World Cups impact analysis in the soccer players transaction and soccer globalization using complex network techniques

Antônio P. S. Alves[1], Lucas G. da S. Felix[2], Vitor E. do Carmo[1], Carlos M. Barbosa[1], Vinícius da F. Vieira[1] and Carolina R. Xavier[1]

[1] Universidade Federal de São João del Rei, Brazil,
[2]Universidade Federal de Minas Gerais, Brazil
{carolinaxavier, vinicius}@ufsj.edu.br

Abstract.

In this paper, we propose an analysis of the relationship between World Cup results and the number of transfers of soccer players of their national teams. For this study, networks are collected, modeled and generated for periods of time before each world cup since 1966. The effects of these events were evaluated by investigating the best and worst teams transfers networks, at each edition of the cups. We also investigated sociological theories that associate globalization to transfer networks in soccer, being able to show through quantitative data, the hypotheses raised and to renew these proposals showing the rise of new markets, such as those from Asia. To carry out the analysis, complex networks and data mining techniques were combined and this evaluation showed that countries that perform many transactions do not necessarily perform well in the world cups. However, part of the countries involved in numerous transfers can have a good performance, standing in good positions after the world cups.

Categories and Subject Descriptors: E.1 [**Data**]: Graphs and networks; H.2.8 [**Database Applications**]]: Data Mining

Keywords: Complex networks, data mining, network analysis, soccer

1. INTRODUCTION

Recently, the sports industry has surrender to data driven analysis in the field and business side, making data analysis a vital part of teams decision-making steps. Thus, with the premise that more information can assist them to win championships, teams, clubs, coaches, managers and athletes search ways to evaluate and improve their performances, given that more winnings implies in increasing the number of fans and financial revenue [Fried and Mumcu 2016].

Data driven solutions are applied in sports from the prediction of matches in Soccer and Martial Mixed Arts (MMA) [Baboota and Kaur 2018; Silva et al. 2015], passing to more analytic evaluations [Vaz de Melo et al. 2012]. In brief, data is everywhere in the sport field [Fried and Mumcu 2016].

Considering a global context, soccer is the most popular sport in the world [Palacios-Huerta 2004]. Different from other regional sports as Baseball in United States and Rugby in Australia, soccer attracts fans in every part of the globe. Due to this fact it has many practitioners and has an appeal to enormous amounts of people to events as FIFA World Cup, continental and intercontinental championships, being often between the most discussed topic in social networks such as Twitter [twi], having one of the most liked photos on Instagram [ins], and most liked pages on Facebook [fac].

In financial terms, soccer performs one of the biggest movement among all sports [glo ;], produced by ticket sells, TV contracts, marketing and merchandising. For instance, only in Europe in 2016/2017 season, soccer has moved approximately 25 billions of euros [Deloitte 2016]. Also, there is a financial compensation attached to clubs, like uniform sells, sponsors, TV channels shares and also the revenue obtained by transfering players [Liu et al. 2016].

The transaction market is extremely influenced by the World Cup. An illustrative example is Kylian Mbappé, which after the 2018 World Cup had an increase in his market value from 120 million euros to 150 million euros. The FIFA World Cup is now the world's second-largest sports event [Baade and Matheson 2004], with ever-growing numbers of viewers, a cumulative audience of 42.5 billion people [Palacios-Huerta 2004] and a value of 1.8 billion of dollars in the year of 2002, raised only by the event organizer. After nearly 30 days of championship, soccer players that outstands in the competition are greatly high-valued.

In this work we investigate the influence of the FIFA World Cup on the transfers of soccer players. Our main interest in this work is to evaluate if top ranked countries in a World Cup had an increase in their number of buys and sells. We also make this evaluation for the bottom ranked countries. Lastly, we analyze and quantify sociological theories about soccer and globalization, which see the transactions of players between countries as a labor force exchange [Maguire 1994; Maguire and Pearton 2000; Poli 2010].

To accomplish this work, and properly investigate sociological theories, we model the transfers between the countries as a graph, and perform several complex network and data mining analysis on the structure. Our results show a globalization effect over the modeled transfer network, since year after year more countries take part of the transaction market. Also we conclude that not necessarily countries which are active negotiators in the market are top ranked in the world cup. However, top positions mostly imply in an increase of the sells of the country in the next years after a world cup.

This work is an extension of Felix et al. [2018], incorporating new analysis of the dataset, enriching the understanding of how teams of important leagues impact the performance of their national teams in world cups. The present work also deepens the discussion about the obtained results and the concepts involved in the proposed methodology and presents a wider range of related works.

2. BASIC CONCEPTS

2.1 Complex networks and centrality measures

According to Barabási [2016], a complex network is a catalog of the components of a system, often called nodes or vertices and the direct interactions between them, called edges. Newman [2003] explains that networks can represent many relationships seen in the real world. Examples include the Internet, the World Wide Web, social networks of knowledge or other connections between individuals, organizations, business, neural, metabolic, food and distribution networks.

One of the most simple and important concepts in Network Science is centrality. Considering the social network context, centrality models interpersonal relationships and the importance of certain nodes in a network. Centrality can be defined under several approaches, as the number of connections a vertex has. Thus, when applied on real world networks, it helps to find the most important relationships in a network.

Several measures based on the network structure have been proposed in the literature, assisting to find the best characterization in networks. Using a certain characteristic, either the number of friends an individual has or the number of important people a person knows, it is possible to determine how important an individual is to the analyzed group. Also, several ideas of how a node or a link could be considered important have emerged, giving rise to different measures, later known as centrality measures, being used in many types of networks, for example, communication, biological, metabolic networks, or any other system that could be modeled as a complex network. Hence, it was noted that such an idea could be extended to the network connections, and it is possible to evaluate the importance of the connections between its nodes.

In the following sections, we will describe the centrality measures used in this work and also what characteristics they take into account and how to calculate them. To obtain the expressions for the

measures of centrality, we adopted a strategy similar to that made by Newman [2009] and by Ronqui [2014].

2.1.1 *Degree centrality.* Degree centrality is possibly the simplest, easiest and most intuitive way to determine centrality. This metric evaluates the importance of a node by analyzing the number of nodes to which it is connected. In other words, the greater the number of nodes connected to it, the greater the importance of the node to the network, therefore, the higher the value assigned to this vertex by the centrality. The degree of a node can be calculated using Equation 1:

$$k_i = \sum_{j=1}^{n} A_{ij} \tag{1}$$

where k_i is the degree of node i, A_{ij} are the elements of the adjacency matrix of the complex network and n is the number of vertices in the complex network. In some cases (and also in this work) one can use a variety of degree centrality, where the value of k_i is divided by the highest possible degree, thus ensuring that the value of the centrality of each node remains between 0 and 1. Equation 2 presents the normalized degree centrality:

$$k_i = \frac{\sum_{j=1}^{n} A_{ij}}{n-1}. \tag{2}$$

It is important to highlight that Equation 2 assumes that the network is a simple graph. Thus, it considers that there are no connections of a node with itself (loops) or parallel connections in the network (which would imply that the maximum degree of each node could be greater than $n-1$). If such connections exist, the expression of Equation 2 would remain valid, however, its values would no longer be limited between 0 and 1. Besides, another fact that can be highlighted is that for directed networks, two degree centralities can be defined. Considering the in-degree, it takes into account how many connections point to a node i and considering the out-degree, it takes into account how many connections are originated from node i to other nodes. To assess which of these two metrics should be considered for directed networks depends intrinsically on what represents the network and what features are being analyzed. As the present work considers only undirected networks, no further details will be given on in-degree and out-degree centralities. A deeper review in degree centrality is performed by Newman [2009].

