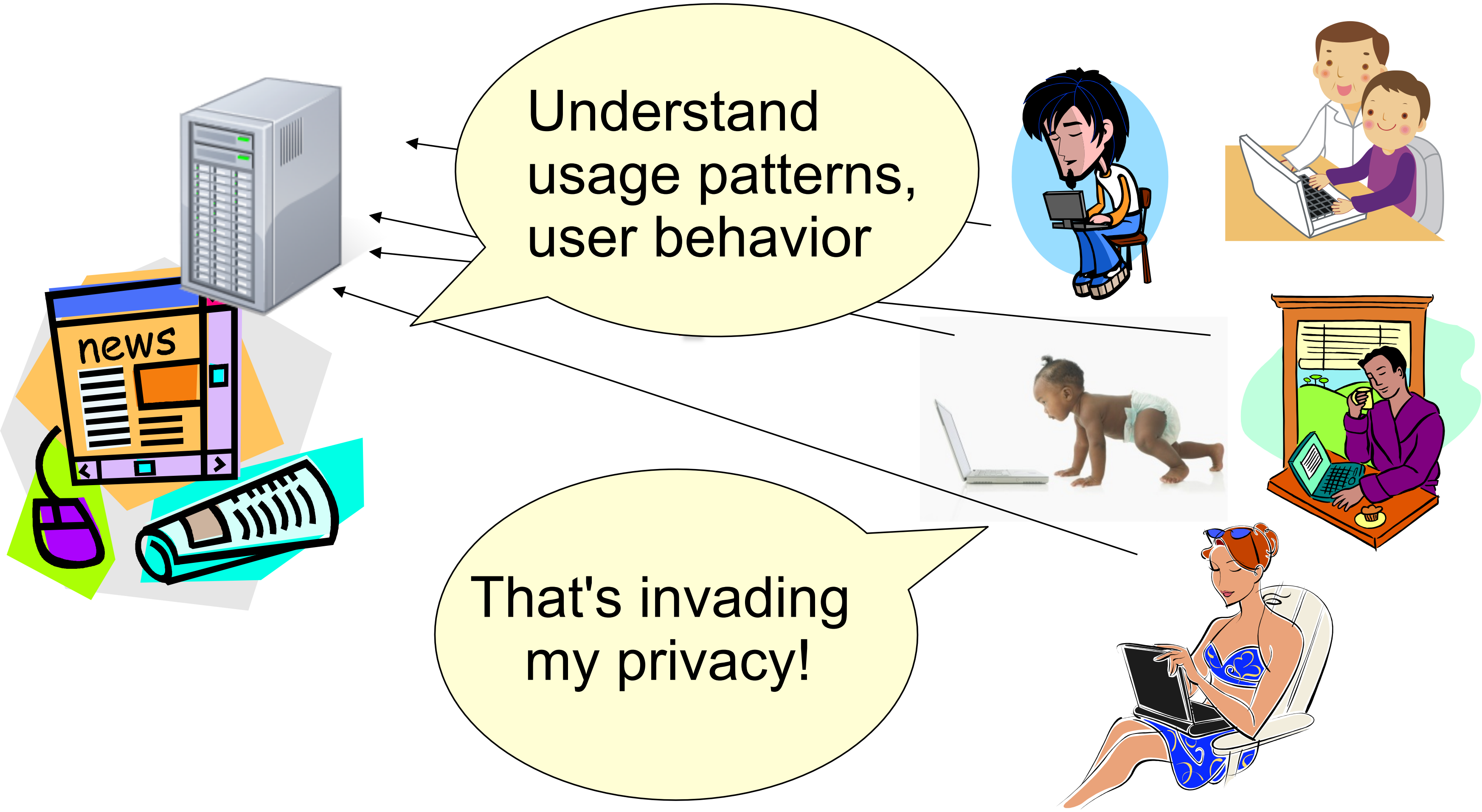
Host Fingerprinting and Tracking on the Web: Privacy and Security Implications

Ting-Fang Yen, *RSA Labs*
Yinglian Xie, Fang Yu, Martin Abadi, *Microsoft Research*
Roger Peng Yu, *Microsoft Corporation*

February 8, 2012

Host-Tracking on the Web



Understand usage patterns, user behavior

That's invading my privacy!

Motivation

- Previous work
  - More elaborate tracking techniques [Eckersley '10, Mayer '09, Kohno et al.'05]
  - Qualitative studies [Krishnamurthy et al.'08,'10]
- How effective are existing approaches? What are the associated privacy risks?

