

Bot Detection Using Mouse Movements

N. S. Afanaseva
Omsk State Transport University
OSTU
Omsk, Russia
dikarevich.ns@gmail.com

P. S. Lozhnikov
Omsk State Technical University
OmSTU
Omsk, Russia
lozhnikov@gmail.com

Abstract— Researching bot attacks and creating possible ways to combat them is a new direction of research in the information security. Bot detection can become an important part of trusted interaction technology, as the relevance of such attacks is constantly increasing. Bots automate processes, simplify interaction with various services and help the user in solving various tasks. However, among the many useful bots, there are also bad bots that negatively impact users and organizations.

This article discusses the advantages of detecting bots using mouse dynamics. An analysis of existing methods for identifying bots is carried out. Publicly available datasets that can be used to identify bots through mouse cursor movement patterns are reviewed.

Keywords— bot; malicious bot; trusted interactions; bot detector; mouse movement; mouse dynamics datasets

I. INTRODUCTION

According to the latest Imperva report, 47.4% of all Internet traffic in 2022 was from Internet bots, which was 5.1% more than in 2021 [1]. Fighting bots is considered to be a cross-industry and cross-functional problem. The ability of bots to perform various actions at a speed and frequency that is unattainable to the average user makes them an ideal tool for malicious activities and attacks.

The possible negative consequences that result from the use of bots are the following:

- unauthorized access to personal data of users;
- spreading misinformation and influencing public opinion through manipulation of information;
- misleading the user by simulating the behavior of a real interlocutor;
- economic damage resulting from distortion of web analytics indicators and assessments of the effectiveness of marketing research;
- increasing the share of speculative trading;

- improving the position of competitors by manipulating network resources or damaging the reputation of other organizations.

These examples show that all services and industries, affected by active digitalization such as “electronic government”, social networks, retail, fintech etc. are a possible target group for the use of bots. The safe functioning of all these spheres is impossible without the formation of an environment of trust, which proves the relevance of bot identification issues. In accordance with changing national requirements bot detection might become an important part of trusted interaction technology, representing an updated set of technologies and methods, that ensure a given level of trusted interaction between subjects of information exchange, implemented in the form of software and hardware-software solutions and ensuring the correct operation of applications and application services in an untrusted environment using modern methods of modeling and assessing instantaneous values of trust levels [2].

Various approaches have been proposed for detecting bots. CAPTCHA (Completely Automated Public Turing Test for Distinguishing Between Computers and Humans) [3], being the most famous Human Identification Test (HIP), is used in most web resources for active detection of bots. CAPTCHA offers the task that is easy to solve for humans but difficult for bots. However, the use of CAPTCHA is currently not an absolute guarantee of tracking bots, since they have learned to bypass this test either with the help of a person or using image recognition methods (OCR). Moreover, with the development of deep learning, it becomes possible to crack various types of CAPTCHAs [4]. Another approach to detecting bots is to use passive monitoring of input data and compare it with human performance (HOP) [5]. Its advantage is the inherent irregularity and complexity that distinguish a human from a bot. The dynamic characteristics of the mouse movement, the use of which is classified as the second group of detection methods, demonstrate significant advantages over known bot detection methods. First, they can provide continuous detection throughout their operation. Secondly, this method does not require the user to perform additional actions, and the

detection can be performed without the user being aware of the work being performed.

II. BOT DETECTION

Dynamic biometric images of a person have become widely used not only in the user authentication, but also in the field of the bot detection. These kinds of images include voice, keyboard handwriting, gestural features, dynamic characteristics of mouse movement etc.

Mouse movement can be viewed as a series of movements (tracks), where each track represents a specific and continuous physical process initiated and completed by the user [6].

Mouse movement analysis is widely used in modern research to detect malicious bots. Malicious bots are autonomous intelligent products designed to defraud, cause harm to end users or organizations as a whole [7]. For the bot recognition, the following advantages of working with the dynamic characteristics of mouse movement are important:

- the prevalence of this device and widespread use in the "human-computer" system;
- mouse movement analysis does not require the use of additional hardware, which saves resources;
- cursor movement is tracked in an unobtrusive and invisible way to the end user [5];
- ease of collecting user behavioral data on a large scale;
- from a user privacy perspective, sharing mouse dynamics data is much less problematic than sharing signatures or voice data.

