How do weak chess engines perform and potentially cheat?

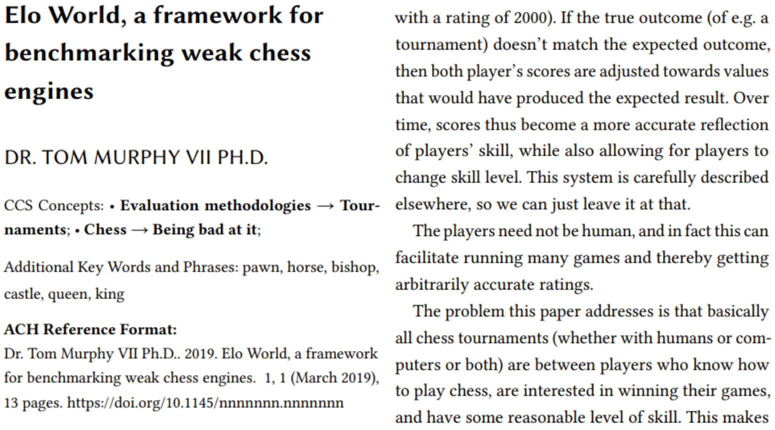
Sports modelling reading group

Clement Lee

2022-11-29

Some weak chess engines

Figure 1: Source: Wikipedia



Elo World, a framework for benchmarking weak chess engines

DR. TOM MURPHY VII PH.D.

CCS Concepts: • **Evaluation methodologies → Tournaments**; • **Chess → Being bad at it**;

Additional Key Words and Phrases: pawn, horse, bishop, castle, queen, king

ACH Reference Format:
Dr. Tom Murphy VII Ph.D.. 2019. Elo World, a framework for benchmarking weak chess engines. 1, 1 (March 2019), 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

with a rating of 2000). If the true outcome (of e.g. a tournament) doesn't match the expected outcome, then both player's scores are adjusted towards values that would have produced the expected result. Over time, scores thus become a more accurate reflection of players' skill, while also allowing for players to change skill level. This system is carefully described elsewhere, so we can just leave it at that.

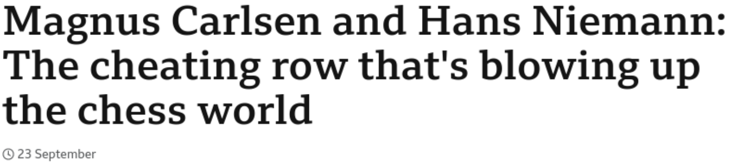
The players need not be human, and in fact this can facilitate running many games and thereby getting arbitrarily accurate ratings.

The problem this paper addresses is that basically all chess tournaments (whether with humans or computers or both) are between players who know how to play chess, are interested in winning their games, and have some reasonable level of skill. This makes

Tom Murphy VII, *Elo world, a framework for benchmarking weak chess engines*, A Record of the Proceedings of SIGBOVIK, 2019.

Motivation

- Modelling performance of chess grandmasters
- Detecting cheating, online or over the board



Magnus Carlsen and Hans Niemann: The cheating row that's blowing up the chess world

23 September

Figure 2: https://www.bbc.co.uk/news/world-63010107

Outline

1. Some basics
2. Modelling performance by outcome
   - + collusion detection
3. Modelling performance by moves
   - + cheating detection

