



UNIVERSITY
OF
GRANADA

# ISP-Enabled Behavioral Ad Targeting without Deep Packet Inspection

Gabriel Maciá-Fernández[1], Yong Wang[2],

Rafael Rodríguez-Gómez[1], Aleksandar Kuzmanovic[3]

[1]University of Granada (Spain)

[2]University of Electronic Science and Technology (China)

[3]Northwestern University (USA)

NORTHWESTERN UNIVERSITY

UNIVERSITY OF GRANADA

# Background

"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

# Background

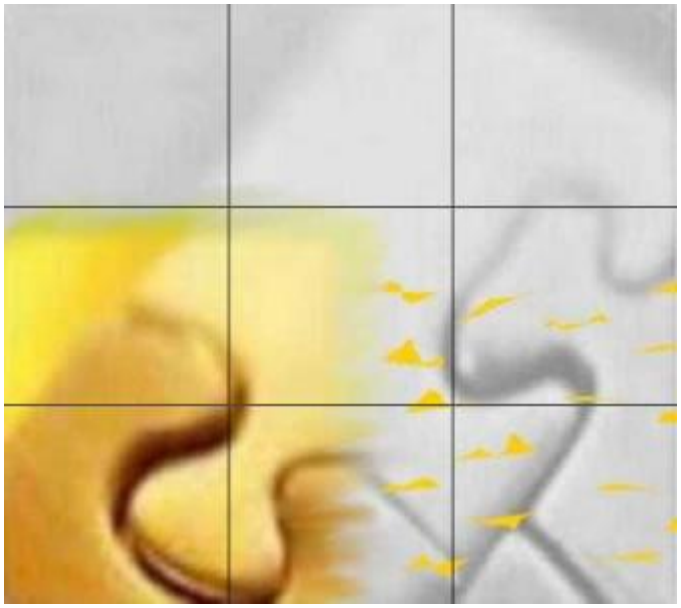
- **Federal Wiretap Act** states a simple prohibition: "*thou shalt not intercept the contents of communications…Violations can result in civil and criminal penalties*"
- This prohibition has clearly been violated by deep packet inspection techniques.
- **Electronic Communications Privacy Act** states that any provider can hand-over non-content records to anyone except the government

NORTHWESTERN UNIVERSITY

UNIVERSITY OF GRANADA

# 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?



NORTHWESTERN UNIVERSITY

UNIVERSITY OF GRANADA

# 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

# The whole process (I)

- **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:
    - Higuest percentage of detected objects
    - Consistent navigation pattern (Link analysis)

# The whole process (IV)

- 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

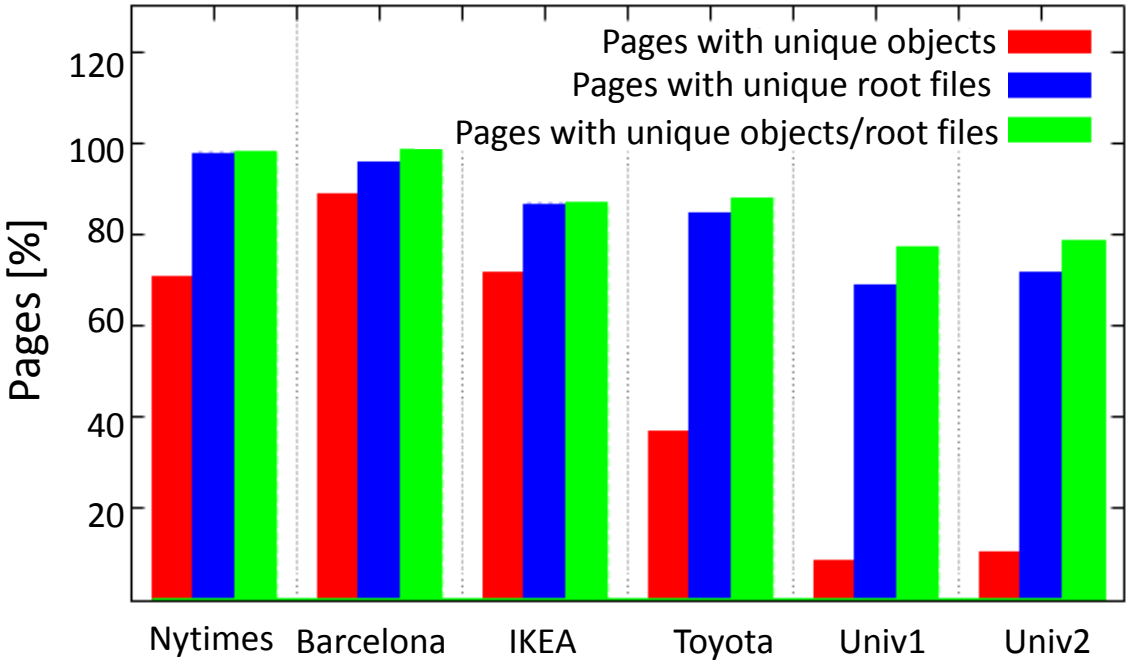
- Experimental setup:
  - 6 different websites for web profiling

