Anomalous diffusion and long-range correlations in the score evolution of the game of cricket

Haroldo V. Ribeiro,[1,2,*] Satyam Mukherjee,[2] and Xiao Han T. Zeng[2]

[1] *Departamento de Física and National Institute of Science and Technology for Complex Systems, Universidade Estadual de Maringá, Maringá, PR 87020, Brazil*

[2] *Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA*

(Received 12 March 2012; published 13 August 2012)

We investigate the time evolution of the scores of the second most popular sport in the world: the game of cricket. By analyzing, event by event, the scores of more than 2000 matches, we point out that the score dynamics is an anomalous diffusive process. Our analysis reveals that the variance of the process is described by a power-law dependence with a superdiffusive exponent, that the scores are statistically self-similar following a universal Gaussian distribution, and that there are long-range correlations in the score evolution. We employ a generalized Langevin equation with a power-law correlated noise that describes all the empirical findings very well. These observations suggest that competition among agents may be a mechanism leading to anomalous diffusion and long-range correlation.

DOI: 10.1103/PhysRevE.86.022102 PACS number(s): 02.50.−r, 05.45.Tp, 89.20.−a

Diffusive motion is ubiquitous in nature. It can represent how a drop of ink spreads in water, how living organisms such as fish [1] or bacteria [2] move, how information travels over complex networks [3], and many other phenomena. One of the most common fingerprints of usual diffusion is the way that the particles or objects in question spread. The spreading can be measured as the variance of the positions of the particles after a certain period of time. For usual diffusion, the variance grows linearly in time. There are two hypotheses underlying this behavior. The first is the absence of memory along the particle trajectory; that is, the actual position of the particle can be approximated by a function of its immediately previous position (Markovian hypothesis). The second is the existence of a characteristic scale for the position increments. When these two assumptions hold, we can show that distribution of the positions will approach a Gaussian profile (central limit theorem).

Naturally, there are situations in nature that do not fit these hypotheses, and consequently, deviations from the usual behavior appear. When this happens, researchers usually report on anomalous diffusion. A well-understood case is when there is no characteristic length for the particle jumps. In this case, the variance is infinity and the distribution of the positions follows a Lévy distribution. Examples of Lévy processes include animals' movement during foraging [4], diffusion of ultracold atoms [5], and systems that are out of thermal equilibrium [6]. The situation is more complex when the diffusive process presents memory. We have many different manners of correlating the particle positions. Depending on this choice, diffusive properties such as the dependence of the variance on time can drastically change. In this context, a typical behavior for the variance is a power-law dependence with an exponent α, where $\alpha < 1$ corresponds to subdiffusion and $\alpha > 1$ to superdiffusion.

Several approaches have been proposed to investigate anomalous diffusion in general. Fractional diffusion equations [7], Fokker-Planck equations [8], and Langevin equations [9] are just a few examples of frameworks used to describe this phenomenon. However, there is a lack of empirical studies aiming to verify situations where these models can be applied and the possible mechanisms that lead to anomalous diffusion. There are a few exceptions, such as the work by Weber, Spakowitz, and Theriot [10], where they showed that the motion of chromosomal loci of two bacterial species is subdiffusive and anticorrelated, as well as the work by Lenz *et al.* [11], where they investigated the role of predation in the motion of bumblebees during foraging.

In this work, we show that the evolution of the scores in the game of cricket can be understood as a diffusive process with scale-invariance properties, anomalous diffusion, and long-range correlations. All these findings are well described by a generalized Langevin equation with a power-law correlated noise. The results presented here suggest that competition among agents may be a mechanism leading to correlation and anomalous diffusion correlation. In the following, we present our data set of scores of cricket matches, a diffusive interpretation of the evolution of these scores, a generalized Langevin equation for modeling the empirical findings, and, finally, some concluding remarks.

