Innovation and nested preferential growth in chess playing behavior

Innovation and nested preferential growth in chess playing behavior

J. I. Perotti[1], H.-H. Jo[1], A. L. Schaigorodsky[2] and O. V. Billoni[2]

[1] *Department of Biomedical Engineering and Computational Science, Aalto University School of Science
P.O. Box 12200, Finland*
[2] *Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba and Instituto de Física
Enrique Gaviola (IFEG-CONICET) - Ciudad Universitaria, 5000 Córdoba, Argentina*

Abstract – Complexity develops via the incorporation of innovative properties. Chess is one of the most complex strategy games, where expert contenders exercise decision making by imitating old games or introducing innovations. In this work, we study innovation in chess by analyzing how different move sequences are played at the population level. It is found that the probability of exploring a new or innovative move decreases as a power law with the frequency of the preceding move sequence. Chess players also exploit already known move sequences according to their frequencies, following a preferential growth mechanism. Furthermore, innovation in chess exhibits Heaps' law suggesting similarities with the process of vocabulary growth. We propose a robust generative mechanism based on nested Yule-Simon preferential growth processes that reproduces the empirical observations. These results, supporting the self-similar nature of innovations in chess are important in the context of decision making in a competitive scenario, and extend the scope of relevant findings recently discovered regarding the emergence of Zipf's law in chess.

Introduction. – In the last decades, the study of complex systems gained interest in the physics community, including studies on financial markets [1], bird flocks [2,3], the human brain [4], cities [5,6], and elections [7] among other subjects. Using methods borrowed from statistical physics, non-linear dynamics, and network theory, researchers have improved the understanding of the structure, properties, and behavior of complex phenomena [8]. However, how complex systems emerge and develop is still an important open question. Complexity results from adaptation processes through which systems acquire innovative properties [9]. Therefore, how and when innovations occur in complex phenomena constitutes an important subject to be understood.

An innovative event occurs when a new idea, product, lexicon, or gene is introduced and eventually exploited by the society or ecosystem. Novelties become innovations only when they bring utility or satisfaction to the members of the system in question. Innovation is necessary for the growth and development of organizations and economies [10], and it is believed to operate at the heart of

natural evolution [9] and language evolution [11,12]. An innovative behavior may enhance the chances of success in competitive scenarios [13]. However, it requires the investment of limited resources in testing new possibilities instead of exploiting well-known solutions. For this reason, it is important to understand how nature and society optimize between the exploration of new possibilities and the exploitation of old ones. In particular, this question is relevant in the context of human decision making when bounded rationality operates [14,15].

The game of chess [16] has been extensively studied in science [17–20], especially for decision making [17,18,21], and it can be used to study innovation phenomena. Chess is a competitive game with a game tree complexity estimated as 10^{120} different move sequences [22]. Because of such huge number, all possible games cannot be fully searched in practice, thus it is necessary to continuously explore new possible moves. This implies that bounded rationality and innovation should operate. In addition, the quantitative characterization of innovation phenomena has been difficult mainly due to the lack of reliable
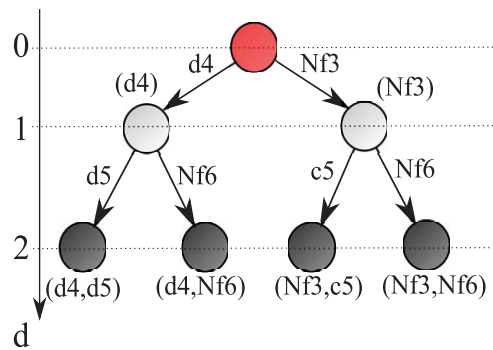
Fig. 1: (Color online) Schematic representation of the chess game tree. The root (red vertex on top) represents the initial position on the board. The white (black) vertices correspond to positions after the white (black) player moved. Each directed edge corresponds to a move. After each move, the game depth d increases by one. Each edge (vertex) is labeled by a move (a move sequence) using algebraic chess notation. Note that the same move, such as Nf6, may appear more than once in different edges.

datasets. This difficulty could be circumvented by the recent appearance of extensive records of chess games.

