氏　　名　DAO THI THU HA

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2367 号

学位授与の日付　2022 年 9 月 28 日

学位授与の要件　複合科学研究科　情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Detection, characterization, and countermeasure of first-party cooperation-based third-party web tracking

論 文 審 査 委 員　主　査　福田　健介
情報学専攻　准教授
計　宇生
情報学専攻　教授
阿部　俊二
情報学専攻　准教授
金子　めぐみ
情報学専攻　准教授
宮本　大輔
東京大学　大学院情報理工学系研究科　准教授

(Form 3)

Summary of Doctoral Thesis

Name in full DAO THI THU HA

Title Detection, characterization, and countermeasure of first-party cooperation-based third-party web tracking

Third-party web tracking has been used for collecting and correlating user browsing behavior. It is becoming more and more ubiquitous, thus this brings an increase in privacy concerns from Internet users. Due to the increasing use of ad-blocking and third-party web tracking protections, tracking providers have introduced new techniques to continue maximizing their profit based on user data. As the recent sophisticated techniques, the third-party tracking providers have leveraged cooperation from the first-party for tracking user activities. Thus, in this dissertation, we focus on the first-party cooperation-based third-party web tracking, including CNAME cloaking-based tracking and PII leakage-based tracking. In particular, third parties have leveraged cooperation from the first-party by using first-party subdomain Canonical Name Record or Alias (CNAME) record in the Domain Name System (DNS), to bypass the filter lists in browsers and extensions that disguise requests to a third-party tracker as first-party ones; they also have leveraged cooperation from the first-party by using user's personally identifiable information (PII) of first-party authentication flows, to create an identification that is a persistent identity. The goals of this dissertation are to perform a first in-depth analysis of the first-party cooperation-based third-party web tracking and develop the countermeasures to protect user privacy against these tracking techniques on the Internet.

In the first half of this dissertation, we detect, characterize, then develop a countermeasure to protect the end-user against the first-party cooperation-based third-party web tracking technique, namely CNAME cloaking-based tracking. This technique misleads web browsers into believing that a request for a subdomain of the visited website originates from this particular website, while this subdomain uses a CNAME to resolve to a tracking-related third-party domain. It thus circumvents the third-party targeting privacy protections. Specifically, we first characterize CNAME cloaking-based tracking by crawling the top pages of the Alexa Top 300,000 sites and analyzing the usage of CNAME cloaking with CNAME blocklist. We also point out that browsers and privacy protection extensions are largely ineffective to deal with CNAME cloaking-based tracking except for Firefox with a developer's version of the uBlock Origin extension. Secondly, we propose a supervised machine learning-based approach to detect CNAME cloaking-based tracking without the on-demand DNS lookup. We show that the proposed approach outperforms well-known tracking blocklists. Finally, to

circumvent the lack of DNS API in Chrome-based browsers, we design and implement a prototype of the supervised machine learning-based browser extension to detect and filter out CNAME cloaking tracking, called *CNAMETracking Uncloaker*. Our evaluation shows that *CNAMETracking Uncloaker* is able to filter out CNAME cloaking-based tracking requests without performance degradation when compared with the vanilla setting on the Chrome browser.

In the second half of this dissertation, we detect, characterize, then develop a countermeasure to protect the end-user against the first-party cooperation-based third-party web tracking technique, namelyPII leakage-based tracking. This technique uses personally identifiable information (PII) to perform cross-site, cross-browser, and cross-device tracking. We document a PII-based tracking ecosystem that leverages user sign-up and sign-in flows on the popular shopping sites from the Tranco Top 10,000 sites. We perform a first in-depth analysis of PII leakage and present a previously unknown persistent web tracking technique based on this data transfer, which enables tracking providers to generate and store a unique persistent identifier for a user on their servers. By measuring the presence of Online Behavioral Advertising (OBA), we confirm that the tracking providers use leaked PII in their advertising strategies for cross-site, cross-browser, and cross-device targeting and personalization. Also, to provide a wider picture of current in-browser privacy protection techniques, we evaluate the effect of browsers and well-known blocklists against PII leakage. Finally, we propose a hybrid approach to detect PII leakage by combining heuristic and supervised machine learning approaches. We show that the proposed approach outperforms well-known tracking blocklists.