Some basics

A standard board

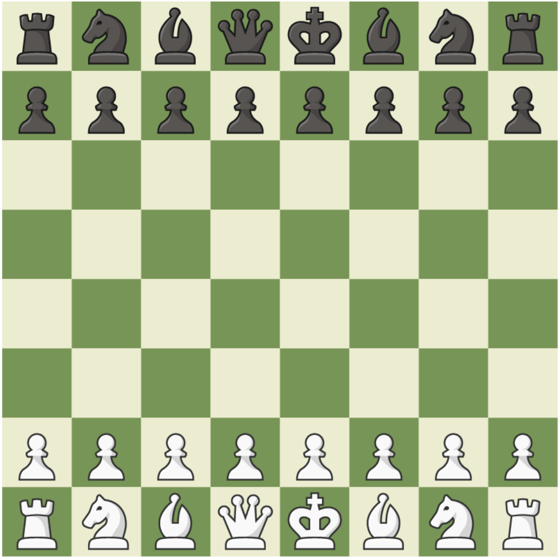


Figure 3: Source: https://www.chess.com/

How the pieces move

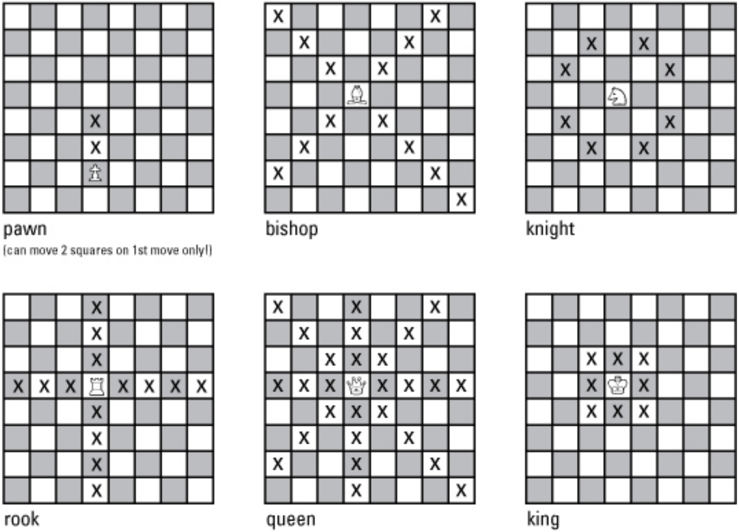
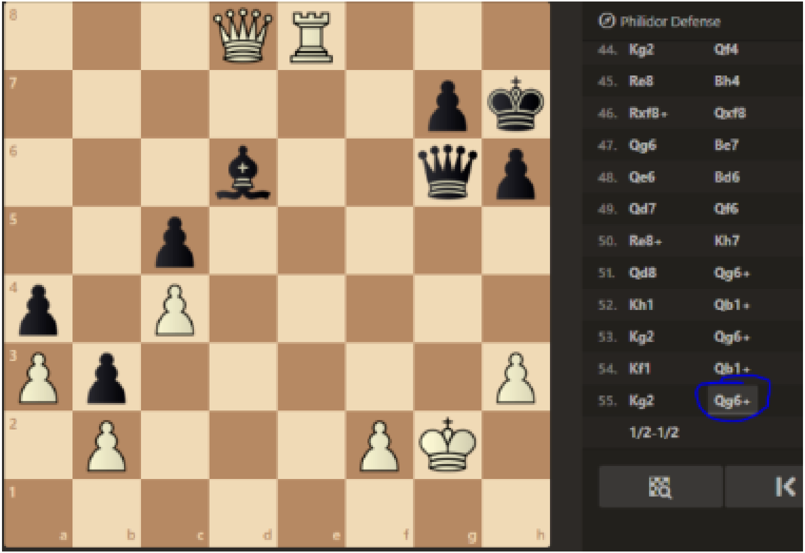


Figure 4: Source: https://www.dummies.com/

Draws can happen



Philidor Defense

44.	Kg2	Qf4
45.	Re8	Bh4
46.	Rxf8+	Qxf8
47.	Qg6	Be7
48.	Qe6	Bd6
49.	Qd7	Qf6
50.	Re8+	Kh7
51.	Qd8	Qg6+
52.	Kh1	Qb1+
53.	Kg2	Qg6+
54.	Kf1	Qb1+
55.	Kg2	Qg6+
	1/2-1/2	

Figure 5: Source: https://support.chess.com/

Most common result in high level games

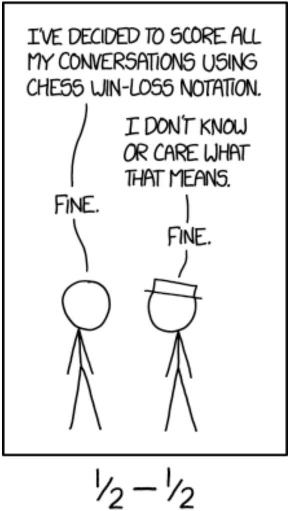
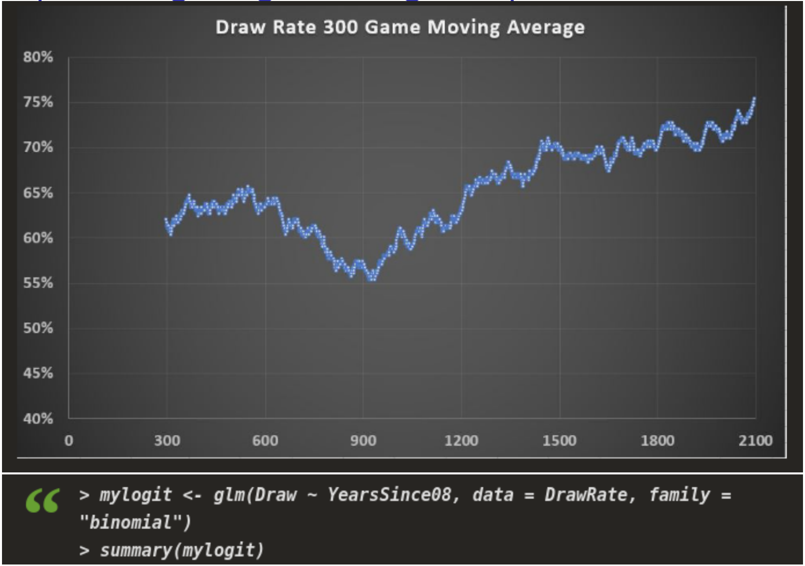