Goals

- Quantify host-tracking information revealed by common identifiers
  - Browser user-agent string (UA)
    - e.g., Mozilla/4.0 (compatible; MSIE6.0; WindowsNT5.1; SV1)
  - IP address
  - Browser cookie
  - User login ID
- Implications of host-tracking
  - Cookie churn study
  - Host mobility study

Data Sources

- Month-long anonymized logs from August 2010
  - Hotmail login events
  - Bing search queries
  - Windows Update logs

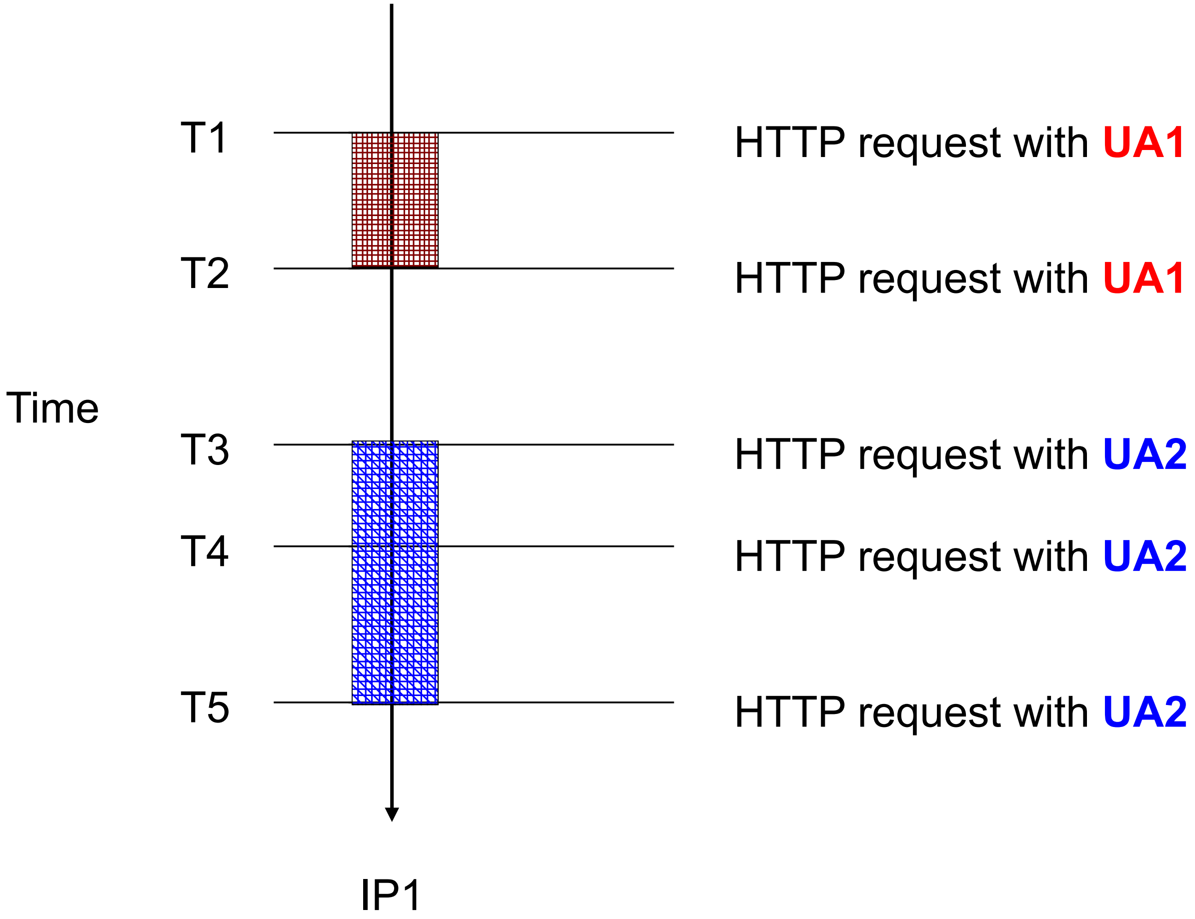
Fingerprints

Validation

Dataset	User-agent info	IP address	Time-stamp	ID	Unique IPs
Hotmail	OS,Browser type	Yes	Yes	User ID	308 Million
Bing	User-agent string (UA)	Yes	Yes	Cookie ID	131 Million
Windows Update	N/A	Yes	Yes	Hardware ID	74 Million

