Exposé for Bachelor Thesis

How well can federated learning be utilized to implement a decentralized and privacy focused bot detection system for websites?

submitted by

Matz-Jona Radloff

Matriculation number 6946325

Study Program: Computer Science

submitted on December 7, 2021

Supervisor: August See

First reviewer: Prof. Dr. Mathias Fischer

Zweitgutachter: N.N.

Abstract

Malicious use of automated bots present an increasing risk to applications in the web. Existing solutions do either not perform well, are not accessible to many providers due to high cost, or disregard modern privacy standards. This work aims to explore the state-of-art as well as provide a proof-of-concept for a basic system that incorporates all of the above criterions. These concerns will be addressed by implementing an open-source library that uses federated learning and its performance will be compared against existing solutions.

Contents

1 Motivation **4**

2 Method **5**
2.1 Concept 5
2.2 Datasets 5
2.3 Machine Learning Model 6
2.4 Evaluation 6

3 Related Work **7**
3.1 Proprietary Solutions 7
3.2 Scientific Work 7

4 Time plan **8**

Bibliography **9**

1 Motivation

In this work, the term bot is referring to software that is automatically performing HTTP(S) requests with the intent of harming the target or reaching another malicous goal. While this threat is nothing new to the web the attack surface has grown significantly over the past year. Especially the increased usage of web interfaces in poorly secured IoT devices and the trend to (re-)implement software as web applications is responsible for this.

The usage of bots can have several goals. DoS attacks aim to overload the target's infrastructure such that it becomes inaccessible for normal use. Carding and Credential stuffing refers to performing payment or login requests to find working credit card numbers and credentials usually obtainend from a data breach. Data scrapers download the website data and can use the data for malicious purposes, e.g. damage SEO or violate copyrights. Content spam includes inserting malicious or polluting data on platforms that allow user generated content. Scalping or inventory hoarding of shopping items can artificially raise prices, damage brands, generate false market forces and create a bad customer experience.

Recent studies show that of 2020's internet traffic 25.6% was fraudulent and automatically generated [Laba] [Labb]. They also show that both the percentage of bot traffic in general as well as malicious bot traffic has increased over time.

Most of the above attacks need to trick the webserver and application backends into performing the request as if it had been initiated by a human. Instead of combating the resulting issues separately, bot detection could potentially mitigate many at once.

A complication in this problem space is the, often desired, requirement for non-malicious bots to be granted normal access. The most prominent example are scraper bots from search engines that need to request websites periodically to build their search indices. A common technique to exploit this requirement is trying to emulate known bot signatures from large search engines, e.g. Googlebot [ADJ18].

Many website operators tend to use solutions that are easy to integrate and perform well. This requires embeding external software which collects user data and sends it to servers of the software vendors. These often do not disclose what exactly happens to the user data and website operators open themselves to additional threats in case of a data breach. Depending on the operating countries of both the websites and software vendors, data privacy regulation might also not allow sharing user data at all or require the operator to document the concrete data transfer in a very detailed and legally complicated way, e.g. in countries falling under the GDPR [Uni]. Because of the above reasons it is desirable to either employ self-hosted software or use a solution that does not require user data transfers.

A promising technique that combines both machine learning and respects modern privacy standards is federated learning [KMR15] [Kon+16].

2 Method

2.1 Concept

The thesis hypothesizes that machine learning systems incorporating federated learning perform better than non-federated counterparts while maintaining user privacy. It assumes that users perform actions on websites using a mouse or touchpad which are hard to fake by attackers trying to emulate human behavior. Raw mouse data is segmented into mouse actions such as mouse move (MM) or mouse move and a click (point and click, PC) similar to [AD19] and their previous paper [AE18]. Multiple results can be averaged to increase detection performance.

2.2 Datasets

Many publicly available datasets exist that contain valid user mouse movement and click data. The following datasets will be aggregated and labeled as human input. Shen et al.'s [SCG12] dataset contains mouse dynamics information from 28 users and 30 sessions per users which each contain around 3000 mouse events. The DFL dataset [AD19] includes 20-30 sessions of 21 users. The Balabit Mouse Dynamics Challenge Data Set [Bal] includes a few longer session and several shorter session which are meant to be used for training and testing respectively. For the purposes of bot detection, both can be used.