I'VE DECIDED TO SCORE ALL MY CONVERSATIONS USING CHESS WIN-LOSS NOTATION.
I DON'T KNOW OR CARE WHAT THAT MEANS.
FINE.
FINE.
½–½

Figure 6: https://xkcd.com/1800/

Draw percentage might have gone up



```
> mylogit <- glm(Draw ~ YearsSince08, data = DrawRate, family =
"binomial")
> summary(mylogit)
```

Figure 7:
https://www.chess.com/article/view/the-draw-rule-is-classical-chess-dead

Another says probably not

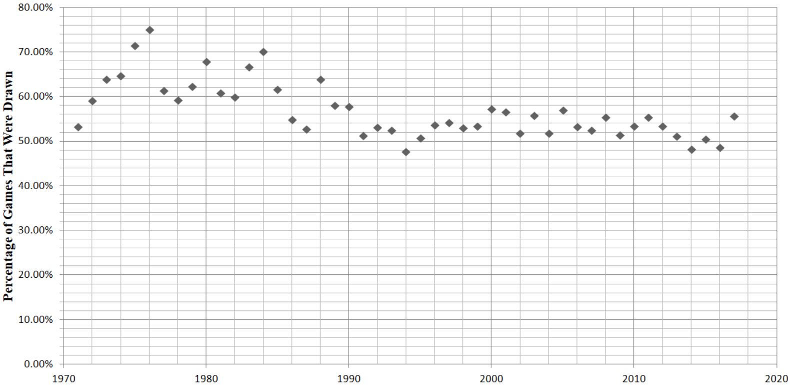


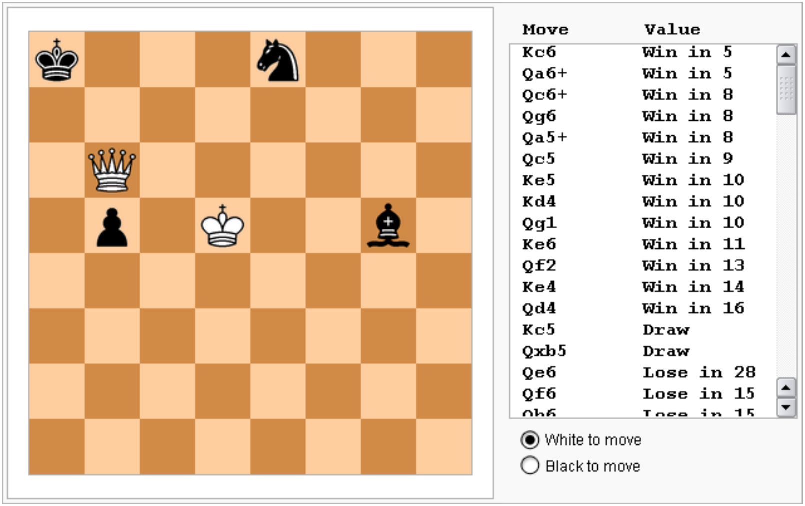
Figure 8: https://en.chessbase.com/post/has-the-number-of-draws-in-chess-increased

Some observations

- Quick comparison:
  - The one says yes: Elo 2750+
  - The one says no: Elo 2600+
  - (One requirement for becoming a GM: Elo 2500+)
- Some recent tournaments reward taking risks
  - 3 points for win, 1 point for draw
  - More points for winning outright instead of winning in tiebreaks
- Super GMs prepare more thoroughly
  - More averse to taking risks, which increase the chance of a decisive result either way
  - With the help of computer engines

Computer engines

Before computer engines



Move	Value
Kc6	Win in 5
Qa6+	Win in 5
Qc6+	Win in 8
Qg6	Win in 8
Qa5+	Win in 8
Qc5	Win in 9
Ke5	Win in 10
Kd4	Win in 10
Qg1	Win in 10
Ke6	Win in 11
Qf2	Win in 13
Ke4	Win in 14
Qd4	Win in 16
Kc5	Draw
Qxb5	Draw
Qe6	Lose in 28
Qf6	Lose in 15
Qh6	Lose in 15

● White to move

○ Black to move

Figure 9: https://en.wikipedia.org/wiki/Endgame_tablebase

There is no 32-piece tablebase (yet)

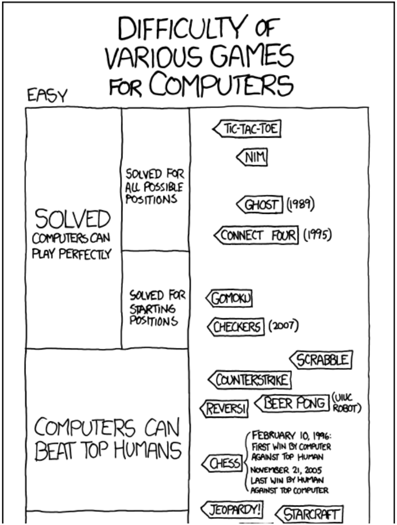

DIFFICULTY OF VARIOUS GAMES FOR COMPUTERS
EASY
SOLVED COMPUTERS CAN PLAY PERFECTLY
SOLVED FOR ALL POSSIBLE POSITIONS
SOLVED FOR STARTING POSITIONS
COMPUTERS CAN BEAT TOP HUMANS
TIC-TAC-TOE
NIM
GHOST (1989)
CONNECT FOUR (1995)
GOMOKU
CHECKERS (2007)
SCRABBLE
COUNTERSTRIKE
REVERSI
BEER PONG (UIUC ROBOT)
CHESS
FEBRUARY 10, 1996: FIRST WIN BY COMPUTER AGAINST TOP HUMAN
NOVEMBER 21, 2005 LAST WIN BY HUMAN AGAINST TOP COMPUTER
JEOPARDY!
STARCRAFT

