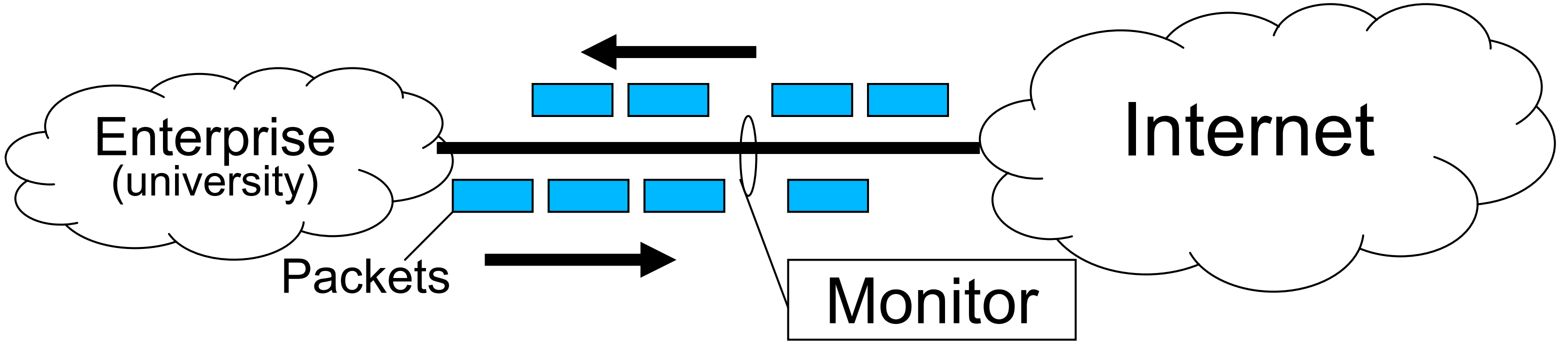
Bruno Ribeiro, Gerome Miklau, Don Towsley
UMass Amherst

Weifeng Chen
California University of Pennsylvania

Analyzing Privacy in Enterprise Packet Trace Anonymization

Motivation



- Packet header traces
  - Used for networking research
  - Many public repositories (UMass, CAIDA, LBNL, …)
- Raw trace may violate user privacy
  - If enterprise IP addresses can be tied to individuals

src address	dest address	src port	dest port	…
14.1.1.1	11.0.0.3	6738	80	…
18.0.0.1	11.0.0.1	2434	22	…
11.0.0.1	20.0.0.3	6913	80	…

Motivation

- Trace repositories
  - Anonymize IP addresses
- Two most widely used schemes
  - Full prefix preservation (Xu et al. , 2001)
  - Partial prefix preservation (Pang et al. 2006)

Original trace

src addr.	dest addr.	src port	dest port	…
14.1.1.1	11.0.0.3	6738	80	…
11.0.0.1	20.0.0.3	7913	22	…

anonymization mapping

Anonymized trace

src addr.	dest addr.	src port	dest port	…
200.0.1.2	128.0.64.2	6738	80	…
128.0.64.0	5.0.4.5	7913	22	…

Adversary

- Adversarial model:
  - De-anonymize enterprise IP addresses in the trace
    1. Probes (scan) enterprise network
    2. Collects similar information from the trace
    - De-anonymizes trace IPs matching (1) with (2)

Outline

- Our contributions
  - New attack on IP anonymization:
    - Attack overview
    - Defined as a tree editing distance problem
  - Worst-case analysis:
    - From a set of trace labels (information)
    - Assesses worst-case attack
- Related work
- Conclusions

Proposed attack overview

- Adversary provides:
  - Labeled tree constructed using anonymized trace
  - Labeled tree constructed from probing enterprise
  - A cost (or distance) function (to deal with “mismatched” labels)

- Our algorithm finds:
  - All de-anonymizations that
    - comply with prefix preservation restrictions
    - and have minimum total cost

