FLOWTRAQ

Identifying Network Users Using Flow-Based Behavioral Fingerprinting

Barsamian, Berk, Murphy

Presented to FloCon 2013

What Is A User Fingerprint?

- Users settle into unique patterns of behavior according to their tasks and interests

- If a particular behavior seems to be unique to one user…
  … and that behavior is observed…
    … can we assume that the original user was observed?

- Affected by population size, organization mission, and the people themselves

Why Fingerprint?

- Basic Research
- Policy Violations and Advanced Security Warning
- Automated Census and Classification

Why Fingerprint?

- Basic Research
  - Change Detection
  - Population Analysis
- Policy Violations and Advance Warning
  - Preliminary heads-up of botnet activity
  - Identify misuse of credentials
- Automated Census and Classification
  - Passive network inventory
  - User count estimation (despite multiple devices)
  - Determination of roles

Background

- Passive and active static fingerprints
  - Operating system identification
    - p0f/NetworkMiner, Nmap
  - Signature-based detection of worms and intrusions
- Dynamic fingerprints
  - Hardware identification
  - Unauthorized device detection[1]
  - Browser fingerprinting[2]
- Increasingly important part of security systems[3]
  - Reinforcing authentication
  - Identifying policy violations

[1] Bratus, et al "Active Behavioral Fingerprinting of Wireless Devices", 2008
[2] http://panopticlick.eff.org
[3] François, et al "Enforcing Security with Behavioral Fingerprinting", 2011

But…

- Difficult to implement, requiring significant expertise not available to many IT departments
- Require unusual or unavailable data
  - Data collection incurs overhead; easier to justify if data is useful for multiple purposes
    - No unitaskers in my shop!
  - Protocol analysis needed
    - Computationally expensive
    - Impinges user privacy
    - Increasingly defeated by encrypted channels and tunnels

Challenge

Make active, adaptive fingerprinting available to the widest possible set of network administrators

- Data requirements
  - Common data source, common data fields
- Processing requirements
  - Can't require major computing resources to create and handle
- Ease of implementation
  - Not just technology, but policy
  - Could search emails and web forms for personally-identifying statistically improbable phrases, but would never fly at most institutions

Why NetFlow Fingerprints?

- NetFlow has very attractive properties to an analyst…
  - Privacy
    - Unintrusive to end users
    - Not affected by encrypted channels
  - Speed
    - Easily-parsed datagrams with fixed fields
    - Bulk of processing taken care of by specialty equipment
  - Scalability
    - Less affected by volume than protocol analyzers
- … but is it up to the task?
  - (Spoiler alert: yes)

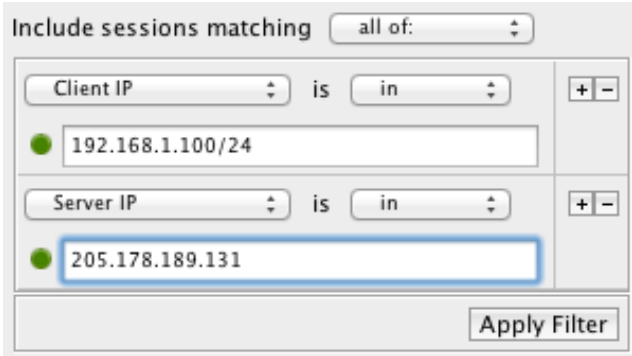
Methodology

After multiple revisions, arrived at the following:

1. Define your parameters
2. Get a list of all the outgoing sessions from that subnet
    1. List of sessions for which client IP is in CIDR block of interest
    2. From that list, extract the destination addresses
3. For each of those destination addresses, do a 'ip-pair' query: (CLNIP==classC && SRVIP=dest).
    1. Count the unique local addresses for each destination
4. Eliminate all of the external addresses that get contacted by more than 1 local address
5. Result is a set of external addresses that are only contacted by ONE client

(CLNIP==*classC*)

(CLNIP==*classC* && SRVIP=*dest*)



FLOWTRAQ

Example Fingerprints

- Individual fingerprints for a user (when that user has one) contain a list of IP addresses that user (and only that user) contacted within the time period
  - One-time connections not included here
- Using the Class C block for the server would compress fingerprints like User B's
  - In this case, would still be unique

