Attribute Inference Attacks in Online Multiplayer Video Games: a Case Study on Dota2

Pier Paolo Tricomi
tricomi.pierpaolo@math.unipd.it
Department of Mathematics
† University of Padua, Italy

Lisa Facciolo
lisa.facciolo@studenti.unipd.it
Department of Mathematics
† University of Padua, Italy

Giovanni Apruzzese
giovanni.apruzzese@uni.li
Hilti Chair of Data and Application Security
University of Liechtenstein

Mauro Conti†
conti@unipd.it
Faculty of EEMCS
Delft University of Technology, NL

ABSTRACT

Did you know that over 70 million of Dota2 players have their in-game data freely accessible? What if such data is used in malicious ways? This paper is the first to investigate such a problem.

Motivated by the widespread popularity of video games, we propose the first threat model for Attribute Inference Attacks (AIA) in the Dota2 context. We explain *how* (and *why*) attackers can exploit the abundant public data in the Dota2 ecosystem to infer private information about its players. Due to lack of concrete evidence on the efficacy of our AIA, we empirically prove and assess their impact in reality. By conducting an extensive survey on ∼500 Dota2 players spanning over 26k matches, we verify whether a correlation exists between a player's Dota2 activity and their real-life. Then, after finding such a link ($p < 0.01$ and $\rho > 0.3$), we ethically perform diverse AIA. We leverage the capabilities of machine learning to infer real-life attributes of the respondents of our survey by using their publicly available in-game data. Our results show that, by applying domain expertise, some AIA can reach up to 98% precision and over 90% accuracy. This paper hence raises the alarm on a subtle, but concrete threat that can potentially affect the entire competitive gaming landscape. We alerted the developers of Dota2.

CCS CONCEPTS

• **Security and privacy**; • **Applied computing** → **Media arts**; • **Computing methodologies** → **Machine learning**;

KEYWORDS

Attribute Inference Attack, Video Games, Dota2, Machine Learning

1 INTRODUCTION

More than 3 billion people played Video Games (VG) in 2021, whose industry is constantly expanding, attracting new players every day [66]. A recent study highlighted that over 71% of participants increased their playtime, and that VG improved their well-being [9]. Within the broad VG landscape, one category stands out: *online multiplayer VG*. These VG allow players to interact with each other in a 'controlled' environment (i.e., the game) that is separated from their private life [62]. Specifically, players can interact in two distinct settings: cooperative or competitive. This paper focuses on the latter, motivated by the rise of the Electronic Sports (E-Sports) panorama, which generated over $1B of revenues in 2021 [66].

In E-Sports, players compete in VG matches [31]. Notable examples of E-Sports VG are Fortnite, ApexLegends, CS:GO, and Dota2. All such VG are addictive (on average, Dota2 players have over 1600 hours of playtime), and have an heterogeneous playerbase [18]. Some individuals "play for fun", e.g., to spend their free-time with friends, or to entertain their audience on streaming platforms [37]. Others, however, "play to win", and their primary aim is improving so that they can participate in (and, perhaps, win) one of the many competitions held regularly. Such competitions have rich prize-pools (up to $40M [2]) which attract thousands of contestants. Indeed, winning matches is difficult due to the highly competitive environment (which is ultimately a zero-sum game [30]), and 'mastering' an E-Sport VG requires constant dedication [29].

Several resources, typically referred to as Tracking Websites (TW), were born to track players' activities on a specific VG. Indeed, an intuitive way to improve is learning from past mistakes, and TW greatly facilitate such process by providing their users (i.e., the players) with detailed statistics of their in-game performance. We provide a screenshot of a TW focused on Dota2 in Fig. 1, showing an overview of the in-game activities of the player "Dendi". Such statistics include, e.g., data of past matches, the days in which the player is more active, their friends; additional information is available by navigating the webpage. Despite the undeniable advantages provided by TW (over 70M of Dota2 players use TW [60]), we observe that *all data retrieved and elaborated by TW is publicly available*: anyone can observe, collect, and use such data.

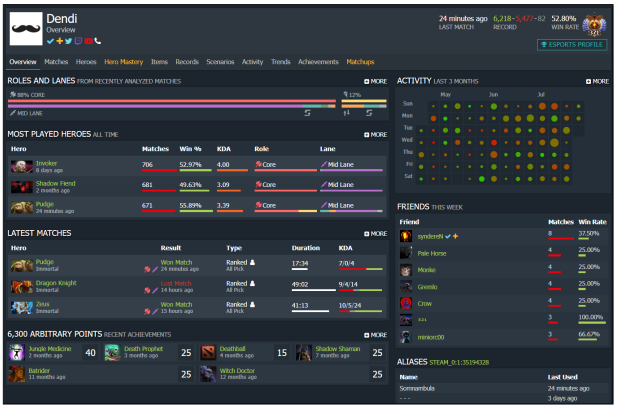


Fig. 1: A TW, showing the statistics of the professional Dota2 player "Dendi" [1]. All such information is constantly updated and publicly accessible: https://dotabuff.com/players/70388657.

We thus ask ourselves: "what if players' in-game data are used *against* them to violate their privacy?" If this were true, then the E-sport setting would be prone to Attribute Inference Attacks (AIA). Such attacks, enabled by the capabilities of Machine Learning (ML), aim to infer private information about a given target (i.e., a player) by using their publicly available data [27]. Although TW report only in-game statistics, we cannot exclude the existence of a link between such data and personal attributes (e.g., gender, age, personality) or even sensitive ones (e.g., health [16, 49])—the latter being outside our scope. Prior research (e.g., [41, 63]) revealed that a correlation exists between the in- and off-game traits of a given player. Surprisingly, however, no study has been carried out within the specific context of Dota2. Such a lack is concerning: the in-game data provided by Dota2 is semantically different from that of other VG. Hence, to this day, it is still *uncertain whether AIA are a threat* to Dota2 players. Consequently, it is also unknown (i) *how* AIA can be carried out and (ii) what is the *impact* of an AIA in the Dota2 context. Inspired by Biggio and Roli [10], we proactively assess the likelihood and the effects of this subtle privacy issue.

Contribution. This paper investigates the threat of AIA against Dota2 players. We begin (§2) by contextualizing the E-Sports ecosystem (with a focus on Dota2) and summarizing the fundamental concepts of AIA (building on related work). Then, we provide four major contributions—which go *beyond the research domain*.

- **A threat model of AIA against Dota2 players** (§3). We describe *how* to (legitimately) launch an AIA to infer private information on players while knowing only their Dota2 handle. We also explain *why* attackers would do so.
- **We prove the existence of correlations between Dota2 players' in-game data and their personal attributes** (§4). By conducting an (informed) survey, we collect in-game and personal data of 484 Dota2 players, corresponding to over 26k matches. We then perform a correlation analysis, showing the existence of statistically significant ($p < 0.01$) and strong (Spearman's $\rho > 0.3$) relationships between in-game (public) and off-game (private) attributes.
- **We proactively evaluate the impact of AIA in Dota2** (§5). We use the data gathered from our survey to (ethically) enact an AIA, and measure its success rate. We develop multiple ML models, by assuming attackers with varying domain expertise on Dota2. We show that even simple AIA can be successful (almost 70% F1-score on gender), and that more sophisticated AIA can further increase such impact (over 75% accuracy on predicting the occupation).
- **We assess AIA that can be staged in practice** (§6). We assume the viewpoint of an attacker with *specific* goals, and elucidate the real-threat of AIA in Dota2 by demonstrating a realistic application of our findings, showing AIA with near-perfect success rate (almost 100% precision).

Finally, we discuss our results, describe some countermeasures, and explain how our AIA can be extended to other E-Sports VG (§7). We then conclude our paper and provide ethical considerations (§8).

Transparency. We release a repository containing exhaustive details on our study, as well as the source code we developed for our analyses—available at: https://github.com/hihey54/Dota2AIA. Finally, we remark that Pier Paolo Tricomi is a top-1% Dota2 player.

Disclaimer. Our paper tackles a delicate privacy issue that potentially touches millions of video-gamers. All our evaluations are conducted ethically [57], but attackers are not bound to such ethics. At the time of writing, the problem is still open.

2 BACKGROUND AND RELATED WORK

Our paper tackles two emerging domains: competitive video games, and attribute inference attacks—which we now summarize.

2.1 The Competitive Video-Game Ecosystem

Competitive *video-games* (VG), and E-Sports in particular, are receiving a lot of attention [66], leading to a constant increase of players all aiming to "reach the top" [36]. To improve their performance, players can analyze their in-game statistics [12]. Such statistics are typically provided by the VG itself, but are limited to a single *match*. Even if most VG allow players to inspect their history, analyses can only be performed on a match-by-match basis. Such limitation was overcome by *Tracking Websites* (TW), which collect and aggregate information pertaining to all matches of a given player(s), providing a comprehensive overview of their activity (cf. Fig. 1). An illustration of such ecosystem is in Fig. 2, which we now describe from the viewpoint of our VG of choice—Dota2.

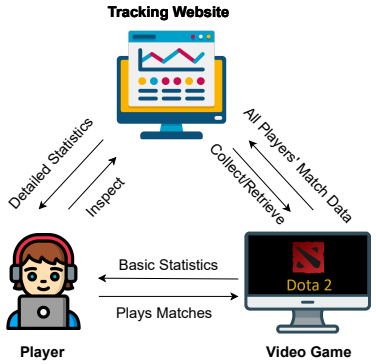


Fig. 2: The E-Sport ecosystem. *Players* engage in *matches* of a *video-game*, which publicly releases data on such matches. These data are collected by *tracking websites*, whose elaborations are made public.

- **Video-Game.** Dota2 is a Multiplayer Online Battle Arena (MOBA) VG. Released in 2013 and available for free, it is one of the most popular VG, counting up to 6M daily players and over 15M monthly players [52]. In a match, two teams of five players fight in real time with a common objective: destroy the enemy team's base before they do it to yours.
- **Players.** Each player in a team has a crucial role in ensuring their team's victory, and such roles are difficult to master. Indeed, Dota2 is extremely competitive: in 2021, its biggest tournament had the largest prize pool in the entire history of VG, amounting to $40M [2]. Such prizes are enticing for players, who continuously strive to get better: every Dota2 player has more than 1600 hours [32] of playtime (on average). It is not surprising, hence, that Dota2 players will resort to any (legitimate) tool to maximize their efficiency.
- **Tracking Websites.** A massive amount of Dota2 players leverage the services provided by TW [17]. Reportedly, some

TW tracked the activities of more than 79M players, aggregating the results of ~3B matches [60]. In our context, TW constantly interact with specific Dota2 APIs to retrieve all historical data pertaining to a player's matches. Before using a TW, a player must explicitly allow Dota2 to share their match details with external sources; however, considering the benefits provided by TW, only few players do not give their consent. Every player (and corresponding Dota2 activity) tracked by a TW is publicly visible on the platform.