- An instance of the *tree edit distance* problem

Full prefix preserving anonymization

- Full prefix preservation
  - If two real addresses share first **X** bits, then
  - the same two anonymized addresses share first **X** bits
- It imposes restrictions on the real IP → Anonymized IP mapping

Labeled trees

Probed tree

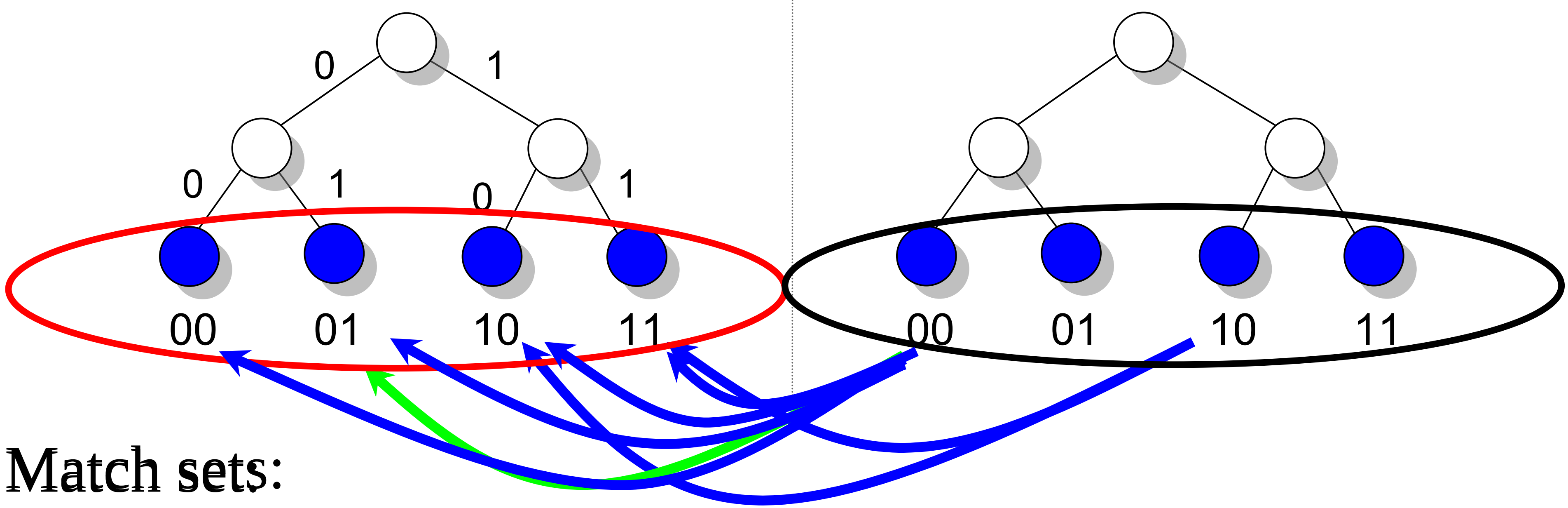
Probed IP leaf labels

- Web server (green)
- Not a Web server (blue)

Trace tree

Trace IP leaf labels

- Traffic on port 80 (green)
- No traffic on port 80 (blue)



