802.11 User Fingerprinting

Jeffrey Pang[1] Ben Greenstein[2] Ramakrishna Gummadi[3]
Srinivasan Seshan[1] David Wetherall[2,4]

[1]CMU [2]Intel Research Seattle [3]USC,MIT [4]University of Washington

Mobicom '07

Some slides borrowed from the
Mobicom 07 presentation
by the owners of the paper

Introduction

- Measurement based paper
- Tracking is worrisome to people, especially the ubiquitous 802.11 network devices.
- Location Privacy in danger because wireless devices disclose our location or identities or both.
- Many other technologies like RFID pose similar threats.
- Inspite of changing parameters, 802.11 devices emit characteristics that make the devices trackable.
- Pseudonyms, temporary unlinkable names were proposed to use to prevent tracking.
- But the results in the paper shows theyre not enough.

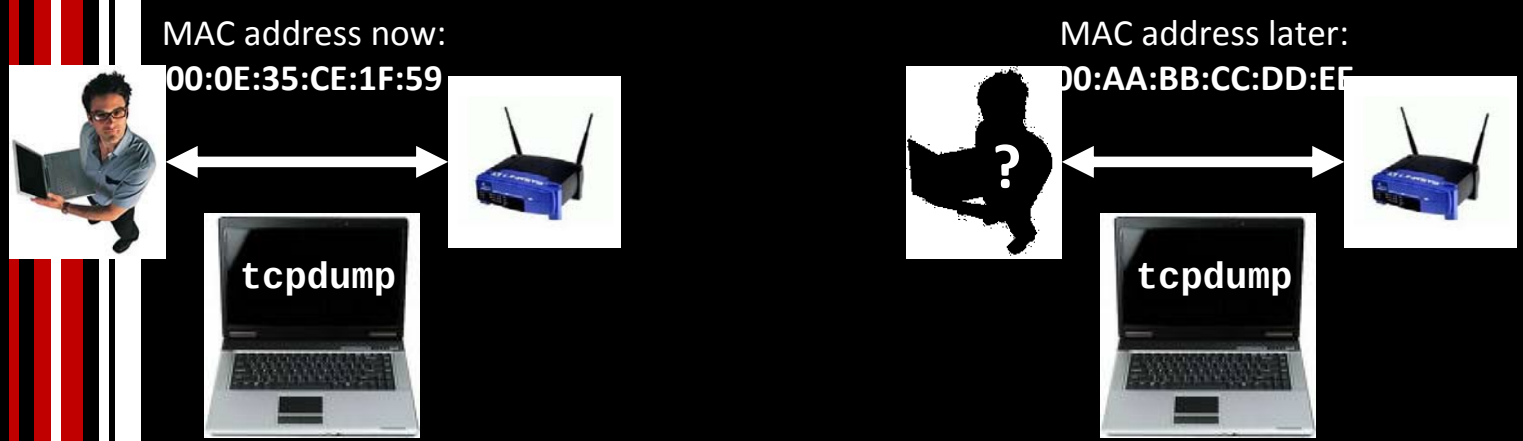
Motivation: The Mobile Wireless Landscape

- A well known technical problem
  - Devices have unique and consistent addresses
  - e.g., 802.11 devices have MAC addresses

  ➔ fingerprinting them is trivial!