The game of cricket is the second most popular sport in the world, after soccer. It is a "bat-and-ball" game (similar to baseball) played between two teams of 11 players. There are three types of cricket games, which differ in length. "Twenty20" (T20) cricket is the shortest one, lasting approximately 3 h; "One Day International" (ODI) cricket lasts almost 8 h; and "Test" cricket is the longest one, taking up to 5 days to finish. The game involves one team batting (their innings) and scoring as many points (runs) as possible and setting up a target for the opponent team. The opponent team comes in to bat and tries to exceed the target. A team's inning is terminated whenever it exceeds the quota of *overs* (six consecutive balls bowled in succession) or when the team has lost 10 *wickets* (wooden stumps used as a target for bowling). The maximum limit is 20 overs for T20 cricket, 50 overs for ODI cricket, and 200 overs for Test cricket.

Surprisingly, the record of a game of cricket (score cards) includes not only the game outcome, but also the event-by-

[*] hvr@dfi.uem.br

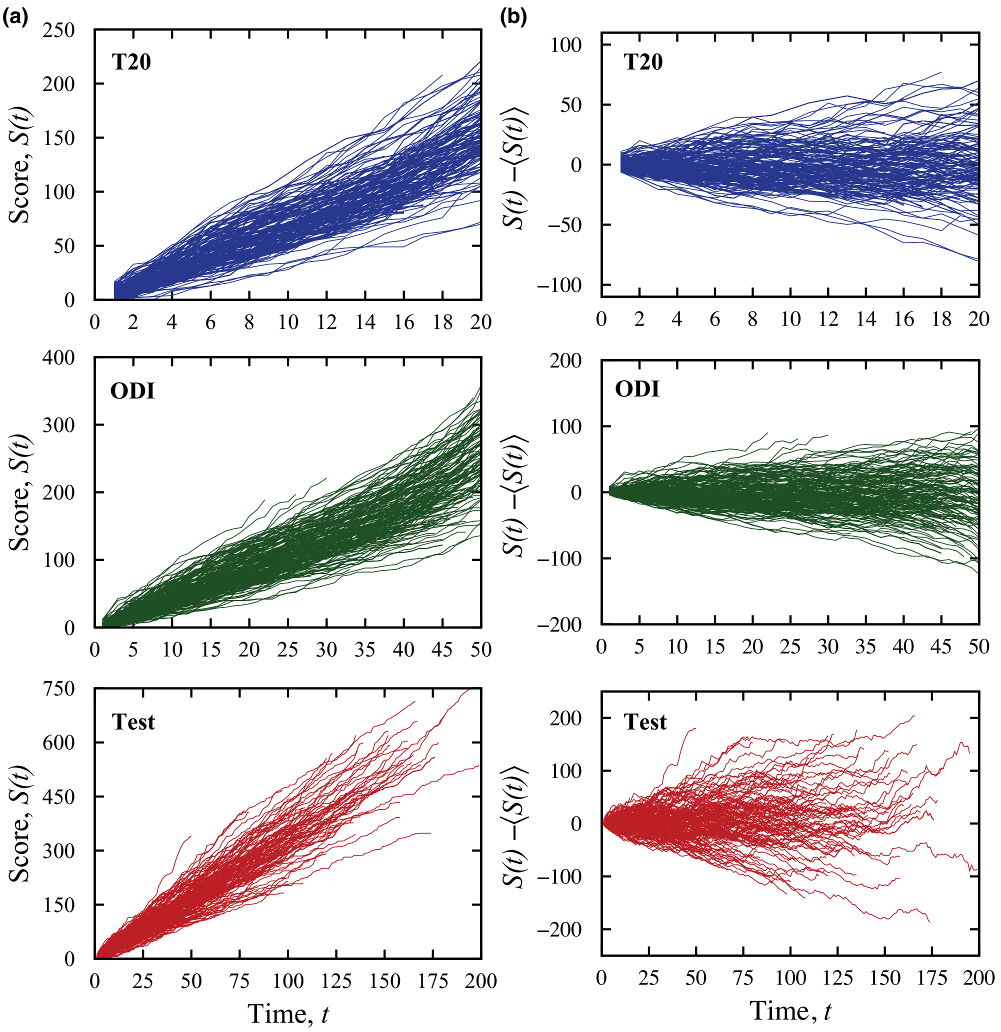

FIG. 1. (Color online) Evolution of scores $S(t)$ for the three types of cricket. To make the notation easier, we have denoted the event-by-event evolution as time evolution. The main difference between these types is the maximum length of the game: 20 time steps for T20, 50 for ODI, and 200 for test. (a) Evolution of the scores of a hundred games selected at random from our database; (b) this evolution after removing the mean tendency to increase, that is, $S(t) - \langle S(t) \rangle$.

event evolution of the scores. We collect the information of scores per over for T20 (2005–2011), ODI (2002–2011), and Test cricket (2002−2011) from the *cricinfo* Web site [12]. Using these data, we create 2144 time series of scores where the time t represents a completed *over*. In Fig. 1(a), we show the temporal dependence of the scores $S(t)$ for 100 games from the three types of cricket selected at random from our database. We note the natural increasing tendency of the scores and also the erratic movement around the mean tendency. For better visualization of these fluctuations, we plot in Fig. 1(b) the scores after subtracting the mean tendency $\langle S(t) \rangle$ from $S(t)$.