In a recent work, Blasius and Tönjes [17] analyzed a dataset of chess games by mapping move sequences or games to vertices in a game tree [22]. In particular, they studied how frequently each vertex in the tree is visited by a game, showing that the distribution of frequencies follows Zipf's law [23,24] over six orders of magnitude. They explained this finding using a multiplicative process with fragmentation that can be treated analytically. In their approach, each game develops following the branches of already explored parts of the game tree, and thus they focused on the dynamics within games. However, the characterization of the growing dynamic of game tree in itself is still an open problem and its characterization is crucial for the understanding of the role of innovation in chess.

In this letter, we investigate the evolution of the game tree to find the underlying processes which describe how innovation works in chess. Questions such as when and where new vertices appear in the already explored game tree, or how the associated frequencies grows, are empirically addressed and answered. These investigations have allowed us to propose a simple generative mechanism reproducing the observations of an extensive chess database.

Results. – We analyze a dataset including around 1.4 million chess games played between 1998 and 2007 in *ChessDB* [25]. Each possible move sequence corresponds to one directed path in a game tree, starting from the root vertex or the initial position of the game. See fig. 1. The moves are represented by edges. Note that there is one-to-one correspondence between move sequences and vertices in the game tree. We grow an initially empty game tree by sequentially adding the chronologically sorted games in the dataset. The depth d of a vertex is its distance from
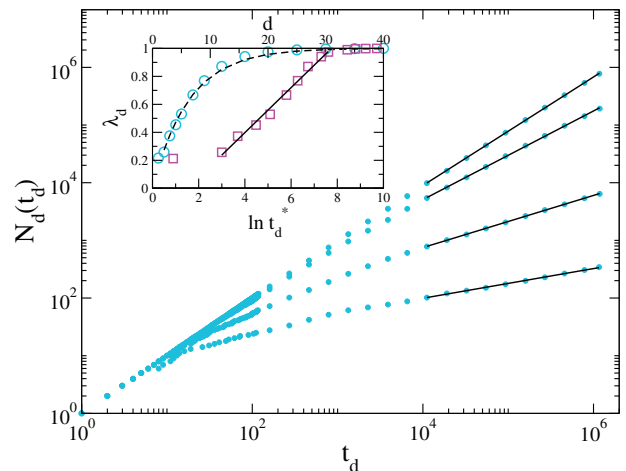

Fig. 2: (Color online) Tree width N_d *vs.* the number of games t_d that reached depth d for the game tree in the chess dataset (cyan circles). We only show the cases of $d = 2, 4, 9, 16$ for clarity and the data is logarithmically binned for $t_d > 100$. Fitted results of the form $N_d = t_d^*(t/t_d^*)^{\lambda_d}$ for large values of t_d are shown in black solid lines. Inset: the estimated values of λ_d are plotted as a function of d (cyan circles) and as a function of the estimated values of $\ln t_d^*$ (magenta squares). The black dashed line corresponds to the fit $\lambda_d = 1 - B^d$ where $B = 0.854 \pm 0.002$ ($R^2 = 0.998$), and the black solid line to $\lambda_d = a + (1/A) \ln t_d^*$ (see eq. (3)), where $a = -0.24 \pm 0.04$ and $1/A = 0.160 \pm 0.006$ ($R^2 = 0.99$).

the root. We define that an innovative event occurs whenever a game generates a new branch in the existing tree. The t-th game may generate a new branch at depth $d_b(t)$ and will end at depth $d_e(t) \geq d_b(t)$. Here t plays a role of ordinal time or time in short. Note that each game may introduce only one new branch or none, where $d_b(t)$ is not defined for the latter case.

