Skill Inference and Chess Cheating Detection from Big Data
Cornell University Applied Math Colloquium

Kenneth W. Regan
University at Buffalo (SUNY)

16 Nov., 2012

\(^1\)Includes joint work with Guy Haworth and GM Bartlomiej Macieja. Sites: http://www.cse.buffalo.edu/~regan/chess/fidelity/ (my homepage links), http://www.cse.buffalo.edu/~regan/chess/ratings/ (not yet linked).
Domain: A set of decision-making situations $t$. Chess game turns
A Predictive Analytic Model

1. **Domain:** A set of decision-making situations $t$. 
   Chess game turns

2. **Inputs:** Values $v_i$ for every option at turn $t$. 
   Computer values of moves $m_i$
A Predictive Analytic Model

1. **Domain**: A set of decision-making situations $t$. Chess game turns

2. **Inputs**: Values $v_i$ for every option at turn $t$. Computer values of moves $m_i$

3. **Parameters**: $s, c, \ldots$ denoting skills and levels. Trained correspondence to chess Elo rating $E$. 

---

Skill Inference and Chess Cheating Detection from Big Data
A Predictive Analytic Model

1. Domain: A set of decision-making situations $t$. Chess game turns
2. Inputs: Values $v_i$ for every option at turn $t$. Computer values of moves $m_i$
3. Parameters: $s, c, \ldots$ denoting skills and levels. Trained correspondence to chess Elo rating $E$
4. Defines fallible agent $P(s, c, \ldots)$. 
A Predictive Analytic Model

1. **Domain:** A set of decision-making situations $t$. Chess game turns.

2. **Inputs:** Values $v_i$ for every option at turn $t$. Computer values of moves $m_i$.

3. **Parameters:** $s, c, \ldots$ denoting skills and levels. Trained correspondence to chess Elo rating $E$.

4. **Defines fallible agent** $P(s, c, \ldots)$.

5. **Main Output:** Probabilities $p_{t,i}$ for $P(s, c, \ldots)$ to select option $i$ at time $t$. 

**Derived Outputs:** Aggregate statistics: move-match $MM$, average error $AE$, \ldots. Projected confidence in intervals for those statistics. Intrinsic Performance Ratings (IPR's).
A Predictive Analytic Model

1. Domain: A set of decision-making situations \( t \).
   Chess game turns

2. Inputs: Values \( v_i \) for every option at turn \( t \).
   Computer values of moves \( m_i \)

3. Parameters: \( s, c, \ldots \) denoting skills and levels.
   Trained correspondence to chess Elo rating \( E \)

4. Defines fallible agent \( P(s, c, \ldots) \).

5. Main Output: Probabilities \( p_{t,i} \) for \( P(s, c, \ldots) \) to select option \( i \) at time \( t \).

6. Derived Outputs:
   - Aggregate statistics: move-match MM, average error AE, \ldots
   - Projected confidence intervals for those statistics.
   - “Intrinsic Performance Ratings” (IPR’s).
Main Principle and Schematic Equation

The probability $\Pr(m_i \mid s, c, \ldots)$ depends on the value of move $m_i$ in relation to the values of other moves.

- Too Simple:

$$\Pr(m_i \mid s, c, \ldots) \sim g(s, c, \text{val}(m_i)).$$

Doesn’t take values of the other moves into account.
Main Principle and Schematic Equation

The probability $\Pr(m_i \mid s, c, \ldots)$ depends on the value of move $m_i$ in relation to the values of other moves.

- Too Simple:

  $$\Pr(m_i \mid s, c, \ldots) \sim g(s, c, \text{val}(m_i)).$$

  Doesn’t take values of the other moves into account.

- Cogent answer—let $m_1$ be the engine’s top-valued move:

  $$\frac{\Pr(m_i)}{\Pr(m_1)} \sim g(s, c, \text{val}(m_1) - \text{val}(m_i)).$$

  That and $\sum_i \Pr(m_i) = 1$ minimally give the Main Principle.
Main Principle and Schematic Equation

The probability $\Pr(m_i \mid s, c, \ldots)$ depends on the value of move $m_i$ in relation to the values of other moves.

