

Bridging the Disconnect Between Web Identity and User Perception

Mike Perry
The Internet
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 some recommendations for making the technical reality of the web match user perception, through both technical improvements as well as user interface cues. By looking at all of the elements of tracking as though they collectively comprise "User Identity", we can make better decisions about improvements to both the technical and the interface aspects of authentication and privacy.

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 more relevant to the current activity, and if possible, more relevant to the current user.

The cost of this 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[10].

The problem is that the revenue model of the web has incentivized companies to find ways to continue to track users against their will, even if those users are attempting to protect themselves through currently available methods. Starting with the infamous "Flash cookies", we have progressed through a seemingly endless arms race of secondary identifiers and tracking information: visited history, cache, font and system data, desktop resolution, keystroke timing, and so on and so forth[5].

These efforts have lead to an even wider disconnect between a user's perception of their privacy and the reality of their privacy. Users simply can't keep up with the ways they are being tracked.

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 tracking and identity.

We only consider implementations that involve privacy-by-design. Privacy-by-policy approaches such as Do Not Track will not be discussed.

2 User Identity on the Web

To properly examine this privacy problem, we must probe into the details of both what a User's perception of their identity is, as well as the technical realities of what goes into web authentication and tracking.

2.1 User Perception of Identity

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 user's perception of their identity on the web is usually a direct function of the mechanisms used for strong authentication for particular sites.

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 likely not aware that this also extends to their activity on other, arbitrary sites that happen to include "Like this on Facebook" buttons or Facebook-sourced advertising content.

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 and private browsing data 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 Technical Reality of Identity

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 is identity? In terms of authentication, it would at first appear to be cookies, HTTP Auth tokens, and client TLS certificates. However, even this begins to break down. High-security websites are already using fingerprinting as an auxiliary second factor of authentication[4], and online data aggregators utilize everything they can to build complete portraits of users' identities[11].

Identity then is a superset of all the authentication tokens used by the browser. It is the ability to link a user's activity in one instance to their activity in another instance, be it across time, or even on a single page due to multiple content origins.

2.3 Identity as Linkability

When 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 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 are what technical people tend to think of first when it comes to private browsing and identity, but they are not the whole story.

Next, we have long-term properties of the browser itself. These include the User Agent String, the list of installed plugins, rendering capabilities, window decoration size, and browser widget size.

Then, we have properties of the computer. These include 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. 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 fingerprinting 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 perfect¹, this metric allows us to prioritize effort at components that have the most potential for linkability.

This metric also indicates that it is beneficial to standardize 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.

3 Matching User Perception with Reality

When the concept of user identity is expanded to cover all aspects of linkability, addressing the problem of the disconnect between user perception and reality becomes clearer. 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.

It is apparent that the user experiences disconnect with the technical realities of the web on two major fronts: the average user does not grasp the privacy implications of the multi-origin model, nor are they given a clear concept of identity to grasp the privacy implications of the union of the trackable components of their browsers.

We will now examine examples of attempts at reducing this disconnect on each of these two fronts.

Note that identity-based approaches and the origin-based approaches are orthogonal. They may be combined, or used independently.

3.1 Origin-Based Approaches

Origin-based approaches seek to improve the technical behavior of the browser to make linkability less implicit and more consent-driven. In short, these approaches seek to make the web behave more like users currently assume it behaves by anchoring browser state to top-level origins as opposed to associating it with arbitrary content elements.

The earliest relevant example of this work is SafeCache[3]. SafeCache seeks to reduce the ability for 3rd party content elements to use the cache to store identifiers. It does this by limiting the scope of the cache to the origin in the url bar. This has the effect that commonly sourced content elements are fetched and cached repeatedly, but this is the desired property. Each of these prevalent content elements can be crafted to include unique identifiers for each user, tracking users who attempt to avoid tracking by clearing cookies.

Mozilla has a wonderful example of an origin-based improvement written by Dan Witte and buried on their wiki[12]. It describes a new dual-keyed origin for cookies, so that cookies would only be transmitted if they matched both the top level origin and the third party origin involved in their creation. This approach would go a long way towards preventing implicit tracking across multiple websites.

Similarly, one could imagine this two-level origin isolation being deployed to improve similar issues with DOM Storage and cryptographic tokens.

Making the origin model for browser identifiers more closely match the user activity and user expectation has other advantages as well. With a clear distinction between 3rd party and top-level cookies, the privacy settings window could have a user-intuitive way of representing the user's relationship with different 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 browser state (cookies, cache, storage, cryptographic tokens) associated with a site simply by removing its favicon from their privacy info panel.

The problem with origin-based approaches is that individually, they do not fully address the entire 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

¹ In particular, the test does not take in all aspects of resolution information. It did not calculate the size of widgets, window decoration, or toolbar size. We believe this may add high amounts of entropy to the screen field. It also did not measure clock offset and other time-based fingerprints. Furthermore, as new browser features are added, this experiment should be repeated to include them.

users to their identities through any aspect of browser state that is not properly compartmentalized to the top level origin and bound to the same rules.