| List of websites (URL) |
|---|
| New York Times (www.nytimes.com) |
| FC. Barcelona (www.fcbarcelona.com) |
| IKEA (www.ikea.com) |
| Toyota (www.toyota.com) |
| University 1 (www.northwestern.edu) |
| University 2 (ceres.ugr.es) |

  - 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

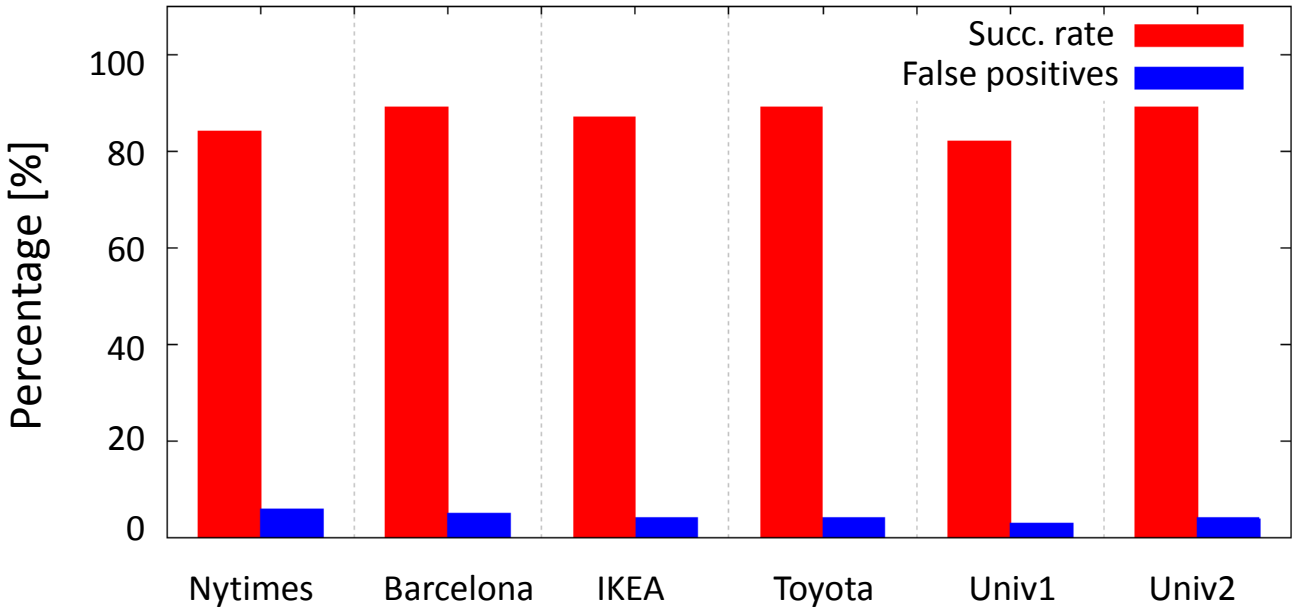
- Site uniqueness results:




  - Uniqueness detection is a powerful feature

# Experimental evaluation

- Basic performance results:




