conclusions

In our research, we have demonstrated that the networks that emerge in the world of football, just like many naturally occurring networks, are small-world networks. Both players, managers, and teams with their transfers form small worlds. Using the information of these graphs, information of historical transfers, and information of FIFA computer games we predicted future player transfers. The model can be used to predict movements into the selected league, but to predict the same for clubs the available data is too sparse. The information extracted from the graphs have a great importance in the predictive models and improve those performances. The most vital features of the transfers were also presented of the top 5 leagues.

Acknowledgment

Project no. 128233 has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the FK_18 funding scheme.

References

1. Newman, M., Barabasi, A.-L., Watts, D. J. (Eds.): The Structure and dynamics of networks. Princeton University Press (2006)
2. Milgram, S.: Psychology Today. 1, 60. (1967)
3. Amaral, L. A. N, Scala, A. , Barthelemy, M., Stanley, H. E.: Classes of small-world network (2000)
4. Barabasi, A.-L. : Linked: How Everything is Connected to Everything Else and What It Means for Business, Science and Everyday Life. Plume Books, New York (2003)
5. Erdos, P., Renyi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad., vol. 5., pp. 17–61 (1960)
6. Amaral, L. A. N, Scala, A. , Barthelemy, M., Stanley, H. E. : Classes of small-world network (2000)
7. Granovetter MS: The strength of weak ties. In: Am J Sociol 78., pp. 1360–1380. (1973)
8. Javier, M. H., Piet, V. M.: Classification of graph metrics (2011)
9. Yamamoto Y, Yokoyama K: Common and Unique Network Dynamics in Football Games. In: PLoS ONE 6(12). (2011)
10. Pena, J. L., Touchette, H..: A network theory analysis of football strategies, in Sports Physics. Proc. 2012 Euromech Physics of Sports Conference, ed C. Clanet., pp. 517–528. (2012)
11. Trequattrini, R., Lombardi, R., Battista, M.: Network analysis and football team performance: a first application (2015)
12. Medina, P., Carrasco, S., Rogan, J. , Montes, F., Meisel, J. D., Lemoine, P., Penas, C. L., Valdivia, J. A.: Is a social network approach relevant to football results? In: Chaos, Solitons and Fractals, vol. 142 (2021)
13. Cintia, P., Rinzivillo, S., Pappalardo, L.: A network-based approach to evaluate the performance of football teams. Workshop on Machine Learning and Data Mining for Sports Analytics, pp. 46–54., Porto, Portugal (2015)
14. Clemente, F. M., Couceiro, M. S., Martins, F. M. L., Mendes, R. S.: Using network metrics to investigate football team players' connections: A pilot study, Motriz, Rio Claro, vol. 20 n.3, pp. 262–271. (2014)
15. Clemente, F. M., Martins F. M. L., Kalamaras, D., Oliveira, J., Oliveira, P., Mendes, R. S.: The social network of Switzerland football team on FIFA World Cup 2014, In Acta Kinesiologica 9, pp. 25–30. (2015)
16. Arriaza-Ardilesa, E., Martín-González, J.M, Zunigac, M. D., Sánchez-Floresd, J., de Saae, Y., García-Mansoe, J.M.: Applying graphs and complex networks to football metric interpretation. In: Human Movement Science 57, pp. 236–243. (2018)
17. David, E., Jon, K.: Networks, Crowds, and Markets: Reasoning about a Highly Connected World. In: Cambridge University Press, pp. 48–50. (2010)
18. Premier League official website, https://www.premierleague.com/players. Last accessed 20 Aug 2021
19. FIFA players, https://www.premierleague.com/players. Last accessed 20 Aug 2021