Internet Jones and the Raiders of the Lost Trackers

An Archaeological Study of Web Tracking from 1996 to 2016

Adam Lerner*  Anna Kornfeld Simpson*  Tadayoshi Kohno  Franziska Roesner

*Joint First Authors in Alphabetical Order
Third Party Web Tracking

1. Third party web tracking is important
2. We should be studying web tracking over time
What is Third Party Web Tracking?

I know that user "d04874f3" has visited these sites:

- tracker.com
- news.google.com
- weather.com
- twitch.tv
- hulu.com
- imdb.com
- foxnews.com
- zappos.com
- honda.com
- jetblue.com
Being watched has chilling effects

Views and edits of privacy-related Wikipedia articles reduced after leaks of NSA surveillance activities*

People donate more for communal coffee when a picture of watching eyes is posted near the coffee**


Web Tracking

1. Third party web tracking is important
2. We should be studying web tracking over time
We should study web tracking over time

1996
2005
2009
2016

Gomer et al. '13
Roesner, Kohno, and Wetherall '12
Eubank et al. '13

Krishnamurthy and Wills '09
Mayer and Mitchell '12
Englehardt et al. '15
We should study web tracking over time


- Gomer et al. '13
- Roesner, Kohno, and Wetherall '12
- Eubank et al. '13
- Gomer et al. '13
- Roesner, Kohno, and Wetherall '12
- Eubank et al. '13
Time Travel
Time Travel
About USENIX:
The Advanced Computing Systems Professional and Technical Association

About SAGE: The System Administrators Guild

USENIX Workshop on Internet Technologies and Systems Call for Papers
Conference on Domain-Specific Languages Call For Papers
2nd Workshop on Electronic Commerce Program Available
1997 USENIX Annual Technical Conference Program Available
Call for Nominations for the USENIX Lifetime Achievement Award and the Software Tools User Group Award

Executive Office:
USENIX Association, 2560 Ninth Street, Suite 215, Berkeley, CA 94710
Phone: 510 528 8649 FAX: 510 548 5738
Email: office@usenix.org

Conference Office:
USENIX Association, 22672 Lambert Street, Suite 613, Lake Forest, CA 92630
Phone: 714 588 8649 FAX: 714 588 9706
Email: conference@usenix.org

webster@usenix.org (Last modified: 10/25/96)
The Wayback Machine
The Wayback Machine

[Membership] [Events] [Publications] [Directory] [Students]
Opening Remarks:
   Greg Rose, Qualcomm

Keynote Address: A Simple Distributed Security Infrastructure
   Ronald L. Rivest, MIT Laboratory for Computer Science

"He who controls the area controls the religion"-Latin

Chair: Brent Chapman, Great Circle Associates

A Secure Environment for Untrusted Helper Applications
   Ian Goldberg, David Wagner, Randi Thomas and Eric A. Brewer, University of California, Berkeley

A DNS Filter and Switch for Packet-filtering Gateways
   Bill Cheswick, Lucent Technologies; Steven M. Bellovin, AT&T Research

Confining Root Programs with Domain and Type Enforcement
Sixth USENIX Security Symposium,  
Focusing on Applications of Cryptography  

July 22-25, 1996  
San Jose, California  

Wednesday, July 24  

Opening Remarks:  
Greg Rose, Qualcomm  

Keynote Address: A Simple Distributed Security Infrastructure  
Ronald L. Rivest, MIT Laboratory for Computer Science  

"He who controls the area controls the religion"  
-Latin  

Chair: Brent Chapman, Great Circle Associates  

A Secure Environment for Untrusted Helper Applications  
Ian Goldberg, David Wagner, Randi Thomas and Eric A. Brewer, University of California, Berkeley  

A DNS Filter and Switch for Packet-filtering Gateways  
Bill Cheswick, Lucent Technologies; Steven M. Bellovin, AT&T Research  

Confining Root Programs with Domain and Type Enforcement  

~11 petabytes of web sites archived!  