  - Mean success rate = 86%
  - False positives rate < 5%

NORTHWESTERN UNIVERSITY

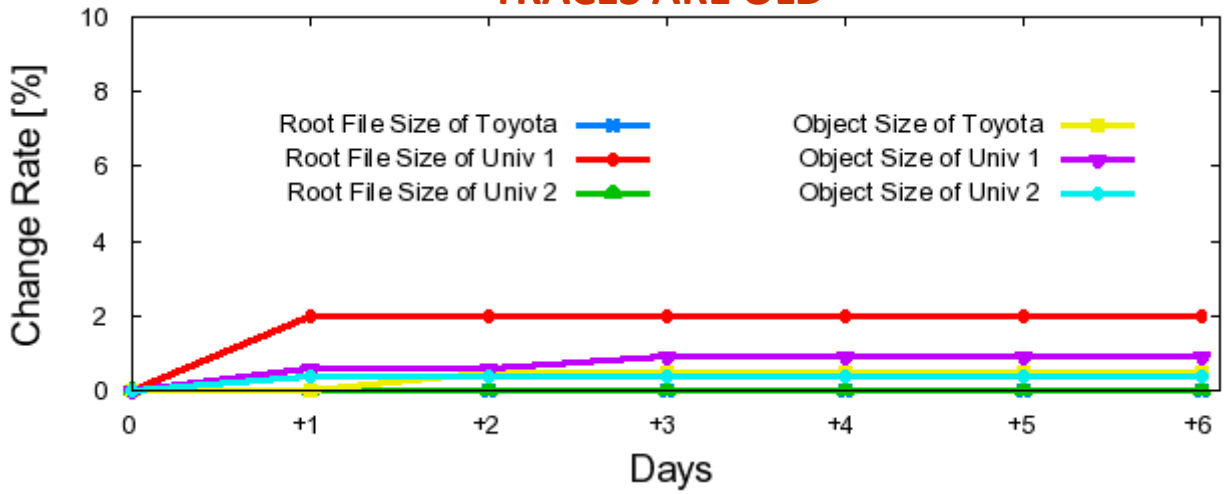
UESTC 1956

UNIVERSITY OF GRANADA

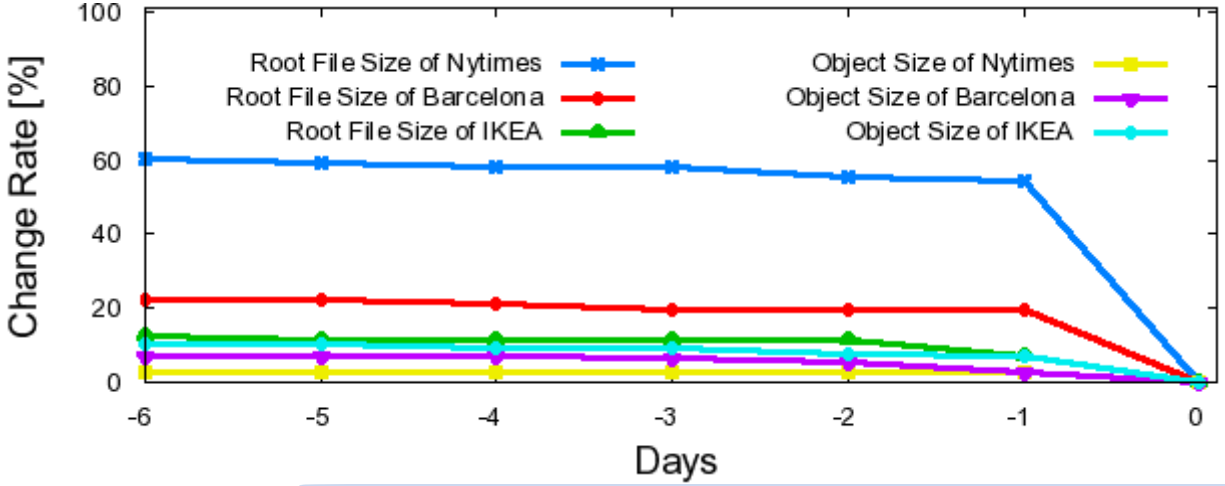
# Experimental evaluation

- Sensibility to **outdated profiles or traces**:

