Creating Privacy Labels for the Web

Owen Kaplan, Logan Brown, Daniel Goldelman, Sebastian Zimmeck

Department of Mathematics and Computer Science, Wesleyan University

privacy-tech-lab

https://www.privacytechlab.org/

Abstract

Since its advent, the internet has become an integral part of modern life. Consumers have grown to expect free-to-use content and technology. Thus, the economics of the web ecosystem rely upon advertising revenue that is sourced from user data. However, for most users it is not transparent what happens to their data while using the internet. **We approach this problem by (1) creating privacy labels by analyzing HTTP communication and (2) surfacing these labels in an easy-to-understand environment to the user.**

We created a proof-of-concept browser extension that identifies and labels privacy practices live as a user browses the web. In our work, we hope to increase the transparency of the web.

Introduction

At the core of the internet is the communication between a client and a server. When a user loads https://example.com on their browser, they initiate HTTP requests to the example.com server, which sends back HTTP responses that contain, among other data, the content of the page.

Request

```
GET / HTTP/1.1
Host: developer.mozilla.org
Accept-Language: fr
```

Response

```
HTTP/1.1 200 OK
Date: Sat, 09 Oct 2010 14:28:02 GMT
Server: Apache
Last-Modified: Tue, 01 Dec 2009 20:18:22 GMT
ETag: "51142bc1-7449-479b075b2891b"
Accept-Ranges: bytes
Content-Length: 29769
Content-Type: text/html

<!DOCTYPE html... (here comes the 29769 bytes of the requested web page)
```

**Example of an HTTP Request/Response (1).
In this example, the server sends back HTML content.**

Loading a website can consist of hundreds of HTTP requests and responses. Most of this communication is not surfaced to the user. Importantly, not all requests are sent to the server that the user originally connected to. For example, when loading https://example.com, it is possible that this site initiates a request to a third party ad server https://adexample.com to show advertising to the user. Our work centers on identifying and labeling these types of privacy practices that are normally not visible to the user.

Interpreting HTTP Communication

Using Firefox's browser API's, notably, the StreamFilter API (2), we are able to analyze HTTP communication as it happens live. Our extension observes the communication, decrypts it if necessary, runs privacy analysis routines, creates labels based on that analysis, and then surfaces these labels to the user.

Our analysis routines use various techniques:

1. **Open source lists of privacy-invasive services**. Existing privacy tools, such as ad blockers, often include lists of ad networks and other privacy-invasive services. We use the Disconnect (3) list to identify those in the traffic we observe.
2. **Regular expressions**. To look for certain kinds of data, we found that regular expression pattern matching works well. For example, if you live in Montclair, an HTTP request may identify your location with the parameter "Montclair."
3. **Resource-specific analysis**. To identify tracking pixels, which are transparent images to ping ad networks, we look for 1x1 or 0x0 images with the word "pixel" in the request URL.
4. **User-specific data.** We also analyze user-specific data, for example, the user's GPS location or Internet Protocol address. To that end, we allow users to enter information themselves they want to flag and then look for it in the web traffic.

From the observed information our extension automatically creates privacy labels, similar to nutrition labels, so the user can quickly and easily understand the result of the privacy analysis.

