COMPUTERS & SECURITY 48 (2015) 58–73

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

On the limits of engine analysis for cheating detection in chess

David J. Barnes[*], *Julio Hernandez-Castro*

School of Computing, The University of Kent, Canterbury, Kent CT2 7NF, United Kingdom

ARTICLE INFO

Article history:
Received 12 July 2014
Received in revised form
14 September 2014
Accepted 10 October 2014
Available online 22 October 2014

Keywords:
Chess
Cheating
Online games
Privacy
Machine assistance
False positives

ABSTRACT

The integrity of online games has important economic consequences for both the gaming industry and players of all levels, from professionals to amateurs. Where there is a high likelihood of cheating, there is a loss of trust and players will be reluctant to participate — particularly if this is likely to cost them money.

Chess is a game that has been established online for around 25 years and is played over the Internet commercially. In that environment, where players are not physically present "over the board" (OTB), chess is one of the most easily exploitable games by those who wish to cheat, because of the widespread availability of very strong chess-playing programs. Allegations of cheating even in OTB games have increased significantly in recent years, and even led to recent changes in the laws of the game that potentially impinge upon players' privacy.

In this work, we examine some of the difficulties inherent in identifying the covert use of chess-playing programs purely from an analysis of the moves of a game. Our approach is to deeply examine a large collection of games where there is confidence that cheating has not taken place, and analyse those that could be easily misclassified.

We conclude that there is a serious risk of finding numerous "false positives" and that, in general, it is unsafe to use just the moves of a single game as *prima facie* evidence of cheating. We also demonstrate that it is impossible to compute definitive values of the figures currently employed to measure similarity to a chess-engine for a particular game, as values inevitably vary at different depths and, even under identical conditions, when multi-threading evaluation is used.

1. Introduction

Online game playing constitutes a large element of recreational Internet usage, with significant sums of money involved by game creators, game hosting sites and those who play the games. Inevitably, cheating is commonplace and often seeks to exploit system vulnerabilities (Yan and Randell, 2009). To combat this, game server's user agreements often include terms such as, "…the Software may include functionality designed to identify software or hardware processes or functionality that may give a player an unfair competitive advantage" (Valve Corporation, 2014). In turn, this can give rise to concerns among users about their privacy being violated by intrusive scanning techniques (Newell, 2014).

Chess is one of the many online games that has become highly vulnerable to cheating in the form of "exploiting

* *Corresponding author.*
E-mail addresses: d.j.barnes@kent.ac.uk (D.J. Barnes), jch27@kent.ac.uk (J. Hernandez-Castro).
http://dx.doi.org/10.1016/j.cose.2014.10.002
0167-4048/

machine intelligence" (Yan and Randell, 2009) since the widespread availability of chess engines on home computers that are easily stronger than the best human players. In the chess community, finding ways to determine whether a player is making their own decisions, or simply playing the choice of a strong program, has become a pressing issue. Interestingly, allegations of cheating are not confined to online play. A number of cheating complaints in over-the-board (OTB) play have recently received a lot of attention in the mainstream news media (chessvibes). This has led to quite a number of changes in tournament-playing conditions — such as the use of metal detectors and the complete ban of mobile phones among players and even spectators, but also occasional requests for full body searches where cheating is suspected (Chess.com, 1152). Indeed, the world governing body of chess, FIDE, has now approved procedures in the formal Rules of Chess that are akin to those of online game servers in terms of their personal intrusiveness: Allowing arbiters to request a full search of bags, clothes and other items in private (FIDE, 2014).

The taint of cheating in both OTB and online chess is bringing bad publicity and discouraging sponsorship and clearly needs addressing if it is not to play a major role in slowing down the wider spread of official online chess tournaments and titles, making monetising the millions of users of online chess servers much harder.

Increased suspicion that cheating might be taking place inevitably leads to a surge in allegations of cheating — whether well founded or not. Where no physical evidence of cheating is available, the primary source for an allegation is usually a demonstration of similarity between a human player's moves and those chosen by a powerful chess engine. Our aim in this paper is to sound a note of caution over the degree to which such a similarity should be taken as *prima facie* evidence of cheating, with the burden of proof then resting on the accused player to demonstrate a negative. We provide evidence of the multiple inherent difficulties and limitations of supporting allegations of chess cheating purely through the use of chess-engine similarity analysis of suspect games, particularly when the sample of such games is small.

Through an extensive analysis of games covering a wide historical period, we conclude that no isolated comparison between played moves and an engine's evaluations can be taken as authoritative evidence of cheating. Among other data, we illustrate our conclusions by highlighting several "false positive" games which, had they not been played well before the current availability of strong chess engines, might have been subject to completely wrong allegations of cheating. We also show how "evidence" to support a cheating case can easily be massaged and cherry-picked by a number of techniques that we describe and analyse in depth.

1.1. *Related work on cheating*

1.1.1. *Chess cheating*

Our work has similarities to that by chess-cheating analysis pioneer Kenneth Regan (Regan's chess page) but also has some significant differences, and is complementary to it. Prof. Regan has published a number of papers in the area (Di Fatta et al., 2009; Haworth et al., 2010; Regan et al., 2012; Regan and Haworth, 2011) and has proposed a set of techniques based on *predictive analytics*. The strength of chess players is measured by the ELO system, originally defined by Dr. Arpad Elo (Elo, 1978). Players gain or lose points depending upon their results, and the number of points won or lost depends on the comparative strength of their opponents. Regan uses a player's ratings before and after a tournament, as well as their performance level within the tournament, to determine whether their move selection is statistically consistent with the historical move selection of similarly rated players, when compared against a strong chess engine's move selection.

Since modern chess engines are rated hundreds of ELO points above the best human player, a tournament performance that is significantly higher than what would normally be expected may be the result of obtaining machine assistance. Regan's is a very interesting approach and, so far, the only available one. The approach has a strong statistical foundation but care is needed in its application, particularly when considering players whose performance is improving rapidly, which is not uncommon among young players, for instance.

It is important to note that we employ a slightly different methodology and way of measurement, and a quite different treatment of the opening moves which becomes apparent in Appendix A. Further differences with the work of Regan are presented where most appropriate through the rest of the paper.

Apart from the seminal works of Regan and his colleagues, the academic or scientific literature on chess cheating is scarce. We should, however, note Friedel's very interesting historical discussion and examples (Friedel, 2001), and some other works that, although not focused on chess cheating, have produced interesting and useful results like those by Guid and Bratko (Guid and Bratko, 2006, 2011; Guid et al., 2008).

An additional obstacle for researchers and progress in this area is that the numerous online chess servers that have developed their in-house techniques for detecting cheating have, in all cases, kept their methodology secret and seem unprepared to disclose any information publicly. This security by obscurity approach, as we have seen in so many other security fields, is destined to fail in the long term.

1.1.2. *Cheating in online games*

In other domains, numerous researchers have worked over the years in detecting cheating in games, particularly in online ones. The most insightful works are those by Jeff Yan and his team (Yan, 2003; Yan and Randell, 2005, 2009; Yan and Choi, 2002; YeungJohn et al., 2006). Most of these focus on massive multiplayer online games (MMOGs) over distributed systems covering, for example, techniques like aimbots, wall-hacking, speedhacks and ghosting — following the naming taxonomy proposed in (Yan and Randell, 2005). It is also worth noting that Yan and Randell added "violation of fairness" to the traditional consequences of security violations (Yan and Randell, 2009).

We believe the strong differences between chess (particularly over the board play) and these MMOGs make this line of related cheating research interesting but of limited relevance to our domain. For example, (Yan, 2003) investigates the security failures of an online Bridge server, dividing them into

two categories: those due to a single player cheats, and those due to collusion between multiple players. The work covers techniques like card eavesdropping, client hacking, the exploitation of bad randomness or opportunistic escaping, illicit information passing, deadlocks, etc.

While not directly involving the exploitation of machine intelligence, it is also worth mentioning the security analysis conducted on the Internet Chess Club (Black et al., 2006), which is one of the most popular online chess servers, where the authors point out numerous security issues with its client/server communication protocol, and show how these could be exploited, not only to expose credit card information but also to gain unlimited powers over an ICC user, including easy wins over unsuspecting rivals.

To conclude this brief introduction to the existing related works, we note an interesting statistical behavioural analysis presented in (Laurens et al., 2007), based on the central hypothesis that players engaged in cheating exhibit behaviour which is significantly distinguishable from normal play. The paper establishes links between cheating detection and intrusion detection, seems promising and general enough to be applicable to other games, though it is initially limited to Half-Life 2.

