The Zipf law of chess openings on 44.8 million Lichess games: the exponent 2 holds, the finite-size laws do not

Qianghao Wu

November 26, 2026

Abstract

Blasius and Tönjes (2009) found that the weights of opening lines in a database of 1.4 million chess games follow a Zipf law with the universal exponent $\alpha = 2$. Their theory also predicts how the database size N changes the depth-dependent exponents: the slope of α_d against the depth d is $1/\ln N$, and α_d crosses 2 at $d_{\rm cr} = 1 + \ln N$. The source tested these finite-size laws at one value of N only. We tested all three predictions out of sample on 44,832,865 rated games from the public Lichess database (2013–2015), with a pre-registered protocol, random subsets from 10^4 games to the full set, and the authors' own branching model as a known-answer control. The pooled exponent is 2.086 at the full N, inside the pre-registered band 2 ± 0.10. The finite-size laws fail. The fitted slope falls with N as $(\ln N)^{-1.76}$, while the same estimator applied to the model gives $(\ln N)^{-0.79}$. The exponent α_d crosses 2 at depth 7.4, while the model gives 17.3. At the source's own sample size the crossing depth is 6.1, against the reported value of about 15. Both failures repeat in every speed, rating and period stratum. A placebo that shuffles moves between games also gives a pooled exponent near 2, so the value 2 alone is not specific to chess. On Lichess, α_d is concave in d: its steps are larger than $1/\ln N$ only in the first 7 or 8 plies. These are statements about online play in 2013–2015; we did not re-test the source's database. This work has not yet been reviewed by experts. The computations and the text were produced by AI models (from Anthropic and OpenAI), working under the author's direction.

Keywords: Zipf's law, chess openings, finite-size scaling, Lichess, pre-registration, replication

1 Introduction

A database of chess games defines a tree. Each node is a sequence of the first d half-moves (plies), and its weight n is the number of games that start with that sequence. Blasius and Tönjes [1] counted these weights in ScidBase, a database of $N = 1.4 \cdot 10^6$ games. They found that the number $S(n)$ of nodes with weight n, pooled over all depths, follows

$$S(n) \sim n^{-\alpha}, \qquad \alpha = 2. \tag{1}$$

They explained this law with a multiplicative fragmentation process: a node of weight n passes the fraction r of its games to a child, and r has a fixed distribution $q(r)$. For any well-behaved $q(r)$, the process gives $S(n) \to N/(\mu n^2)$ for $n/N \to 0$, with $\mu = \langle -\ln r \rangle$ (their Eq. 8). They measured an arcsine distribution, $q(r) = 2/(\pi\sqrt{1-r^2})$, and found it to be depth-independent (their Eq. 2).

The same theory predicts finite-size effects. At depth d the weights follow $S_d(n) \sim n^{-\alpha_d}$ with

$$\alpha_d = (1 - \beta) + \frac{d-1}{\ln N}, \tag{2}$$

where β is close to 0 (their Eq. 6). So α_d rises linearly in d, and the slope $1/\ln N$ falls when the database grows. The depth where α_d crosses 2 is

$$d_{\mathrm{cr}} = 1 + (1 + \beta) \ln N, \tag{3}$$

their Eq. 9. The source writes: "Applied to the chess database with $N = 1.4 \cdot 10^6$ we obtain $d_{\mathrm{cr}} \approx 15$" [1]. The value 15 requires the natural logarithm ($1 + \ln N = 15.15$).

