Large-scale Analysis of Chess Games with Chess Engines: A Preliminary Report

Mathieu Acher, François Esnault

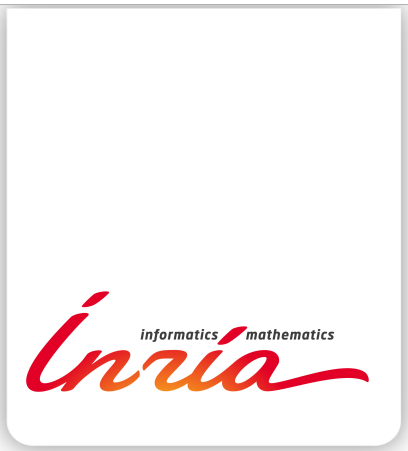
TECHNICAL REPORT

N° 479

April 2016

Project-Teams DiverSE

ISSN 0249-0803 ISRN INRIA/RT--479--FR+ENG



Large-scale Analysis of Chess Games with Chess Engines: A Preliminary Report

Mathieu Acher*, François Esnault*

Project-Teams DiverSE

Technical Report n° 479 — April 2016 — 16 pages

Abstract: The strength of chess engines together with the availability of numerous chess games have attracted the attention of chess players, data scientists, and researchers during the last decades. State-of-the-art engines now provide an authoritative judgement that can be used in many applications like cheating detection, intrinsic ratings computation, skill assessment, or the study of human decision-making. A key issue for the research community is to gather a large dataset of chess games together with the judgement of chess engines. Unfortunately the analysis of each move takes lots of times. In this paper, we report our effort to analyse almost 5 millions chess games with a computing grid. During summer 2015, we processed 270 millions unique played positions using the Stockfish engine with a quite high depth (20). We populated a database of 1+ tera-octets of chess evaluations, representing an estimated time of 50 years of computation on a single machine. Our effort is a first step towards the replication of research results, the supply of open data and procedures for exploring new directions, and the investigation of software engineering/scalability issues when computing billions of moves.

Key-words: chess game, data analysis, artificial intelligence

* Inria/IRISA, University of Rennes 1, France

RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Analyse large échelle de parties d'échecs avec des moteurs: un rapport préliminaire

Résumé : La force des moteurs d'échecs et l'existence de nombreuses parties d'échecs ont attiré l'attention des joueurs d'échecs, des scientifiques de la donnée, et des chercheurs au cours des dernières années. Les meilleurs moteurs fournissent un jugement autoritaire qui peut être utilisé dans de nombreuses applications comme la détection de triches, le calcul d'une force intrinsèque, l'évaluation d'aptitudes, ou l'étude de prises de décision par des humains. Un problème important pour la communauté de chercheurs est de collecter un large ensemble de parties d'échecs avec les jugements des moteurs. Malheureusement l'analyse de chaque coup peut prendre énormément de temps. Dans ce rapport, nous décrivons notre effort pour analyser près de 5 millions de parties d'échecs avec une grille de calcul. Durant l'été 2015, nous avons analysé 270 millions de positions uniques issues de parties réelles en utilisant le moteur Stockfish avec une assez grande profondeur (20). Nous avons construit une base de données d'évaluation d'échecs de plus de 1 tera-octet, représentant un temps estimé de 50 années sur une machine seule. Notre travail est une première étape vers la réplication de résultats de recherche, la mise à disposition de données ouvertes et de procédures pour explorer de nouvelles directions, et l'étude de problèmes de génie logiciel pour calculer des milliards de coups.

Mots-clés : jeu d'échecs, analyse de données, intelligence artificielle

Large-scale Analysis of Chess Games with Chess Engines: A Preliminary Report

Mathieu Acher and François Esnault

April 28, 2016

1 Introduction

Millions of chess games have been recorded from the very beginning of chess history to the last tournaments of top chess players. Meanwhile chess engines have continuously improved up to the point they cannot only beat world chess champions but also provide an authoritative assessment [7–9, 14]. The strengths of chess engines together with the availability of numerous chess games have attracted the attention of chess players, data scientists, and researchers during the last three decades. For instance professional players use chess engines on a daily basis to seek strong novelties; chess players in general confront the moves they played to the evaluation of a chess engine for determining if they do not miss an opportunity or blunder at some points.

From a scientific point of view, numerous aspects of the chess game have been considered, being for quantifying the complexity of a position, assessing the skills, ratings, or styles of (famous) chess players [2, 3, 5, 11, 12, 15], or studying the chess engines themselves [4, 6]. Questions like "Who are the best chess players in history?" can potentially have a precise answer with the objective (and hopefully optimal) judgement of a chess engine. So far numerous applications have been considered, such as methods for detecting cheaters [1, 10], the computation of an intrinsic rating or the identification of key moments chess players blunder [13].

A key issue for the research community is to gather a large dataset of chess games together with the judgement of chess engines [2]. For doing so, scientists typically need to analyze millions of games, moves, and combinations with chess engines. Unfortunately it still requires lots of computations since (1) there are numerous games and moves to consider while (2) chess engines typically need seconds for fully exploring the space of combinations and thus providing a precise evaluation for a given position. As a result and due to the limitation of computing storage or power, chess engines have been executed on a limited number of games or with specific parameters to reduce the amount of computation.