Includes JavaScript! HTTP headers!  
CSS! Images! Hosted on web servers!
The Wayback Machine enables us to perform retrospective web measurements.
This Talk

A 20 year longitudinal study of third party web tracking, performed using archival web data
Abstract
Abstract

Insight: Web archives exist (e.g., Wayback Machine)
Abstract

**Insight:** Web archives exist (e.g., Wayback Machine)

**Challenges:** Web archives are sometimes 1. incomplete & 2. inconsistent
Abstract

Insight: Web archives exist (e.g., Wayback Machine)

Challenges: Web archives are sometimes 1. incomplete & 2. inconsistent

Tool: TrackingExcavator
Abstract

**Insight:** Web archives exist (e.g., Wayback Machine)

**Challenges:** Web archives are sometimes 1. incomplete & 2. inconsistent

**Tool:** TrackingExcavator

**Validation:** Comparison of archival measurements to ground truth
Abstract

Insight: Web archives exist (e.g., Wayback Machine)

Challenges: Web archives are sometimes 1. incomplete & 2. inconsistent

Tool: TrackingExcavator

Validation: Comparison of archival measurements to ground truth

Measurement Results: A picture of web tracking over the last 20 years
How does Third Party Web Tracking Work?

A **first-party** is a domain to which a user goes intentionally, by typing a URL or clicking a link.

A **third-party** is a domain whose content is embedded in a first-party web page.
How does Third Party Web Tracking Work?

A **first-party** is a domain to which a user goes intentionally, by typing a URL or clicking a link.

A **third-party** is a domain whose content is embedded in a first-party web page.

If a third party is able to **link together** a subset of a person's **browsing history**, we call this ability **3rd Party Web Tracking**.
How does Third Party Web Tracking Work?

tracker.com
How does Third Party Web Tracking Work?
How does Third Party Web Tracking Work?
How does Third Party Web Tracking Work?
How does Third Party Web Tracking Work?
How does Third Party Web Tracking Work?
How does Third Party Web Tracking Work?

Browsing profile for user 789:
theonion.com

Set this cookie: id=789
How does Third Party Web Tracking Work?

Browsing profile for user 789:
theonion.com

Set this cookie: id=789

tracker.com:
How does Third Party Web Tracking Work?

Browsing profile for user 789: theonion.com
How does Third Party Web Tracking Work?

Browsing profile for user 789:
theonion.com

tracker.com

cnn.com
How does Third Party Web Tracking Work?

Browsing profile for user 789:
theonion.com

tracker.com

cnn.com
How does Third Party Web Tracking Work?

Browsing profile for user 789:
theonion.com
How does Third Party Web Tracking Work?

Browsing profile for user 789: theonion.com

My cookie is: id=789

tracker.com
How does Third Party Web Tracking Work?

Browsing profile for user 789: theonion.com, cnn.com

My cookie is: id=789
How does Third Party Web Tracking Work?

Browsing profile for user 789: theonion.com, cnn.com

My cookie is: id=789
How does Third Party Web Tracking Work?

Browsing profile for user 789:
theonion.com, cnn.com
Contributions

- **TrackingExcavator**, our Tool
- **Challenges** of Retrospective Measurement
- **Validation** of Archival Data vs. Ground Truth
- **Measurement Results**: A 20 Year Picture of Web Tracking
Contributions

- **TrackingExcavator**, our Tool
- **Challenges** of Retrospective Measurement
- **Validation** of Archival Data vs. Ground Truth
- **Measurement Results**: A 20 Year Picture of Web Tracking
Our Tool, TrackingExcavator, and the Challenges of Retrospective Measurements
TrackingExcavator

(1) Input Generation
- List of top sites (from Alexa or historical sources)
- Timestamp
- Wayback URL Generation Script
- List of archive.org URLs with appropriate timestamp

(2) Data Collection
- Browser (Chrome)
- Browser Extension
- Browse automatically
- Measurement events
- Database
- Prevent Wayback "bubble" escapes