- Match sets:
  - 00 maps to {00, 01, 10, 11}
  - 10 maps to {00, 01, 10, 11}

Imperfect information

Probed tree

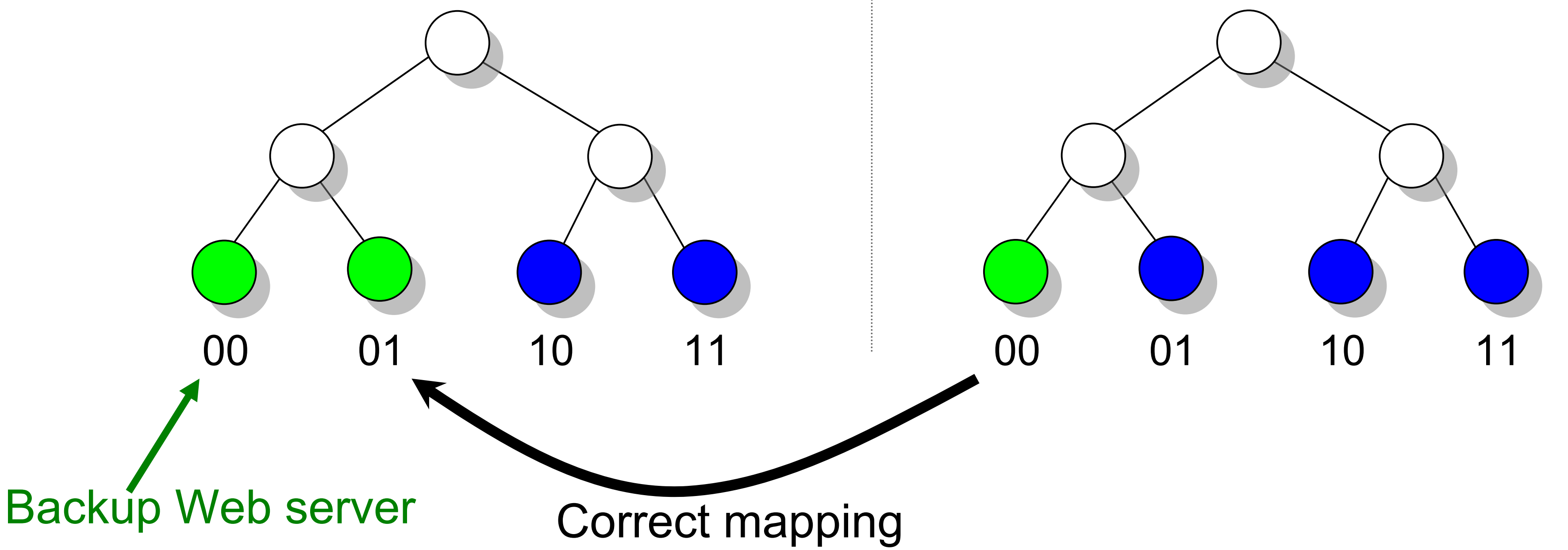
Probed IP leaf labels

- Web server (green)
- Not a Web server (blue)

Trace tree

Trace IP leaf labels

- Traffic on port 80 (green)
- No traffic on port 80 (blue)

