Part 1: tracking by edge providers

Online tracking has evolved into an arms race. Sneaky tracking techniques that are difficult for consumers to see and control are increasingly being deployed.

One such technique is canvas fingerprinting. It works by drawing an invisible image on the page and reading it back as a sequence of bits. Due to tiny system differences, this bit string acts as a device identifier. Our 2014 study found canvas fingerprinting prevalent on over 5% of websites.

We find that measurement changes behavior. Days after our study was released and the attendant publicity, the largest providers of canvas fingerprinting stopped the practice.

This is because measurement removes information asymmetry between trackers and the rest of the web. In fact, in many cases publishers themselves are unware of the third-party tracking on their own sites.

In our 2016 study we find that news sites have the most trackers on average, and adult sites the least.

Part 2: ISPs and privacy

We find that the prevalence of web encryption is low. The percentage of traffic that is encrypted is not the right choice of metric since it is skewed by video statistics, especially Netflix.

However, when we look at the top 500 websites (as ranked by Alexa) in various categories, a very different picture emerges. 84% of health websites are HTTP-only, 86% of shopping websites, and 97% of news websites.

It is also important to ask what percentage of web pages visited by a typical user are encrypted. We use Twitter links as a proxy for web browsing behavior. We find that of a random sample of around 1,500 Twitter links, 67% are unencrypted and 33% are encrypted.

ISP access to user information is comprehensive. On unencrypted pages, URLs often reveal highly sensitive information. URLs are part of application headers. Accessing URLs doesn't require scanning the entire contents of packets.

Further, even on encrypted pages ISPs may be able to infer the URLs being visited due to traffic analysis or side channel attacks. There are a few computer science research papers on this topic.

To drive home the point that ISPs have a unique, and in some ways greater, visibility into a customer's web browsing than web tracking companies, another of our research findings is relevant: only 3 companies (Google, Facebook, and Twitter) have third parties that are present on over 10% of web pages. In contrast, ISPs see 100% of unencrypted web page URLs.

Similarly, users have much more control over tracking by web third parties than over tracking by ISPs. Our research shows that tools such as Ghostery are largely effective at blocking prominent third parties, but these tools do nothing to stop data collection on the wire.

Part 3: de-identification.

Re-identification has a habit of surprising us. After Latanya Sweeney's re-identification of Mass. hospital records, the lesson drawn was to consider "quasi-identifiers" such as ZIP code as having potential for re-identification. But later studies such as the Netflix Prize re-identifiability demonstration have shown that any piece of data can potentially be used for re-identification.

Re-identifiability has been demonstrated repeatedly: with location data, credit card data, social n/w structure, writing style, coding style, typing cadence, genetic data, and so forth. Another lesson from CS research: In the vast majority of cases, longitudinally linked data cannot be effectively anonymized.

Narayanan, Huey, & Felten recently proposed a precautionary approach to data privacy, where the burden of proof that the released data will resist re-identification attempts falls on those releasing the data.

We recommend that non-longitudinally-linked and summary data present far fewer privacy concerns than de-identified micro-data, so there should be exceptions/carve-outs for the former type of data.

Papers cited

Acar, Eubank, Englehardt, Juarez, Narayanan, Diaz; The Web Never Forgets: Persistent Tracking Mechanisms in the Wild.

Swire, et al., Online Privacy and ISPs: ISP Access to Consumer Data is Limited and Often Less than Access by Others (2016)
http://www.iisp.gatech.edu/sites/default/files/images/online_privacy_and_isps.pdf

Englehardt, et al., Cookies That Give You Away: The Surveillance Implications of Web Tracking (2015) http://www.www2015.it/documents/proceedings/proceedings/p289.pdf

Greenwald, M., A Comprehensive Privacy Analysis of Fitness Tracker Companies (2015)

Miller, et al., I Know Why You Went to the Clinic: Risks and Realization of HTTPS Traffic Analysis (2014) https://www.petsymposium.org/2014/papers/Miller.pdf
Danezis, Traffic Analysis of the HTTP Protocol over TLS http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.3893&rep=rep1&type=pdf
Englehardt, S. and Narayanan, N., Online tracking: A 1-million-site measurement and analysis (2016), http://randomwalker.info/publications/OpenWPM_1_million_site_tracking_measurement.pdf

Sweeney, L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (2002). https://epic.org/privacy/reidentification/Sweeney_Article.pdf

Narayanan, A., & Shmatikov, V. Robust de-anonymization of large sparse datasets. (2008) https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

de Montjoye, Y., et al. "Unique in the crowd: The privacy bounds of human mobility." (2013).
de Montjoye, Y., et al. "Unique in the shopping mall: On the reidentifiability of credit card metadata." (2015).
Narayanan, A, and Shmatikov, V. "De-anonymizing social networks." (2009).
Narayanan, A., et al. "On the feasibility of internet-scale author identification." (2012).
Caliskan-Islam, A., et al. "De-anonymizing programmers via code stylometry." (2015).
Monrose, F., & Rubin, A. D.. Keystroke dynamics as a biometric for authentication. (2000)
Gymrek, M., et al. Identifying personal genomes by surname inference. (2013).

Narayanan, A., Huey, J., & Felten, E. W. A Precautionary Approach to Big Data Privacy (2016). http://randomwalker.info/publications/precautionary.pdf