Bridging the Disconnect Between Web Privacy and User Perception

Mike Perry
mikeperry@torproject.org

Abstract

There is a huge disconnect between how users perceive their online presence and the reality of their relationship with the websites they visit. This position paper explores this disconnect and provides recommendations for making the technical reality of the web match user perception. We frame the core technical problem as one of "linkability"—the level of correlation between various online activities that the user naturally expects to be independent. We look to address the issue of unexpected linkability through both technical improvements to the web's origin model, as well as through user interface cues about the set of accumulated identifiers that can be said to comprise a user's online identity. We argue that without approaching the problem from both of these directions, true privacy-by-design is simply not a possibility for average web users.

1 Introduction

The prevailing revenue model of the web is an appealing one. Web users receive unfettered, frictionless access to an extensive variety of information sources in exchange for viewing advertising. This advertising is more valuable if each advertisement is relevant to the current activity, and if possible, relevant to the current user.

The cost of this incentive structure is that user privacy on the web is a nightmare. There is ubiquitous tracking, unseen partnership agreements and data exchange, and surreptitious attempts to uncover users' identities against their will and without their knowledge. This is not just happening in the dark, unseemly corners of the web. It is happening everywhere [9].

The efforts towards ever increasing amounts of web tracking have led to a growing disconnect between users' perception of their privacy and the reality of their privacy. Users simply can't keep up with the ways they are being tracked [5]. When users are being coerced into ceding data about themselves without clear understanding or consent (and in fact, in many cases despite their explicit attempts to decline to consent), serious moral issues begin to arise.

To understand and evaluate potential solutions and improvements to this status quo, we must explore the disconnect between user experience and the way the web actually functions with respect to user tracking.

To this end, the rest of this document is structured as follows: First, we compare the average user's understanding of web tracking to what actually is happening technically behind the scenes, and note the major disconnects. We then examine solutions that bridge this disconnect from two different directions, corresponding to the two major sources of disconnect[1]. The first direction is improving user cues and browser interface to suggest a coherent concept of identity to users by accurately reflecting the set of unique identifiers they have accumulated. The second direction is improving the linkability issues inherent in the multi-origin model of the web itself. Both of these directions must be pursued to provide users with the ability to properly use the web in a privacy-preserving way.

2 User Privacy on the Web

To properly examine the privacy problem, we must probe the average users' perception of what their "web identity" is, and compare their perceptions to the technical realities of web authentication and tracking.

2.1 User Perception of Privacy

Instinctively, users define their privacy in terms of their identity: in terms of how they have interacted with a site in order to inform it of who they are. Typically, the users' perception of their identity on the web is usually a direct function of the identifiers used for strong authentication for particular sites.

[1] We only consider privacy-by-design approaches. Privacy-by-policy approaches such as Do Not Track will not be discussed.

For example, users expect that logging in to Facebook creates a relationship in their browsers when facebook.com is present in the URL bar, but they are typically not aware that this relationship also extends to their activity on other, arbitrary sites that happen to include "Like this on Facebook" buttons or Facebook-sourced advertising content [9].

Many, if not most, users expect that when they log out of a site, their relationship ends and that any associated tracking should be over. Even users who are aware of cookies can be prone to believing that clearing the cookies related to a particular site is sufficient to end their relationship with that site.

Neither of these beliefs has any relation to reality.

2.2 The Technical Reality of Privacy

The technical reality of the web today is that users are usually wrong about their authentication status with respect to a particular site, and are almost always oblivious to the relationship between content elements of arbitrary pages. The default experience is such that all of this data exchange is concealed from the user.

So then what comprises a user's web identity for tracking purposes? At first glance, it appears to be limited to cookies, HTTP Auth tokens, and client TLS certificates. However, this identifier-based approach breaks down quickly on the modern web. High-security websites are already using fingerprinting as a second factor of authentication [4], and data aggregators utilize everything they can to build complete portraits of users' identities [10].

Despite what the user may believe, her actual web identity then is a superset of all stored identifiers and authentication tokens used by the browser. This identity can link a user's activity in one instance to her activity in another instance, be it across time, or even on a single page due to common prevalent third-party content origins.

