In Plain Sight: Online Tracking and Profiling

July 2014

Natasa Milic-Frayling
Principal Researcher
Microsoft Research
Web and third parties

User awareness of online tracking
FIRST PARTY: New York Times
Third parties

FIRST PARTY: New York Times

THIRD PARTIES: providers of content, ads, analytics
Third party tracking

Tracking user visits to first party websites

- Facilitated through co-operation with the first party
  - Websites receive income from a third-party advertising network
- Independently, by exploiting security vulnerabilities
  - cross site scripting (XSS)
  - cross site request forgery (CSRF)

- Web cookies are a common tracking mechanism
- Users can also be tracked by
  - data stored in their web browser cache
  - HTML5 local-storage
  - E-Tag data
  - Flash locally-stored objects (LSOs)
  - the long-lived unique IDs provided by many mobile devices.
Tracking through cookies

User Study: Lifting the Lid on Cookies

Analysis of User Tracking Networks through Search

Propagation of cookies through Social Networks
Microsoft IE

Internet Options

General | Security | Privacy | Content | Connections | Programs | Advanced
---|---|---|---|---|---|---

Settings

Select a setting for the Internet zone.

Medium

- Blocks third-party cookies that do not have a compact privacy policy
- Blocks third-party cookies that save information that can be used to contact you without your explicit consent
- Restricts first-party cookies that save information that can be used to contact you without your implicit consent

Sites | Import | Advanced | Default
---|---|---|---

Location

Never allow websites to request your physical location

Clear Sites

Pop-up Blocker

Turn on Pop-up Blocker

Settings

InPrivate

Disable toolbars and extensions when InPrivate Browsing starts

OK | Cancel | Apply
---|---|---

Firefox

Options

General | Tabs | Content | Applications | Privacy | Security | Sync | Advanced
---|---|---|---|---|---|---|---

Tracking

- Tell sites that I do not want to be tracked
- Tell sites that I want to be tracked
- Do not tell sites anything about my tracking preferences

Learn More

History

Firefox will: Use custom settings for history

- Always use private browsing mode
- Remember my browsing and download history
- Remember search and form history
- Accept cookies from sites
  - Accept third-party cookies: Always
  - Keep until: they expire
  - Clear history when Firefox closes

OK | Cancel | Help
---|---|---

Location Bar

When using the location bar, suggest: History and Bookmarks

Exceptions... | Show Cookies... | Settings...
Microsoft IE

Internet Options

Select a setting for the Internet zone.

Medium

- Blocks third-party cookies that do not have a corresponding privacy policy
- Blocks third-party cookies that save information that will be used to contact you without your explicit consent
- Restricts first-party cookies that save information that can be used to contact you without your explicit consent

Settings

Select a setting for the Internet zone.

Block All Cookies

- Blocks all cookies from all websites
- Cookies that are already on this computer cannot be read by websites

Locations

- Never allow websites to request your personal location
- Location

Pop-up Blocker

- Turn on Pop-up Blocker
- Settings

InPrivate

- Disable toolbars and extensions when InPrivate Browsing starts

OK Cancel
Microsoft IE

Advanced Privacy Settings

You can choose how cookies are handled in the Internet zone. This overrides automatic cookie handling.

Cookies

- Override automatic cookie handling

First-party Cookies
- Accept
- Block
- Prompt
- Always allow session cookies

Third-party Cookies
- Accept
- Block
- Prompt
- Always allow session cookies
Microsoft IE

Internet Options

General | Security | Privacy | Content | Connections | Programs | Advanced

Settings

Select a setting for the Internet zone.

Medium
- Blocks third-party cookies that do not have a compact privacy policy
- Blocks third-party cookies that save information that can be used to contact you without your explicit consent
- Restricts first-party cookies that save information that can be used to contact you without your implicit consent

Sites | Import | Advanced | Default

Location
Never allow websites to request your physical location
- Clear Sites

Pop-up Blocker
- Turn on Pop-up Blocker
- Settings

InPrivate
- Disable toolbars and extensions when InPrivate Browsing starts

OK | Cancel | Apply
Lifting the Lid on Cookies

User Study: Lifting the Lid on Cookies

Research topic:
• Designs to increase the user awareness and understanding of tracking activities

Framework:
• Communication-Human Information Processing (C-HIP) model of warning effectiveness

Research method
• Technical probe: Browser extensions to Comparison of browser extensions.
• Quantitative analysis of the tracking activities.