Methodology

- Create “binding windows” for each fingerprint



T1 HTTP request with **UA1**

T2 HTTP request with **UA1**

Time

T3 HTTP request with **UA2**

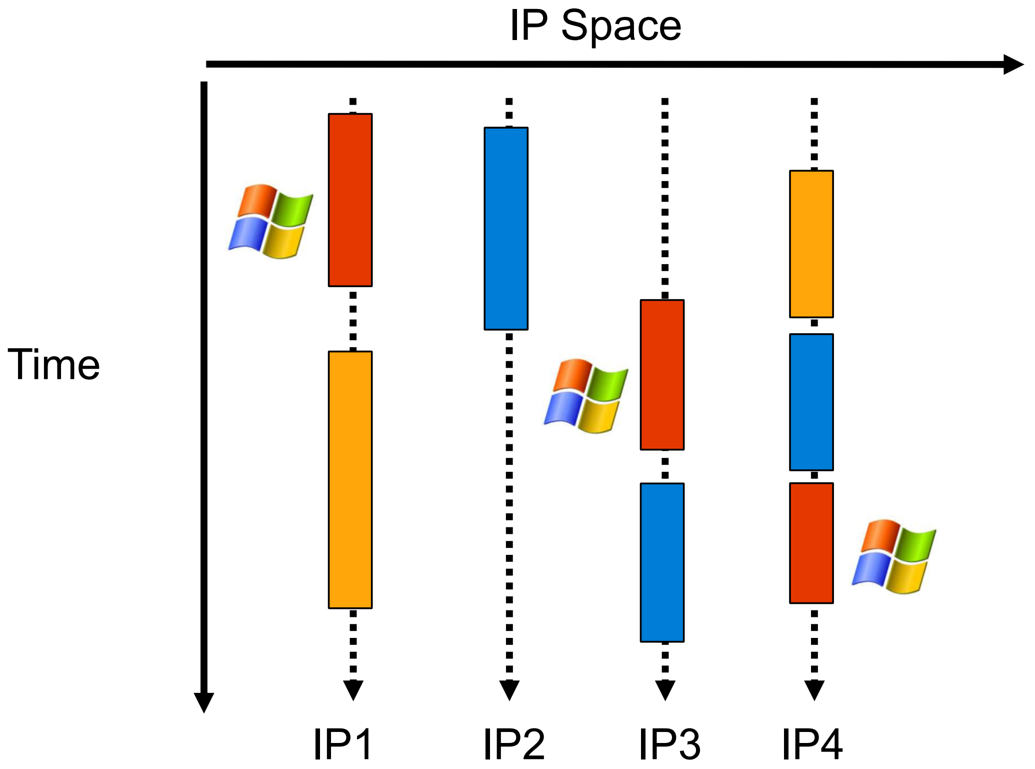
T4 HTTP request with **UA2**

T5 HTTP request with **UA2**

IP1

Methodology (cont'd)

- Construct host-tracking graph
- Validate with Windows Update logs



Metric

- Precision
  - Percentage of fingerprints corresponding to one hardware ID
- Recall
  - Percentage of hardware IDs corresponding to one fingerprint

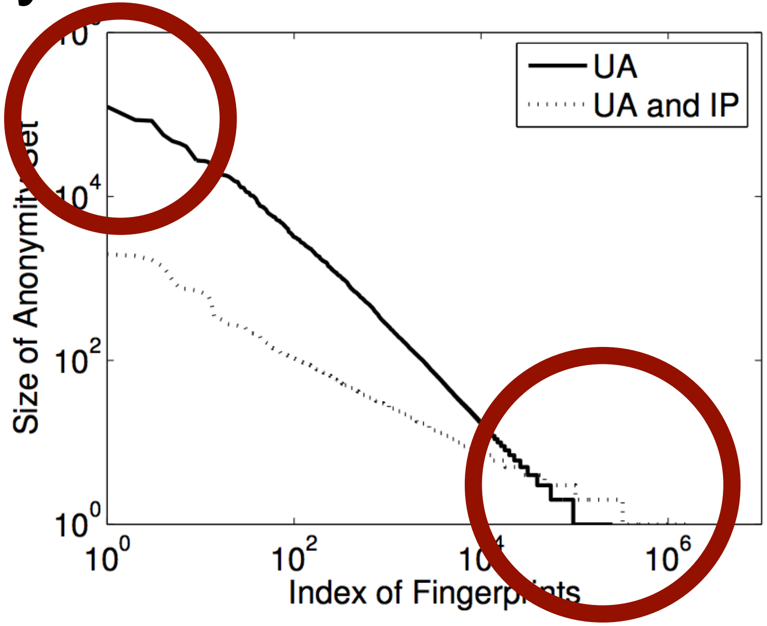
Host-Tracking Results

Identifiers	Precision (%)
User-agent string (UA)	62.01%
UA, IP address	80.62%
UA, /24 IP *prefix*	79.33%
Browser cookie	82.35%
User login ID	92.82%