Therefore, instead of viewing the user's identity as the sum of her identifiers, or as her relationship to individual websites, it is best to view it as the ability to link activity from one website to activity in another website. We will call this property "user linkability".

2.3 User Privacy as Linkability

In terms of what the user actually expects, user privacy is more accurately modeled as the level of linkability between subsequent actions on the web—not just the sum of her unique identifiers and authentication tokens.

When privacy is expanded to cover all items that enable or substantially contribute to linkability, a lot more components of the browser are now in scope. We will briefly enumerate these categories of components.

First, the obvious properties are found in the state of the browser: cookies, DOM storage, cache, cryptographic tokens and cryptographic state, and location. These identifiers are what technical people tend to think of first when it comes to user identity and private browsing, but they are not the whole story.

Next, we have long-term properties of the browser and the computer. These include the User Agent string, the list of installed plugins, rendering capabilities, window decoration size, browser widget size, desktop size, IP address, clock offset and timezone, and installed fonts.

Finally, linkability also includes the properties of the multi-origin model of the web that allow tracking due to partnerships and ubiquitous third-party content elements. These include the implicit cookie transmission model, and also explicit click referral and data exchange partnerships.

2.4 Developing a Threat Model

Unfortunately, just about every browser property and functionality is a potential linkability target. In order to properly address the network adversary on a technical level, we need a metric to measure linkability of the various browser properties that extend beyond any stored origin-related state.

The Panopticlick project by the EFF provides us with exactly this metric [2]. The researchers conducted a survey of volunteers who were asked to visit an experiment page that harvested many of the above components. They then computed the Shannon entropy of the resulting distribution of each of several key attributes to determine how many bits of identifying information each attribute provided.

While not perfect[2], this metric allows us to prioritize our efforts on the components that have the most potential for linkability. It also shows the benefits of standardizing on implementations of fingerprinting resistance where possible. More implementations using the same defenses means more users with similar fingerprints, which means less entropy in the metric. Similarly, uniform feature deployment leads to less entropy.

3 Matching User Perception with Reality

For users to have privacy, and for private browsing modes to function, the relationship between a user and a site must be understood by that user.

Users experience disconnect with the technical realities of the web on two major fronts: the average user is not given a clear concept of browser identity to grasp the privacy implications of the union of the linkable components of her browser, nor does she grasp the privacy implications of the multi-origin model and how identifiers are transmitted under this model.

Both of these areas of disconnect must be fully addressed in order for the average user to have the technical level of privacy that they intuitively expect.

3.1 Conveying Identity to the User

The first major disconnect that prevents users from achieving true privacy-by-design is that most browsers do not provide any cues to the user to indicate that their current set of accumulated linkable state comprise a single, trackable web identity that can be changed or cleared.

We believe that the browser UI should convey a sense of persistent identity prominently to the user in the form of a visual cue. This cue can either be an image, graphic or theme (such as the user's choice of Firefox Persona [7]), or it can be a text area with the user's pseudonym. This idea of identity should then be integrated with the browsing experience. Users should be able to click a button to get a clean slate for a new identity, and should be able to log into and out of password-protected stored identities, which would contain the entire state of the browser.

To this user, the Private Browsing Mode would be no more than a special case of this identity UI—a special identity that they can trust not to store browsing history information to disk. Such a UI also more explicitly captures what is going on with respect to the user's relationship to the web.

Of the major private browsing modes, Google Chrome's Incognito Mode comes the closest to conveying this idea of "identity" to the user, and its implementation is also simple as a result. The Incognito Mode window is a separate, stylized window which clearly conveys that an alternate identity is in use in this window, which can be used concurrently with the non-private identity. The better UI appears to lead to less mode error (in which the user forgets whether private browsing is enabled) than other browsers' private browsing modes [1].

The Mozilla Weave project [8] appears to be proposing an identity-oriented method of managing, syncing, and storing authentication tokens, and also has use cases described for multiple users of a single browser. It is the closest idea on paper to what we envision as the way to bridge user assumptions with reality.