In summary, cheating in the gaming community is widespread, particularly where financial rewards for players are available. The detection and prevention of cheating are important features for the viability of gaming communities and their hosts but grounds for allegations must always be based on a robust and credible analysis.

Our aim with this article is not primarily to offer an alternative to Regan's approach but to show the limitations of the sort of naive analysis that is commonly followed by non-experts. This should move people to act with extreme caution in the easy levelling of cheating allegations. Alternatively, you can view our contribution in this article as a set of difficult questions that need to be answered by any cheating-detection technique: how to detect, deal with and remove — or at least take fully into account — the large number of "false positives" that occur naturally, so that cheating accusations are sound and well-founded in the future.

2. Materials and methods

As sources for the games to be analysed, and for the classification of opening moves, we used the ChessBase cbmega database (ChessBase) covering the 19^{th} and 20^{th} centuries, and Mark Crowther's TWIC archive (Crowther) containing games from 1997 to 2013. We used Stockfish v3.0 (Romstad et al.,) for our analysis because it is one of the strongest open source and freely available chess engines, but the same methodology could be applied using any UCI-compatible chess engine (Kahlen, 2004).

The move text of the PGN (Edwards, 1994) format of a game was converted from standard-algebraic notation (SAN) to long-algebraic notation via pgn-extract (Barnes), as this is the format used in UCI. We wrote a piece of software, which we refer to hereafter as "the analyser" which translated the moves of a game to a series of UCI *position* commands. Each position command was then passed by the analyser to the chess engine to be evaluated to the required depth. The analyser received back the engine's evaluations and wrote them to a file in XML format (Fig. 1). We developed a separate program, based around an XML parser, to read the XML files and compute the metrics we describe in Section 2.1. The source code of both the analyser and metrics tool are available online.[1]

The chess engine was run in MultiPV (i.e., multi-best-line) mode and set to return evaluations for the five best moves in each position (but the analyser allows the number of moves to be varied as required). Both opening books and endings table bases were turned off. When the engine did not consider the actual move played in the position to be one of the five best, we detected this and forced its evaluation. As shown in Fig. 1, the engine's output consists of either an integer evaluation of the value for each move (expressed as a positive or negative number of centipawns) or "mate in N".

In contrast to our selective approach, Regan uses an exhaustive search of each position but eliminates branches having an evaluation greater than 300 centipawns. Our handling of opening moves, which is new and different from any previously used in the related literature, is covered in detail on Section 2.2.

2.1. Metrics

The most obvious test to use when comparing moves by a human player against those picked by a chess engine is the percentage of the human moves matching those preferred by the engine. Regan calls this the "move matching percentage" (MM). The conventional conclusion derived from a 100% match would be that the human played the moves suggested by the engine and, therefore, was self-evidently cheating. What is less obvious is what an 80% or a 90% match means, and that is where Regan uses rating-related statistical analysis to help the inquiry. Players have ratings (similar to rankings) based on past performance, which can be used to spot outlying improvements in performance that could be the result of machine assistance.

We determine a similar metric to MM, although ours is calculated slightly differently so we refer to it as the "coincidence value" (CV) to avoid any confusion.

> Coincidence Value (CV) is a figure between 0 and 1 representing the proportion of non-book moves chosen by a player with the same evaluation as the engine's preferred move.

CV differs from MM in that the latter is the percentage of played moves that are identical to the engine's preferred move, whereas CV is the percentage of those *having the same evaluation*.

Our CV values are then, by definition, always greater-than or equal-to the corresponding MM values. Our justification for the difference is that the identification of whether a particular move is better than another when the evaluations are the same is an arbitrary decision that may well vary from run to run. Section 3.2 discusses in more detail the issues

[1] http://www.cs.kent.ac.uk/~djb/chessplag/.

```
<game>
<tags>
<tag name = "Event" value = "New York" />
<tag name = "Site" value = "New York" />
<tag name = "Date" value = "1857.??.??" />
<tag name = "Round" value = "?" />
<tag name = "White" value = "Kennicott, Hiram" />
<tag name = "Black" value = "Morphy, Paul" />
<tag name = "Result" value = "0-1" />
<tag name = "BookDepth" value = "29" />
</tags>
<moves>
e2e4 e7e5 g1f3 b8c6 d2d4 e5d4 f1c4 f8c5 f3g5 g8h6 g5f7
h6f7 c4f7 e8f7 d1h5 g7g6 h5c5 d7d6 c5b5 h8e8 e1g1 e8e4
b5d5 e4e6 c1g5 d8e8 f2f4 f7g7 f4f5 g6f5 d5f5 e6g6 g5f6
g7g8 f5f4 c8h3 f6g5 e8e3 f4e3 d4e3 g2h3 g6g5 g1h1 e3e2
f1e1 c6d4 b1a3 a8e8 0-1
</moves>
<analysis bookDepth = "29" searchDepth = "18" variations = "5">
<move>
<played>g6f5</played>
<evaluation move = "g6f5" value = "137" />
<evaluation move = "e6f6" value = "-218" />
<evaluation move = "h7h6" value = "-327" />
<evaluation move = "a7a6" value = "-387" />
<evaluation move = "c8d7" value = "-412" />
</move>
```

Fig. 1 – Sample of partial XML output from an engine's analysis. The moves are in long-algebraic format and associated with each move `played` are the `evaluation` values of the best alternatives.

with the repeatability of evaluations over multiple runs. In addition, choosing a second or third ranked (but equally good) move might potentially be a good strategy employed by a cheater for easily bypassing naive rank-based detection schemes without resorting to play weaker chess moves. An extension of this detection-avoidance technique would be for a player occasionally to choose sub-optimal moves that are still a good second or third choice of the engine, leading to MM/CV values below 100%. MM or CV are clearly too crude a measure to reliably detect cheating on their own.

Regan, therefore, also determines the "average error" (AE) as a second metric.

> Average error (AE) is the mean difference in evaluation between the best move and the played move for non-book moves, expressed in centipawns.

The difference will always be ≤ 0 but we will use the phrase "low error" to mean a value close to zero. The idea is that a very low AE might be indicative of cheating even in the face of an MM/CV that does not appear to give cause for concern.

We also use this metric. However, we note that evaluation differences involving mating moves are not easily expressed as a number of centipawns in the form they are returned by a UCI engine. For instance when:

- The engine finds a shorter forced mate than the one played.
- The engine finds a forced mate but the played move does not lead to forced mate.
- The engine finds a non-mating move but a weaker move leading to forced mate was played.

While it could be argued that it should be possible to ascribe a centipawn-equivalent to each of these cases, we prefer to omit such instances from the otherwise entirely numerical analysis. The only exception is the occurrence of alternative forced mates of the same length via different moves, which we score as a difference of zero.

2.2. *Treatment of opening moves*

The opening moves of a chess game inevitably repeat the moves of games that have been played in the past, and well-trodden move sequences are known as an "opening book". Such repetitions can easily last twenty moves or more and high-level players usually stick to familiar, well understood opening lines. Opening lines are classified under 3-letter "ECO codes."

We do not include book moves in the engine's analysis on the grounds that they tell us little of use in determining coincidence. Given the wide period of time we were covering, it is important to differentiate the opening knowledge a player at the beginning of the 20th century would have from one at the end of the century. Therefore, in order to differentiate opening moves derived from book knowledge from

those requiring analytical thought, we built a historical database of board positions from all the games available to us (c. 7M) and not just those that we analysed for this study (c. 120K).

The resulting database contains around 87 million different positions. Each entry consists of a hashed Forsyth–Edwards Notation (FEN) position[2] and the date at which that position was first encountered in a game in our compilation. The database is structured as a set of approximately 500 tables, one per 3-character ECO classification. A game is entered into the database by determining its ECO code and then generating the FEN positions for the first 20 moves.

Each FEN position in the first 20 moves of a game is then checked against the database table for the game's overall ECO classification. If the position has not been seen before, then this game's date is entered in the table as the first known occurrence of that position. If the position has been met before and this game pre-dates the date of the existing entry, then the new date replaces the old one, otherwise the table is not changed for that position.

For each game analysed in this study, its overall ECO code was determined and its FEN positions generated in a similar way. The positions were then looked up in the corresponding table of the database in order to determine how much of the game was known theory at the time the game was played. An important point to note here is that only positions that were encountered in games *prior to the date at which the game was played* are used in determining the book-depth of the game. The book depth was then added to the game as a pseudo-PGN tag — *BookDepth* — which was passed on to the analyser (see Fig. 1). One limitation of the approach is that games that are given different ECO codes, but transpose to the same position will not be compared.[3]

Our approach differs from Regan's general method of considering the first 8 moves of each game to be book. However, when analysing particular games in detail, Regan does appear to identify the move at which a novelty is first played and begin his analysis at that point. See, for example, his detailed analysis of the games played by Ivanov at the Zadar Open in 2012 (Regan Letter) which were subject to allegations of cheating. In Appendix A we compare our own identification of the starting point for non-book moves in those games with his.