We start by investigating how the mean value of the scores depends on time [Fig. 2(a)]. These plots reveal that the mean score $\langle S(t) \rangle$ grows linearly in time for the three types of game. The only difference is in the rate of growth, which is 6.4 ± 1.0 for T20, 3.9 ± 1.0 for ODI, and 3.3 ± 1.0 for Test. The different values show that the overall performance of teams is related to the length of the game. T20 cricket (which lasts ~3 h) and ODI cricket (which lasts ~8 h) have the highest rates, indicating that the players work hard to score as many points as they can, while in Test cricket the players may prefer to reserve their efforts, since Test matches are quite long.

Next we characterize the spreading process by evaluating the variance of the scores as a function of time [Fig. 2(b)]. We show the variance $\sigma^2(t) = \langle [S(t) - \langle S(t) \rangle]^2 \rangle$ in a log-log plot, where we observe a nonlinear increase in $\sigma^2(t)$. By least squares fitting a linear model to these log-log data, we find a superdiffusive regime, that is, $\sigma^2(t) \propto t^\alpha$ with $\alpha \approx 1.3$ for the three types of cricket. This intriguing feature suggests that the competition within the game may drive the scores to spread more rapidly than a regular Brownian motion.

Another interesting question is whether the distribution of the scores is self-similar and whether these distributions follow a particular functional form. To answer this question, we calculate the cumulative distribution functions of the scores for each time step. Figure 3(a) shows these distributions for several values of t and for the three types of cricket. We note the shift of the distributions towards positive values and the increase in