Figure 10: https://xkcd.com/1002/ (cropped)

Human chess is not dead because

- Chess is not completely solved
- Weak chess engines make mistakes (esp. under time pressure)

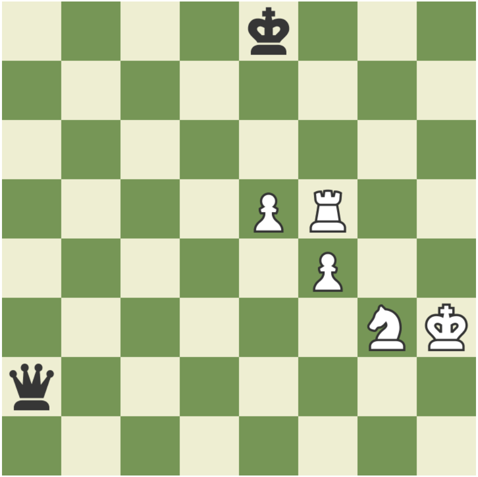
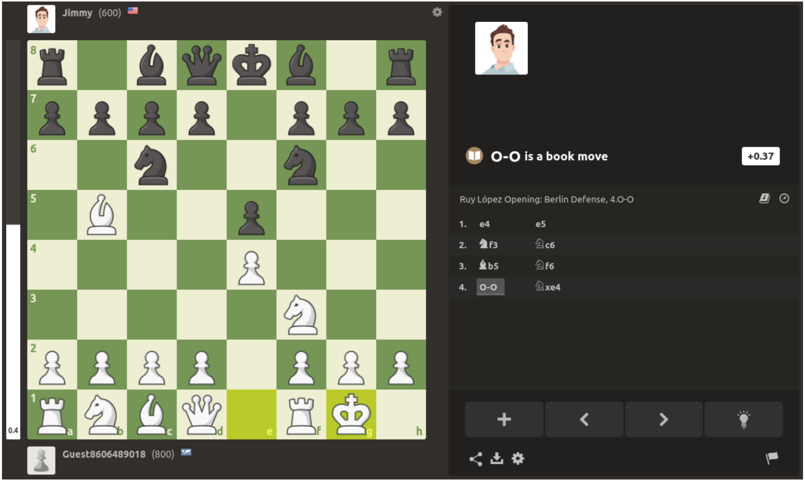


Figure 11: Move 130, Game 6, FIDE World Chess Championship 2021

So how do computer engines work?

- Evaluate the *position*
- Assign a value ≈ the number of (unpassed) pawns



Jimmy (600)

O-O is a book move +0.37

Ruy López Opening: Berlin Defense, 4.O-O

1.	e4	e5
2.	♞f3	♘c6
3.	♝b5	♘f6
4.	O-O	♘xe4

Guest8606489018 (800)