Currently, there are several bot detectors that have their own characteristics and are aimed at solving specific problems.

The bot detector BeCAPTCHA-Mouse, introduced in the study [8], analyzes the full mouse movement trajectory rather than individual characteristics (track duration, average speed, cursor position changes). BeCAPTCHA-Mouse is trained on data obtained through neuromotor modeling of mouse dynamics. This type of detector does not advocate for the abandonment of the existing reCAPTCHA tool but serves as a complement, utilizing additional information about mouse movement dynamics. In this work, two methods for creating synthetic mouse trajectories are presented to enhance the training and evaluation of bot detection methods: the first is based on heuristic functions, and the second is based on generative adversarial networks, which synthesize trajectories based on input Gaussian noise.

Rahman R. U. and Tomar D. S. [9] proposed ten basic characteristics such as input source, degree of mouse click pressure, horizontal scroll amount, vertical scroll amount, horizontal scroll speed, vertical scroll speed and entropy of time spent moving between two pages, as well as the time on page and typing speed to analyze the behavior of bots in web applications. The developed model is based on two machine

learning algorithms which are K-medoid algorithm and naive Bayes one.

Chong P., Elovici Y., Binder A [10] pioneered the use of deep learning to detect bots based on mouse movement characteristics. They proposed to apply two-dimensional convolutional neural networks to the dynamic characteristics of mouse movement by representing time series of sequential tracks. In their work, they used the difference in cursor positions (dx , dy) and the speed values between two adjacent track points (dx/dt , dy/dt). This approach has proven its effectiveness and was further developed in the research [11] of Antal M. and Fejér N.

When detecting mouse-based bot activity using sequential learning [12], mouse movement is represented as a numerical vector (dx , dy , dx/dt , dy/dt), upon which deep learning is employed for bot detection. This study utilizes deep learning models with long-term memory (LSTM), a one-dimensional convolutional neural network model (1D-CNN), and a hybrid deep learning model combining a convolutional neural network with long-term memory (CNN+LSTM). A similar approach is used in reference [13]. However, in this case, the input data for the 1D-CNN consists of images obtained through a two-step transformation of mouse movement parameters. In the first step, spatial characteristics such as the distance between two adjacent points and the mouse's trajectory are mapped to an image. Then, color information is used to represent the kinematic information of each point.

The detection of bots replicating sessions of real users is continues to be a separate issue. There are scripts that not only capture the overall user behavior statistics but also record complete sessions of individual users on a page [14]. Counteracting such bots is discussed in a number of works [8,15,16] The paper[16] considers user authentication rather than bot identification. The research [15] deals with blog bots, and the study [8] takes into account the presence of websites where the generation of repetitive sessions by real users, such as news and banking websites, is possible.

III. MOUSE DYNAMICS DATASETS

Two types of data analysis sets are used in the above-mentioned works. The first one is public datasets, which are posted in the public domain and can be used by anyone; the second one is self-collected data. Moreover, the second type of datasets can be created in two ways: users perform a specific task using the mouse (managed environment method), or users are not instructed and they independently move the mouse cursor (unmanaged environment method).

The most commonly used datasets include Balabit [17], Bogazici [18], Attentive Cursor [19], SapiMouse [20], Chao Shen [21] and DFL [22], ReMouse [8]. A brief description of each dataset is given below.

The Balabit dataset [17], published in 2016, is classified as an unsupervised dataset and includes cursor position and track time information for 10 users connected to a remote server. During data collection, the users were asked to carry out their normal daily activities. Mouse events contain the following

data: timestamp, button pressed, mouse state, and mouse pointer coordinates. The main purpose of collecting the Balabit dataset was to find out how the engaged users use their mouse in order to be able to protect them from unauthorized use of their accounts. The training and testing data are represented as sessions in the dataset. However, test sessions are much shorter than training sessions.

The Bogazici dataset [18], published in 2021, also falls into the category of unsupervised datasets and includes the mouse usage behavior patterns of 24 users collected within a month. Data collection participants were selected from different positions within the same company to obtain different patterns of user behavior when interacting with different programs and tools in an office environment. Each user's computer was loaded with a specially designed program that collected the user's mouse movements without being tied to a specific task or interfering with the user's normal daily activities. The data set consists of the mouse action type, timestamp, spatial coordinates, state, and application window name.