**TRACES ARE OLD**




**PROFILES ARE OLD**



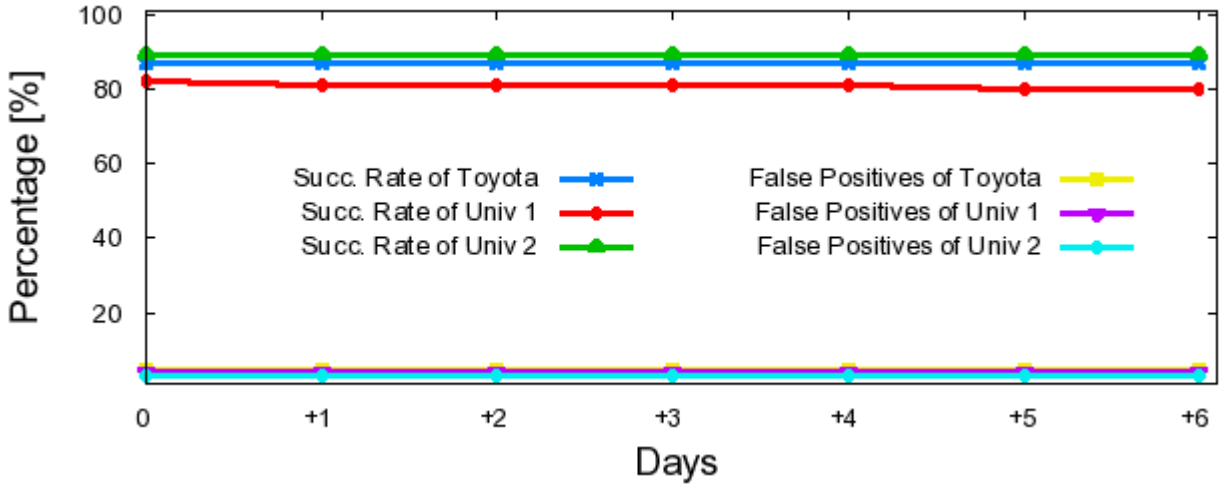
NORTHWESTERN UNIVERSITY

UNIVERSITY OF GRANADA

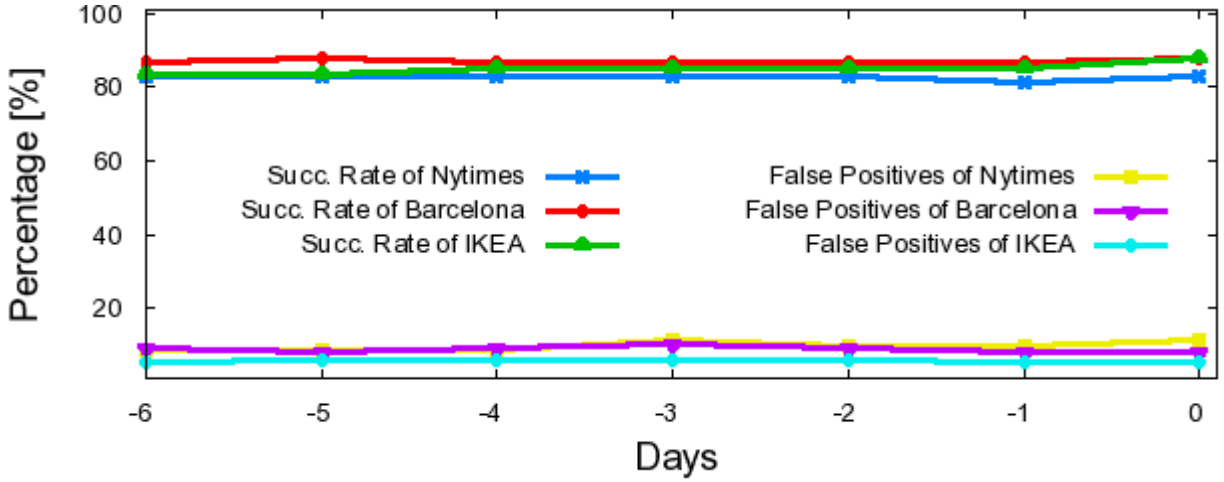
# Experimental evaluation

- Sensibility to **outdated profiles or traces**:

**TRACES ARE OLD**




**PROFILES ARE OLD**



# Experimental evaluation

- Different **browsing scenarios**:

| Scenario | | Success rates | False positives |
|---|---|---|---|
| Pipelining | Disabled | 89% | 4% |
| | Enabled | 88% | 4% |
| Cache | Disabled | 90% | 4% |
| | Enabled | 89% | 4% |
| Type of navigation | Sequential | 89% | 4% |
| | Parallel-two | 74% | 7% |
| | Parallel-four | 63% | 8% |

# 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%

# Experimental evaluation

- 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.



UNIVERSITY
OF
GRANADA

# Thank you for your attention