Identifiers and Online Tracking

Submission to W3C Workshop on Web Tracking and User Privacy
April 28-29 2011, Princeton, NJ, USA

Ashkan Soltani, Independent Researcher and Consultant

March 24, 2011

I. INTRODUCTION

In the active discussions around 'Do Not Track', there seems to be some debate around what constitutes "tracking" and what consumer should expect when they signal that they do not want to be tracked online.

Proposed definitions for opting out of tracking range from companies agreeing to not collect or retain information resulting from online interaction to more obtuse definitions such as not serving personalized ads to these users, but still allowing for data collection. Some definitions provide stronger privacy protections to consumers (albeit with potential burdensome technical requirements for the ad networks) while others will still enable companies to collect (and perhaps even monetize) users' data, even if they have indicated they don't want to be tracked.

In this short position paper, I propose a potential alternative approach to framing tracking that enables companies to engage in measurable online advertisement while providing the most important privacy protections articulated by advocates. This approach focuses primarily on the active removal of persistent identifiers that are used to correlate browsing activity over multiple sessions or multiple websites.

2. CURRENT DEFINITIONS OF TRACKING

There are various definitions what it means to opt-out of 'tracking' but I will summarize them into to primary camps:

A. Do Not Track = Do Not Use For Behavioral Advertising

Under the current system of ad network opt out cookies, consumers can opt out of the use of their data. That is, when a user 'opts-out', companies continue to track the consumer and even build a profile. However, they pledge to not to use this information for targeting although little is known about the secondary uses of this data, such as resale to other companies. This is the least privacy preserving option and arguably even a worse outcome for consumers since they have even less visibility to the data collection that is occurring yet do not receive the benefit of relevant ads. They still pay the privacy "cost", but receive none of the benefits.

B. Do Not Track = Do Not Collect or Retain

Others are pushing for a Do Not Track system requiring companies theoretically delete all information received through third party transactions from consumers indicating that they do not wish to be tracked.

While this would certainly ensure that no private data would be stored by the third party, this implementation is has been criticized by website operators as being overly burdensome or difficult to implement. In order to comply with this definition in the **strictest** sense, they would be required to potentially configure all of the networking equipment and web servers they operate to not log data or delete it immediately. Load balancers, networking switches, routers and SSL accelerators would potentially all need to be modified to ‘respect’ the header and not log the browser request since most network infrastructure is built to log requests by default.

Furthermore, definitions in this category carve out multiple exceptions that allow collection and retention of data for specific uses, such as proving security, verification of ad impressions, or fraud detection. These exceptions will likely need to be crafted carefully and updated frequently in order to allow site operators to reliably serve content and innovate while still adhering to what most consumers expect when they request to ‘not be tracked’.

3. DO NOT TRACK = DO NOT IDENTIFY?

Much of the third party tracking that occurs online hinges on the presence of unique persistent identifiers which allow ‘trackers’ to identify individual users or devices across multiple visits to the same or different websites. These identifiers can be of the form of browser cookies although recent advances have given ways to other methods of identification, such as device fingerprinting and/or persistent storage outside of a browser’s direct control.

Under this proposal, companies that agree to respect the Do Not Track signal could voluntarily make a best faith effort to strip **any** unique identifiers associated with the user/browser/client device as part of a web transaction after the transaction has occurred. The remaining data can be retained assuming that it doesn’t later prove to be identifiable based on existing ‘best practices’ in identification.

This approach is good for business and consumers as it would allow businesses to collect and use data about how their websites are being used while preventing the creation of profiles. Fewer exemptions would need to be created since traffic management, fraud detection, and verification of impressions could all occur without relying on the uniquely identification of a individual device or browser persistently.