2.1.2 *Eigenvector centrality.* Proposed by Bonacich [1987], eigenvector centrality extends the concept of degree centrality as follows: a node is important to the network if it has many connections with other nodes, or if it is connected to nodes that are themselves important. In this way, eigenvector centrality takes into account not only the connections of a node i but how many connections its neighbors have. Considering a social network, a person is important if he/she has many friends or if he/she knows some people with many contacts, or if he/she is somewhere between the two situations.

To calculate the eigenvector centrality, imagine a process in which, at the beginning, all vertices i have a centrality $x_i = 1$. In this way we can calculate the centralities x'_i of all the vertices as the sum of the centralities of all their neighbors:

$$x'_i = \sum_{j=1} A_{ij} x_j \tag{3}$$

where A_{ij} are the elements of the adjacency matrix. It is possible to note that Equation 3 can also be written using matrix notation as $x' = Ax$, where x is the vector with the elements x_i. By repeating the process of Equation 3 t times, we get:

$$x(t) = A^t x(0) \tag{4}$$

where $x(t)$ is the vector with the centralities for all nodes after t iterations and $x(0)$ is the initial value assigned to each node. $x(0)$ can be written as a linear combination of the eigenvectors v_i of the

adjacency matrix, so that:

$$x(0) = \sum_i c_i v_i. \tag{5}$$

Replacing Equation 5 in Equation 4, Equation 6 can be written:

$$x(t) = A^t \sum_i c_i v_i = \sum_i c_i k_i^t v_i = k_1^t \sum_i c_i \left[\frac{k_i}{k_1}\right]^t v_i \tag{6}$$

where k_i are the eigenvalues of the adjacency matrix A and k_1 is the largest eigenvalue. For a large enough number of iterations, the values of $x(t)$ will get into a stationary situation where all the values of its components will no longer vary. Thus, in the limit $t \longrightarrow \infty$ we get $x(t) \longrightarrow c_1 k_1 v_1$. Therefore, it can be said that the value of eigenvector centrality in the case where the values of centrality stop changing can be written as

$$Ax = k_1 x, \tag{7}$$

which is the eigenvector centrality proposed by Bonacich [1987]. From Equation 7 it can be observed that the centrality of node i depends on the centrality of all its neighbors, what can be written as Equation 8:

$$x_i = k_1^{-1} \sum_j A_{ij} x_j. \tag{8}$$

2.1.3 *Betweenness centrality.* The idea of betweenness centrality was proposed in 1977 by Freeman [1977] and consists in evaluating the importance of a node in the transmission of messages or events between the other nodes, or, equivalently, how it is in the path between the network if they want to exchange information. Equation 9 defines betweenness centrality for vertices.

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}} \tag{9}$$

where n_{st}^i is the number of shortest paths between vertices s and t that passes through vertex i and g_{st} is the total number of shortest paths between vertices s and t. If the network is composed of more than one connected component, the sum considers only the nodes belonging to the same component of node i, considering that in most cases, the comparison of betweenness centrality between nodes of different components is impossible, since there are no paths between nodes of different components.

2.1.4 *Closeness centrality.* Closeness centrality was defined by Freeman [1979] in 1979 and it is another example of centrality that uses the information of the distance between vertices. It measures how far each vertex is close to the others which are given by the geodesic distance from one vertex to all other nodes in the network. A detailed review of the closeness centrality can be found in [Newman 2009]. The goal of this measure is to assess the extent to which a node is distant from the others. Thus, nodes that have a lower average distance compared to others will receive a high value for closeness centrality. Besides, such nodes must be considered important in a complex network because of their influence, since the information present in them reaches the other elements of the network in a shorter time than the others. Closeness centrality can be calculated by Equation 10:

$$C_i = \frac{1}{l_i}, \text{ where } l_i = \frac{1}{n-1} \sum_{j(\neq i)} d_{ij}, \tag{10}$$

where n represents the total number of nodes in the network and d_{ij} is the length of the shortest path between nodes i and j, thus l_i represents the average length of the shortest distances between i and all other nodes in the network. It is important to note that C_i is defined as the inverse of this mean

value so that closeness centrality holds the same pattern as the other measures, where nodes with a larger value for centrality (and therefore the smallest mean geodesic distance) are the most central.

This measure also presents some issues, specially when the considered network has more than one connected component. Therefore, this work only considers the largest connected component of each of the networks, in cases where the networks have more than one. For details on some solutions for the case of several components, please refer to [Newman 2009].

2.1.5 *PageRank centrality.* Another measure of centrality based on eigenvectors is the PageRank centrality, applied to directed networks. The PageRank centrality [Page et al. 1999] was initially proposed to explore the hyperlinks structure of the web, and thereby evaluate the relevance of the pages. This is the technology that undergoes Google's search engine, which generates lists of useful pages from an index of pages that match the search done by the user. Langville and Meyer [2006] perform a very complete study of PageRank, showing general concepts of search engines, crawling, information retrieval, and the mathematical concepts involved in the calculation this measure.

The calculation of PageRank centrality is based on the propagation of a centrality value proportional to the number of outgoing edges. In this way, vertices that have a high out-degree propagate only a small part of centrality to the others, even if the centrality is high. The calculation of PageRank is presented in Equation 11.

$$R_p(i) = \alpha \sum_j \frac{A_{j,i}}{d_j} R_p(j) + \rho \tag{11}$$

where α and ρ are constants and d_j is the out-degree of vertex v_j if the degree is greater than zero, and $d_j = 1$, if the out-degree is zero.

2.2 Principal Component Analysis for dimension reduction

Principal Component Analysis (PCA) is a well-known method used for dimensionality reduction, which works by identifying sets of uncorrelated variables that explain most of the variability of the data. In algebraic terms, we are interested in smaller rank matrices that allow us to explain the original data and reconstruct them as closely as possible.

PCA has several attractive features [Tan 2018]. First, it tends to identify the stronger patterns in the data. Thus, PCA can be used as a pattern identification technique. Second, many times the variability of the data can be captured by a small fraction of the total set of dimensions. As a result, dimensionality reduction using PCA can result in data of relatively low dimensions and it may be possible to apply techniques that do not work well with high-dimensional data. Third, since the noise in the data is weaker than the standards, the reduction of dimensionality can eliminate much of the noise. This is beneficial for both data mining and other data analysis algorithms.

The idea behind PCA is to find a new set of dimensions (attributes) that best capture the variability of the data. More specifically, the first dimension is chosen to capture as much variability as possible. The second dimension is orthogonal to the first and, subject to this restriction, captures as much of the remaining variability as possible, and so on.

3. RELATED WORK

The use of complex network techniques for machine learning and data analysis purposes has gained great attention in the last years. Silva and Zhao [2016] present an overview of the theme and in other works, they apply the concepts of complex networks for data classification as in [Carneiro and Zhao 2018; Cupertino et al. 2018].

Several works related to soccer have been presented for decades, involving works that evaluate the

sport from a different perspectives, from a sociological to a computational and physical education point of view.