Our objective is to propose an open infrastructure for the large-scale analysis of chess games. We hope to consider more players, games, moves, chess engines, parameters (e.g., the depth used by a chess engine), and methods for processing the overall data. With the gathering of a rich and large collection of chess engines' evaluations, we aim to (1) replicate state-of-the-art research results [2] (e.g., on cheat detection or intrinsic ratings); (2) provide open data and procedures for exploring

*Inria/IRISA, University of Rennes 1, France

new directions; (3) investigate software engineering/scalability issues when computing millions of moves; (4) organize a community of potential contributors for fun and profit.

In this paper, we report our recent effort to analyse almost 5 millions chess games with a computing grid. During summer 2015, we processed 270 millions of unique played positions using the Stockfish [14] chess engine with a quite high depth (20). Overall we populated a database of 1+ tera-octets of chess evaluations, representing an estimated time of 50 years of computation on a single machine.

Data analysts or scientists can use the dataset as well as the procedures to gather novel insights, revisit existing works, or address novel issues. The lessons and numbers of our experience report can also be of interest for launching other large-scale analysis of chess games with other chess engines, games, and settings.

2 Dataset and Chess Games

We gathered 4.78 million unique games publicly available on some Web repositories. In this section we report on some properties of the dataset.

From a technical point of view, we parsed Portable Game Notation (PGN) files with a Java parser[1] and pgn-extract[2]. As a preprocess we notably eliminated a significant number of duplicated games with pgn-extract. We used the R programming language to produce plots and results.

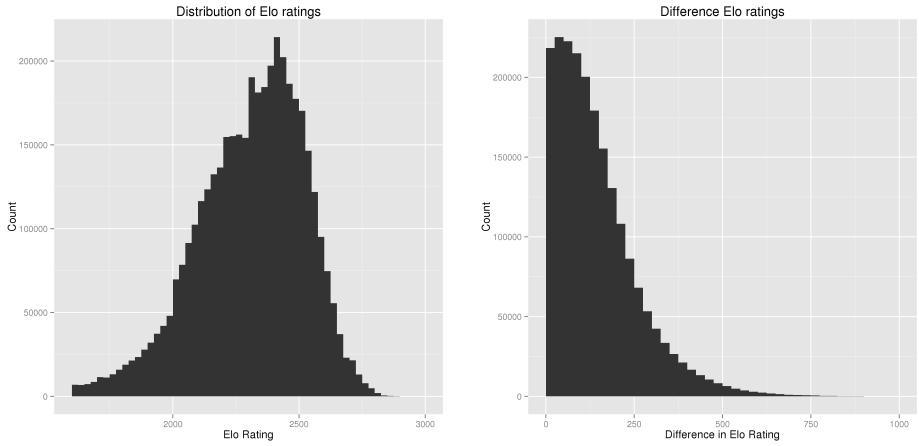
Interestingly we can confront our results to a previous attempt by Randal Olson that explored a dataset of over 650,000 chess tournament games (see a series of blog posts[3]). We consider similar properties (such as distribution of Elo ratings) as well as additional ones (such as most represented events). In our case we have much more games and it is worth comparing our numbers to Olson's results. Our objective was to have a better understanding of the quality of the dataset, since some factors can have a negative impact on their exploitations by chess engines.

2.1 Elo Ratings

The Elo rating system is intensively used in chess for calculating the relative skill levels of players. Figure 1a gives the distribution of Elo ratings thanks to PGN headers. It should be noted that the Elo rating was adopted by the World Chess Federation (FIDE) in 1970. As such oldest games do not appear in Figure 1.

Players with >2000 Elo rating are considered as experts in Chess; between 2200 and 2400, players are national masters; above 2400, international masters and above 2500 international grand masters. Figure 1a shows that most records come from games with players >2200 Elo, with a large proportion played by international masters, grand masters, and top players. In general, the difference of Elos between two opponents is quite close (<100), see Figure 1b. It is in line with what observed in chess tournaments (e.g., round-robin or swiss systems).

[1] https://github.com/jvarsoke/ictk
[2] https://www.cs.kent.ac.uk/people/staff/djb/pgn-extract/
[3] http://www.randalolson.com/2014/05/24/chess-tournament-matches-and-elo-ratings/



(a) Distribution of Elo ratings (b) Differences of Elo

Figure 1: Distribution and differences of Elo

2.2 Ply per Games

Our 4.78 million unique games have a mean of 80 ply per game (40 moves). Figure 2 shows that the number of moves follows two trends: With <200 Elo difference, the number of ply[4] per game varies between 78 and 85. It is not clear why there is a slight increase. We can formulate the assumption that some games end early with a draw in case the Elo difference is closed. With a larger Elo difference, the number of Ply starts decreasing until 60 ply per game. It is quite intuitive since strongest players gain the upper hand quickly over their opponents.