Equations (2) and (3) make claims about how the tree changes with N. The source tested them at one N only. Maslov's Viewpoint on the paper questioned the pooling of all depths into one distribution and suggested a comparison by player rating [2]. Later work used the same 1.4-million-game database (ChessDB). Perotti, Jo, Schaigorodsky and Billoni modelled the growth of the tree with a nested preferential-growth model [3]. Schaigorodsky, Perotti and Billoni fitted $\alpha_1, \ldots, \alpha_4 = 1.10, 1.29, 1.47, 1.59$ and studied memory effects [4]. Chacoma and Billoni [5] used 3000 FIDE games and 4000 Lichess games. They drew n^{-2} lines over the popularity distributions and compared the trend of α_d visually with the slope $1/\ln N \approx 0.11$ at $N = 10^4$. Cheng et al. [6] compared Zipf exponents across rating groups in a Kaggle chess data set and found that they differ. De Marzo and Servedio [7] studied the network of players and openings on 472,183 Lichess blitz games. None of these works measures α_d or d_{cr} at several values of N. None tests Eq. (2) or Eq. (3) out of sample. We found no such test in the citation lists of the source (80 citing works in OpenAlex, 63 in Semantic Scholar) or in searches of arXiv, Zenodo and GitHub (searches of November 24, 2026).

This paper. The Lichess open database [10] holds every rated game played on the site. Its 36 monthly files for 2013–2015 contain 44.8 million games with moves. That is 32 times the source's N. We fixed the tests, the estimator and the decision thresholds in a protocol before we computed any statistic on these games. Table 1 gives the outcome. The universal exponent is supported. The two finite-size laws are refuted.

2 Data and protocol

Data. We used the 36 files `lichess_db_standard_rated_YYYY-MM.pgn.zst` for 2013-01 to 2015-12 (6.9 GB compressed, CC0). Their SHA-256 sums are in the result folder. The game count published with the files [10] is 45,043,493. Of these, 210,628 games (0.47%) have no moves. The other 44,832,865 games form the data set.

Pre-registration. The protocol was committed to the project's git repository on 2026-11-25 at 12:15 CET, before any node weight was computed. The first outcome files were committed at 12:52. The protocol was not changed later. Sections 4 and 5 report analyses that were not pre-registered, and we mark them as such.

Game key and counting. A game is reduced to its first 40 plies in standard algebraic notation (SAN), without move numbers, comments, `!` and `?`. Transpositions are different nodes, as in the source. We sort the keys and take the longest common prefix of each pair of neighbours; one pass then gives $S_d(n)$ for $d = 1, \ldots, 40$. This is the source's method. A random subset of games keeps the sorted order, so subsets need no new sort.

Test	Rule	Lichess (model)	Decision
T1: $\alpha = 2$	$\lvert\alpha_{\text{pool}} - 2\rvert \le 0.10$	$\alpha_{\text{pool}} = 2.086$	supported
T2: slope $1/\ln N$	$\lvert\gamma - \gamma_{\text{model}}\rvert \le 0.5$	$\gamma = -1.756 \pm 0.058$ (-0.788)	refuted
	$\lvert\rho - \rho_{\text{model}}\rvert \le 0.5\rho_{\text{model}}$	$\rho = 1.602$ (1.041); rel. diff. 0.54	
T3: $d_{\text{cr}} = 1 + \ln N$	$s \ge 0.5$, $\lvert\Delta d_{\text{cr}}\rvert \le 4$	$s = 0.400$ (0.896); $d_{\text{cr}} = 7.42$ (17.33)	refuted

Table 1: Pre-registered tests at $N_{\text{full}} = 44{,}832{,}865$. Values in brackets come from the authors' model through the same estimator. $\pm$ is the standard deviation of γ over the 5 subset seeds. Eq. (3) with $\beta = 0$ gives $d_{\text{cr}} = 18.62$ at this N.

Estimator. We used the source's method, least squares on a log-binned histogram. The bin edges are the distinct values of $\lfloor 2^{k/2} \rfloor$. A bin enters the fit if its upper edge is at most $N/100$ and it holds at least 5 nodes. The exponent is undefined if fewer than 4 bins enter. The source does not state its fit range; this one was fixed in the protocol.

Statistics. α_{pool} is the estimate on the pooled $S(n)$. $b(N)$ is the least-squares slope of α_d on d for $d = 3, \ldots, \lfloor \ln N \rfloor$. $\rho = b \ln N$; Eq. (2) predicts $\rho = 1$. γ is the slope of $\ln b$ on $\ln \ln N$ across the N grid; Eq. (2) predicts $\gamma = -1$. d_{cr} is the first depth with $\alpha_d \ge 2$, interpolated linearly. The N grid is $10^4, 10^{4.5}, \ldots, 10^{7.5}$, the source's $1.4 \cdot 10^6$, and the full set. For each N below the full set we drew 5 random subsets (seeds 0–4) and report the mean.

