An R database of 1.12 million chess opening sequences: A 43-year time-series of game-play from 42,644 Chess players linked in annually-resolved game-playing networks

Cody T. Ross and Bret Behiem

Max Planck Institute for Evolutionary Anthropology. Department of Human Behavior, Ecology and Culture. Deutscher Platz 6. 04103 Leipzig.

Abstract

A central focus in the study of cultural evolution is the relative importance of individual and social information in driving the diffusion of behavioral and symbolic variants. In observational social learning research, however, empirical tests of quantitative models are limited by a dearth of long-term, individual-level data. Here we present an archive of longitudinal records from the game of Chess, which will allow researchers to study the year-by-year evolution of opening strategies among a large multi-national population of players. These records will allow for analyses of within-individual skill development, diffusion models of strategies over time, network-structured time-series modeling of social learning, and measurement of effects motivated by other theoretical models of social learning strategies, including payoff- and frequency-dependent heuristics. These records provide a unique window into a long-term record of human behavior, often in high stakes competitions.

Keywords: Chess, Social Learning, Social Networks, Cultural Evolution

Specifications Table

Subject area	Evolution and ecology, anthropology, psychology.
More specific subject area	Cultural evolution, social learning, social networks.
Type of data	Electronic records of Chess game movement patterns and associated player data.
How data was acquired	Source data come from a database of .pgn text files published by Ed Schröder [1].
Data format	.RData, .R code
Experimental factors	Raw published text files [1] on Chess moves were processed and cleaned to generate a useable database of Chess opening move sequences and associated player data.
Experimental features	The R database contains information on the date of each game, the names of the players, the Elo ratings of the players, the outcome of the game, and the opening sequence of moves. The database structure facilitates time-series and social network analysis.
Data source location	NA.
Data accessibility	All data and code are included in the Supplementary Materials and will be maintained at: `https://github.com/ctross/chessbase`

Value of the data

- This database structures and cleans downloadable plaintext data on Chess opening move sequences, and will allow researchers to model the data using standard statistical tools.

- This database provides quantitative, individual-level data on the frequency of opening move variants in Chess over a long course of time, and will be useful for research on the dynamics of cultural change and individual and social learning (e.g., as in [2]).

- Since the pair-wise relationships of game-play between individuals is explicit in this database, it contains the information needed to model social learning through social networks, and will allow researcher interested in social learning to test if focal players update their move repertoires as a function of the performance or frequencies of play of the opening move repertoires of players in their direct and local social networks.

Data

The main database presented here is based on the compilation of Chess games called Million Base 2.2, published by Ed Schröder [1]. The original data-set was provided in a plaintext markup language specific to Chess software, limiting access and potential for research use. We present a transformed version of this data in more common digital formats, and provide processing software which can reproduce our work on similarly formatted Chess data. We also exclude cases that are difficult to use for common social learning research questions, remove erroneous data, and provide some standardization of individual identity, improving the data quality.

After running the set of R scripts included in the Supplementary Materials, a single R object containing the cleaned data records is exported. Table 1 provides an example of the resulting data structure. Each row in the data-base represents a single game, and includes: 1) the year and month (if available) in which the game was played, 2) the names of each player, 3) the outcome of the game (1 = white wins, 2 = a draw, 3 = black wins), 4) the Elo [3] ratings of each player, and 5) the opening sequence of moves (i.e., the first move played by white, M1; black's first response, M2; the second moved played by white, M3; up to the response of black to white's fifth move, M10).

Table 1: Example data structure. Records are composed of: 1) a unique game ID, 2) the year and month in which the game was played, 3) the names of the players, 4) the outcome of the game (1 = white wins, 2 = a draw, 3 = black wins), 5) the Elo rankings of the players, and 6) the opening sequence of the first ten moves of the game.

