Forecasting Soccer Market Tendencies Using Link Prediction

Lucas G. S. Félix[1], Carlos M. Barbosa[1], Iago A. Carvalho[2], Vinícius da F. Vieira[1], and Carolina Ribeiro Xavier[1(✉)]

[1] Department of Computer Science, Universidade Federal de São João del Rei, São João del Rei, MG, Brazil
lucasgsfelix@gmail.com, cmagnobarbosa@gmail.com, {carolinaxavier,vinicius}@ufsj.edu.br

[2] Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
iagoac@dcc.ufmg.br

Abstract. Soccer is the most popular sport in the world and due its popularity, soccer moves billions of euros over the years, in most diverse forms, such as marketing, merchandising, TV quotas and players transfers. As example, in the 2016/2017 season, only England has moved about 1.3 billion of euros only in players transfers. In this work, it is performed a study of the transfer market of players and, to do so, players transfer data were gathered from the website Transfermarkt and transfers were modeled as a graph. In order to perform this study, different Complex Networks techniques were applied, such as Link Prediction, Community Detection and Centrality Analysis. Link Prediction it was possible to conceive a network of a forecast market for the 2022 World Cup using temporal data. Through the generated network it was possible to notice changes of the market and the rise of Asian countries, given a notion of how the market is going to behave in the next years.

Keywords: Complex networks · Data Mining · Sports analytics

1 Introduction

Soccer is the most popular sport in the world [14,21], and due to this enormous popularity it attracts millions of practitioners and billions of spectators for large events as FIFA World Cup, continental and intercontinental championships. Thus, soccer generates a high financial flow produced by ticket sells, TV contracts, marketing and merchandising. A clear illustration of the huge amount of money in this market can be observed in England where, in a single season 2016/2017, moved almost 1.3 billions of euros [4].

In between the various financial compensation that a team can explore, players transference between clubs stands out. This process makes possible to sell and buy athletes, consequently giving a team the opportunity to get a certain

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12249, pp. 663–675, 2020.
https://doi.org/10.1007/978-3-030-58799-4_48

quantity of money in a short time. Those transactions are cataloged and categorized in the website Transfermarket[1], which contains a considerable quantity of transfer data, classifying buy and sell transaction, as well as loan transfers among teams.

Due to these facts, the goal of this work is to study and to analyze the properties of the network of soccer players transfers between countries present in the World Cup. We choose the World Cup because it is the second biggest sport event in a global scale [1], besides that, the World Cup is one of the biggest influence in the player's transfer market [6]. To perform this evaluation, we applied Complex Networks techniques.

From the perspective of Complex Networks, studied in the field of study as Network Science [23], data is modeled as graphs in order to be analyzed. Our methodology applies Link Prediction, Community Detection, and Centrality analysis. By the application of Link Prediction we could forecast the chances of a vertex A to connect to a vertex B and then investigate the community structure of the network by Community Detection, to view sub-markets and the countries that are present in these clusters. Using Centrality measures we could observe which nodes are most important for the network and consequently which countries are most important for the football market.

Through our methodology, we could observe the tendency of soccer players market concerning other countries that dominate this market. However, it is possible to perceive changes in sub-markets, evidencing the rise of Asian countries driven mostly by China.

2 Related Work

Soccer has been a source of study over the year in many areas as Physical Education, where researchers evaluate the biomechanics of the sport [12], methods for sportive training and efficiency [10], sports medicine and nutrition [19]. Another research area analyzes soccer by a sociological or anthropological vision. Those works make qualitative analysis over topics as, international migration of work forces [16], cultural transformation induced by soccer [22], or even social questions, as racism [11], and violence [24].

Few works in the literature seek to understand, analyze and evaluate the soccer players transactions market. In those works, we highlight the works of [21] and [9]. The work of Huerta et al. [21] evaluated the players transfer applying temporal behavioral statistical analysis in the English Premier League. The work of Frick et al. [9] analyzed the European players transactions empirically, evaluating many features present in soccer athletes as salary and career time.