Figure 12: A positive value here means white is better

Advantage over the moves

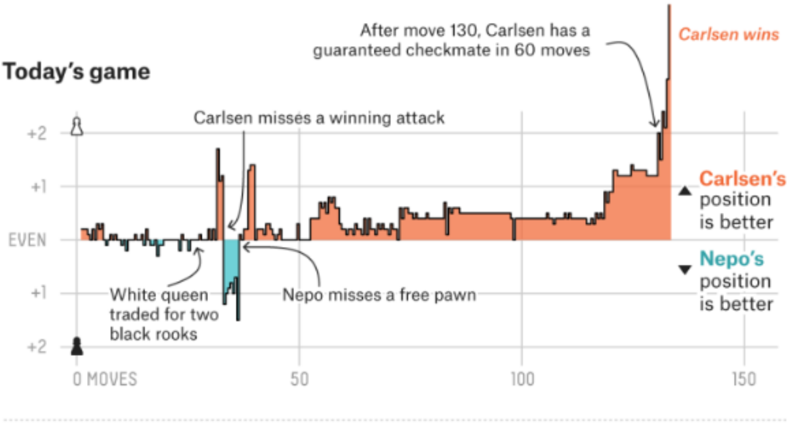


Figure 13: Source: https://fivethirtyeight.com

Modelling the performance by outcome and collusion detection

Another xkcd picture

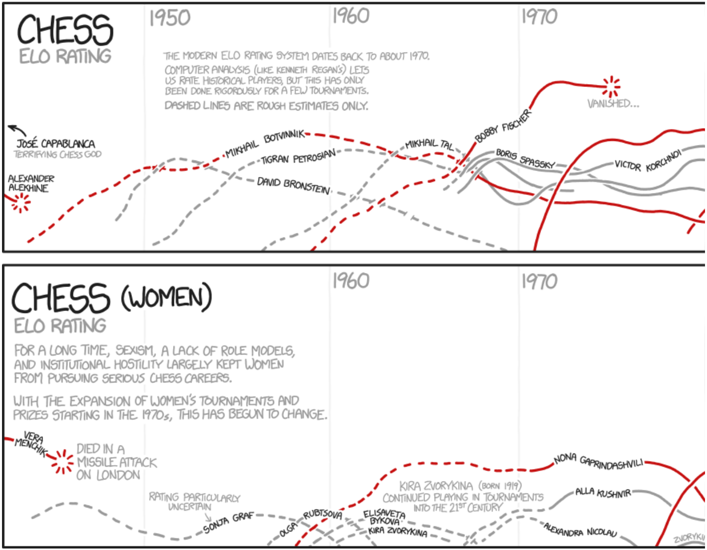


Figure 14: https://xkcd.com/1392/

Inflation or getting stronger?

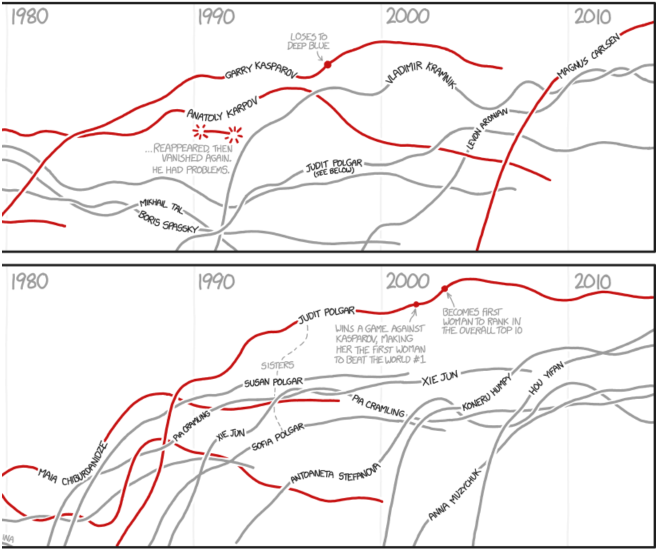


Figure 15: https://xkcd.com/1392/

Divinsky and Keene (1989)

- Book: *Warriors of the Mind*
- Data: games between only 64 of the greatest players
- Bradley-Terry model
  - a draw counts half of a win

Henery (1992)

- Criticises Divinsky and Keene (1989)
  - of the use of selective data
  - of the treatment of draws
- Noted systematic increase in the proportion of draws
- Thurstone-Mosteller model
  - Allows differing draw proportions of the players
- Suggests other potentially important factors
  - The length of game (number of moves)
  - Age of the players

Caron and Doucet (2012)

- Bradley-Terry Model
    - Also allows ties
- Focus on the efficient Gibbs sampler upon data augmentation
- No ranking results presented
    - Only plots of ACF and test set RMSE

Trying to unify the models

- The variables
  - Y_i: performance of i-th player
  - $X = Y_i - Y_j$
- The parameters
  - For Y_i: $\lambda_i > 0$, or equivalently $\mu_i = \log \lambda_i$
  - Draw thresholds: $\delta_{ij} \geq 0$, $\delta_{ji} = -\delta_{ij}$
  - σ additionally for Thurstone-Mosteller
- Scenarios
  - i beats $j \quad \Leftrightarrow \quad X > \delta_{ij}$
  - i draws $j \quad \Leftrightarrow \quad \delta_{ji} < X < \delta_{ij}$
- When $\delta_{ij} = \delta_{ji} = 0$, no draws

If Y_i is normally distributed with mean μ_i

$$\begin{aligned} X &\sim \mathrm{N}(\mu_i - \mu_j, \sigma) \\ \Pr(i \text{ beats } j) &= \Pr(X > \delta_{ij}) \\ &= \Pr\left(\frac{X - (\mu_i - \mu_j)}{\sigma} > \frac{\delta_{ij} - (\mu_i - \mu_j)}{\sigma}\right) \\ &= \Pr\left(\frac{X - (\mu_i - \mu_j)}{\sigma} \le \frac{(\mu_i - \mu_j) - \delta_{ij}}{\sigma}\right) \\ &= \Phi\left(\sigma^{-1}(\mu_i - \mu_j - \delta_{ij})\right) \end{aligned}$$

