Zipf's Law in the Popularity Distribution of Chess Openings

Bernd Blasius[1] and Ralf Tönjes[2]

[1]*ICBM, University Oldenburg, 26111 Oldenburg, Germany*
[2]*Institut für Physik, Universität Potsdam, 14415 Potsdam, Germany and Ochadai Academic Production, Ochanomizu University,
Tokyo 112-8610, Japan*

We perform a quantitative analysis of extensive chess databases and show that the frequencies of opening moves are distributed according to a power law with an exponent that increases linearly with the game depth, whereas the pooled distribution of all opening weights follows Zipf's law with universal exponent. We propose a simple stochastic process that is able to capture the observed playing statistics and show that the Zipf law arises from the self-similar nature of the game tree of chess. Thus, in the case of hierarchical fragmentation the scaling is truly universal and independent of a particular generating mechanism. Our findings are of relevance in general processes with composite decisions.

Decision making refers to situations where individuals have to select a course of action among multiple alternatives [1]. Such processes are ubiquitous, ranging from one's personal life to business, management, and politics, and have a large part in shaping our life and society. Decision making is an immensely complex process and, given the number of factors that influence each choice, a quantitative understanding in terms of statistical laws remains a difficult and often elusive goal. Investigations are complicated by the shortage of reliable data sets, since information about human behavior is often difficult to be quantified and not easily available in large numbers, whereas decision processes typically involve a huge space of possible courses of action. Board games, such as chess, provide a well-documented case where the players in turn select their next move among a set of possible game continuations that are determined by the rules of the game.

Human fascination with the game of chess is long-standing and pervasive [2], not least due to the sheer infinite richness of the game. The total number of different games that can be played, i.e., the game-tree complexity of chess, has roughly been estimated as the average number of legal moves in a chess position to the power of the length of a typical game, yielding the Shannon number $30^{80} \approx 10^{120}$ [3]. Obviously only a small fraction of all possible games can be realized in actual play. But even during the first moves of a game, when the game complexity is still manageable, not all possibilities are explored equally often. While the history of successful initial moves has been classified in opening theory [4], not much is known about the mechanisms underlying the formation of fashionable openings [5]. With the recent appearance of extensive databases, playing habits have become accessible to quantitative analysis, making chess an ideal platform for analyzing human decision processes.

The set of all possible games can be represented by a directed graph whose nodes are game situations and whose edges correspond to legal moves from each position

(Fig. 1). Every opening is represented by its move sequence as a directed path starting from the initial node. We will differentiate between two game situations if they are reached by different move sequences. This way the graph becomes a game tree, and each node σ is uniquely associated with an opening sequence.

Using a chess database [6] we can measure the popularity n_σ or weight of every opening sequence as the number of occurrences in the database. We find that the weighted game tree of chess is self-similar and the frequencies $S(n)$ of weights follow a Zipf law [7]

$$S(n) \sim n^{-\alpha} \qquad (1)$$

with universal exponent $\alpha = 2$. Note, the precise scaling in the histogram of weight frequencies $S(n)$ and in the cumulative distribution $C(n)$ over the entire observable range [Fig. 2(a)]. Similar power law distributions with universal exponent have been identified in a large number of natural, economic, and social systems [7–15]—a fact which has come to known as the Zipf or Pareto law [7,8]. If we count only the frequencies $S_d(n)$ of opening weights n_σ after the first d moves we still find broad distributions consistent with power law behavior $S_d(n) \sim n^{-\alpha_d}$ [Fig. 2(b)]. The exponents α_d are not universal, however, but increase linearly with d [Fig. 2(b), inset]. The results are robust: similar power laws could be observed in different databases and other board games, regardless of the considered game depth, constraints on player levels or the decade when the games were played. Stretching over 6 orders of magnitude, the here-reported distributions are among the most precise examples for power laws known today in social data sets.