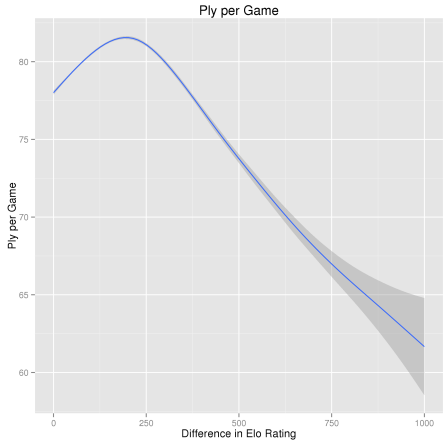


Figure 2: Ply per Game w.r.t. difference in Elo rating

[4]A ply refers to one turn taken by one of the chess players

The number of ply in some games can be suspicious or presents limited interests:

- Klip,H (2305) - Bottema,T (2205) 1. e4 f6 2. d4 g5 3. Qh5# 1-0 (1990)
- Landa,K (2678) - Grall,G (1812) 1. e4 e5 2. Bc4 Bc5 3. Qh5 Nf6 4. Qxf7# 1-0 (2007)
- Strekelj,V (1843) - Kristovic,M (2328) 1. f3 e5 2. Kf2 d5 3. Kg3 Bc5 4. Nc3 Qg5# 0-1 (2011)

On the other hand, we have very long games, for example:

- Sapin - Hyxiom (2003-01-01) 600 plies
- Felber,J (2150) - Lapshun,Y (2355) (1998-09-05) 475 plies
- Sanal,V (2286) - Can,E (2476) (2012-03-29) 456 plies

2.3 Elo Difference and Winning Chance

The difference in Elo rating strongly predicts the winner of the game. With a difference of 100, the chance for the higher player to win is 70% – not counting draws. The graph of Figure 3 is in line with the probabilities and expected outcomes of the Elo rating system. In general our dataset provides PGN headers information (Elo, winner) we can trustfully exploit in the future. We have also some exceptional records/anecdots:

- **Vera Gonzalez,J (1551)** — Hernandez Carmenates,Hold (2573): 1-0 (Elo difference: 1022)
- **Freise,E (2018)** — Anand,V (2794): 1-0 (Elo difference: 776)
- **Sikora,J (1833)** — Movsesian,S (2710): 1-0 (Elo difference: 877)

2.4 Colors and Percentage Win

Figure 4a shows that the percentage games won by white players depends of Elo Rating. Below 2200 Elo, chances to lose are higher (even with white colors) because such players face to better opponents and simply have less chance to win. On the other side, the percentage to win for a 2500 white player is 70% – not counting draws. Specifically, we have 120K games with white player having a Elo rating >2600: They win more than 55K games, draw 50K and lose only 15K games. Figure 4b shows the importance of having white pieces. For low Elo ratings, the draw rarely happens whereas it happens half the time for stronger players. All these results are coherent with what is usually observed in chess.

2.5 First moves and Openings

Figure 5a shows that the period 1900 – 1960 was more prone to fluctuations/experiments for the choice of the first moves since the openings theory was not developed. d4, e4, Nf3, and c4 are the most popular first moves to start a game (see Figure 5b). It should be noted that for the period 1960 – 2015 the top first moves slightly change over the time and tend to stabilize. It is consistent with existing databases and current practices.

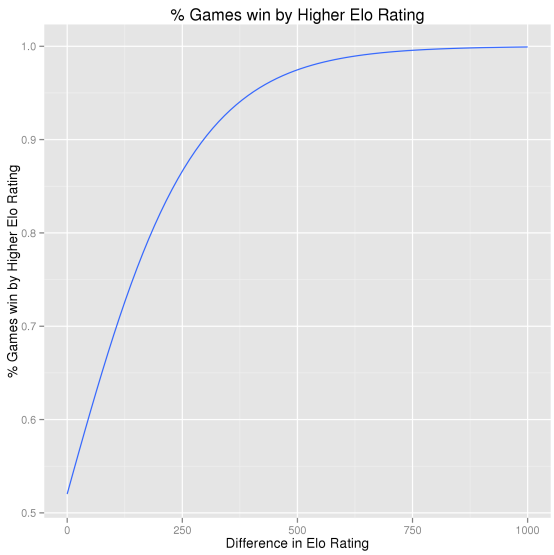
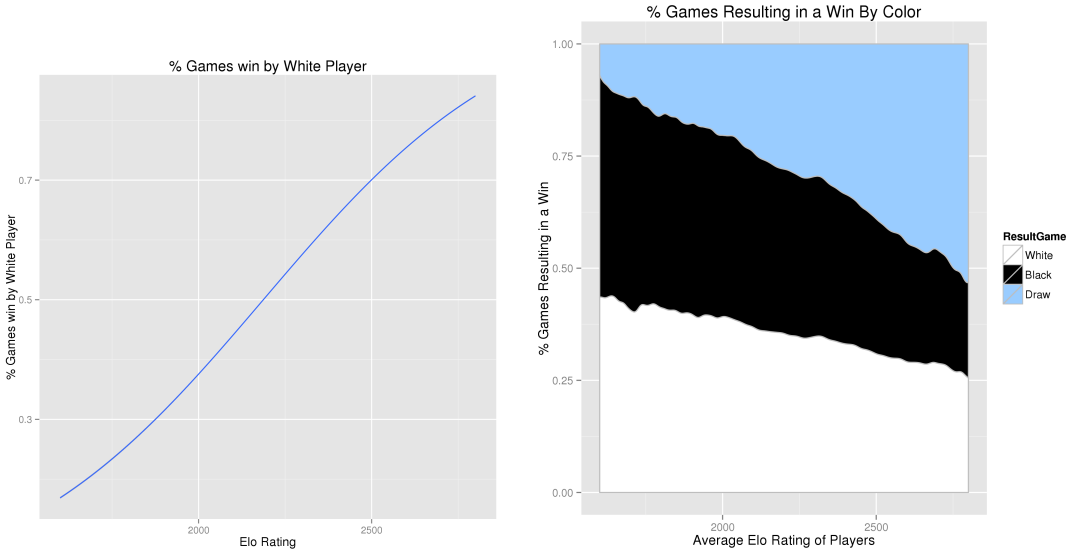