Such context begs the question: "**why are TW publicly releasing players' data?**" The answer is: "because players themselves want such data to be public." Indeed, such availability allows players to:

- browse other players' statistics, so as to learn how the game is played by top-players;
- increase their visibility to professional organizations, which can hire them if they show good performance;
- share their activity with friends, teammates, or even unknown players that paired up with them;
- climb TW-specific rankings (e.g., players who get most wins with a given character).

Simply put, players benefit from their in-game data being publicly released by TW—thereby exposing players to the threat of AIA.

2.2 Attribute Inference Attacks

We summarize the fundamentals of Attribute Inference Attacks (AIA), and then highlight the research gap motivating our paper.

2.2.1 AIA in a nutshell. The underlying goal of AIA is inferring *private* information on a given target by exploiting *publicly available* data on such target. For example, an attacker can use the (public) ratings posted on a video streaming platform by a given user to infer their (private) gender [64]. Such inference can be done leveraging the predictive capabilities of Machine Learning (ML): By training a ML model on a representative dataset, and then providing such ML model with some user's public data, the ML model will output the personal attributes of such user. We remark that AIA are semantically different than membership inference attacks (e.g., [34, 72]), whose goal is inferring information on the ML model's training set.

AIA are becoming problematic due to the lack of education of most internet users, who publicly share their data while overlooking (or ignoring) the corresponding risks (e.g., [15, 33, 46]). For instance, most data published on social networks can be easily retrieved via OSINT [6] and then used to setup an AIA. Indeed, most prior research considers the ecosystem of social networks, due to the ease of retrieving information linking public data with private attributes: Goelbeck et al. [26] infer personality traits of social media users. Jurgens et al [35] consider Twitter, and predict the location of the users based on their tweets. More recently, Gong et al. [28] focus on Google+ users, whereas Zhang et al. [71] consider, e.g., YouTube, and predict users' gender (above 70% F1-score) based on their historical activity. Similarly, [51] focus on Facebook, showing that the gender can be predicted (~80% accuracy) by analyzing the usage of emojis. (The authors of [38] also consider Facebook, and infer sensitive data which is outside our scope). Other examples are [14, 21, 64, 69]. All such works show that AIA can be enacted in the real world, representing a subtle privacy risk.

2.2.2 Motivation: AIA and Video Games. Surprisingly, no efforts consider AIA exploiting (public) VG data to infer players' (private) attributes—to the best of our knowledge. As shown in §2.1, the competitive VG ecosystem (and especially the one of Dota2) is particularly prone to the risk of AIA. A trace of such exposure is provided by the few works analyzing the correlation between the players' in-game behaviour and their personal life—albeit for VG of different genres. For instance, Oggins et al. [50] highlighted that MMORPG players have a similar in- and off-game behaviour. Martinovic et al. [41] reasoned on how such similarity can be used by producers of MMORPG. For instance, some players' physical traits can be inferred from their in-game avatar—which tends to be alike [48]. In this context, Likarish et al. [40] analyzed the in-game avatars to predict the age of the corresponding player; whereas Symborski et al. [61] predicted the gender. Besides physical characteristics, some researches also studied personality indicators. Spronck et al. [58] found correlations between personality traits of 36 players and their playing-style. The only paper we are aware of that considers a competitive VG is [63], showing correlations between Battefield3 players' in-game data and some of their personality traits.

Most related studies on VG (i) did not consider MOBA—which are our focus; and (ii) adopted the perspective of the producers of the VG—i.e., they assumed the availability of in-game data that was not publicly available [19, 56]. The latter is crucial: a *real* attacker is unlikely [5] to have access to a company's databases—especially in domains with a high market share, such as (competitive) VG. Granted: such studies showed that correlations exist between players' in- and off-game characteristics, *but in different VG*. No paper, however, investigated: (i) whether a correlation exists also in Dota2; and, if it exists, (ii) 'how' and 'to what extent' it can be exploited in the Dota2 ecosystem by real attackers—who are not omnipotent. The only work that considers a similar setting as ours is [17], but it focused on recognizing the play-style of Dota2 players across different accounts—which is an objective orthogonal to ours. To the best of our knowledge, we are the first to investigate AIA in VG.

3 DOTA2 ATTRIBUTE INFERENCE ATTACKS

Our primary contribution is the first threat model for feasible AIA against Dota2 players. We describe 'how' AIA can be staged in the Dota2 ecosystem (§3.1); and 'why' attackers would do so (§3.2).

3.1 Proposed Threat Model

Our AIA is mostly tailored for players who actively engage in *competitive* Dota2 matches. (Some Dota2 players do not "play to win", and hence are less likely to use TW.) For simplicity, we assume that a player only owns a single 'handle' (e.g., "Dendi" in Fig. 1 is the handle of the player "Danil Ishutin"), which is used to retrieve data from any public source (e.g., tracking websites).

Formal Definition. We describe the viewpoint of our considered attacker according to the following four criteria [10]:

- *Goal:* The attacker wants to infer the personal attributes of a set of players whose real identity is completely private.
- *Knowledge:* The attacker knows the handles of a set of players, and is well-aware of the Dota2 ecosystem.
- *Capability:* The attacker can only access and retrieve data that is either publicly available, or that users are willing to share (e.g., social networks, public surveys).

- *Strategy:* The attacker first (legitimately) gathers information associating players' in-game data with their respective personal attributes. Then, the attacker trains a ML model to perform AIA against players whose personal information is completely private, i.e., by only using their (known) handle.

We implicitly assume that the targeted players enabled in-game data sharing with external sources (e.g., TW). We stress that the attacker shall *not* perform any data breach to obtain the desired private information—an attacker will never launch an AIA otherwise.

Practical scenario. We present in Fig. 3 an illustration of our threat model, which is divided in three stages: *prepare*, *infer*, *exploit*.

(1) *Prepare.* First (left of Fig. 3), the attacker must collect a representative dataset associating Dota2 players' in-game data (e.g., daily matches played, win/loss ratio) with the corresponding ground truth (e.g., the players' gender).
(2) *Infer.* Then (middle of Fig. 3), the attacker uses the harvested dataset to train a ML model, which is the tool to carry out the AIA. The inference is done by providing public in-game information on a target player (obtainable, e.g., by querying a TW with the handle of a player) as input to the ML model.
(3) *Exploit.* Finally (right of Fig. 3), the attacker benefits by either stalking a victim (targeted AIA), or by profiting from the inferred attributes (an indiscriminate AIA).

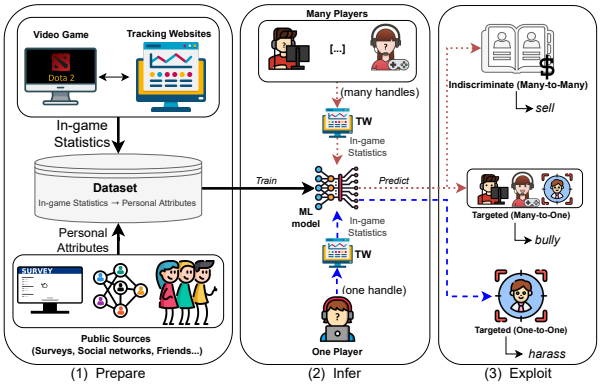


Fig. 3: Overview of our proposed AIA against Dota2 players. Public information is used to infer personal (private) attributes. Besides privacy violations, attackers can harass or bully their victims, or profit from all the inferred attributes.

3.2 Feasibility of AIA in Dota2

Any attack is theoretically possible, and several papers (e.g., [7]) advocate to always consider worst-case scenarios. Nonetheless, we argue that our proposed AIA are not only possible "in theory", but also likely to occur "in practice" due to their high feasibility [5]. Indeed, real attackers have a cost-benefit mindset [68]. In our case, AIA will be launched only if an attacker finds them easy to setup (in terms of cost and risk), and if they lead to tangible benefits.

In particular, we focus the attention on three aspects—each pertaining to a given stage of our exemplary use-case, namely: acquiring the dataset to train the ML model (i.e., the *capabilities* of the attacker); improving the performance of such ML model (i.e., the *knowledge* of the attacker); and how a successful AIA can be exploited (i.e., the *goal* of the attacker).[1]

- **Data Harvesting.** Obtaining *public* in-game data of multiple players (i.e., the "features") is straightforward in Dota2: it is simply necessary to go to a TW[2] and retrieve all information related to a set of players. In contrast, obtaining the corresponding personal attributes (i.e., the "labels") may appear harder, as such information is typically kept *private*. Unfortunately, this is not the case in the Dota2 ecosystem.[3] For instance, the real identity of many players (e.g., professionals or streamers) is well-known. Moreover, it is possible to search for a given handle on popular search-engines and inspect the results. For example, a given player may use the same handle also on social media; some people even announce their handle on public forums to facilitate establishment of partnerships. Alternatively it is also possible to conduct surveys in which interviewees must input their handle, as well as some inconspicuous private information (e.g., gender, age). For instance, two large surveys were carried out in 2016 and 2021, *receiving 30k and 8k responses* respectively, by simply posting announcements on popular boards [22].
- **Refining the ML model performance.** Even if an attacker can acquire a suitable training dataset, it is unlikely that such dataset can yield a proficient ML model from the start—hence, *naive* attackers will hardly be successful in their AIA. *Expert* attackers, however, can use their superior knowledge on the Dota2 scene to improve the success rate of their AIA. In our evaluation (§5) we will show some pre- and post-processing techniques that boost the predictive performance of the ML model. Given that attackers interested in our AIA are well-aware of how Dota2 works, this characteristic further aggravates the threat of AIA.
- **Exploiting AIA**. We identify three ways in which an attacker can benefit from AIA in Dota2. (We will consider all of these ways in our evaluation.) First, they can launch an *indiscriminate* 'many-to-many' AIA, i.e., by using many handles (belonging to many players) to infer the respective personal attributes; such attributes can then be sold[4] to any potential buyer—e.g., dark web, or even to ad-companies which want to send customized ads [55]). Second, they can launch a *targeted* 'one-to-one' AIA by inferring the attributes of just one player—e.g., after losing a match, an attacker can launch an AIA against a player of the opposing team and harass them [17]). Third, they can launch a *targeted* 'many-to-one' AIA by inferring the attributes of a set of (many) players, and then finding a (single) player within such set that meets some criteria—e.g., finding an underage player and then bully them [25, 47, 54].

Finally, we observe that the results of the two surveys [22] showed similar trends despite the 5 year timespan. Such stability may suggest that even **data collected many years prior can still be used to enact successful AIA**. Considering the high likelihood of such

[1] We observe that our threat model is significantly different from the one in [42].
[2] We observe that abundant information is also available directly from Dota2, hence TW are not strictly required (we will discuss this in §7).
[3] Zhang et al. [71] also state that ground-truth harvesting is easy in today's landscape.
[4] This is a popular strategy adopted by some real companies [45].

a threat, we embrace Biggio and Roli's [10] recommendation: we must proactively assess the impact of AIA in Dota2.