- This is what Henery (1992) used

Another distribution function

- Replace $\Phi(z)$ by $e^z/(e^z+1)$
  - σ set to 1 (essentially removed)

$$
\begin{aligned}
\Pr(i \text{ beats } j) &= \frac{e^{\mu_i-\mu_j-\delta_{ij}}}{e^{\mu_i-\mu_j-\delta_{ij}}+1} \\
&= \frac{e^{\mu_i}}{e^{\mu_i}+e^{\mu_j}e^{\delta_{ij}}} \\
&= \frac{\lambda_i}{\lambda_i+\lambda_j\theta_{ij}}
\end{aligned}
$$

- $\theta_{ij} = e^{\delta_{ij}} \geq 1$ as $\delta_{ij} \geq 0$
- This is what Caron and Doucet (2012) used

Some modelling considerations

- Incorporating the rise of engines?
    - A single changepoint
    - Age of the players
    - Date the game was played
- Draw probability depends on the players?
    - In Henery (1992), $\delta_{ij} = \delta_i + \delta_j$
    - In Caron and Doucet (2012), $\delta_{ij} = \delta$

Hankin (2020)

- Bradley-Terry Model
  - Incorporates 1) a “draw monster” & 2) white’s advantage

$$
\begin{aligned}
\Pr(i \text{ beats } j \text{ with white pieces}) &= \frac{\lambda_i + \omega}{\lambda_i + \lambda_j + \omega + \theta} \\
\Pr(i \text{ beats } j \text{ with black pieces}) &= \frac{\lambda_i}{\lambda_i + \lambda_j + \omega + \theta} \\
\Pr(i \text{ draws } j) &= \frac{\theta}{\lambda_i + \lambda_j + \omega + \theta} \\
\sum_i \lambda_i + \omega + \theta = 1
\end{aligned}
$$

- CRAN package hyper2
- Data: World Chess Championship 1963

Incorporating collusion allegation

- Fischer claimed that Keres, Petrosian and Geller colluded
- A different draw monster θ^* for the concerned players
  - No mention of how the "sum to 1" constraint is dealt with
- Testing equality with that of the remaining players
- Also applied to another data (interzonal Stockholm 1962)
  - Soviet players drew more than the rest

Modelling strength by moves and cheating detection

A bit of history

Figure 16: Hans Niemann admitted to cheating on chess.com in 2020

The alleged coach/mentor

Figure 17: Maxim Dlugy, suspected of & admitted to online cheating twice

World Champion

Figure 18: Carlsen dominated classical chess for over 10 years

Sinquefield Cup 2022

- Saint Louis Chess Club, Missouri
- 10 players, single round robin, over the board

Figure 19: Carlsen & Niemann played in the 3rd round

Aftermath

- 2022-09-04: Carlsen lost with white pieces to Niemann
  - Carlsen's 53-game unbeaten streak in classical over the board tournaments ended
  - Niemann's live Elo rating surpassed 2700 for the first time
- 2022-09-05: Carlsen withdrew from the tournament, while organisers put in more anti-cheating measures
- 2022-09-21: Carlsen said he's "impressed by Niemann's play" and that Dlugy "must be doing a great job"
- 2022-09-26: Carlsen issued a statement accusing Niemann of cheating
- 2022-10-04: Chess.com issued a report and banned Niemann
- 2022-10-20: Niemann filed a $100 million defamation lawsuit against Carlsen, Chess.com and some others

Circumstantial evidence for a statistical anomaly

- Niemann, in a post-match interview, claimed that he prepared the line (sequence of moves) after watching a Carlsen game
  - But he remembered the place and year incorrectly
  - Some found him nervous in the interview and fumbled explaining the lines
- Some questioned how Niemann miraculously prepared this particular line
- No evidence was found that
  - the line was leaked to Niemann
  - Niemann cheated during the game
- Niemann denied ever cheating over the board

Statistical framework

- Is there a way of statistically inferring if somebody cheated or not?
- Frequentist: test H_0: Niemann didn't cheat, using games data
- Bayesian: calculate $\Pr(\text{Niemann cheated}|\text{games data})$
- Required:
  - a model that describes how moves are made with or without cheating
  - understanding of the behaviour / pattern of cheating
  - understanding of how human plays in general

The human factor

- Time pressure
- Not remembering / mixing up a line
- Surprising opponent with a sub-optimal move
- **Fallibility**

Ken Regan

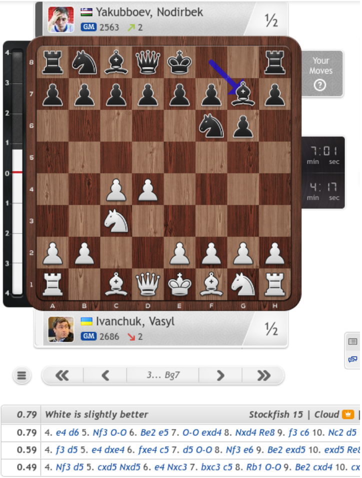
Figure 20: Credit: Sinna Nasseri for TIME

Di Fatta, Haworth, and Regan (2009)