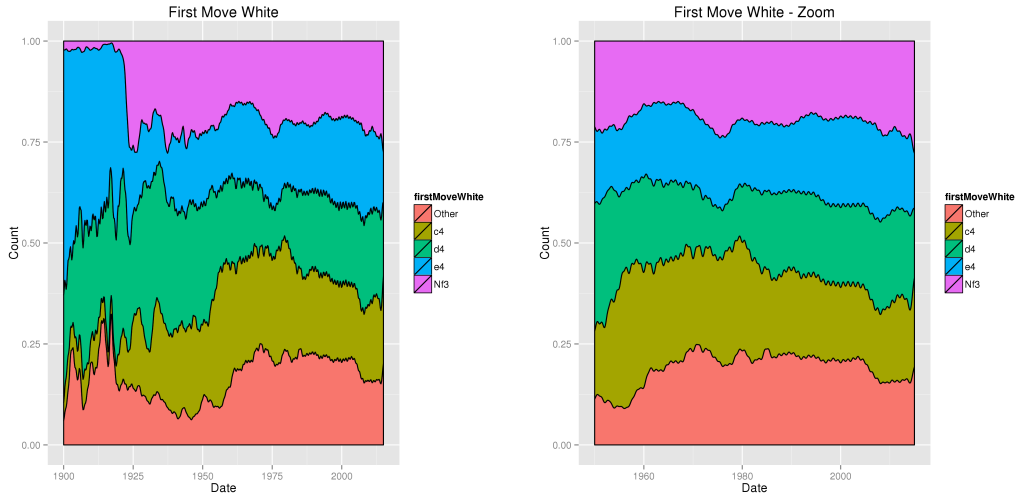
Figure 3: Elo difference and games win (similar probabilities graph/outcomes of Elo rating system)



(a) White players, Elo rating, and games win

(b) Games win, color, and Elo

Figure 4: Colors and percentage win



(a) (b)

Figure 5: First moves (white) depending on the dates

2.6 Distribution of Date Match

Before 1950s, only interesting games are recorded and saved for posterity. Now, amateurs and professionals can easily share PNG files (see, e.g., TWIC). The number of records explodes in the 2000s. A noticeable curve is observed in Figure 6: 81 plys in 1950s, 72 plys in 1970 and about 84 today. We have 679K games recorded after 2010 and the average is 82 ply per game.

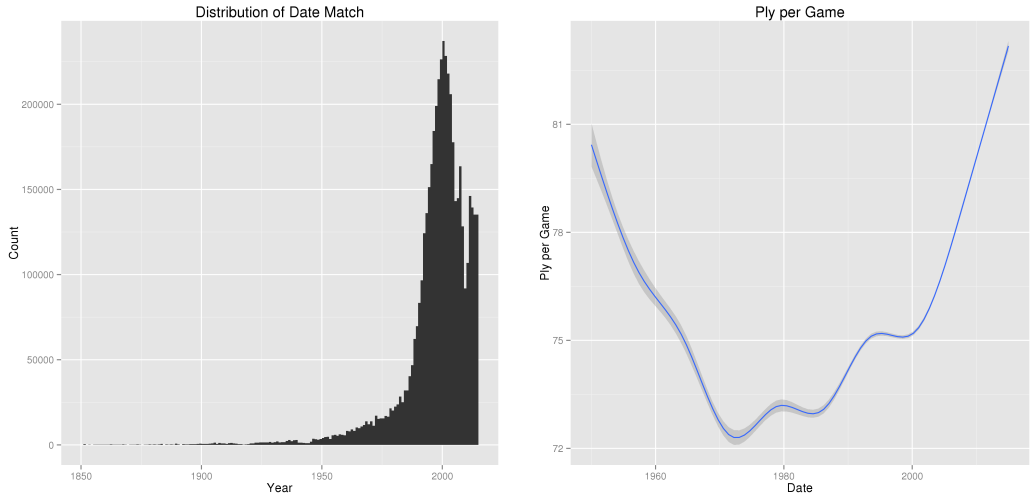
Figure 7a and Figure 7b show that the percentage win by white players tends to decrease steadily even if the first-move still provides a small advantage.

