Security and Fraud Exceptions Under Do Not Track

Christopher Soghoian
Center for Applied Cybersecurity Research, Indiana University

Position Paper for W3C Workshop on Web Tracking and User Privacy
28/29 April 2011, Princeton, NJ, USA

Introduction

As the debate over Do Not Track continues to evolve, the most important issue under discussion is the definition of tracking. Generally speaking, advertising networks and other firms that engage in online tracking wish for this definition to be as narrow as possible, while privacy advocates are pushing for a broad definition with few exceptions.

Even among many privacy advocates, there seems to be a general acceptance that companies should be able to engage in some forms of tracking and data collection in order to protect against fraud and security related threats.[1]

The fraud and security issues are particularly challenging, because many companies are unwilling to publicly disclose how much data they need, or how they use it, for fear of tipping off those who would misuse the information. As a result, we are forced to take these companies at their word, without the means to independently verify that they do in fact legitimately need the information they are tracking, and that they need to retain it for as long as they are doing so.

Unfortunately, this exception has the very real potential to swallow the rule. For example, in 2008, primarily in response to strong pressure from European privacy regulators, Yahoo! announced a bold new policy of only retaining identifiable log data for search and other services for 90 days. However, the company keeps a second set of identifiable logs for six months, which it uses for fraud and security related purposes. Although Yahoo! will not reveal how much data is kept in this alternate set of logs, these files can of course be obtained by law enforcement agencies wishing to learn how users interact with Yahoo!'s site, long after the primary database of logs have been anonymized.

There are of course different privacy concerns related to Yahoo!'s first party collection of search query information and data collected by third party advertising networks. While law enforcement agencies have shown a keen interest in search queries, I am not aware of any advertising network that has received queries from law enforcement agencies for information about users' web browsing activities. Nevertheless, the Yahoo! example does serve to demonstrate that data retained for security and fraud purposes can seriously undermine the effectiveness of an

[1] The IETF draft proposal for Do Not Track by Mayer *et. al.* includes security and fraud related exceptions to the definition of tracking. Likewise, the DNT scoping proposal published by the Center for Democracy and Technology includes an exception for “Data collection required by law and for legitimate fraud prevention purposes.”

otherwise privacy-preserving data retention policy, particularly when companies are unwilling to reveal what data is being retained and how long they are keeping it.

First party activities are not considered “tracking”

When the average consumer, regulator or policy maker is told that some kinds of tracking are necessary for purposes of fraud and security, the argument of course sounds reasonable. No one wants to allow fraud or hacking, particularly given that consumers ultimately pay the cost, for example, in the form of higher credit card interest rates and transaction fees.

However, when you actually enumerate the most common examples of tracking for fraud prevention, it quickly becomes clear that most of them do not fall under the any of the definitions of tracking under consideration, even without the fraud and security exceptions. Consider the following scenarios:

- A consumer logging into their online bank account and then paying a bill.
- A consumer clicking on a Facebook “Like” button while visiting a blog.
- A consumer conducting a Google search, and then clicking on one of the search results.
- A consumer clicking form a merchant’s shopping cart to Paypal, where they authenticate and then pay for a product.
- A consumer purchasing an item at Amazon or Walmart’s online store.

In all of these scenarios, the consumer is interacting with a website in a first party manner. No one is suggesting that PayPal should not be able to track a user when they visit their site, that Facebook should not be able to log the clicking of Like buttons to protect against click-jacking,[2] or that Bank of America should not be able to use first party Flash cookies as part of its SiteKey two-factor authentication system.

Next, consider the legitimate desire (and obligation) of third party ad networks to protect against click fraud, in which a malicious first party publisher generates fraudulent clicks for ads displayed on its own site. These ad clicks are a first party activity (or should be considered as such), because the moment a user clicks on the ad network’s banner advertisement, the user is knowingly interacting with that company’s servers. Certainly, as soon as the click is processed, the user will leave the publisher’s website and be taken to the website of the advertiser, who can now drop cookies into the user’s browser as a first party.[3]

[2] Click-jacking issues aside, Facebook logging the clicking of the Like button seems to be a clear first party interaction. However, Facebook logging the display of the Like button before it is clicked is almost certainly a third party interaction, and should be prohibited when the user has enabled Do Not Track.

[3] There are of course advertisements that a consumer can interact with without leaving the publisher’s site. As such, there edge cases that are worthy of further discussion. An example raised by Ashkan Soltani is that clicking the mute button on an auto-playing ad should not be considered a first party interaction.

By excluding these legitimate activities from the definition of tracking, only a few third party forms of tracking remain for us to consider.

Tracking for the purpose of detecting advertising impression fraud

Many advertising networks deliver advertisements under a pay for impression (CPM) model. In order to bill their clients, the advertisers, they need to be able to demonstrate that the 1000 ad impressions were delivered to 1000 different users, and not the same user clicking the reload button 1000 times. This is currently done by giving users unique tracking cookies, and logging impressions.