While we do not claim that our database is exhaustive nor necessarily fully historically accurate, we do consider this to be a much better way to exclude book moves from analysis than simply assuming an arbitrary cut-off point for all games. This novel approach should also give a more accurate evaluation of the coincidence levels of similar games that were played at quite different historical periods.

[2] FEN concisely encodes game state, including whether castling is still permitted for each side and how many moves have occurred since the last pawn capture since this can be used to declare a draw.

[3] An example of this is noted in the discussion of game 2 in Appendix A.

2.3. *Historical analysis*

As the starting point for our study we randomly selected from the cbmega database (ChessBase) around 70,000 games for the period up to 1950 and 50,000 for the period from 1950 to 2005. We analysed these games to the modest search depth of 8.[4]

Since we were primarily interested in identifying games with low AE and high CV, we selected all games with $AE \leq 13$ for analysis at depth 10, along with a 10% random selection from the remaining games in order to reduce the risk of missing interesting ones through the relatively low search depth. This gave us 25,000 games at depth 10. We then repeated this process for depths 12–22 by progressively narrowing the upper limit of AE, although we dropped the random sampling after depth 10. This selection process roughly halved the number of games analysed at each successive depth until we had around 250 at depth 22.[5]

3. Results

Figs. 2 and 3 show the ranges of AE and CV values produced from our sample at depth 8; both are plotted against the number of non-book moves played. We excluded all games where fewer than 10 non-book moves were played. Each game generates two points — one for each player — which contributes to the symmetry in Fig. 3.

Even at this low analysis depth there are several clear trends that we see repeated at larger depths:

- Games where one player has a high AE tend to be shorter than those with a low AE. This is unsurprising as earlier blunders or weak moves tend to lead either to resignation or rapid superiority for the opponent.
- Games with a high CV ($\geq$0.8) tend to be relatively short, although there are still several shown here extending as far as 70 moves after the opening.
- Several games have a CV of 1.0, although all those shown here (except one) are of fewer than 20 moves. If played today, such games would be prime candidates for false allegations of cheating.

Another apparent effect, clearly shown in Fig. 3 is what we can call "human-probabilistic fatigue": as the length of the game increases, its CV shows a remarkable decrease which is consistent with human players becoming increasingly tired and committing more mistakes. This happens in combination with the common fact that more challenges need to be solved, so the likelihood of error-free games probabilistically decays after each move, but more markedly so in the case of human

[4] The computational facilities available to us when we started this work did not allow us to go any deeper with such a large number of games.

[5] We should note here that the assumption implicit in this selection process — that low CV/high AE games at low search depth will not give high CV/low AE values at high search depths (and vice versa) — is not entirely well founded. It was purely used as an arbitrary mechanism to reduce the sample size at increasing depths.

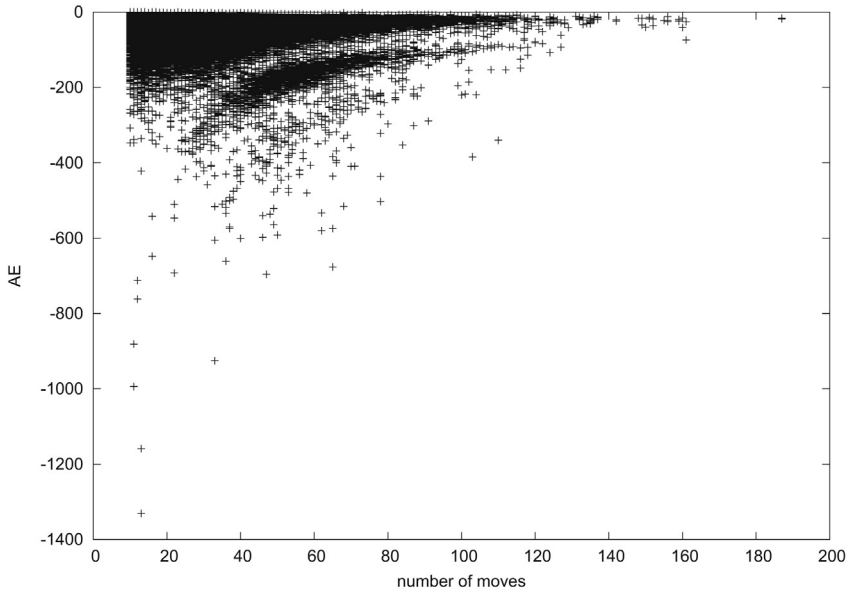

Fig. 2 – AE at depth 8.

players. This would clearly not be the case where chess-engines are used, so long games involving cheating would likely appear as outliers in this diagram. On the other hand, short high-CV games are relatively common, most likely as a result of home opening or early-middlegame preparation. We plan to use these findings in the future as the basis for an easy rule to quickly characterise suspect games.

For comparison with the games analysed to depth 8, Figs. 4 and 5 show the range of AE and CV values at search depth 20 plotted against the number of non-book moves. The number of games is, of course, considerably smaller as a result of our progressive filtering process. At this depth, games where both players had a large AE have not persisted. Note that there are no games beyond 40 moves where either player has an AE worse than −30 and it can be seen that there are still a considerable number of games with a very small AE over a large number of moves.

Fig. 6 shows the region of Fig. 4 for which AE values range from −15 to 0. The fact that both players tend to have a similarly low AE and high CV in longer games can clearly be seen towards the right-hand side of each plot. In particular, note the pair of points around (91,−3) on the AE plot, whose CV values are 0.85 and 0.82. Yet this 100-move game between Carames and Fedorovsky was played in 1965 (Appendix B.1), so it clearly did not involve any engine assistance. The link between game length and high accuracy from both players

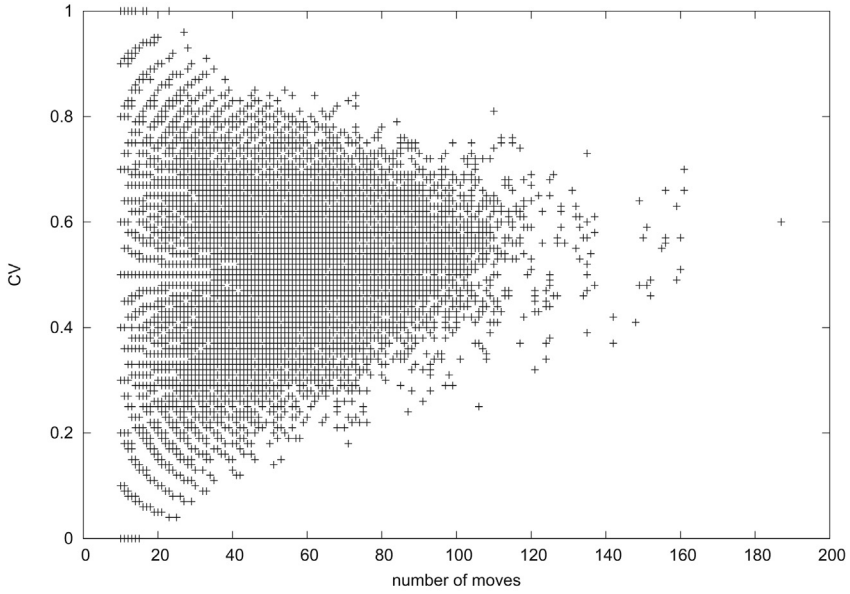

Fig. 3 – CV at depth 8.

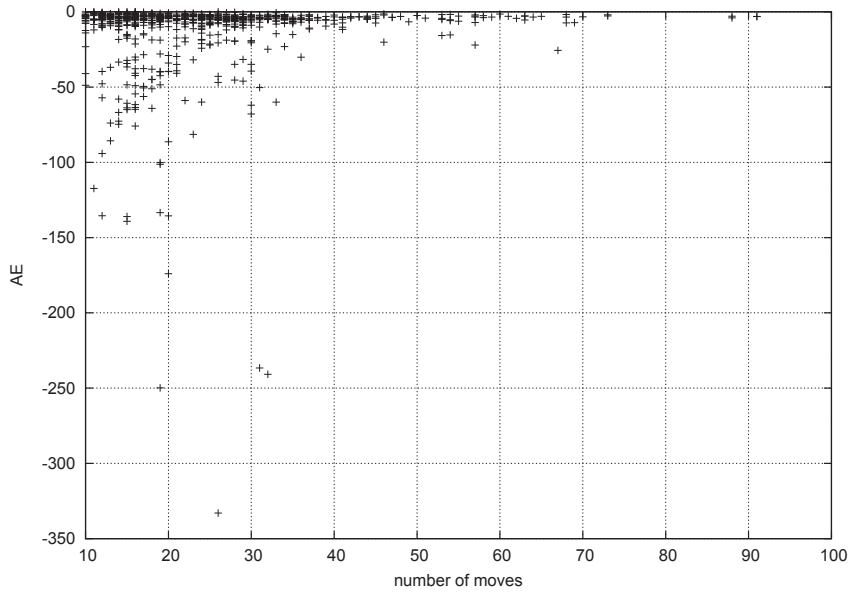

Fig. 4 – AE at depth 20.

clearly reduces the significance of coincidence and accuracy in support of cheating allegation in long games, unless levelled equally at both players.