We conclude by emphasizing the research contributions made by this thesis and present some open research problems. We first highlight the practical implication of our work to researchers, browser vendors, and Internet users. We think that this work will stimulate follow-up works in the research community and lead to web browser improvements. We also think that this work increases Internet user awareness regarding privacy. In addition, we identify a number of possible research directions, including measurements, perspectives, and recommendations to improve transparency on the World Wide Web.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　DAO THI THU HA

Title
論文題目　　Detection, characterization, and countermeasure of first-party cooperation-based third-party web tracking

本学位論文は、ウェブにおけるファーストパーティの協調に基づくサードパーティによるユーザトラッキングに関するものである。取り扱うトラッキング技術として、(1) DNS CNAME レコードに基づく CNAME Cloaking に基づく手法、(2) メイルアドレス等の個人情報に基づく手法に着目している。これらの手法は、従前よりその存在は知られていたものの、その検出・解析・防御手法に関する知見は得られていないため、これらを明らかにすることで、ユーザのプライバシ漏洩を防ぐ技術の発展に資することを目的としている。

論文は 6 つの章から構成される。第 1 章では研究の背景と目的、第 2 章では既存のウェブにおけるユーザトラッキング技術に関する関連研究について説明している。

続いて第 3 章では DNS CNAME レコードを使用した CNAME Cloaking によるユーザトラッキング技術の検出・解析・防御について述べている。従来のサードパーティトラッキングではウェブページに明示的にサードパーティのドメイン名が現れることから、ブロックリスト等を用いてユーザトラッキングを検出することが可能である。しかし、DNS によるホスト名の変換を用いることで、ファーストパーティのドメインと見せかけた、サードパーティによるユーザトラッキングが可能となる。この CNAME Cloaking によるユーザトラッキング手法は従前より知られている技術ではあるが、どの程度のウェブサイトで使用されているかについては明らかとなっていなかった。出願者は、著名なウェブサイトのリストの上位 30 万サイトについてホームページのデータ収集を行い、解析の結果、約 0.8% のウェブサイトでこの技術が使用されていることを明らかにした。さらに、各種ウェブブラウザおよびブラウザ拡張がこの技術を検出可能であるかを調査したところ、DNS API を持つ Firefox のみが検出可能であり、ユーザシェアを考慮すると、この技術を効率的に検出することが困難であることを明らかにした。次に、クロールしたデータより特徴量を抽出し、教師あり機械学習手法を適用することで、CNAME Cloaking を検出する技術の提案・評価を行っている。実験結果より、従来技術と比べて高い検出性能が得られることが示された。この学習モデルをブラウザで使用するために、ブラウザ拡張の設計・実装・評価を行い、低オーバヘッドでユーザトラッキングを検出･防御できることを明らかにした。

第 4 章では PII (Personally Identifiable Information)に基づくユーザトラッキング技術の検出・解析について述べている。この技術では、サービス開始時に登録したユーザ情報（メイルアドレス等）を URL に付加することによって、ユーザトラッキングを実現しているが、実際の利用状況については明らかとなっていなかった。出願者は、307 の人気のあるショッピングサイトにユーザ登録を行い調査したところ、35%のサイトにおいてこのトラッキング技術が使われていることを明らかにした。同様に実証実験によって異なるブ

ラウザ間・異なるデバイス間でトラッキングが行われることを確認している。さらに、本技術によるトラッキングを検出するために、ヒューリスティックスおよび教師あり機械学習を組み合わせた手法の提案・評価を行っている。実験結果より、ブロックリストによる検出手法に比べて高い検出性能が得られることを示した。

第 5 章では、3 章・4 章の結果についての考察、手法の限界、各ステークホルダに対する提言などの議論を行っている。第 6 章では結論ならびに今後の課題について述べている。

公開発表会では博士論文の章立てにしたがって発表が行われた。その後行われた論文審査会及び口述試験では、審査委員からの質疑に適切に回答がなされた。

質疑応答の後に審査委員会を開催し、審査委員で議論を行った。審査委員会では、出願者の博士研究の学術レベルが十分高く、実用的な意義も有していることが評価された。

以上を要するに本学位論文は、ウェブにおけるファーストパーティの協調によるサードパーティユーザトラッキングの検出・普及度の解明・防御に関して重要な知見を示したものであり、研究分野の発展に貢献しているという点で学術的な価値が大きい。また、本学位論文の成果は、学術雑誌論文 1 件、フルペーパ査読付国際会議論文 3 件として発表され、社会的な評価も得ている。以上の理由より、審査委員会では、本学位論文が学位の授与に値すると判断した。