Known-answer control. The same estimator does not return Eq. (2) exactly on a finite curve. So the reference for each test is the source's own process passed through the same estimator. We simulated $2 \cdot 10^6$ chains $n_d = \max(1, \text{round}(n_{d-1} r_d))$ from $n_0 = N$, with r_d drawn from the arcsine law, for each N of the grid, and converted the chain counts into expected node counts.

Decision rules. T1: supported if $\lvert\alpha_{\text{pool}} - 2\rvert \le 0.10$ at the full N. T2: refuted if $\lvert\gamma - \gamma_{\text{model}}\rvert > 0.5$ or $\lvert\rho - \rho_{\text{model}}\rvert > 0.5\,\rho_{\text{model}}$ at the full N; supported if both differences are at most half of these limits. T3: refuted if the slope s of d_{cr} on $\ln N$ is below 0.5 or above 1.5, or if d_{cr} differs from the model by more than 4 at the full N. The finite-size theory is refuted if T2 or T3 is refuted.

Controls. Held-out strata at $N = 10^6$: Bullet, Blitz and Classical games; terciles of the mean rating (cuts 1534 and 1732); and 2013–2014 against 2015. A placebo: among games with a full 40-ply key, take $1.4 \cdot 10^6$ games and permute each ply column independently across games (5 seeds). This keeps the move frequencies at each ply and removes all dependence between moves.

3 Results

T1: the pooled exponent. At the full N, $\alpha_{\text{pool}} = 2.086$. T1 is supported. The pooled exponent falls toward 2 as N grows: 2.61 at 10^4 and 2.09 at $4.5 \cdot 10^7$ (Table 2). The model shows the same finite-size trend, from 2.41 to 2.04. A Clauset–Shalizi–Newman fit [8] with the `powerlaw` package [9], at $N = 1.4 \cdot 10^6$ (seed 0), gives $x_{\min} = 16$ (KS distance 0.0025), $\alpha = 2.024$ and 62,030 nodes in the tail. The likelihood ratio does not prefer a lognormal ($R = 0.98$, $p = 0.33$) or a truncated power law ($R = -0.72$, $p = 0.47$). The pooled weights are nested, so these p-values are descriptive only.