Cookies and Search

Tracking networks based on the http referral header
Search and Tracking

Observations:
UI features facilitate the value exchange between individuals and services
Value exchange is unclear

Research:
• Data analysis of search results and tracking companies associated with them
• Uncover the characteristics of the tracking network and model the value exchange between the consumer and services
Data set – Search Queries

- **KDD Cup 2005 Challenge**
- **800 queries with assigned categories by three assessors**
- **Categorization involved three level categories:**
  - Top level: Computers (8), Entertainment (9), Information (8), Living (18), Online Community (6), Shopping (6), Sports (11)
  - 67 Second and Third level categories
- **Selected queries with higher label agreement among assessors:** 662 queries

<table>
<thead>
<tr>
<th>Category Label</th>
<th>Num of SearchQueries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shopping\Stores &amp; Products</td>
<td>101</td>
</tr>
<tr>
<td>Information\Local &amp; Regional</td>
<td>95</td>
</tr>
<tr>
<td>Information\Companies &amp; Industries</td>
<td>60</td>
</tr>
<tr>
<td>Living\Health &amp; Fitness</td>
<td>49</td>
</tr>
<tr>
<td>Living\Car &amp; Garage</td>
<td>41</td>
</tr>
<tr>
<td>Information\Law &amp; Politics</td>
<td>40</td>
</tr>
<tr>
<td>Living\Travel &amp; Vacation</td>
<td>39</td>
</tr>
<tr>
<td>Living\Fashion &amp; Apparel</td>
<td>37</td>
</tr>
<tr>
<td>Information\Science &amp; Technology</td>
<td>36</td>
</tr>
<tr>
<td>Living\Finance &amp; Investment</td>
<td>34</td>
</tr>
<tr>
<td>Living\Food &amp; Cooking</td>
<td>33</td>
</tr>
<tr>
<td>Information\Education</td>
<td>30</td>
</tr>
</tbody>
</table>
## Data set – Retrieved Documents

<table>
<thead>
<tr>
<th>Market</th>
<th>Bing API Identifier</th>
<th>Google Search Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>India</td>
<td>en-IN</td>
<td><a href="http://www.google.co.in">www.google.co.in</a></td>
</tr>
<tr>
<td>South Africa</td>
<td>en-ZA</td>
<td><a href="http://www.google.co.za">www.google.co.za</a></td>
</tr>
<tr>
<td>United Kingdom</td>
<td>en-UK</td>
<td><a href="http://www.google.co.uk">www.google.co.uk</a></td>
</tr>
<tr>
<td>United States</td>
<td>en-US</td>
<td><a href="http://www.google.com">www.google.com</a></td>
</tr>
</tbody>
</table>
Tracking network uncovered through Google search in the India Search Market. Shows one giant connected component.
Tracking network uncovered through Google search in the India Search Market.
Clustered.