If needed, additional datasets are available, e.g.:

https://figshare.com/articles/dataset/Mouse_Behavior_Data_for_Static_Authentication/5619313

https://www.uvic.ca/ecs/ece/isot/datasets/?utm_medium=redirect&utm_source=/engineering/ece/isot/datasets/&utm_campaign=redirect-usage#section0-3

Data labeled as bot input will be generated using three methods. The first naive approach interpolates linearly between the start and end point of randomly generated movements. The second method adds noise and the third uses both B-spline interpolation and noise. More complicated simulations exist but their implementations are not publicly available. [Hu+17] [Naz03] The three methods used depict a reasonable choice of a basic attack that tries to evade detection by fake mouse movement.

A percentage of the aggregated data will be split and used as test data while the rest is used for training.

2.3 Machine Learning Model

Many different machine learning models are suitable for binary classification. Hu et al. [Hu+17] compare different classifiers in a similar context with Random Forest and Multilayer Perceptron performing the best. If time allows it, the thesis will compare multiple classification methods against each other and only use one otherwise.

The input features will consist of the (probably normalized) x- and y-coordinates as well as a time value for each mouse event. Additional features will be engineered similar to [AE18], such as mean, standard deviation, minimum and maximum value of path tangent, horizontal, vertical and overall velocity, acceleration, jerk, angular velocity. Additionally the type of action, length of the movement and time needed to complete the action will be used.

A system for pre-processing the different datasets will be developed.

The thesis will use Tensorflow with and without federated learning with a Random Forest keras model. Tensorflow is one of the most widely used machine learning frameworks and the integration of its runtime into distributed learners in the form of website backends or even client devices is feasible. The thesis will not include the actual integration into such systems. A simulated environment of multiple Tensorflow instances that have access to a secure communications channel will be developed.

2.4 Evaluation

The bot datasets will be split by their generation method such that different combinations of data are used per learning participant. The human datasets will also be split into subsets containing random samples from all or only specific datasets. The thesis will primarily compare the performance of the individual classifiers with and without federated learning with the hypothesis that the more learning results of different input data is combined, the better the overall performance.

3 Related Work

3.1 Proprietary Solutions

https://datadome.co/
https://www.perimeterx.com/products/bot-defender/
https://www.imperva.com/products/advanced-bot-protection-management/
https://www.fastly.com/products/cloud-security/bot-protection
https://www.cloudflare.com/products/bot-management/
https://developers.google.com/recaptcha/docs/v3
https://www.hcaptcha.com/

3.2 Scientific Work

The paper [Li+21] introduces a federated learning approach similar to the goals of this work but differs in the specific use case and implementation. Their system focuses on the detection of IoT (Internet of Things) devices which are easily hacked and turned into zombies. These zombies are commonly used in DDoS (Distributed Denial of Service) attacks which their strategy tries to make not feasible to perform. They also develop their own iterative model averaging based method "gated recurrent unit" (GRU) which is optimized for their specific use case.

Other works related to DDoS mitigation that also incorporate machine learning include Farivar *et al.* [Far+20], Liu *et al.* [LLW19] and Hussain *et al.* [Hus+21].

Many of the above solutions use the actual IP network traffic to extract relevant information to be used as input parameters for their models and are intended to be run in a server environment. In comparison clients' browser environments in the web offer a much greater amount and variety of user information that might be very useful to differentiate between a valid user and a malicious bot.

The work of [Pap+21] outlines the problems and privacy-realted concerns really well and tries to solve a very similar problem but focuses on mobile devices. The authors run a pre-trained machine learning model on the user's device. To avoid local changes to the model a cryptographic proofis generated that is verified on a server.

Among others, the works of Shen et al. [SCG12] and Antal et al. [AD19] [AE18] show the viability of using mouse and trackpad actions to verify the authenticity of users but privacy concerns often stand in the way of using such a method in practice.

4 Time plan

My planned steps in roughly two week intervals:

1. Start implementing data aggregator and feature extractor
2. Extract features of one dataset
3. Implement bot data generator (only naive type)
4. Build the Tensorflow model, verify that it works and that it can classify the data correctly
5. Integrate Federated learning and build the simulated environment of multiple distributed learners
6. Aggregate the remaining datasets and extend the bot data generation
7. Formulate and perform the experiments to test the thesis' hypothesis
8. Process and visualize the results

Bibliography

[AD19] Margit Antal and Lehel Denes-Fazakas. *User Verification Based on Mouse Dynamics: a Comparison of Public Data Sets*. In: *2019 IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. 2019, pp. 143–148. DOI: 10.1109/SACI46893.2019.9111596.