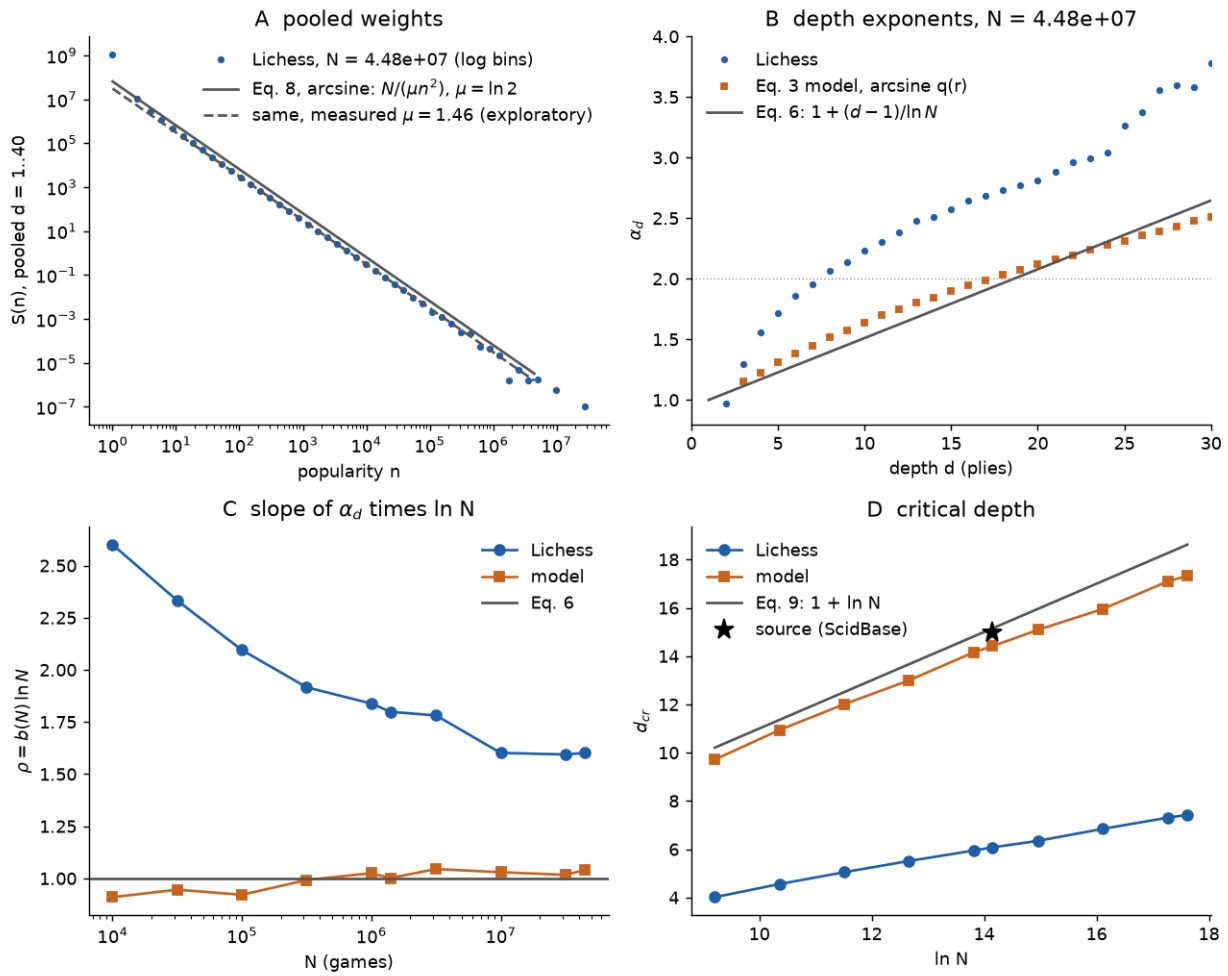


Figure 1: (A) Pooled $S(n)$ at the full N, with Eq. 8 of the source for the arcsine value $\mu = \ln 2$ (solid) and for the measured $\mu = 1.46$ (dashed; exploratory, Section 5). (B) α_d against d for the data, the model and Eq. (2). (C) $\rho = b \ln N$ against N; Eq. (2) predicts 1. (D) d_{cr} against $\ln N$, with the source's value 15 at $N = 1.4 \cdot 10^6$.

T2: the slope of α_d. The estimator recovers Eq. (2) on the model: ρ_{model} is between 0.91 and 1.04 on the whole grid. On Lichess, ρ is 2.60 at 10^4 and 1.60 at the full N. The slope falls faster than $1/\ln N$: $\gamma = -1.756$ (SD 0.058 over 5 seeds) against -0.788 for the model. Both T2 limits are exceeded. T2 is refuted.

T3: the crossing depth. On the model, d_{cr} grows with slope 0.90 in $\ln N$, close to Eq. (3). On Lichess the slope is 0.40, and $d_{\mathrm{cr}} = 7.42$ at the full N, against 17.33 for the model and 18.62 for Eq. (3). At the source's $N = 1.4 \cdot 10^6$, Lichess gives $d_{\mathrm{cr}} = 6.1$; the source reports about 15 for ScidBase. T3 is refuted.