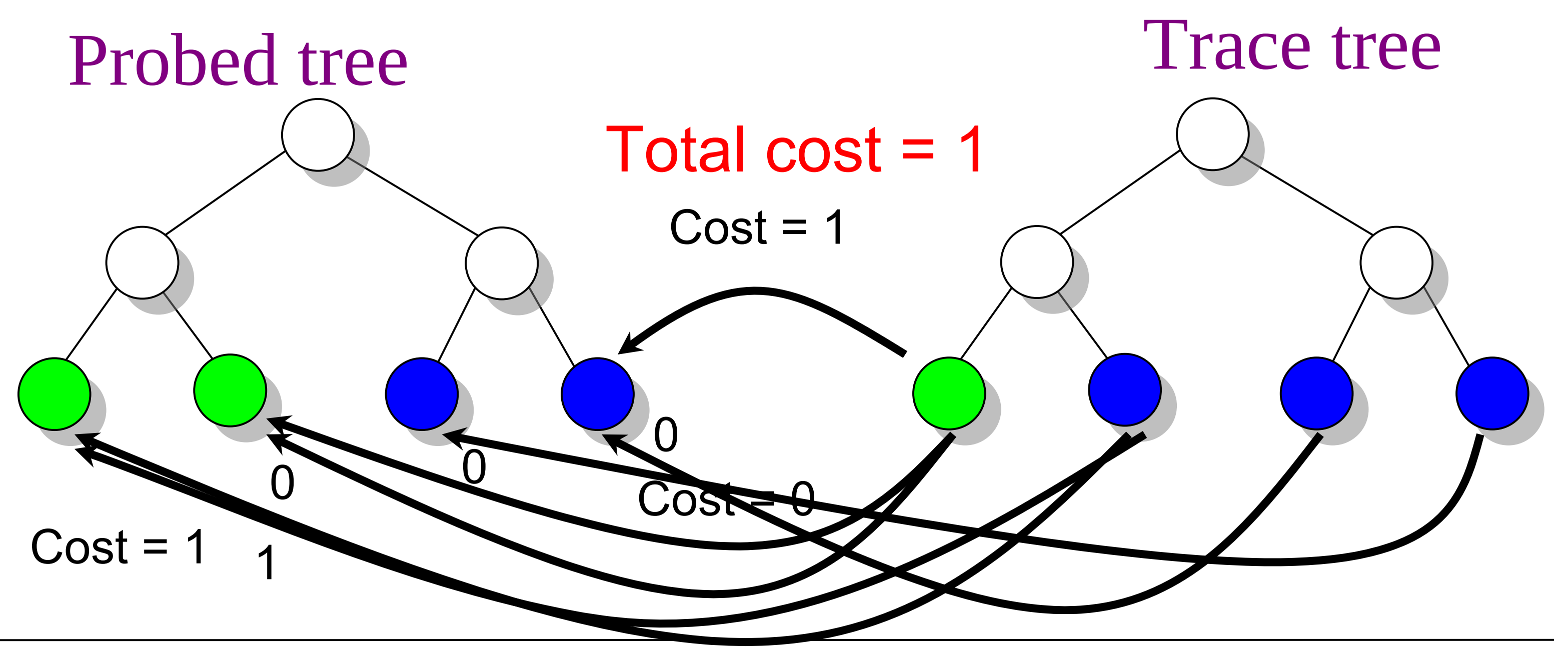


- Other sources of imperfect labels: Dynamic IP addresses, host shutdown, etc.

Mapping costs