In this field of study, other works with more computationally focused approaches can be fond [5,6,25], applying network metrics to evaluate the players transfer market. The work of Xiao et al. [25] pointed out only a few properties of the transactions market, using data from 2011 to 2015. The general aim of the

[1] www.transfermarkt.com.

work was not to analyze the market, but to evaluate the success of a team given its transfers. In previous works, we used Complex Networks and Data Analysis techniques to analyze the teams presents in the 2018 FIFA World Cup [5] and proposed a method to join centrality metrics [6]. Besides that, we explored sociological vision of soccer, showing theories previously raised by other authors in a quantitative way [6].

The methodology proposed in this work differs from those found in other works because it aims at the evaluation of future consumers market, considering Link Prediction techniques.

3 Methodology

The objective of this work is to evaluate the future soccer market, in specific the countries that can be be present in the next FIFA World Cup. To make those analysis Complex Networks and Data Mining techniques are applied. Our approach is composed by four steps: (*i*) Data Gather and Treatment, which consists in the development of an automatic method to capture data in websites; (*ii*) Network Modeling, in which we model the network transfers between countries as graphs; (*iii*) Link Prediction, when a future network is generated to analyze and preview countries and connections between these countries (links) that can be in the 2022 FIFA World Cup; and (*iv*) Network Analysis, which applies Complex Networks measures to evaluate the network and validate our approach.

3.1 Data Collect and Treatment

To perform the proposed study, we developed a crawler for the website Transfermarkt (See footnote 1). This website consists of a vast database with information related to soccer: statistics, championships standings and, specifically concerning to this work, data about players transfers. To process the data gathering, we developed a crawler and a parser. These two tools aims to clean and pre-process the data. At the end of the process, the data is in a semi-structured format.

The data gathered ranges from 1962 to 2017 and we considered the 250 most valuable transactions per year per tactical position (goalkeeper, defenders, left-back, right-back, midfielder, forwards). It is important to mention that, for some years, the number of transfers in the website does not reach 250. The final data collection consists of about 28 thousand players transfers.

From Fig. 1, it is possible to observe the number of transfer gathered by position and period. Also, it is possible to notice that the number of transfers is increased year by year. This aspect reflects the expansions of the player's transactions market as shown in [6].

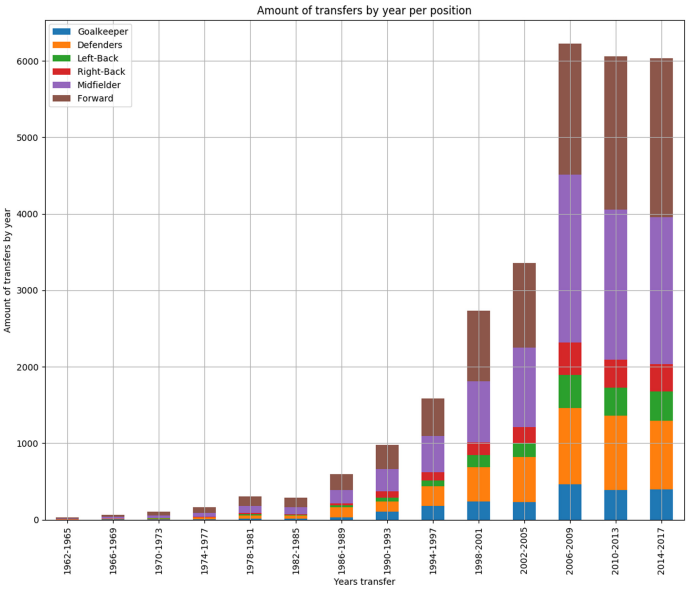

Fig. 1. Number of transfer gathered by position and period

3.2 Network Modeling

After the data gathering, the networks were modeled in a way that each country is a vertex and a directed edge exists between two nodes if a transference between these countries occurs. In the proposed model, an outgoing edge represents a sell transaction, and an incoming edge represents a buy transfer. The weight of the links is defined as the sum of the number of transfers occurred between one country and another. Previous works [5,6] also modeled the edge weight as the sum of the value of transactions between two countries, however, this approach is not considered in this work, due to the temporal feature of our modeling. Thus, it was possible to model the network avoiding distortions caused by the variations of athletes prices over the years.