Held-out strata. Table 3 gives the strata at $N = 10^6$. No stratum meets the pre-registered replication rule. In all strata, α_{pool} is 2.155–2.167. The full set gives 2.166 at the same N, so this is the finite-size trend of T1 and not a difference between strata. The T2 and T3 failures repeat in

N	α_{pool}	b	ρ	d_{cr}	model ρ	model d_{cr}
10^4	2.609	0.282	2.60	4.01	0.91	9.72
10^5	2.299	0.182	2.09	5.05	0.92	12.01
10^6	2.166	0.133	1.84	5.95	1.03	14.16
$1.4 \cdot 10^6$	2.155	0.127	1.80	6.07	1.00	14.41
10^7	2.105	0.099	1.60	6.85	1.03	15.96
$4.48 \cdot 10^7$	2.086	0.091	1.60	7.42	1.04	17.33

Table 2: Statistics at selected N (mean of 5 random subsets below the full set). $1 + \ln N$ is 10.21, 12.51, 14.82, 15.15, 17.12 and 18.62 on these rows.

Stratum	games	α_{pool}	ρ	d_{cr}
Bullet	14,337,880	2.155	2.16	5.69
Blitz	18,281,202	2.164	1.70	6.27
Classical	11,582,480	2.161	1.86	6.10
Rating $<$ 1534	14,901,926	2.164	2.08	5.75
Rating 1534–1731	14,983,580	2.162	1.89	6.02
Rating $\geq$ 1732	14,942,928	2.158	1.68	6.38
2013–2014	14,964,152	2.167	1.89	5.89
2015	29,868,713	2.164	1.84	5.97

Table 3: Held-out strata, $N = 10^6$ games each, mean of 5 subsets.

every stratum: $\rho \geq 1.68$ against $\rho_{\text{model}}(10^6) = 1.03$, and $d_{\text{cr}} \leq 6.4$ against 14.2. The deviation is smallest for high ratings and Blitz and largest for Bullet and low ratings.

Placebo. The placebo uses the 36,705,863 games with a full 40-ply key. On the real games ($1.4 \cdot 10^6$, mean of 5 seeds) we get $\alpha_{\text{pool}} = 2.155$, $b = 0.128$, $\rho = 1.81$ and $d_{\text{cr}} = 6.07$. On the shuffled games, α_{pool} is 2.137–2.138, b is 0.75–0.79 and ρ is 10.6–11.1 over the 5 seeds. So the shuffled tree also gives a pooled exponent near 2. This agrees with the source's argument that any well-behaved branching process gives $\alpha = 2$. It also means that $\alpha \approx 2$ alone does not show anything specific to chess. The shuffle makes b about 6 times larger. In the shuffled set α_d is defined at only 6 depths, so we use only this ratio and not the exact shuffled value.

4 Robustness of T2 and T3 (not pre-registered)

The fit range of the estimator is our choice, because the source does not state one. An independent robustness script repeats T2 and T3 with five fit ranges, new random subsets (seeds 9000–9002, 3 per N) and a separately written estimator. Table 4 gives the result. The first row uses the protocol range. Its $\gamma = -1.82$ differs from the pre-registered -1.756 only because the subsets are new; ρ and d_{cr} at the full N use all games and are equal in both analyses.

T3 is refuted in all five fit ranges. The slope of d_{cr} on $\ln N$ is 0.37–0.43 (model 0.78–1.12). T2 changes to "inconclusive" for one range (lowest weight 4: $|\gamma - \gamma_{\text{model}}| = 0.499$, relative ρ difference 0.33).

The reason is the shape of α_d. On Lichess it is concave in d, not linear. At the full N (protocol range), $\alpha_2, \ldots, \alpha_9$ are 0.97, 1.30, 1.55, 1.72, 1.86, 1.95, 2.06 and 2.14. After that, the steps are

lowest n	highest n	γ	ρ	d_{cr}	T2	T3
1 (protocol)	$N/100$ (protocol)	-1.82 (-0.78)	1.60 (1.05)	7.42 (17.33)	refuted	refuted
2	$N/100$	-1.56 (-0.79)	1.47 (1.06)	7.38 (17.40)	refuted	refuted
4	$N/100$	-1.31 (-0.81)	1.43 (1.07)	7.31 (17.18)	inconclusive	refuted
1	$N/1000$	-1.89 (-0.98)	1.73 (1.00)	7.83 (18.10)	refuted	refuted
2	$N/10$	-1.59 (-0.86)	1.45 (1.05)	7.38 (17.40)	refuted	refuted

Table 4: T2 and T3 for five fit ranges at the full N; model values in brackets. Decisions use the pre-registered thresholds.