Takeaway: Attackers can – cheaply and legitimately – use many methods to setup an AIA, which can be exploited in various ways to violate Dota2 players' privacy.

4 PRELIMINARY ASSESSMENT

A prerequisite for a successful AIA is the existence of relationships between the players' in-game data, and their corresponding personal attributes [70]. We recall (§2.2.2) that past research found some correlations—but *in different VG* (e.g., Battlefield3 [63]).

Hence, as our second contribution, we now investigate whether there is some evidence hinting that AIA "can be successful in Dota2". To this purpose, we perform an extensive survey on real Dota2 players (§4.1 and §4.2), and analyze the correlation coefficient between their in-game statistics and personal attributes (§4.3).

4.1 Collection of personal attributes (survey)

We conduct a survey to collect the handles of Dota2 players, together with their personal attributes.

Method. The handle consists in the Steam ID of each player. For the personal attributes, we consider: gender, age, occupation, purchase_habits, as well as the "Big Five" personality traits [65]). Such attributes are those typically envisioned by past research (e.g., [26, 51, 69, 71]); the only exception is purchase_habits, which is an 'original' attribute that we propose due to the given Dota2 context, in which players typically purchase "cosmetics" to embellish their characters. Nevertheless, all such personal attributes represent information that is *not available* from any resource linked with Dota2: hence inferring such information without the explicit consent of the corresponding player represents a privacy violation.[5] Our survey entailed 10 questions used to determine the personality traits [53]; 4 questions which explicitly referred to the remaining four attributes considered in this paper; as well as one question for the country. We also included 10 questions, which served both as 'attention checks', but also for verifying the authenticity of the answers (e.g., we asked "what is your favorite Dota2 hero?" and we verified on a TW whether the answer was genuine).[6] Overall, the survey began in Oct. 2019 and ended in Dec. 2019. In this timeframe, we hosted our survey on a website, whose link was distributed on many online social media platforms such as Facebook, Reddit, Discord, and Telegram. Upon landing on the survey's website, participants had to login with their Steam account (via OpenID), thereby ensuring that all personal attributes were correctly linked to the actual player.

Analysis. We received 625 answers from 62 different countries. We filtered out: 18 invalid answers; 43 participants who were not visible on any TW; and 78 inactive players (i.e., less than 5 games in the last month). Thus, our sample size consists in 484 players. Despite being far smaller than the overall amount of Dota2 players, such number still allows to draw statistically significant result. Indeed, we are above the minimum sample size of 384 required by setting a confidence level of 95%, a margin of error of 5%, population proportion of 50%, and a population size of 7 million [39]. We report in Table 1 the considered personal attributes, as well as their class-distribution in our population. We grouped age in three bins (similarly to [14]): very young/underage, young adults, and over 25 (our 'oldest' respondent was 38); the frequencies for purchase_habits are never, less than once a month (rarely), and monthly or more often (regularly); for occupation, we consider a student as unemployed. Since our survey quantified each personality trait as an integer [0–100], we group such values into three categories (similarly to [11]) differentiating low, middle, or high scores.

Table 1: Personal attributes considered in our study. Our population is of 484 Dota2 players. The distribution resembles the one in [22].

Private Attribute	*Description*	*Classes Distribution*
gender	Gender at birth	*Female*: (4.96%), *Male*: (95.04%)
age	Current age	*13–18*: (13.43%), *19–24*: (53.72%), *25–38*: (32.85%)
occupation	Whether a player is employed or not	*No*: (57.44%), *Yes*: (42.56%)
purchase_habits	Frequency of in-game purchases	*Never*: (10.54%), *Rarely*: (61.16%), *Regularly*: (28.30%)
openness	Inventive/curious (high) vs. consistent/cautious (low)	*Low*: (19.22%), *Medium*: (24.38%), *High*: (56.40%)
conscientiousness	Efficient/organized (high) vs extravagant/careless (low)	*Low*: (39.46%), *Medium*: (23.97%), *High*: (36.57%)
extraversion	Outgoing/energetic (high) vs. solitary/reserved (low)	*Low*: (47.31%), *Medium*: (21.07%), *High*: (31.62%)
agreeableness	Friendly/compassionate (high) vs. critical/rational (low)	*Low*: (20.87%), *Medium*: (19.42%), *High*: (59.71%)
neuroticism	Sensitive/nervous (high) vs. resilient/confident (low)	*Low*: (53.51%), *Medium*: (19.21%), *High*: (27.27%)

Validation. By observing Table 1, we can see that some classes may present a high imbalance, such as gender or age. However, our class-distribution is strikingly similar to those of the surveys carried out in previous years [22]: specifically, we focus on the largest survey from 2016, whose sample size was of 29,351. Let us make some exemplary comparisons, so as to *validate* all our subsequent analyses: if our population significantly differs from the 'real' one, then we cannot claim that the threat is 'real'. According to [22], *male* players are 96%, which match our results of 95%. The same can be said for age: according to [22], minors represent 15% of the population (ours is 13.4%), whereas young adults are 66% (ours is 54%), with over 25 being 20% (ours is 33%). (Small differences are due to slightly different thresholds for the bins). For occupation, the unemployed are 67% in [22] (ours is 57%).

Summary: from our survey, we derive that: our population (i) is representative of the Dota2 community, and (ii) is large enough to derive statistically significant conclusions. Moreover, our survey also shows that (iii) the Dota2 community is willing to participate in online surveys—representing one of the means an attacker can use to harvest players' private information for a (real) AIA.

We will use $\mathcal{A}$ to indicate the dataset containing the (personal) *attributes* of our 484 players—collected via our ethical survey.[7]

4.2 Collection of in-game statistics (TW)

Once we obtained the handles of the participants, we retrieved their in-game statistics via public Tracking Websites.

Method. Our TW of choice is OpenDota because it provides free APIs[8] usable to retrieve in-game statistics. We used two APIs:

- `player`, which, given a handle, returns some summary statistics (e.g., win/loss ratio) of the corresponding player, as well as the list of matches[9] played by such a player;
- `matches`, which, given the identifier of a match (obtained from the `player` API), returns all information on that specific match (e.g., kills, deaths, assists).

[5] Even purchase_habits is not public: a player may have many "cosmetics", which can have been *gifted*; moreover, a single purchase may include *more* than a single "cosmetic", which can also be obtained via "bundles".

[6] Our repository includes the full questionnaire. Some questions found therein asked for other (non-sensitive) information that do not pertain to this paper.

[7] We never attempt at inferring additional (private) information of our respondents.

[8] OpenDota API: https://docs.opendota.com/

[9] For simplicity, we only considered the matches played in the previous 30 days since making each API call (i.e., from December 2019 to January 2020).

We report in Table 2 the information returned by our invoked API. Some fields are provided as lists, which include additional entries. For example, `matches_chat` includes all messages exchanged by the two opposing teams during a Dota2 match. For a detailed explanation of all fields, we refer the reader to the official documentation.

Table 2: Data returned by the `player` and `matches` OpenDota APIs.

Type	*Field*	*Type*	*Field*	*Type*	*Field*
num	player_rank_tier	num	match_human_players	list	match_radiant_team
bool	player_plus	num	match_lobby_type,	list	match_dire_team
list	player_matches	list	match_objectives	num	match_skill
num	match_match_id	list	match_picks_bans	list	match_players
num	match_barracks_status_dire	num	match_positive_votes	num	match_patch
num	match_barracks_status_radiant	list	match_radiant_gold_adv	num	match_region
list	match_chat	num	match_radiant_score	list	match_all_word_counts
list	match_cosmetics	bool	match_radiant_win	list	match_my_word_counts
num	match_dire_score	list	match_radiant_xp_adv	num	match_throw
list	match_draft_timings	num	match_start_time	num	match_comeback
num	match_duration	list	match_teamfights	num	match_loss
num	match_first_blood_time	num	match_tower_status_dire	num	match_win
num	match_game_mode	num	match_tower_status_radiant		

Overall, after querying the `players` API for each of the 484 players, we found out that our population participated in 26241 matches during the considered timeframe. Therefore, we invoked the `matches` API on all these entries.

Preprocessing. By applying original feature engineering techniques on the data retrieved from OpenDota, we distill additional knowledge to assist in our analysis. Such techniques involve both 'traditional statistics', but also our own 'domain expertise' on Dota2.

- *Traditional Statistics.* The most straightforward operation involves computing some aggregated metrics on the details of each match played by a given player (e.g., average match length). We also perform some more refined operations. For instance, the `players` API does not directly provide the playtime trend of a given player, but such information can be computed by using the results from `matches`: by inspecting the dates of the matches played, we can identify, e.g., which day of the week a given player is most likely to play Dota2.
- *Domain Expertise.* By applying knowledge on the Dota2 context, we further increase the amount of information usable for our analysis. As an example, we inspect all chat messages to determine whether players use words that are typical of Dota2 slang (e.g., "cd", "b", "rat", "smurf", "gank"). We provide in Appendix A an additional description of how we computed the features related to `match_chat`.

Overall, we compute over 300 features—all of which are novel in the context of AIA.[10] Such features identify three datasets: $\mathcal{P}$, focused on the players, containing 484 samples, each described by 187 features; $\mathcal{M}$, focused on the matches, containing 26241 samples, each described by 137 features; and $\overline{\mathcal{M}}$, containing 11117 samples and 160 features, which is a 'distilled' version of $\mathcal{M}$. In particular, $\overline{\mathcal{M}}$ differs from $\mathcal{M}$ in two ways: First, we address the problem of the highly imbalanced distribution of $\mathcal{M}$ in terms matches-per-player (some players in $\mathcal{A}$ have only 5 matches in $\mathcal{M}$, while others have hundreds); we thus reduce the potential bias by randomly sampling[11] at most 30 matches for each player. Second, we augment the features in $\mathcal{M}$ with those derived with our domain knowledge; the intention is determining how much of an impact our intuitions (resembling those of an attacker) have on all our experiments.

[10] A complete description of all our considered features is provided in our repository.
[11] To mitigate the effects of randomness, we create 20 versions of $\overline{\mathcal{M}}$ and will use all of these for our experiments, averaging the results.

4.3 Correlation between Dota2 in-game statistics and personal attributes

We can now objectively determine whether a relationship exists between Dota2 players' in-game statistics and their personal attributes. This step is crucial to provide a theoretical foundation supporting the effectiveness of AIA in this context.

Method. We perform a correlation analysis between our three dataset containing in-game statistics, and the dataset containing corresponding personal attributes. Inspired by [26], we compute the correlation between each feature of $(\mathcal{P}|\mathcal{M}|\overline{\mathcal{M}})$, with each feature of $\mathcal{A}$. To conduct a rigorous analysis, for each pair of features we compute: (i) the *statistical significance* of the correlation—measured with a p-value; and (ii) the corresponding *strength of the relationship*—whose measure varies depending on the chosen correlation metric. We consider two metrics [3]: *Cramer's V* for categorical variables; *Spearman's* ρ for numerical variables. We remark that low p denotes strong significance (we set $p < 0.01$ as default threshold), whereas strong relationships are denoted by high absolute values of the corresponding metric (ranging between 0 and 1).