To model the networks, the period of time considered was divided in intervals of 4 years, and then, the transfers were joined aiming to summarize the transactions realized between a World Cup and another. This approach was considered for the network because the years of World Cup tend to be renewal years for the teams. In this sense, it is considered that the period that a defines the 2018 FIFA World Cup, for instance, ranges from 2014 to 2017.

3.3 Link Prediction

Link Prediction is a Complex Networks technique that assists in understanding many dynamic elements, as Social Networks, and can be used as a tool to analyze the evolution of these networks [13]. Thus, to preview the tendencies of

a link between the nodes of the modeled networks, Jaccard Similarity Measure was applied. This technique assumes high values to pairs of nodes that share a great proportion of common neighbors concerning the total number of neighbors [15]. In other words, the larger the quantity of common neighbors between two vertexes, the larger is the probability of these two nodes share an edge. Mathematically, the Jaccard measure between two sets A and B is defined by Eq. 1.

$$J(A, B) = \frac{|a \bigcap b|}{|a \bigcup b|}. \tag{1}$$

3.4 Network Analysis

To perform analysis over the network, we applied different techniques, as community detection and centrality measures. The Community Detection algorithm aims at the identification of cohesive groups in a network. In this work three different algorithms to find partitions in the graph were considered: (*i*) Multilevel [2]; (*ii*) Eigenvector [17]; (*iii*) Fastgreedy [3]. All these methods has, as basic principle, the optimization of modularity [18], which evaluates how good is a partition of a graph. In this way, the larger the modularity value, the better is network partition, and more consistent are the formed communities. Using these different approaches, it is possible to find different community partitions. A brief description of the methods for community detection is preseted as follows:

- **Multilevel:** This algorithm works in an agglomerative way, and performs two steps to divide the graph. In the first step, it is considered that every node is a community which swap to more convenient communities, until there is no modularity gain. In the second phase, it builds a new network considering each community as a node and perfoms community joins, similarly to the first step [2].
- **Eigenvector:** This divisive algorithm works based in a Laplacian Matrix (Modularity Matrix) to make the graph partition and identify the communities. This way, the algorithm manipulates the Modularity Matrix, and calculates the eigenvector associated to the largest eigenvalue to perform a bisection. This process is repeated until there is no modularity gain with a new division [17].
- **Fastgreedy:** This algorithm is also agglomerative and considers, initially, the nodes as unary communities and then perform convenient joins. The algorithm stops when there is no communities joins that cause a positive modularity gain [3].

Centrality measures were also explored in order to investigate the relevance of the nodes of the graph. In the context of this work the analysis corresponds to the importance of a country to the player's transfer market. Three different approaches were investigated in this study: (*i*) Closeness [8], (*ii*) Betwenness [7], (*iii*) Pagerank [20], which can be defined as follows:

- **Closenness:** Evaluates the geodesic distance between a node and the others reachable nodes considering the analyzed vertex [8].
- **Betweenness:** Evaluates the fraction of shortest paths between all pair of nodes. The most central vertex will be the node with the most significant fraction of shortest paths passing through it [7].
- **Pagerank:** Evaluates the importance of a node by assessing the importance of its neighbours [20].

For Betweenness and Closenness centrality, the weight of the edges were reversed. Such action was needed because these methods are based on shortest paths and using a traditional edge weight approach the precision over the main countries was altered. Thus, using the approach where the edge weight is $\frac{1}{w_i j}$, where $w_i j$ is the weight between the nodes i and j, the relevance of a node is more clearly revealed.