From a computational perspective, a lot of works consider data-based approaches for a wide range of purposes such as optimization of a team [Payyappalli and Zhuang 2019], evaluation of players [Cotta et al. 2016; Matano et al. 2018; Pelechrinis and Winston 2018], analysis of the transfer network [Felix et al. 2018; Liu et al. 2016].

The work of Payyappalli and Zhuang [2019] proposes a mathematical optimization model that presents clubs with recommendations of the best possible transfers. To do this, the authors use skill data, salary and market value of a player and recommend the best player based on the needs of the team, as well as calculate player price restrictions.

In the works of Cotta et al. [2016], Matano et al. [2018] and Pelechrinis and Winston [2018], data from the FIFA electronic game, developed by *EA Sports*, are used in order to study soccer. The work of Cotta et al. [2016], uses a data set from the game for power analysis of soccer and evaluates new styles of soccer as the *tiki-taka* deployed by the Barcelona team.

The work of Matano et al. [2018] applies a variation of a statistical regression technique that measures the contribution of each player to victory while controlling the quality of teammates and opponents called Adjusted Plus-Minus (APM). This technique is widely applied in sports such as hockey and basketball. Thus, the work applies the variation of the APM together with the data of FIFA to evaluate the performance of a player in a soccer match.

The work of Pelechrinis and Winston [2018] develops a framework that estimates how much each player contributes to the victory of a team. For this, the authors use data from 20 thousand European championship games along with data from the FIFA game, helping them to predict the probability of a team to win and to predict the values of player. However, this study does not apply real data to predict the value of soccer players, using a metric proposed to estimate the transfer value and salary of the athlete.

The works of Liu et al. [2016], Felix et al. [2018], as well as the current work, use network metrics to evaluate the player transactions market. The work of Liu et al. [2016], points out only some properties of the built network, having as main objective to analyze the success of a team according to their transfers. A previous work ([Felix et al. 2018]) applies complex network metrics to analyze the top ranked countries and communities formed by teams present at the 2018 World Cup.

From social and economic perspectives, some works ([Lee and Taylor 2005; Kaplanski and Levy 2010; Baade and Matheson 2004]) analyze the economic impact of the world cup in different ways, with positive and negative reflections on the countries participating in the event. In the work [Lee and Taylor 2005], the economic impact of the 2002 World Cup, specifically related to tourism is evaluated. In [Baade and Matheson 2004] the authors analyze whether to host the world cup causes more losses or profits to a city due to the massive investments in infrastructure that are necessary to carry out the event. In the work [Kaplanski and Levy 2010] the authors evaluate how the results of a country in the world cup can positively or negatively affect the local stock market.

In some works [Maguire 1994; Maguire and Pearton 2000; Poli 2010], soccer transfers are discussed as the effect of globalization. However, it is worth to mention that only the work of Poli [2010] considers quantitative data on European leagues in its methodology. The objective of the study is to verify if the increase of the international flow reflects in a spatial diversification of migratory routes or if there are privileged channels of recruitment of the countries origin and destination. On the other hand, Maguire and Pearton [2000] and Maguire [1994] analyze the impact of migration on sport, the first work being more focused on the development of European players and the second one on analyzing the labor flow not only in soccer but in all sports.

Some related works that directly deal with soccer transactions as central theme can be highlighted,

as the works of Palacios-Huerta [2004] and Frick [2007], which apply different methodologies for the evaluation of the transfer market. The study of Palacios-Huerta [2004] assesses soccer transactions through temporal behavioral statistical analyses, investigating only English leagues and giving an economic view of the sport. On the other hand, the work of Frick [2007], investigates the market for athlete transfers in Europe empirically, evaluating aspects not considered in our analyses such as player salary and player career time.

4. DATA COLLECT AND MODELLING

The current work presents a study of soccer transfers and their impact on the world cup, applying complex network techniques with data mining and analysis principles. In order to perform these tasks, it was first necessary to collect data to identify the transactions between soccer teams. For this, the data from the Transfermarkt website [1] was gathered. Transfermarkt is a large dataset with diverse information related to soccer, including statistics, championship standings and data related to the transactions of players, particularly important to the study conducted in this work.

Data on external transactions were collected from 1962 to 2017. For each year, the 250 most important transactions in terms of monetary value by position were collected. However, it is worth noting that for some years the number of transactions on the site has not reached 250, including cases in which the number of transactions in a given position was equal to zero. Figure 1 shows the number of transfers made in each period by positions. At the same time, it is possible to see that the fraction of players of midfield and attackers that have been transferred are superior to other positions, as it is possible to observe the growth of the number of transactions per year.

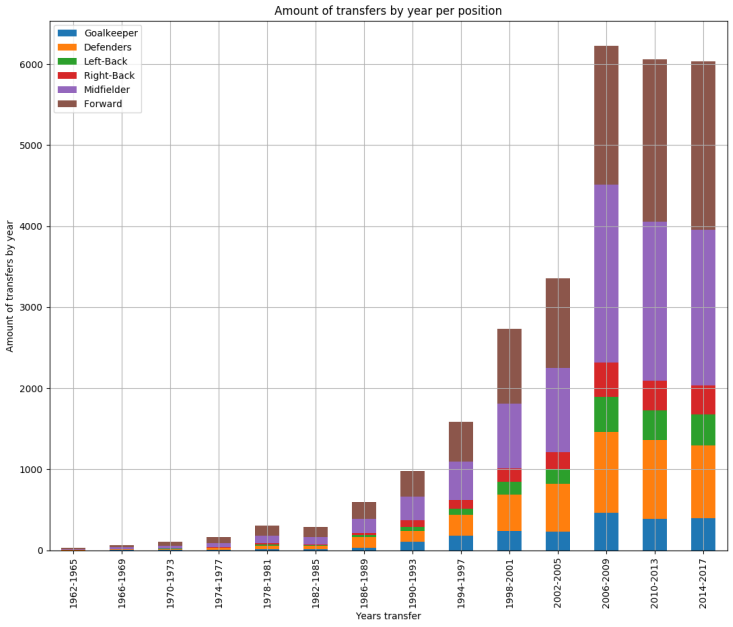


Fig. 1. Amount of transactions in each year.

[1] transfermarkt.com

After obtaining the data, the networks were modeled. In our study, we consider a set of vertices V, where each $v \in V$ is a country, a node v_i has a directed link to v_j if the countries i and j have already made a transaction. Each edge $e \in E$ has a weight $W_{i,j}$ that represents the amount of the transfer where a player p has left a country i and gone to a country j.

To model the networks, a 4-year interval was considered, since the transactions were grouped to summarize the transactions carried out between one world cup and another. This approach was considered to better model the networks in the context of the FIFA World Cup. Thus, an evaluation was made considering that in the first half of the event year a team is already assembled and ready to play in the competition. In the second year, a team that does not perform well uses this time to remodel e assemble a new group. So, it can be argued that the period that define the world cup of 2018, for example, are the years from 2014 to 2017.

After modeling the networks, some properties of the structure could be calculated, specially for the analysis presented in Section 5. These properties are exhibited as long they were used for the analysis of the networks.