For each innovative event occurring at depth $d_b(t)$, the tree width or number of vertices N_d at depth d evolves as $N_d(t) = N_d(t-1) + 1$ if $d_b(t) < d \leq d_e(t)$ or $N_d(t) = N_d(t-1)$, otherwise. Let $t_d(t)$ be the number of games that reached at least depth d after t games have been played. From now on, the variable t is omitted if not necessary. As shown in fig. 2, we find that

$$N_d(t_d) \simeq \begin{cases} t_d, & t_d \ll t_d^*, \\ t_d^*(t_d/t_d^*)^{\lambda_d}, & t_d \gg t_d^*, \end{cases} \qquad (1)$$

where t_d^* is a crossover value of t_d. The exponent λ_d, characterizing the innovation rate, saturates exponentially with d as

$$\lambda_d = 1 - B^d \qquad (2)$$

with $B \simeq 0.85$ (fig. 2, inset). The scaling $N_d \sim t_d^{\lambda_d}$ for $t_d \gg t_d^*$ corresponds to Heaps' law commonly found in the vocabulary growth of literary corpora or languages [26–28]. Moreover, it is found that the crossover t_d^* exponentially grows with the Heaps' exponent as

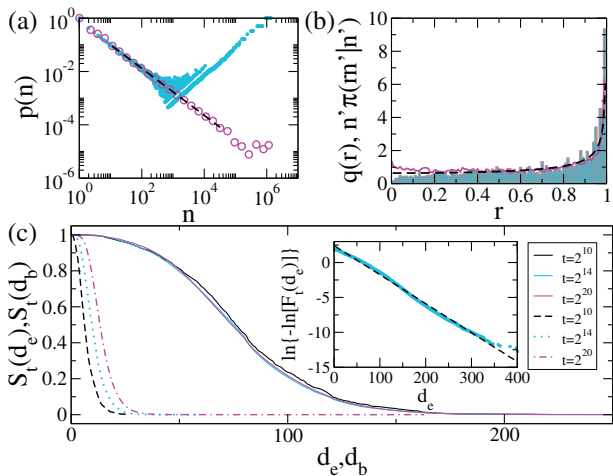$$t_d^* \sim \exp(A\lambda_d). \qquad (3)$$

Fig. 3: (Color online) (a) The probability of generating a new branch, $p(n)$, as a function of the vertex frequency n (cyan filled circles). The log-binned data (magenta circles) is fitted by $p(n) \sim n^{-\nu}$ (black dashed line) with $\nu = 0.881 \pm 0.009$ ($R^2 = 0.9998$). (b) Empirical probability density function, $q(r)$, of the frequency ratio $r = n/n'$ (cyan histogram) and its analytical expression, $q(r) = 2/[\pi\sqrt{1-r^2}]$ (black dashed line), compared with the rescaled form $n'\pi(rn'|n')$ of the conditional growth probability distribution, $\pi(n|n')$ (magenta continuous line). (c) Fraction $S_t(d_e)$ ($S_t(d_b)$) of games with game lengths (branching depths) larger or equal to d_e (d_b) after t games have been played are plotted as solid lines (broken lines). Different colors indicate different values of t. The curves of $S_t(d_e)$ collapse into one, while the curves of $S_t(d_b)$ depend on t. Inset: fit of a Gumbel distribution for maxima $F_t(d_e) = 1 - S_t(d_e) = \exp(-\exp(-(d_e - \mu_e)/\beta_e))$ (black dashed line) with $\mu_e = 58.2 \pm 0.2$ and $\beta_e = 24.1 \pm 0.1$ ($R^2 = 0.995$) for the largest-t case (cyan circles).

Double-scaling behavior in eq. (1) was also found by Gerlach and Altmann [28] in the context of vocabulary growth. They showed that this double-scaling is associated with a double-scaling in Zipf's law [29–33]. We confirm this double-scaling in the depth-dependent Zipf's laws in our dataset (not shown).

