Complex network study of Brazilian soccer players

Roberto N. Onody* and Paulo A. de Castro†
Departamento de Física e Informática, Instituto de Física de São Carlos, Universidade de São Paulo, C.P.369, 13560-970 São Carlos-SP, Brazil
(Received 11 March 2004; revised manuscript received 4 June 2004; published 29 September 2004)

Although being a very popular sport in many countries, soccer has not received much attention from the scientific community. In this paper, we study soccer from a complex network point of view. First, we consider a bipartite network with two kinds of vertices or nodes: the soccer players and the clubs. Real data were gathered from the 32 editions of the Brazilian soccer championship, in a total of 13 411 soccer players and 127 clubs. We find a lot of interesting and perhaps unsuspected results. The probability that a Brazilian soccer player has worked at N clubs or played M games shows an exponential decay while the probability that he has scored G goals is power law. Now, if two soccer players who have worked at the same club at the same time are connected by an edge, then a new type of network arises (composed exclusively by soccer player nodes). Our analysis shows that for this network the degree distribution decays exponentially. We determine the exact values of the clustering coefficient, the assortativity coefficient and the average shortest path length and compare them with those of the Erdös-Rényi and configuration model. The time evolution of these quantities are calculated and the corresponding results discussed.

DOI: 10.1103/PhysRevE.70.037103 PACS number(s): 89.75.Hc, 02.50.−r, 89.75.Fb

In the past few years there has been a growing interest in the study of complex networks. The boom has two reasons—the existence of interesting applications in several biological, sociological, technological and communications systems and the availability of a large amount of real data [1–4].

Social networks are composed by people interacting with some pattern of contacts like friendship, business or sexual partners. One of the most popular works in this area was carried out by Milgram [5] who first arrived to the concept of the "six degrees of separation" and small-world. Biological networks are those built by nature in its indefatigable fight to turn life possible: the genetic regulatory network for the expression of a gene [6], blood vessels [7], food webs [8] and metabolic pathways [9]. Technological or communications networks are those constructed by man in its indefatigable fight to turn life good: electric power grid [4,10], airline routes [10], railways [11], internet [12] and the World Wide Web [13].

A network is a set of vertices or nodes provided with some rule to connect them by edges. The degree of a vertex is defined as being equal to the number of edges connected to that vertex. In order to characterize a network, six important quantities or properties can be calculated [1,2]: the degree distribution, the clustering coefficient, the assortativity coefficient, the average shortest path length, the betweenness and the robustness to a failure or attack. The first four quantities appear in this work and their meanings are explained below.

In this report, we study a very peculiar network: the Brazilian soccer network. Using the information at our disposal [14], a bipartite network is constructed with two types of vertices: one composed by 127 clubs (teams) and the other formed by 13 411 soccer players. They correspond to the total number of clubs and soccer players that have sometime participated in the Brazilian soccer championship during the period 1971–2002 [15]. Whenever a soccer player has been employed by a certain club, we connect them by an edge.

Figure 1 shows the number of clubs N_C versus G/M, which stands for goals by match and it is equal to the total number of goals (G) scored by a club, divided by the number of matches (M) disputed by that club. Clearly, the data are well fitted by a Gaussian curve centered at $G/M \simeq 1.03$. The Brazilian club with the best index is São Caetano $(G/M = 1.73)$ and the worst is Colatina with $G/M = 0.22$.

Figure 2 plots the degree distributions for each kind of vertex of the bipartite network: players and clubs. The player probability $P(N)$ exhibits an exponential decay with the player degree N. Naturally, N corresponds to the number of