To analyze the first placed national teams of the World Cup, data was gathered from the Wikipedia [2]. From Wikipedia, it was collected the list of players of the champions and last placed teams, considering data about ages and teams they have played until the evaluated World Cup. Data on where first placed athletes played when they were champions was collected in the year of each World Cup edition since 1966 - to be compatible with the analyzed transfer network. Thus, of all cups from 1966, the information related to the 22 players of the three top ranked teams and the three bottom ranked teams in the world cup of each edition was collected. To gather this data a parser was implemented, responsible to get, clean and pre-process all data, given that it was originally in an unstructured data format.

For the analysis of these players' distribution, we counted the countries for which they competed, to obtain the statistics of these countries in the national teams. From this dataset, we were able to construct 14 tables with the locations where players from any specific team and any specific year acted.

In this work, the centrality measures of PageRank, Betweenness, Closeness, Eigenvector, and Degree were employed. Each of these metrics has a way to define the main vertices of the network, as detailed in Section 2. Lastly, Principal Component Analysis (PCA) was applied in our methodology to perform dimensionality reduction over a set of centrality measures, generating a single rank. With this rank, we evaluate if countries vary their importance in the soccer market or if this importance is stable through the years. The choice of this technique was based on the advantages pointed out by Han et al. [2011] concerning other algorithms, particularly their ability to deal with ordered and unordered data, being able to have good results even with sparse data.

5. NETWORK ANALYSIS

5.1 Soccer Globalized

According to Beck [2018], globalization can be defined as a process where national sovereignty is crossed and undetermined by international actors varying powers, orientations, identities, and networks. In the case of soccer, these actors are the athletes who cross national barriers to be able to take chances in clubs from other countries [Poli 2010]. When the collected data are analyzed, it can be seen that the number of countries increases from year to year, showing the reduction of borders between the countries for a Soccer player.

[2]wikipedia.com

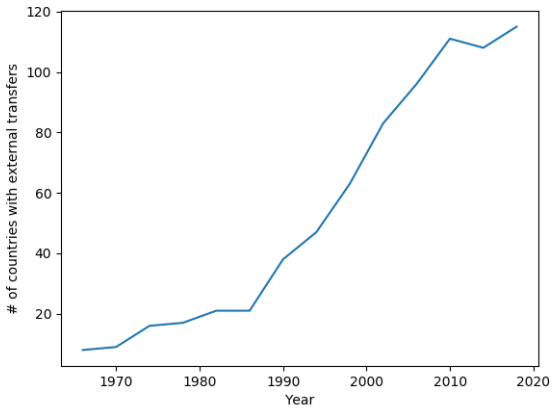


Fig. 2. Amount of countries involved in transactions per year.

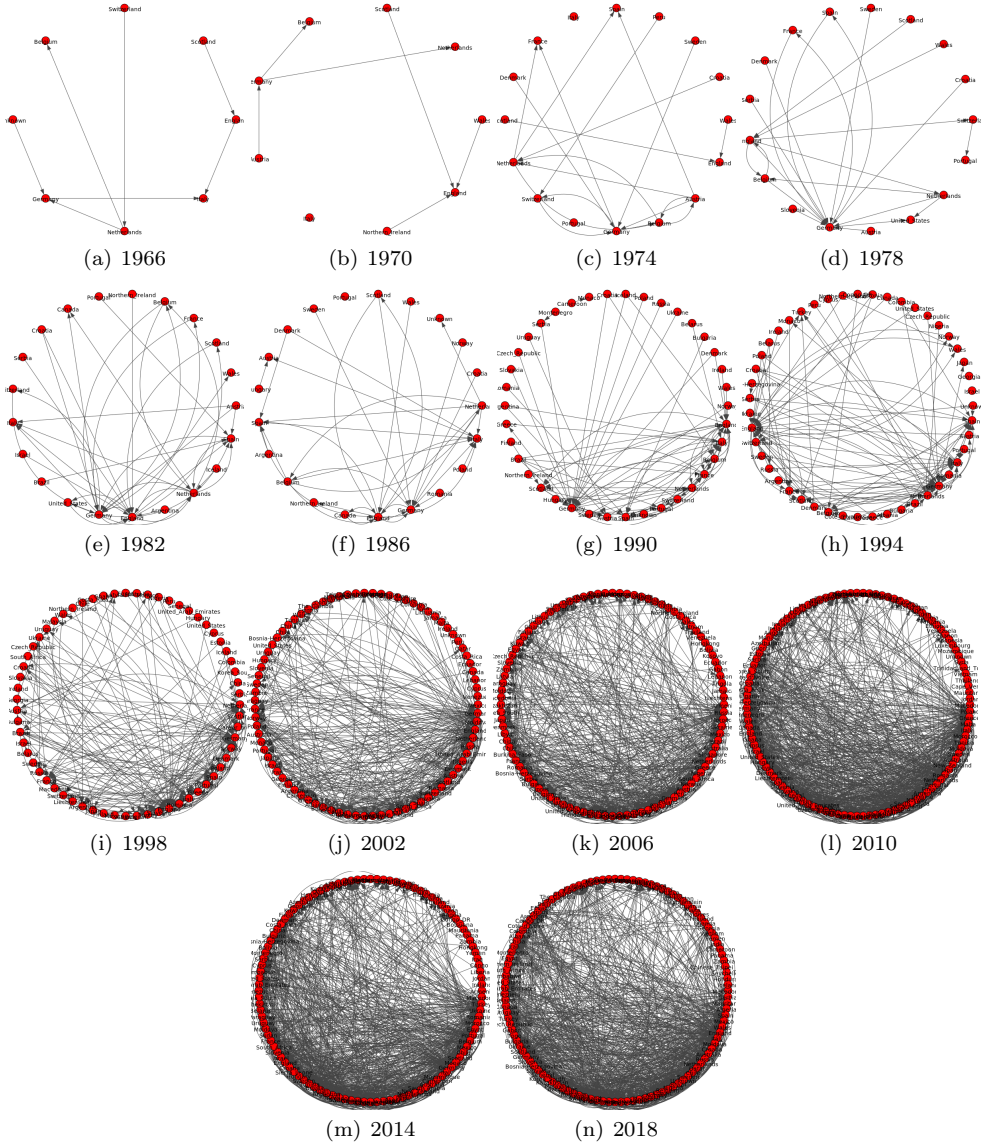
From Figure 2, it can be observed that in the first years of the studied period the number of transfers and countries involved is much lower when compared with current times. One can justify this fact simply by stating that the dataset is incomplete and it has no data because of the difficulty of accessing information at the time, however, this is a phenomenon previously studied by sociology that demonstrates that globalization can be understood through soccer [Poli 2010; Maguire and Pearton 2000; Maguire 1994]. And as the number of transactions made by these countries has increased, so has the variety of countries participating in the transactions. This may vary according to the properties of the networks over time.

Figure 3 shows 14 networks generated by the collected data. It is easy to see the growth of the number of vertices (countries) and the edges (transactions) from one Cup to another.

Year	Degree assortativity	Density	Diameter	Reciprocity	Clustering coefficient
1966	-0.091	0.17	3	0	0
1970	-0.707	0.08	2	0	0
1974	-0.093	0.08	5	0.32	0.24
1978	-0.512	0.08	4	0.26	0.16
1982	-0.630	0.10	4	0.52	0.23
1986	-0.362	0.09	6	0.39	0.32
1990	-0.403	0.06	5	0.28	0.26
1994	-0.301	0.05	5	0.29	0.23
1998	-0.270	0.05	5	0.38	0.27
2002	-0.255	0.06	6	0.38	0.35
2006	-0.162	0.07	6	0.34	0.40
2010	-0.116	0.08	5	0.43	0.45
2014	-0.079	0.08	6	0.41	0.46
2018	-0.081	0.07	5	0.45	0.50

Table I. Properties of the World Cup networks.