The remaining of the analysis was made evaluating the properties of the network as density, diameter, clustering coefficient, reciprocity, and assortativity degree. These measures will be presented as they appear in Sect. 4.

4 Results

In this section the results obtained by the application of the proposed methodology are presented. First, we described how the network was developed. Next, we report and analyze the obtained experimental results by Complex Network techniques.

4.1 Network Development

To assembly the graph used in this work, a process that could guarantee the reliability of the modeled network was necessary. To make so, several preparation and preprocessing steps were performed. These steps were fundamental due to the long term characteristics of the data, ranging over decades, which could bring some inconsistencies if not properly treated. Some countries can have a huge number of transactions in one year, and a smaller number in another, showing the dynamicity of the soccer market. Thus, to generate the 2022 player's market network, it was necessary to assure that countries present in these graphs are really actives in the soccer market. First, to model the network in a way that the outliers were not present in the graph, the vertex that has just a few links over all networks were withdrawn of the graph, i.e., countries that did not maintain transactions over the years, only having exceptional transfers, was removed from the graph in a temporal way.

The proposed methodology considers as reference a network modeled in the subsequent years to 2002. Thus, it was possible to guarantee that a vertex present in the network of 2022, has as nodes, countries that are present in transference in 2002, 2006, 2010, 2014 and 2018. As previously shown in Sect. 3, the proposed methodology modeled the networks in periods of 4 years.

2002 was selected as the reference year we used the maximization of the number of vertex and networks modeled as criterion. From Fig. 2, it is possible to see that starting from the year 2002, we could achieved a reasonable number of networks and nodes, being a good trade-off between the number of networks and the number of nodes.

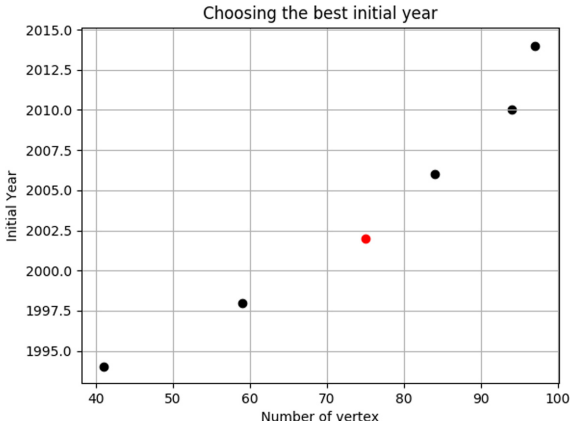

Fig. 2. Initial number of vertex by year, considering a intersection between the years

By the evaluation of Fig. 2, it can be noticed that the number of nodes is increasing along the years. This happens due to the soccer globalization that gave the opportunity to countries that does not have much tradition in soccer to sell and buy players [6]. As an example, there is China that in recent years have become a big negotiator in the soccer market.

After identifying the nodes that will form the network of 2022, it was defined that the graph which could serve as a base is the one from 2018. The 2018 graph was selected because the soccer market have small changes over the years, thus, to define this changes from 2018 to 2022 it was applied a Link Prediction technique, considering the Jaccard Coefficient. As the Jaccard Metric is calculated for every pairs of vertices in the graph, it was necessary to limit the minimum value to link two nodes, as a threshold value. Thus, it is was possible to avoid connections between vertices that have a small and we avoid to create a very dense graph. So, to define a threshold to Jaccard Metric it was used an approach based on the elbow method. The elbow method is used to find the ideal number of clusters in an unsupervised machine learning technique. However, we have adapted the metric, using it to define the point that gives the better accuracy for the forecast made by Link Prediction.