User A	8475 total sessions
aaa.93.185.143	38
bbb.175.78.11	44
ccc.22.176.46	42
ddd.28.187.143	37

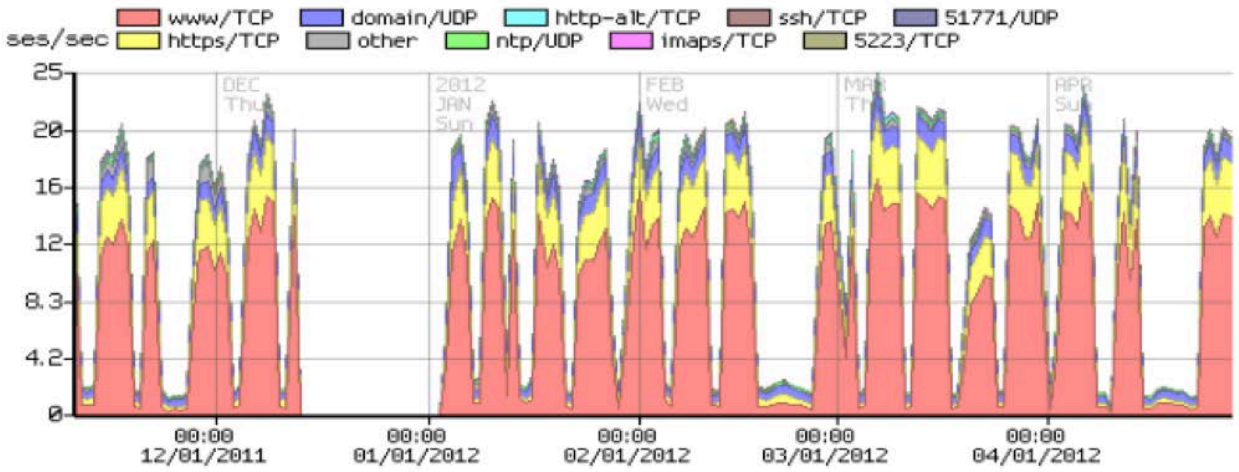
User B	661 total sessions
eee.87.169.51	93
eee.87.160.30	34
eee.87.169.50	37

Parameters

- Definition of local network
  - Select the smallest network of interest
  - May be worth fingerprinting wired and wireless networks separately, to account for users with both desktops and wireless devices
- Time frame
  - Shorter-term profiles faster to create
  - Longer-term profiles less transitory
- Destination subnet
  - When filtering on each destination, using a slightly wider subnet can reduce the computing impact of content distribution networks
- Top N vs. All
  - Cutting off the list of servers with very few sessions improves scalability
  - Potential reduced fingerprint list

Data Source Characterization

- Knowing your source helps determine optimal parameters
- Educational environment with a mix of wireless and wired infrastructure
- Inherent “life spans” to fingerprints
  - Large turnover each year
  - “Mission” changes every term
  - Gaps in data (scheduled breaks) confound ability to detect gradual change