Results. We report in Fig 4 the correlation between $\mathcal{P}$ and $\mathcal{A}$ as measured by the ρ metric. For each numerical variable in $\mathcal{A}$, we report the top-3 variables[12] of $\mathcal{P}$ (as measured by ρ), all of which obtain $p < 0.01$. We can see that age is correlated with kills, probably because younger players have an aggressive playstyle. A strong correlation exists between purchase_habits and (i) cosmetics_prices, i.e., the money spent by a player in skins; and (ii) special messages (i.e., hero_msg and counter_thank_msg) that can be unlocked with a paid subscription. Moreover, extroversion is highly correlated to chat usage (i.e., rank_chat and ratio_chat_msg); whereas agreeableness to wins in unranked games (i.e., normal_win). Interestingly, neuroticism is correlated with denies (a unique mechanic of Dota2), openness to the type of selected heroes, and conscientiousness is low for players that play on Thursdays. Although not shown in Fig. 4 (because they are categorical features), we also mention high correlation between the gender of the player and the gender of the most played heroes (which is common in cooperative VG [61]); whereas the occupation is strongly correlated to paid subscriptions.

> **Takeaway.** A correlation exists between Dota2 players' in-game data and their personal attributes. Our finding demonstrates the risk of AIA in Dota2.

We report in Appendix B an extended analysis on the correlations between $(\mathcal{M}|\overline{\mathcal{M}})$ and $\mathcal{A}$. Our repository includes all heatmaps.

5 PROACTIVE EVALUATION OF AIA IN DOTA2

Our preliminary assessment provides evidence that AIA against Dota2 can be successful. Hence, as our third contribution, we set out to proactively evaluate the impact of such AIA. To this purpose, we use the data derived from our survey (described in §4.1 and §4.2) to perform various ethical and controlled AIA.

Specifically, we find instructive to study three diverse AIA, each requiring different amounts of preparation. First, we consider the most *simple* way to carry out an AIA, i.e., by using only the aggregated data of each player (§5.1). Second, we evaluate the success

[12] We remark that $\rho > 0.1$ is a valid signal indicator for orthogonal tasks [44].

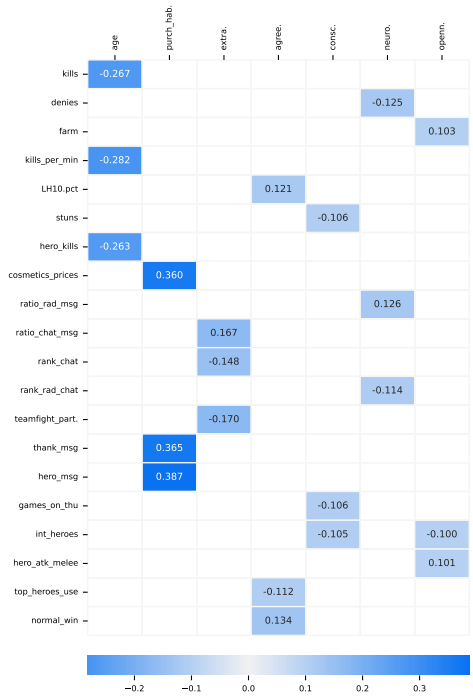


Fig. 4: Top-3 Spearman's ρ between $\mathcal{P}$ and $\mathcal{A}$ (at $p < 0.01$). Higher absolute values denote stronger correlation, while the sign indicates the direction of the correlation.

rate of AIA that use information derived from just *one match* (§5.2). Third, we analyze *sophisticated* AIA in which the attacker leverages all their expertise to maximize their impact (§5.3). Finally, we perform a reflective exercise by discussing the general context of AIA in light of the results achieved in research (§5.4). We also perform a statistical validation of all our results in Appendix C.

Common Setup. We always adhere to our threat model (§3.1). The attacker knows the handle of one or more players, and uses such handle to retrieve in-game data from TW, which are then provided as input to an ML model for inference. Moreover, we also assume that the attacker gathered the private attributes for training the ML model via a survey (i.e., the one described in §4.1). Indeed, as evidenced by [22], thousands of Dota2 players willingly participate in game-related surveys. For ethical reasons, we do not violate our respondents' privacy by performing OSINT, or crawl their social media profiles (which are both viable means that an attacker can – legitimately – use to improve their AIA).

5.1 Simple AIA (aggregated player data)

The underlying principle of these AIA is that they only use the information contained in $\mathcal{P}$, i.e., which aggregates the statistics of all matches played by any given player. Such information is simple to compute, but is lossy. For instance, the average_match_length includes the duration of all matches, and inevitably leads to oversimplifications. However, due to their simplicity, such AIA are feasible to stage and it is important to assess their impact.

Testbed. For these experiments, we merge $\mathcal{P}$ with $\mathcal{A}$, generating a single dataset containing 484 samples, each described by 187 features (from $\mathcal{P}$) and associated to 9 attribute labels (from $\mathcal{A}$). To develop the ML model for the AIA, we consider four ML algorithms: Logistic Regression (*LR*), Decision Trees (*DT*), Random Forest (*RF*), and Neural Networks (*NN*). We validate our results through a nested stratified 10-fold cross-validation, during which we also apply feature selection and hyperparameter optimization for each considered ML model. Finally, to address the imbalance of some target attributes (e.g., age), we apply well-known under- and over-sampling techniques [13, 67] (as also recommended in [7]).

Impact. We report in Table 3 the results of the simple AIA. Rows denote the target attributes, whereas columns denote the considered ML algorithms; the rightmost column refers to a 'Dummy' stratified classifier (simulating a random guess) which we use as baseline for comparison. Cells report the predictive macro F1-score (and standard deviation) across all our trials.

Table 3: Impact of the *simple* AIA (based on $\mathcal{P}$) as measured by the F1-score. Rows report the attributes and columns our ML models (boldface denotes the best model for a given attribute).

	LR	*DT*	*RF*	*NN*	*Dummy*
gender	64.97±10.9	59.71±12.7	50.91±5.33	**67.24±13.4**	51.62±10.9
age	40.47±6.30	39.38±8.76	**44.08±6.17**	28.06±7.59	32.21±5.70
occup.	53.23±7.22	47.44±8.34	56.08±7.88	**59.89±7.15**	43.76±9.56
purch.	32.05±10.1	31.74±4.53	**34.40±8.20**	32.17±7.19	31.20±6.26
open.	28.94±5.94	**40.76±6.80**	32.6±7.77	30.89±7.60	29.59±2.04
consc.	26.52±5.65	33.87±8.78	**34.27±5.60**	23.83±8.18	33.23±8.94
extrav.	30.15±7.53	36.16±5.14	**36.49±5.56**	28.59±5.95	32.27±7.01
agreeab.	29.46±6.29	**34.11±8.58**	33.68±6.25	24.54±9.43	33.39±7.35
neurot.	32.38±6.56	**40.76±6.80**	32.6±7.74	31.6±8.30	30.07±4.46

From Table 3, we observe that at least one of our models always outperforms the baseline. The *NN* achieves remarkable performance (almost 70% F1-score) to predict gender, whereas occupation is correctly predicted with almost 60% F1-score. In contrast, some attributes are very difficult to predict, such as purchase_habits for which the performance hardly goes 3% above the baseline. We can conclude that such simple AIA can be effective in some cases, but real attackers can easily improve the success rate by considering additional information—as we will show in §5.3.

5.2 One-match AIA (ablation study)

We now assess the effects of AIA carried out by using the statistics of just *a single match*. This scenario can be considered as either a best-case or a worst-case depending on the viewpoint. Indeed, we can expect that using only one match to predict the personal attributes may yield poor results—which is a best-case for the defender. However, if such an AIA is successful, it would turn into a worst-case because the attacker can infer the private attributes with limited information (e.g., less queries to the TW API).

Moreover, we consider two attackers: an 'expert' attacker that uses their *domain expertise* to distill additional knowledge from the single match; and a 'naive' attacker that does not do so. Hence, the results of the 'naive' attacker can serve as an *ablation study*, allowing to gauge the effects of domain expertise in AIA.

Testbed. To simulate the 'naive' attacker, we merge $\mathcal{M}$ with $\mathcal{A}$. Hence, for each of the 26241 matches in $\mathcal{M}$ (described by 137 features), we append the 9 attributes of $\mathcal{A}$. For the 'expert' attacker, we merge $\overline{\mathcal{M}}$ (having 11117 matches, each with 160 features) with $\mathcal{A}$, because $\overline{\mathcal{M}}$ is augmented with Dota2 domain knowledge. We

consider the same ML algorithms as in the simple AIA (i.e., *RF*, *LR*, *NN*, *DT*). We then train and test ML models by adopting a split of 80:20 (such split is done on the basis of the unique players in $\mathcal{M}$ (or $\overline{\mathcal{M}}$) to avoid overfitting); we reserve 10% of the training set for validation purposes. Finally, we repeat all our experiments 20 times to account for the random sampling of $\overline{\mathcal{M}}$.

Impact. We report the results in Table 4; for simplicity, we only consider the models using *RF*, because they consistently outperformed all the others. The three columns show the F1-score obtained by the 'naive' (left) and 'expert' (middle) attackers, as well as that of a 'Dummy' classifier (right) that simulates a coin-toss.

Table 4: Impact of the *one-match* AIA (F1-score). Columns refer to the 'naive' attacker (using $\mathcal{M}$), 'expert' attacker (using $\overline{\mathcal{M}}$), and the Dummy (random guess). The expert attacker is always superior.

	Naive attacker (ablation study)	Expert attacker (domain knowledge)	Dummy (baseline)
gender	49.03±0.18	**58.47**±5.21	49.75±0.55
age	**43.72**±2.66	**56.82**±3.01	33.28±0.46
occup.	49.42±4.56	**68.42**±1.90	49.87±0.89
purch.	**35.61**±5.06	**49.71**±3.85	33.37±0.53
open.	32.26±3.75	**43.73**±2.96	33.48±0.41
consc.	29.49±3.63	**46.11**±3.20	32.88±0.62
extrav.	32.33±2.47	**46.82**±1.96	33.25±0.56
agreeab.	33.62±2.28	**45.36**±3.37	34.09±0.46
neurot.	27.39±4.78	**46.60**±2.72	33.65±0.58

From Table 4, we can see that the 'naive' attacker cannot successfully predict 8 out of 9 attributes, because the F1-score is always comparable (or even inferior) than the Dummy classifier. The only exception is the age attribute, for which the F1-score is 10% superior (albeit still hardly usable). We also note that such results are inferior to those of the simple AIA (cf. Table 3). From a defender's viewpoint, these results may appear encouraging. Unfortunately, the 'expert' attacker is much more successful, with 10–20% improvements over the Dummy classifier. Notably, occupation reaches ~70% F1-score (up from 49%), whereas gender almost 60% (up from 49%). Such results prove that using domain knowledge of Dota2 substantially improves the success of AIA. What is surprising is that such AIA require the statistics of *a single match* (i.e., just one API query).