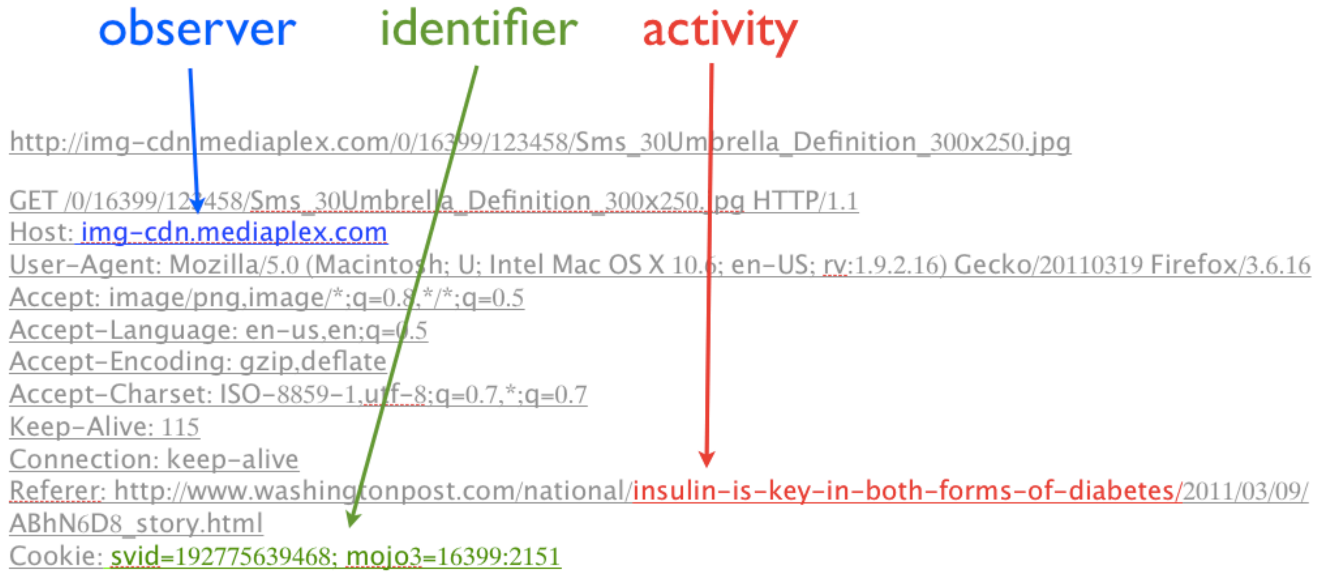
Much like a secret ballot: everyone gets the benefit of voting and the votes are tallied accurately, but no one can tell who voted for whom.

4. POTENTIAL IMPLEMENTATION

Third party tracking consists of 3 key components, present in nearly every connection your browser makes:

OBSERVER: the third party site that is tracking your activity
IDENTIFIER: unique descriptors that allow the 3rd party site to uniquely track you
ACTIVITY: i.e the URL of the 1st party site you're viewing (often the referrer url)

Consider the following snippet of data generated by viewing a page on the **WashingtonPost.com** about **insulin** which included a third party advertisement from **Mediaplex.com**:



```
http://img-cdn.mediaplex.com/0/16399/123458/Sms_30Umbrella_Definition_300x250.jpg

GET /0/16399/123458/Sms_30Umbrella_Definition_300x250.jpg HTTP/1.1
Host: img-cdn.mediaplex.com
User-Agent: Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.6; en-US; rv:1.9.2.16) Gecko/20110319 Firefox/3.6.16
Accept: image/png,image/*;q=0.8,*/*;q=0.5
Accept-Language: en-us,en;q=0.5
Accept-Encoding: gzip,deflate
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
Keep-Alive: 115
Connection: keep-alive
Referer: http://www.washingtonpost.com/national/insulin-is-key-in-both-forms-of-diabetes/2011/03/09/
ABhN6D8_story.html
Cookie: svid=192775639468; mojo3=16399:2151
```

In this request, the **observer** *img-cdn.mediaplex.com* is able to observe that a browser with cookie **identifier** *svid=192775639468* viewed the page *insulin and diabetes* page (**activity)** on the washingtonpost.com website.

Upon repeat activity, the **img-cdn.mediaplex.com** can correlate multiple visits of washingtonpost.com into a browsing profile keyed off of their cookie: **svid=192775639468**.
if the user visits other websites which display third party advertising from **img-cdn.mediaplex.com**, then Mediaplex can correlate this activity across these sites as well, based on the same unique cookie id.