Unfortunately, all current private browsing modes protect only against adversaries with access to the local computer and fail to deal with linkability against network adversaries (such as advertising networks) [1], claiming that the latter is outside their threat model[3]. If the user is given a new identity that is still linkable to the previous one due to shortcomings of the browser, the approach has failed as a privacy measure.

Therefore, in addition to isolating explicit disk-based identifiers and browser state, an attempt should be made to obfuscate or alter the biggest culprits in terms of the entropy linkability metric mentioned in Section 2.4.

[2] In particular, Panopticlick did not measure all aspects of resolution information. It did not calculate the size of widgets, window decoration, or toolbar size. We believe these resolution-related properties may add high amounts of entropy to the resolution component. It also did not measure clock offset and other time-based fingerprints. Furthermore, as new browser features are added, the experiment should be repeated to measure them.

[3] The primary reason given to abstain from addressing a network adversary is IP-address linkability. However, we believe this argument to be a red herring. Users are quite capable of using alternate Internet connections, and it is common practice for ISPs (especially cellular IP networks) to rotate user IP addresses daily, to discourage users from operating servers and to impede the spread of malware.

However, not all linkability sources have viable solutions under an identity-isolation approach, and moreover, identity-isolation approaches fail to protect the user against linkability due to ubiquitous third party content elements that track them across nearly all sites as soon as they log into any one site [9].

3.2 Improving the Origin Model

The other primary source of disconnect between user expectations and reality on the web is the origin model that governs cookie and other identifier transmission. The model allows unique, globally linkable identifiers to be transmitted for arbitrary content elements on any page, and such elements can be sourced from anywhere without user interaction or awareness. This property enables popular advertising and content distribution networks to have near-omniscient visibility into all user activity retroactively after any level of authentication takes place with a cooperating partner site.

This identifier transmission model is fundamentally flawed when viewed from the perspective of meeting the expectations of the user.

Industry has so far resisted changes to the identifier transmission model due to compatibility concerns and inertia. However, the disconnect is so severe and the associated tracking is so pervasive that some level of temporary breakage must be tolerated to improve the status quo. Because of the retroactive nature of the linkability of cookies and other identifier storage, and because of the invisible and pervasive nature of these partnerships, privacy-by-design is essentially impossible to provide to the average user without addressing this issue.

However, the behavior of identifiers and linkable attributes can be improved to make linkability less implicit and more consent-driven without the need for cumbersome interventionist user interface, and with minimal damage to existing content. Where explicit identifiers exist, they should be tied to the pair of the top-level origin and the third-party element origin. Where linkability attributes exist, they can be obfuscated on a per-origin basis.

The work done by the Stanford Applied Crypto Group shows that it is relatively straightforward to isolate the browser cache to specific top-level origins, effectively binding identifiers hidden in cached elements to the pair of top-level and third-party origin [3]. Commonly sourced third-party content elements are then fetched and cached repeatedly, but this is necessary to prevent linkability: each of these content elements can be crafted to include an identifier unique to each user, thus tracking even users who clear normal cookies.

The Stanford group correctly observed that individually, origin model improvements do not fully address the linkability problem unless the same restriction is applied uniformly to all aspects of stored browser state, and all other linkability issues are dealt with. Behind-the-scenes partnerships can easily allow companies to continue to link users to their activity through any linkable aspect of browser state that is not properly compartmentalized to the top level origin and bound to the same rules as all other linkable attributes.

Along these lines, the Mozilla development wiki describes an origin model improvement for cookie transmission written by Dan Witte [11]. He describes applying this same dual-keyed origin to cookies, so that cookies would only be transmitted if they match both the top-level origin and the third-party origin involved in their creation. Dan observed minimal breakage to popular sites, and where breakage did occur, alternative approaches that do not violate the new model were readily available to web designers and often already in use.

Similarly, this two-level dual-keyed origin isolation can be deployed to improve similar issues with DOM Storage and cryptographic tokens, so that these identifiers are sent only if both the top-level and the third-party origins match. This dual-origin policy should be considered a must for all future origin-bound identifiers.