Motivation: The Mobile Wireless Landscape

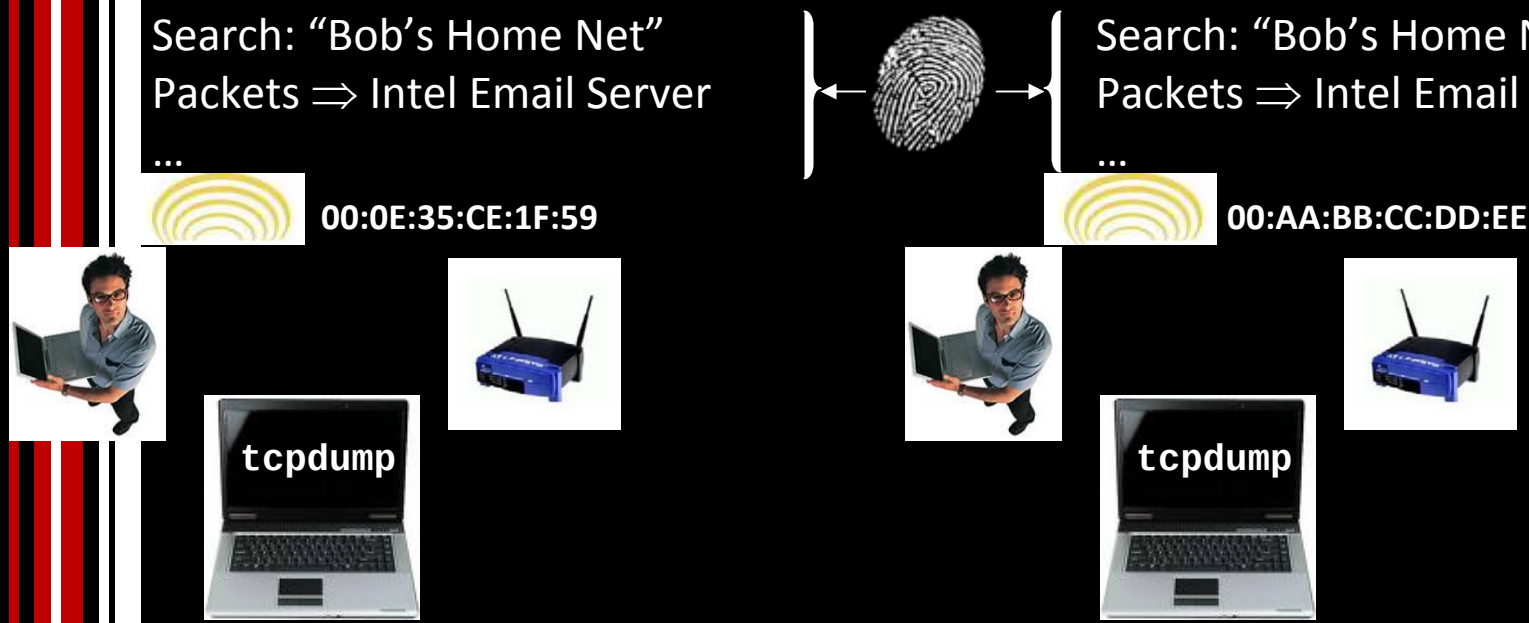
- The widely proposed techical solution
  - **Pseudonyms**: Change addresses over time
    - 802.11: Gruteser ’05, Hu ’06, Jiang ’07
    - Bluetooth: Stajano ’05
    - RFID: Juels ‘04
    - GSM: already employed



Motivation: The Mobile Wireless Landscape

The results show: **Pseudonyms are not enough**

- *Implicit identifiers*: identifying characteristics of traffic
- Parameters like IP address of your frequently used email
- E.G., most users identified with 90% accuracy in hotspots



Contributions

- Four Novel 802.11 Implicit Identifiers
- Automated Identification Procedure
- Evaluating Implicit Identifier Accuracy

Implicit Identifiers

- netdest pairs
- SSID
- Broadcast packets
- MAC protocol fields

The Implicit Identifier Problem

- How significantly do implicit identifiers erode privacy?.....lets see by example….
- A signal trace obtained at 2004 SIGCOMM conference is used.
- MAC address is hashed to provide anonymity…… equivalent to a pseudonym

- A device automatically searches for preferred networks first and hence from the SSID users could be identified.
- For example, a user's laptop searched for network names like "MIT", "roofnet"…. The user must be from Cambridge,MA!!
- SSID probes with unique names make the job easier. E.g. "therobertmorris"
- Another user used BitTorrent to download. The MAC address in the data packets was hashed but he accessed the same SSH and IMAP server every hour and was the only one to do so at SIGCOMM….hence IDENTIFIED!!

- Implicit identifiers are many times exposed by design flaws
- Identifying information is exposed at the higher layers of network stack as they are not adequately masked
- Identifying information during service discovery is not masked
- Rectifying these shortcomings will come at a high cost.