In order to understand the underlying mechanism of the evolution of the game tree, we characterize the innovation and exploitation processes. Let $n(t)$ be the frequency of occurrence of a given move sequence after t games were played. For the innovation process, the probability $p(n)$ that a game reaching a vertex with frequency n generates a new branch turns out to be

$$p(n) \simeq n^{-\nu}, \qquad (4)$$

where $\nu \simeq 0.88$ (fig. 3(a)). For the exploitation process, we study the conditional probability, $\pi(n|n')$, that a game follows an existing edge from a vertex with frequency n' to one of its child vertices which has frequency n. We measure $\pi(n|n')$ as the games traverse the edges of the growing tree. For instance, fig. 3(b) (continuous magenta line) shows $\pi(n|n' = 100)$ obtained from moves with depths

$d \geq 5$ in order to assert an adequate statistics. Similar results are obtained for other values of n' (not shown). We found that $\pi(n|n')$ is an homogeneous function satisfying

$$\pi(n|n') = \frac{1}{n'} q\left(\frac{n}{n'}\right), \qquad (5)$$

where $q(r)$ is the probability density function of frequency ratios, $r = n/n'$ (fig. 3(b), cyan histogram). The probability density, $q(r)$, is measured along the growing process of the tree. Moves ending in vertices with $n < 100$ or $d < 5$ are discarded from the statistics in order to avoid discretization effects in the shape of $q(r)$. The functional form $q(r) = 2/[\pi\sqrt{1-r^2}]$ (fig. 3(b), black dashed line) was previously determined by Blasius and Tönjes [17] measuring the values of r as the games traverse the already grown tree. As $q(r)$ is an increasing function of r, branches with larger frequencies are played more frequently, and therefore it corresponds to a preferential growth process. However, it is different from the typical case where the preferential growth/attachment probability grows linearly with the frequency [34–38]. We remark that the scaling behaviors of eqs. (4) and (5) have been measured over the whole tree, and therefore corresponds to the behavior of a depth independent process. This, and the functional form of $\pi(n|n')$, evidences the self-similar nature of the evolution of the game tree.

Next, we consider the effect of the growing stage of the tree on the statistical properties of the game length d_e. The fraction $S_t(d_e)$ of games that did not end until length d_e follows a Gumbel distribution for maxima [39], and it is independent of the number of games that have been played (fig. 3(c), inset). On the other hand, the fraction $S_t(d_b)$ of games that did not branch until depth d_b does depend on t (fig. 3(c)). Therefore, it turns out that game lengths are almost independent of the growing stage of the tree.

Now we introduce theoretical considerations in order to understand the mechanism that generates the game tree. We assume that games are infinite in length and therefore $t_d = t$ for all d. This assumption is justified by our previous observation about the statistical independence between the game lengths and the tree growth. From now on, we consider the asymptotic behavior for large t. According to a depth-dependent mean-field approach, the branching factor K_d at depth d satisfies

$$K_d(t) = N_{d+1}(t)/N_d(t) \sim t^{B^d(1-B)}, \qquad (6)$$

and the average frequency per vertex at depth d is given by

$$n_d = t/N_d(t) \sim t^{B^d}. \qquad (7)$$

Combining eqs. (6) and (7) we obtain

$$K_d \sim n_d^{1-B}. \qquad (8)$$

The branching factor at depth d grows sub-linearly with the frequency n_d. The derivative of the branching factor

with respect to t leads to

$$\frac{dK_d}{dt} \sim n_d^{-B} \frac{dn_d}{dt} \sim p(n_d) \frac{dn_d}{dt}. \qquad (9)$$

In the last expression we assumed that the branching factor increases whenever a game arriving at the vertex generates a new branch. The arrival of new games at the vertex at depth d occurs at rate dn_d/dt, and the generation of new branches with probability $p(n_d)$. Therefore, the approximate relation, $\nu \simeq B$, is obtained. We should remark that our theoretical considerations correspond to a mean-field approximation, in the more general case ν and B might be different. However, as we will show with our numerical simulations, the approximation, $\nu \simeq B$, is good enough when $B \simeq 0.85$ and it is robust under model variations. The fact that the growth exponent ν is independent of d is consistent with a self-similar growth process where at each vertex the same stochastic mechanism operates. Note that the innovation rate per vertex, dK_d/dt, is not constant as opposed to the constant growth rate in the standard formulation of the Yule-Simon preferential growth process [34,35].