With a clear association between third-party cookies and their top-level origin due to double-keying, it becomes easier to provide the user with more intuitive control over site identifiers, and thus with more control over their actual relationship to particular sites. For example, the privacy settings window could have a user-intuitive way of representing the user's relationship with different top-level origins, perhaps by using only the 'favicon' of that top-level origin to represent all of the browser state accumulated by that origin. The user could delete the entire set of top-level origin state (cookies, cache, storage, cryptographic tokens) simply by removing the favicon from her privacy info panel.

Linkability based on fingerprintable browser properties is also amenable to improvement under this model. In particular, one can imagine per-origin plugin loading permissions, per-origin limits on the number of fonts that can be used, and randomized window-specific time offsets.

While these approaches are in fact useful for bringing the technical realities of the web closer to what the user assumes is happening, they must be deployed uniformly, with a consistent top-level origin restriction model. Uniform deployment will take significant coordination and standardization efforts. Furthermore, even an improved origin model cannot protect against large multi-service top-level origins. Therefore, both origin improvements and identity-isolation approaches are necessary.

4 Conclusions

The appeal of the web's prevailing revenue model and the difficulties associated with altering browser behavior have lulled us into accepting user deception as the norm for web use. The average user completely lacks the understanding needed to grasp how web tracking is carried out. This disconnect is so extreme that it raises moral issues about the level of consent actually involved in web use and associated tracking.

In fact, standardization efforts realized this problem early on but failed to create feasible recommendations for improving the situation. RFC 2965 governing HTTP State Management mandated in Section 3.3.6 that third-party origins must not cause the browser to transmit cookies unless the interaction is "verifiable" and readily apparent to the user [6]. In Section 6, it also strongly suggested that informed consent and user control should govern the interaction of users to tracking identifiers.

Without changes to both browser behavior and browser interface, informed consent is simply not possible on today's web. The lack of informed consent makes it impossible to expect privacy-by-design approaches to function properly. We cannot expect users who do not even understand the basic properties of these tracking mechanisms to effectively use privacy mechanisms to avoid, opt out of, or decline such tracking. Therefore, browser interface and behavior must be brought in line with user expectations.

References

1. Gaurav Aggrawal, Elie Bursztein, Collin Jackson, and Dan Boneh. An analysis of private browsing modes in modern browsers. In *Proc. of 19th Usenix Security Symposium*, 2010.
2. Peter Eckersley. How unique is your web browser? In *Proceedings of the 10th international conference on Privacy enhancing technologies*, PETS'10, pages 1–18, Berlin, Heidelberg, 2010. Springer-Verlag.
3. Collin Jackson and Dan Boneh. Protecting browser state from web privacy attacks. In *In Proceedings of the International World Wide Web Conference*, pages 737–744, 2006.
4. Jennifer Valentino-DeVries. Evercookies and Fingerprinting: Are Anti-Fraud Tools Good for Ads?, 2010. `http://blogs.wsj.com/digits/2010/12/01/evercookies-and-fingerprinting-finding-fraudsters-tracking-consumers/`.
5. Julia Angwin and Jennifer Valentino-DeVries. Race Is On to 'Fingerprint' Phones, PCs, 2010. `http://online.wsj.com/article/SB10001424052748704679204575646704100959546.html`.
6. D. Kristol and L. Montulli. Http state management mechanism. IETF RFC 2965, October 2000. `http://www.rfc-editor.org/rfc/rfc2965.txt`.
7. Mozilla. Personas. `https://mozillalabs.com/personas/`.
8. Mozilla. The Weave Account Manager. `https://wiki.mozilla.org/Labs/Weave/Identity/Account_Manager`.
9. Arnold Roosendaal. Facebook Tracks and Traces Everyone: Like This! *SSRN eLibrary*, 2010.
10. Emily Steel. Online Tracking Company RapLeaf Profiles Users By Name, 2010. `http://online.wsj.com/article/SB10001424052702304410504575560243259416072.html`.
11. Dan Witte. `https://wiki.mozilla.org/Thirdparty`.