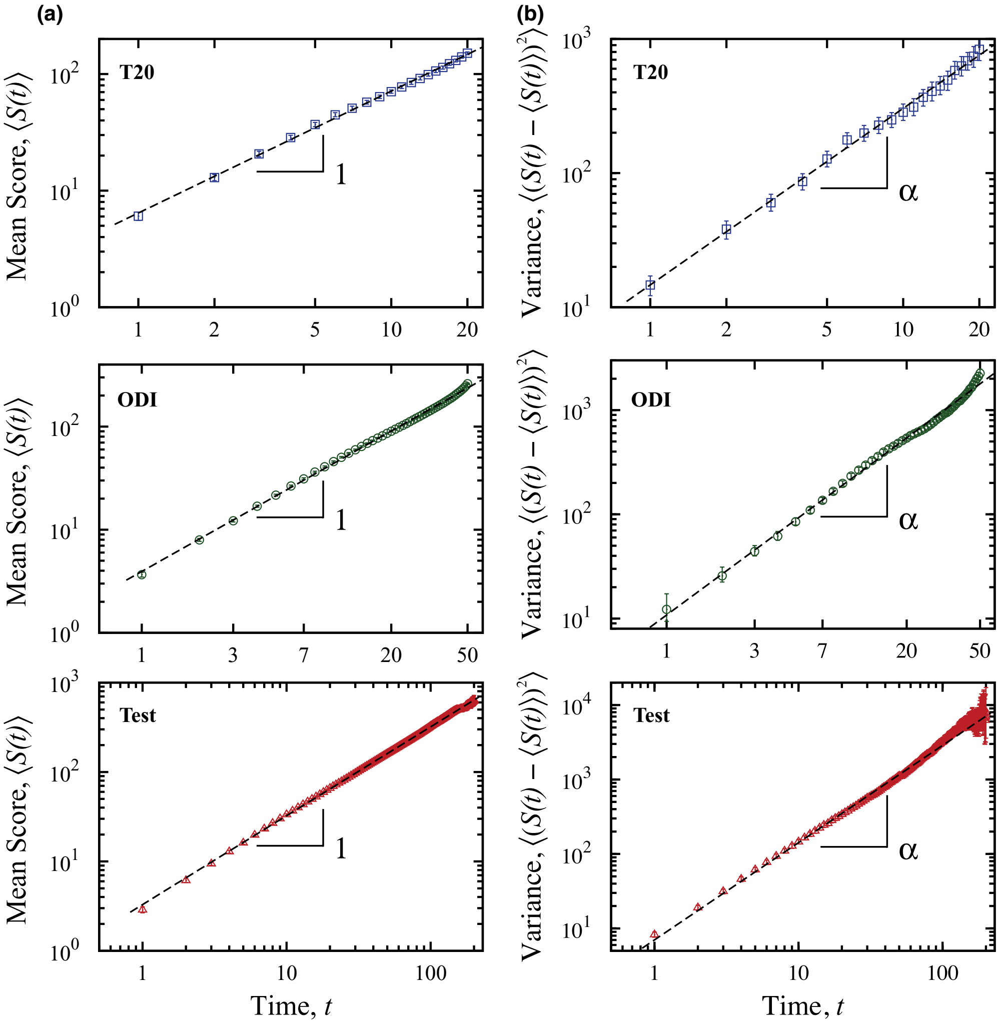

FIG. 2. (Color online) Anomalous diffusion of the scores. (a) Mean value of the scores as a function of time, $\langle S(t)\rangle$, for the three types of cricket. Dashed lines are linear fits to each data set. We find the mean values to grow linearly with time. (b) Spreading of the score trajectories measured as the variance $\sigma^2(t) = \langle [S(t) - \langle S(t)\rangle]^2\rangle$ versus time. Dashed lines are power-law fits to the variance, where we find the exponents $\alpha = 1.32 \pm 0.02$ for T20, $\alpha = 1.31 \pm 0.02$ for ODI, and $\alpha = 1.30 \pm 0.02$ for Test cricket. Since $\alpha > 1$, the diffusive process underlying the evolution of scores is superdiffusive. The error bars are 95% confidence intervals obtained via bootstrapping [13].

the distribution width. Moreover, these semilog plots indicate that the distributions are close to normal distributions.

To check the normality and self-similarity, we evaluate the distribution of the normalized scores $\xi(t) = \frac{S(t)-\langle S(t)\rangle}{\sigma(t)}$, where $\langle S(t)\rangle$ is the mean value of the score and $\sigma(t)$ is the standard-deviation. As shown in Fig. 3(b), the distributions exhibit a good collapse and a profile that is very close to a Gaussian distribution. These results are also supported by the insets in Fig. 3(b), where we plot the p-values for the Pearson chi-square test as a function of time. We note that the normality is rejected for small values of $t \lesssim 90$ due to the discrete nature of $S(t)$ and also the asymmetry in the score system. After enough time ($t \sim 90$), the p-values are >0.1 and we can not reject the Gaussian hypothesis in Test cricket.