As seen in (Fig. 1) for each node σ the weights of its subtrees define a partition of the integers $(1 \ldots n_\sigma)$. The assumption of self-similarity implies a statistical equivalence of the branching in the nodes of the tree. We can thus define the branching ratio distribution over the real interval $r \in [0, 1]$ by the probability $Q(r|n)$ that a random pick
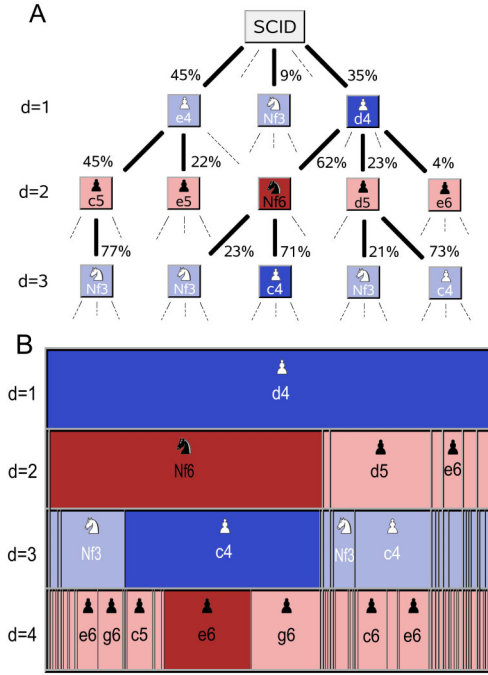
FIG. 1 (color online). (a) Schematic representation of the weighted game tree of chess based on the SCIDBASE [6] for the first three half moves. Each node indicates a state of the game. Possible game continuations are shown as solid lines together with the branching ratios r_d. Dotted lines symbolize other game continuations, which are not shown. (b) Alternative representation emphasizing the successive segmentation of the set of games, here indicated for games following a 1.d4 opening until the fourth half move $d = 4$. Each node σ is represented by a box of a size proportional to its frequency n_σ. In the subsequent half move these games split into subsets (indicated vertically below) according to the possible game continuations. Highlighted in (a) and (b) is a popular opening sequence 1.d4 Nf6 2.c4 e6 (Indian defense).
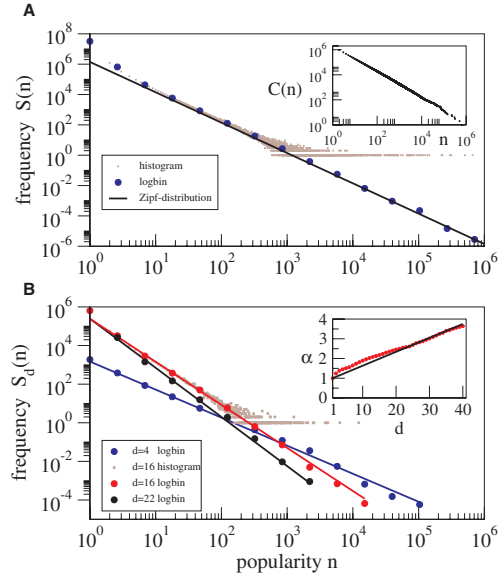

FIG. 2 (color online). (a) Histogram of weight frequencies $S(n)$ of openings up to $d = 40$ in the Scid database and with logarithmic binning. A straight line fit (not shown) yields an exponent of $\alpha = 2.05$ with a goodness of fit $R^2 > 0.9992$. For comparison, the Zipf distribution Eq. (8) with $\mu = 1$ is indicated as a solid line. Inset: number $C(n) = \sum_{m=n+1}^{N} S(m)$ of openings with a popularity $m > n$. $C(n)$ follows a power law with exponent $\alpha = 1.04$ ($R^2 = 0.994$). (b) Number $S_d(n)$ of openings of depth d with a given popularity n for $d = 16$ and histograms with logarithmic binning for $d = 4$, $d = 16$, and $d = 22$. Solid lines are regression lines to the logarithmically binned data ($R^2 > 0.99$ for $d < 35$). Inset: slope α_d of the regression line as a function of d and the analytical estimation Eq. (6) using $N = 1.4 \times 10^6$ and $\beta = 0$ (solid line).

The form of the branching ratio distribution suggests that in the case of chess there is no preferential growth process involved, but something entirely different which must be rooted in the decision process during the opening moves of a chess game [5].

In the following we show that the asymptotic Zipf law in the weight frequencies arises independently from the specific form of the distribution $q(r)$, and hence, the microscopic rules of the underlying branching process. Consider N realizations of a general self-similar random segmentation process of N integers, with paths $(\sigma_0, \sigma_1, \dots)$ in the corresponding weighted tree. In the context of chess each realization of this process corresponds to a random game from the database of N games (e.g., dark shading in Fig. 1). The weights $n_d = n_{\sigma_d}$ describe a multiplicative random process

$$n_d = N \prod_{i=1}^{d} r_i, \qquad n_0 = N, \qquad (3)$$

where the branching ratios $r_d = n_d/n_{d-1}$ for sufficiently large n_d are distributed according to $q(r)$ independent of d. For lower values of n_d the continuous branching ratio distribution is no longer a valid approximation and a