Table I shows the main properties related to all generated networks. To visualize these properties basic complex network metrics were considered. Among these metrics **density** comprises the ratio between the number of edges of the graph G and the number of edges of a complete graph G' with the same number of vertices, **diameter** is the greatest distance between the vertices of a graph G, **reciprocity** is the measure of the possibility of the vertices of a directed graph G being mutually connected, **degree assortativity** is the measure of how the vertices of a graph G tend to have



(a) 1966 (b) 1970 (c) 1974 (d) 1978

(e) 1982 (f) 1986 (g) 1990 (h) 1994

(i) 1998 (j) 2002 (k) 2006 (l) 2010

(m) 2014 (n) 2018

Fig. 3. Graph Transfers between World Cups.

connections with vertices of similar degree, **Cluster coefficient** measures the probability that two neighbors of a vertex are connected (the ratio of the triangles and connected triples in the graph).

It is possible to notice that the reciprocity of the network has the highest value of approximately 52 % in the year 1982. It can be observed from these values that most networks have shown small reciprocity, which in the studied context, indicates the presence of two distinct roles in the networks: the consumer countries, which generally do not buy in the same rate that they sell, thus providing a continuous flow of labor forces to regions with more money and prestige; and the producer countries, which usually only buy players from the producer countries. Thus, as stated by Poli [2010], companies do not look only for consumer markets and raw materials in southern countries, but also in the workforce.

From the analysis of the network diameter, it is possible to see the "degrees of separation"between

the vertices, considering the increasing number of vertices in the generated networks, as illustrated in Figure 2. Compared to this metric, we see that these values of diameter are not increasing over the last years, with the highest values between 5 and 6, showing that even with the network growth the distance between the countries present in the network remains the same. The analysis of density allows us to identify sparse networks since the highest value is 0.17 (in 1966) when the number of countries was very low, which shows that to obtain a complete graph, with edges between every pairs of nodes, we would need to highly increase the number of edges. In the context of soccer, this shows that many countries make connections with few countries, transforming this market into a business where few countries are very connected, and most of these very connected vertices are extremely consuming countries that are looking for new talented players. The assortativity of the networks confirms that countries with many connections tend to connect to countries with few connections, all networks presented disassortativity connections.

The clustering coefficient shows that the number connected triangles increased over the years. In 1966, the probability of two neighbors of a vertex to be connected was 0, but in 2018 this probability is raised to 50%. Combining this result to the density results allows us to observe that there is a probability of existence of big hubs in these networks.

Year	Max. degree	Max output strength	Max input strength
1966	3(Germany)	4.0(Scotland)	5.0(Italy)
1970	3(Germany)	2.0(Scotland)	4.0(England)
1974	7(Germany)	7.0(Scotland)	8.0(England)
1978	14(Germany)	10.0(Netherlands)	17.0(Germany)
1982	17(Germany)	21.0(Netherlands)	28.0(England)
1986	13(England)	13.0(England)	25.0(England)
1990	24(Germany)	34.0(England)	44.0(England)
1994	29(England)	65.0(England)	182.0(England)
1998	50(Germany)	140.0(Germany)	324.0(England)
2002	70(Germany)	210.0(England)	610.0(England)
2006	61(Germany)	322.0(Brazil)	510.0(England)
2010	77(Germany)	478.0(Brazil)	776.0(England)
2014	73(Russia)	400.0(Spain)	722.0(Italy)
2018	80(Italy)	438.0(England)	888.0(England)

Table II. Max. Degrees and strengths.

Table II shows other important measures. **Max. degree** is the number of edges of the vertex with the greatest number of connections of a graph. In the studied networks, this node corresponds to the country with more commercial partners each year. **Max. output strength** is the sum of the weights of the outgoing edges of the vertex V of a graph G and **Max. input strength** is the sum of the weights of the incoming edges for the vertex V of a graph G. These measures in the networks mean the country that respectively exports and imports more players each year.

When analyzing Table II it is possible to observe how much it tells us about soccer in general, since these metrics indicate countries that make many transactions of sale and many transactions of purchase. The country that has the maximum output value of all networks is Brazil (478) in 2010 showing that it is a producer country, with its role in the market in the sale of players to large countries, such as England representing the country that has the maximum input weight (888), showing that this is a country with consumer role in the market. In general European teams figure out in all main positions, showing the great investment that these countries have made in their national championships.

Considering these aspects, we see that it is possible to observe the globalization and theories developed by sociology in soccer networks. These can show us beyond the workflow for large consumer countries present in Europe and new growth markets like China and Saudi Arabia, as well as producing

countries, also called farm markets, which provide athletes for richer leagues. Besides, it is possible to visualize the expansion of the number of vertices of the networks, which shows us the expansion of the market as a whole. Finally, the properties of the networks give us a variety of information that underlies, strengthens, details and updates the sociological theories raised, since in general, these works are out of date because they do not consider the rise of the Asian market, especially China which has brought monetary proposals with values above the European market, even though the teams do not have the same prestige as the major European clubs.

5.2 Impact of transactions in World Cups

Considering the historical panorama of transactions carried out in World Cups, when the transfers made by countries that were in the first three positions and the last three positions are analyzed, it was found that in approximately 75 % of the cases, top-ranked countries in the competition manage to increase the number of transactions from one cup to another. On the other hand, bottom-ranked countries show a decrease in the number of transactions, where in 71 % there is no increase in this number. It is worth noting that the number of sales transactions of 84 % of the countries increased, thus enhancing the national player market.

Figures 4 and 5, show the transactions behavior of the best and the worst placed countries in a world cup, respectively. It is possible to see that countries that win the cup mostly increase the transfers of players in the next four years, while countries that lose continue to have a small transfer rate through the years. Another clear result that can be observed in these Figures, specially in Figure 4, is the increase of the amount of transfer through years, as discussed in Section 5.1.