White circles – Web sites
Dark and light red circles – trackers and ad brokers

White circles – Web sites
Dark and light red circles – trackers and ad brokers
<table>
<thead>
<tr>
<th>Metric</th>
<th>Google</th>
<th>Bing</th>
<th>Baidu</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tracking network</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US</td>
<td>5958</td>
<td>5850</td>
<td>5958</td>
</tr>
<tr>
<td>UK</td>
<td>6171</td>
<td>6638</td>
<td>6171</td>
</tr>
<tr>
<td>S. Af.</td>
<td>5991</td>
<td>5938</td>
<td>5991</td>
</tr>
<tr>
<td>IN</td>
<td>6000</td>
<td>6321</td>
<td>6000</td>
</tr>
<tr>
<td><strong>Nodes N(G)</strong></td>
<td>67739</td>
<td>79214</td>
<td>473</td>
</tr>
<tr>
<td><strong>Edges E(G)</strong></td>
<td>73374</td>
<td>81015</td>
<td>4868</td>
</tr>
<tr>
<td><strong>Unique edges E'(G)</strong></td>
<td>70411</td>
<td>80171</td>
<td>1117</td>
</tr>
<tr>
<td><strong>Clustering coeff.</strong></td>
<td>0.1958</td>
<td>0.2105</td>
<td>0.1685</td>
</tr>
<tr>
<td><strong>Avg. node degree</strong></td>
<td>8.7959</td>
<td>8.8721</td>
<td>4.7230</td>
</tr>
<tr>
<td><strong>Connected comp.</strong></td>
<td>405</td>
<td>358</td>
<td>12</td>
</tr>
<tr>
<td><strong>Giant component</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nodes N(GC) /N(G)</strong></td>
<td>92%</td>
<td>93%</td>
<td>97%</td>
</tr>
<tr>
<td><strong>Edges E'(GC) /E'(G)</strong></td>
<td>99.8%</td>
<td>99.8%</td>
<td>99.6%</td>
</tr>
<tr>
<td></td>
<td>Google</td>
<td>Bing</td>
<td>Baidu</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>----------------</td>
</tr>
<tr>
<td><strong>Tracking network</strong></td>
<td>US</td>
<td>UK</td>
<td>S. Af.</td>
</tr>
<tr>
<td>Nodes N(G)</td>
<td>5958</td>
<td>6171</td>
<td>5991</td>
</tr>
<tr>
<td>Edges E(G)</td>
<td>67739</td>
<td>73374</td>
<td>70411</td>
</tr>
<tr>
<td>Unique edges E’(G)</td>
<td>26203</td>
<td>26552</td>
<td>25763</td>
</tr>
<tr>
<td>Clustering coeff.</td>
<td>0.1958</td>
<td>0.1947</td>
<td>0.1993</td>
</tr>
<tr>
<td>Connected comp.</td>
<td>405</td>
<td>381</td>
<td>398</td>
</tr>
</tbody>
</table>

**Giant component**

<table>
<thead>
<tr>
<th></th>
<th>Google</th>
<th>Bing</th>
<th>Baidu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes N(GC) /N(G)</td>
<td>92%</td>
<td>93%</td>
<td>93%</td>
</tr>
<tr>
<td>Edges E’(GC) /E’(G)</td>
<td>99.8%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
</tbody>
</table>
Small World Property of the Tracking Network

Comparison of the synthetic (Watt-Strogatz random model) and the observed tracking network, based on the average path and the clustering coefficient. Tracking network follows the small world network closely, for the rewiring probability $p=0.2$, (8.91 average node degree)
Data set – Baidu Search Queries

- 10 popular queries are published daily by Baidu
- Collected 98 popular queries
- No categories available
Tracking network uncovered through the Baidu search in China.

White circles – Web sites
Dark and light red circles – trackers and ad brokers
Baidu.com and qq.com cover most of the network.
Probability of Being Exposed to a Tracking Company

All top 10 trackers will be tracking you with:

\[ p > 90\% \] after clicking on 20 search results
\[ p > 99\% \] after viewing 30 search results

Trackers in the search results of Google in the US Search market
Trackers in the search results of Bing in the US Search market
<table>
<thead>
<tr>
<th>Label</th>
<th>Number of Logs</th>
<th>TPs w/ Cookies</th>
<th>TPs w/ No Cookies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shopping\Stores &amp; Products</td>
<td>785</td>
<td>2.79 (3.57)</td>
<td>3.65 (3.37)</td>
</tr>
<tr>
<td>Information\Local &amp; Regional</td>
<td>726</td>
<td>2.16 (4.26)</td>
<td>3.44 (4.27)</td>
</tr>
<tr>
<td>Info\Companies &amp; Industries</td>
<td>459</td>
<td>2.88 (4.02)</td>
<td>3.79 (4.28)</td>
</tr>
<tr>
<td>Living\Health &amp; Fitness</td>
<td>362</td>
<td>2.33 (3.54)</td>
<td>3.42 (3.42)</td>
</tr>
<tr>
<td>Living\Car &amp; Garage</td>
<td>286</td>
<td>3.11 (4.05)</td>
<td>3.84 (3.98)</td>
</tr>
<tr>
<td>Information\Law &amp; Politics</td>
<td>298</td>
<td>0.44 (1.26)</td>
<td>1.23 (1.42)</td>
</tr>
<tr>
<td>Living\Travel &amp; Vacation</td>
<td>301</td>
<td>3.10 (4.85)</td>
<td>3.19 (2.93)</td>
</tr>
<tr>
<td>Living\Fashion &amp; Apparel</td>
<td>289</td>
<td>3.37 (3.87)</td>
<td>3.91 (3.31)</td>
</tr>
<tr>
<td>Information\Science &amp; Tech..</td>
<td>271</td>
<td>1.77 (3.16)</td>
<td>2.07 (2.35)</td>
</tr>
<tr>
<td>Living\Finance &amp; Investment</td>
<td>245</td>
<td>3.08 (3.68)</td>
<td>3.99 (4.28)</td>
</tr>
</tbody>
</table>
Spread of tracking through Twitter