5.3 Sophisticated AIA

We now assess AIA launched by a sophisticated attacker who, alongside using their domain expertise during pre-processing, exploits post-processing methods to further improve the AIA success rate.

Intuition. We build from the one-match results of the the 'expert' attacker (§5.2). Then, we leverage the fact that a given Dota2 player (i.e., the one targeted by the attacker) typically plays many matches. It is reasonable to assume that said player exhibits a *stable behaviour* across all such matches. Indeed, taken individually, a single match may not capture the true behaviour of a given player, thereby leading an ML model to make a wrong prediction; however, by considering the predictions of the *same* ML model to *many* matches (from the same player), the stable behaviour (i.e., the desired attribute) of the targeted player is more likely to emerge. For example, a player that has 'high' openness may not show such trait in every single match; but such trait may emerge by (independently) analyzing more matches, and aggregating the results.

Testbed. We use the ML models trained with $\overline{\mathcal{M}}$ using the *RF* algorithm. Then, we provide as input to such models an increasing amount of matches from the same targeted player: specifically, we consider from 1 up to 30 matches (if available), which are randomly sampled (from the test portion of $\overline{\mathcal{M}}$). Then, for each attribute in $\mathcal{A}$, we take the predictions (provided as probabilities) of the ML model for all such matches, and we average all such predictions, choosing the one with the higher value.[13] To reduce bias, we repeat all such experiments 20 times for each different variant of $\overline{\mathcal{M}}$; and, we repeat the draw of the chosen matches 1000 times.

Impact. The results of our sophisticated AIA are shown in Fig. 5, showing accuracy (y-axis) as a function of the matches analyzed by the ML model (x-axis). Lines correspond to the target attributes; shaded areas show the standard deviation. We do not report gender because the highly unbalanced population would inflate the results.

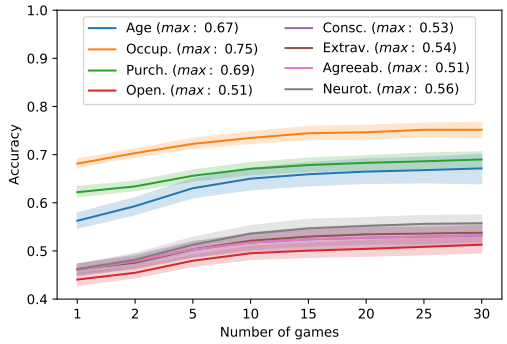


Fig. 5: Impact of Sophisticated AIA. We post-process the predictions of the ML model over multiple matches of the same targeted player.

From Fig. 5 we can see that the accuracy increases as more matches are analyzed. For example, predicting the occupation goes from 68% up to 75% after 15 matches. Similarly, age goes from 58% up to 65%. What makes these results concerning is that retrieving the information on such extra matches requires little effort by the attacker, because (i) it is free and (ii) it can be automatized.

5.4 Reflection: AIA in research and in practice

As a reflective exercise, we report in Table 5 the results (according to a given 'Metric') obtained by some prior works attempting to predict the same attributes considered in our paper (we exclude purchase_habits because it is novel). We stress that Table 5 is not meant to be a way to compare our AIA with previous ones, since we are the first to consider the Dota2 setting (§2.2.2). Moreover, past works envision (i) different classes having (ii) different distributions for each attribute—making any comparison unfair.

From Table 5, we can see that – from a general viewpoint – obtaining high performance (e.g., overall *accuracy*) is difficult for some attributes. However, the real threat of AIA lies in the fact that they can be customized: although precisely inferring, e.g., the age of *all* individuals among a population may be unfeasible, it is different when the objective is more specific. For instance, an attacker may

[13] E.g.: we want to predict the occupation (which is binary) of a player by analyzing 4 matches. The ML model analyzes 4 matches and outputs 4 probabilities, e.g., {0.1, 0.2, 0.8, 0.2} (i.e., values below/above 0.5 denote employment/unemployment). We assign the class after averaging the probabilities, thereby 'filtering' the noise (i.e., the 0.8).

Table 5: Results of prior work on AIA. Cells denote the value of a given 'Metric' for each of the attributes considered in our paper.

Prior Work	Metric	gend.	age	occup.	open.	consc.	extrav.	agreeab.	neurot.
Goelbeck [26]	MAE	–	–	–	0.09	0.10	0.14	0.11	0.13
Weinsberg [64]	AUC	0.84	–	–	–	–	–	–	–
Al [4]	Acc.	0.80	0.80	–	–	–	–	–	–
Chen [14]	AUC	0.82	0.61	–	–	–	–	–	–
Fang [23]	Acc.	0.80	0.73	0.25	–	–	–	–	–
Bunian [11]	Acc.	–	–	–	0.58	0.60	0.58	0.58	0.58
Yo [69]	Acc.	0.70	0.80	0.70	–	–	–	–	–
Mei [43]	MAE	–	0.09	–	–	–	–	–	–
Pijani [51]	F1	0.83	–	–	–	–	–	–	–
Zhang [71]	F1	0.74	0.38	0.13	–	–	–	–	–
Eidizadehakhcheloo [21]	AUC	0.95	0.98	–	–	–	–	–	–

want to identify just a specific group of people (e.g., children—see §3.2), and they can tweak their ML models for this exact purpose.

A positive message. Our paper tackles an open issue, and our ultimate goal is to cast light on a real[a] problem—and not to aggravate such problem. Hence, for the sake of responsible research, we will now showcase only a few 'practical' AIA, having near-perfect success rate.

[a]The problem is real, and we demonstrated it. Our survey resembles Dota2 population (§4.1), the statistical analysis proves the existence of correlations (§4.3) and our evaluation shows improvements over the baselines (§5).

6 PRACTICAL AIA (THE TRUE THREAT)

Insofar, the objective of our AIA was always to infer *each* class by independently considering every attribute. According to our threat model (§3.2), such AIA conformed to the targeted 'one-to-one' category: given *any* player, infer (*all* of) their attributes. The results (in §5), despite being arguably serious, may not induce real attackers to launch most of such AIA (aside from, perhaps, those on occupation): some players exhibit traits that are difficult to infer.

However, attackers can also launch two other categories of AIA, which can yield 'devastating' results while being surprisingly simple to carry out. As our fourth and last contribution, we now elucidate the effects of some indiscriminate 'many-to-many' AIA (§6.1), and of some targeted 'many-to-one' AIA (§6.2).

6.1 Indiscriminate 'many-to-many' AIA

Let us assume an attacker whose goal is to sell the inferred attributes to the black market. Such an attacker may want to advertise their data as being "most likely correct"; put differently, the attacker wants to ensure that the inferred information is "unlikely to be completely incorrect", thereby accepting some margin of error.

Method. We use exactly the same setup as in the 'sophisticated' AIA (§5.3), where the inference is done after analyzing multiple matches. However, for these AIA, we assume an attacker who is satisfied as long as the prediction is not completely wrong. For instance, assume that a player has 'high' openness (cf. Table 1): we consider the AIA to be successful if the probability associated to 'high' is either at the first or second place among all the possible classes (three in this case). A similar scenario describes an AIA in which the attacker wants to find, e.g., a player who is "likely to be open" (i.e., has 'high' openness either at the first or second place).

Impact. We report the results of these AIA (after using 30 matches) in the central column of Table 6, in which rows denote the attributes (we exclude those that only have two classes, as it would be unfair to include them); the leftmost column denotes the accuracy obtained by the sophisticated AIA (cf. Fig. 5), whereas the rightmost column denotes the improvement (as a flat difference). From Table 6, we can see a big jump in predictive accuracy with respect to Fig. 5. For instance, inferring age reaches 89% accuracy, whereas purchase_habits goes from 65% to 96% accuracy. Remarkably, this method is the only one that provides usable results for agreeableness and openness, both with ~80% accuracy. Despite bearing some intrinsic margin of errors (because the predicted class is not guaranteed to be the exact one), an attacker can still benefit from such imprecision, making these AIA a tangible threat.[14]

Table 6: Indiscriminate 'many-to-many' AIA (mid column). Compared to the baseline (cf. Fig. 5), the accuracy substantially increases.

	Sophisticated AIA (30 matches)	Indiscriminate AIA (30 matches)	Improvement
age	67.15±6.87	**89.15**±4.66	+22.00%
purch.	68.99±3.81	**96.13**±2.86	+27.14%
open.	51.30±3.87	**77.86**±3.39	+26.56%
consc.	53.24±4.88	**80.19**±4.12	+26.95%
extrav.	53.78±3.90	**81.51**±4.40	+27.73%
agreeab.	50.71±4.65	**76.84**±5.59	+26.13%
neurot.	55.74±3.88	**80.64**±4.02	+24.90%

6.2 Targeted 'many-to-one' AIA

We now assume an attacker who wants to find players that present specific traits among a large population, e.g., finding very young players. In these cases, the attacker would train their ML models to maximize the *precision* on a given class, so as to minimize the amount of false positives. Although a similar strategy inevitably leads to a reduced *recall*, this is not an issue in reality: the attacker is not interested in, e.g., "finding *all* young players" (which is an unfeasible objective), but rather "finding a subset of those players that are *guaranteed* to be young". Such scenario is even more problematic than the previous ones, especially given that a low recall is not an issue when the population counts millions of players.

Targets. We consider an attacker that is interested in identifying four "vulnerable" groups of players[15]. Specifically: "very young" (age=*13–18*), "purchasers" (purchase_habits=*Rarely* ∨ *Regularly*), and "introverts" (extraversion=*Low.*) Moreover, we also consider an attacker that attempts an 'intersectional' AIA, wherein the targeted group conforms to two specific classes of two *distinct* attributes. In this case, the attacker wants to pinpoint "purchasers & workers" (occupation=*Yes*, and purchase_habits=*Rarely*∨*Regularly*), which could be ideal to identify players to which advertise new products—because such players tend to make purchases, and are likely to have the economical resources for doing so (as they have a job).

Testbed. We adopt a similar setup of the sophisticated AIA (§5.3), i.e., we use $\overline{\mathcal{M}}$ as dataset, and evaluate the performance of our ML models as they analyze increasingly more matches of the same player, and then averaging the output probabilities. The crucial difference, however, lies in the problem formulation, which now reflects a *binary classification* setting: the objective is predicting the targeted class, and anything outside of such class is irrelevant.

[14]We provide a statistical validation of these results in Appendix C.
[15]There are over 8000 possible combinations of all our classes, and investigating all of them is clearly unfeasible and outside our scope.