2.7 Comparison with Olson's dataset

Compared to Olson's results, we observe similar properties like:

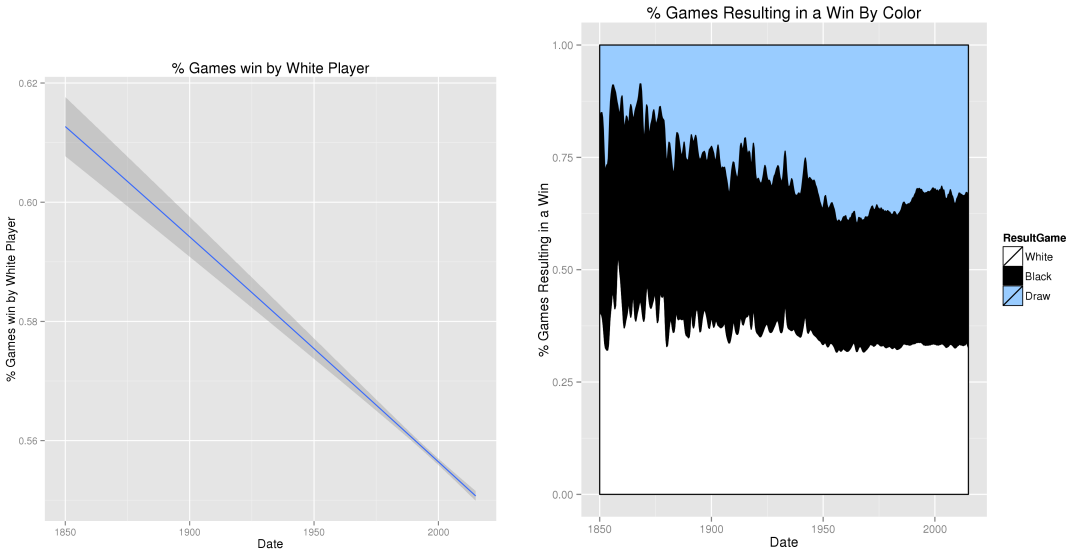
- the distribution of Elo ratings and difference of Elo ratings between two opponents
- the importance of white and first-move advantage
- the increase of draws when chess experts are involved
- the proportion of first moves and openings
- the fact Elo ratings tend to predict game outcome

We also observe some differences; we comment two of them here. First, the number of moves depending on difference Elo Rating (see Figure 2, page 5): Olson's dataset does not exhibit an increase curve between 0 and 200 difference Elo rating. Moreover the number of ply is slightly higher all along the graph. Second, the proportion of games win by white players depending on



(a) Years of games (b) Ply per game depending on the date

Figure 6: Games and years/dates



(a) Percentages games win by white player depending on the date (b) Percentages games win by color depending on the date

Figure 7: Percentages games win by color w.r.t. dates

the date (see Figure 7a, page 9). We concur with the conclusion that having white pieces is an advantage for all periods. However we also observe a linear decrease that is not apparent in Olson's dataset, especially for old periods.

Our conclusion is that there is no fundamental difference, i.e., the number of games considered in the two datasets can explain such differences.

2.8 Summary

Appendix B provides further results related to kinds of moves (promoted rates, queen side castling rates, etc.). Several other information can be extracted from the dataset as well. Our results so far suggest that (1) the database contains numerous interesting games with rather strong players; (2) headers information such as Elo rating or results of the games are coherent. Though some games can certainly be removed or corrected, we consider the dataset is representative of existing chess databases and consistent with chess practices and trends. We also obtain similar properties than in other datasets or databases. Finally, the number of games (almost 5 millions) is significant.

3 Large-scale Analysis of Chess Games with Chess Engines

We have now a better understanding of our database of chess games. We consider that the properties of the dataset are reassuring and justify the analysis of all games by chess engines – our original motivation. In a sense our next objective is to produce a *dynamic* analysis of chess games (by opposition to a static analysis as we made in previous section). The exploitations of the chess engines analysis (for blunder detection, players' ratings, etc.) are left as future work. In this section and preliminary report, we describe how we performed a large-scale analysis.

3.1 Analysis process

270 millions FEN positions. We encoded each game position using the Forsyth-Edwards Notation (FEN) notation. Some positions are equals (including the context leading to the position): a FEN encoding allows us to detect such equality. The underlying idea is that a chess engine is executed only one time per equal position. We also observe that some positions are theoretical openings and quite well-known, presenting limited interests for an analysis: we used Encyclopaedia of Chess Openings (ECO) code to detect such positions. Our dataset originally exhibits 380 millions positions. By exploiting FEN encoding and ECO classification, we only had to consider 270 millions positions. With these simple heuristics, we drastically reduced the number of games to analyze with chess engines.