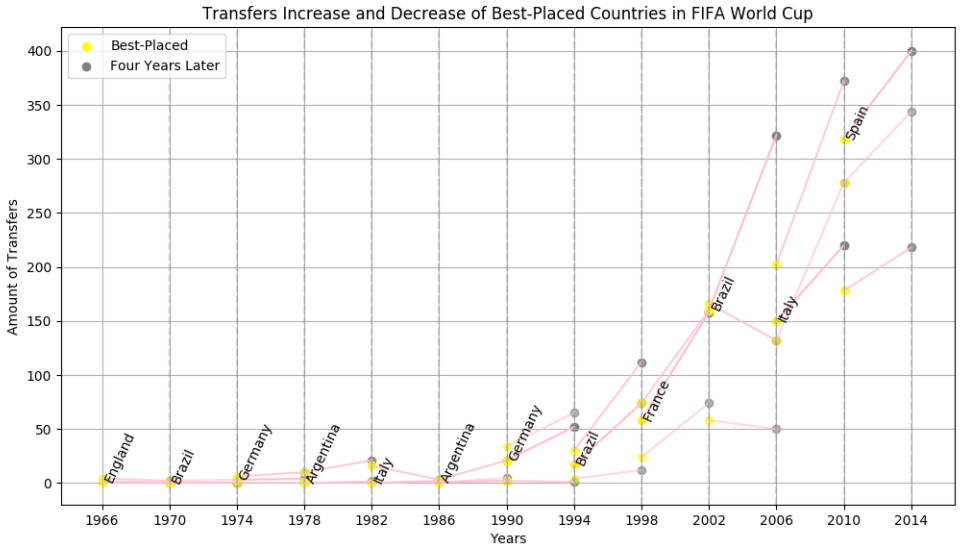


Fig. 4. Transfers behavior of the Best-Placed Countries in World Cup

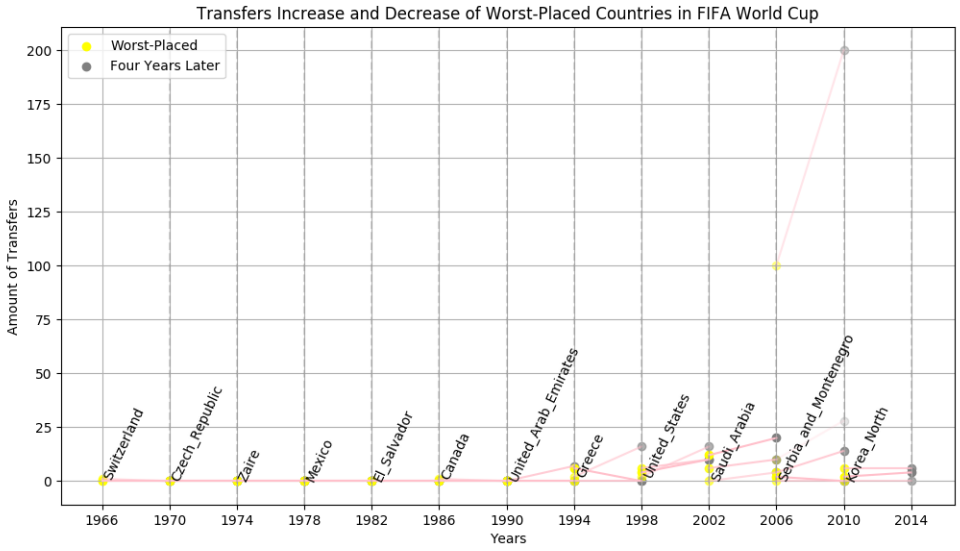


Fig. 5. Transfers behavior of the Worst-Placed Countries in World Cup

In order to investigate how the importance of a country impacts its performance in the network, several rankings were defined, considering centrality measures available in the *igraph* library [3] for *python*. However, considering that the different centrality measures define the importance of a node from different perspectives and with the objective to define an unique centrality ranking for each of the fourteen considered networks, Principal Component Analysis (PCA) was applied. Thus, the dimensionality of rankings can be reduced, providing a single ranking that describe the importance of the countries for the network combining different centrality measures.

After applying this method, centrality rankings were defined to each network. These rankings were compared using Spearman correlation in order to verify whether countries that are central in one year tend to remain relevant in the other World Cup or these countries are changing over time.

Year	**Correlation top # 8**	**Correlation**
1966-1970	0.047	-0.238
1970-1974	0.380	-0.116
1974-1978	0.238	0.071
1978-1982	-0.261	0.208
1982-1986	0.833	0.035
1986-1990	0.190	0.125
1990-1994	0.595	-0.045
1994-1998	0.785	-0.079
1998-2002	0.714	0.025
2002-2006	0.309	0.144
2006-2010	0.285	0.167
2010-2014	-0.547	0.114
2014-2018	-0.071	-0.005

Table III. Correlation year from each year to the previous one.

[3] igraph.org

By analyzing Table III it is possible to realize that the correlation values vary a lot over the years, even if only the top 8 positions (the minimum number of countries in all networks - 1966) is considered, which means that the generated rankings are not constant, so the transaction market tends to greatly vary from one cup to another. Table III also allows us to identify the rise of Asian countries like China and Saudi Arabia in recent years. However, it is also noticeable that the first positions remain between some European countries, such as Italy, England, Spain, France, and Germany. Comparing the correlation in top #8 countries with all the list, one can see that the full list is much more variable then the top-ranked list. It can be also noticed that the list is increasing through the years and the positions of countries change all the time, showing a lack of correlation in the full list.

When Table III is associated to the results of the World Cups, in order to verify if the importance of a country in the network directly implied good placings in the cup, it is possible to verify that in the great majority of the competitions, at least one country that was among the first positions of the ranking had a good performance in the event, finishing in the top three.

Year	**First places in the World Cup**	**First places in the ranking**
1966	England, Germany, Portugal	Germany, Netherlands, Italy
1970	Brazil, Italy, Germany	England, Germany, Belgium
1974	Germany, Netherlands, Poland	Germany, Netherlands, Belgium
1978	Argentina, Netherlands, Brazil	Germany, England, Netherlands
1982	Italy, Germany, Poland	England, Germany, Netherlands
1986	Argentina, Germany, France	England, Germany, Italy
1990	Germany, Argentina, England	England, Germany, Italy
1994	Brazil, Italy, Switzerland	England, Germany, Italy
1998	France, Brazil, Croatia	England, Germany, Spain
2002	Brazil, Germany, Turkey	England, Germany, Italy
2006	Italy, France, Germany	England, Germany, Russia
2010	Spain, Netherlands, Germany	England, Germany, Spain
2014	Germany, Argentina, Netherlands	Italy, England, Germany