Before attempting to evaluate the tracking activities necessary to combat ad impression fraud, two important factors should first be considered:

- Apple's Safari browser has long blocked the setting of third party cookies by default. Even so, ad networks still monetize the impressions generated by the millions of consumers using Apple's products.
- An adversary seeking to engage in impression fraud can always delete, modify or refuse to accept cookies. A such, ad networks cannot trust cookies sent to them by adversaries.

These two factors mean that many advertising networks already detect and prevent impression fraud without the benefit of cookies or other unique identifiers.

It would seem rather illogical to permit ad networks to continue to use unique cookies to track users who have expressed a strong desire to not be tracked, when these ad networks already have to make do without giving cookies to millions of Safari users who have expressed no privacy preference at all. As such, I think there is a strong argument to be made that advertising networks should be prohibited from tracking users via cookies or other locally stored unique identifiers when a user has expressed a desire to not be tracked (this could be enforced via legislation, or preferably, by the browsers refusing third party cookies or at least making them session only).

With regards to logs kept by ad networks, the sensitive information is not really the user's IP address, but the information contained in the referring header revealing the first party site that the user was visiting when the advertisement was displayed. In some cases, the privacy concerns could be addressed by redacting the path portion of the URL (webmd.com vs webmd.com/cancer/). However, in other cases, the domain name itself would be sufficient to raise privacy concerns (for example, a gay dating website, or a website focused on a specific medical disease). Because some URLs raise greater privacy issues than others, the only automated way to protect this information reliably would be to redact the entire referring URL.

However, it is likely that advertisers wish to know, even with just aggregate numbers, which specific URLs are generating the most impressions (and clicks) in their advertising campaigns.

As such, the most practical solution to protecting privacy with regard to impression log data may be to use a combination of front end data anonymization (for example, hashing IP addresses) and relatively short retention times.

Tracking for security purposes

Third parties, like first party sites, have a legitimate interest in protecting the security of their systems. This includes detecting and protecting against denial of service attacks and intrusions.

In the case of denial of service attacks, logging can be used to detect large numbers of requests from the same IP address, although this is less useful when the attack is distributed among a large pool of IP addresses. It is unclear though why long data retention periods are necessary to protect against such attacks. Furthermore, if a particular IP address is not generating traffic above some reasonable threshold, it is not even clear why logs are necessary at all.

In order to protect against intrusions and other sophisticated attacks, companies obviously want to know how a potential attacker is interacting with their servers. Of course, hackers do not identify themselves as intruders beforehand, and so sites must log every single request in order to later determine which particular requests were associated with a hacking attempt.

While it would be unwise to try and dictate what data third parties can and should collect in order to protect their systems against skilled attackers, it is worth noting that all companies face the problem of security breaches and denial of service attacks. As such, there isn't likely to be any particular "secret sauce" specific to protecting third party sites from attack. Unlike sector-specific attacks such as click fraud and ad impression fraud, it should be possible to have a relatively open discussion about the data retention and tracking necessary to reasonably protect against the general security threats faced by all firms.

A word on fingerprinting

The use of browser fingerprinting presents a unique problem to those concerned about user privacy. First, users do not know when their browsers are being fingerprinted, and second, users often are not given a way to opt out, at least when fingerprinting is used for fraud prevention.

As a baseline requirement, fingerprinting should be disclosed, when conducted by first or third parties. Not only can consumers not easily determine that a site is collecting a fingerprint of their browser, but few companies will confirm their own use of these technologies, even when directly queried by privacy advocates.[4]

[4] Employees at one prominent first party company would not comment on their own use of fingerprinting technology when I asked. Such silence is disgraceful, and suggests that these companies know they are engaged in a practice that would cause outrage among consumers and legislators if disclosed.

If first parties wish to fingerprint browsers, they should be required to clearly and prominently notify users that it is occurring. This does not mean the website needs to reveal which specific data points are collected and analyzed, but simply that the website is collecting information about the user’s browser that will be used to identify them the next time they visit.

Third parties should be prohibited from using fingerprinting technology, preferably at all times, and at least when a user has enabled a Do Not Track setting in their browser. While there may be legitimate scenarios in which this third party collected information may benefit first parties who wish to protect themselves from fraud, the covert collection of data by these third parties raises far too many privacy issues. Third party fingerprinting is still new enough that it can be quietly killed off without seriously disrupting the market. Now is the time to do, before large numbers of first parties become dependent upon this highly problematic source of tracking data.

Conclusion

The development of Do Not Track policies and technologies promise to deliver a significant increase in privacy protection for the average user. Of course, as the industry continues to remind us, there are some legitimate forms of tracking, and some of these relate to the prevention of fraud and protection of site security.

As is also the case in the area of national security, there is a great risk that those wishing to abuse their powers may hide their otherwise improper behavior behind the shroud of “security.”

Technologists and regulators should be highly skeptical regarding companies’ claims of security and fraud, at least when they are unwilling to reveal the exact data they need to track, and how long they wish to keep it. Many such claims cannot, and will not stand up to reasonable analysis.