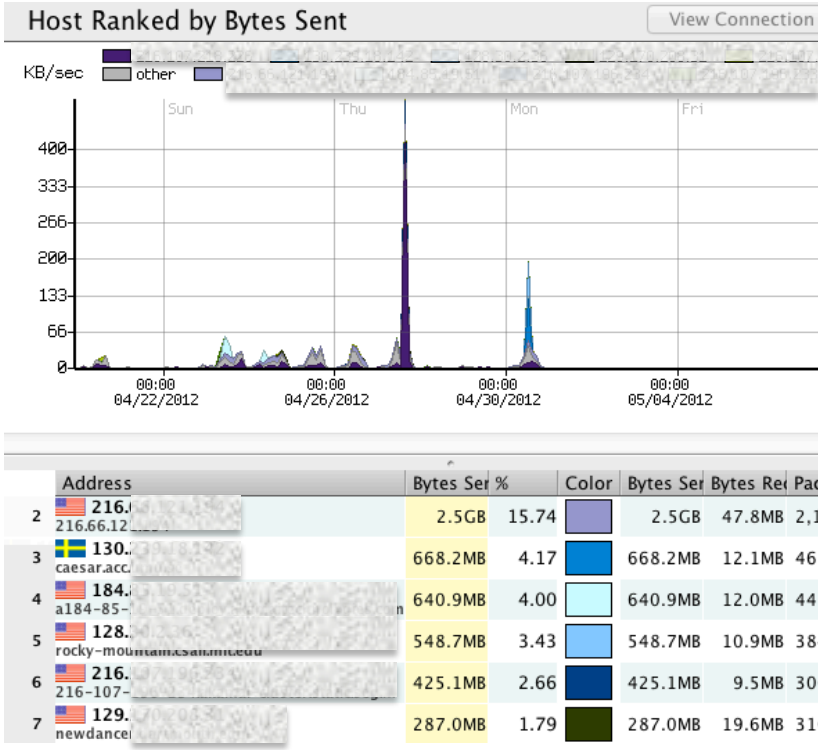


FLOWTRAQ

Select Outbound Requests

- Get a list of top servers by destination

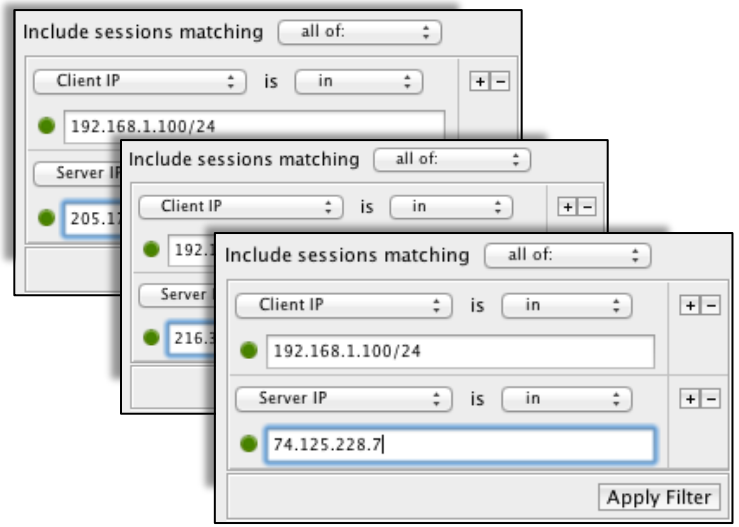
- How do you define “outbound” and why?
  - Anything outside examined subnet? Outside organization?
  - Presumption that use of internal resources not identifying?
    - Mostly true, but what about private servers?



FLOWTRAQ

Select Pairs

- For each server in Top N list, get the list of clients that contacted it
- Filter to reduce computation?
  - Select only ports of interest (HTTP)
    - Avoiding BitTorrent makes for stronger profiles
  - Filter out known-common networks (Akamai, Google)
  - Include only servers with more than some minimum number of sessions



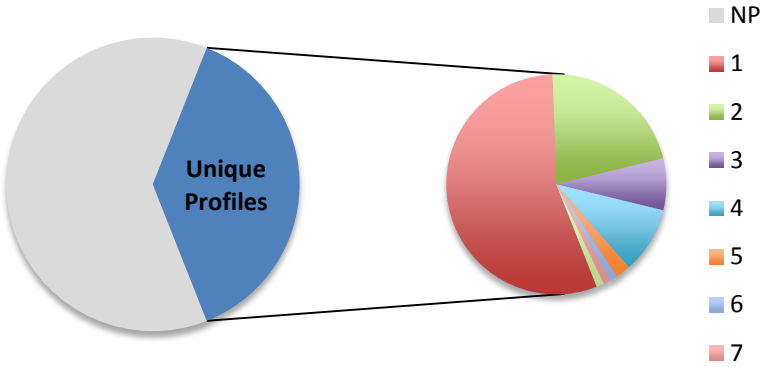
FlowTraq

Compile Fingerprints

- At this stage we have a list of those servers that have only been contacted by one client
  - Potentially pre-filtered for significance (e.g. minimum number of sessions, removed trivial connects such as BitTorrent, etc)
- Create for each client a list of servers
  - Optionally: ranked by percent of client's total traffic (requires second query for each client, increasing total fingerprint time, but providing context and significance measure)
- Each list is a basic but functional fingerprint of that client
  - Sessions to one of those servers in future traffic indicates likely link to that fingerprinted user
    - Primary: that user generated that traffic (on the original device or not)
    - Secondary: that user is connected directly to the user who generated that traffic