From Fig. 3 it is possible to observe the accuracy rate of the networks given a threshold. To configure an accuracy to the modeled networks it was considered that every network has an ground truth, except the 2018 network. To illustrate that, consider the that 2002 network has as ground truth the 2006 graph, and

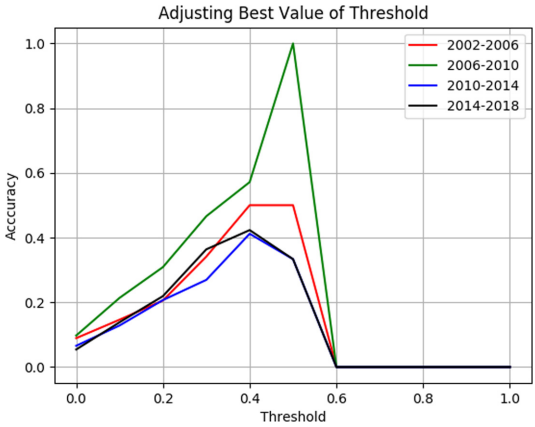

Fig. 3. Adjusting the best value of threshold

the applied Link Prediction technique helps to get closer to all right answers. This way, it was considered that the bigger is the accuracy, the better is the prediction.

To define the ideal threshold t_i, we considered an approach where t_i value is the mean between all k-threshold t_k found before, given by Eq. 2, where n is the number of networks. This way, an edge is added to the graph if Jaccard Coefficient is bigger than 0.425. In the end, the 2022 final network have 75 nodes and 1117 edges.

$$\frac{\sum(t_k)}{n} \tag{2}$$

$$\frac{(\sum(X_i))/n + \sum(Y_j)/m}{2} \tag{3}$$

To define the weights of the edges in the graph, Eq. 3 was considered, where X and Y are the linked countries, i represents sell transactions, j represents buy transactions, n is the number of sells of country X and m is the number of buys of country Y.

4.2 Centrality Analysis

Network centrality analysis gives the importance of a vertex in a graph, being an fundamental tool in the study context since through it is possible to identify the main countries in the player's transactions market. In this work, three different metrics were evaluated, as described in Sect. 3: Closeness, Betweenness, and Pagerank. After the application of the metrics three different sorted lists were obtained. From Table 1 it is possible to observe the 10 first positions. Also, from Fig. 4 it is possible to see 2022 network with the main vertex identified by the Betweenness centrality.

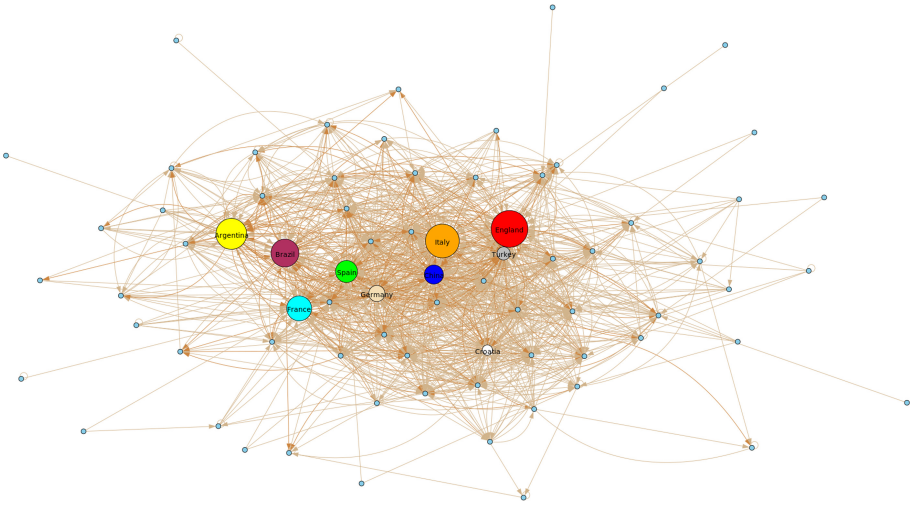

Fig. 4. Main countries identified by the Betweenness centrality.

By analyzing Tables 1 and 2, it is possible to see small, but significant changes in the rankings. As example we can see the rise of countries like Turkey and Romania. Besides, there is the presence of China between the main nodes in all three metrics, due to the expansion of his market as a potential buyer, as Saudi Arabia.