Stockfish chess engine depth=20, multipv=1. We used Stockfish [14] (version 6) to analyze all FEN positions. Stockfish is an open source project and a very strong chess engine – one of the best at the time of writing [9]. Other researchers have also considered Stockfish in prior works. We used the UCI protocol to collect analysis. Importantly we set the depth to 20. It is a quite high depth (prior works usually set lower depths because of the computation cost). We also set multipv[5] to 1.

[5]UCI chess engines like Stockfish support multi best line or k-best mode. It means they return the k best moves/lines. When k=1, Stockfish returns the best line and evaluation. The increase of multipv is possible and has practical interests (see, e.g., [2]), but it has also a computational and storage cost.

Structuring data. We used a simple database schema (see Figure **??**, page 13). PGN headers informations are stored and structured for retrieving games, positions, players, etc. For each position (FEN), we associate the score and log (multipv=1) computed by Stockfish. We developed several proof-of-concepts to validate the schema: `https://github.com/ChessAnalysis/chess-analysis-database`. For instance, we can gather all positions of a game and depict the scores' evolution.

Distributing the computation. Our experiments suggested that it takes about 6 seconds to analyze a FEN on a basic machine. The use of a single machine was simply not an option since we have to analyze 270 millions position. It would require $270 * 10^6 * 6 = 1620 * 10^6$ seconds, 450K hours, 18K days, and around 50 years of computation. The third step of our process was thus to distribute the computation on a cluster of machines. We used IGRIDA[6], a computing grid available to research teams at IRISA / INRIA, in Rennes. The computing infrastructure has 125 computing nodes and 1500 cores.

Computational and storage cost. We split the FEN positions for distributing the computation on different nodes. We processed in batch (without user intervention) and the analysis was incremental. We used in average 200+ cores during night and day during 2 months. We gathered around 1,5 tera-octets of data (FEN logs).

3.2 Conclusion and Future Work

Our experience showed that it is practically feasible to analyze around 5 millions chess games with state-of-the-art chess engines like Stockfish.

Our next step is naturally to process and exploit data for either replicating existing works or for investigating new directions [2]. The amount of analysis we have collected is superior to what have been considered so far in existing works: We expect to gain further confidence in existing results or methods (e.g., for players' ratings, blunders detection, influence of depth engines, etc.). Another direction is to collect more data. We used Stockfish with depth=20 and multipv=1 (single-pv mode). It has some limitations; for example, some methods/applications require multipv mode. We can rely on other chess engines. We can also use different settings for Stockfish (e.g., a higher depth and multipv). A possible threat is that the computation and storage cost can be very important.

We believe more data can help to better assess methods based on chess engines. From this perspective, we are happy to share our results with scientists, chess experts or simply data hobbyists. We hope to confront methods or interpretations and have different perspectives on data. More information can be found online: `https://github.com/ChessAnalysis/chess-analysis`

Our long-term goal is to better understand the underlying beauty and complexity of chess thanks to the incredible skills of contemporary chess engines. The rise of computing power, software, and chess data can also be seen as an opportunity to address very old questions in the fields of cognitive and computer science (e.g., artificial intelligence, computational complexity, data and software engineering).

Acknowledgement. We would like to thank our colleagues at DiverSE (`http://diverse.irisa.fr`) for their discussions and feedbacks.

[6] `http://igrida.gforge.inria.fr/`

References

[1] David J. Barnes and Julio Hernandez-Castro. On the limits of engine analysis for cheating detection in chess. *Computers and Security*, 48:58 – 73, 2015.

[2] T. Biswas, G. Haworth, and K. Regan. A comparative review of skill assessment: Performance, prediction and profiling. In *14th Advances in Computer Games conference*, 2015.

[3] Tamal T. Biswas and Kenneth W. Regan. Quantifying depth and complexity of thinking and knowledge. In *ICAART 2015 - Proceedings of the International Conference on Agents and Artificial Intelligence, Volume 2*, pages 602–607, 2015.

[4] Diogo R. Ferreira. The impact of search depth on chess playing strength. *ICGA Journal No. 2*, 36(2), june 2013.

[5] Matej Guid, Aritz Pérez, and Ivan Bratko. How trustworthy is crafty's analysis of world chess champions. *ICGA journal*, 31(3):131–144, 2008.

[6] Guy McCrossan Haworth et al. Gentlemen, stop your engines! *ICGA Journal*, 30(3):150–156, 2007.

[7] Feng-Hsiung Hsu. *Behind Deep Blue: Building the computer that defeated the world chess champion*. Princeton University Press, 2002.

[8] Matthew Lai. Giraffe: Using deep reinforcement learning to play chess. *CoRR*, abs/1509.01549, 2015.

[9] CCRL 40/40 Rating List. `http://www.computerchess.org.uk/ccrl/4040/`, 2015.

[10] K. Regan. A computer program to detect possible cheating in chess `http://www.nytimes.com/2012/03/20/science/a-computer-program-to-detect-possible-cheating-in-chess.html?_r=0`, 2012.

