Private Information Disclosure from Web Searches

Claude Castelluccia[1], Emiliano De Cristofaro[2], Daniele Perito[1]
[1]INRIA Rhone Alpes, Montbonnot, France [2]University of California - Irvine
http://planete.inrialpes.fr/projects/private-information-disclosure-from-web-searches

The Session Hijack Attack

1. Listen passively on a unencrypted channel (e.g., wi-fi hotspot).
2. Capture an HTTP Authentication Cookie.
3. Replay captured cookie to impersonate a user.

Session Hijacking is very well-known and can be prevented by transmitting authentication cookies only over HTTPS.

Google Authentication

- All Google services are accessible with a single set of credentials.
- Authentication to all services is done on the *Accounts* page (on HTTPS).
  - Services considered "sensitive" (e.g., Gmail) are then served on HTTPS.
  - All other services (including Google Search) are then served on HTTP.

Hint: A service using HTTP cookie-based authentication is sensible to Session Hijacking.

Session Hijacking on Google

We monitored the INRIA network for 2-weeks, and explored the percentage of successfully hijacked accounts for the following Google services:

Type of information leaked	Corresponding service	Accessible Accounts
Blogs followed on Reader	Reader	15%
Address book	Contacts	87%
Maps search history	Maps	79%
Default address on Maps	Maps	5%
Financial portfolio	Portfolio	1%
First/Last name	Maps profile	75%
Bookmarks	Bookmarks	27%

Google Web History

- Available at `www.google.com/history`.
- It is an *opt-out* service that stores *all* Web searches conducted by a signed-in user.
- **Google's Protection**: To prevent session hijacking, signed-in users are asked to re-enter their credentials to access Web History.

Figure: An example of a user's Google Web History.

Google Personalized Services

- **Personalized Results.** Search results are personalized based on user search history.
- **Personalized Suggestions.** User receives **max 3** *as-you-type* keyword suggestions based on previously conducted searches.

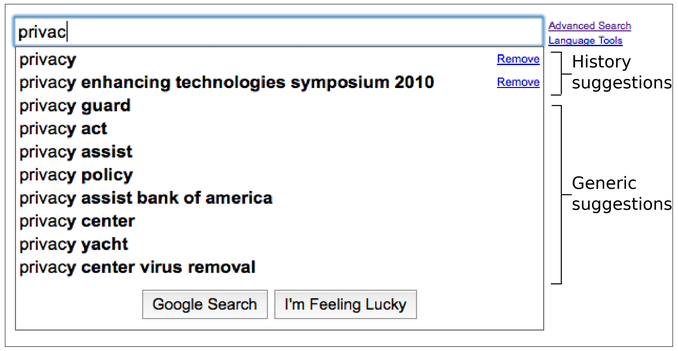


Figure: An example of Google Personalized Suggestions.

Hint: Authentication to Google Search is implemented using HTTP cookies.

The Historiographer

The Historiographer is an inference attack that reconstructs a user's Web History and circumvents Google's stricter access control policy.

1. Capture the HTTP authentication cookie used by the victim accessing Google Search.
2. Replay the cookie to conduct searches while impersonating the victim.
3. Exploit the personalized suggestions to discover searches previously conducted by the victim.
   - Suggestions start to be sent already after 2- or 3-letter keystrokes.
4. Reconstruct the victim's Web History.

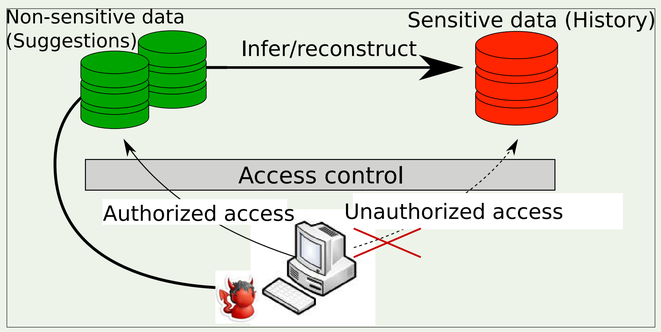


Figure: The Historiographer inference attack.

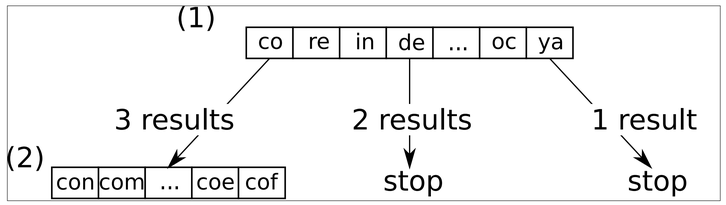
The attack is much more powerful than monitoring the victim's network: it potentially retrieves all searches conducted by the victim, even if done from different networks and different computers.

Google's Countermeasures

After receiving a preliminary report, Google disabled the personalized suggestions for several weeks. Later on, Google started encrypting back-end server requests associated with personalized suggestions. Note, however, that searches conducted from mobile phones are still vulnerable. Additional information and discussion is available on the project page.

Tree Algorithm

- We use a tree-based algorithm to optimally trade-off the attack's accuracy with the number of requests.
- We only simulate the keystrokes of the most frequent 2-letter prefixes.
- Only if 3 suggestions are returned, there may be additional entries in the Web History, thus, we add another letter, as in the example below.



Within a few hundred simulated keystrokes, one can retrieve a conspicuous portion of Web History.

Experimental Results

We ran the Historiographer on 2-weeks *anonymized* traffic from INRIA internal network, a TOR exit node, and on 10 volunteers.

1. **Number of potential victims**. Percentage of users searching on Google while signed-in and with Web History enabled: $\sim 45\%$.
2. **Historiographer's accuracy**. Percentage of entries recovered from the Web History (on volunteers): $\sim 60\%$ with ~ 400 requests.

The Privacy Threats

- Many Google services are still vulnerable to simple session hijacking.
- Several studies have showed that user searches are to be considered highly sensitive: the Historiographer may severely endanger user privacy.
- Data exposed by the Web History may be combined with other publicly available data or location-based services.
- Personalized Services (personalized searches, but also personalized results) leak private information and must be provided over encrypted channels.

Although focused on specific features of Google, we do not aim to attack Google – other search engines, e.g., Bing, also presents similar vulnerabilities. We highlight the general problem of protecting privacy of sensitive data when using mixed architecture with both secure and insecure connections.