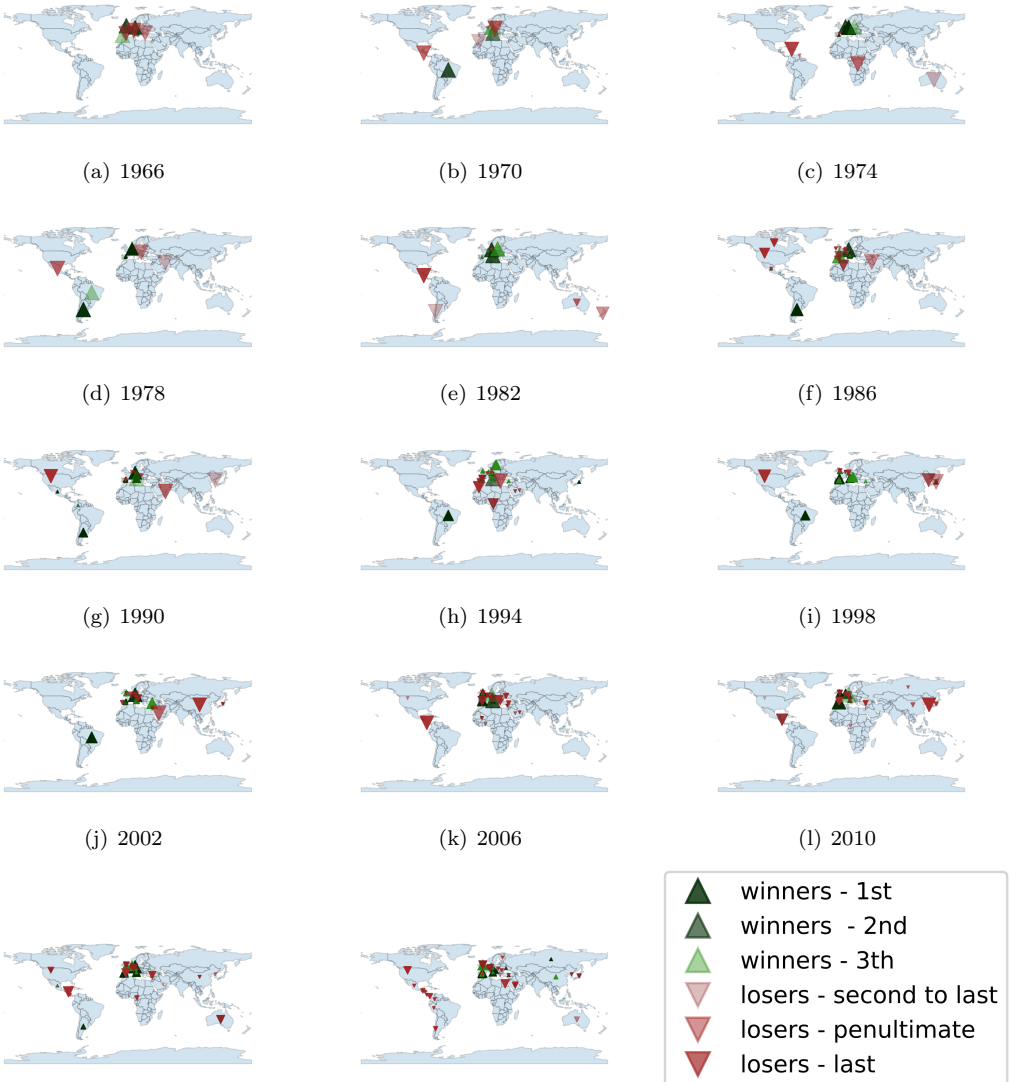
Table IV. First positions at each World Cup edition and at each network ranking.

It is possible to notice from Table IV that the incidence of countries present in the column of main countries is almost constant, although it is worth noting that the country that appeared for several years as most central, England, has won only one competition, in 1966, having a good place again only in 1990, finishing in third place. This shows that the number of transactions carried out by a country does not necessarily induces its position in the world cup. It is also worth noting that the investment made in players by a country reflects on the quality of teams for a national and international championship, but does not reflect on a national team strong enough for major continental and world competitions.

5.3 Impact of the most important leagues in the World Cup

Table IV shows that at least one of the top-ranked countries finished the current World Cup in the first positions. However, from the analysis of the country where the world champions played, it was noticed that the effect of the transactions has an inverse relation with the performance of the national teams. This is because these players were in countries that made a lot of investment, forming an extremely competitive national league and with a very high quality, which implies that players are playing at a high level.

An overview of all World Cups since 1966 can be illustrated by Figure 6, where the triangles are placed in countries from which players were acting in the occasion of their invitation to the national team to play the World Cup. Green triangles represent the top-ranked teams and red triangles represent the bottom-ranked teams. The sizes of the triangles are proportional to the numbers of players in each country.



(a) 1966 (b) 1970 (c) 1974

(d) 1978 (e) 1982 (f) 1986

(g) 1990 (h) 1994 (i) 1998

(j) 2002 (k) 2006 (l) 2010

(m) 2014 (n) 2018 (o) Legend

Fig. 6. Distribution of soccer players for each national team and each World Cup.

It can be seen that from the first analyzed Cup until the 1982 Cup, the three best ranked teams and the three worst ranked teams had the players of their national teams acting in the country of origin, which corroborates the data observed in Figure 1 and allows us to validate the assumption that in the early years analyzed, transactions of players between countries were uncommon. For example, the 1970 champion team, the Brazilian team was formed only by players who played in Brazil. The Brazilian national team of 2002, also a world champion, was formed mostly by players who played in Brazil, but also by players who competed in England, Italy and Spain. That is, the phenomenon of globalization, as Beck [2018] suggests, had not taken shape in this period. However, a change can be observed in 1986, when a greater distribution of players of the worst placed teams can be observed and in 1990, when the players of the best and worst teams of that cup begin to act in countries different from their origin. Over the years, the globe, as a whole, has become the stage for the players of the champions and the last places. In the last Cup, for instance, players were acting on all continents.

Europe, as a major reference to world soccer, is a clear example of this situation. Until the 1990s, there was a great hegemony of the place where its players acted. Italy of 1982, for example, had all its players playing in their own country, while from the 1990s onwards, European countries that finished in the first position (the exception is, again, Italy, in 2006) had their players scattered among other countries, especially in Europe itself. This can be explained by the consolidation of globalization and also, in the case of Europe, by the formation of the bloc of the European Union (EU) in 1993, which destroys the physical borders between bloc countries and allows a greater movement of goods, money, and people.

Year	Country	Percentage of players
1966	**England**	100
1970	**Brazil**	100
1974	**Germany**	95.45
	Spain	4.55
1978	**Argentina**	95.45
	Spain	4.55
1982	**Italy**	100
1986	**Argentina**	68.18
	Italy	13.63
	France	4.54
	Spain	9.09
	Mexico	4.54
1990	**Germany**	77.27
	Italy	22.72
1994	**Brazil**	50
	Spain	13.63
	Germany	13.63
	Italy	9.09
	France	9.09
	Japan	4.54
1998	**France**	45.45
	Italy	31.81
	England	13.63
	Germany	4.54
	Spain	4.54
2002	**Brazil**	56.52
	Italy	17.39
	Spain	13.04
	France	8.70
	Germany	4.35
2006	**Italy**	100
2010	**Spain**	86.96
	England	13.04
2014	**Germany**	69.57
	England	17.39
	Italy	8.70
	Spain	4.35
2018	**France**	39.13
	Spain	26.09
	England	21.74
	Germany	8.70
	Italy	4.34

Table V. Percentage of players of the champion team (bold) playing in the best leagues.

Thus, it has been observed that in the last editions, the top-ranked teams have a significant number of players playing exactly in the countries that have invested the most during the studied period, which makes these leagues the fundamental components to understand the performance of the world

cup national teams. For example, the 2010 world champion, Spain, had players playing championships in Spain and England, who are the same countries that were among the most invested countries in the period of this cup. In 1966, on the other hand, it was not very common for players to compete in leagues from other countries, but, as indicated in Table V, we can see that the share of players playing in other championships has increased over time. In fact, from the 1986 World Cup onwards it was possible to see that the teams had their players playing in countries that invested most in transactions, which is consistent with the sociological phenomena of globalization in soccer [Poli 2010; Maguire and Pearton 2000; Maguire 1994]. Therefore, it can be noticed that the countries that invest more participate in the positive performance of selections that had significant portions of players who played in these.

6. CONCLUSIONS AND FUTURE WORKS

This work presented a study of soccer player transfer networks carried out in the four years before each world cup between 1966-2018 and analyzed their relationship with the FIFA World Cup using complex network and data mining techniques. To perform this task, the methodology considered data collected from the Transfermarkt and Wikipedia.

From this study it was possible to confirm that soccer has followed the globalization of the world market with a global market of players transfers, promoting a great movement of labor forces. It is known that not only soccer stars who leave their countries in search of a better life, farm countries (producers of talent), as Brazil, provides players to clubs around the world, and many thrive and some even work in national teams from other countries.