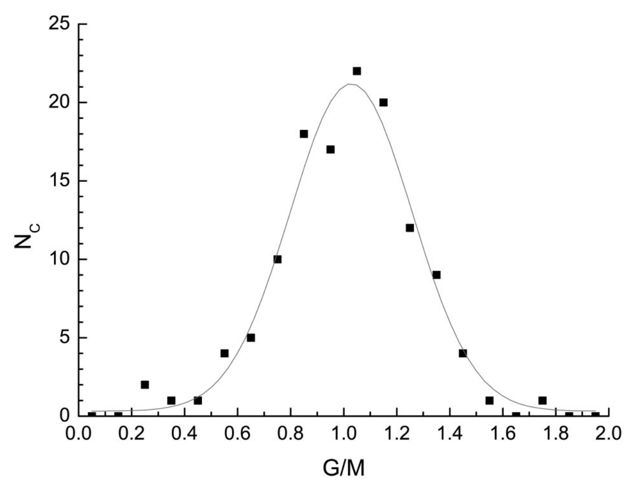

FIG. 1. Histogram of the number of clubs against the number of goals scored by match. Bins of size 0.1 were used. The full line corresponds to the fitted Gaussian curve. The average number of goals by match is equal to 1.00.

*Electronic address: onody@if.sc.usp.br
†Electronic address: pac@if.sc.usp.br

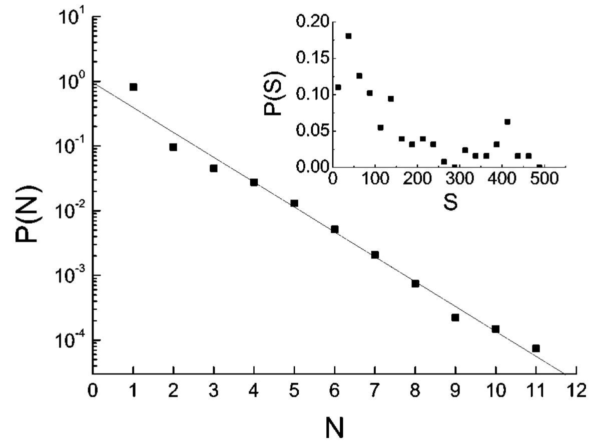

FIG. 2. Probability $P(N)$ that a player has worked for N clubs. The full line corresponds to the fitted curve $P(N) \sim 10^{-0.38N}$. So, it is 190 times more probable to find someone who has played for only two clubs than for eight clubs. The inset is the degree distribution $P(S)$ for the clubs.

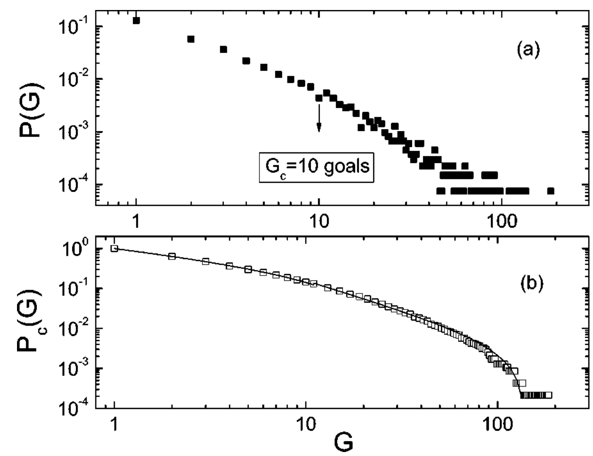

FIG. 4. (a) The goals probability $P(G)$ that a player has scored G goals. Choosing randomly, a Brazilian soccer player has ten times less chance to have scored 36 goals than 13 goals. The player with the highest score is Roberto Dinamite with $G=186$. (b) The corresponding cumulative probability distribution $P_c(G)$.

clubs in which a player has ever worked. We find the average $\overline{N}=1.37$. The most nomad player is Dadá Maravilha [16] with $N=11$. The inset shows the club probability $P(S)$ as a function of the club degree S. Regrettably, its form cannot be inferred may be because the small number of involved clubs.