- **“Skill rating by Bayesian inference”**
- The moves are the data
- Compare against a benchmark i.e. computer engine
- Probabiliity of making a particular move depends on:
  - The top candidate moves suggested by engine
  - The values associated with these moves
  - The strength λ
  - Some other (hyper)parameters

$$\Pr(\text{candidate move } i) \propto (v_{\max} - v_i + K)^{-\lambda}$$

Toy example



0.79	White is slightly better Stockfish 15 | Cloud
0.79	4. e4 d6 5. Nf3 O-O 6. Be2 e5 7. O-O exd4 8. Nxd4 Re8 9. f3 c6 10. Nc2 d5
0.59	4. f3 d5 5. e4 dxe4 6. fxe4 c5 7. d5 O-O 8. Nf3 e6 9. Be2 exd5 10. exd5 Re
0.49	4. Nf3 d5 5. cxd5 Nxd5 6. e4 Nxc3 7. bxc3 c5 8. Rb1 O-O 9. Be2 cxd4 10. cx

Figure 21: Source: https://chess24.com

- $v_{\max} = 0.79, \quad K = 0.1$
- Probability of e4 $\propto$

$$(0.79 - 0.79 + 0.1)^{-\lambda}$$

- Probability of f3 $\propto$

$$(0.79 - 0.59 + 0.1)^{-\lambda}$$

- Probability of Nf3 $\propto$

$$(0.79 - 0.49 + 0.1)^{-\lambda}$$

- Larger $\lambda \Rightarrow$ more able to pick out best move

Bayesian inference

- Iteratively apply the Bayesian rule
- $\mathcal{F}_n$: moves 1 to $n \Rightarrow \mathcal{F}_n = \{F_{n-1}, \text{move } n\}$

$$\begin{aligned}\pi(\lambda|\mathcal{F}_1) &= \pi(\lambda|\mathcal{F}_0, \text{move } 1) \propto \Pr(\text{move } 1|\lambda)\pi(\lambda) \\ \pi(\lambda|\mathcal{F}_2) &= \pi(\lambda|\mathcal{F}_1, \text{move } 2) \propto \Pr(\text{move } 2|\lambda, \mathcal{F}_1)\pi(\lambda|\mathcal{F}_1)\end{aligned}$$

- They did this so that they can obtain (the distribution of) λ over the moves, instead of just one at the end of the game

Consistent with Elo

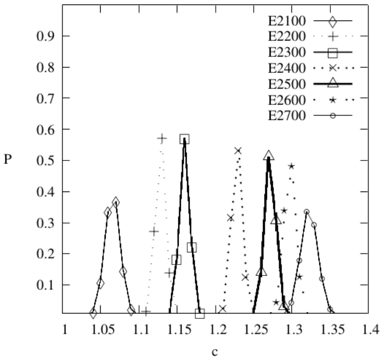
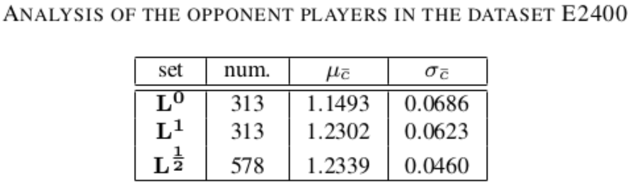


Figure 22: Figure 1 of Di Fatta, Haworth, and Regan (2009)

Compare skill rating

- Between players of similar Elo rating



ANALYSIS OF THE OPPONENT PLAYERS IN THE DATASET E2400

set	num.	$\mu_{\bar{c}}$	$\sigma_{\bar{c}}$
$\mathbf{L^0}$	313	1.1493	0.0686
$\mathbf{L^1}$	313	1.2302	0.0623
$\mathbf{L^{\frac{1}{2}}}$	578	1.2339	0.0460

Figure 23: Table 3 of Di Fatta, Haworth, and Regan (2009)

Consistent with increasing draw percentage?

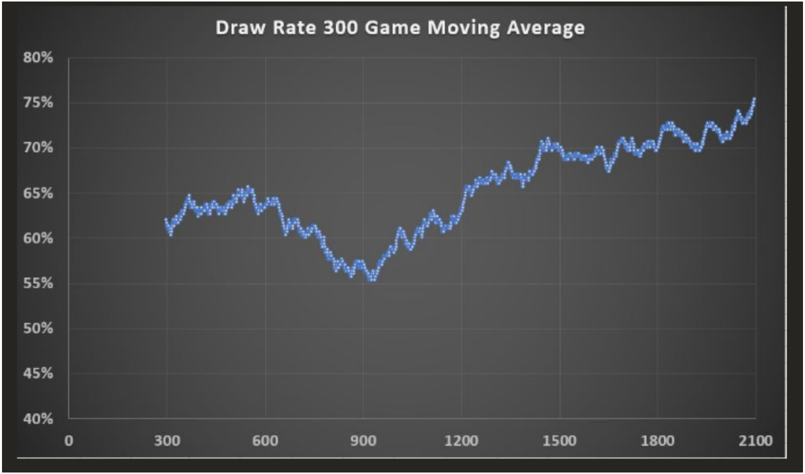


Figure 24: Previously shown figure on increasing draw percentage

- Different ranges of Elo rating though

Two other papers

- Haworth, Regan, and Di Fatta (2010)
- Haworth, Biswas, and Regan (2015)
- Not much being offered on cheating detection apart from **Move Matching**
  - Percentage of matches with computer engine moves
  - This is what some focus on when accusing Niemann of cheating
- I'm still none the wiser