Based on the theoretical considerations, we propose a generative mechanism consisting of nested preferential growth processes. We grow an ensemble of one hundred game trees each generated by one million games. Then, we calculate λ_d as a function of d and t_d^* from the average of $N_d(t)$, in order to compare the results with the empirical case (see fig. 4). Each simulation starts from a tree with the root vertex only, and games are added one by one. If the $(t + 1)$-th incorporated game reaches a vertex v with frequency $n_v(t)$, the vertex v generates a new child vertex with probability $p(n_v(t))$. Otherwise, with probability $1 - p(n_v(t))$, the game continues to one of the existing child vertices of v. We use two types of preferential growth probabilities. A non-linear preferential growth according to eq. (5), and a linear preferential growth given by $\pi(n|n') \propto n$. More specifically, when there is not a branching event, a move from a vertex v to one of its child vertices u is performed with probability, $\pi(n_u|n_v) = q(n_u/n_v)/[\sum_{u'} q(n_{u'}/n_v)]$, for the non-linear case, and with probability, $\pi(n_u|n_v) = n_u/n_v$, for the linear case. After the game traverse the tree, the vertex frequencies in the corresponding path are increased by one. In the simulations, we choose the value of the parameter ν that provides the best prediction for $B = 0.85$ in order to reproduce the empirical case. For the non-linear preferential growth probability, the best prediction of B occurs at $\nu = 0.95$, and in the linear case at $\nu = 0.85$ (inset in fig. 4, plus signs and circles, respectively). As we previously pointed out, in both cases the approximation, $B \simeq \nu$, holds. Moreover, the scaling behavior between the crossover point t_d^* and λ_d properly reproduce the empirical case (inset in fig. 4, crosses and squares, respectively). In the absence of preference, *i.e.* $\pi(n|n')$ independent of n, the simulation results deviates significantly from the empirical case (not shown). Finally, the proposed mechanism
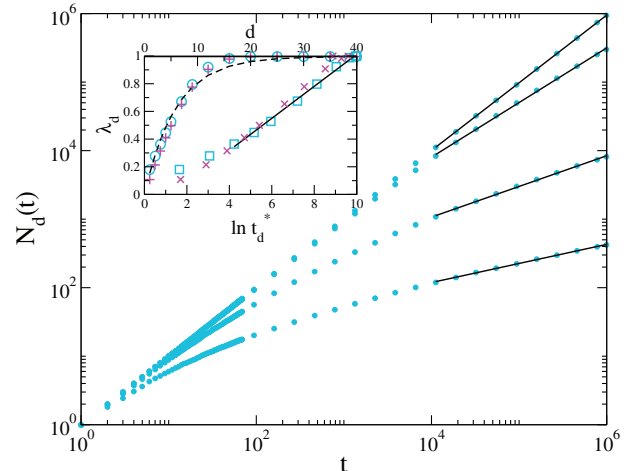

Fig. 4: (Color online) Simulation results for the models implementing the nested preferential growth mechanism. Tree width, N_d, is plotted as a function of the number of games, t, and depths, $d = 2, 4, 9, 16$, for the model with linear preferential growth probability (cyan circles). The data is logarithmically binned for $t > 100$. The black solid lines are fitted results of the form $N_d = t_d^*(t/t_d^*)t^{\lambda_d}$. Inset: the estimated values of λ_d are plotted as a function of d for the non-linear (linear) preferential growth probability using magenta plus signs (cyan circles), and as a function of the estimated values of $\ln(t_d^*)$ using magenta crossings (cyan squares). The black dashed line corresponds to $\lambda_d = 1 - B^d$ with $B = 0.846 \pm 0.005$ ($R^2 = 0.98$), and the black solid line to $\lambda_d = a + (1/A)\ln t_d^*$, where $a = -0.15 \pm 0.02$ and $1/A = 0.116 \pm 0.003$ ($R^2 = 0.996$). Both are fits to the curves of the linear preferential growth case.

turns out to be robust against variations of its details. For example, the replacement of eq. (4) by fragmentation processes [17] or the introduction of noisy selection of child vertices [30] leads to qualitatively the same results (not shown).

