Tracking Internet Hosts Using Unreliable IDs

Yinglian Xie, Fang Yu, and Martín Abadi
Microsoft Research, Silicon Valley

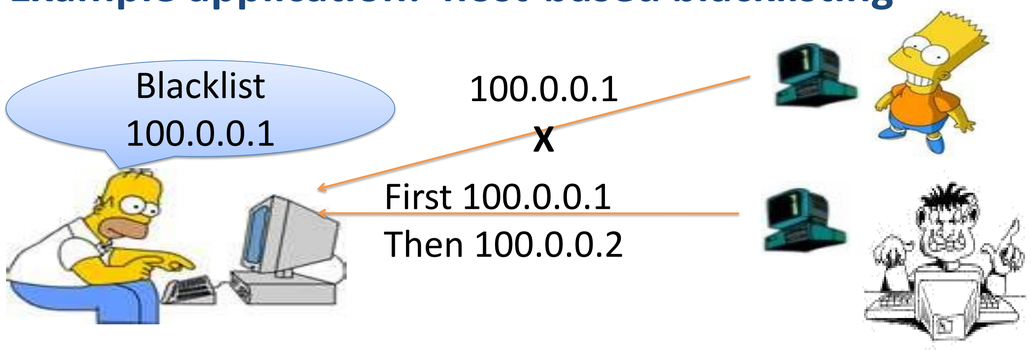
Accountability is weak on the Internet

Open and anonymous ➔ weak accountability

- IP addresses are not unique, fixed identifiers (because of dynamic IP addresses, proxies, and NATs).

➔ It is hard to identify who is responsible for traffic.

Example application: host-based blacklisting

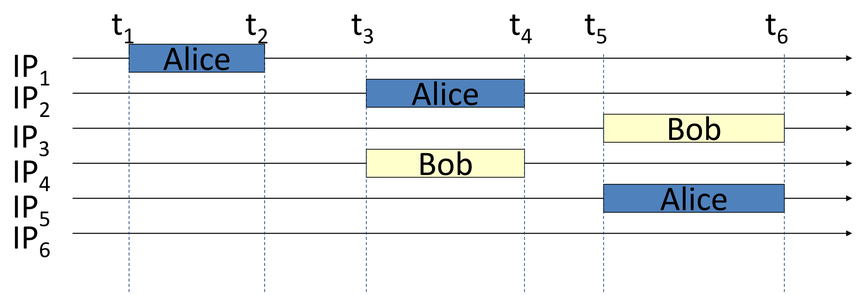


We should block attack traffic by host rather than by IP address

Goal: Inferring host-IP bindings

- **Track hosts more reliably in spite of dynamic IP addresses**
- **Explore applications that require identifying hosts over time**
  - **E.g., security and data-mining**

Our approach: Use application IDs and events to track hosts



Alice's host
IP_1: [t1, t2]
IP_2: [t3, t4]
IP_5: [t5, t6]

Bob's host
IP_4: [t3, t4]
IP_3: [t5, t6]

Challenges:

- No 1-1 mappings between hosts and IDs
- Dynamic IP addresses, proxies and NATs
- Malicious IDs

Problem formulation

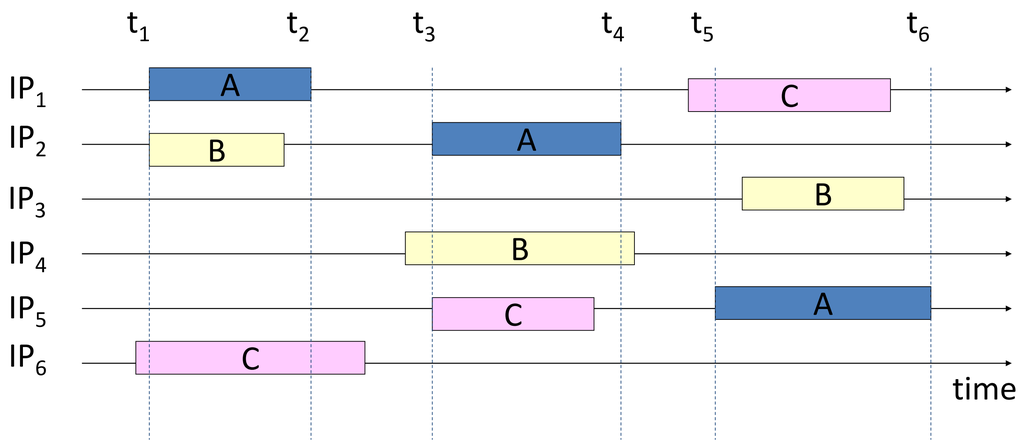
- **Host-tracking graph**

Input events

e_1: <u_1, IP_1, t_1>
e_2: <u_2, IP_1, t_2>
e_3: <u_3, IP_4, t_3>
… …
e_n: <u_4, IP_6, t_5>