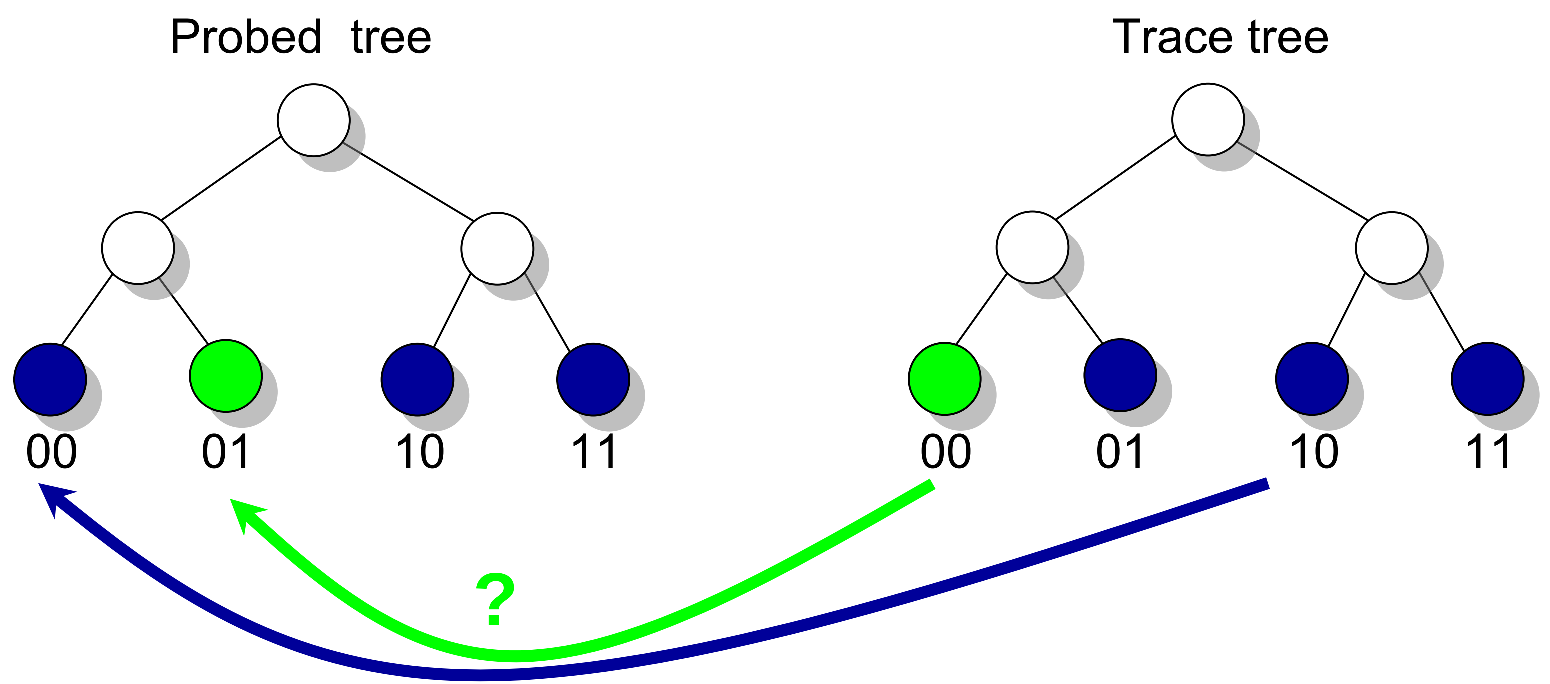
- Assign a cost to map two IPs with different labels
  - Is **zero** if labels are equal
- Mapping cost
  - Sum of all individual costs