Discussion. – The self-similar nature of the game tree and its generative mechanism implies a lack of typical scales of the innovation phenomena in chess. Equation (4) indicates that there are no vertices with particularly large frequencies after which further innovation becomes impossible. In other words, in chess there is no winning strategy, but there is always a possibility for innovative solutions to be introduced. Moreover, the observed preferential growth mechanism suggests that the exploitation also works in a self-similar way.

Our findings in chess exhibit similarities to vocabulary growth in language evolution. The crossovers in the Heaps' law in eq. (3) have a direct interpretation in vocabulary growth. According to Gerlach and Altmann [28], such crossover is rooted in the existence of two types of words, core words and non-core words, due to a separation of time scales in the language evolution process. This suggests the existence of core and non-core move sequences in the game tree. However, notice that the dataset do not contain games from the beginning of chess-playing, but

from the year 1998. Therefore, the initial linear growth of $N_d(t)$ might be the consequence of non-realistic innovations due to random fluctuations for small values of t. Nevertheless, our simulations shown in fig. 4 do not implement a delayed beginning of the measurement process of the tree growth, and still exhibit a transition between two regimes predicting the right relationship between the crossover point t_d^* and λ_d (eq. (3)). Moreover, we simulated a delayed start of the measurement process by using the last 10^6 games generated from a simulation with 10^7 games. We found no significant differences with the results in fig. 4 (not shown). Therefore, the crossover in the Heaps' law and the existence of core and non-core moves are intrinsic properties of the the tree evolution. For the long time behavior, it has been found that the Heaps' exponent is larger in languages with a larger degree of inflection, where through declination and conjugation several words may be generated from root words [30]. This is consistent with our results if we make an analogy between the degree of inflection and a depth in the tree, because λ_d grows with d. Thus, we provide further evidence about the origin of the non-universal Zipf's exponents [33].

Conclusions. – In this work we studied how innovations are introduced into chess games by analyzing the evolution of the game tree. In our picture, move sequences are in one-to-one correspondence with the vertices in the tree. The probability that a new innovation event occurs at a vertex decays as a power law of the frequency at which the vertex is reached. Already known move sequences are played or exploited according to their frequencies, following a preferential growth process. Our model is consistent with previous results on the static properties of the already explored game tree [17], and introduces important clues about its growing dynamics. We found striking similarities between the evolution of the chess game tree and vocabulary growth. Based on our empirical findings, we proposed a generative mechanism that reproduces the observations and is robust with respect to variations of its details. All these findings provide insights into innovation phenomena in the context of decision making.

$$* * *$$

REFERENCES

[1] POTTERS M., CONT R. and BOUCHAUD J.-P., *Europhys. Lett.*, **41** (1998) 239.

[2] TONER J. and TU Y., *Phys. Rev. E*, **58** (1998) 4828.

[3] BIALEK W., CAVAGNA A., GIARDINA I., MORA T., SILVESTRI E., VIALE M. and WALCZAK A. M., *P. Natl. Acad. Sci. U.S.A.*, **109** (2012) 4786.

[4] EGUÍLUZ V. M., CHIALVO D. R., CECCHI G. A., BALIKI M. and APKARIAN A. V., *Phys. Rev. Lett.*, **94** (2005) 018102.

[5] ENCARNACAO S., GAUDIANO M., SANTOS F. C., TENEDORIO J. A. and PACHECO J. M., *Sci. Rep.*, **2** (2012).

[6] BETTENCOURT L. M. A., *Science*, **340** (2013) 1438.