- Common identifiers can track hosts well, particularly in combination
- Prefix-preserving anonymization is not enough

Host-Tracking Results (cont'd)

- Browser anonymity set



- Entropy
  - UA: **11.59** bits
  - UA+IP: **20.29** bits
  - Installed browser plug-ins, screen resolution, timezone, system fonts, and user-agent strings [Eckersley et al.'10]: **18.1** bits

Application: Cookie Churn Study

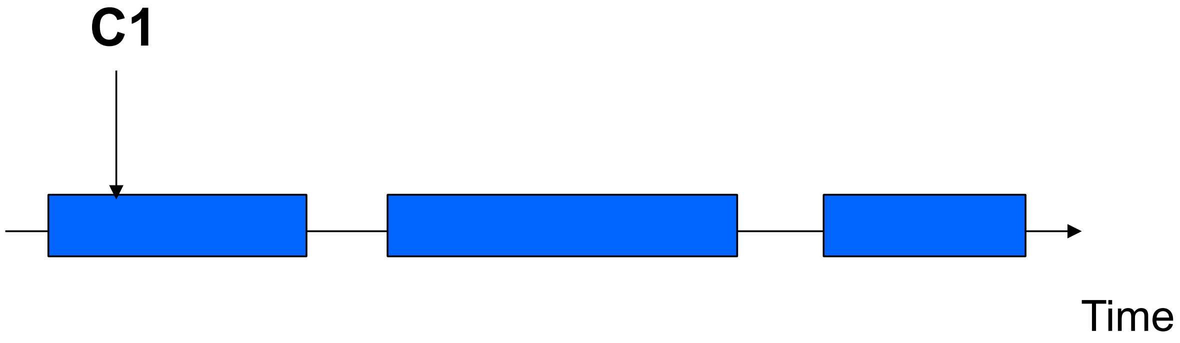
- Cookie IDs are unreliable
- 82% new cookie IDs never returned within the month!
- Apply host-tracking results

: Identify returning clients

: Learn caveats of clearing cookies

Cookie Churn Study

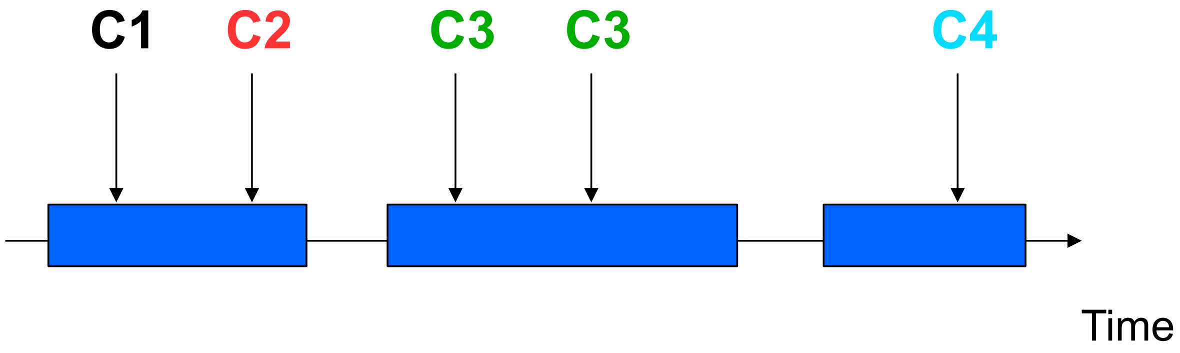
- Overlap HTTP requests with host-tracking graph
- For bindings associated with a user ID...