- Too Simple:

$$\Pr(m_i \mid s, c, \ldots) \sim g(s, c, val(m_i)).$$

Doesn’t take values of the other moves into account.

- Cogent answer—let $m_1$ be the engine’s top-valued move:

$$\frac{\Pr(m_i)}{\Pr(m_1)} \sim g(s, c, val(m_1) - val(m_i)).$$

That and $\sum_i \Pr(m_i) = 1$ minimally give the Main Principle.

- Much Better answer (best?): Use $\frac{\log(1/\Pr(m_1))}{\log(1/\Pr(m_i))}$ on LHS.
Main Principle and Schematic Equation

The probability \( \Pr(m_i \mid s, c, \ldots) \) depends on the value of move \( m_i \) in relation to the values of other moves.

- Too Simple:

  \[
  \Pr(m_i \mid s, c, \ldots) \sim g(s, c, \text{val}(m_i)).
  \]

  Doesn’t take values of the other moves into account.

- Cogent answer—let \( m_1 \) be the engine’s top-valued move:

  \[
  \frac{\Pr(m_i)}{\Pr(m_1)} \sim g(s, c, \text{val}(m_1) - \text{val}(m_i)).
  \]

  That and \( \sum_i \Pr(m_i) = 1 \) minimally give the Main Principle.

- Much Better answer (best?): Use \( \log(1/\Pr(m_1)) \) on LHS.

- Needs Multi-PV analysis—already beyond Guid-Bratko work.

- Single-PV data on millions of moves shows other improvements.
The Data

- Over 1 million moves of 50-PV data: 62GB
The Data

- Over 1 million moves of 50-PV data: 62GB
- Over 20 million moves of Single-PV data: 22 GB
The Data

- Over 1 million moves of 50-PV data: 62GB
- Over 20 million moves of Single-PV data: 22 GB
- = 42 million pages of text data at 2k/page.
- All taken on two quad-core home-style PC’s. Is this “Big Data”?
The Data

- Over 1 million moves of 50-PV data: 62 GB
- Over 20 million moves of Single-PV data: 22 GB
- ≈ 42 million pages of text data at 2k/page.
- All taken on two quad-core home-style PC’s. Is this “Big Data”?

![Image of a person with a gun]
“Big-Data” Aspects
“Big-Data” Aspects

1. **Synthesis of two different kinds of data.**
   - Single-PV data acts as scientific control for Multi-PV data.
   - Covers almost entire history of chess.
   - Shows large-scale regularities.
“Big-Data” Aspects

1. **Synthesis of two different kinds of data.**
   - Single-PV data acts as scientific control for Multi-PV data.
   - Covers almost entire history of chess.
   - Shows large-scale regularities.

2. **Model design decisions based on large data.**
   - Logarithmic scaling law
   - “58%-42% Law” for probability of equal-value moves
   - Choice of fitting methods
“Big-Data” Aspects

1. Synthesis of two different kinds of data.
   - Single-PV data acts as scientific control for Multi-PV data.
   - Covers almost entire history of chess.
   - Shows large-scale regularities.

2. Model design decisions based on large data.
   - Logarithmic scaling law
   - “58%-42% Law” for probability of equal-value moves
   - Choice of fitting methods

3. Scientific discovery beyond original intent of model.
   - Human tendencies (different from machine tendencies...)
   - Follow simple laws...
Better, and Best?

Need a general function $f$ and a function $\delta(i)$ giving a *scaled-down* difference in value from $m_1$ to $m_i$.

$$\frac{f(\Pr_E(m_i))}{f(\Pr_E(m_1))} = g(E, \delta(i)).$$

**Implemented** with $f = \log$ and log-log scaling, as guided by the data.

**Best model?** Let *weights* $w_d$ at different *engine depths* $d$ reflect a player’s depth of calculation. Apply above equation to evals at each depth $d$ to define $\Pr_E(m_i, d)$. Then define:

$$\Pr_E(m_i) = \sum_d w_d \cdot \Pr_E(m_i, d).$$

This accounts for moves that *swing* in value and idea that weaker players prefer weaker moves. *In Process Now.*
Why Desire Probabilities?