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

So, 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. This may take significant coordination and standardization efforts.

3.2 Identity-Based Approaches

We will now discuss what we call the identity-based approaches to privacy. These approaches, whether explicitly or implicitly, all model the user's web identity as the entirety of the user's state for all origins.

The key advantage of identity-based approaches is that they can be simpler than origin-based approaches when used to improve the privacy problem on their own.

While the earliest example of an identity-based approach is our own work on Torbutton[9], Torbutton deserves poor marks for both simplicity and usability[8]. Torbutton attempts to isolate the user's non-Tor activity from their Tor activity, effectively providing the user with a blank slate for their Tor activity, but optionally allowing them to toggle between these two identities.

Firefox Private Browsing Mode is similar, in that it allows users to switch between their normal browsing and a "private" clean slate.

Both Firefox PBM and Torbutton suffer from usability issues, primarily because this concept of separate browsing identities is not properly conveyed to the user. In Firefox's case, this usability issue is apparent through the quantity of mode error observed in the review of Private Browsing Modes by Dan Boneh et al[1]. In Torbutton's case, the issues appear more severe. We've informally observed that users have tremendous difficulties remembering which tabs were Tor-related and which were non-Tor related, and we've also observed issues with mode error.

Both of these approaches are exceedingly complex: they deal with every aspect of browser state individually. This development effort however does enable Firefox and Torbutton to provide the user with great fine-grained control.

Google Chrome's Incognito Mode comes the closest to conveying this idea of "Incognito identity" to the user, and the implementation is also simpler as a result. The Incognito Mode window is a separate, stylized window that clearly conveys an alternate identity is in use for this window, which can be used concurrent to the non-private identity. This appears to lead to less mode error (where the user forgets their private browsing state) compared to other browsers.

The implementation of Incognito is as a virtualized in-memory profile, which allows them to achieve protection against history storage issues with minimal effort. It also allows them to tweak browser properties and permissions specifically for this profile.

The Mozilla Weave project appears to be proposing an identity-based method of managing, syncing, and storing authentication tokens, and also has use cases described for multiple users of a single browser[7]. It is the closest idea on paper to what we envision as the way to bridge user assumptions with reality.

We believe that the user interface of the browser should convey a sense of persistent identity prominently to the user in the form of a visual cue. This cue can either be an abstract image, graphic or theme (such as the user's choice of Firefox Persona[6]), or it can be a text area with the user's current favored 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 in and out of password-protected stored encrypted identities, which would contain the entire state of the browser. This is the direction the Tor Project intends to head in with the Tor Browser Bundle[8].

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.

However, all current private browsing modes fall short of protecting against a network adversary and fail to deal with linkability against a network adversary[1], claiming that it is outside their threat model². If the user is given a new identity that is still linkable to the previous one due to shortcomings of the browser, this approach has failed as a privacy measure.

Linkability solutions within the identity framework would be similar to the origin-based solutions, except they would be properties of the entire browser or browser profile, and would be obfuscated only once per identity switch.

4 Conclusions

There is a demand for private browsing, and we believe that solid private browsing modes can be created. In order to do this, we need solid analysis of the threat models involved, and we need standardization for many aspects of defense.

However, there is currently a huge disconnect between user privacy and identity due to both the multi-origin nature of the web, and the failure of browsers to adequately convey a sense of identity to the user. It is possible to bridge this disconnect both by addressing the issues with the multi-origin model, as well as providing the user with an explicit representation of their web identity, and with control over this identity.

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. Mozilla. Personas. <https://mozillalabs.com/personas/>.
7. Mozilla. The Weave Account Manager. https://wiki.mozilla.org/Labs/Weave/Identity/Account_Manager.
8. Mike Perry. To toggle, or not to toggle: The end of torbutton. <https://lists.torproject.org/pipermail/tor-talk/2011-April/020077.html>.
9. Mike Perry. The Torbutton Design Document, 2011. <https://www.torproject.org/torbutton/en/design/>.
10. Arnold Roosendaal. Facebook Tracks and Traces Everyone: Like This! *SSRN eLibrary*, 2010.
11. Emily Steel. Online Tracking Company RapLeaf Profiles Users By Name, 2010. <http://online.wsj.com/article/SB10001424052702304410504575560243259416072.html>.
12. Dan Witte. <https://wiki.mozilla.org/Thirdparty>.

² The primary reason given to abstain from addressing the network adversary is IP address linkability. However, we believe this to be a red herring. Users are quite capable of using alternate Internet connections, and it is common practice for ISPs in many parts of the world to rotate user IP addresses daily, to discourage servers and to impede the spread of malware. This is especially true of cellular IP networks.