Also, it was possible to notice that the incidence of countries identified as the central ones considering the ranking of transfer of players is almost constant. However, it is worth noting that a country that showed itself for several years as the main actor, England, only won one competition, at 1966, repeating a good performance only in 1990, when it finished in third place. In spite of this, the world champion teams had a significant percentage of players that competed in teams that lead the ranking of transfers. This helps to explain how Germany, who also always figured among the main countries for the studied period, was present in 10 podiums of the 13 competitions analyzed. Thus, Germany is a country that has several rich clubs that import players from around the world to strengthen the national and continental championships, making its athletes, which mostly play within the country, to play in competitive and high-level leagues, ensuring good placements in the world cups.

Brazil was not ranked among the top countries in the ranking of transfers in the top three positions, although it is a country that is, for most of the ranking measures, among the top 10, often being the country with the highest output weight, and has a significant portion of its athletes playing in the most competitive leagues. It is also noted that there is a growth of transactions for the majority of countries well placed in the world cup and a decrease in transfers to countries in the worst positions.

As future works, we intended to extend the studies in the networks performing an investigation on the communities in an attempt to identify the change of commercial partners according to the results of each World Cup. Also, another network modeling can be considered over the dataset, making teams as nodes and considering a different set of network centrality measures.

Acknowledgement. The authors thank to the funding agencies CNPq and FAPEMIG for the financial support.

REFERENCES

Here are the most retweeted sports tweets of 2016. https://www.si.com/extra-mustard/2016/12/06/sports-twitter-most-retweeted-tweets-moments-2016. Accessed: 2019-08-20.

Here are the most retweeted sports tweets of 2016. https://en.wikipedia.org/wiki/List_of_most-liked_Instagram_posts. Accessed: 2019-08-27.

List of most-followed facebook pages. https://en.wikipedia.org/wiki/List_of_most-followed_Facebook_pages. Accessed: 2019-08-27.

Soccer falls short from being the sport with the highest revenue. https://www.thenewbarcelonapost.com/en/soccer-falls-short-from-being-the-sport-with-the-highest-revenue/. Accessed: 2019-08-29.

Soccer global dominance in three simple charts. https://www.mic.com/articles/91009/soccer-s-global-dominance-of-sports-in-3-simple-charts. Accessed: 2019-08-29.

Baade, R. A. and Matheson, V. A. The quest for the cup: Assessing the economic impact of the world cup. *Regional Studies* 38 (4): 343–354, 2004.

Baboota, R. and Kaur, H. Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, 2018.

Barabási, A.-L. *Network science*. Cambridge university press, 2016.

Beck, U. *What is globalization?* John Wiley & Sons, 2018.

Bonacich, P. Power and centrality: A family of measures. *American journal of sociology*, 1987.

Carneiro, M. G. and Zhao, L. Organizational data classification based on theimportanceconcept of complex networks. *IEEE Transactions on Neural Networks and Learning Systems* 29 (8): 3361–3373, Aug, 2018.

Cotta, L., de Melo, P., Benevenuto, F., and Loureiro, A. A. Using fifa soccer video game data for soccer analytics, 2016.

Cupertino, T. H., Carneiro, M. G., Zheng, Q., Zhang, J., and Zhao, L. A scheme for high level data classification using random walk and network measures. *Expert Systems with Applications* vol. 92, pp. 289–303, 2018.

Deloitte. Annual review of football finance, June 2016.

Felix, L., Barbosa, C. M., Carvalho, I. A., Vieira, V. F., and Xavier, C. R. Uma análise das seleções da copa utilizando uma rede de transferências de jogadores entre países. *Brazilian Workshop on Social Network Analysis and Mining*, 2018.

Felix, L., Barbosa, C. M., Vieira, V. F., and Xavier, C. R. 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. *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, 2018.

Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry*, 1977.

Freeman, L. C. Centrality in social networks conceptual clarification. *Social networks* 1 (3): 215–239, 1979.

Frick, B. The football players' labor market: Empirical evidence from the major european leagues. *Scottish Journal of Political Economy* 54 (3): 422–446, 2007.

Fried, G. and Mumcu, C. *Sport analytics: A data-driven approach to sport business and management*. Taylor & Francis, 2016.

Han, J., Pei, J., and Kamber, M. *Data mining: concepts and techniques*. Elsevier, 2011.

Kaplanski, G. and Levy, H. Exploitable predictable irrationality: The fifa world cup effect on the u.s. stock market. *Journal of Financial and Quantitative Analysis* 45 (02): 535–553, 2010.

Langville, A. N. and Meyer, C. D. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2006.

Lee, C.-K. and Taylor, T. Critical reflections on the economic impact assessment of a mega-event: the case of 2002 fifa world cup. *Tourism Management* 26 (4): 595 – 603, 2005.

Liu, X. F., Liu, Y.-L., Lu, X.-H., Wang, Q.-X., and Wang, T.-X. The anatomy of the global football player transfer network: Club functionalities versus network properties. *PLOS ONE* 11 (6): 1–14, 06, 2016.

Maguire, J. Preliminary observations on globalisation and the migration of sport labour. *The Sociological Review* 42 (3): 452–480, 1994.

Maguire, J. and Pearton, R. The impact of elite labour migration on the identification, selection and development of european soccer players. *Journal of Sports Sciences* 18 (9): 759–769, 2000. PMID: 11043901.

Matano, F., Richardson, L. F., Pospisil, T., Eubanks, C., and Qin, J. Augmenting adjusted plus-minus in soccer with fifa ratings. *arXiv preprint arXiv:1810.08032*, 2018.

Newman, M. *Networks: an introduction*. Oxford University Press, 2009.

Newman, M. E. The structure and function of complex networks. *SIAM review* 45 (2): 167–256, 2003.

Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: bringing order to the web., 1999.

Palacios-Huerta, I. Structural changes during a century of the world's most popular sport. *Statistical Methods and Applications* 13 (2): 241–258, 2004.

Payyappalli, V. M. and Zhuang, J. A data-driven integer programming model for soccer clubs' decision making on player transfers. *Environment Systems and Decisions*, 2019.

Pelechrinis, K. and Winston, W. Positional value in soccer: Expected league points added above replacement. *arXiv preprint arXiv:1807.07536*, 2018.

Poli, R. Understanding globalization through football: The new international division of labour, migratory channels and transnational trade circuits. *International Review for the Sociology of Sport* 45 (4): 491–506, 2010.

Ronqui, J. R. F. *Medidas de centralidade em redes complexas: correlações, efetividade e caracterização de sistemas*. M.S. thesis, Instituto de Física de São Carlos, Universidade de São Paulo, 2014.

Silva, L. A., Messias, J., Moro, M. M., de Melo, P. O. V., and Benevenuto, F. Algoritmos de aprendizado de maquina para predicao de resultados das lutas de mma. *Brazilian Symposium on Databases*, 2015.

Silva, T. C. and Zhao, L. *Machine learning in complex networks*. Vol. 2016. Springer, 2016.

Tan, P.-N. *Introduction to data mining*. Pearson Education India, 2018.

Vaz de Melo, P. O., Almeida, V. A., Loureiro, A. A., and Faloutsos, C. Forecasting in the nba and other team sports: Network effects in action. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6 (3): 13, 2012.