*Tracking network uncovered by analyzing URL sharing in Twitter*
Re-defining the Social Contribution

Observations:
Design of specific social media encourages interaction and rewards specific behaviours. However, the contribution needs to be assessed by considering the entire ecosystem, not the individual service alone.

Research:
• Sharing URLs in Twitter
• How to measure the social contribution when individuals’ actions can affect exposure to tracking of others.
## Twitter DATA

**hashtags.org on 05/01/2013**

### TOPICS
- U.S. Politics
- TV/Entertainment
- Music
- General
- Business
- Tech
- Education
- Environment
- Social Change
- Astrology

**twitaholic.com/top100/followers/ on 05/01/2013**

### TWITTER USERS
- BarackObama
- KimKardashian
- britneyspears
- ladygaga
- BrunoMars
- NICKIMINAJ
- Cristiano
- Oprah
- instagram
- rihanna
- JLo
- shakira
- jtimberlake
- taylorswift13
- justinbieber
- TheEllenShow
- KAKA
- twitter
- katyperry
- YouTube

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total tweets</th>
<th>Original Tweets with URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOPICS</td>
<td>5,364,905</td>
<td>499,228</td>
</tr>
<tr>
<td>TOP USERS</td>
<td>7,914,188</td>
<td>153,029</td>
</tr>
</tbody>
</table>
### URLs Distribution in the Top User Dataset

<table>
<thead>
<tr>
<th>Domain</th>
<th>URLs</th>
<th>%</th>
<th># Third Parties</th>
</tr>
</thead>
<tbody>
<tr>
<td>youtube.com</td>
<td>13,918</td>
<td>18.84%</td>
<td>12</td>
</tr>
<tr>
<td>tumblr.com</td>
<td>8,927</td>
<td>17.02%</td>
<td>246</td>
</tr>
<tr>
<td>youtu.be</td>
<td>5,929</td>
<td>10.15%</td>
<td>1</td>
</tr>
<tr>
<td>instagr.am</td>
<td>5,343</td>
<td>6.40%</td>
<td>0</td>
</tr>
<tr>
<td>peopleschoice.com</td>
<td>1,511</td>
<td>3.25%</td>
<td>37</td>
</tr>
<tr>
<td>twitpic.com</td>
<td>1,493</td>
<td>2.49%</td>
<td>17</td>
</tr>
<tr>
<td>twitlonger.com</td>
<td>1,442</td>
<td>2.41%</td>
<td>23</td>
</tr>
<tr>
<td>twitter.com</td>
<td>1,085</td>
<td>1.89%</td>
<td>27</td>
</tr>
<tr>
<td>facebook.com</td>
<td>896</td>
<td>1.76%</td>
<td>24</td>
</tr>
<tr>
<td>tl.gd</td>
<td>744</td>
<td>1.56%</td>
<td>0</td>
</tr>
</tbody>
</table>

*Most tweeted Web domains, measured by the appearance of URLs in tweets*
## Network Analysis

<table>
<thead>
<tr>
<th>Network</th>
<th>Dataset</th>
<th># Nodes</th>
<th># Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOCIAL</td>
<td>TOPICS</td>
<td>151,624</td>
<td>214,327</td>
</tr>
<tr>
<td>SOCIAL</td>
<td>TOP USERS</td>
<td>286,389</td>
<td>300,697</td>
</tr>
</tbody>
</table>
Connected components, showing a dominant sub-graph of the Twitter community that emerged in the Politics topic.
SOCIAL NETWORK

