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**;

KEYWORDS

Attribute Inference Attack, Video Games, Dota2, Machine Learning

ACM Reference Format:
Pier Paolo Tricomi, Lisa Facciolo, Giovanni Apruzzese, and Mauro Conti. 2023. Attribute Inference Attacks in Online Multiplayer Video Games: a Case Study on Dota2. In *Proceedings of Thirteenth ACM Conference on Data and Application Security and Privacy (CODASPY '23).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3577923.3583653

CODASPY '23, April 24–26, 2023, Charlotte, NC, USA

ACM ISBN 979-8-4007-0067-5/23/04.
https://doi.org/10.1145/3577923.3583653

IMPORTANT. This document is meant to extend our main paper: the numbering of Figures and Tables will resume from the one in the main paper; some parts of this Appendix will refer to the sections (§), Tables and Figures of the main paper (using the corresponding numbering). GitHub page: https://github.com/hihey54/Dota2AIA

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*[1] 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[2], 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

[1]More info here: https://dota2.fandom.com/wiki/Chat_Wheel
[2]For instance: https://dota2freaks.com/glossary/

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[3] 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).

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).

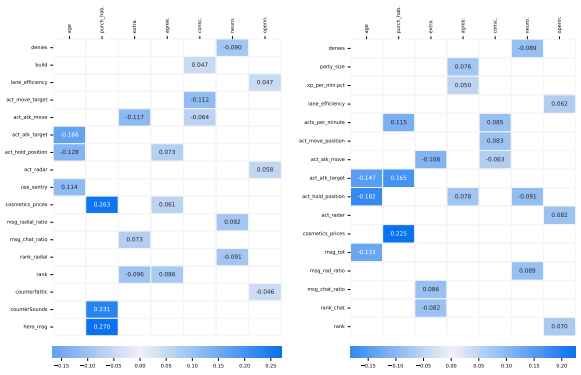
Dataset	*Metric*	α	gend.	age	occ.	purch.	extr.	agree.	consc.	neur.	open.
$\mathcal{M}$	Cram.	<0.01	17	17	15	18	13	18	17	16	13
	Cram.	0.05	18	19	15	18	14	19	18	19	14
	Cram.	0.1	18	19	17	19	15	19	19	19	16
	Spear.	0.01	–	88	–	51	44	52	22	70	36
	Spear.	0.05	–	95	–	65	57	59	35	85	50
	Spear.	0.1	–	99	–	73	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.

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.

[3] A thorough description of all our correlation analyses is provided in our repository.



(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.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).

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,

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

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 ($p < 0.00001$) .