Regan's more recent analysis

- **Raw outlier index**, based on move matching
  - 50: perform at Elo rating
  - < 50: worse than Elo rating
  - > 50: better than Elo rating
- Niemann's performance doesn't show anomaly
- This requires Elo rating being accurate & up-to-date

A smart cheater

- Just cheat at some key moves
- Only cheat at one or two games in a tournament
- Regan's methods at best **help** to keep chess fair

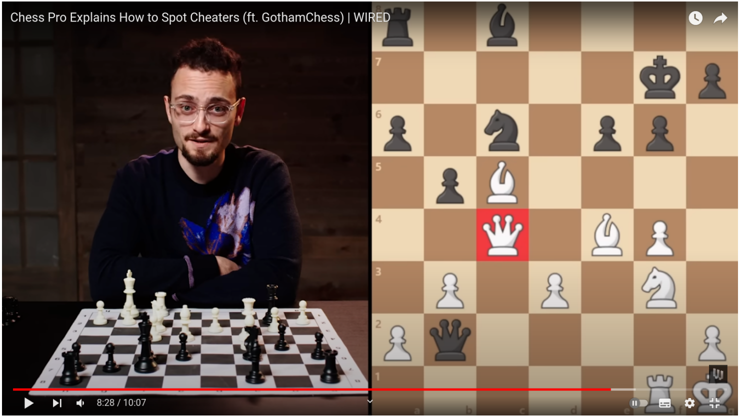
Coming back to the basics

- People mixing up the two

$$
\begin{aligned}
&\Pr(\text{cheat}|\text{win}) \\
= \; &\frac{\Pr(\text{cheat})\,\Pr(\text{win}|\text{cheat})}{\Pr(\text{cheat})\,\Pr(\text{win}|\text{cheat}) + (1 - \Pr(\text{cheat}))\,\Pr(\text{win}|\text{not cheat})}
\end{aligned}
$$

- High prior Pr(cheat) by some
- Pr(win|not cheat) by strength
  - Either outcome or skill rating
- Pr(win|cheat) difficult to determine
  - Need to understand how to do it over the board

Last night



Chess Pro Explains How to Spot Cheaters (ft. GothamChess) | WIRED
8:28 / 10:07
7
6
5
4
3
2
1
a
b
c
d
e
f
g

Figure 25: https://www.youtube.com/watch?v=P4LnwRHGIHg

(More) final thoughts

- Different techniques to cheat online and over the board
  - Unfair Pr(cheat) for Niemann?
- Same / similar instantaneous outcome
  - One of the top moves
  - Elevated λ?
- "A human would not play this kind of moves"
  - **Style** of player from historical games
  - Moves closer to computer engine or their own style

Bibliography I

Caron, François, and Arnaud Doucet. 2012. "Efficient Bayesian Inference for Generalized Bradley-Terry Models." *Journal of Computational and Graphical Statistics* 21 (1): 174–96. https://doi.org/10.1080/10618600.2012.638220.

Di Fatta, Giuseppe, Guy McC Haworth, and Kenneth W Regan. 2009. "Skill Rating by Bayesian Inference." In *2009 Ieee Symposium on Computational Intelligence and Data Mining*, 89–94. https://doi.org/10.1109/CIDM.2009.4938634.

Divinsky, Nathan, and Raymond Keene. 1989. *Warriors of the Mind: A Quest for the Supreme Genius of the Chess Board*. Brighton: Hardinge Simpole.

Hankin, Robin K S. 2020. "A Generalization of the Bradley-Terry Model for Draws in Chess with an Application to Collusion." *Journal of Economic Behavior and Organization* 180: 325–33. https://doi.org/10.1016/j.jebo.2020.10.015.

Bibliography II

Haworth, Guy, Tamal Biswas, and Ken Regan. 2015. "A Comparative Review of Skill Assessment: Performance, Prediction and Profiling." In *Advances in Computer Games*, edited by Aske Plaat, Jaap van den Herik, and Walter Kosters, 135–46. Cham: Springer International Publishing.

Haworth, Guy, Ken Regan, and Giuseppe Di Fatta. 2010. "Performance and Prediction: Bayesian Modelling of Fallible Choice in Chess." In *Advances in Computer Games*, edited by H. Jaap van den Herik and Pieter Spronck, 99–110. Berlin, Heidelberg: Springer Berlin Heidelberg.

Henery, Robert J. 1992. "An Extension to the Thurstone-Mosteller Model for Chess." *Journal of the Royal Statistical Society: Series D (the Statistician)* 41: 559–67.