GameID	Year	Month	White	Black	Result	WhiteElo	BlackElo	M1	M2	...	M10
85bbdf29a360b	1982	1	TalM	VanderWielJ	1	2605	2470	c4	e6	...	Ba5
98c779e412b0c	1982		TalM	VanderWielJ	1	2610	2520	c4	Nf6	...	c5
9f918a596dd53	1982	1	TalM	ChandlerM	3	2605	2470	e4	c6	...	Bg6
78c93075b3a91	1982	10	TalM	RubinettiJ	1	2610	2415	e4	e5	...	Be7
b7c2146d818b3	1982	1	TalM	SunyeNetoJ	1	2605	2475	e4	c5	...	Nf6
1090dbb7f923e	1982		TalM	GheorghiuF	2	2610	2550	e4	c5	...	Qc7
6700eb627b15c	1982		TalM	SaxG	2	2610	2560	e4	c5	...	g6
489e65fa48669	1982		TalM	QuinterosM	2	2610	2520	e4	c5	...	e6
6c1f2a80daa61	1982	1	TalM	KavalekL	2	2605	2590	e4	c5	...	d6

We note a few facts about the composition of the data-set over time (see Fig. 1) that might color future inferential models. First, the data is densest in the 1990s and 2000s, both in terms of games played, and unique players (Fig. 1a). Second, the average Elo rating of players has declined slightly with time in this data-set, and variation in the Elo ratings of players has increased (Fig. 1b). This effect is not due to decreasing skill with time, but rather to a larger number of games of lower ranked players being included in more recent years. The frequency of both wins and losses by white have remained fairly constant with time (Fig. 1c).

Materials and Methods

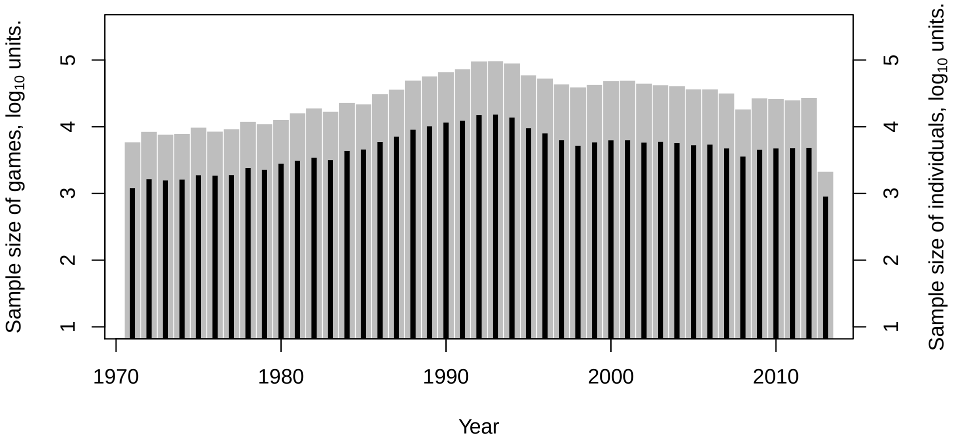
The Million Base 2.2 data-set of pgn-formatted Chess games was downloaded in late 2017 [1]. We first designed a script `Parse.R` (see Supplementary Materials) to parse the original text files, and extract the essential bits of information: 1) the year and month in which each game was played, 2) the Elo rating of each player, 3) the names of the players, 4) the outcome of the game, and 5) the opening sequence of moves. This script creates machine-readable json and csv formatted versions of the database.

Next, we clean and compile these initial records into an R database using a second script `Clean.R` (see Supplementary Materials). We remove records with suspiciously low Elo scores, records where players played against themselves, and records where impossible opening moves were played, as each of these cases suggests data entry errors.