A very amazing result comes out when we determine the probability $P(M)$ that a soccer player has played a total of M games (disregarding by what club). There is an elbow [see Fig. 3(a)] or a critical value at $M_c=40$ for the semi-log plot of $P(M)$. As there is a lot of scatter, we also determine the corresponding cumulative distribution $P_c(M)$ [1]. The latter distribution is very well fitted by two different exponentials: $P_c(M)=0.150+0.857\ 10^{-0.042M}$ for $M<40$ and $P_c(M)=0.410\ \ 10^{-0.010M}$ for $M>40$, as it is shown in Fig. 3(b). This implies that the original distribution $P(M)$ has also two exponential regimes with the *same* exponents [1]. The existence of the threshold M_c probably indicates that, after a player has found some fame or notoriety, it is easier to him to keep playing soccer. Surpassing this value is like the player has gained some kind of "stability" in his job.

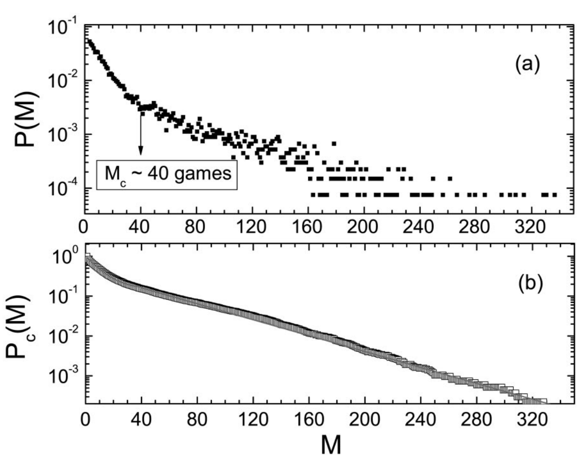

FIG. 3. (a) Game probability $P(M)$ versus the number M of disputed matches. (b) Cumulative distribution $P_c(M)$ built from $P(M)$. Fittings appear as full lines.

As the goals are the quintessence of soccer, we determine the goal's probability $P(G)$ that a player has scored G goals in the Brazilian championships. The result is shown in Fig. 4(a). Here again we find an intriguing threshold at $G_c=10$ separating regions with apparently two distinct power law exponents. Such a kind of behavior has already been found in the context of scientific collaborations network [17]. To verify if this threshold really exist (since the tail is, once more, very scattered) we calculate the corresponding cumulative distribution $P_c(G)$, plotted in Fig. 4(b). The curve we get resembles that one found in the network of collaborations in mathematics [see Fig. 3.2(a) of Ref. [1]] and it may correspond to a truncated power-law or possibly two separate power-law regimes. We have tried to fit $P_c(G)$ with truncated power laws. However, our best result was obtained using two power laws with different exponents: $P_c(G)=-0.259+1.256\ G^{-0.500}$ for $G<10$ and $P_c(G)=-0.004+4.454\ G^{-1.440}$ if $G>10$. Notice that the existence of additive constants in the power laws spoils the expected straight line characteristic in the log-log plot. Moreover, as the cumulative distribution does not preserve power law exponents, it follows that $P(G) \sim G^{-1.5}$ and $P(G) \sim G^{-2.44}$ for $G<10$ and $G>10$, respectively. We conjecture that the origin of this threshold can be simply explained by the structure, position or distribution of the players in the soccer field. Circa two thirds of the eleven players form in the defense or in the middle field. Players in these positions usually score less than those of the attack.

From the bipartite network (of players and teams), one can construct unipartite network composed exclusively by the soccer players. If two players were at the *same team* at the *same time*, then they will be connected by an edge. Let us call the resulting network the Brazilian soccer player (BSP) network. With this merging, we get a time growing network that reflects acquaintances and possible social relationships between the players. Similar merging has already been done

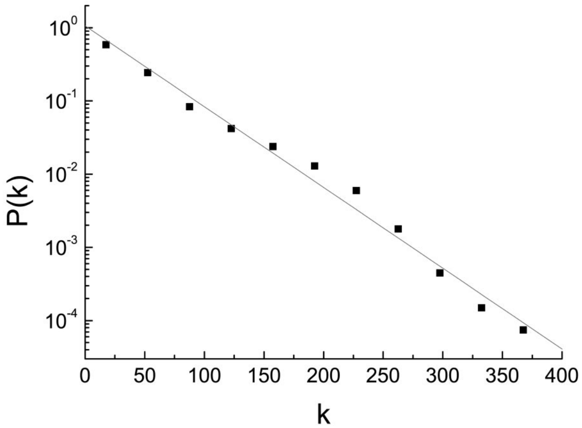

FIG. 5. Degree distribution of the BSP network. The fitting curve (full line) has the exponential form $P(k) \sim 10^{-0.011k}$.