In the controlled environment dataset of 2020 [19], users were presented with a web search task aimed at determining areas of attention and demographic information. The authors recorded the real-life behavior of approximately 3,000 people performing a transactional Web search task. The collected information includes the following: mouse cursor position, timestamp, event name.

The dataset [20], collected in 2020, was also obtained through a controlled method. It comprises data on mouse movement from 120 users (92 males and 28 females, ranging from 18 to 53 years old). The participants were asked to perform four different actions, each involving geometric shapes on a web page, including left and right mouse clicks, as well as dragging actions. For each participant, two files were associated, corresponding to one- and three-minute sessions, respectively. The dataset includes the following information: mouse cursor position, event type (movement, dragging, click, or release), and corresponding timestamp.

The Chao Shen dataset consists of mouse dynamics information from 28 users, collected over a period of two months [21]. Each session includes approximately thirty minutes of mouse activity for each user. In the dataset, each mouse operation was represented as features including action type, application type, screen area, window position, and their respective timestamps. The dataset was collected for the purpose of continuous user authentication.

The DFL dataset [22] was collected in 2018 in an uncontrolled environment and includes the following information about user mouse actions: timestamp, button (left, right, none), state (movement, click, release, drag), and coordinates.

The ReMouse dataset, collected in 2023, contains information on mouse dynamics from 100 users residing in different countries and using various devices. One distinctive feature of this dataset is the presence of repeated sessions, which were obtained as users performed identical tasks, which is a repetitive sequence of steps. This dataset includes

information about mouse cursor positions, cursor movement speed, and application window size.

IV. DISCUSSION OF RESULTS

The existing research in the field of bot detection confirms the relevance of the issue. It is worth noting that the current methods and datasets have specific characteristics and limitations in their application. For specialized tasks, there is a need to create additional datasets that include mouse movement dynamics. The most suitable datasets for our further study are Balabit and Bogazici.

V. CONCLUSION

The active use of bots creates serious threats to the security of users' personal data, public opinion, the economy and the competitiveness of various services and organizations. Comprehensive measures are needed to combat bots at cross-industry and cross-functional levels. This is the only way to ensure the security and stability of online services in the context of the active use of bots. Successfully combating bots will not only reduce the negative consequences of their activities, but will also increase the level of user trust in the online environment as a whole, which is one of the tasks of trusted interaction, namely the development of technologies that ensure trusted interaction of elements of the digital environment in open systems.

Modern technologies and methods of protection against bots must be improved and adapted to effectively combat threats. This study represents an in-depth analysis of existing approaches for bot detection. A list of publicly available datasets on mouse dynamics, which were used for bot detection, is also provided. This information can be used to make an informed choice of a dataset for further research on bot detectors.

ACKNOWLEDGMENT

This research was carried out with the financial support of Ministry of Digital Development, Communications and Mass Communications of the Russian Federation (Ministry of Digital Development of Russia), agreement No. 40469-07/23-K dated June 30, 2023.

REFERENCES

- [1] 2022 Imperva Bad Bot Report [Electronic source]. URL: <https://www.imperva.com/resources/resource-library/reports/bad-bot-report/> (reference date: 01.09.2023)
- [2] Competence Center of the National Technology Initiative "Trusted Interaction Technologies" [Electronic source]. URL: https://nti2035.ru/technology/competence_centers/tdv.php (reference date: 01.09.2023).
- [3] What is recaptcha? [Electronic source]. URL: <https://www.google.com/recaptcha/about/> (reference date: 03.09.2023).
- [4] Stark, F., Hazirbas, C., Triebel, R. and Cremers, D. (2015, October). "Captcha recognition with active deep learning." In Workshop new challenges in neural computation (Vol. 2015, p. 94).
- [5] Leiva, L. A., Arapakis, I., & Iordanou, C. (2021, March). "My mouse, my rules: Privacy issues of behavioral user profiling via mouse tracking.