(3) Data Analysis
- Data Analysis Scripts
- Including:
  - Cookie simulation
  - Tracking type detection
  - Timestamp range filtering

(4) Data Visualization
- Graph Creation Scripts
TrackingExcavator

(1) Input Generation
- List of top sites (from Alexa or historical sources)
  - Timestamp
  - Wayback URL Generation Script
  - List of archive.org URLs with appropriate timestamp

(2) Data Collection
- Browser (Chrome)
- Browser Extension
- Prevent Wayback "bubble" escapes
- Browse automatically
- Measurement events

(3) Data Analysis
- Data Analysis Scripts
- Including:
  - Cookie simulation
  - Tracking type detection
  - Timestamp range filtering

(4) Data Visualization
- Graph Creation Scripts

Input Munging
TrackingExcavator

(1) Input Generation
- List of top sites (from Alexa or historical sources)
- Timestamp
- Wayback URL Generation Script
- List of archive.org URLs with appropriate timestamp

(2) Data Collection
- Browser (Chrome)
- Browser Extension
- Data Collection Scripts
- Measurement events
- Prevent Wayback "bubble" escapes

(3) Data Analysis
- Data Analysis Scripts
- Including:
  - Cookie simulation
  - Tracking type detection
  - Timestamp range filtering

(4) Data Visualization
- Graph Creation Scripts

Input Munging

Automatic Browsing
TrackingExcavator

(1) Input Generation
- List of top sites (from Alexa or historical sources)
  - Timestamp
  - Wayback URL Generation Script
  - List of archive.org URLs with appropriate timestamp

(2) Data Collection
- Browser (Chrome)
- Browser Extension
- Prevent Wayback “bubble” escapes
- Browse automatically
- Measurement events
- Database

(3) Data Analysis
- Data Analysis Scripts
  - Including:
    - Cookie simulation
    - Tracking type detection
    - Timestamp range filtering

(4) Data Visualization
- Graph Creation Scripts

Input Munging
Automatic Browsing
Tracking Excavator

(1) Input Generation
- List of top sites (from Alexa or historical sources)
- Timestamp
- Wayback URL Generation Script
- List of archive.org URLs with appropriate timestamp

(2) Data Collection
- Browser (Chrome)
- Browser Extension
- Measurement events

(3) Data Analysis
- Prevent Wayback "bubble" escapes
- Data Analysis Scripts
- Including:
  - Cookie simulation
  - Tracking type detection
  - Timestamp range filtering

(4) Data Visualization
- Graph Creation Scripts

Input Munging

Automatic Browsing

Data Analysis & Visualization
What Makes this Hard?

- Cookies don't work in the archive
- Incomplete archives
- Inconsistent archives ("anachronisms")
Cookies don't work in the archive

Set-cookie: id=789

becomes...

X-Archive-Orig-Set-cookie: id=789
TrackingExcavator

Cookie simulation

Input Munging
Automatic Browsing
Data Analysis & Visualization

(1) Input Generation
- List of top sites (from Alexa or historical sources)
- Timestamp
- Wayback URL Generation Script
- List of archive.org URLs with appropriate timestamp

(2) Data Collection
- Browser (Chrome)
- Browser Extension
- Measurement events
- Prevent Wayback “bubble” escapes

(3) Data Analytics
- Data Analysis Scripts
- Cookie simulation
- Timestamp range filtering

(4) Data Visualization
- Graph Creation Scripts
What Makes this Hard?

- Cookies are weird in archival environments
- Incomplete archives
- Inconsistent archives ("anachronisms")
Incomplete Archives

Set-cookie: id=789
Set-cookie: oatmeal=raisin
Set-cookie: chocolate=chip
http://www.craigslist.com/hou/4.1.19981203211156.038ab8

Latest
Show All

Hrm.