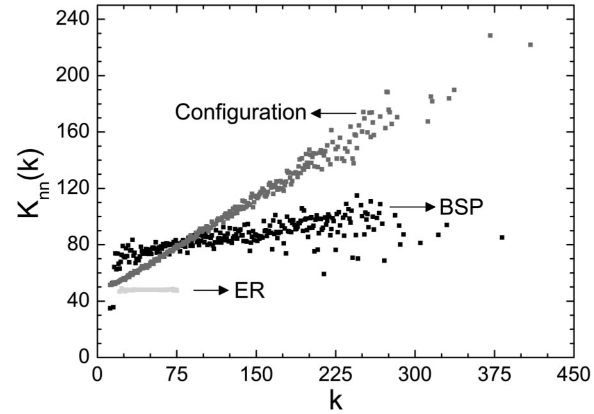

FIG. 6. Nearest-neighbors average connectivity for the ER, BSP, and Configuration model.

for the bipartite networks: actor-film [4,10], director-firm [18] and scientist-paper [17].

In the year 2002, the BSP network had 13 411 vertices (the soccer players) and 315 566 edges. The degree probability distribution $P(k)$ can be easily calculated and we obtain an average degree $\overline{k}=47.1$. The result is plotted in Fig. 5.

Many others quantities of the BSP network can be precisely evaluated. One can measure, for example, what is the probability C_i that the first neighbors of a vertex i are also connected. The average of this quantity over the whole network gives the clustering coefficient C, a relevant parameter in social networks. We find $C=0.79$, which means that the BSP is a highly clustered network. At this point, it is very interesting to compare the BSP network results with those of random graphs as the Erdös-Rényi (ER) model [19] and the configuration model [1,20]. We simulated an ER network (with the same size as the BSP) in which the vertices are connected with a probability equal to 0.00351, which gives, approximately, the same number of edges for both networks. Still keeping the network size, we also simulated the configuration model using the fitting curve of Fig. 5 as the given degree distribution. We see from Table I, that both ER and the configuration model have a small clustering coefficient as it would be expected for random networks.

TABLE I. In the first column, v is the number of vertices, e is the number of edges, $\overline{k}$ is the mean connectivity, C is the clustering coefficient, A is the assortativity coefficient, and D is the average shortest path length.

	BSP	Erdös-Rényi	Configuration
v	13 411	13 411	13 411
e	315 566	315 443	345 294
$\overline{k}$	47.1	47.1	51.5
C	0.790	0.004	0.008
A	0.12	0.00	0.46
D	3.29	2.84	2.85

The assortativity coefficient A [1] measures the tendency of a network to connect vertices with the same or different degrees. If $A>0$ ($A<0$) the network is said to be assortative (disassortative) and not assortative when $A=0$. To determine A, we first need to calculate the joint probability distribution e_{jk}, which is the probability that a randomly chosen edge has vertices with degree j and k at either end. Thus,

$$A = \frac{1}{\sigma_q^2}\sum_{jk} jk(e_{jk} - q_j q_k), \quad (1)$$

where $q_k=\Sigma_j e_{jk}$ and $\sigma_q^2=\Sigma_k k^2 q_k-(\Sigma_k kq_k)^2$. The possible values of A lie in the interval $-1 \leq A \leq 1$. For the BSP network, we find $A=0.12$ so it is an assortative network. This value coincides with that of the Mathematics coauthorship [21] and it is smaller than that of the configuration model ($A=0.46$). The explanation is very simple: although the vertices of the configuration model are in fact randomly connected, the given degree distribution constraint generates very strong correlations. This can be measured by the nearest-neighbors average connectivity of a vertex with degree k, $K_{nn}(k)$ [22], which is plotted in Fig. 6.

