ISP-Enabled Behavioral Ad Targeting without Deep Packet Inspection

Gabriel Maciá-Fernández\textsuperscript{1}, Yong Wang\textsuperscript{2}, Rafael Rodríguez-Gómez\textsuperscript{1}, Aleksandar Kuzmanovic\textsuperscript{3}

\textsuperscript{1}University of Granada (Spain)
\textsuperscript{2}University of Electronic Science and Technology (China)
\textsuperscript{3}Northwestern University (USA)
“Online advertising is a $20 billion industry that is growing rapidly...”

- Google, Yahoo, AOL dominate the online advertising market
- Online advertisers v.s. ISPs
- ISP started deploying deep packet inspection techniques to track and collect user browsing behavior
Federal Wiretap Act states a simple prohibition: “thou shalt not intercept the contents of communications...Violations can result in civil and criminal penalties”

This prohibition has clearly been violated by deep packet inspection techniques.

Electronic Communications Privacy Act states that any provider can hand-over non-content records to anyone except the government
Our challenge

- Is it possible to recover user browsing patterns only from the limited information provided by TCP headers?
- How accurately?
- How scalable would this approach be?
Our Approach

Profile websites: collect information about web pages from websites

Trace analysis: in a tapping point, extract web browsing communication features from traces

Detection: Correlate the information from the two sources to detect the web pages actually accessed by clients
Website profiling (crawling websites):

- For every web page in a site, we record:
  - Size in bytes (plain/compressed) of root file and all embedded objects
  - Location of objects (internal / external)
  - List of embedded objects
  - List of links
The whole process (II)

Web browsing features analysis from traces:

- Obtain traces in a tapping point
- Filter and separate web traffic from every source IP to any destination
- Estimate the size and the location of the downloaded objects:
  - Web pages delimited by a time threshold: 1 second
  - Downloaded objects delimited by PUSH flag
The whole process (III)

- **Detection algorithm basics:**
  - Find the web page in the website profile that best matches the sizes and locations of the objects detected in the trace

- **Details:**
  - Unique objects or root files lead to direct detection
  - Separate comparison for root files and objects
  - Ambiguities are clarified by selecting pages with:
    - Highest percentage of detected objects
    - Consistent navigation pattern (Link analysis)
Sources of error:

- Estimation of the objects size: cookies, chunk size information
- Dynamic website behavior
- Browsing behavior: pipelining, caching, parallel browsing
- Spurious requests
Experimental evaluation

Experimental setup:
- 6 different websites for web profiling

List of websites (URL)

<table>
<thead>
<tr>
<th>Website</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FC. Barcelona</td>
<td><a href="http://www.fcbarcelona.com">www.fcbarcelona.com</a></td>
</tr>
<tr>
<td>IKEA</td>
<td><a href="http://www.ikea.com">www.ikea.com</a></td>
</tr>
<tr>
<td>Toyota</td>
<td><a href="http://www.toyota.com">www.toyota.com</a></td>
</tr>
<tr>
<td>University 1</td>
<td><a href="http://www.northwestern.edu">www.northwestern.edu</a></td>
</tr>
<tr>
<td>University 2</td>
<td>ceres.ugr.es</td>
</tr>
</tbody>
</table>

- We crawl a subset of 2000 pages for each website
- We generate quasi-random walks on each website with 100 pages and obtain TCP level traces
Experimental evaluation

Site uniqueness results:

- Uniqueness detection is a powerful feature
Experimental evaluation

Basic performance results:

- Mean success rate = 86%
- False positives rate < 5%
Sensibility to outdated profiles or traces:

**Experimental evaluation**

**TRACES ARE OLD**

<table>
<thead>
<tr>
<th>Change Rate [%]</th>
<th>Days</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td>+2</td>
</tr>
<tr>
<td></td>
<td>+3</td>
</tr>
<tr>
<td></td>
<td>+4</td>
</tr>
<tr>
<td></td>
<td>+5</td>
</tr>
<tr>
<td></td>
<td>+6</td>
</tr>
</tbody>
</table>

**PROFILES ARE OLD**

<table>
<thead>
<tr>
<th>Change Rate [%]</th>
<th>Days</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-6</td>
</tr>
<tr>
<td></td>
<td>-5</td>
</tr>
<tr>
<td></td>
<td>-4</td>
</tr>
<tr>
<td></td>
<td>-3</td>
</tr>
<tr>
<td></td>
<td>-2</td>
</tr>
<tr>
<td></td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Experimental evaluation

- Sensibility to outdated profiles or traces:

  **TRACES ARE OLD**

  ![Graph showing traces are old](image)

  **PROFILES ARE OLD**

  ![Graph showing profiles are old](image)
## Experimental evaluation

### Different browsing scenarios:

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Success rates</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pipelining</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disabled</td>
<td>89%</td>
<td>4%</td>
</tr>
<tr>
<td>Enabled</td>
<td>88%</td>
<td>4%</td>
</tr>
<tr>
<td>Cache</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disabled</td>
<td>90%</td>
<td>4%</td>
</tr>
<tr>
<td>Enabled</td>
<td>89%</td>
<td>4%</td>
</tr>
<tr>
<td>Type of navigation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequential</td>
<td>89%</td>
<td>4%</td>
</tr>
<tr>
<td>Parallel-two</td>
<td>74%</td>
<td>7%</td>
</tr>
<tr>
<td>Parallel-four</td>
<td>63%</td>
<td>8%</td>
</tr>
</tbody>
</table>
Experimental evaluation

Scaling the website profile:

- From 2000 to 9200+ web pages crawled in Toyota
- 78% of pages have either unique size objects or unique root files

The success rate reduces from 89% to 81%
The false positives increase from 4% to 8%
Experiments in the wild:

- Logged visited URIs and timestamps for 17 volunteers
- User navigation replayed and traces saved
- Top 41 websites crawled

Success rate of 85%
False positive ratio is 9%
Conclusions

- We are able to recover web browsing patterns **without inspecting payload**
- Our detection algorithm achieves detection rates around 86% with **low false positives** (<5%)
- The methodology is also **scalable** and **resilient** to a wide number of error sources: outdated information (profiles or traces), pipelining, caching, different types of navigation, etc.
Thank you for your attention