Computer Communications 210 (2023) 273–284

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Early detection of new web tracking methods across 1.5 million sites

Ismael Castell-Uroz *, Óscar Sánchez-de-Mingo, Pere Barlet-Ros

The Broadband Communications Research Center of the Universitat Politècnica de Catalunya, Barcelona, Spain

ARTICLE INFO

Keywords:
Web tracking
Code signature
Fingerprinting
Network graph

ABSTRACT

Current web tracking practices pose a constant threat to the privacy of Internet users. As a result, the research community has recently proposed different tools to combat well-known tracking methods. However, the early detection of new, previously unseen tracking systems is still an open research problem. In this paper, we present *TrackSign+*, a novel approach to discovering new web tracking methods. The main idea behind *TrackSign+* is the use of code fingerprinting to identify common pieces of code shared across multiple domains. To detect tracking fingerprints, *TrackSign+* builds a novel *4-mode network graph* that captures the relationship between domains, URLs, online resources, and code fingerprints. We evaluated *TrackSign+* with the 1.5M most popular Internet domains, including more than 45M web resources from almost 77M HTTP requests. Our results show that our method can detect new web tracking resources with high precision (over 92%). *TrackSign+* was able to detect more than 300k new trackers, 800k new tracking resources, and 4.5M new tracking URLs, not yet detected by most popular pattern lists at the time. Finally, we also validated the effectiveness of *TrackSign+* with more than 20 years of historical data from the Internet Archive.

1. Introduction

Although online privacy has attracted much attention in recent years (e.g., [1,2]), the fast evolution [3] and inherent complexity of the Internet [4] make it very difficult to develop effective privacy protection methods. Web tracking – a collection of techniques developed to identify users across multiple domains, browsers, and devices – is the main tool used by web services to compile large amounts of personal data about their users [5]. Previous works have shown that the collected information is used for many purposes, such as targeted advertisement [6] and search customization [7], but also for more obscure practices, including price discrimination [8], credit scoring [9] or personal financial assessment [10].

Over the last decade, the research community has made great efforts to combat web tracking (e.g., [11–20]). Although some of these works have succeeded in the detection of well-known tracking techniques, the early detection of new, previously unseen web tracking methods still remains an open research problem. Nowadays, the only reliable way to discover new tracking systems falls to human experts with the daunting task of analyzing millions of websites. Fortunately, there are ways to narrow down the search to a manageable number under specific circumstances. For instance, experts can study the characteristics introduced by new web standards and programming languages that could potentially be exploited by new tracking methods. Although the community has developed some privacy measurement tools to facilitate this task (e.g. [11,21]), the entire process requires a high degree of expertise and is both hard and time-consuming.

In this work, we seek to transform this blind chase for new web tracking mechanisms into a guided hunt that can lead to results in a much faster and more effective way. The intuition behind our proposal is based on the observation that: (*i*) there are relatively few tracking approaches, but they are shared across many different domains, and (*ii*) different Internet resources, including those using the same or similar tracking methods, share some fragments of code (e.g., JavaScript API calls used to perform the actual tracking). If we could focus our lens on relevant pieces of code with those characteristics from all the resources available on the Internet, there would be a high chance they belong to tracking systems. Once found, we could explore the web again, looking for other resources that share these small pieces of labeled code to automatically classify them.

Based on this observation, we present *TrackSign+*, a new web tracking discovery system that automatically crawls the Internet in search of small pieces of code that are shared across multiple domains. To this end, we use a file partition method based on Rabin Fingerprinting [22], which allows us to split Internet resources (e.g., HTML and JavaScript files) in an unambiguous way based on their content. To distinguish between fingerprints with a higher probability of belonging to tracking systems from other shared code (e.g., JavaScript libraries), *TrackSign+* builds a *4-mode network graph* of the Internet, which captures the relationship between the computed fingerprints and the resources, URLs

* Corresponding author.
E-mail address: ismael.castell@upc.edu (I. Castell-Uroz).

https://doi.org/10.1016/j.comcom.2023.08.017
Received 17 January 2023; Received in revised form 6 August 2023; Accepted 17 August 2023
Available online 21 August 2023