from the numbers $1 \dots n$ is in a subset of size smaller or equal to rn. Taking n to infinity $Q(r|n)$ may have a continuous limit $Q(r)$ for which we find the probability density function (PDF) $q(r) = Q'(r)$. If the limit distribution $q(r)$ of branching ratios exists it carries the fingerprint of the generating process. For instance, the continuum limit of the branching ratio distribution for a Yule-Simon preferential growth process [13] in each node of the tree would be $q(r) \sim r^\beta$, where $\beta < 0$ is a model specific parameter. On the other hand, in a k-ary tree where each game continuation has a uniformly distributed random *a priori* probability the continuum limit corresponds to a random stick breaking process in each node, yielding $q(r) \sim (1 - r)^{k-2}$. For the weighted game tree of chess $q(r)$ can directly be measured from the database [Fig. 3(a)]. We find that $q(r)$ is remarkably constant over most of the interval but diverges with exponent 0.5 as $r \rightarrow 1$, and is very well fitted by the parameterless arcsine distribution

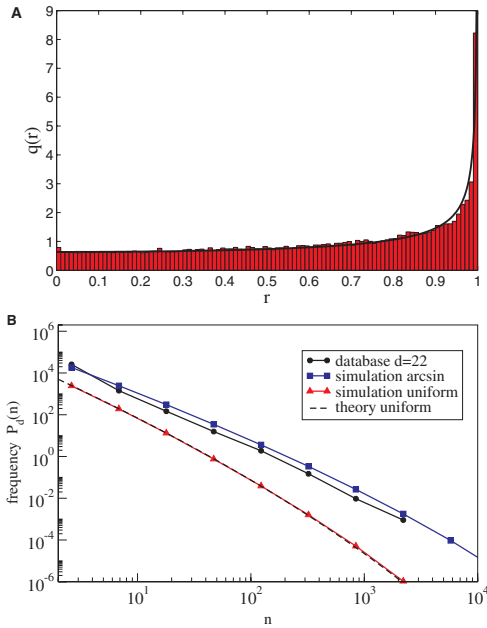$$q(r) = \frac{2}{\pi\sqrt{1 - r^2}}. \qquad (2)$$

FIG. 3 (color online).   (a) Probability density $q(r)$ of branching
ratios r sampled from all games in the Scid database with a bin
size of $\Delta r = 0.01$ and arcsine distribution Eq. (2) (black solid
line). Every edge of the weighted game tree, from nodes of size
n_{d-1} to n_d, contributes to the bin corresponding to $r = n_d/n_{d-1}$
with weight r. We disregarded clusters with $n_d < 100$ so that, in
principle, a cluster could contribute to any of the bins. We found
$q(r)$ to be depth independent. (b) Distribution of opening popu-
larities $S_d(n)$ for $d = 22$ obtained from the Scid database (black)
and from a direct simulation of the multiplicative process
Eq. (2), with branching ratios $q(r)$ taken from a uniform or
arcsine distribution. Further indicated is the theoretical result
Eq. (5) (dashed line). Similar results are obtained for other
values of d.

node of weight one has at most one subtree; i.e., the state
$n_d = 1$ is absorbing.

To calculate the PDF $p_d(n)$ of the random variable n_d
after d steps it is convenient to consider the log-
transformed variables $\nu = \log(N/n)$ and $\rho = -\log r$.
The corresponding process $\{\nu_d\}$ is a random walk $\nu_d =
\sum_{i=1}^{d} \rho_i$ with non-negative increments ρ_i and its PDF
$\pi_d(\nu)$ transforms as $n\, p_d(n) = \pi_d(\nu)$. An analytic solution
can be obtained for the class

$$q(r) = (1 + \beta)r^\beta, \qquad 0 \leq r \leq 1, \qquad (4)$$

of power law distributions, which typically arise in prefer-
ential attachment schemes. In this case the jump process ν_d
is Poissonian and distributed according to a gamma distri-
bution $\pi_d(\nu) = \frac{(1+\beta)^d}{(d-1)!}\nu^{d-1}e^{-(1+\beta)\nu}$. After retransforma-
tion to the original variables and noting that from the
probability $p_d(n)$ for a single node at distance d to the
root to have the weight n one obtains the expected number
$S_d(n)$ of these nodes in N realizations of the random
process as $S_d(n) = N p_d(n)/n$, and in particular

$$S_d(n) = \frac{(1 + \beta)^d}{N(d-1)!}\left(\log\frac{N}{n}\right)^{d-1}\left(\frac{N}{n}\right)^{1-\beta}. \qquad (5)$$

