Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

DESIGN AND IMPLEMENTATION OF HUMAN-BEHAVE BOT FOR REALISTIC WEB BROWSING ACTIVITY GENERATION

Akhil Vuputuri
Aris Cahyadi Risdianto
Ee-Chien Chang

School of Computing
National University of Singapore (NUS)
13 Computing Drive, SINGAPORE 117417

ABSTRACT

Cybersecurity experiments/exercises in a testbed require benign traffic to ensure a more effective and rigorous experiments/exercises process. This traffic is usually generated by humans, to camouflage the attack traffic by producing different activities/tasks in the testbed. Some agents can generate this traffic, but unfortunately, most simple agents are predictable, and their actions are easily distinguished from human behavior. We propose distributional models of human activity datasets to mimic replicate specific human actions. These models can be employed in a human agent (i.e., bot) controlled by our orchestrator to generate a realistic human activity generation that is easy to deploy and scale. This paper discusses our modeling of real users web browsing data into well-known distributions and how to fed them into our bot. The results show that our bot can produce realistic web browsing activities similar to real-human behavior.

1 INTRODUCTION

As the complexity of generation scenarios may change dramatically or the size of the exercise environment needs to scale up very quickly, a human-based or hard-coded tools scenario can not achieve scalability or efficiency in producing background traffic. This paper adopted the well-known bot concept from a BotNet (Kokkonen et al. 2015). However, a large number of bots execute complex human activities orchestrated by a single controller. The bot is designed as an independent software component with logic packed (i.e., containerized) for web browsing activity. It is fed with decision-making based on distribution models constructed from real human data to mimic real user's activity. The summary of contributions are:

1. Model the dataset from real users dataset into well-known distribution of users' browsing behaviors.
2. Propose a unique web browsing bot implementation to take the model as input.
3. Show proposed bot can generate realistic activity that are very similar to the users' behavior.

2 REALISTIC WEB BROWSING ACTIVITY GENERATION

To mimic the web browsing activity generated by a human, we model the web browsing activity dataset. We select six features of human web browsing behaviours (i.e., normal inter-keystroke interval, password inter-keystroke interval, page reading rate, probability of first-time visit, funds deposit, and transfer amount) from five datasets, How We Type (Feit et al. 2016), Keystroke Dynamics (Killourhy and Maxion 2009), Natural Stories Corpus (Richard Futrell and Edward Gibson and Hal Tily and Idan Blank and Anastasia Vishnevetsky and Steven T. Piantadosi and Evelina Fedorenko) and Daily website visitors (Nau 2020), PaySim Mobile Money Simulator (Lopez-Rojas et al. 2016). The overall design of realistic browsing activity generation, including data preparation, measurement results, and comparison, is shown in 1. As an example of web browsing activity generation, a terracotta bank web application (Cummings 2021) is used with different workflows such as login/logout, account deposit, and balance transfer.

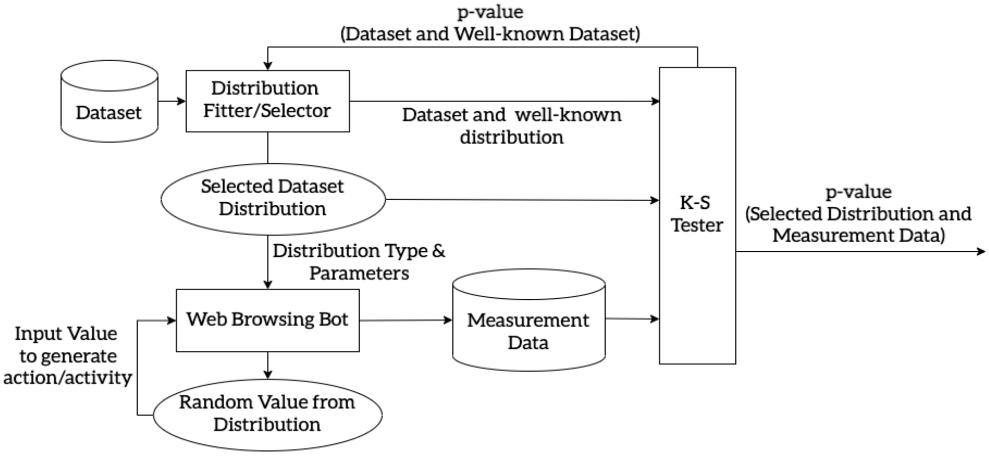
p-value
(Dataset and Well-known Dataset)
Dataset
Distribution Fitter/Selector
Dataset and well-known distribution
Selected Dataset Distribution
K-S Tester
p-value
(Selected Distribution and Measurement Data)
Distribution Type & Parameters
Web Browsing Bot
Measurement Data
Input Value to generate action/activity
Random Value from Distribution

Figure 1: Overall design of browsing activity generation using human behavior dataset.

3 BOT ACTIVITY MEASUREMENT AND ANALYSIS

We compare the distribution of the original data and generated data for each of the six features using K-S (Kolmogorov-Smirnov) test. Table 1 shows how ”realistic” our bot is in generating web browsing activity. Given that the p-values are all greater than 0.45, even at a 45% level, we cannot reject the null hypothesis (i.e., generated data sample came from the distribution of the original datasets).

Table 1: Distribution distance between datasets and measurement data.

Features	K-S Test Statistic	p-value
Password Keystroke Interval	0.014372759287635595	0.5650549040480954
Normal Keystroke Interval	0.01351518257560047	0.6435859689093139
Reading Time	0.015360888676002937	0.47855259177692344
Probability of First-Time Visit	0.012407546140967673	0.7448894171799325
Funds Transfer Amount	0.014301810387064517	0.5714668079068523
Funds Deposit Amount	0.012958882639986402	0.6948866263149966

REFERENCES

Cummings, Josh 2021. “Terracotta Bank”. https://github.com/terracotta-bank/terracotta-bank. accessed August 25th, 2021.

Feit, A. M., D. Weir, and A. Oulasvirta. 2016. “How We Type: Movement Strategies and Performance in Everyday Typing”. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, 4262–4273. New York, NY, USA: Association for Computing Machinery.

Richard Futrell and Edward Gibson and Hal Tily and Idan Blank and Anastasia Vishnevetsky and Steven T. Piantadosi and Evelina Fedorenko. “The Natural Stories Corpus”. https://arxiv.org/abs/1708.05763. accessed October 14th, 2022.

Killourhy, K. S., and R. A. Maxion. 2009. “Comparing Anomaly-Detection Algorithms for Keystroke Dynamics”. In *2009 IEEE/IFIP International Conference on Dependable Systems Networks*, 125–134.

Kokkonen, T., T. Hämäläinen, M. Silokunnas, J. Siltanen, M. Zolotukhin, and M. Neijonen. 2015. “Analysis of Approaches to Internet Traffic Generation for Cyber Security Research and Exercise”. In *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*, edited by S. Balandin, S. Andreev, and Y. Koucheryavy, 254–267. Cham: Springer International Publishing.

Lopez-Rojas, E., A. Elmir, and S. Axelsson. 2016. “PaySim: A financial mobile money simulator for fraud detection”. In *28th European Modeling and Simulation Symposium, EMSS, Larnaca*, 249–255. Dime University of Genoa.

Nau, Bob 2020. “Daily website visitors (time series regression)”. https://www.kaggle.com/bobnau/daily-website-visitors/version/1. accessed January 14th, 2022.