Initial Results

- Of ~250 users, profiles could be created representing
  - 38% of users
  - 53% of total traffic
- Breakdown by profile length (# servers in profile):
  1. 51 users (55.4% of profiles)
  2. 20 users (21.7%)
  3. 7 users (7.6%)
  4. 9 users (9.8%)
  5. 2 users (2.2%)
  6. 1 users (1.1%)
  7. 1 users (1.1%)
  8. 1 users (1.1%)

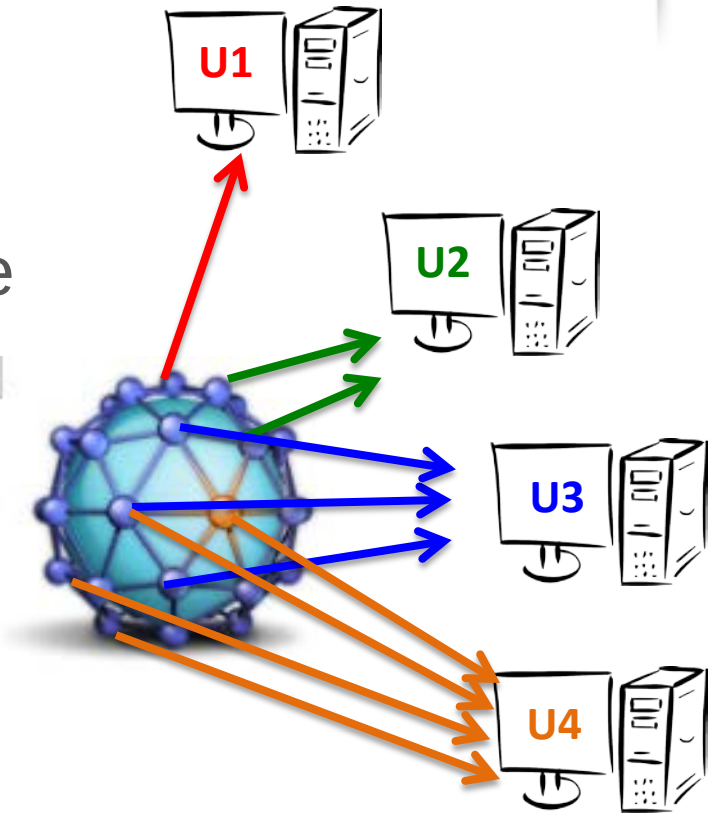


(i.e. 51 users each contacted 1 host unique to them, and one user contacted 8 hosts that nobody else did)

FLOWTRAQ

Uniqueness Levels

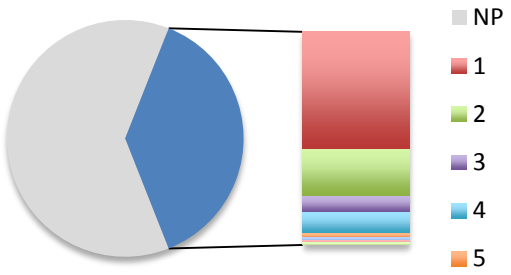
- By relaxing uniqueness requirement, more users can be fingerprinted
  - Tradeoff: Certainty vs. breadth
- Nomenclature
  - The more clients that share a host, the higher the U number



- What is lost in ability to pinpoint users, is gained in insight into shared task/interest
- Some profiles non-unique
  - Same user at different IP addresses?