Identity mapping

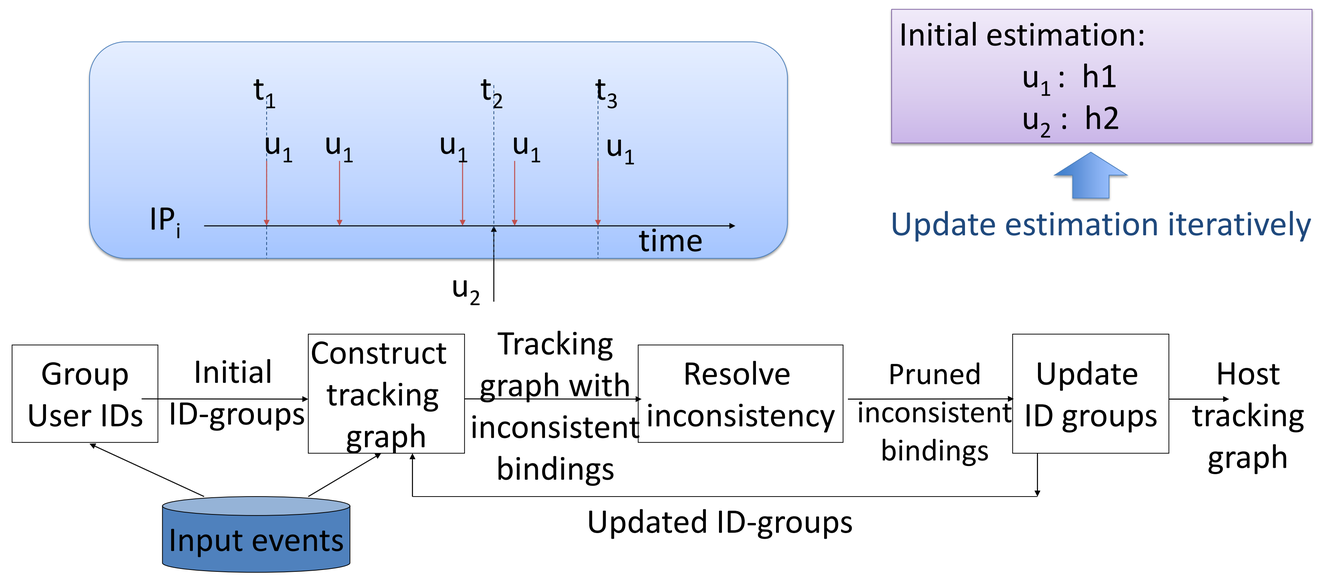
A ⟷ u_1, u_2
B ⟷ u_3
C ⟷ u_4
… …



Host-tracking graph

Our goal : maximize tracked events and tracked IDs

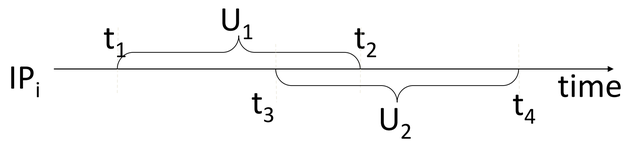
Methodology overview



Grouping IPs

- Consider the probability of two random IDs appearing together

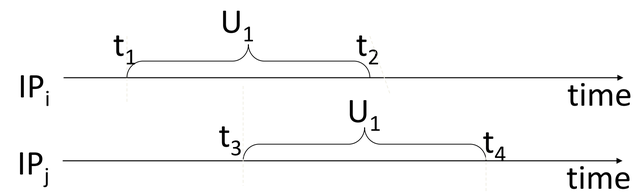
Resolve inconsistencies



Conflict bindings

→Proxy identification

→ Guest removal

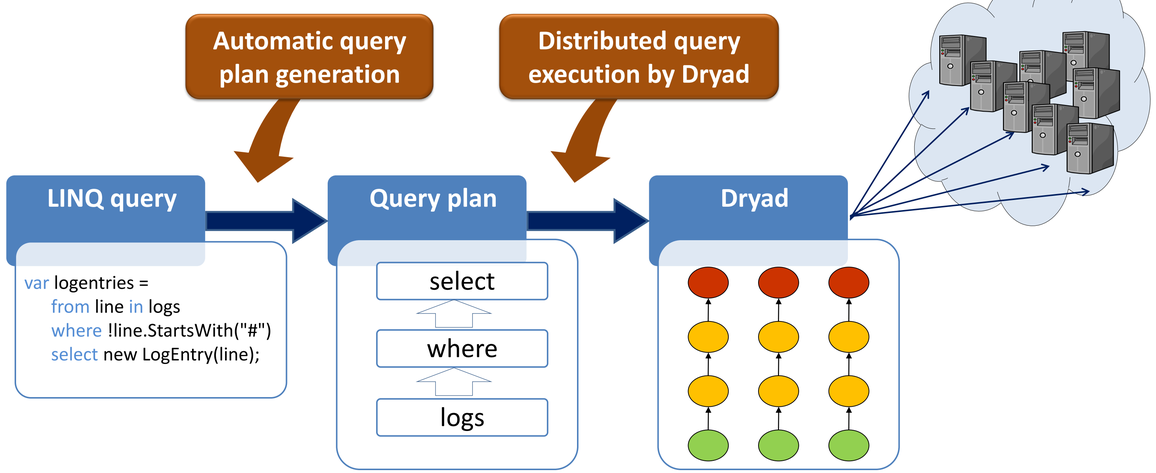


Concurrent bindings

→ Group splitting

Implementation with Hotmail data

On Dryad and DryadLINQ



Applications

- ✓ Calibrate cookie churns
- ✓ Estimate host population more accurately
- ✓ Build normal-user profiles
- ✓ Host-aware blacklists – block by **host** rather than **IP**
  - ❑ Post-mortem forensics