We now focus on a correlation analysis to answer whether the score evolution is a Markovian process. To investigate this hypothesis, we select all games from Test cricket that are longer than 120 time steps, totaling 431 games. For this subset, we calculate the time series of the score increments $\Delta S(t) = S(t+1) - S(t)$. Next we employ detrended fluctuation analysis (DFA) to obtain the Hurst exponent h. DFA [15,16] consists of four steps. (i) First, we define the profile $Y(i) = \sum_{k=1}^{i} \Delta S(t) - \langle \Delta S(t)\rangle$. (ii) Next we cut $Y(i)$ into $N_n = N/n$ nonoverlapping segments of size n, where N is the length of the series. (iii) For each segment, a local polynomial trend (here we have used a linear function) is calculated and subtracted from $Y(i)$, defining $Y_n(i) = Y(i) - p_\nu(i)$, where $p_\nu(i)$ represents the local trend in the νth segment. (iv) Finally, we evaluate the root-mean-square fluctuation function $F(n) = [\frac{1}{N_n}\sum_{\nu=1}^{N_n}\langle Y_n(i)^2\rangle_\nu]^{1/2}$, where $\langle Y_n(i)^2\rangle_\nu$ is the mean-square value of $Y_n(i)$ over the data in the νth segment. For self-similar time series, the fluctuation function $F(n)$ displays a power-law dependence on the time scale n, that is, $F(n) \sim n^h$, where h is the Hurst exponent. Intriguingly, we find that the Hurst exponent does not depend on the game and that it has a mean value equal to $h = 0.63 \pm 0.01$ [Fig. 4(a)]. This result shows that there is long-range memory in the score evolution, and therefore it is a non-Markovian process.