The functions $S_d(n)$ are strongly skewed and can exhibit
power law like scaling over several decades. A logarithmic
expansion for $1 < n \ll N$ shows that they approximately
follow a scaling law $S_d(n) \sim n^{-\alpha_d}$ with exponent

$$\alpha_d = (1 - \beta) + \frac{1}{\log N}(d - 1). \qquad (6)$$

The exponent α_d is linearly increasing with the game depth
d and with a logarithmic finite size correction which is in
excellent agreement with the chess database [Fig. 2(b),
inset]. Power laws in the stationary distribution of random
segmentation and multiplicative processes have been re-
ported before [9] and can be obtained by introducing slight
modifications, such as reflecting boundaries, frozen seg-
ments, merging, or reset events [16–18]. In contrast, the
approximate scaling of $S_d(n)$ in Eq. (5) is fundamentally
different, as our process does not admit a stationary distri-
bution. The exponents α_d increase due to the finite size of
the database.

As shown in Fig. 3(b) we find excellent agreement
between the weight frequencies $S_d(n)$ in the chess database
and direct simulations of the multiplicative process, Eq. (3)
using the arcsine distribution Eq. (2). If the branching
ratios are approximated by a uniform distribution $q(r) =
1$ the predicted values of $S_d(n)$ are systematically too
small, since a uniform distribution yields a larger flow
into the absorbing state $n^* = 1$ than observed in the data-
base. Still, due to the asymptotic behavior of $q(r)$ for $r \rightarrow
0$, this approximation yields the correct slope in the log-log
plot so that the exponent α_d can be estimated quite well
based on Eq. (6) with $\beta = 0$.

By observing that $S_d(n)$ in Eq. (5) is the dth term in a
series expansion of an exponential function, we find the
weight distribution in the whole game tree as $S(n) =
\sum_d S_d(n)$ to be an exact Zipf law. For branching ratio
distributions $q(r)$ different from Eq. (4) the weight fre-
quencies are difficult to obtain analytically. But using
renewal theory [19] the scaling can be shown to hold
asymptotically for $n \ll N$ and a large class of distributions
$q(r)$. For this, note that the random variable $\tau(\nu) =
\max(d : \nu_d < \nu)$ is a renewal process in ν. The expectation
$\mathbf{E}[\tau(\nu)]$ is the corresponding renewal function related to
the distributions of the ν_d as $\sum_{d=1}^{\infty} \mathrm{Prob}(\nu_d < \nu) =
\mathbf{E}[\tau(\nu)]$. If the expected value $\mu = \mathbf{E}[\rho] = \mathbf{E}[-\log r]$ is
finite and positive [e.g., for the distribution (4) $\mu =
1/(1 + \beta)$], the renewal theorem provides

$$\lim_{\nu \to \infty} \frac{d}{d\nu}\mathbf{E}[\tau(\nu)] = \frac{1}{\mu}. \qquad (7)$$

Thus, we obtain $\lim_{\nu \to \infty}\sum_{d=1}^{\infty} \pi_d(\nu) = \frac{1}{\mu}$ and finally

$$\lim_{(n/N) \to 0} S(n) = \frac{N}{\mu n^2}. \qquad (8)$$

Thus, the multiplicative random process [Eq. (3)] with any
well behaving branching ratio distribution $q(r)$ on the
interval [0, 1] always leads to an asymptotically universal
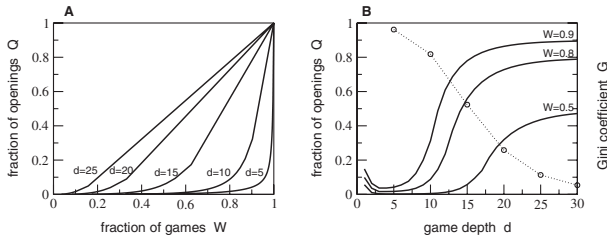scaling for $n \ll N$ [compare also the excellent fit of Eq. (8)

FIG. 4. Inequality [11,21] of the distribution $S_d(n)$. (a) Proportion W of games that is concentrating in the fraction Q of the most popular openings, for several levels of the game depth d. (b) Q as a function of d for three different values of W (solid lines) and Gini coefficient $G = 1 - 2 \int_0^1 Q(W) dW$ as a function of game depth (dotted line).

to the chess data in Fig. 2(a)]. In [15] the same Zipf-law scaling was found for the sizes of the directory trees in a computer cluster. The authors propose a growing mechanism based on linear preferential attachment. Here we have shown that the exponent $\alpha = 2$ for the weight distribution of subtrees in a self-similar tree is truly universal in the sense that it is the same for a much larger class of generating processes and not restricted to preferential attachment or growing.