4.3 Community Analysis

To identify sub-markets that perform a huge volume of transference Community Detection was considered. Thus cohesive groups in the network, communities,

Table 1. Top-10 countries for each metric considered in the 2022 network.

Position	Closeness	Betweenness	PageRank
1	England	England	England
2	Italy	Italy	Italy
3	Spain	Argentina	Germany
4	Germany	Brazil	Spain
5	Turkey	France	Turkey
6	France	Spain	France
7	Wales	China	China
8	Romania	Germany	Brazil
9	Argentina	Turkey	Saudi Arabia
10	China	Croatia	Romania

Table 2. Top-10 countries for each metric considered in the 2018 network.

Position	Closeness	Betweenness	PageRank
1	England	Italy	England
2	Italy	England	China
3	Spain	Argentina	Italy
4	Germany	Brazil	Germany
5	Turkey	Germany	Spain
6	France	France	Turkey
7	Wales	Spain	France
8	China	Portugal	Saudi Arabia
9	Argentina	Croatia	Belgium
10	Portugal	Turkey	Brazil

could be identified. The nodes in this subgroups have as main characteristic the fact that their nodes have a larger number of inside connections (intracluster) that outside links (extracluster).

This analysis is motivated by the fact that the soccer players transference market is an embracing market, where there are free exchanges of players, and some consumption patterns can observed. Those consumption patterns are given by some countries that are used as a source of talents do rich European markets. This way, to identify this sub-markets, three different methods of community detection were considered, as presented in Sect. 3: Multilevel, Eigenvector and Fastgreedy. From Table 3, it is possible to see the modularity and number of communities results obtained by each of the algorithms.

Considering that the applied methods to identify communities aim to maximize modularity and evaluating Table 3, it is possible to notice tha Multilevel and Eigenvector methods offer the same modularity value and the same number of communities. Due to its lower computational cost, Multilevel method was selected for further analysis.

By analyzing Table 4, it is possible to see the generated communities by the Multilevel method with more than one node. Communities with only one node were removed from this table because these communities do not bring any additional information to our analysis. Table 4 highlights the countries considered of rich leagues [25], except for Spain which was identified in an unary community and removed.

In previous works [5,6], it has been shown that soccer player's market is dominated by European countries, where each one dominates its own niche of trades, being present in distinct communities with countries of less relevance.

Besides that, as this is predicted network, it is possible to notice tendencies in the market, where the rich countries are searching for better markets at the moment. However, the changes in the market is bigger in the Asian countries, where China is becoming a country as powerful as European countries.

Table 3. Modularity found by each algorithm.

Algorithm	Modularity	# of communities
Eigenvector	0.53	35
Fastgreedy	0.32	29
Multilevel	0.53	35

Table 4. Communities with more than one vertex formed by the Multilevel method.

Communities	Countries
1	Wales, **England**, Scotland
2	**Italy**, Moldova, Croatia, Australia, Slovenia
3	Qatar, Belgium, Greece, **France**, Cyprus, Israel, Monaco
4	Netherlands, Sweden, Norway, Denmark, Serbia, Russia, Poland, Slovakia, Bosnia-Herzegovina, Ukraine, Hungary,Georgia
5	Austria, Japan, Switzerland, **Germany**
6	Romania, Bulgaria
7	Czech Republic, Turkey
8	Saudi Arabia, Tunisia, Egypt
9	Mexico, Chile, Brazil, Paraguay, Portugal, Uruguay, Ecuador United Arab Emirates, United States, Colombia, Argentina
10	South Korea, China

5 Conclusion

In this work we proposed a study about the tendencies of the soccer players transfer market. To make so different complex networks strategies were applied, where we identify some of these changes in soccer world scenario for the next World Cup. Among these possible transformations, a more clear protagonism of Asian countries in the transfer market could be observed. However, there will be still a visible dominance of Europe in this market.