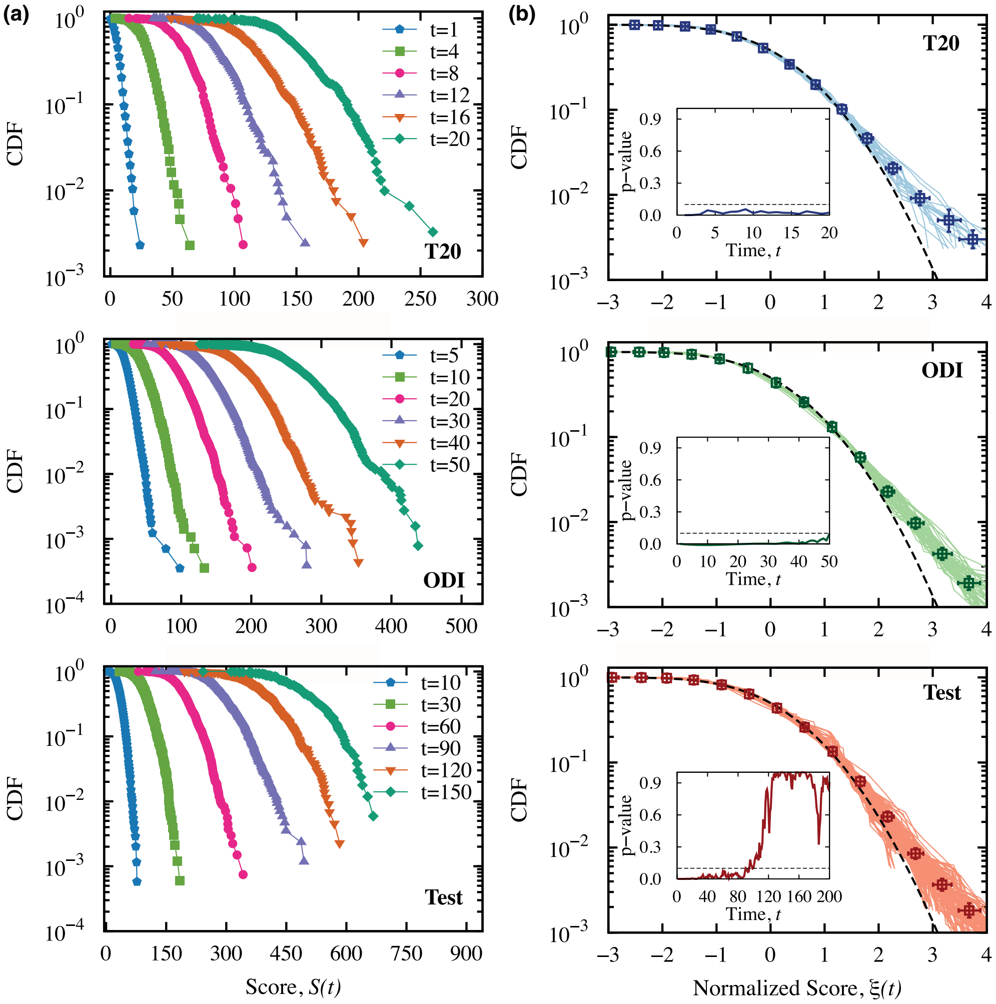

FIG. 3. (Color online) Scale invariance of the scores. (a) Evolution of the cumulative distribution function (CDF) for the three types of cricket and for different values of time t. Note that the distributions shift towards positive values of the score and that the width of the CDFs increases. (b) Scale invariance of the scores. We evaluate the CDF using normalized scores $\xi(t) = \frac{S(t)-\langle S(t)\rangle}{\sigma(t)}$, where $\langle S(t)\rangle$ is the mean value of the scores and $\sigma(t)$ is the square root of the variance of the scores. Note that the good collapse of the distributions indicates that the scores present scaling properties, i.e., after normalization they follow the same universal distribution. In these plots, the solid lines are the CDFs for each value of t and the symbols are the averaged values of these CDFs. The error bars are 95% confidence intervals obtained via bootstrapping [13]. We note, further, that these distributions are very close to a normalized Gaussian distribution (dashed lines). Insets: p-values for the Pearson chi-square test [14] as a function of time. The dashed line is the threshold 0.1 for rejecting the Gaussian hypothesis. Note that the normality is rejected for small values of t because of the discrete values of $S(t)$ and also the asymmetric initial condition of the diffusive process. After enough time ($t \sim 90$), we cannot reject the Gaussian hypothesis in Test cricket.

Moreover, the value of $h > 0.5$ indicates the existence of a persistence behavior in the scores' increments, that is, positive values are followed by positive values and negative values are followed by negative values much more frequently than by chance.

To model the previous empirical findings, we consider the following generalized Langevin equation to describe the score evolution of Test cricket:

$$\frac{d^2S(t)}{dt^2} + \int_0^t \lambda(t-\tau)\frac{dS(\tau)}{d\tau}d\tau + K = \xi(t).$$

Here, $\lambda(t-\tau)$ is the retarded effect of the frictional force, K is a drift constant, and $\xi(t)$ represents a Gaussian stochastic force. Because we know that long-range correlations are present in our system, we consider that $\xi(t)$ is also power law correlated, that is, $\langle\xi(0)\xi(t)\rangle \sim t^{-\alpha}$. We also assume that $\lambda(t) \propto \langle\xi(0)\xi(t)\rangle$ in order to satisfy the fluctuation-dissipation theorem [17]. This equation was presented in Refs. [9], [18], and [19] for $K = 0$ and it can be solved by using the Laplace transform. Indeed, after some calculations, we can show that the mean score is linear in time $\langle S(t)\rangle \sim t$, the variance obeys a power-law relationship $\langle[S(t) - \langle S(t)\rangle]^2\rangle \sim t^{-\alpha}$, and the distribution of the scores is Gaussian. Remarkably, these are exactly the same features that our empirical data present (see Figs. 2 and 3).