- edge: re-tweeted
- node size: number of followers
- node colour: number of re-tweets
<table>
<thead>
<tr>
<th>Third Party Domain</th>
<th>% of URLs with Third Parties</th>
<th>Third Party Domain</th>
<th>% of Users who Tweeted</th>
</tr>
</thead>
<tbody>
<tr>
<td>google-analytics.com</td>
<td>65.46%</td>
<td>google-analytics.com</td>
<td>66.19%</td>
</tr>
<tr>
<td>facebook.com</td>
<td>56.14%</td>
<td>doubleclick.net</td>
<td>58.02%</td>
</tr>
<tr>
<td>google.com</td>
<td>53.59%</td>
<td>google.com</td>
<td>57.83%</td>
</tr>
<tr>
<td>twitter.com</td>
<td>50.24%</td>
<td>gstatic.com</td>
<td>57.08%</td>
</tr>
<tr>
<td>gstatic.com</td>
<td>48.58%</td>
<td>googlesyndication.com</td>
<td>51.52%</td>
</tr>
<tr>
<td>chartbeat.net</td>
<td>44.62%</td>
<td>googleadservices.com</td>
<td>48.92%</td>
</tr>
<tr>
<td>chartbeat.com</td>
<td>40.31%</td>
<td>facebook.com</td>
<td>45.18%</td>
</tr>
<tr>
<td>youtube.com</td>
<td>38.32%</td>
<td>googleusercontent.com</td>
<td>43.80%</td>
</tr>
<tr>
<td>doubleclick.net</td>
<td>38.32%</td>
<td>youtube.com</td>
<td>36.24%</td>
</tr>
<tr>
<td>facebook.net</td>
<td>36.17%</td>
<td>youtube-nocookie.com</td>
<td>35.86%</td>
</tr>
</tbody>
</table>

*LEFT:* Percentage of URLs that are associated with, i.e., ‘refer to’ the specific third party

*RIGHT:* Percentage of users that tweeted about Web sites, i.e., URLs who are associated with the specific third parties.
<table>
<thead>
<tr>
<th>Network</th>
<th>Dataset</th>
<th># Nodes</th>
<th># Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SOCIAL</strong></td>
<td>TOPICS</td>
<td>151,624</td>
<td>214,327</td>
</tr>
<tr>
<td></td>
<td>TOP USERS</td>
<td>286,389</td>
<td>300,697</td>
</tr>
<tr>
<td><strong>TRACKING</strong></td>
<td>TOPICS</td>
<td>25,044</td>
<td>174,840</td>
</tr>
<tr>
<td></td>
<td>TOP USERS</td>
<td>10,474</td>
<td>66,609</td>
</tr>
</tbody>
</table>
Business Ecosystem

Ad Demand

ADVERTIZERS

Advertising Agency (ad source)

BIDDERS

CONSUMERS

Third party data (about consumers)

Ad Supply

WEB SITES (PUBLISHERS)

ADS

Tracking cookies

Advertising Network (ad server)

SUPPLIERS

BROKER

Real-time bidding

Personal data disclosure
Business Ecosystem

Ad Demand

ADVERTIZERS

Advertising Agency (ad source)

BIDDERS

Ad Supply

CONSUMERS

WEB SITES (PUBLISHERS)

ADS

WEB SITE

THIRD PARTY ONLY

TRACKER/AD EXCHANGE

DUAL ROLE

Third party data (about consumers)

Advertising Network (ad server)

SUPPLIERS

AD EXCHANGE SERVICE

Real-time bidding

BROKER

Personal data disclosure

Tracking cookies
new slogan

WYSIWYG → WY’nsI WYP

What You Don’t See is What You Pay

• Interaction with applications and services through UI is captured and forms a digital footprint that is used for (behavioural) profiling

• Protocols used to enable communication between the PC and Web services enable device fingerprinting and user tracking.
research investigation

Designs of computing systems lack transparency about the personal data capture and data flow.

**Implications:**

We are unable to make informed decisions and assume responsibility for our own actions, and ensure we do not harm others.

We have been stripped of the ability to determine our own self within the (digital) society.

Information is collected by first and third parties. We have no say and no control over what is collected, to whom and how information is presented, and how it is used.
Thank you

Natasa Milic-Frayling
natasamf@microsoft.com