about 0.05 to 0.09. Eq. (2) predicts a straight line with slope $1/\ln N = 0.057$. Thus the value of b depends on the depth window, and the protocol window $3 \le d \le \lfloor \ln N \rfloor$ grows with N, which adds to γ. In the fixed window $d = 3, \ldots, 9$, $\gamma = -1.13$ (model -0.46) and $\rho = 2.38$ (model 1.29); T2 is still refuted by the same thresholds. In the window $d \ge 8$ (lowest weight 2), $\rho = 1.03$ (model 0.92). So the slope at depths 8 and more is close to $1/\ln N$, and the excess slope is in the first 7 or 8 plies.

The published exponents for ChessDB show the same feature. The values $\alpha_1, \ldots, \alpha_4$ (1.10, 1.29, 1.47 and 1.59) of Schaigorodsky et al. [4] have steps 0.19, 0.18 and 0.12, while $1/\ln N = 0.071$ at $N = 1.4 \cdot 10^6$. This is our reading of their numbers. Their fit method differs from ours, so it is not a test.

T3 fixes $\beta = 0$, as the source does for $d_{\mathrm{cr}} \approx 15$. With a free β, the measured slope $s = 0.40$ equals Eq. (3) with $\beta \approx -0.6$. But Eq. (2) with the same β predicts a straight line that starts at $\alpha_2 \approx 1.66$. The data give $\alpha_2 = 0.97$ and a concave curve. So no single β makes Eq. (2) and Eq. (3) agree with the data together. This is a reading of the numbers above, not a pre-registered test.

5 Exploratory: branching ratios (not pre-registered)

We measured the branching ratios r as in Fig. 3A of the source (parents with $n \ge 100$, each edge weighted by r). The mean $\mu = \langle -\ln r \rangle$ is 1.47 at $N = 1.4 \cdot 10^6$ and 1.46 at $N = 10^{6.5}$. The arcsine law gives $\mu = \ln 2 = 0.69$. At $N = 1.4 \cdot 10^6$ the value is highest at ply 2 (2.06) and falls with depth (1.16–1.26 at plies 12–14). So Lichess games split faster than the arcsine law, and the split rate depends on depth. With the measured $\mu = 1.46$, Eq. 8 of the source matches the amplitude of the pooled $S(n)$ (Fig. 1A).

To separate the depth dependence from correlations along a game, we shuffled the measured ratios across games within each depth and rebuilt the chains. This surrogate gives $d_{\mathrm{cr}} = 6.4$ ($N = 1.4 \cdot 10^6$) and 6.8 ($N = 10^{6.5}$), close to the data (6.1 and 6.4). It gives $\rho = 1.15$ and 1.25, between the model (1.0) and the data (1.8). One reading: the early crossing follows from the fast, depth-dependent splitting, and the excess slope comes partly from correlations along the path of a game. No pre-registered test supports this reading.

6 How the result was checked

- **Full rerun.** `checks/rerun.py` checks the SHA-256 of the 36 raw files (it downloads missing files), parses them, builds the sorted key set, runs all pre-registered stages, the robustness script and the exploratory analysis, draws Fig. 1, and compares 1950 numbers with `data/expected.json` (relative tolerance 10^{-6}). The script refuses a work folder that already holds parsed or built data, so each run starts from the raw files. Three from-zero runs before

release (one by the solving agent, one by the verifying agent, one on the released folder) printed `1950/1950 numbers match`.

- **Counting code.** `checks/test_counting.py` counts 10^5 synthetic games with the sort/LCP code and with a direct prefix dictionary. The histograms are equal at all 40 depths, for the full set and for a subset.
- **Independent count.** A second program, written separately by the verifying agent, parses the raw PGN text with its own SAN regular expression, hashes each prefix and counts with `numpy.unique`. On all 36 months it finds the same 44,832,865 games, $S_d(n)$ equal to ours at depths 1 to 12, and $d_{\mathrm{cr}} = 7.4214$ with its own estimator. Depths 13 to 40 were checked independently on one month only.
- **Protocol order.** The protocol commit precedes all outcome files, and the protocol file has no later change.