- Allows to *predict* the $N$ of agreements with any sequence of moves $m^t_*$ over game turns $t$, not just computer’s first choices:

$$N = \sum_t \Pr_E(m^t_*).$$

- and it gives confidence intervals for $N$.

- Also predicts *aggregate error* ($\text{AE}$, scaled) by

$$e = \sum_t \sum_i \delta(i) \cdot \Pr_E(m^t_i).$$

Comparing $e$ with the *actual* error $e'$ by a player over the same turns leads to a “virtual Elo rating” $E'$ for those moves.

- **IPR** $\equiv$ “Intrinsic Performance Rating.”
The Turing Pandolfini?

- **Bruce Pandolfini** — played by Ben Kingsley in “Searching for Bobby Fischer.”
- 25th in line for throne of Monaco.
- Now does “**Solitaire Chess**” for Chess Life magazine:
  - Reader covers gamescore, tries to guess each move by one side.
  - E.g. score 6 pts. if you found 15.Re1, 4 pts. for 15.h3, 1 pt. for premature 15.Ng5.
  - Add points at end: say 150=GM, 140=IM, 120=Master, 80 = 1800 player, etc.
- Is it scientific?
- With my formulas, *yes*—using *your* games in *real* tournaments.
Training Sets: **Multi-PV** analyze games with both players rated:

- **2490–2510**, all three times
- **2390–2410**, (lower sets have over 20,000 moves)
- **2290–2310**, (all sets elim. moves 1–8, moves in repetitions,
- **2190–2210**, (and moves with one side > 3 pawns ahead)
- Down to **1590–1610** for years 2006–2009 only.
- **2600-level set done for all years since 1971.**
Training the Parameters

- Formula $g(E; \delta)$ is really

$$g(s, c; \delta) = \frac{1}{e^{x^c}} \text{ where } x = \frac{\delta}{s}.$$  

- $s$ for *Sensitivity*: smaller $s \equiv$ better ability to sense small differences in value.

- $c$ for *Consistency*: higher $c$ reduces probability of high-$\delta$ moves (i.e., blunders).

- Full model will have parameter $d$ for depth of calculation.
For each Elo $E$ training set, find $(s, c)$ giving best fit.

Can use many different fitting methods...

- Can compare methods...
- Whole separate topic...
- Max-Likelihood does poorly.

Often $s$ and $c$ trade off badly, but $E' \sim e(s, c)$ condenses into one Elo.

Strong linear fit—suggests Elo mainly influenced by error.
Some IPRs—Historical and Current

- Magnus Carlsen:
  - 2983 at London 2011 (Kramnik 2857, Aronian 2838, Nakamura only 2452).
  - 2855 at Biel 2012.

- Bobby Fischer:
  - 2921 over all 3 Candidates’ Matches in 1971.
  - 2650 vs. Spassky in 1972 (Spassky 2643).
  - 2724 vs. Spassky in 1992 (Spassky 2659).


- Paul Morphy: 2344 in 59 most impt. games, 2124 vs. Anderssen.

- Capablanca: 2936 at New York 1927.

- Alekhine: 2812 in 1927 WC match over Capa (2730).

Sebastien Feller Cheating Case

- Cyril Marzolo confessed 4/2012 to cheating most moves of 4 games. On those 71 moves:
  - Predicted match% to Rybka 3 depth 13: 60.1% ± 10.7%
  - Actual: 71.8%, z-score 2.18 (Barely significant: rumor says he used Firebird engine.)
  - AE test more significant: \( z = 3.37 \) sigmas.
  - IPR on those moves: 3240.
- On the other 5 games: actual < predicted, IPR = 2547.
- Paris Intl. Ch., July 2010: 3.15 sigmas over 197 moves, IPR 3030.
- Biel MTO, July 2010: no significant deviation, alleged cheating on last-round game only.
What is a Scientific Control?

- If I say odds are 2,000-to-1 against Feller’s performance being “by chance,” then I should be able to show 2,000 other players who did not match the computer as much.