A. Do Not Use For Behavioral Advertising

Currently, some third party trackers allow the user to opt-out of tracking. However, this definition of 'opt-out' varies from third party to third party. While some websites allow users to opt-out of tracking by deleting or masking their cookies they still are able to identify users based on other factors such as IP address or Flash cookies.

In the above example, Mediaplex may allow a user to delete their **svid** cookie, but is still capable of profiling them based on other identifiers such as IP address or browser fingerprint.

While not typically apparent to the user, Mediaplex's systems would have an internal identifier they are utilizing.

B. Do Not Collect or Retain

Conversely, if this user wants to opt-out of this tracking completely, based on the 'Do Not Collect or Retain' definition, we could require that Mediaplex delete all of the log and profile data associated with the web request above. This is effective for consumers that don't wish to be tracked, but would likely make it difficult for Mediaplex to keep a record of this ad impression for accounting purposes.

As such, multiple exemptions may need to be created to allow third parties to retain browsing information in order to provide their basis accounting and security which ultimately goes against what consumers may expect when they believe they're not being 'tracked'.

C. Do Not Identify

Instead of asking Mediaplex to log no data at all, we could potentially request that third party websites strip *any persistent unique identifiers* from requests from consumers indicating that they do not wish to be tracked. In this case, that would mean stripping the unique cookie id although it could mean stripping other identifiers if they occurred in other portions of the request, such as the URL or Referrer header. If other identifiers, such as a browser fingerprint, are utilized on the back-end, the company would also be required to remove these as well. This 'stripping' can occur immediately or after a reasonable amount of time (i.e 24hrs) to facilitate processing of the transaction, though this is something that still needs to be worked through.

This approach would allow websites to collect information for the purpose of ad impressions, anti-fraud, security, and other purposes that are not user-specific. In fact, this is actually the current practice of many big advertisers who delete identifiers in log data as a result of that interaction, including but not limited to IP address, cookies, referrers, etc.

Participating websites could make a good faith effort to employ best practices in de-identification of their data based on evolving research in the field. Since these websites are the ones typically creating these unique persistent identifiers, they are in the best position to determine which information needs to be removed in order to make the data impervious to profiling.

D. Added Benefit for Monitoring Compliance

While all of these definitions require participation by website operators, the 'Do Not Identify' approach has the added benefit that allows web browsers or browser extensions to monitor web traffic and help identify any unique identifiers, such as cookies or URL parameters that are embedded in the content from sites that the user has signaled 'Do Not Track'. This could indicate that this company may be engaging in unauthorized or accidental tracking. Browser fingerprinting or obfuscated identifiers are obviously still possible by rogue trackers, but this issue exists in the ‘do not track = do not collect/retain’ context too, i.e we’re trusting the 3rd party websites to actually comply.

E. Mixed First/Third Party Interactions

Finally, this approach allows companies that operate simultaneously in first and third party context to comply with Do Not Track with no significant advantage to those that simply have third party presence, such as traditional ad networks. Companies that the user has a first-party relationship with, such as social networks or video sites, would still be able to serve personalized third-party content, such as social widget or ‘over 18’ video content, as per normal based on the identifier that was created during a first-party visit. However, the third party social network or video site could still be required to strip any unique identifiers in the subsequent tracking data recorded by these passive third party impressions.

Once the user takes direct action with a third party object, such as clicking a ‘Share Widget’ this could potentially convert the interaction into a first party user experience and fall outside of the scope of Do Not Track. However ‘forced’ first party interactions, such as auto-playing of an embedded third party video that the user must dismiss with should still be covered.

5. CONCLUSION

While much discussion and clarification is needed to properly define what companies should do to comply with Do Not Track, focusing on identifiers *could be* a simple approach to reducing unwanted tracking/profiling while still enabling companies to engage in measurable online advertisement. Companies in the space can then innovate on ways to provide ads and services in a reliable way that does not infringe on a users desire to not be tracked/profiled.