To this purpose, we first merge all players that do not belong to the targeted class (i.e., the "positive") into a single class (i.e., the "negative"). Then, for each target, we train a (binary) classifier by using the precision as optimization metric (whereas in the sophisticated AIA, we used the macro F1-score). We find the best models and hyper-parameters using a validation set having players never seen at training time, simulating that the attacker can use only data that has gathered. The good results achieved on the validation set (combined with our correlation findings described in §4.3) suggest that the attack is feasible, and would incentivize the attackers to launch it in reality. Last, we evaluate the best models on the test set, having players not included in either the training or validation sets. For each targeted attribute, we repeat all these procedures five times to reduce bias and account for randomness.

Impact. We report in Fig. 6 the *precision* in identifying the targets as a function of the matches analyzed by the ML models.

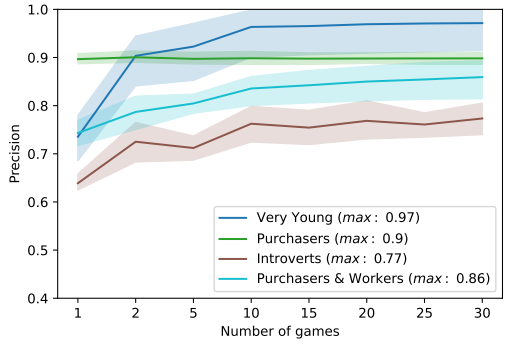


Fig. 6: Targeted 'many-to-one' AIA. We train our ML models by maximizing the *precision* on a single targeted class. Such AIA are very effective after analyzing ~10 matches for each player in the test-set.

It immediately stands out that we obtained much 'dangerous' results than in any of the previously considered scenarios. For instance, by analyzing 10 matches, our ML models can detect "very young" with almost perfect precision. Obviously, this comes at the cost of a low recall, which was about 47% after 30 matches.[16] Moreover, our models ably detect "purchasers" after a single match, achieving a stable 90% precision—surprisingly exhibiting also a recall of 98% after 30 matches (not shown in Fig. 6), suggesting that purchasing indicators are well defined, and the mistakes happened probably when users are gifted expensive items. The models devoted to "introverts" achieve 76% precision (and 73% recall) after with 30 matches, indicating that players belonging to this group have many characteristics in common. Finally, for the 'intersectional' AIA focusing on "purchasers & workers", the models obtain 86% precision (and 47% recall) after 30 matches, suggesting that roughly half of such players exhibit distinctive traits.

Takeaway. Attackers with *specific* goals can easily setup AIA that are highly successful, thereby confirming the exposure of Dota2 players to such privacy threat.

[16]Roughly speaking, we detected half of the "very young", but with no mistakes—i.e., the ML model found ~5 guaranteed "very young" out of ~81 players in the test-set.

7 DISCUSSION

Our proactive evaluation showed that AIA can be highly successful in Dota2. A legitimate observation is that our experiments consider a small subset of all Dota2 players. However, our population still allows to derive statistically significant results (see §4.1). Another observation is that we (ethically) simulated an AIA by collecting personal attributes through a survey (instead of, e.g., scraping social networks [28]). However, as explained in §3.2, Dota2 players are willing to participate in similar surveys (even when promoted by random users [22]). Hence, our (ethical) AIA represents a feasible scenario for an attacker, and our results are statistically significant. Finally, there exist infinite ways in which an attacker can use the collected data to carry out AIA; yet, those considered in our paper confirm our point, i.e., that AIA are a threat to the Dota2 playerbase.

We now discuss some possible mitigations (§7.1), and explain how our threat model can be applied to other E-Sports (§7.2).

7.1 Countermeasures to AIA in Dota2

Our AIA are rooted in the fact that players' in-game statistics are publicly obtainable from TW. The most obvious countermeasure would be denying public access to all such statistics *from the VG itself*. Unfortunately, players are the ones (implicitly) asking for such public availability (see §2.1). Alternatively, Dota2 developers can use our analyses to make the features with stronger correlation to some attributes to be impossible to compute with public data; however, attacker are free to derive also other features—potentially with stronger correlations with (also) other attributes.

It is hence difficult to find a 'general' mitigation that preserves the functionalities of TW while ensuring players' privacy. Yet, in an attempt to reduce the feasibility of an AIA, we propose two countermeasures. **(1)** *TW could allow players to select 'what content' is public.* For instance, a player can have only their last few matches to be visible by anyone. This solution has two drawbacks. First, if the statistics of *other* players in the same match are visible, an attacker could still launch an AIA—albeit at a higher cost, because they need to retrieve the information from the other players (of which they need to know the handle). **(2)** *TW could allow user to choose 'who' can see their profiles.* For instance, two players could browse each other's statistics only if they are friends *within the VG*—which is a different environment than the TW (e.g., Fig. 1 shows the friends within the TW). Such a countermeasure requires, however, a deep cooperation between TW and the VG. Alternatively, visibility can be granted *upon request*.

Unfortunately, both countermeasures impair the use of TW to learn from others players, because their matches would be hidden. The only exception are professional players, whose profiles can be public since they are less likely to be targeted AIA in the first place.

Summary: Countermeasures against AIA present tradeoffs. Our paper will hopefully inspire the search for a cost-effective solution.

7.2 Extension to other E-Sports

Our threat model can cover also other VG beyond Dota2. Indeed, we observe that our AIA necessitates access to in-game statistics, which are mainly retrievable through TW. However, **the existence of TW is not a strict requirement**. In fact, TW elaborate statistics and replays by directly interacting with the VG itself—because it is the

Table 7: Overview of E-Sports VG. Numbers are taken from various sources [17, 20, 32, 52, 59].

	Release Year	*Genre*	*Monthly Players*	*Concurrent Players Avg*	*Playtime Avg (Hours)*	*Age Range (PEGI rec.)*	*Tournament Revenue*	*Exemplary TW*	*Replay System*	*Max Players per Lobby*
League of Legends	2009	MOBA	127 M	700 K	832 H	11–50 (12+)	$93 M	lolprofile.net	Yes	10
CS:GO	2012	FPS	34 M	560 K	611H	13–40 (18+)	$134 M	csgostats.gg	Yes	18
Rocket League	2016	Sport	90 M	25 K	315 H	6–35 (3+)	$18 M	rltracker.pro	Yes	8
Fortnite	2017	Battle Royale	270 M	4 M	1800 H	6–54 (12+)	$121 M	fortnitetracker.com	Yes	100
PUBG	2018	Battle Royale	510 M	200 K	356 H	12–55 (16+)	$45 M	pubg.op.gg	Yes	100
Apex Legends	2019	Battle Royale	118 M	195 K	91 H	8–37 (16+)	$10 M	apex.tracker.gg	No	60
Dota2	2013	MOBA	15 M	450 K	1700 H	12–50 (12+)	$283 M	opendota.com	Yes	10

VG that makes such data publicly available. Therefore, an attacker could harvest these information and elaborate them autonomously. Obviously, the amount of effort required in this scenario is much higher than relying on a TW, but an AIA would still be feasible (especially for targeted 'one-to-one' AIA).

Let us summarize the panorama of other E-Sports VG, for which we provide an overview in Table 7. All these VG present at least one TW akin to those of Dota2 TW. Moreover, for all these VG, the in-game details of a player are public *by default* (except for Dota2 and CS:GO), and they often have replay system which could relax the requirement of a TW. We remark, however, that the other requirement for a successful AIA is the existence of a relationship between players' in-game statistics and personal attributes. Although there is no proof (yet) of the existence of such relationship in other contexts, we believe in its existence. In fact, many Dota2 features can be found in the other VG. Examples are the kill/death/assist ratio, paid subscription plans and cosmetics, chat usage, or information about the play-time. Finally, we highlight that players' of some VG (e.g., Fortnite) are children, increasing the risk of AIA [25, 47].

8 CONCLUSION

We addresses the problem of Attribute Inference Attack (AIA) in competitive video-games (VG), with a focus on Dota2. We observe that Dota2 players are naturally exposed to AIA due to the abundant in-game statistics that are publicly available. Based on this observation, we propose a threat model of AIA in Dota2, and (ethically) evaluate its impact. Our results demonstrate that with little preparation and domain expertise, attackers can predict the personal attributes of Dota2 players with high success (e.g., near-perfect precision). Countermeasures to such AIA are unfeasible due to tradeoffs that would disrupt the entire Dota2 ecosystem.

By elucidating this subtle threat, which can affect also players of other VG, this work will hopefully inspire the development of effective mitigations (either by the VG producers, or by the TW administrators), therefore fostering an increased privacy of video gamers (who should be made aware of such risk).

Ethical Considerations

Our institutions do not require any formal IRB approval to carry out the experiments described herein. Nonetheless, our survey and corresponding evaluation are all performed by adhering to the guidelines of the Menlo report [8]. All interviewees were informed that their responses would be used for research purposes. Our questionnaire does not ask for sensitive data, or for private details such as name or address. We never released our dataset publicly (not even in anonymised form). All participants are also aware of the email address to contact should they be willing to have their entry removed from our dataset. Since our user-base is located in Europe, we also strictly complied with the GDPR, and all underage participants were located in countries which allowed their participation in research surveys without explicit parental consent [24]. For our AIA, we always infer the attributes that the participants of our survey willingly provided to us, hence there is no privacy violation. We do not attempt to infer personal attributes of players who did not participate in our survey (i.e., we do not collect in-game data of randomly chosen Dota2 players, and use such data to infer their private information). The attributes we infer are non-sensitive.

Acknowledgement. We thank the Hilti Corporation for funding, and the AISec PC for the constructive feedback.

REFERENCES

[1] [n.d.]. Dendi - Liquipedia Dota2 Wiki. https://liquipedia.net/dota2/Dendi.
[2] 2022. The International. https://liquipedia.net/dota2/The_International.
[3] Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turkish journal of emergency medicine* (2018).
[4] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Proc. AAAI Int. Conf. Web Social Media.*
[5] Giovanni Apruzzese, Mauro Andreolini, Luca Ferretti, Mirco Marchetti, and Michele Colajanni. 2021. Modeling realistic adversarial attacks against network intrusion detection systems. *ACM Digital Threats: Research and Practice* (2021).
[6] Giovanni Apruzzese et al. 2022. The Role of Machine Learning in Cybersecurity. *ACM Digital Threats: Research and Practice* (2022).
[7] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. 2022. Dos and don'ts of machine learning in computer security. In *USENIX Security.*
[8] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. 2012. The Menlo report. *IEEE Security & Privacy* (2012).
[9] Matthew Barr and Alicia C. Stewart. 2022. Playing Video Games During the COVID-19 Pandemic and Effects on Players' Well-Being. *Games & Culture* (2022).
[10] Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition* 84 (2018), 317–331.
[11] Sara Bunian, Alessandro Canossa, Randy Colvin, and Magy Seif El-Nasr. 2017. Modeling individual differences in game behavior using HMM. In *Artif. Intell. Interact. Digit. Entert. Conf.*
[12] Chadlantis. 2019. How to Improve in Any Video Game. https://medium.com/@chadlantistv/how-to-improve-in-any-video-game-7d0efe5ed053.
[13] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.
[14] Terence Chen, Roksana Boreli, Mohamed-Ali Kaafar, and Arik Friedman. 2014. On the effectiveness of obfuscation techniques in online social networks. In *Int. Priv. Enhancing Techn. Symp.*
[15] Yuan Cheng, Jaehong Park, and Ravi Sandhu. 2013. Preserving user privacy from third-party applications in online social networks. In *Int. Conf. World Wide Web.*
[16] European Commission. [n.d.]. Sensitive Data. https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data_en.
[17] Mauro Conti and Pier Paolo Tricomi. 2020. PvP: Profiling Versus Player! Exploiting Gaming Data for Player Recognition. In *Int. Conf. Inf. Secur.*

[18] Christina Cough. 2020. Share of gamers who want to become professional gamers in the future worldwide in 2020, by gender. https://www.statista.com/statistics/1132968/professionals-gamers-gender/. Accessed: June, 2022.