However, this example does illustrate that high-CV and low-AE play is perfectly possible for human players over a large number of moves. This naturally leads us to consider the issue of "false positives".

3.1. *False positives*

The Carames–Fedorovsky game illustrates that great care must be taken on assuming that high CV/MM and low AE values are clear indicators of cheating. In this section we want to present a number of further games that, given the period at which they were played, can safely be discarded as not involving any computer-based cheating. We call this collection of games the "false positives", because they would likely trigger alarms in any automated cheating detection mechanism based simply on move accuracy or correlation when measured or compared with computer moves. Given the sampling technique we have used, our false-positive collection is in no way exhaustive, and it could easily be enlarged from the 92 games it now contains with access to more powerful computational resources than we had available. We believe it is interesting to present and discuss some of these games in more detail.

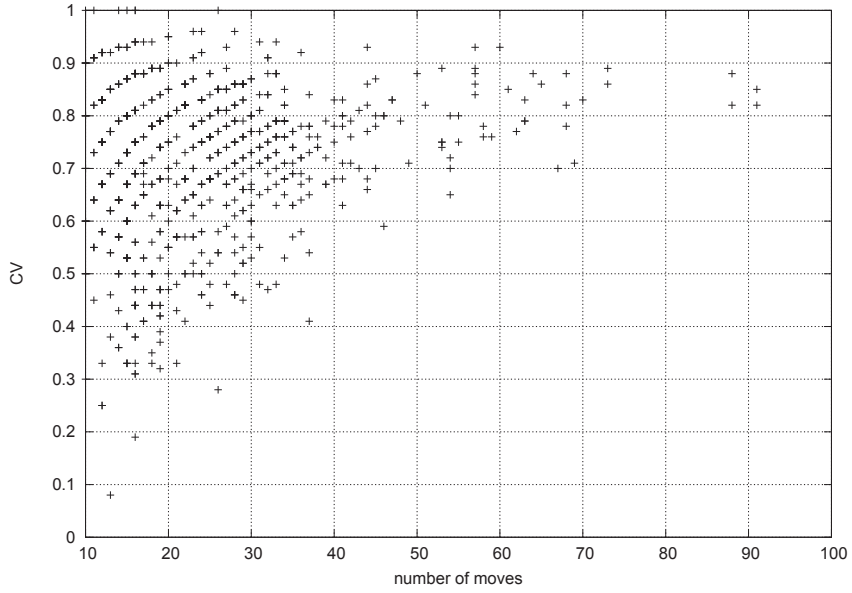

Fig. 5 – CV at depth 20.

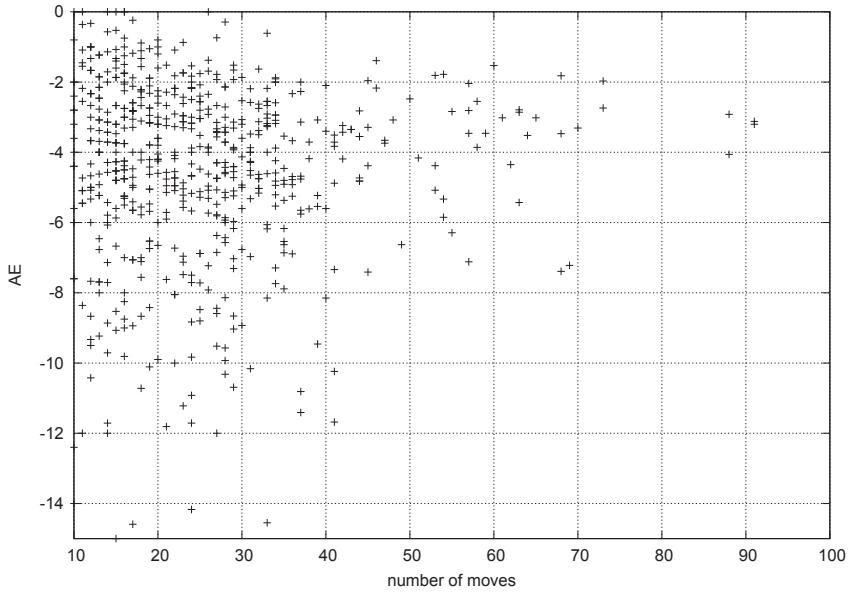

Fig. 6 – AE at depth 20 in the range −15 to 0.

Chronologically, we can start with the effort by the legendary Paul Morphy with black against Hiram Kennicott in 1857 (Appendix B.2). Our analysis of historical book depth found that the first 29 half-moves were not new, and were probably well known to both players. After that point, and in a very tactical position, Morphy played ten "perfect" new moves that ensured he won the game smoothly. At depth 18, Stockfish matches 100% of these moves.

Is this strange? Well, if Morphy had been playing today he would very likely be accused of cheating, because in the same year he posted another "perfect" performance, with black and blindfolded no less. The game was played against John William Schulten in New York and included 13 new consecutive perfect moves (analysis at depth 16). If this were not enough, not much later, in 1859 we see the same player defeat Augustus Mongredien in their seventh match game in Paris, which he won convincingly by 7.5–0.5. Another exceptional performance by Morphy happened in 1866, when he defeated C. Muarian in New Orleans. This game showed 12 consecutive new perfect moves from Morphy and is another 100% match with Stockfish's suggestions at depth 18.

One can argue that all these games have in common that they were quite short, and that finding 10 perfect moves in succession should be rare but not so worrying in the case of one of the all time greats like Morphy. We would disagree, at least partially, with that view.

The game between Weiss and Burille given in Appendix B.3 is a good case in point. It was played in the 6th USA Congress in New York between Miksa (Max) Weiss (the co-winner of the Congress, with Chigorin) and the position after 13 half-moves had been encountered before in our historical database. From that point on, Mr. Weiss won the game by playing 26 consecutive moves that exactly match our engine's choices at depth 20. This is a much larger series of perfect moves than that we saw in the Morphy examples. Surely due to computer cheating, except for the fact that the year was 1889.

Among our admittedly limited search for similar "false positive" games our list of possible "cheaters" include names like Anderssen, Zukertort, Alekhine, Euwe, Spielmann, Capablanca, Marshall, Steinitz, Rubinstein, Gruenfeld, Flohr, Keres, Botvinnik, Reshevsky, Pachman, Unzicker, Petrosian, etc. But also fairly unknown players like Liebert, Book, Prandstetter, Garcia Vera, Raud, Tyroler, Prochazka, Ekstrom, Turover, Gilg, Santa Cruz, Dely, and many others. All of these played moves that gave a CV of 1.0 at one or more depths.

A potential explanation that might be offered for this perfect accuracy phenomenon is that perhaps all the related games were highly tactical, with a very limited number of available options, which could have helped in finding the right move every time. This is a good attempt to explain and characterise these "false positive" games, but it is not without its faults as the game between Browne and Timman (Appendix B.4) demonstrates. This is a relatively long game of 40 moves, played with perfect accuracy by Walter Browne for 23 moves, after the divergence from our book record at move 17. While the game is tactical, Browne also has to take positional decisions of roughly equivalent value at (at least) moves 24, 26, 31, 32, and 37.

Should we accuse Walter Browne of cheating? Of course not. First of all, this is only a single game, and no cheating accusation should be based on a single game because, as we have already seen, such perfect sequences happen at all levels of the game, and in roughly all types of game and position.

Only when there is some sort of supporting external evidence and this happens in a number of games (and still very cautiously, as we saw Morphy and others had multiple perfect games through the years) can we start to give any cheating accusation some credibility.

