Publisher-Centric Machine Learning-Based Solution for Click Fraud

Gihan Saminda Pathirage

MS23003378

A THESIS

SUBMITTED TO

SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

December 2024

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Dr. Kalpani Manathunga

Approved for MSc. Research Project:

MSc. Programme Co-ordinator, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign: Gihan Saminda Pathirage

Date: …… 13/09/2024……..

ABSTRACT

Publisher-Centric Machine Learning-Based Solution For Click Fraud

Gihan Saminda Pathirage

MSc. in Information Technology

Supervisor: Dr. Kalpani Manathunga

December 2024

Invalid traffic and click fraud present significant challenges in online advertising, impacting advertising metrics and causing substantial financial losses across the digital advertising ecosystem. While advertisers have access to various protective solutions and receive protection from advertising networks, publishers face limited options for detecting and preventing fraudulent activities on their websites. This gap in publisher-side protection creates a critical area for investigation and development of practical solutions. This research presents an effective publisher-side solution: the Ad Click Fraud Protector (ACFP), an open-source WordPress plugin that detects and prevents click fraud and invalid traffic. The research methodology involved studying browser fingerprinting approaches by collecting browser fingerprints from legitimate users and bots, distinguished through firewall rules and honeypots.

Experimental analysis identified six key browser fingerprinting attributes that effectively distinguish between legitimate and fraudulent traffic. These findings informed the development of the ACFP plugin, which incorporates additional security measures for enhanced protection. Testing of the plugin on two AdSense publisher accounts demonstrated its effectiveness in reducing invalid clicks, minimizing invalid traffic, and decreasing revenue deductions due to invalid clicks. The results show that publishers can effectively protect their ad accounts from penalties and deductions through browser fingerprint-based traffic filtering. This research provides publishers with an accessible, open-source solution for combating click fraud while contributing to the theoretical understanding of browser fingerprinting effectiveness in fraud detection. Additionally, it establishes a framework for future development in publisher-side protection systems.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my supervisor, Dr. Kalpani Manathunga, for her invaluable guidance and support throughout this research journey. Her mentorship has been instrumental in shaping this work, and her encouragement kept me motivated during challenging times. Her commitment to addressing my questions and providing detailed guidance throughout the entire process has been truly remarkable.

I extend my sincere appreciation to Dr. Prasanna Haddala, Dr. Dilshan De Silva, Dr. Dharshana Kasthurirathna, and Dr. Nathali Silva for their insightful feedback during my proposal and progress presentations. I am also grateful to Mr. Jagath Wickramarathne for his enlightening lecture series on research methodology and to all the academic staff at SLIIT who supported me during my academic journey. I would also like to acknowledge my fellow students at SLIIT for their friendship and support throughout this endeavor.

A special note of gratitude goes to my family for their unconditional love, patience, and understanding throughout this challenging journey. Additionally, I wish to acknowledge the various online communities, particularly on Reddit, for providing valuable resources and engaging in constructive discussions that enriched my research.

This would not have been possible without the collective support of everyone mentioned above. Their contributions have been instrumental in bringing this work to fruition.

TABLE OF CONTENTS

List of Figures

List of Tables

Abbreviations

Abbreviation	Full Form
ACFP	Ad Click Fraud Protector
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
API	Application Programming Interface
AWS	Amazon Web Services
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
CPM	Cost Per Mille
CPC	Cost Per Click
CPA	Cost Per Action
CSS	Cascading Style Sheets
DDoS	Distributed Denial of Service
DMP	Data Management Platform
DSP	Demand-Side Platform
GDPR	General Data Protection Regulation
GIVT	General Invalid Traffic
HDR	High Dynamic Range
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IP	Internet Protocol
JSON	JavaScript Object Notation
JWT	JSON Web Token
LSH	Locality-Sensitive Hashing
LSTM	Long Short-Term Memory
ML	Machine Learning
PNG	Portable Network Graphics
PPC	Pay Per Click
RGB	Red Green Blue
ROI	Return on Investment
RTB	Real-Time Bidding
SIVT	Sophisticated Invalid Traffic
SME	Small and Medium-sized Enterprise
SSP	Supply-Side Platform
SVG	Scalable Vector Graphics
UI	User Interface
UTC	Coordinated Universal Time
VPN	Virtual Private Network
VPS	Virtual Private Server
WAF	Web Application Firewall
WebGL	Web Graphics Library