Example:



Proposed attack

- All minimum cost mappings (over the whole network)
  - Because it is prefix-preserving
    - Every de-anonymization limits future de-anonymizations



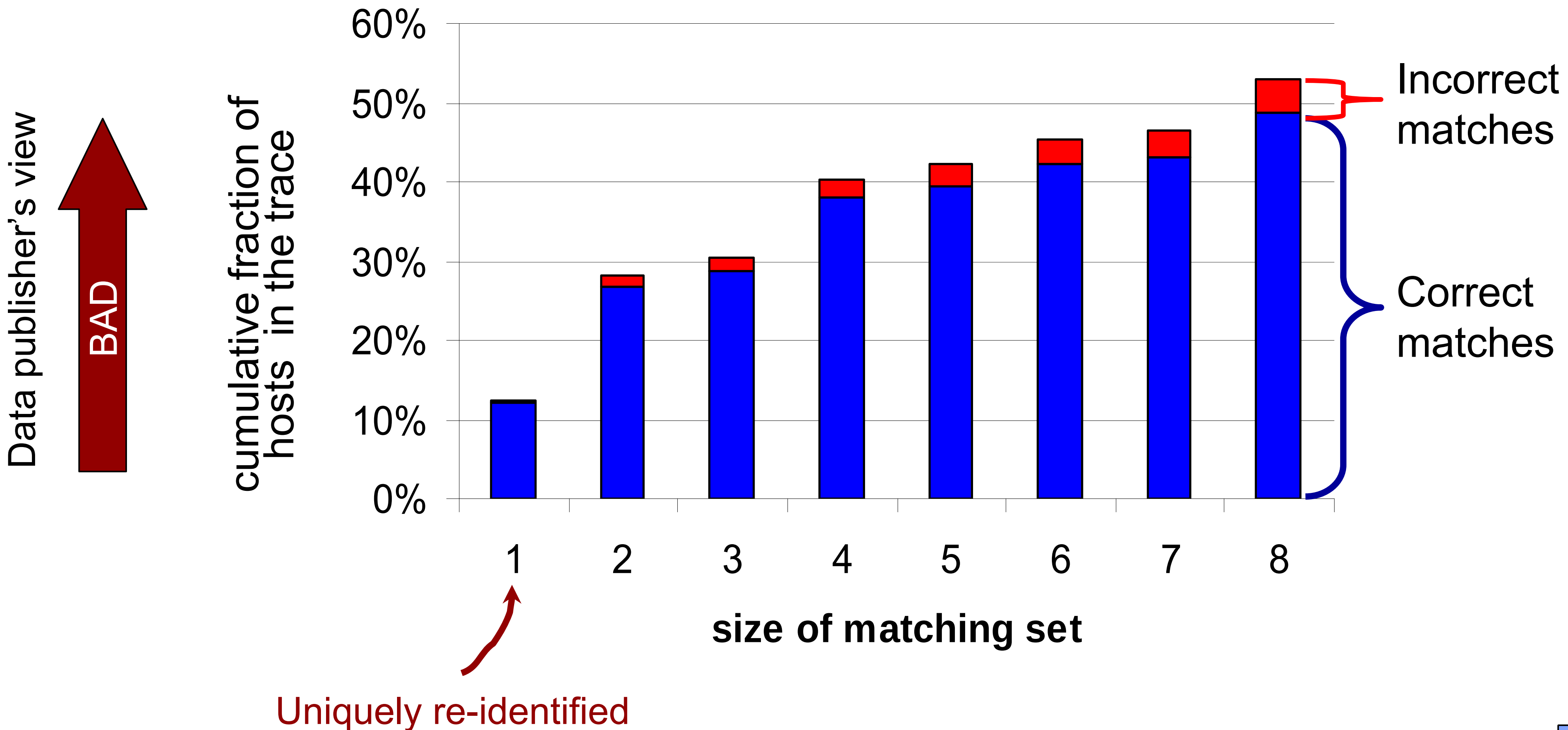
- And our algorithm is fast
  - 10 seconds (on this laptop) for all mappings of a network with 2^{16} addresses

Experiment

- Network: class B (64K addresses)
- Labels
  - “Active host”
  - Active ports: FTP, SSH, Telnet, E-mail, Time, DNS, Web, POP3, SOCKS
- Trace IP labels
  - “Active host” label – recorded any outgoing traffic
  - “Active ports” – Recorded traffic from ports 80, 22, ….
- Probed IP labels
  - Probed over all network
    - “Active host” label – PING
    - “Active ports” – TCP SYN ACK reply from ports 80, 22, …
- Naïve cost function: **Zero** is labels are equal, **one** otherwise

Experiment results

- Trace collected: 2007, June 18th (9097 active IPs)
- Network probed: 2007, June 18th