The idea of false positives should be easy to grasp for most chess players, and particularly for professional players. Unfortunately this seems not always to be the case, as the unfortunate incident involving Igor Kurnosov in the 2009

Aeroflot Tournament seems to illustrate. In this tournament, one of the largest and most prestigious of the calendar, top seed GM Shakhriyar Mamedyarov accused his Russian opponent Igor Kurnosov of computer-assisted cheating and dropped out of the event after resigning their game in round six.

In the game, given in Appendix B.5, a known position is reached after 16 White moves (not after 12. d5 as implicitly claimed by Mamedyarov, but only after 16…Qd6, the novelty with respect to Rodshtein–Margvelashvili, Budva, 2003). This means that Kurnosov has played only 6 "perfect" moves (as seen by Stockfish at depths 16 and 20, but not depths 12, 24 and 26, for example).

Such a coincidence is unremarkable, particularly between strong professional players. This game would not have even made our list of "false positives" due to Kurnosov having played fewer than 10 new perfect moves at any depth level, and is order of magnitudes less *suspicious* that any of the previously shown cases. Fortunately, Chief arbiter Geurt Gijssen decided it was not a convincing claim but nevertheless searched Kurnosov's jacket ("a pack of cigarettes, a lighter and a pen").

Our conclusion is that this infamous cheating complaint is totally without merit and should have been discarded on the spot. Regan has also studied this game and, through completely different methods, reached the same conclusion that there is not evidence at all supporting any cheating allegations (Regan).[6]

3.2. *Reliability and repeatability*

In Section 3.1 we noted that in the Kursonov case we saw agreement with the disputed moves at depths 16 and 20, but not at other depths. This lack of agreement from a single engine at different but similar depths raises interesting questions of both reliability and repeatability when calculating metrics such as MM/CV and AE. Since in the absence of objective physical evidence the similarity of a player's moves to those of an engine is usually the starting point for most allegations, we sought to investigate how definitive any particular set of AE and CV values might be, and how repeatable they are under both similar and different conditions.

UCI engines have the capability to run in either single or multi-threaded mode, configured via a *Threads* option. On a multi-core machine, using multiple threads for game analysis is clearly an attractive option since it allows either a particular depth to be reached more quickly or greater depth to be covered within a limited time.

We selected 11 games from our sample that had shown a particularly high CV value at depth 14 for at least one of the players, and re-analysed them multiple times at depths between 8 and 22 using either 1, 8 and 16 threads.

3.2.1. *Repeatability in single-threaded mode*

When an engine is run in single-threaded mode, our observations were that its evaluations for moves were identical over multiple runs, and even across different machines, assuming all other configuration was identical. Variation was only evident at different search depths. Figs. 7 and 8 show the results from analysing the 11 games using only a single thread at depths between 8 and 22. Each column represents evaluations for a single player's moves in a particular game at the different depths. Note that two AE points for player 1 (−298 at depth 20 and at −285 depth 22) and 1 for player 20 (−169 at depth 20) are not shown in order not to distort the scale for the majority of the points.

Aside from those of players 1 and 20, most of the move sets exhibit little substantial variation in AE across all move depths (e.g., a range of 3.3 for player 15 and 3.5 for player 10), but 10 or more points is fairly common. A spread is more evident in the CV where several values that might be considered suspiciously close to 1.0 at some depths are much more comfortably around 0.8 in others.

3.2.2. *Repeatability in multi-threaded mode*

For multi-threaded mode we used either 8 or 16 threads on machines with multiple cores. The machines were only being used for the task of analysis. In contrast to single-threaded mode, repeat runs at a single depth rarely resulted in identical AE or CV values. We observed exactly the same features whether we used 8 or 16 threads, and there appeared to be no significance in the particular number of threads used. Figs. 9 and 10 show the variation we observed in the same sample of 11 games at depth 22 over four runs using 16 threads. As with Fig. 7, we have avoided including the scores of players 1 and 20 in Fig. 9 in order to avoid distorting the scale. Both figures include the equivalent values from the runs in single-threaded mode for comparison purposes.

Multiple identical runs at a single depth result in variations in both AE and CV values. A range of 0.1 in CV values is not uncommon, for instance, and this feature is present at all depths. While most of the AE variations are relatively small, note that player 7's AE values range from −4.16 to −59.56. In addition, the values omitted from Fig. 9 for players 1 and 20 exhibit even greater AE variations: −101 to −310 for player 1 and −103 to −617 for player 20. The values in single-threaded mode at the same depth are −285 and −98, respectively. What is the reason for such wide variations under identical conditions?

At each stage of a position's analysis, the move tree will be searched in the order determined by the engine's developer (e.g., depth-first or breadth-first). Unless some degree of randomness is deliberately introduced into the time-constrained, non-exhaustive search process the search will be deterministic in a single-threaded implementation, and always give the same result. This is why randomness is commonly used in solving search-based optimisation problems, for instance, in order to introduce variation in an attempt to escape local minima over multiple runs.

In contrast, a multi-threaded implementation in a multi-core environment naturally introduces a degree of non-determinism into the search process through the way the threads are scheduled and managed. The nature of multi-threaded scheduling means that, while each thread will receive a broadly similar share of the available computing resources, equality and synchronicity are not guaranteed. As a consequence, a thread undertaking the search of a particular line at any one point may receive either a greater or lesser time on different runs of the engine, but rarely exactly the

[6] Unfortunately GM Kurnosov passed away recently in tragic circumstances http://en.chessbase.com/post/ruian-gm-igor-kurnosov-dies-in-car-accident-120813.

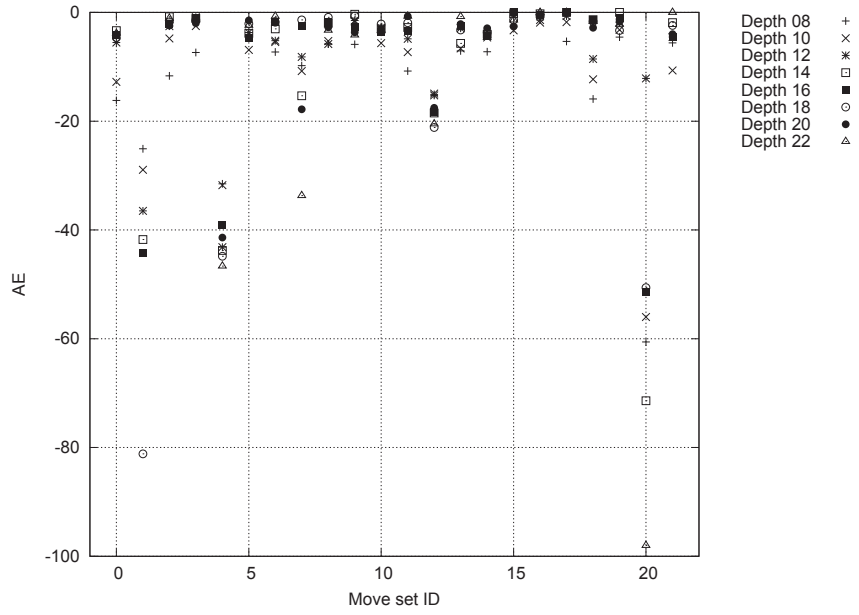

Fig. 7 – AE values in the range [−100,0] for 11 games (22 player performances) at depths 8–22 using a single thread.

same amount. It should also be appreciated that the search depth set on an engine is generally only a lower bound. Engines will often search a little deeper when they find a promising line that looks likely to give a more accurate evaluation.

One way to explain the extreme variation we have observed for the moves of players 1 and 20 at the greater depths, therefore, is that occasionally a single evaluation thread gets a slightly greater share of computation time, hits lucky and finds a forcing line of moves that it pursues beyond the pre-set search depth. For instance, in the case of player 20: already in an inferior position, his penultimate move was scored as −1296, −1632, −8904 and −9034 on the four runs (−1571 in single-threaded mode). The larger two values suggest a much higher degree of confidence that the move played will definitely lead to a loss. Such large values will clearly have a very big impact on the average error for the game. In other circumstances, a single such value within a long game could turn a suspiciously small AE value into an innocuous looking larger one.