[7] FORTUNATO S. and CASTELLANO C., *Phys. Rev. Lett.*, **99** (2007) 138701.

[8] AMARAL L. and OTTINO J., *Eur. Phys. J. B*, **38** (2004) 147.

[9] WAGNER A., *The Origins of Evolutionary Innovations* (Oxford University Press, Oxford) 2011.

[10] FAGERBERG J., MOWERY D. C. and NELSON R. R., *The Oxford Handbook of Innovation* (Oxford University Press, Oxford) 2005.

[11] BOBDA A. S., *World Engl.*, **13** (1994) 245.

[12] WILDGEN W., *The Evolution of Human Language: Scenarios, Principles, and Cultural Dynamics* (John Benjamins, Amsterdam) 2004.

[13] CHRISTENSEN C., *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail* (Harvard Business School Press, Boston, Mass.) 1997.

[14] SIMON H. A., *Administrative Behavior* (Macmillan, New York) 1947.

[15] ARTHUR W. B., *Am. Econ. Rev.*, **84** (1994) 406.

[16] MURRAY H. J. R., *A History of Chess* (Benjamin Press, Northampton, Mass.) 1986.

[17] BLASIUS B. and TÖNJES R., *Phys. Rev. Lett.*, **103** (2009) 218701.

[18] SIGMAN M., ETCHEMENDY P., SLEZAK D. F. and CECCHI G. A., *Front. Neurosci.*, **4** (2010).

[19] RIBEIRO H. V., MENDES R. S., LENZI E. K., DEL CASTILLO-MUSSOT M. and AMARAL L. A., *PLoS ONE*, **8** (2013) e54165.

[20] SCHAIGORODSKY A. L., PEROTTI J. I. and BILLONI O. V., *Physica A*, **394** (2014) 304.

[21] DE LAFUENTE V., *Front. Neurosci.*, **5** (2011).

[22] SHANNON C. E., *Philos. Mag.*, **41** (1950) 256.

[23] ZIPF G. K., *Human Behavior and The Principle of Least Effort* (Addison-Wesley, Cambridge, Mass.) 1949.

[24] NEWMAN M. E., *Contemp. Phys.*, **46** (2005) 323.

[25] Original data at http://chessdb.sourceforge.net. Filtered data at http://perottijuanignacio.com.ar.

[26] HEAPS H. S., *Information Retrieval: Computational and Theoretical Aspects* (Academic Press, New York, NY) 1978.

[27] SERRANO M. Á., FLAMMINI A. and MENCZER F., *PLoS ONE*, **4** (2009) e5372.

[28] GERLACH M. and ALTMANN E. G., *Phys. Rev. X*, **3** (2013) 021006.

[29] VAN LEIJENHORST D. C. and VAN DER WEIDE T. P., *Inf. Sci.*, **170** (2005) 263.

[30] ZANETTE D. and MONTEMURRO M., *J. Quant. Linguist.*, **12** (2005) 29.

[31] CATTUTO C., BARRAT A., BALDASSARRI A., SCHEHR G. and LORETO V., *Proc. Natl. Acad. Sci. U.S.A.*, **106** (2009) 10511.

[32] Lü L., Zhang Z.-K. and Zhou T., *PLoS ONE*, **5** (2010) e14139.

[33] Eliazar I., *Physica A*, **390** (2011) 3189.

[34] Yule G. U., *Philos. Trans. R. Soc. B*, **213** (1925) 21.

[35] Simon H. A., *Biometrika*, **42** (1955) 425.

[36] De Solla Price D., *J. Am. Soc. Inf. Sci*, **27** (1976) 292.

[37] Barabási A.-L. and Albert R., *Science*, **286** (1999) 509.

[38] Jeong H., Néda Z. and Barabási A.-L., *Europhys. Lett.*, **61** (2003) 567.

[39] de Haan L. and Ferreira A., *Extreme Value Theory: An Introduction* (Springer-Verlag, New York) 2006.