7 Limits

- **Population.** Lichess in 2013–2015 is online play, mostly Bullet and Blitz, with mean ratings near 1500–1800. ScidBase holds mostly master and club games. The refutation concerns Eqs. (2) and (3) as general laws of opening trees. It does not show that the source's measurements on ScidBase are wrong. We had no ScidBase data and did not re-test them. Even the top rating tercile and Classical games give $d_{\mathrm{cr}} \approx 6.1$–$6.4$ at $N = 10^6$, and no stratum here is comparable to master play.
- **Estimator.** The fit range is ours. T3 does not depend on it in five ranges. T2 is refuted in four of five ranges and inconclusive in one.
- **β.** T3 tests Eq. (3) with $\beta = 0$ only.
- **Statistics.** The CSN p-values treat nested weights as independent. Sections 4 and 5 are not pre-registered.
- **Earlier work.** Chacoma and Billoni [5] report that the trend of α_d on their mixed set of 10^4 games agrees with slope 0.11. On Lichess subsets of 10^4 games we get $b = 0.28$. Their data (FIDE games and Lichess games above 2500 and below 1500, all classical) and their fits differ from ours. We did not find the cause of the difference.
- **Keys.** We did not check the SAN text for legality. We assume that the Lichess server records only legal moves. Games with no moves (0.47%) are dropped.

8 Conclusion

On 44.8 million Lichess games the pooled weights of opening lines have exponent 2.09, which supports the universal $\alpha = 2$. A shuffled placebo also gives an exponent near 2, as the theory of the source predicts for any branching process. The finite-size laws of the same theory fail on this data: the slope of α_d falls faster than $1/\ln N$, α_d is concave in d, and α_d crosses 2 at about depth 7, not near $1 + \ln N \approx 18$. Both failures hold in every stratum and, for d_{cr}, in every fit range.

References

[1] B. Blasius and R. Tönjes, "Zipf's law in the popularity distribution of chess openings," *Phys. Rev. Lett.* **103**, 218701 (2009). doi:10.1103/PhysRevLett.103.218701; arXiv:0704.2711v2.

[2] S. Maslov, "Power laws in chess," *Physics* **2**, 97 (2009). doi:10.1103/Physics.2.97.

[3] J. I. Perotti, H.-H. Jo, A. L. Schaigorodsky and O. V. Billoni, "Innovation and nested preferential growth in chess playing behavior," *EPL* **104**, 48005 (2013). doi:10.1209/0295-5075/104/48005.

[4] A. L. Schaigorodsky, J. I. Perotti and O. V. Billoni, "A study of memory effects in a chess database," *PLoS ONE* **11**, e0168213 (2016), Fig. 1 caption. doi:10.1371/journal.pone.0168213.

[5] A. Chacoma and O. V. Billoni, "Emergent complexity in the decision-making process of chess players," *Sci. Rep.* **15**, 23234 (2025). doi:10.1038/s41598-025-06335-9.

[6] Y. Cheng, Z. Fang, D. Xu, Y. Dong, X. Hu and J. Qian, "Multi-scaling laws and long-term correlations in chess game records popularity," *J. East China Normal Univ. (Nat. Sci.)* 2025(3), 137–146. doi:10.3969/j.issn.1000-5641.2025.03.016.

[7] G. De Marzo and V. D. P. Servedio, "Quantifying the complexity and similarity of chess openings using online chess community data," *Sci. Rep.* **13**, 5327 (2023). doi:10.1038/s41598-023-31658-w.

[8] A. Clauset, C. R. Shalizi and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.* **51**, 661–703 (2009). doi:10.1137/070710111.

[9] J. Alstott, E. Bullmore and D. Plenz, "powerlaw: a Python package for analysis of heavy-tailed distributions," *PLoS ONE* **9**, e85777 (2014). doi:10.1371/journal.pone.0085777.

[10] Lichess, *lichess.org open database*, standard rated games 2013-01 to 2015-12, `https://database.lichess.org/`, CC0, accessed November 25, 2026.