Regan uses single-threaded mode for the sake of repeatability of results. While it is reasonable to have some form of repeatability to serve as the basis for discussion and further investigation, our concern is that its use could create the

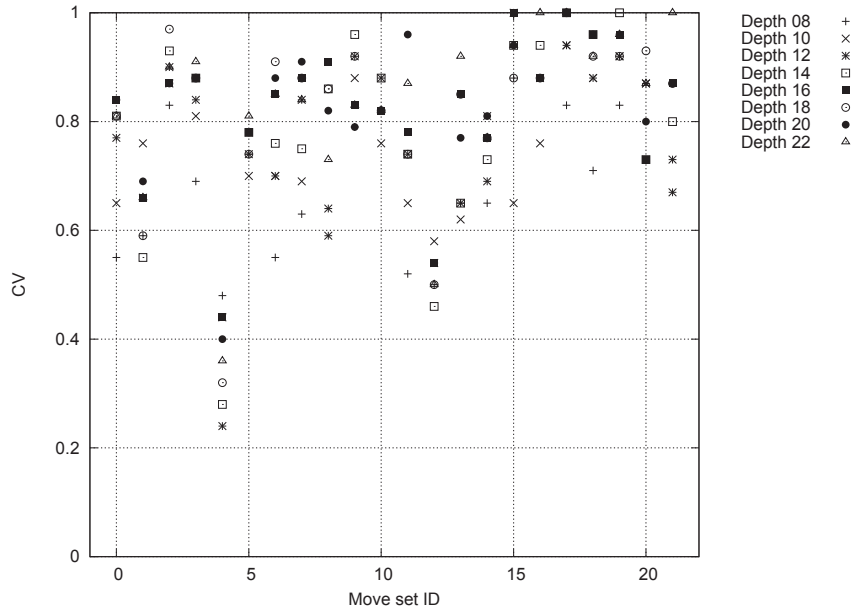

Fig. 8 – CV values for 11 games (22 player performances) at depths 8–22 using a single thread.

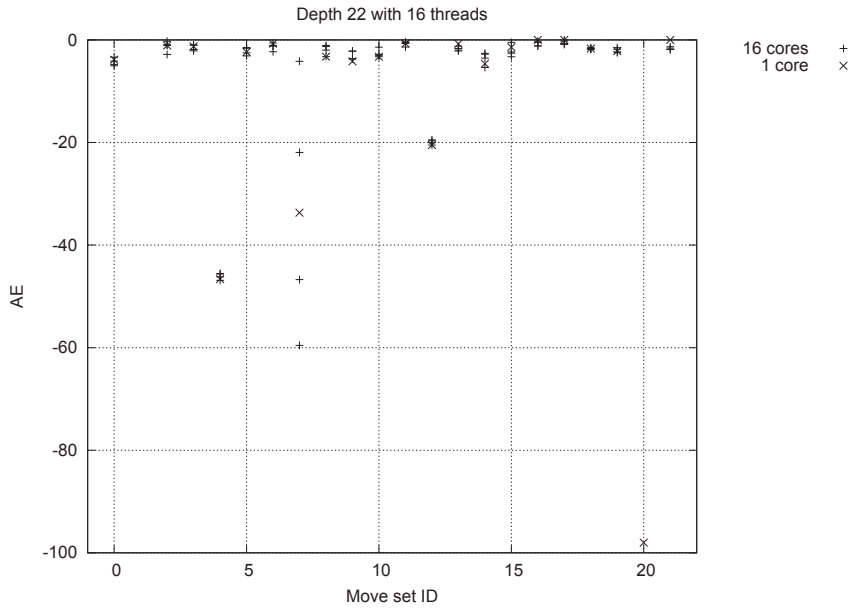

Fig. 9 – AE values for 11 games (excluding players 1 and 20) over four runs at depth 22 using 16 threads. (Single-core values at depth 22 also shown for comparison)

misleading impression that a particular set of evaluations are, in some way, definitive for a game. The point of the discussion in this section is that a particular pair of AE and CV values cannot be considered definitive without taking into account how representative they are within a range of values that could be produced for the same game under either slightly different conditions (such as a different analysis depth) or even the same conditions (in a multi-threaded environment). It would be naive to assume that anyone using an engine for the purposes of cheating would limit themselves to single threaded mode, when the advantages of faster and deeper analysis with multi-threaded mode are obvious. Furthermore, different chess-engines — of which there are many — introduce an additional variability in the scoring of moves, further strengthening the point that no definitive AE and CV values can be authoritatively ascribed to a game.

These variations are, in many ways, highly inconvenient for those seeking to detect cheating using comparisons such as the ones we have outlined. On the other hand there is also a risk in that these effects could be abused in order to "massage" evidence to support a cheating accusation: the accuser could run the game over different engines, at different depths and in

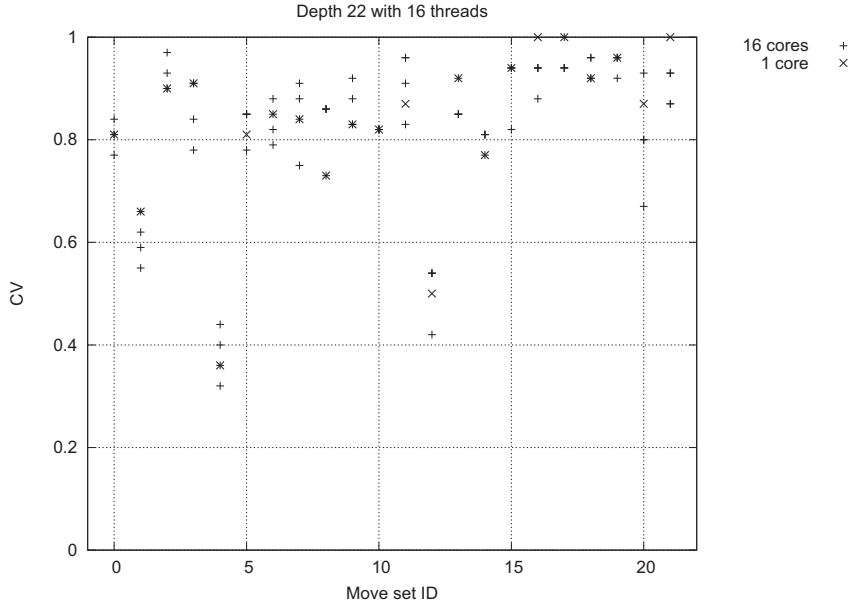

Fig. 10 – CV values for 11 games (22 player performances) over four runs at depth 22 using 16 threads. (Single-core values at depth 22 also shown for comparison)

multi-threaded mode and present as evidence the more "incriminating" values found from the large set of different results obtained. Most chess players and arbiters are unfortunately completely unaware of this potentially misleading feature of multi-threaded evaluation, thus increasing the likelihood that such malpractice would be successful.

4. Conclusions

Cheating in online games violates the trust of players and the fairness of the game and the environment. Where there is prize money involved, the results are, in effect, an act of theft. In seeking to combat cheating, online gaming organisations regularly require participants to accept conditions of use that affect an individual's privacy, and could be exploited either intentionally or unintentionally by those organisations. Chess is one online game that is subject to these same pressures, and we have noted that even participants in over-the-board chess may have to subject themselves to body searches if subject to an allegation of cheating.

We have explored the most obvious means of attempting to demonstrate that a player has cheated at chess: comparing the moves they played to those suggested by a strong chess program. Using a large source of games beyond question, we have identified a considerable number of games where the human player selected identical moves to those of a modern engine, but this demonstrates nothing more than accurate play. Clearly, great care must be taken in assuming that a 100% (or close to it) match is *prima facie* evidence of cheating.

Furthermore, we have also demonstrated the even more basic and worrying fact that it is impossible to compute definitive values of MM, CV and AE for a particular game. Their values for a single game inevitably vary at different analysis depths and even under identical conditions when multi-threading evaluation is used. To this already high variability we can add that of using different but similarly strong chess engines to complete a picture that should raise some doubts about the soundness of using these measures too blindly.

The remarkable fact is that even if a valid accusation were to be based on the correct identification of the engine, the search depth actually used, and all other engine configurations, it could well be the case that, in a multi-threaded setting, a 100% correspondence would not be demonstrated.

However, our study has shown that general trends do exist in the relationship between game length and player accuracy: The longer a game is played, the less likely it is that high MM/CV and low AE values are insignificant. These trends could be the basis of future anti-cheating tools but, as we have conclusively shown, they should still be used with extreme care to avoid false positives and biases that lead to wrong accusations.

Rather than being entirely negative, another way to see these conclusions is as a sanity check against frivolous and hasty accusations and to provide thresholds for further investigation, such as recommending a body search. More work is clearly needed with a larger sample of games to establish exactly where the length/coincidence borderline should be drawn for a single game. Once this has been done, it should be relatively easy to integrate both the analysis and sanity-checking processes into a software tool to support tournament arbiters in dealing with allegations, and to allow both players involved to independently verify that the decisions taken have been fair and unbiased.

We plan to extend this research to a much larger collection of games and later on to set up a cloud service providing this analysis service (from PGN input) to automatically help tournament organisers, arbiters and players. We ultimately want to help to stop both chess cheating and the cheating paranoia by offering FIDE, arbiters, players and the general public a transparent set of rules and open source, freely available tools to take founded, sound decisions. We do not believe in security by obscurity, and hence our emphasis in publishing the methods publicly and offering the source code of our tools.