Furthermore, we calculate the autocorrelation function of the score "velocity" $\langle V(0)V(t)\rangle \sim t^{\alpha-2}$, where $V(t) = dS(t)/dt$ and $\alpha \neq 1$. Note that this derivative corresponds to the score increments $\Delta S(t) = S(t+1) - S(t)$ in the discrete case. Thus, the Langevin equation also predicts the existence

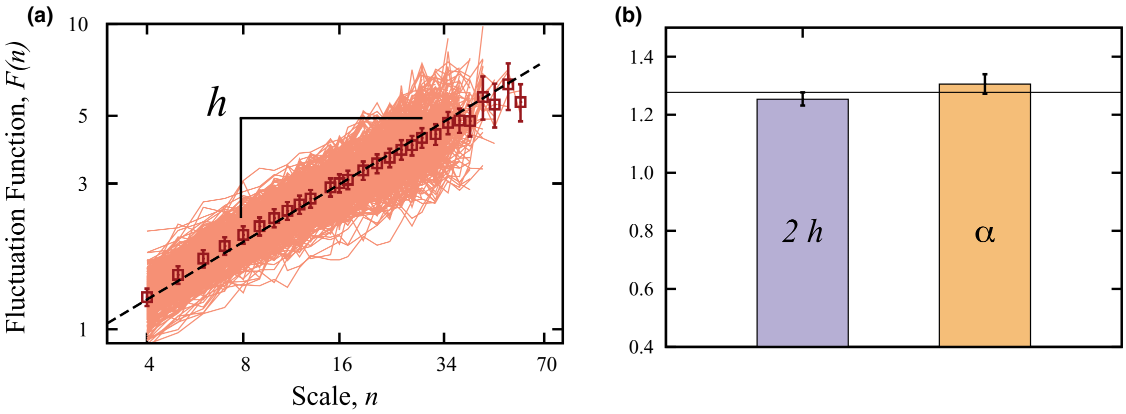

FIG. 4. (Color online) Long-range correlations in scores. (a) Detrended fluctuation analysis of score increments $\Delta S(t) = S(t+1) - S(t)$. We show the fluctuation functions $F(n)$ versus the scale n (solid lines) for all games of Test cricket that are longer than 120 units of time (431 games). Note that $F(n)$ follows a power law, where the exponent h is the Hurst exponent. We estimate the mean value of the Hurst exponent to be $h = 0.63 \pm 0.01$, and the dashed line is a power law with this exponent. We find the average value of the Pearson linear correlation coefficient to be 0.89 ± 0.02, which enhances the quality of the power-law relationships. Symbols represent average values of the fluctuation functions, and error bars are standard errors of the means. (b) Comparison of the model prediction, that is, $\alpha = 2h$, for Test cricket. The left bar shows the empirical value of $2h$ and the right bar shows the value of α. Error bars are 95% confidence intervals and the horizontal line is the upper limit of the confidence interval for $2h$. We note the existence of overlap in the confidence intervals, indicating that the relation $\alpha = 2h$ holds.

of long-range memory in the score increments. We can check this prediction by observing that the power-law exponent of the autocorrelation function is related to the Hurst exponent [16], which consequently leads to a relationship between the Hurst and the diffusive exponent $\alpha = 2h$. Figure 4(b) shows a bar plot that compares the value of $2h$ (left bar) with the value of α (right bar) for Test cricket. We note that these values are close to each other and that there is overlapping between the confidence intervals. Therefore, the relationship $\alpha = 2h$ applies.