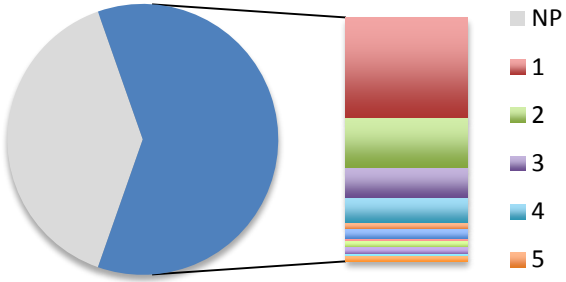
U1-U4 Profile Lists

U1 Profiles



38% of users, 53% of traffic

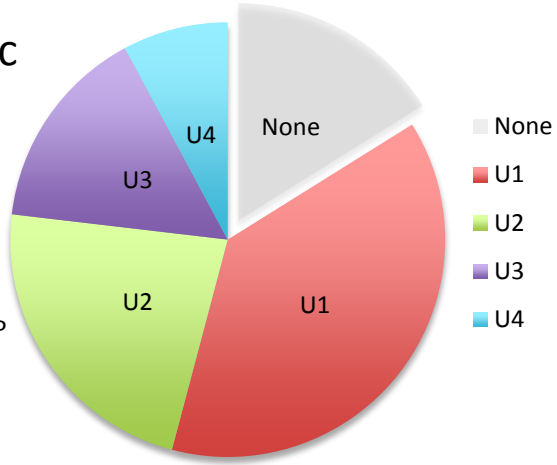
U2 Profiles



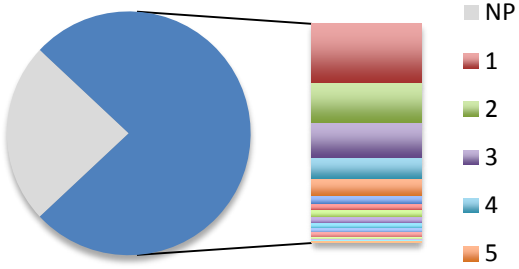
60% of users, 78% of traffic

12 non-unique users

Membership



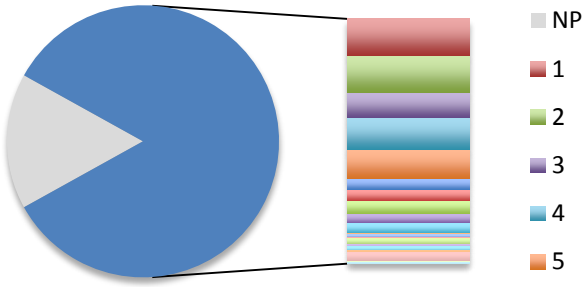
U3 Profiles



75% of users, 89% of traffic

10 non-unique users

U4 Profiles



83% of users, 93% of traffic

10 non-unique users

Variance Over Time

- Variability from month to month is observed
- Month 1

Uniqueness	% of users	% of traffic
U1	38%	53%
U2	60%	78%
U3	75%	89%
U4	83%	93%

- Month 2

Uniqueness	% of users	% of traffic
U1	46%	80%
U2	60%	92%
U3	69%	96%
U4	75%	98%

Results and Lessons Learned

- This represents a first step toward making simple flexible fingerprinting widely available
  - NetFlow is an ideal data source
- Able to fingerprint users comprising majority of network traffic in relatively unrestricted environment
- Uniqueness Levels
  - U1 profiles are more significant
  - U4 profiles cover far more of the population
  - Keeping track of them in parallel allows us the best of both worlds

Take-Home

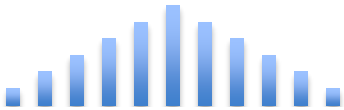
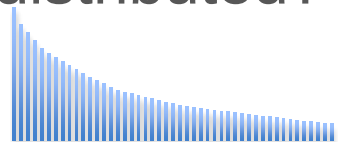
- NetFlow, with its benefits to privacy, ease, and scalability, can be used to produce simple user fingerprints
  - Several types are possible; we went with the simplest plausible type
- Unique site accesses represent one such fingerprint type
  - Intuitive and easy to grasp
  - Adjustable to the level of desired uniqueness
- More sophisticated fingerprints are expected to be more useful still

Next Steps, Short-Term

- Room to grow within NetFlow collection regime:
  - Refine by port/protocol
  - Aggregate content distribution networks
- Make better use of ground truth
  - Newer version of software allows searching on MAC address, to quickly check when fingerprint appears to change or duplicate
  - Determine whether there are substantive differences between wireless and wired networks
    - Number of individuals with identifiable fingerprints
    - Fingerprint stability

Next Steps, Long-Term

- Learning Period Estimation
  - What constitutes a baseline?
- Long-Term Stability
  - How much do these fingerprints change over time?
  - What can be learned from those changes?
  - How are fingerprint lives distributed?


vs


- Autonomous Operation
  - Can fingerprint creation and tuning be automated?

    … to the point of using them for auto-remediation?

For Additional Information…

- For a copy of these slides and the whitepaper, or to evaluate the fingerprinting tool, visit us at:
  - http://www.flowtraq.com/research/FloCon2012.html
- We would be happy to address any questions or comments
  - abarsam@flowtraq.com
  - vberk@flowtraq.com
  - jmurphy@flowtraq.com