Wayback Machine doesn't have that page archived.
The University of Washington Department of Computer Science & Engineering includes roughly thirty five faculty members, forty staff members, two hundred seventy five graduate majors (150 in the Doctoral program and 125 in the Professional Masters Program), and four hundred twenty five undergraduate majors. Ranked among the top ten programs in the nation, we are active in most of the principal areas of the field, and are engaged in a broad range of interdisciplinary initiatives. We are located in Beautiful Seng Hall, at the center of the UW campus in Seattle, Washington – increasingly prominent as a national and international technology center in software, biotech, the health sciences, and other fields.

See information on the Paul G. Allen Center for Computer Science & Engineering -- a new home for CSE, and the University of Washington's top institutional priority.
The University of Washington Department of Computer Science & Engineering includes roughly thirty five faculty members, forty staff members, two hundred seventy five graduate majors (150 in the Doctoral program and 125 in the Professional Masters Program), and four hundred twenty five undergraduate majors. Ranked among the top ten programs in the nation, we are active in most of the principal areas of the field, and are engaged in a broad range of interdisciplinary initiatives. We are located in Beautiful SIng Hall at the center of the UW campus in Seattle, Washington – increasingly prominent as a national and international technology center in software, biotech, the health sciences, and other fields.

See information on the Paul G. Allen Center for Computer Science & Engineering -- a new home for CSE, and the University of Washington's top institutional priority.
Fortunately, we can still count requests to missing resources!
http://doubleclick.net/search

Latest
Show All

Page cannot be displayed due to robots.txt.

See doubleclick.net robots.txt page. Learn more about robots.txt.
Robots.txt files

http://doubleclick.net/robots.txt:

User-agent: *
Disallow: /search
Allow: /search/about
Disallow: /sdch
Disallow: /groups
Disallow: /index.html?
Disallow: /?
Allow: /?hl=
Disallow: /?hl=*
 Allow: /?hl=*&
Disallow: /?hl=*&gws_rd=ssl$
Disallow: /?hl=*&&gws_rd=ssl$
Allow: /?gws_rd=ssl$
Allow: /?pt1=true$
Disallow: /imgres
...

Robots.txt files

http://doubleclick.net/robots.txt:

User-agent: *
Disallow: /search
Allow: /search/about
Disallow: /sdch
Disallow: /groups
Disallow: /index.html?
Disallow: /?
Allow: /?hl=
Disallow: /?hl=*&
Allow: /?hl=*&gws_rd=ssl$
Disallow: /?hl=*&*&gws_rd=ssl
Allow: /?gws_rd=ssl$
Allow: /?pt1=true$
Disallow: /imgres
...

The Wayback Machine respects these preferences retroactively!
TrackingExcavator

Input Munging

Automatic Browsing

Data Analysis & Visualization

(1) Input Generation
List of top sites (from Alexa or historical sources)
Timestamp
Wayback URL Generation Script
List of archive.org URLs with appropriate timestamp

(2) Data Collection
Browser (Chrome)
Browse automatically
Measurement events

(3) Data Analysis
Data Analysis Scripts
Including:
- Cookie simulation
- Tracking type detection
- Timestamp range filtering

(4) Data Visualization
Graph Creation Scripts
Inconsistent Archives
Palin's path from city hall to governor's mansion

When she played basketball in high school, Sarah Palin, the soon-to-be Republican vice presidential nominee, earned the nickname “Sarah Barracuda” for her fierce competitiveness. In the 21 months she has served as governor of Alaska, no one is suggesting she’s lost her fighting edge. full story

- Palin ready to lead? (Q) | Analysts: Her task
- Palin to call for reform in RNC speech
- Will her gender sway women to Palin?