- (show “Control” site on Internet)

- But note—if I have many more performances, say over 20,000, then I should expect to see higher match % by non-cheating players! “Littlewood’s Law”

- (show)

- To be sure, stats must combine with other evidence.

- (show “Parable of the Golfers” page)

- Aside from cheating, what does this tell us about humanity?
1. Perception Proportional to Benefit

How strongly do you perceive a difference of 10 kronor, if:

- You are buying lunch and a drink in a pub. (100 Kr)
- You are buying dinner in a restaurant. (400 Kr)
- You are buying an I-pod. (1000 Kr)
- You are buying a car. (100,000 Kr)

For the car, maybe you don’t care. In other cases, would you be equally thrifty?

*If you spend the way you play chess, you care maybe 4× as much in the pub!*

(show pages)
2. Is Savielly Tartakover Right?

The winner is the player who makes the next-to-last blunder.

- We like to think chess is about Deep Strategy.
- This helps, but is it statistically dominated by blunders?
- Recent Examples:
  - USA-Russia and USA-China matches at 2012 Olympiad.
  - Gelfand-Anand 2012 Rapid playoff.
- My Average Error (AE) stat shows a tight linear fit to Elo rating.
- Full investigation will need ANOVA (analysis of variance).
3. Procrastination...

- (Show graph of AE climbing to Move 40, then falling.)
  - King’s Indian: 12. Bf3!? then 13. Bg2 N (novelty)
  - “Grischuk was already in some time pressure.”
- IPR for Astana World Blitz (cat. 19, 2715) **2135**.
- IPR for Amber 2010+2011 (cat. 20+21): **2545**.
- Can players be coached to play like the young Anand?
4. Human Skill Increasing Over Time?

- In 1970s, two 2700+ players: Fischer and Karpov. In 1981: none!
- Sep. 2012 list, 44 2700+ players. Rating Inflation?
- My results:
- 2600 level, 1971–present:
  - Can argue 30-pt. IPR difference between 1980’s and now.
  - Difference measured at 16 pts. using 4-yr. moving averages, 10-year blocks.
  - Explainable by faster time controls, no adjournments?
- Single-PV AE stat in all Cat 11+ RR since 1971 hints at mild deflation.
- Moves 17–32 show similar results. Hence not just due to better opening prep?
- Increasing skill consistent with Olympics results.
5. Variance in Performance, and Motivation?

- Let’s say I am 2400 facing 2600 player.
- My expectation is 25%. Maybe:
  - 60% win for stronger player.
  - 30% draw.
  - 10% chance of win for me.
- In 12-game match, maybe under 1% chance of winning if we are random.
- But my model’s intrinsic error bars are often 200 points wide over 9–12 games.
- Suggests to take event not game as the unit.
- How can we be motivated for events? (show examples)
6. Are We Reliable?

- One blunder in 200 moves can “ruin” a tournament.
- But we were reliable 99.5% of the time.
- Exponential $g(s, c)$ curve fits better than inverse-poly ones.
- Contrary to my “Black Swan” expectation.
- But we are even more reliable if we can use a computer...
- (show PAL/CSS Freestyle stats if time).
7. Not Just About Chess?

- Only chess aspect of entire work is the evaluations coming from chess engines.
- No special chess-knowledge, no “style” (except as reflected in fitted $s, c, d$).
- General Problem: Converting Utilities Into Probabilities for colordarkredfallible agents.
- Framework applies to multiple-choice tests, now prevalent in online courses.
- Alternative to current psychometric measures?
- Issue: Idea of “best move” at chess is the same for all human players, but “best move” in sports may depend on natural talent.
Conclusions

- Lots more to do!
- Can use helpers!
  - Run data with other engines, such as Stockfish.
  - Run more tournaments.
  - Run to higher depths—how much does that matter?
- Spread word about general-scientific aspects; fight gullibility and paranoia over cheating.
- Deter cheating too.
- Learn more about human decision making.
- Thus the Turing Tour comes back to the human mind.
- Thank you very much for the invitation.