Experimental Setup

- The Adversary
  - Service providers and large monitoring networks are the biggest threat.
  - Network monitoring softwares like “tcpdump” enables any lay man to track with just an 802.11 device like laptop.
- The Environments
  - Public networks such as hot spots.
    - Unencrypted link layer
    - Access control employed at higher layers with MAC address filtering
    - Identifying features in network link layer and physical layer are visible to the eavesdropper
  - Home networks
    - High density of access points in urban areas
    - Employ link layer encryption
    - Authorized users are known and small in number
    - Eavesdropper can still view the payloads of data packets, frame sizes, timing
  - Enterprise networks
    - Devices authorized
    - Less diversity in the behavior of wireless cards

- Monitoring scenario
  - Assume that users use different pseudonyms for each session in each of the networks
  - Hence explicit identifiers cannot link their sessions
  - The authors define a traffic sample to be one user's network traffic observed during one hour
  - Assume that the adversary is able to obtain training samples either before or during the monitoring period from the person being tracked.

- Evaluation Criteria
  - ***Did this traffic sample come from user U?***
  - ***Was user U here today?***
- Wireless Traces
  - "*sigcomm*" a 4 day trace from monitoring point in 2004 SIGCOMM conference
  - "*ucsd*" a trace of all 802.11 traffic in U.C Sand Diego's computer science building during one day
  - "*apt*" a 19 day trace monitoring all networks in an apartment building

Implicit identifiers

- Results show:
  - Many identifiers are effective at distinguishing users while others are useful for distinguishing groups of users
  - A non-trivial fraction of users are trackable using one highly discriminating identifier
  - On an average only 1 to 3 samples are enough to leverage identifiers to full effect
  - At least one implicit identifier accurately identifies users over multiple weeks

Network destinations

- “netdests” is a set of IP<address, port> pairs that are known to be common to all users
- This set is unique to each user.
- An adversary can obtain network address in any wireless network inspite of link layer encryption or VPN. No application or network layer security mechanism such as IPSec would mask this identifier

SSID Probes

- SSID of a netwoork is added to the networks list when a client first associates with the network.
- The client sends probe requests to find if it is in the vicinity of its preferred networks
- Probes are never encrypted because they occur before association and key agreement
- Some SSIDs are more distinguishing than others which makes it useful many times.

Broadcast packet sizes

- Many applications broadcast packets to advertise their existence to machines on the local network
- These packets contain naming information
- In the observed traces, NetBIOS advertisements and filemaker and Microsoft office *bcasts* were found
- DHCPP requests and power management beacons are common to all users hence not included in the bcasts set.

MAC protocol fields

- Specific combination of 802.11 protocol fields visible in the MAC header that distinguish a wireless users card, driver and configuration
- For example:
  - More fragments
  - Retry
  - Power management
  - Order bits
  - Authentication algorithms
  - Supported transmission rates

Implicit Identifier Summary

802.11 Networks:	Public	Home	Enterprise
Network destinations	✓		
SSIDs in probes	✓	✓	✓
Broadcast pkt sizes	✓	✓	✓
MAC protocol fields	✓	✓	

Identifying even if devices have WEP/WPA encryption

Visible even with WPA2 and drivers

- More implicit identifiers exist
  - ➔ Results presented establish a lower bound

Automated Identification Procedure

- Many potential tracking applications:
  - Was user X here today?
  - Where was user X today?
  - What traffic is from user X?
  - When was user X here?
  - Etc.

Build a *profile* from training samples:
First collect some traffic known to be from user X and from random others