[19] Anders Drachen, Christian Thurau, Rafet Sifa, and Christian Bauckhage. 2014. A comparison of methods for player clustering via behavioral telemetry. *arXiv preprint arXiv:1407.3950* (2014).

[20] Esport Earnings. 2022. Top Games Awarding Prize Money. https://www.esportsearnings.com/games. Accessed: July, 2022.

[21] Sanaz Eidizadehakhcheloo, Bizhan Alipour Pijani, Abdessamad Imine, and Michaël Rusinowitch. 2021. Divide-and-Learn: A Random Indexing Approach to Attribute Inference Attacks in Online Social Networks. In *IFIP Ann. Conf. Data Appl. Secur. Privacy*.

[22] Into The Breach Esports. 2021. r/DotA2 Demographic Survey. https://www.docdroid.net/ZeJTLar/rdota2-demographics-report-2021-pdf. Accessed: June, 2022.

[23] Quan Fang, Jitao Sang, Changsheng Xu, and M Shamim Hossain. [n.d.]. Relational user attribute inference in social media. *IEEE T. Multimedia* ([n. d.]).

[24] European Union Agency for Fundamental Rights. 2014. Child participation in research. https://fra.europa.eu/en/publication/2019/child-participation-research.

[25] Meg Fryling, Jami Lynn Cotler, Jack Rivituso, Lauren Mathews, and Shauna Pratico. 2015. Cyberbullying or normal game play? Impact of age, gender, and experience on cyberbullying in multi-player online gaming environments: Perceptions from one gaming forum. *J. Inf. Syst. Appl. Res.* (2015).

[26] Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. In *CHI—Human Factors in Computing Systems*. 253–262.

[27] Neil Zhenqiang Gong and Bin Liu. 2016. You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. In *25th USENIX Security Symposium (USENIX Security 16)*. 979–995.

[28] Neil Zhenqiang Gong and Bin Liu. 2018. Attribute inference attacks in online social networks. *ACM T. Privacy Secur.* (2018).

[29] Mark D Griffiths. 2017. The psychosocial impact of professional gambling, professional video gaming & eSports. *Casino & Gaming International* (2017).

[30] Wenbo Guo, Xian Wu, Sui Huang, and Xinyu Xing. 2021. Adversarial policy learning in two-player competitive games. In *Int. Conf. Machin. Learn.*

[31] Juho Hamari and Max Sjöblom. 2017. What is eSports and why do people watch it? *Internet research* (2017).

[32] howlongis.io. 2022. Dota 2 Playtime. https://howlongis.io/app/570/Dota+2.

[33] Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. 2015. Face/off: Preventing privacy leakage from photos in social networks. In *Proc. ACM CCS*.

[34] Ismat Jarin and Birhanu Eshete. 2021. Pricure: privacy-preserving collaborative inference in a multi-party setting. In *Proc. ACM Workshop Secur. Privacy Analytics*.

[35] David Jurgens, Tyler Finethy, James McCorriston, Yi Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proc. Int. AAAI Conf. Web Social Media*.

[36] Panicos Karkallis, Jorge Blasco, Guillermo Suarez-Tangil, and Sergio Pastrana. 2021. Detecting video-game injectors exchanged in game cheating communities. In *Europ. Symp. Res. Comp. Secur.*

[37] Mehdi Kaytoue, Arlei Silva, Loïc Cerf, Wagner Meira Jr, and Chedy Raïssi. 2012. Watch me playing, i am a professional: a first study on video game live streaming. In *Proc. Int. Conf. World Wide Web*.

[38] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proc. Nat. Academy Sciences* (2013).

[39] J. Kotrlik and C. Higgins. 2001. Organizational research: Determining appropriate sample size in survey research. *Inf. Tech. Learn. Perf. J.* (2001).

[40] Peter Likarish, Oliver Brdiczka, Nicholas Yee, Nicholas Ducheneaut, and Les Nelson. 2011. Demographic Profiling from MMOG Gameplay. In *11th Privacy Enhancing Technologies Symposium. Waterloo, Canada*. Citeseer.

[41] Dragana Martinovic, Victor Ralevich, Joshua McDougall, and Michael Perklin. 2014. "You are what you play": Breaching privacy and identifying users in online gaming. In *Proc. IEEE Ann. Int. Conf. Priv. Secur. Trust.*

[42] Shagufta Mehnaz et al. 2022. Are your sensitive attributes private? Novel model inversion attribute inference attacks on classification models. In *USENIX Security*.

[43] Bo Mei, Yinhao Xiao, Ruinian Li, Hong Li, Xiuzhen Cheng, and Yunchuan Sun. [n.d.]. Image and attribute based convolutional neural network inference attacks in social networks. *IEEE T. Netw. Sci. Eng.* ([n. d.]).

[44] Gregory J Meyer, Stephen E Finn, Lorraine D Eyde, Gary G Kay, Kevin L Moreland, Robert R Dies, Elena J Eisman, Tom W Kubiszyn, and Geoffrey M Reed. 2001. Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist* 56, 2 (2001), 128.

[45] Hooman Mohajeri Moghaddam, Gunes Acar, Ben Burgess, Arunesh Mathur, Danny Yuxing Huang, Nick Feamster, Edward W Felten, Prateek Mittal, and Arvind Narayanan. 2019. Watching you watch: The tracking ecosystem of over-the-top tv streaming devices. In *Proc. ACM Conf. Comp. Commun. Secur.*

[46] Joshua Morris, Sara Newman, Kannappan Palaniappan, Jianping Fan, and Dan Lin. 2021. "Do you know you are tracked by photos that you didnt take: large-scale location-aware multi-party image privacy protection. *IEEE TDSC* (2021).

[47] BBC News. 2019. Fortnite predator 'groomed children on voice chat'. https://www.bbc.com/news/technology-46923789. Accessed: June 2022.

[48] Kristine L Nowak and Christian Rauh. 2005. The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. *Journal of Computer-Mediated Communication* 11, 1 (2005), 153–178.

[49] US Department of the Treasury. [n.d.]. Sensitive Personal Data. https://home.treasury.gov/taxonomy/term/7651.

[50] Jean Oggins and Jeffrey Sammis. 2012. Notions of video game addiction and their relation to self-reported addiction among players of World of Warcraft. *International Journal of Mental Health and Addiction* 10, 2 (2012), 210–230.

[51] Bizhan Alipour Pijani, Abdessamad Imine, and Michaël Rusinowitch. 2020. You are what emojis say about your pictures: language-independent gender inference attack on Facebook. In *Proc. ACM Symp. Appl. Comp.*

[52] Active Player. 2022. Live Player Count and Statistics. https://activeplayer.io/.

[53] Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality* 41, 1 (2007), 203–212.

[54] Olivia Richman. 2020. Hashinshin responds to accusations of grooming a minor. https://win.gg/news/hashinshin-responds-to-accusations-of-grooming-a-minor/. Accessed: June 2022.

[55] Bruce Schneier. 2015. *Data and Goliath: The hidden battles to collect your data and control your world*. WW Norton & Company.

[56] Rafet Sifa, Anders Drachen, and Christian Bauckhage. 2018. Profiling in games: Understanding behavior from telemetry. *Social interactions in virtual worlds: An interdisciplinary perspective* (2018).

[57] Jonathan M Spring, Tyler Moore, and David Pym. 2017. Practicing a science of security: a philosophy of science perspective. In *New Secur. Paradig. Workshop*.

[58] Pieter Spronck, Iris Balemans, and Giel Van Lankveld. 2012. Player profiling with fallout 3. In *Artif. Intell. Interactive Dig. Entertainment Conf.*

[59] Steam. 2022. An ongoing analysis of Steam's concurrent players. https://steamcharts.com/. Accessed: July, 2022.

[60] Stratz. 2022. Accounts and matches analyzed by STRATZ. https://stratz.com/welcome. Accessed: July, 2022.

[61] Carl Symborski, Gary M Jackson, Meg Barton, Geoffrey Cranmer, Byron Raines, and Mary Magee Quinn. 2014. The use of social science methods to predict player characteristics from avatar observations. In *Predicting real world behaviors from virtual world data*. Springer, 19–37.

[62] Anne Clara Tally, Yu Ra Kim, Katreen Boustani, and Christena Nippert-Eng. 2021. Protect and Project: Names, Privacy, and the Boundary Negotiations of Online Video Game Players. *Proc. ACM Human-Comp. Inter.* (2021).

[63] Shoshannah Tekofsky, Jaap Van Den Herik, Pieter Spronck, and Aske Plaat. 2013. Psyops: Personality assessment through gaming behavior. In *In Proceedings of the International Conference on the Foundations of Digital Games*. Citeseer.

[64] Udi Weinsberg, Smriti Bhagat, Stratis Ioannidis, and Nina Taft. 2012. BlurMe: Inferring and obfuscating user gender based on ratings. In *Proceedings of the sixth ACM conference on Recommender systems*. 195–202.

[65] Jerry S Wiggins. 1996. *The five-factor model of personality: Theoretical perspectives*. Guilford Press.

[66] Tom Wijman. 2021. The Games Market and Beyond in 2021: The Year in Numbers. https://newzoo.com/insights/articles/the-games-market-in-2021-the-year-in-numbers-esports-cloud-gaming. Accessed: June 2022.

[67] Dennis L Wilson. 1972. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics* 3 (1972), 408–421.

[68] Kelce S Wilson and Müge Ayse Kiy. 2014. Some fundamental cybersecurity concepts. *IEEE access* (2014).

[69] Take Yo and Kazutoshi Sasahara. 2017. Inference of personal attributes from tweets using machine learning. In *IEEE Int. Conf. Big Data*.