[11] Kenneth W. Regan. Depth of satisficing `https://rjlipton.wordpress.com/2015/10/06/depth-of-satisficing/`, october 2015.

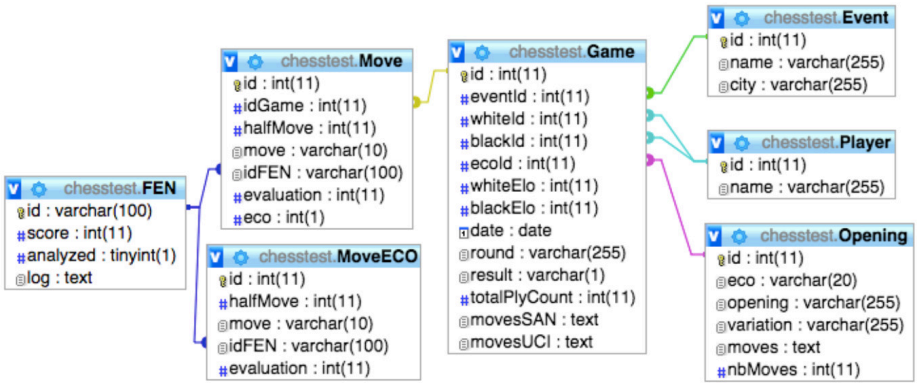
[12] Kenneth W. Regan, Tamal Biswas, and Jason Zhou. Human and computer preferences at chess. In *8th Multidisciplinary Workshop on Advances in Preference Handling (MPREFS 2014), associated to AAAI 2014*, 2014.

[13] Kenneth Wingate Regan and Guy McCrossan Haworth. Intrinsic chess ratings. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, 2011.

[14] Stockfish. `https://stockfishchess.org/`. last access: april 2016.

[15] Han LJ Van Der Maas and Eric-Jan Wagenmakers. A psychometric analysis of chess expertise. *The American journal of psychology*, pages 29–60, 2005.

A Database schema



B Misc: checkmates, captured pieces, promoted rates, and kinds of moves

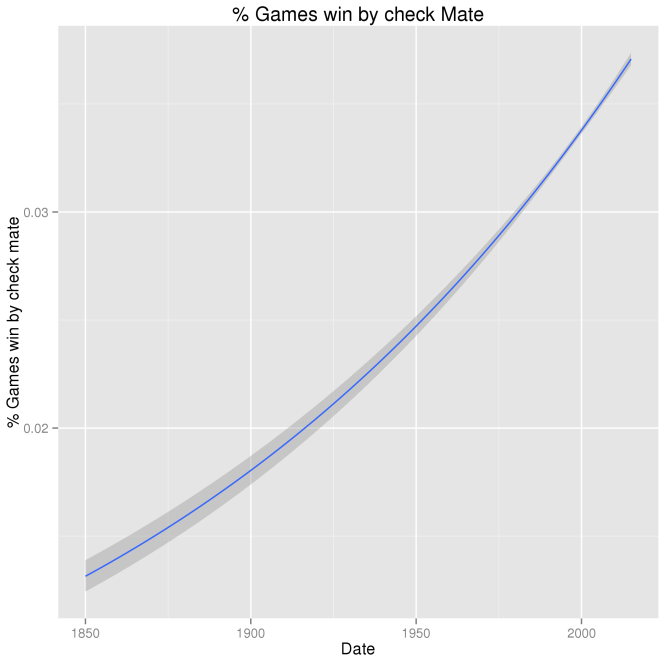


Figure 8: There are only a few checkmates during games because players typically resign when they realize the defeat is near. The slight increase is surprising and deserves more investigations (e.g., a possible explanation is the inclusion of recent rapid games like Blitz in the dataset).