Finally, we also determine the average shortest path length D between a given vertex and all the others vertices of the network. Taking the average of this quantity for all BSP network vertices, we get $D=3.29$. In analogy with social networks, we can say that there are *3.29 degrees of separation* between the Brazilian soccer players or, in other words, the BSP network is a small-world.

We can also study the time evolution of the BSP network. We have verified that this network is broken in many clusters in 1971 and 1972, after that there is only one component. In Table II, we observe an increasing mean connectivity $\overline{k}$. We can think of two reasonings for that: the player's professional life is turning longer and/or the player's transfer rate between teams is growing up. On the other hand, the clustering coefficient is a time decreasing function. Also in this case, there may be two possible explanations: the player's transfer rate

TABLE II. Temporal evolution of some quantities of the BSP network. The meanings of the first column are the same as those of Table I.

	1975	1980	1985	1990	1995	2002
v	2490	6420	8797	10 329	11 629	13 411
e	48 916	128 424	181 293	219 968	254 371	315 566
$\overline{k}$	39.3	40.0	41.2	42.6	43.7	47.1
C	0.84	0.83	0.82	0.81	0.80	0.79
A	0.02	0.06	0.06	0.07	0.08	0.12
D	3.17	3.35	3.39	3.27	3.28	3.29

between national teams and the exodus of the best Brazilian players to foreigner teams (which has increased, particularly, in the last decades).

Naturally, this kind of movement diminishes the cliques probabilities. From Table II, we also see that the BSP network is becoming more assortative with time. This seems to indicate the existence of a growing segregationist pattern, where the player's transfer occurs, preferentially, between teams of the same size. Finally, the average shortest path length values may suggest that it is size independent but, most probably, this conclusion is misled by the presence of only some few generations of players in the growing BSP network.

We hope that the work presented here may stimulate further research on this subject. Some opened questions are, for instance, whether the results obtained for the Brazilian soccer held for different countries or, perhaps, for different sports.

We acknowledge financial support by the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo). We are very grateful to Marli Schapitz de Castro and Henrique Jota de Paula Freire for helping us to manipulate the database.

[1] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[2] S. N. Dorogovtsev and J. F. F. Mendes, in *Evolution of Networks* (Oxford University Press, Oxford, 2003).
[3] P. S. Dodds, R. Muhamad, and D. J. Watts, Science **301**, 827 (2003).
[4] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[5] S. Milgram, Psychol. Today **1**, 60 (1967).
[6] T. I. Lee *et al.*, Science **298**, 799 (2002).
[7] G. B. West, J. H. Brown, and B. J. Enquist, Nature (London) **400**, 664 (1999).
[8] J. Camacho, R. Guimerá, and L. A. N. Amaral, Phys. Rev. Lett. **88**, 228102 (2002).
[9] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L Barabási, Science **297**, 1551 (2002).
[10] L. A. N. Amaral, A. Scala, M. Bartheélémy, and H. E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **97**, 11149 (2000).
[11] P. Sem *et al.*, Phys. Rev. E **67**, 036106 (2003).
[12] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).
[13] A.-L. Barabási, R. Albert, and H. Jeong, Physica A **272**, 173 (1999).
[14] Data were gathered from the CD-ROM released by the magazine Placar (Editora Abril, Brazil, 2003).
[15] The first official Brazilian soccer championship was disputed in 1971 and since then it is managed by the CBF (Confederação Brasileira de Futebol).
[16] We preferred to use the nicknames so that the fans can easily recognize their idols. Few people knows that Pelé's name is Édson Arantes do Nascimento.
[17] M. E. J. Newman, Phys. Rev. E **64**, 016131 (2001).
[18] G. F. Davis, M. Yoo, and W. E. Baker, Strat. Org. **1** (3), 301 (2003).
[19] P. Erös and A. Rényi, Publ. Math. (Debrecen) **6**, 290 (1959).
[20] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[21] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[22] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Phys. Rev. Lett. **87**, 258701 (2001).