[70] Shijie Zhang, Hongzhi Yin, Tong Chen, Zi Huang, Lizhen Cui, and Xiangliang Zhang. 2021. Graph embedding for recommendation against attribute inference attacks. In *Proc. Web Conf.*

[71] Yifei Zhang, Neng Gao, and Junsha Chen. 2020. A practical defense against attribute inference attacks in session-based recommendations. In *IEEE Int. Conf. Web Serv.*

[72] Da Zhong, Haipei Sun, Jun Xu, Neil Gong, and Wendy Hui Wang. 2022. Understanding Disparate Effects of Membership Inference Attacks and their Countermeasures. In *Proc. ACM AsiaCCS*.

A EXTRACTION OF CHAT FEATURES

We explain how we used the chat of Dota2 for our AIA. More information is available in our repository.

Motivation. Analyzing chat messages can reveal substantial information on a player. For instance, younger players may use more slang (age). Provocative messages can relate to nervous (neuroticism) or energetic (extraversion) players. Friendly (agreeableness) players use good-behaviour messages. Efficient (conscientiousness) players could use more tactics message, and openness could relate to messages sent at the start of a match. The gender could be affected by several types of custom messages. Finally, some hero messages must be purchased, therefore relating to occupation and purchase_habits.

Context. During a match, two chat channels exist simultaneously: a *team* chat, reserved for each team; and a *global* chat, visible to all players. Moreover, players have the possibility to setup two *chat-wheels*[17] by choosing from a set of pre-defined messages—whose purpose is to facilitate sending of commonly used messages. In particular, each player has a *general* chat-wheel (which is the same for all matches) and a *hero-specific* chat wheel (which is fixed for each hero). Both Dota2 and TW make public all messages sent in the *global* chat, as well as all those sent with the chat-wheel (even if they are sent in the *team* chat). Therefore, we use our domain expertise to extract meaningful features from both types of chat.

Global chat. We gathered lists of common English words (English is the default language for Dota2 jargon) denoting laughs, gaming/Dota2/online slang, bad/good behavior, and provocative messages. To create such lists, we explored websites (e.g., Dota2 forums[18], urban dictionary), manually inspected thousands of match chats, and leveraged our Dota2 expertise. Next, we counted the occurrences of such words in the player's messages. We also searched for messages containing only '?' (in Dota2 is highly provocative), counted the number of '?', '!', and capital letters (they express astonishment or anger), the number of early-game messages (usually sent to make noise or interact with the other team), and after-kill messages (used to complain, provoke, taunt).

Chat Wheels. Messages from the chat-wheel allow to distinguish if a player is communicating in the *global* (which is public) or in the *team* chat (which is not publicly available). For example, a 'laugh' can be sent either in the *global* or in the *team* chat: such difference is captured in some of our chat-wheel features. Nevertheless, such features entail tactical, laughs, deny, and good behavior messages. Moreover, we extracted which of them were 'sounds', or sprays left on the ground.

B ADDITIONAL CORRELATION ANALYSES

Let us expand our analysis in §4.3 with additional[19] evidence.

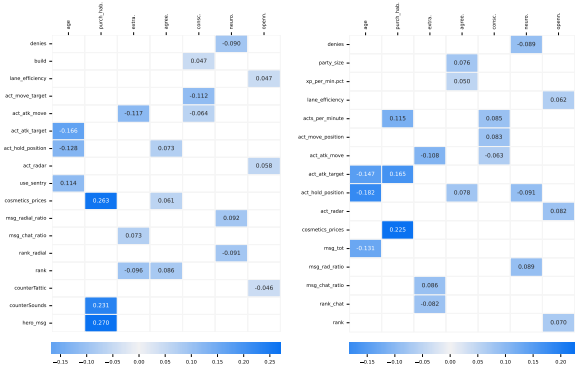
Given the high number of features that describes our datasets ($\mathcal{M}$, $\overline{\mathcal{M}}$, $\mathcal{P}$), we report in Table 8 the number of significant correlations at different p-values level, for both Cramer and Spearman indexes. From Table 8, we derive that many significant correlations exist for all our datasets, suggesting that ML models would be able to learn and infer private attributes starting from in-game statistics. Such a finding motivates our decision to consider AIA that use all our datasets (i.e., $\mathcal{P}$ in §5.1, $\mathcal{M}$ and $\overline{\mathcal{M}}$ in §5.2).

[17] More info here: https://dota2.fandom.com/wiki/Chat_Wheel
[18] For instance: https://dota2freaks.com/glossary/
[19] A thorough description of all our correlation analyses is provided in our repository.

Table 8: Significant Correlations at different p-values in our three datasets. Each column reports a personal attribute in $\mathcal{A}$. Rows denote how many features in each dataset (either $\mathcal{M}$, $\overline{\mathcal{M}}$ or $\mathcal{P}$) achieve p below the target α (i.e., the correlations are statistically significant).

Dataset	*Metric*	α	gend.	age	occ.	purch.	extr.	agree.	consc.	neur.	open.
$\mathcal{M}$	Cram.	<0.01	17	17	15	18	13	18	17	16	13
	Cram.	0.05	18	19	15	18	14	19	18	19	14
	Cram.	0.1	18	19	17	19	15	19	19	19	16
	Spear.	0.01	–	88	–	51	44	52	22	70	36
	Spear.	0.05	–	95	–	65	57	59	35	85	50
	Spear.	0.1	–	99	–	73	62	67	43	87	59
$\overline{\mathcal{M}}$	Cram.	<0.01	16	12	12	11	15	10	10	14	8
	Cram.	0.05	18	17	18	15	17	11	14	15	11
	Cram.	0.1	18	17	18	15	18	14	15	20	13
	Spear.	0.01	–	95	–	43	53	38	25	60	27
	Spear.	0.05	–	104	–	63	65	54	40	82	47
	Spear.	0.1	–	108	–	69	73	64	53	90	58
$\mathcal{P}$	Cram.	<0.01	2	1	2	1	0	0	0	1	0
	Cram.	0.05	3	3	3	1	0	0	1	1	0
	Cram.	0.1	4	3	3	1	0	0	1	2	1
	Spear.	0.01	–	69	–	11	13	2	0	2	0
	Spear.	0.05	–	97	–	16	27	13	8	22	4
	Spear.	0.1	–	110	–	26	47	26	16	44	14

In Figs. 7, we report the Top-3 Spearman's correlation between $\mathcal{A}$ and both $\mathcal{M}$ (Fig. 7a) and $\overline{\mathcal{M}}$ (Fig. 7b). We observe similar strengths (e.g., compare purchase_habits with cosmetics_prices). However, personality traits tend to have low strength ($\rho < 0.1$) for both of these datasets—suggesting that AIA may not be very successful at predicting such attributes.



(a) Correlations between $\mathcal{M}$ and $\mathcal{A}$. (b) Correlations between $\overline{\mathcal{M}}$ and $\mathcal{A}$.

Fig. 7: Top-3 Spearman significant correlation (p-value < 0.01)

C STATISTICAL VALIDATION

We now validate the results obtained by our AIA described in §5 and §6. Specifically, our goal is verifying whether our techniques achieve a performance that can be considered to be "statistically equivalent" to a given baseline. If such statement is found to be true, then it means that any performance difference is irrelevant; otherwise, it means that one method is better/worse than the other.

C.1 Methodology: two-sample Student t-test

We rely on a two-sample t-test, the result of which is a p-value which, if superior to a given target α, can be used to accept a given *null hypothesis*. Specifically, we set our target α=0.05, and we set

our null hypothesis as "the technique T_1 is equal to the technique T_2". Let us explain what T_1 and T_2 consist in by describing all of the statistical tests we perform.

Simple AIA (§5.1). We set T_1 to be the performance (F1-score) achieved by the 'Dummy' classifier (our baseline); whereas T_2 is the *best* ML model for each considered attribute (i.e., the bold values in Table 3). We hence consider the corresponding values (i.e., average and std. dev.) from Table 3, and the number of samples for both T_1 and T_2 is 10 (because we use stratified 10-fold cross-validation). We perform these tests 9 times—one for each attribute in Table 3.

One-match AIA (§5.2). Here, we perform two tests. First, we set T_1 to be the performance (F1-score) of the 'Dummy' classifier (our baseline), and T_2 is the performance of the 'Naive' attacker (leftmost column in Table 4). Then, we consider the same T_1, but consider T_2 to be the performance of the 'Expert' attacker (middle column in Table 4). The number of samples for all these T is 20 (because we repeat these experiments 20 times to account for the random sampling of $\overline{\mathcal{M}}$). We perform these tests 9 times—one for each attribute in Table 4.

Indiscriminate AIA (§6.1). Here, we consider T_1 to be the performance (accuracy) of the 'sophisticated AIA' (leftmost column in Table 6), whereas T_2 represents the performance of the 'indiscriminate AIA' (central column in Table 6). The number of samples for both T_1 and T_2 is 20 (because we perform the draw 20 times). We perform these tests 7 times—one for each attribute in Table 6.

C.2 Results

We report the results of all our tests in Table 9. Specifically, since we perform 34 comparisons in total, we report the amount of times that a given null hypothesis (in the mid-left column) must be rejected (i.e., when $p < \alpha$, mid-right column).

Table 9: Statistical Validation of our results. We report the amount of tests in which the null hypothesis must be rejected (because $p < \alpha$).

Table	Null Hypothesis (T_1=T_2)	# Reject	Total
Table 3	Dummy Classifier = Best Model	5	9
Table 4	Dummy = Naive Attacker	4	9
	Dummy = Expert Attacker	9	9
Table 6	Sophisticated = Indiscriminate	7	7

From Table 9, we can see that there are cases in which our null hypothesis must be accepted, i.e., a given technique is statistically equivalent to the corresponding baseline. Unsurprisingly, these cases entail the 'simple AIA' (§5.1) and the ablation study (§5.2).

- **Simple AIA.** There are 4 cases in which "Dummy Classifier = Best Model", corresponding to the attributes: purchase_habits, conscientiousness, extraversion, agreeableness. In these cases, our 'simple AIA' provide a negligible performance improvement over the baseline; whereas the improvement is statistically significant for the remaining 5 attributes.
- **Ablation Study.** There are 5 cases in which "Dummy Classifier = Naive Attacker", corresponding to the attributes: occupation, purchase_habits, openness, extraversion, agreeableness. In these cases, the Naive Attacker has the same effectiveness as a coin-toss; furthermore, such an attacker is even *worse* (statistically) than the Dummy classifier for conscientiousness and for neuroticism. The encouraging part of these results is that if an attacker could use only a single match (and is not knowledgeable about Dota2 to derive $\overline{\mathcal{M}}$), then their AIA would be not very effective.

In contrast to the above, however, our null hypothesis is *always rejected* for the "sophisticated AIA = indiscriminate AIA", and for the "Dummy = Expert Attacker", thereby showing that **our 'advanced' methods are always *statistically superior* to the corresponding baseline**—and by a huge margin ($p < 0.00001$) .