Besides that, it can be also noticed a large Jaccard Coefficient connecting rich countries, given that these countries have sufficient money to perform transactions with any country in the world. Also, our analysis identifies that there is a common node change in the communities, showing a high variation of the market.

As future works, we plan to use other techniques of Link Prediction and heuristics that with prior knowledge could help us to increase accuracy.

References

1. Baade, R.A., Matheson, V.A.: The quest for the cup: assessing the economic impact of the world cup. Reg. Stud. **38**(4), 343–354 (2004)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. **2008**(10), P10008 (2008)
3. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**, 066111 (2004)
4. Deloitte, L.L.P.: Annual review of football finance, June 2016
5. Felix, L., Barbosa, C., Carvalho, I., Vieira, V., Xavier, C.: Uma análise das seleçóes da copa utilizando uma rede de transferências de jogadores entre países. In: CSBC 2018 - VII BraSNAM, July 2018
6. Felix, L., Barbosa, C., Vieira, V., Xavier, C.: Análise do impacto das copas do mundo no mercado de transações de jogadores de futebol e da globalização do futebol utilizando técnicas de redes complexas. In: ENIAC 2018 - VII KdMIle, October 2018
7. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry 35–41 (1977)
8. Freeman, L.C.: Centrality in social networks conceptual clarification. Soc. Netw. **1**(3), 215–239 (1978–1979)
9. Frick, B.: The football players' labor market: empirical evidence from the major European leagues. Scott. J. Polit. Econ. **54**(3), 422–446 (2007)
10. González-Badillo, J.J., Pareja-Blanco, F., Rodríguez-Rosell, D., Abad-Herencia, J.L., del Ojo-López, J.J., Sánchez-Medina, L.: Effects of velocity-based resistance training on young soccer players of different ages. J. Strength Conditioning Res. **29**(5), 1329–1338 (2015)
11. Jarvie, G.: Sport, racism and British society: a sociological study of England's elite male Afro/Caribbean soccer and rugby union players. In: Sport, racism and ethnicity, pp. 79–102. Routledge (2003)
12. Lees, A., Asai, T., Andersen, T.B., Nunome, H., Sterzing, T.: The biomechanics of kicking in soccer: a review. J. Sports Sci. **28**(8), 805–817 (2010)
13. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. **58**(7), 1019–1031 (2007)
14. Liebig, J., Rhein, A.V., Kastner, C., Apel, S., Dorre, J., Lengauer, C.: Large-scale variability-aware type checking and dataflow analysis (2012)
15. Lü, L., Zhou, T.: Link prediction in complex networks: a survey. Physica A **390**(6), 1150–1170 (2011)
16. Magee, J., Sugden, J.: "The World at their Feet" professional football and international labor migration. J. Sport Soc. Issues **26**(4), 421–437 (2002)
17. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**, 036104 (2006)
18. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
19. Osgnach, C., Poser, S., Bernardini, R., Rinaldo, R., Di Prampero, P.E.: Energy cost and metabolic power in elite soccer: a new match analysis approach. Med. Sci. Sports. Exerc. **42**(1), 170–178 (2010)
20. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web. Technical report 1999–66, Stanford InfoLab, November 1999. previous number = SIDL-WP-1999-0120

21. Palacios-Huerta, I.: Structural changes during a century of the world's most popular sport. Stat. Methods Appl. **13**(2), 241–258 (2004)
22. Redhead, S.: Post-Fandom and the Millennial Blues: The Transformation of Soccer Culture. Routledge, Abingdon (2002)
23. Strogatz, S.H.: Exploring complex networks. Nature **410**(6825), 268 (2001)
24. Taylor, I.: 0n the sports violence question: soccer hooliganism revisited. In: Sport, Culture and Ideology (RLE Sports Studies), p. 152 (2014)
25. Liu, X.F., Liu, Y.L., Lu, X.H., Wang, Q.X., Wang, T.X.: The anatomy of the global football player transfer network: club functionalities versus network properties. PLoS One **11**(6), e0156504 (2016)