Acknowledgements

We acknowledge the support of the University of Kent's Faculty of Sciences Research Fund, grant 5590-2013, for this work.

Appendix A. Opening Novelties by Borislav Ivanov in the 2012 International Zadar Open

In his open letter to the ACP Board Members (Reganc), Kenneth Regan identified the opening novelties in the games played by Ivanov in each round of the 2012 Zadar Open and discarded all moves before them from his analysis. The use of our opening database described in Section 2.2 identified slightly different points for the novelties, most of which are the result of transposition. The primary reason for the differences is that he uses an opening book based on players with an ELO rating of 2300+, whereas we use all games available to us from our game sources.

For these particular games, once we had identified the putative book depth from the database, we then conducted an exhaustive search of all games based on the final book position in order to find those matching. As illustrated under game 2, this has the advantage of picking up identical positions reached via transposition in games that have been given different ECO codes.

The differences between Regan's and our identification of the opening novelties are as follows:

- **Game 1:** Regan gives 15. Bh4 as the novelty. We found 13 games played between 1994 and 2012 that had also reached this position. These include Detter v Roeder, Cappelle la Grande 1994 and Pokorny v Rykalin, Czech Open, Pardubice 2006. 15 … Nfd7 was the novelty we identified as previous continuations had been: Nh7, g5 and Nh5.
- **Game 2:** Regan gives 9… Be7. We found 3 games in which the position after 10 … O–O had been reached earlier, one of which (Benedetto v Vidmar, Villa Ballester 2004) reached an identical position after 12. Nxd5 but then continued exd5. So the novelty we identified is 12… cxd5.

  Our database search did not initially pick up the length of this coincidence as the games were given different ECO codes (A13 and D45).

- **Game 3:** Regan gives 10… b6 and we found no games to contradict this.
- **Game 4:** Regan gives 15. h4. We found 1 game in which the position after 18. Ke2 had been reached by transposition (Dziuba v Socko, Najdorf Memorial Open, Warsaw 2009), so 18 … Rad8 is the novelty.
- **Game 5:** Regan gives 10. Bc4. We found 1 game in which the position after 13. O–O had been reached (Puschmann v Szabo, Hungarian Championship 1997), so 13 … f5 is the novelty.
- **Game 6:** Regan gives 9. Qc1. We found 1 game in which the position after 10. O–O was reached (Geller v De Castro, Skopje Olympiad 1972), so the novelty is 10 … Bb7.
- **Game 7:** Regan gives 10. g3. We found 2 games in which the position after 11 … Nd7 was reached, after which 12. Qd2 was played, so 12. O–O is the novelty.
- **Game 8:** Regan gives 8. Nf3. We found 3 games in which the position after 10. g3 was reached, so 10 … Be7 was the novelty, Nc6 and Qb5 having previously been played.
- **Game 9:** Regan gives 11. h4. We found 2 games in which the position after 13. Qf2 was reached so 13 … O–O is the novelty, Nc4 and Qa5 having previously been played.

While the differences we describe are relatively small, we note that players using widely-available commercial and free game databases for their opening preparation or engine opening books may well not limit themselves to only games played at the highest level.

We feel, therefore, that it is important to eliminate from comparison those moves that could easily be known to a player whose play is called into question.

Appendix B. Sample Games

Games discussed in detail in the body of the paper.

Appendix B.1. Carames–Fedorovsky, Buenos Aires, 1965

A game that is beyond question but with unusually high CV values for both players over such a long game. At depth 20 with Stockfish 3.0, the CV values were 0.82 and 0.85, respectively (AE −3.12 and −3.20), and at depth 22 they are even higher at 0.86 and 0.89 (AE −3.09 and −2.52).

```
[Event "Buenos Aires Comunicaciones"]
[Site "Buenos Aires"]
[Date "1965.??.??"]
[Round "1"]
[White "Carames, Luis"]
[Black "Fedorovsky, Rafael"]
[Result "1/2-1/2"]
[BookDepth "18"]

1. e4 c6 2. d4 d5 3. Nc3 dxe4 4. Nxe4 Bf5 5. Ng3 Bg6 6. Nf3
Nd7 7. c3 e6 8. Qb3 Qb6 9. Qxb6 axb6 10. Bf4 Ngf6 11. Nh4
Nd5 12. Nxg6 hxg6 13. Bd2 b5 14. Bd3 N7f6 15. Ne4 Nxe4 16.
Bxe4 Bd6 17. g3 Kd7 18. f4 Ra7 19. a3 Nb6 20. Ke2 Nc4 21.
Bc1 Be7 22. Rb1 Nd6 23. Bd3 Rh3 24. Kf3 Ra8 25. Bf1 Rh5 26.
h4 Rah8 27. Kg2 Ra8 28. Be2 Rhh8 29. Be3 Nc4 30. Bf2 Bd6 31.
a4 Rxa4 32. b3 Ra2 33. Bxc4 bxc4 34. bxc4 Rb8 35. c5 Bc7 36.
h5 gxh5 37. Rxh5 Ke7 38. Rb3 g6 39. Rh1 b5 40. cxb6 Rxb6 41.
Rhb1 Rxb3 42. Rxb3 Rc2 43. Kf3 Kd6 44. Ra3 f6 45. Ke3 g5 46.
Be1 gxf4+ 47. gxf4 Rh2 48. Bg3 Rh3 49. Kf2 f5 50. c4 Bd8 51.
Kg2 Rh7 52. Be1 Rb7 53. Ra4 Bf6 54. Bb4+ Kd7 55. Bc5 Bd8 56.
Ra7 Rxa7 57. Bxa7 Bc7 58. Kf3 Kc8 59. Ke3 Kb7 60. Bc5 Ka6
61. Bb4 Kb6 62. Bd2 c5 63. Kd3 Kc6 64. Be3 cxd4 65. Kxd4
Bb6+ 66. Kd3 Bc7 67. Kd4 Bb6+ 68. Kd3 Bc7 69. Bd2 Kc5 70.
Be3+ Kc6 71. Bc1 Kc5 72. Ba3+ Kc6 73. Ke3 Bb6+ 74. Kf3 Bd4

75. Ke2 Bb6 76. Be7 Bc7 77. Ke3 Bb6+ 78. Kd3 Bc7 79. Ke3
Bb6+ 80. Kf3 Bc5 81. Bf6 Bg1 82. Ke2 Bb6 83. Be5 Bg1 84. Kd3
Kc5 85. Bc3 Bf2 86. Bd2 Bg1 87. Be1 Kc6 88. Bd2 Kc5 89. Ba5
Kc6 90. Bd8 Bc5 91. Bf6 Bd6 92. Bg5 Bc5 93. Bh6 Bb6 94. Bg7
Bc7 95. Bh6 Bb6 96. Kc3 Be3 97. Kd3 Bb6 98. Bg5 Bc5 99. Bf6
Bd6 100. Bg5 Bc5 1/2-1/2
```

Appendix B.2. Kennicott-Morphy, New York, 1857

At depth 18, Stockfish matches 100% of Morphy's final 10 moves.

```
[Event "New York"]
[Site "New York"]
[Date "1857.??.??"]
[Round "?"]
[White "Kennicott, Hiram"]
[Black "Morphy, Paul"]
[Result "0-1"]
[BookDepth "29"]

1. e4 e5 2. Nf3 Nc6 3. d4 exd4 4. Bc4 Bc5 5. Ng5 Nh6 6. Nxf7 Nxf7
7. Bxf7+ Kxf7 8. Qh5+ g6 9. Qxc5 d6 10. Qb5 Re8 11. O-O Rxe4 12.
Qd5+ Re6 13. Bg5 Qe8 14. f4 Kg7 15. f5 gxf5 16. Qxf5 Rg6 17. Bf6+
Kg8 18. Qf4 Bh3 19. Bg5 Qe3+ 20. Qxe3 dxe3 21. gxh3 Rxg5+ 22. Kh1
```