Recent News

- Poll measures race in three key states
- Dems blast Lieberman, say he lied to delegates
- Ticker: McCain greets Palin's daughter, fiance
- Martin: Keep politics out of Bristol Palin issue
- Biographer: America, meet Sarah Palin
- Report.com: Can a mom of 5 be VP?
- Feffer: Should McCain consider different VP?
- CNNMoney: Auto sales plunge
- Gunman’s bloody trail leads to six bodies
- Caylee's grandma disputing evidence
- Napa: Residents won't be turned away
- CNNMoney: Gulf oil production scrambling back
- Hanna rooms Caribbean, may hit U.S.
- Pakistan protests 'coalition air attack'
- The cartoonist beloved by GIs and regular guys
- 8 dogs die in hot truck during lunch break
- Review: Google’s Chrome needs more polish
- New Marilyn Monroe film scenes found
- Beyonce’s sister aims for path of her own
- CNN Wire: Paraguay coup plot alleged

All news from the past 24 hours: all most popular

Republican National Convention

- Bernstein: Democrats better take note
- Bush, McCain still an uneasy alliance
- Interactive: How conventions work
- McCain arrives in Minnesota for convention
- Key players: Triva, The Forum, In Depth

Secrets to a healthy heart
 Palin’s path from city hall to governor’s mansion

When she played basketball in high school, Sarah Palin, the soon-to-be Republican vice presidential nominee, earned the nickname “Sarah Barracuda” for her fierce competitiveness. In the 21 months she has served as governor of Alaska, no one is suggesting she’s lost her fighting edge. full story

- Palin ready to lead? [OS] | Analyst: Her task
- Palin to call for reform in RNC speech
- Will her gender sway women to Palin?

SECRET TO A HEALTHY HEART

Block non-archival web requests
Input Munging

Automatic Browsing

Data Analysis & Visualization

Timestamp filtering
Contributions

✓ **TrackingExcavator**, our Tool
✓ **Challenges** of Retrospective Measurement
  ● **Validation** of Archival Data vs. Ground Truth
  ● **Measurement Results**: A 20 Year Picture of Web Tracking
Validation of Archival Data
Validation of Archival Data vs. Ground Truth

UW NSDI 2012 paper* (and follow-up measurements) provided a taxonomy of web tracking and ground truth data over the last 5 years.

Archival trends reflect ground truth trends
Contributions

✓ TrackingExcavator, our Tool
✓ Challenges of Retrospective Measurement
✓ Validation of Archival Data vs. Ground Truth

● Measurement Results: A 20 Year Picture of Web Tracking
Measurement Results
Measurement Results

"Web Tracking has Increased Over Time"

1. Number of trackers
2. Type of tracking employed
3. Coverage of top trackers
Measurement Results

"Web Tracking has Increased Over Time"

1. Number of trackers
2. Type of tracking employed
3. Coverage of top trackers
Measurement Results

"Web Tracking has Increased Over Time"

1. Number of trackers
2. Type of tracking employed
3. Coverage of top trackers
The power of individual trackers has grown over time. Individual companies know more about your browsing history than they did before.
Individual companies' choices are important to your privacy

Transparency is important
The "Worst" Days of the Web
The Worst Days of the Web

The "Worst" Days of the Web

Popups enable stronger tracking
The "Worst" Days of the Web
The "Worst" Days of the Web

2004: Internet Explorer starts popup-blocking by default
The "Worst" Days of the Web

30 third-party popups
(top 500 sites)

The choices of browser manufacturers, even those unrelated to tracking, influence the way trackers behave significantly

2003 - 2004
We should study web tracking over time

Gomer et al. '13
Roesner, Kohno, and Wetherall '12
Eubank et al. '13

Krishnamurthy and Wills '09
Mayer and Mitchell '12
Englehardt et al. '15


This Work
Adam, Anna, Yoshi, & Franzi say: "Thanks!"

**Insight**: Web archives exist (Wayback Machine)

**Challenges**: Web Archives are sometimes 1. incomplete & 2. inconsistent

**Tool**: TrackingExcavator

**Validation**: Comparison of archival measurements to ground truth

**Measurement Results**: A 20 year picture of third party web tracking

https://trackingexcavator.cs.washington.edu