In summary, we have studied the score evolution of the game of cricket as a diffusive process. Our analysis reveals that the mean score grows linearly in time, while the variance of the scores has a power-law dependence in time with a superdiffusive exponent. We show that the scores are statistically self-similar and follow a universal distribution approximated by a Gaussian. By using DFA, we point out that this diffusive process is non-Markovian since the scores increments are long range correlated. It is noteworthy that the persistent long-range memory present in the diffusive process can be related to the "hot hand" phenomenon in sports. Since the seminal work of Gilovich *et al.* [20] there has been an historical debate on whether "success breeds success" or "failure breeds failure" in the scoring process of many sports [21,22]. Here, the long-range persistent behavior in the score evolution not only indicates the existence of this phenomenon in cricket, but also suggests that this phenomenon can act over a very long temporal scale. Because of the long-range memory, we proposed to model the empirical findings using a generalized Langevin equation driven by a power-law-correlated stochastic force. The correlation in the noise term induces the faster-than-regular spreading of the diffusive process and also gives rise to correlations in the score increments. The results of this model show that there is a simple relation between the diffusive exponent α and the Hurst exponent h, which we have verified to hold in the empirical data. We are optimistic that the discussion presented may be applied to other sports, where new analyses can reveal more complex diffusive patterns to be compared with the increasing number of theoretical results on anomalous diffusion.

We thank the *cricinfo* Web site for making the data set used here publicly available. Fruitful discussions with members of the Amaral Lab are also gratefully acknowledged. H.V.R. is grateful to CAPES for financial support under Grant No. 5678-11-0.

[1] G. T. Skalski and J. F. Gilliam, Ecology **81**, 1685 (2000).
[2] T. Ishikawa, N. Yoshida, H. Ueno, M. Wiedeman, Y. Imai, and T. Yamaguchi, Phys. Rev. Lett. **107**, 028102 (2011).
[3] J. L. Iribarren and E. Moro, Phys. Rev. Lett. **103**, 038702 (2009).
[4] N. E. Humphries *et al.*, Nature (London) **465**, 1066 (2010).
[5] Y. Sagi, M. Brook, I. Almog, and N. Davidson, Phys. Rev. Lett. **108**, 093002 (2012).
[6] F. Bardou, J. Bouchaud, A. Aspect, and C. Cohen-Tannoudji, *Lévy Statistics and Laser Cooling* (Cambridge University Press, Cambridge, 2002).
[7] R. Metzler and J. Klafter, Phys. Rep. **339**, 1 (2000).
[8] S. A. Trigger, J. Phys. A **43**, 285005 (2010).

[9] S. C. Lim and S. V. Muniandy, Phys. Rev. E **66**, 021114 (2002).

[10] S. C. Weber, A. J. Spakowitz, and J. A. Theriot, Phys. Rev. Lett. **104**, 238102 (2010).

[11] F. Lenz, T. C. Ings, L. Chittka, A. V. Chechkin, and R. Klages, Phys. Rev. Lett. **108**, 098103 (2012).

[12] www.espncricinfo.com; accessed February 2012.

[13] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, New York, 1993).

[14] P. E. Greenwood and M. S. Nikulin, *A Guide to Chi-Squared Testing* (Wiley, New York, 1996).

[15] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E **49**, 1685 (1994).

[16] J. W. Kantelhardt, E. Koscielny-Bunde, H. H. A. Rego, S. Havlin, and A. Bunde, Physica A **295**, 441 (2001).

[17] R. Kubo, Rep. Prog. Phys. **29**, 255 (1966).

[18] K. G. Wang and C. W. Lung, Phys. Lett. A **151**, 119 (1990).

[19] K. G. Wang, Phys. Rev. A **45**, 833 (1992).

[20] T. Gilovich, R. Vallone, and A. Tversky, Cognit. Psychol. **17**, 295 (1985).

[21] G. Yaari and S. Eisenmann, PLoS ONE **6**, e24532 (2011).

[22] G. Yaari and S. Eisenmann, PLoS ONE **7**, e30112 (2012).