Appendix B.3. Weiss-Burille, New York, 1889

At depth 20, 26 moves by Weiss matched those chosen by Stockfish 3.0.

```
[Event "?"]
[Site "New York"]
[Date "1889.??.??"]
[Round "1"]
[White "Weiss, Miksa"]
[Black "Burille, Constant Ferdinand"]
[Result "1-0"]
[BookDepth "13"]

1. e4 e5 2. Nf3 Nc6 3. Bb5 Nf6 4. O-O Nxe4 5. d4 Be7 6. Qe2 Nd6
7. Bxc6 dxc6 8. dxe5 Nf5 9. Rd1 Bd7 10. e6 fxe6 11. Ne5 Bd6 12.
Qh5+ g6 13. Nxg6 Ng7 14. Qh6 Nf5 15. Qh3 Rg8 16. Qxh7 Rg7 17. Qh5
Qf6 18. Nh4+ Ke7 19. Nxf5+ Qxf5 20. Qxf5 exf5 21. Nc3 Rh8 22. g3
Rgh7 23. Bg5+ Kf8 24. h4 f4 25. Ne4 Bg4 26. Rd2 Bf5 27. Nxd6 cxd6
28. Rxd6 fxg3 29. fxg3 Rd7 30. Rf6+ Rf7 31. Rf1 Rxf6 32. Bxf6 Rh5
33. Bg5 1-0
```

Appendix B.4. Browne–Timman, Wijk aan Zee, 1980

Walter Browne plays 23 moves with a perfect match to those of Stockfish 3.0.

```
[Event "Hoogovens"]
[Site "Wijk aan Zee"]
[Date "1980.01.??"]
[Round "2"]
[White "Browne, Walter S"]
[Black "Timman, Jan H"]
[Result "1-0"]
[BookDepth "33"]

1. d4 Nf6 2. Nf3 g6 3. c4 Bg7 4. g3 O-O 5. Bg2 d6 6. Nc3 Nc6 7.
O-O a6 8. d5 Na5 9. Nd2 c5 10. Qc2 Rb8 11. b3 b5 12. Bb2 Bh6 13.
f4 bxc4 14. bxc4 e5 15. dxe6 Bxe6 16. Nd5 Bxd5 17. cxd5 Ng4 18.
Nb3 Nxb3 19. axb3 Qb6 20. Qc3 c4+ 21. Kh1 f6 22. Bh3 Nf2+ 23.
Rxf2 Qxf2 24. Be6+ Kh8 25. Qxc4 Qb6 26. Bd4 Qxb3 27. Qxb3 Rxb3
28. Rxa6 g5 29. fxg5 Bxg5 30. Rxd6 h5 31. Rc6 h4 32. d6 Rb1+ 33.
Kg2 Rd1 34. e3 hxg3 35. hxg3 Rxd4 36. exd4 f5 37. Bxf5 Rxf5 38.
Rc5 Rxc5 39. dxc5 Kg7 40. c6 1-0
```

Appendix B.5. Mamedyarov–Kurnosov, Moscow, 2009

Subject to an allegation by Mamedyarov of cheating, only 6 moves were played by Kurnosov beyond those seen in previous games according to our analysis of the opening. Coincidence to engine-selected moves is insignificant under these circumstances.

```
[Event "Aeroflot Open"]
[Site "Moscow RUS"]
[Date "2009.02.22"]
[Round "6"]
[White "Shakhriyar Mamedyarov"]
[Black "Igor Kurnosov"]
[Result "0-1"]
[BookDepth "31"]

1. d4 Nf6 2. c4 g6 3. f3 d5 4. cxd5 Nxd5 5. e4 Nb6 6. Nc3 Bg7 7.
Be3 O-O 8. Qd2 Nc6 9. O-O-O f5 10. h4 fxe4 11. h5 gxh5 12. d5 Ne5
13. Bh6 Nec4 14. Qg5 Rf7 15. Bxc4 Nxc4 16. Rd4 Qd6 17. Bxg7 Rxg7
18. Qxh5 Qf4+ 19. Kb1 Bf5 20. fxe4 Bg4 21. Nge2 Qd2 0-1
```

REFERENCES

Barnes David J. pgn-extract: A portable game notation manipulator. http://www.cs.kent.ac.uk/~djb/pgn-extract/.

Black John, Cochran Martin, Gardner Ryan. A security analysis of the internet chess club. IEEE Secur Priv January 2006;4(1):46–52.

ChessBase. Mega database. http://www.chessbase.com/.

Chess.com. Official statement on the Ivanov story. http://www.chess.com/news/official-statement-on-the-ivanov-story-1152.

chessvibes. French cheating: disciplinary committee says guilty. http://www.chessvibes.com/reports/french-cheating-disciplinary-committee-says-guilty.

Crowther Mark. The week in chess. http://theweekinchess.com/.

Di Fatta Giuseppe, McC Haworth Guy, Regan Kenneth W. Skill rating by bayesian inference. In: Computational intelligence

and data mining, 2009. CIDM'09. IEEE Symposium on. IEEE; 2009. p. 89–94.
Edwards Steven J. Portable game notation (PGN). 1994. http://www6.chessclub.com/help/PGN-spec.
Elo Arpad. The rating of chessplayers, past and present, vol 3. Batsford; 1978.
FIDE. Fide laws of chess taking effect from 1st July 2014. 2014. http://rules.fide.com/images/stories/downloads/draft_laws_of_chess_1.7.2014.pdf.
Friedel F. Cheating in chess. Advances in computer games, 9. Maastricht, The Netherlands: Institute for Knowledge and Agent Technology (IKAT); 2001. p. 327–46.
Guid Matej, Bratko Ivan. Computer analysis of world chess champions. ICGA J 2006;29(2):65–73.
Guid Matej, Bratko Ivan. Using heuristic-search based engines for estimating human skill at chess. ICGA J 2011;34(2):71–81.
Guid Matej, Pérez Aritz, Bratko Ivan. How trustworthy is crafty's analysis of world chess champions. ICGA J 2008;31(3):131–44.
Haworth Guy, Regan Ken, Fatta Giuseppe Di. Performance and prediction: bayesian modelling of fallible choice in chess. In: Advances in computer games. Springer; 2010. p. 99–110.
Kahlen Stefan-Meyer. Universal chess interface (UCI). 2004. http://wbec-ridderkerk.nl/html/UCIProtocol.html.
Laurens Peter, Paige Richard F, Brooke Phillip J, Chivers Howard. A novel approach to the detection of cheating in multiplayer online games. In: Engineering complex computer systems, 2007. 12th IEEE international conference on. IEEE; 2007. p. 97–106.
Newell Gabe. Valve, vac, and trust. 2014. http://www.reddit.com/r/gaming/comments/1y70ej/valve_vac_and_trust/.
Regan Kenneth Wingate, Haworth Guy McCrossan. Intrinsic chess ratings. AAAI; 2011.
Regan Kenneth W, Macieja Bartlomiej, Haworth Guy McC. Understanding distributions of chess performances. In: Advances in computer games. Springer; 2012. p. 230–43.
Regan Kenneth W. Analysis of Mamedyarov vs Kurnosov. http://www.cse.buffalo.edu/~regan/chess/fidelity/M-Kaccuse/M-Kresults.txt.
Regan Kenneth W. Kenneth W. Regan's chess page. http://www.cse.buffalo.edu/~regan/chess/.
Regan Kenneth W. Letter and report to the association of chess professionals. http://www.cse.buffalo.edu/~regan/chess/fidelity/ACPcover-and-report.pdf.
Romstad Tord, Costalba Marco, Kiiski Joona. Stockfish chess engine. http://stockfish.org/.
Valve Corporation. Steam subscriber agreement. 2014. http://store.steampowered.com/subscriber_agreement/.
Yan Jeff. Security design in online games. In: Computer security applications conference, 2003. Proceedings. 19th Annual. IEEE; 2003. p. 286–95.
Yan Jianxin Jeff, Choi Hyun-Jin. Security issues in online games. Electron Libr 2002;20(2):125–33.
Yan Jeff, Randell Brian. A systematic classification of cheating in online games. In: Proceedings of 4th ACM SIGCOMM workshop on network and system support for games. ACM; 2005. p. 1–9.
Yan Jeff Jianxin, Randell Brian. An investigation of cheating in online games. IEEE Secur Priv 2009;7(3):37–44.
Yeung S, Lui John CS, Liu Jiangchuan, Yan Jeff. Detecting cheaters for multiplayer games: theory, design and implementation. In: Proc IEEE CCNC, vol 6; 2006. p. 1178–82.

David J. Barnes is a lecturer in Computer Science at the University of Kent. His research interests include the teaching of introductory programming, software testing and biological modelling. He is author of a major text book on introductory Java programming using the BlueJ IDE. He is a chess player to club-level standard and author of the widely used chess processing software, pgn-extract.

Julio Hernandez-Castro is a lecturer in Computer Security at the University of Kent. His research interests range from Cryptology to Steganography & Steganalysis, including Computer & Network Security, Computer Forensics, CAPTCHAs, RFID Security, the application of Non-Standard techniques to Cryptology and Quantum Information Processing. He is an active club-level chess player.