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/© 2014 Elsevier Ltd. All rights reserved.

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