[ADJ18] Nilani Algiryage, Gihan Dias, and Sanath Jayasena. *Distinguishing Real Web Crawlers from Fakes: Googlebot Example*. In: *2018 Moratuwa Engineering Research Conference (MERCon)*. 2018, pp. 13–18. DOI: 10.1109/MERCon.2018.8421894.

[AE18] Margit Antal and Elod Egyed-Zsigmond. *Intrusion Detection Using Mouse Dynamics*. In: *CoRR* abs/1810.04668 (2018). arXiv: 1810.04668. URL: http://arxiv.org/abs/1810.04668.

[Bal] Balabit. *Releasing the Balabit Mouse Dynamics Challenge Data Set*. URL: https://medium.com/balabit-unsupervised/releasing-the-balabit-mouse-dynamics-challenge-data-set-a15a016fba6c (visited on 12/02/2021).

[Far+20] Faezeh Farivar et al. *Artificial Intelligence for Detection, Estimation, and Compensation of Malicious Attacks in Nonlinear Cyber-Physical Systems and Industrial IoT*. In: *IEEE Transactions on Industrial Informatics* 16.4 (2020), pp. 2716–2725. DOI: 10.1109/TII.2019.2956474.

[Hu+17] Shujie Hu et al. *Deceive Mouse-Dynamics-Based Authentication Model via Movement Simulation*. In: *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*. Vol. 1. 2017, pp. 482–485. DOI: 10.1109/ISCID.2017.134.

[Hus+21] Bilal Hussain et al. *Deep Learning-Based DDoS-Attack Detection for Cyber–Physical System Over 5G Network*. In: *IEEE Transactions on Industrial Informatics* 17.2 (2021), pp. 860–870. DOI: 10.1109/TII.2020.2974520.

[KMR15] Jakub Konečný, Brendan McMahan, and Daniel Ramage. *Federated Optimization: Distributed Optimization Beyond the Datacenter*. In: *CoRR* abs/1511.03575 (2015). arXiv: 1511.03575. URL: http://arxiv.org/abs/1511.03575.

[Kon+16] Jakub Konečný et al. *Federated Optimization: Distributed Machine Learning for On-Device Intelligence*. In: *CoRR* abs/1610.02527 (2016). arXiv: 1610.02527. URL: http://arxiv.org/abs/1610.02527.

[Laba] Imperva Threat Research Lab. *Bad Bot Report 2020: Bad Bots Strike Back*. URL: https://www.imperva.com/blog/bad-bot-report-2020-bad-bots-strike-back/ (visited on 11/12/2021).

[Labb] Imperva Threat Research Lab. *Bad Bot Report 2021: The Pandemic of the Internet*. URL: https://www.imperva.com/resources/resource-library/reports/bad-bot-report/ (visited on 11/12/2021).

[Li+21] Jianhua Li et al. *FLEAM: A Federated Learning Empowered Architecture to Mitigate DDoS in Industrial IoT*. In: *IEEE Transactions on Industrial Informatics* (2021), pp. 1–1. DOI: 10.1109/TII.2021.3088938.

[LLW19] Chi Harold Liu, Qiuxia Lin, and Shilin Wen. *Blockchain-Enabled Data Collection and Sharing for Industrial IoT With Deep Reinforcement Learning*. In: *IEEE Transactions on Industrial Informatics* 15.6 (2019), pp. 3516–3526. DOI: 10.1109/TII.2018.2890203.

[Naz03] Akif Nazar. *Synthesis & Simulation of Mouse Dynamics*. In: (2003). URL: https://dspace.library.uvic.ca/bitstream/handle/1828/308/Thesis-v19.pdf?sequence=1&isAllowed=y.

[Pap+21] Panagiotis Papadopoulos et al. *ZKSENSE: a Privacy-Preserving Mechanism for Bot Detection in Mobile Devices*. In: *Proceedings on Privacy Enhancing Technologies*. On the Internet: Privacy Enhancing Technologies Symposium, July 2021, pp. 1–23.

[SCG12] Chao Shen, Zhongmin Cai, and Xiaohong Guan. *Continuous authentication for mouse dynamics: A pattern-growth approach*. In: *IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2012)*. 2012, pp. 1–12. DOI: 10.1109/DSN.2012.6263955.

[Uni] European Union. *General Data Protection Regulation*. URL: %7Bhttps://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679%7D (visited on 11/12/2021).