There are direct implications of our theory to general composite decision processes, where each action is assembled from a sequence of d mutually exclusive choices. What in chess corresponds to an opening sequence, may be a multivariate strategy or a customized ordering in other situations. The question how such strategies are distributed is important for management and marketing [20]. One consequence of our theory is that in a process of d composite decisions the distribution $S_d(n) \sim n^{-\alpha_d}$ of decision sequences, or strategies, which occur n times shows a transition from low exponents $\alpha_d \leq 2$, where a few strategies are very common, to higher exponents $\alpha_d > 2$, where individual strategies are dominating. This is due to the divergence of the first moment in power laws with exponents smaller than 2 [11]. From [Eq. (6)] the critical number d_{cr} of decisions at which this transition occurs depends logarithmically on the sample size N and on the leading order β of $q(r)$ near zero as

$$d_{cr} = 1 + (1 + \beta) \log N. \tag{9}$$

Applied to the chess database with $N = 1.4 \times 10^6$ we obtain $d_{cr} \approx 15$ [see also Figs. 2(b), inset, and 4]. This separates the database into two very different regimes: in their initial phase ($d < d_{cr}$) the majority of chess games are distributed among a small number of fashionable openings (for $d = 12$, for example, 80% of all games in the database are concentrated in about 23% of the most popular openings), whereas beyond the critical game depth rarely used move sequences are dominating such that in aggregate they comprise the majority of all games (Fig. 4). Note, that this result arises from the statistics of iterated decisions and does not indicate a crossover of playing behavior with increasing game depth.

Our study suggests the analysis of board games as a promising new perspective for statistical physics. The enormous amount of information contained in game databases, with its evolution resolved in time and in relation to an evolving network of players, provides a rich environment to study the formation of fashions and collective behavior in social systems.

[1] H. Simon, *Administrative Behaviour* (Macmillan, New York, 1947); I. L. Janis and L. Mann, *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment* (Free Press, New York, 1977).
[2] H. J. R. Murray, *A History of Chess* (Oxford University Press, London, 2002), reprint.
[3] C. E. Shannon, Philos. Mag. **41**, 256 (1950).
[4] *The Encyclopedia of Chess Openings A–E*, edited by A. Matanovic (Chess Informant, Beograd, Serbia, 2001), 4th ed.
[5] M. Levena and J. Bar-Ilan, Computer Journal (UK) **50**, 567 (2007).
[6] Here we present results based on ScidBase (http://scid.sourceforge.net) with $N = 1.4 \times 10^6$ recorded games. Each game was uniquely coded by a string (up to $d = 40$ half moves) and the game strings were sorted alphabetically, so that all games following the same move sequence up to a given game depth d were grouped together in clusters. Popularities were obtained by counting the cluster sizes n_d.
[7] G. K. Zipf, *Human Behaviour and the Principle of Least-Effort* (Addison-Wesley, Cambridge, 1949).
[8] V. Pareto, *Cours d'Economie Politique* (Droz, Geneva, 1896).
[9] D. Sornette, *Critical Phenomena in Natural Sciences* (Springer, Heidelberg, 2003), 2nd ed.
[10] M. Mitzenmacher, Internet Math. **1**, 226 (2004).
[11] M. E. J. Newman, Contemp. Phys. **46**, 323 (2005).
[12] J. Willis and G. Yule, Nature (London) **109**, 177 (1922).
[13] H. A. Simon, Biometrika **42**, 425 (1955).
[14] R. A. K. Cox *et al.*, J. Cult. Econ. **19**, 333 (1995); S. Redner, Eur. Phys. J. B **4**, 131 (1998); A. L. Barabasi and R. Albert, Science **286**, 509 (1999); R. L. Axtell, Science **293**, 1818 (2001); X. Gabaix *et al.*, Nature (London) **423**, 267 (2003).
[15] K. Klemm *et al.*, Phys. Rev. Lett. **95**, 128701 (2005).
[16] H. Kesten, Acta Math. **131**, 207 (1973); D. Sornette Phys. Rev. E **57**, 4811 (1998); D. Sornette and R. Cont, J. Phys. I **7**, 431 (1997); S. C. Manrubia and D. H. Zanette, Phys. Rev. E **59**, 4945 (1999).
[17] P. L. Krapivsky *et al.*, Phys. Rev. E **61**, R993 (2000).
[18] J. R. Banavar *et al.*, Phys. Rev. E **69**, 036123 (2004).
[19] W. Feller, *An Introduction to Probability Theory and Its Applications* (John Wiley & Sons, New York, 1971), Vol. 2.
[20] C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More* (Hyperion, New York, 2006).
[21] J. L. Gastwirth, Rev. Econ. Stat. **54**, 306 (1972).