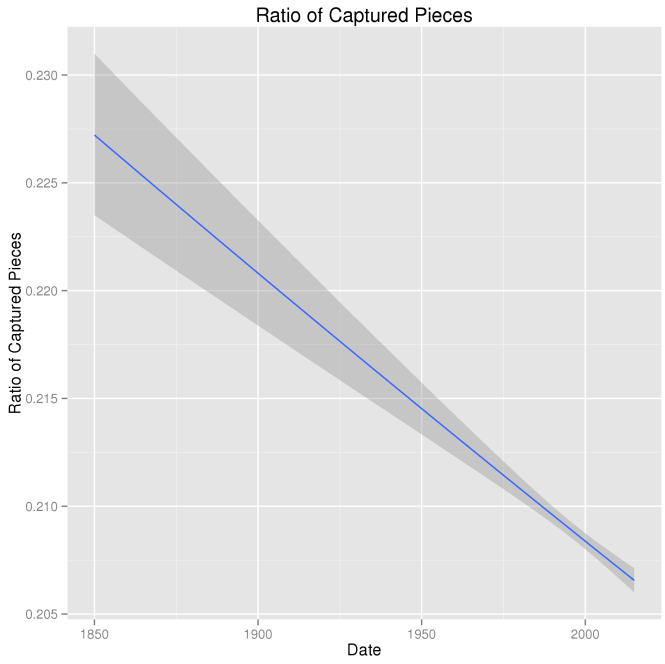


Figure 9: The ratio of captured pieces during a game has slowly decreased

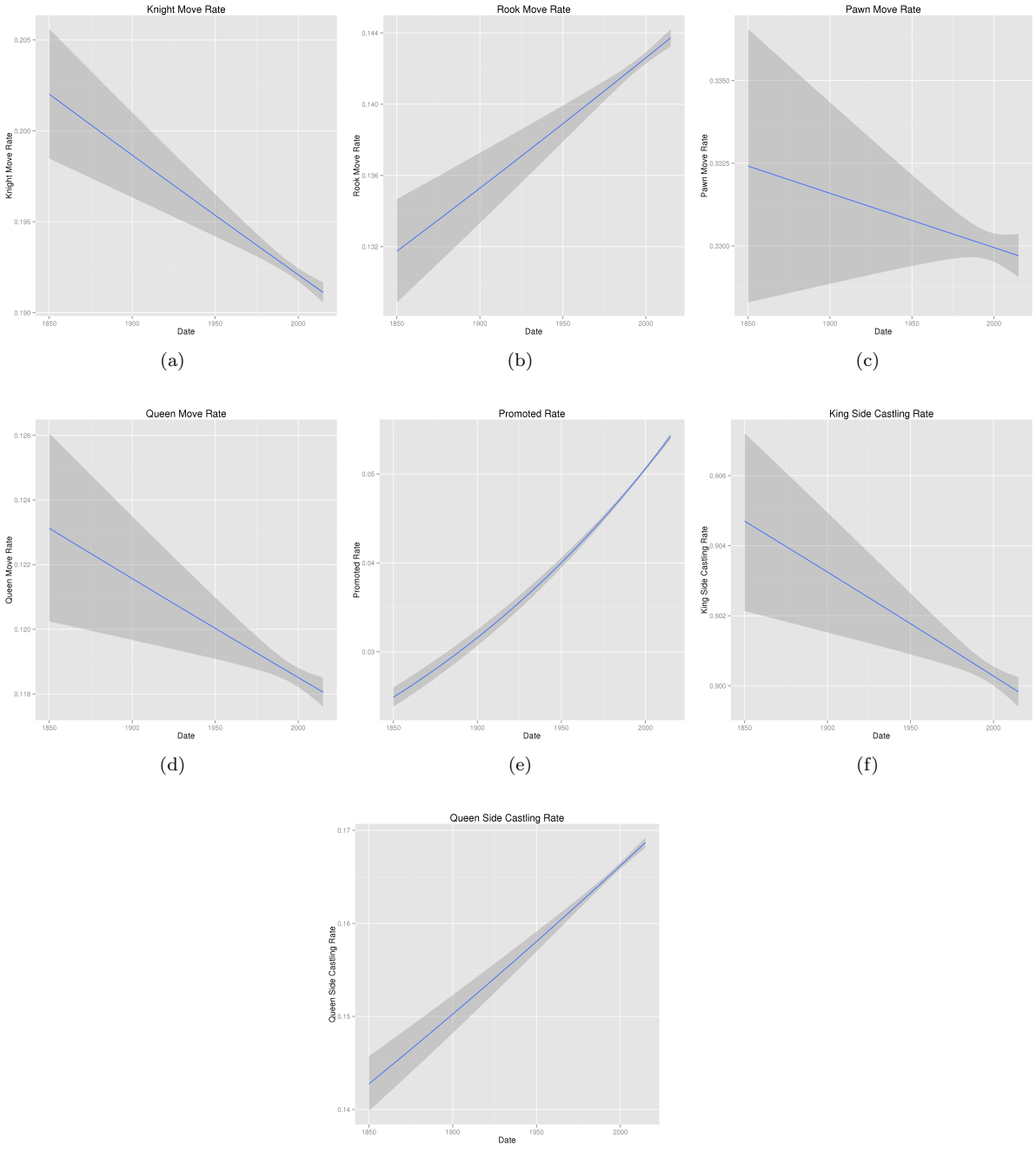
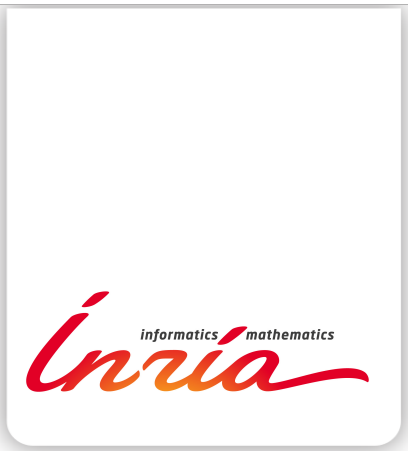


Figure 10: Pieces' rates



RESEARCH CENTRE
RENNES – BRETAGNE ATLANTIQUE

Campus universitaire de Beaulieu
35042 Rennes Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-0803