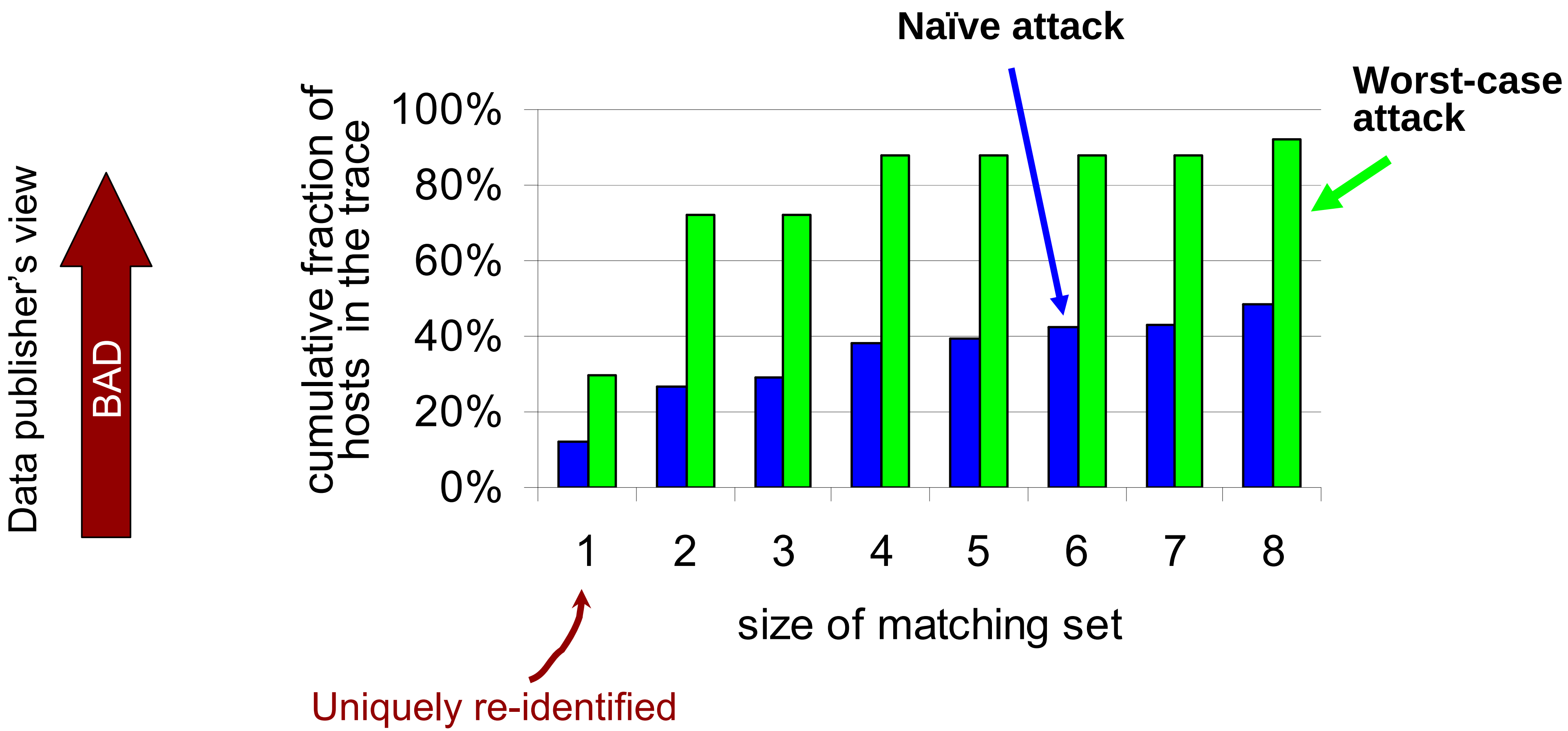
Worst-case analysis

- Given a labeled trace tree
- Find best de-anonymization

- We provide an algorithm that
  - Obtains worst attack matching set size
    - For each IP address in the trace
    - For any label mismatch cost function
    - For any labeled probed tree