Privacy Label Categories and Types

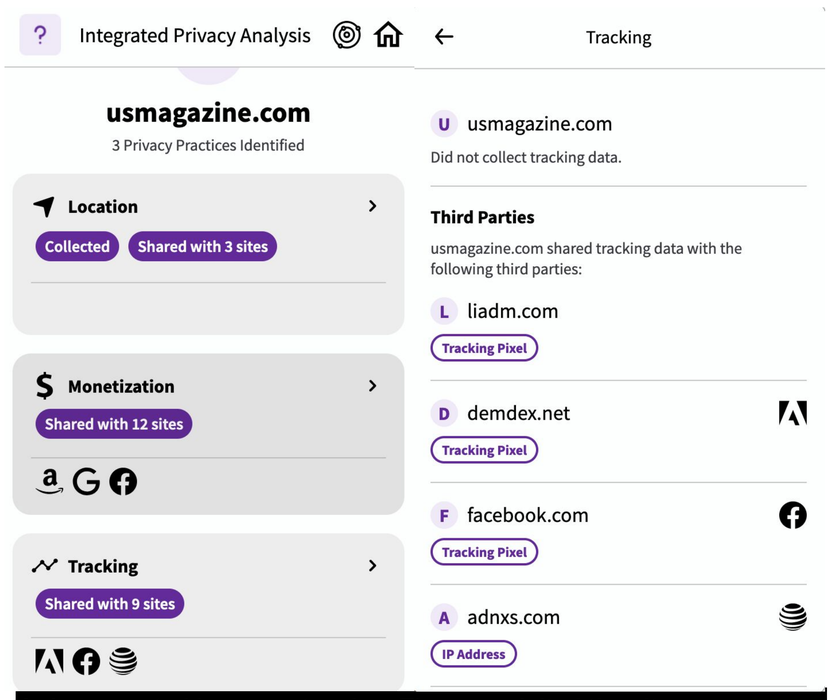
Our extension creates the following labels:

1. **Monetization**
   - Advertising
   - Analytics
   - Social Networking
2. **Tracking**
   - Tracking Pixel
   - IP Address
   - Browser Fingerprinting
3. **Location**
   - Coarse Location
   - Fine Location
   - ZIP Code
   - Street Address
   - City
   - State
4. **Watchlist**
   - Phone Number
   - Email Address
   - Encoded Email Address
   - Custom Keywords

Sites and their Privacy Labels

Labels are created for first and third party websites. A first party website is a site the user enters in the URL bar. A third party site is an ad network or other site that the user does not intentionally visit but that is part of the first party site as it is loaded.

Our extension groups labels by first parties and also shows the third party sites' privacy practices. We are able to identify third party sites by checking a site against the user's browser history to see if they actually visited that site.



Labels created by our extension. The user visited the site usmagazine.com, which collects the user's location data and allows twelve third party sites to monetize it and nine third party sites to track the user. The tracking can happen through tracking pixels.

Overview Privacy Labels

In addition to the privacy labels for the individual sites our extension also provides a summary of the created privacy labels.

On the Overview page users get a quick summary of the labels our extension has created giving them a higher-level picture of the footprint of their internet activity. For example, a total of 60 sites monetized the users data as shown above.

Surfacing Context to the User

Our extension surfaces the privacy analysis results such that the user gets an understanding in which context a certain privacy practice occurred.

In this example the user's coordinates are part of a larger HTTP request, which appears to cover the user's location with different parameters.

Evaluating Usability

One of the major obstacles for users to adopt privacy technologies is usability. In particular, it is challenging to comprehensively and succinctly convey privacy information to average users who are not privacy experts. We believe that privacy labels can help overcoming this challenge. We ask:

How do privacy labels help users understand how their data is used on the web?

Conclusion

Increasingly, privacy labels are being deployed in order to better inform users how their data is used. In controlled ecosystems, such as iOS, each app is now required to have privacy labels (4). However, the web ecosystem does not have comparable privacy labels. Its lack of central governance, scale, and dynamic nature make the automatic creation of privacy labels for the web challenging. In our work we are making inroads towards that goal by automating privacy analysis and label creation techniques and packaging them into a proof-of-concept browser extension.

References

1. https://developer.mozilla.org/en-US/docs/Web/HTTP/Messages
2. https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/API/webRequest/StreamFilter
3. https://github.com/disconnectme/disconnect-tracking-protection
4. https://developer.apple.com/app-store/app-privacy-details/

Acknowledgements

We are grateful to Wesleyan University's Department of Mathematics and Computer Science and the National Science Foundation (Grant #2055196) for their support. Conclusions reached or positions taken are our own and not necessarily those of the supporters, its trustees, officers, or staff.