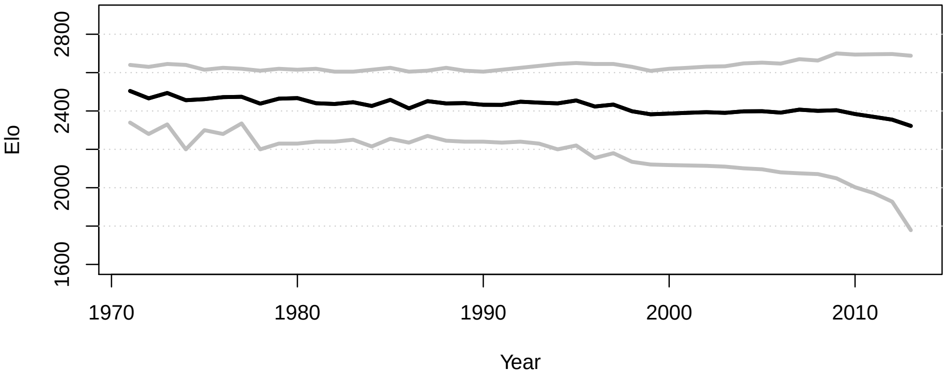
Finally, we conduct some thinning of the original data-set using a third script `Thin.R` (see Supplementary Materials). Here, we produce a final data set by: 1) removing games by players who never received an Elo ranking at any point in the data-base (in most cases, researchers will want to account for player skill), 2) we remove historic games from prior to 1971 (the data is much sparser over this time period, and games during this period are completely devoid of Elo information), and 3) we remove games played by players with fewer than 5 recorded games as white (there is little hope for time series analysis with such few points, and these records might reflect higher levels of data entry errors, like misspellings of names). The likelihood of some of these cases being based on name misspelling was suggested by estimates of pairwise Levenshtein distance [4] between those players with $\geq$5 and $<$5 games—a density peak at small difference values was apparent. Lastly, 4) we conduct one additional check for name collisions—we calculate the standard deviation on year (σ_Y) for all game records linked to an individual, and plot this as a function of games played. In cases where a single player name is associated with games played at distinct time periods (i.e., a few games in the 1970s and a few games in the late 2000s) the value of σ_Y will be large. We spot-checked cases of large values of σ_Y, and then reviewed external data sources on these individuals' playing history to investigate if the divergent playing periods were real, or likely artifacts of a name collision. We found that around 0.25 of these cases reflected likely collisions, so we removed all games from players with $\sigma_Y > 13$ and less than 50 games played. While we were able to remove many likely collisions using these measures, other collisions might still be present.

The completed database includes games played between 1971 and 2013. There are

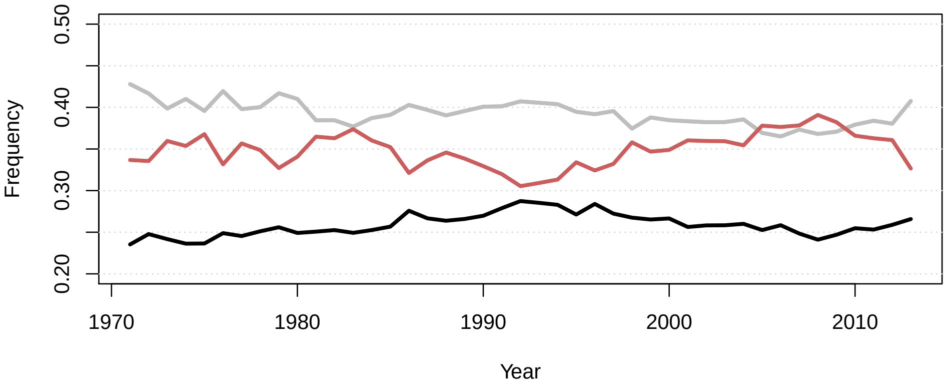
Figure 1: Time-series of the descriptive statitics of the data-set.



(a) Games by year (grey), and individuals by year (black), in the data-set.



(b) Mean Elo rating (black) with 90% intervals (grey) across time in the data-set.



(c) Frequency of wins (grey), draws (red), and losses (black) by white across years in the data-set.

1,121,900 game records from 42,644 unique players (as white). The raw data, text processing code, and cleaned data are included in the Supplementary Materials released with this paper, and will be maintained on GitHub, https://github.com/ctross/chessbase.

References

[1] E. Schröderr, Million Base 2.2 (2012).
URL http://www.top-5000.nl/pgn.htm

[2] B. A. Beheim, C. Thigpen, R. Mcelreath, Strategic social learning and the population dynamics of human behavior: the game of go, Evolution and Human Behavior 35 (5) (2014) 351 – 357.

[3] A. E. Elo, The rating of chessplayers, past and present, Arco Pub., 1978.

[4] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet physics doklady, Vol. 10, 1966, pp. 707–710.