- " In Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (pp. 51-61).
- [6] Katerina, T., & Nicolaos, P. (2018). "Mouse behavioral patterns and keystroke dynamics in End-User Development: What can they tell us about users' behavioral attributes?". *Computers in Human Behavior*, 83, pp. 288-305.
- [7] Afanasyeva, N. S. "Malicious Bots in the Modern World: Analysis, Consequences, and Possible Countermeasures" / N. S. Afanasyeva, D. A. Elizarov, P. S. Lozhnikov // *Information Security in the Digital Economy: Proceedings of the XIX Scientific and Practical Conference (within the framework of the X Plenum of the Regional Division of the Federal Educational and Methodological Association in the System of Higher Education for the Aggregated Group of Specialties and Training Directions 10.00.00 "Information Security" for the Siberian and Far Eastern Federal Districts (SibROUMO))*, Ulan-Ude, June 07-11, 2023. – Novosibirsk: Siberian State University of Telecommunications and Informatics, 2023. – pp. 84-92.
- [8] Sadeghpour, Shadi, and Natalija Vljajic. "ReMouse Dataset: On the Efficacy of Measuring the Similarity of Human-Generated Trajectories for the Detection of Session-Replay Bots." *Journal of Cybersecurity and Privacy* 3.1 (2023): pp.95-117.
- [9] Rahman, Rizwan Ur, and Deepak Singh Tomar. "A new web forensic framework for bot crime investigation." *Forensic Science International: Digital Investigation* 33 (2020): 300943.
- [10] Chong, Penny, Yuval Elovici, and Alexander Binder. "User authentication based on mouse dynamics using deep neural networks: A comprehensive study." *IEEE Transactions on Information Forensics and Security* 15 (2019): pp.1086-1101.
- [11] Antal, M.; Fejér, N.; Buza, K. SapiMouse "Mouse dynamics-based user authentication using deep feature learning." In *Proceedings of the 2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, Timisoara, Romania, 19–21 May 2021; pp. 61–66.
- [12] Niu, H., Chen, J. and Zhang, Z. Cai, Z. (2021). "Mouse Dynamics Based Bot Detection Using Sequence Learning." In *Biometric Recognition: 15th Chinese Conference, CCBR 2021, Shanghai, China, September 10–12, 2021*, Proceedings 15 (pp. 49-56). Springer International Publishing.
- [13] Wei, A., Zhao, Y., & Cai, Z. (2019). "A deep learning approach to web bot detection using mouse behavioral biometrics." In *Biometric Recognition: 14th Chinese Conference, CCBR 2019, Zhuzhou, China, October 12–13, 2019*, Proceedings 14 (pp. 388-395). Springer International Publishing.
- [14] Popular Web Analytics Services Illegally Collect Personal Data. [Electronic source]. URL: <https://www.anti-malware.ru/news/2017-11-23-3/24884> (reference date: 09.09.2023).
- [15] Chu, Zi, Steven Gianvecchio, and Haining Wang. "Bot or human? A behavior-based online bot detection system." *From Database to Cyber Security: Essays Dedicated to Sushil Jajodia on the Occasion of His 70th Birthday* (2018): pp. 432-449.
- [16] Solano, J., Lopez, C., Rivera, E., Castelblanco, A., Tengana, L., and Ochoa, M. (2020, November). "Scrap: synthetically composed replay attacks vs. adversarial machine learning attacks against mouse-based biometric authentication." In *Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security* (pp. 37-47).
- [17] Antal, M., & Fejér, N. (2020). "Mouse dynamics based user recognition using deep learning." *Acta Universitatis Sapientiae, Informatica*, 12(1), pp. 39-50.
- [18] Kılıç, Arjen Aykan, Metehan Yıldırım, and Emin Anarım. "Bogazici mouse dynamics dataset." *Data in Brief* 36 (2021): 107094.
- [19] Leiva, Luis A., and Ioannis Arapakis. "The Attentive Cursor Dataset." *Frontiers in Human Neuroscience* 14 (2020): 565664.
- [20] Antal, M. Sapimouse. Python. 2021. [Electronic source]. URL: <https://github.com/margitalant68/sapimouse> (reference date: 05.09.2023).
- [21] Shen, C., Cai, Z., & Guan, X. (2012, June). "Continuous authentication for mouse dynamics: A pattern-growth approach." In *IEEE/IFIP international conference on dependable systems and networks (DSN 2012)* (pp. 1-12). IEEE.
- [22] Karim, Masud, and Md Hasanuzzaman. "A Study on Mouse Movement Features to Identify User." (2020).