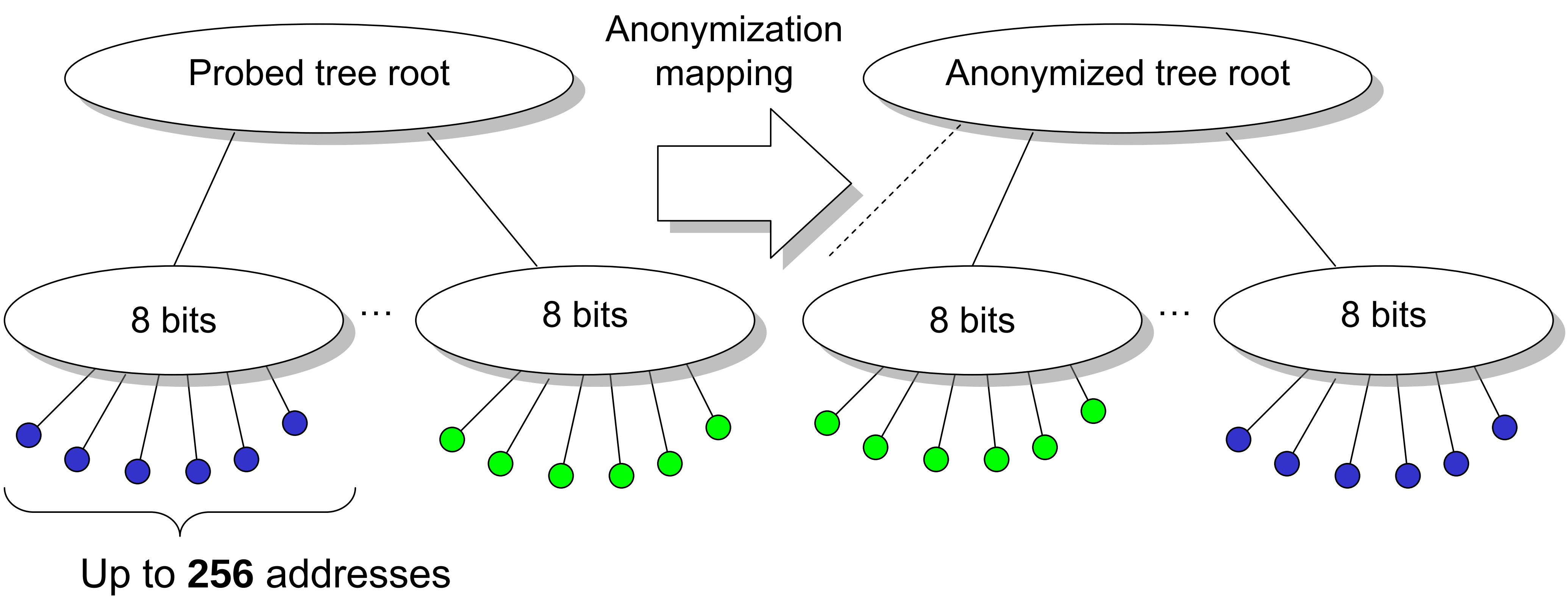
Worst-case experiment

- Full prefix preservation
- June 18th experiment



Partial prefix preservation

- Does not retain part of the address structure
- Used in *Pang et al., 2006*
- Solution also formulated as an instance of the *tree edit distance* problem



Partial vs. Full prefix preservation

- Intuition: Partial is **much** safer than full prefix preservation

Worst case:

st case:
prefix
servation

Data publisher's view

cumulative fraction of
hosts in the trace

Worst-case analysis (II)

Uniquely re-identified

- **Full prefix preservation: 2713 active IP addresses in the trace**
- **Partial prefix preservation: 113 active IP addresses in the trace**

Partial prefix preservation is safer but not completely safe

Related work

- *"Playing Devil's Advocate: Inferring Sensitive Information from Anonymized Traces"*, Scott Coull, Charles Wright, Fabian Monrose, Michael Collins and Michael Reiter, NDSS 2007
  - An attack on partial prefix preservation

- *"Taming the Devil: Techniques for Evaluating Anonymized Network Data"*, Scott Coull, Charles Wright, Fabian Monrose, Angelos Keromytis and Michael Reiter, NDSS 2008
  - Comes right after this talk ☺

Conclusions

- Attack
  - Include global mapping restrictions
  - An instance of the tree edit distance problem
  - Indicates that full prefix preservation has flaws
    - Impact of late probing on the de-anonymization
- Worst-case analysis
  - Can help future anonymization schemes
    - A tool for data publishers
  - Experiments indicate that:
    - Partial is much safer than full prefix preservation
      - But still not completely safe

Thanks

- **Jim Kurose**, UMass Amherst
- **Edmundo de Souza e Silva**, Federal University of Rio de Janeiro
- **Kyoungwon Suh**, Illinois State University
- **Anonymous NDSS'08 reviewers**
- **Neils Provos**, Google Inc.