- Hypothesis: User left service

Cookie Churn Study

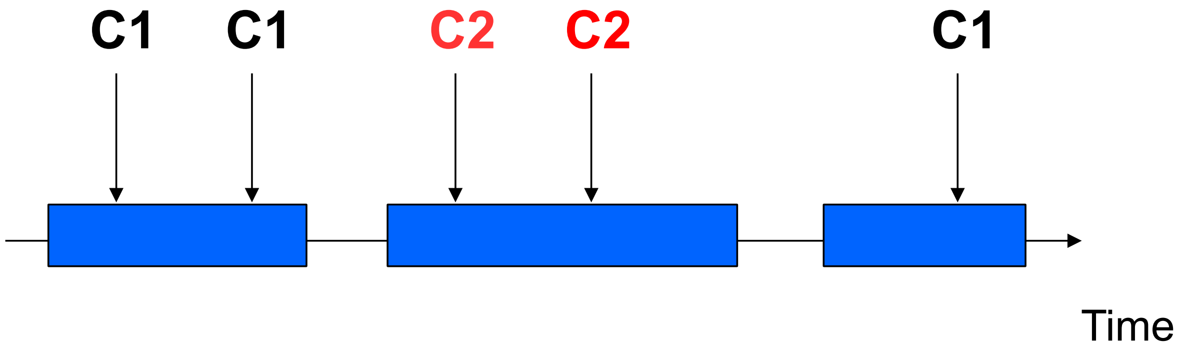
- For bindings associated with a user ID...



- Hypothesis: User clears cookies

Cookie Churn Study

- For bindings associated with a user ID...



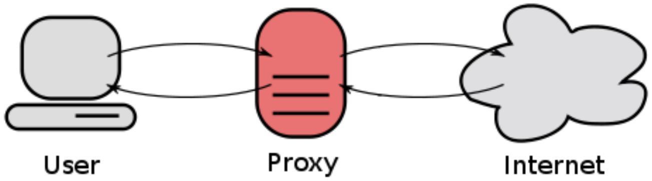
- Hypothesis:
  - Same UA → Private browsing modes
  - Different UA → Multiple browsers, or NAT/proxy

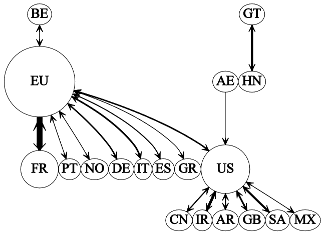
Cookie Churn Results

- 88% one-time cookie IDs are returning users
- 33% users likely clear cookies or utilize private-browsing modes
- **Lesson: Clearing cookies may not be enough**
  - **Utilize proxies or NATs, private browsing, and modify default UA string**

Application: Host Mobility Patterns

- What are the general host mobility patterns?

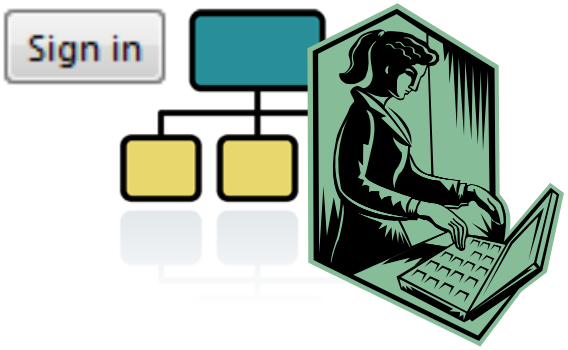


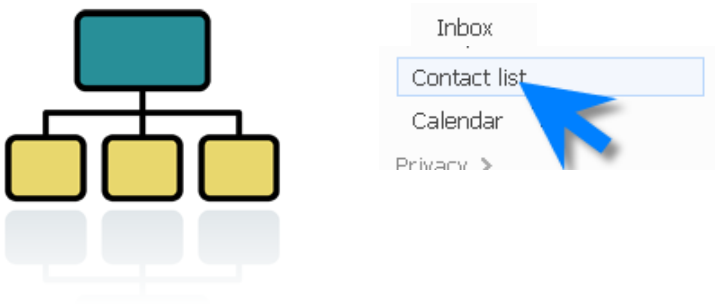


- Anomalous activities outside the norm?
  - e.g., anonymous routing

Detecting Cookie-Forwarding Attacks

- Suspicious activities in Hotmail





- Cannot be explained by general mobility patterns
  - Uni-directional movement
  - Src/Dest domains different from general host mobility
  - No geographic locality

Cookie-Forwarding Bot Users

- One IP address logging in for multiple users, who then appear from 9 network domains
- Over 75,000 such user accounts
- Attackers avoiding spam-detection?

Conclusion

- Large-scale, quantitative study on host-tracking using common identifiers
- Privacy and security implications:
    - Clearing cookies may not be enough –- should also modify default UA string, utilize proxies/NATs, private browsing, anonymous routing
    - Aggregated information can detect malicious events