  - ❑ Real-time blacklists (Tracklist)

Coverage and Accuracy

Coverage: tracked events / total events

- ❑ 75% - 80%

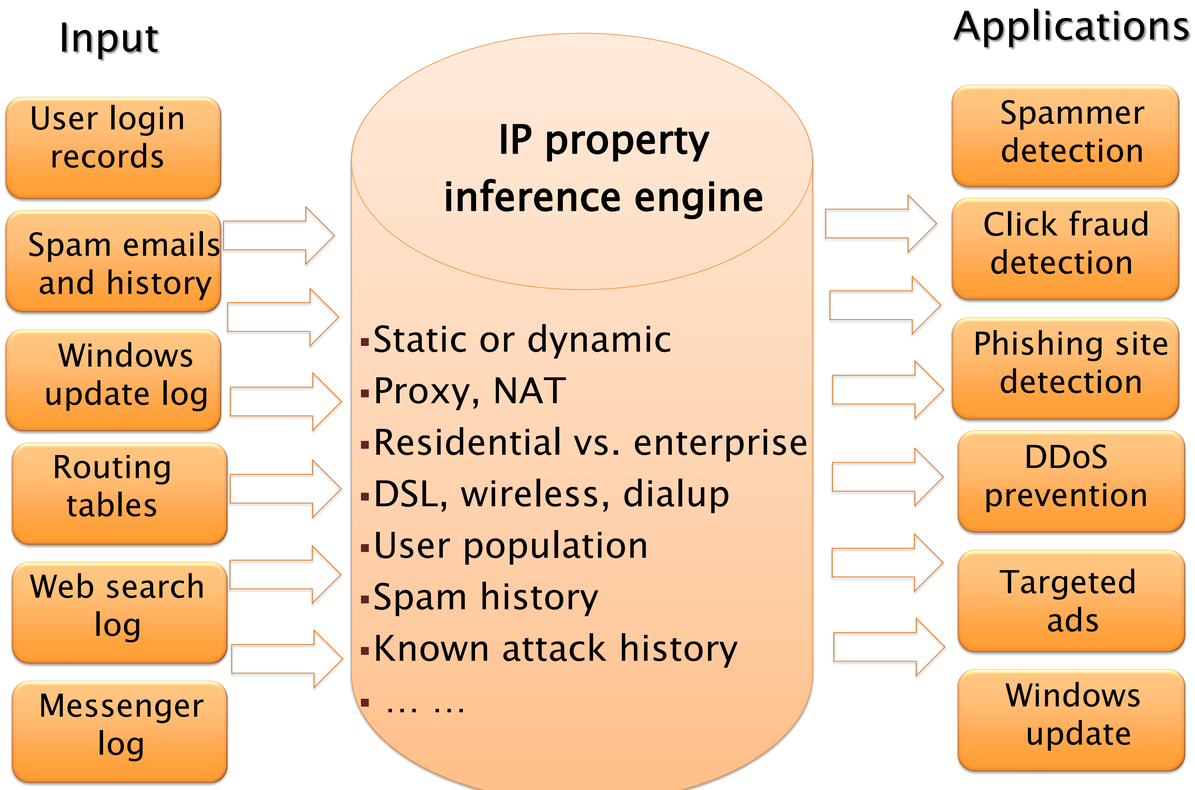
Evaluate accuracy using Windows update data

- ❑ 92% - 96%

	# of blocked users	False positives
IP blacklist / infinitely	44.70 million	52.8%
IP blacklist / one hour	27.94 million	34.1%
Tracklist / one hour	16.01 million	4.9%
Tracklist with profile/one hour	14.27 million	0.1%

Seed data: 5.6 million bot-accounts detected by BotGraph in one month

Network security with IP intelligence



Collaboration with WLSP (Jason Atlas, Geoff Hulten, Ivan Osipkov), Hotmail (Hersh Dangayach, Eliot Gillum, Krish Vitaldevara), Bing (Fritz Behr, David Soukal, Roger Yu, Zijian Zheng), and Messenger (Steve Miale)

Our work