Sample Classification Algorithm

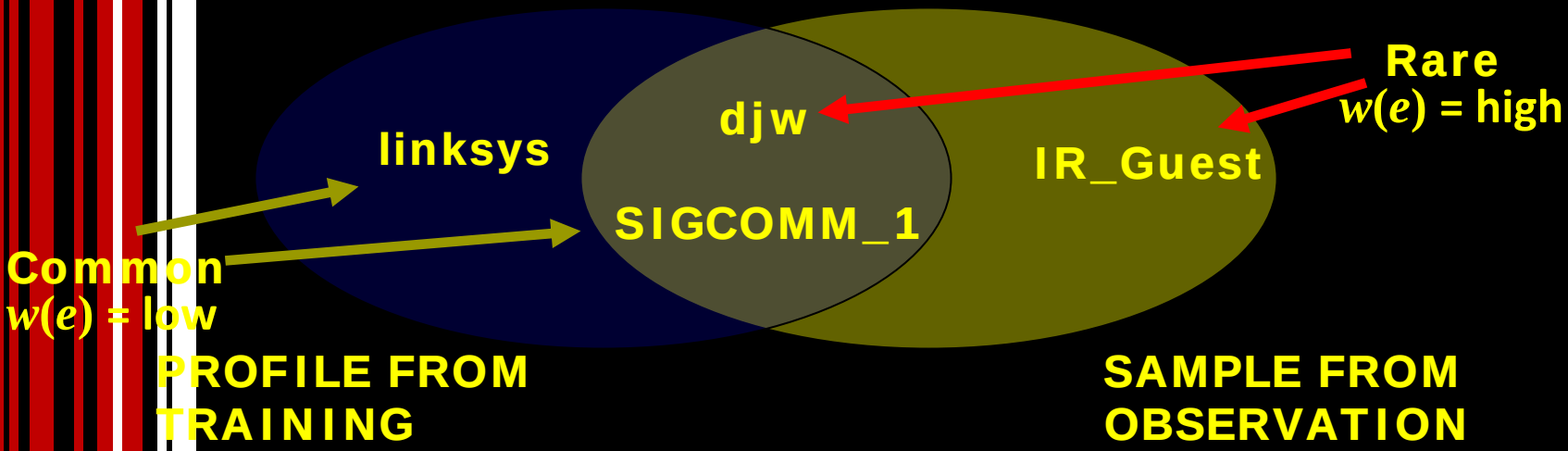
- Core question:
  - Did traffic sample s come from user X?
- A simple approach: naïve Bayes classifier
  - Derive probabilistic model from training samples
  - Given s with features F, answer "yes" if:

    Pr[s from user X | s has features F] > T

    for a selected threshold T.
  - F = feature set derived from implicit identifiers

Sample Classification Algorithm

Deriving features F from implicit identifiers



$$feature_U(s) = \frac{\sum_{e \in Profile_U \cap Set_s} w(e)}{\sum_{e \in Profile_U \cup Set_s} w(e)}$$

Evaluating Classification Effectiveness

- Simulate tracking scenario with wireless traces:

	sigcomm		ucsd		apt	
	training	validation	training	validation	training	validation
Duration (hours)	37	54	10	11	119	345
Total Samples	1974	3391	587	1240	638	1473
Frames Per Sample (median)	289	284	1227	1128	57	92
Total Users	377	412	225	371	97	196
Profiled Users	337	337	153	153	39	39
Samples Per Profiled User (mean)	5.5	9.1	3.1	4.7	14.7	32.2
Users Per Hour (mean)	53	64	59	113	5	4

Table 1—Summary of relevant workload statistics and parameters. The duration reports only hours with at least one active user.

  - Split each trace into training and observation phases

Question: Is observation sample s from user X?

Evaluation metrics:

- **True positive rate (TPR)** = ???
  Fraction of user X's samples classified correctly
- **False positive rate (FPR)** = 0.01
  Fraction of other samples classified incorrectly

Measure TPR

Fix T for FPR

Pr[s from user X | s has features F] > T

- Q: Did this sample come from user U?

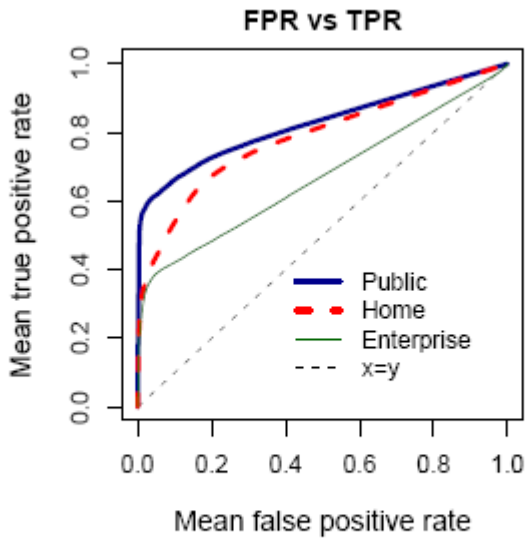
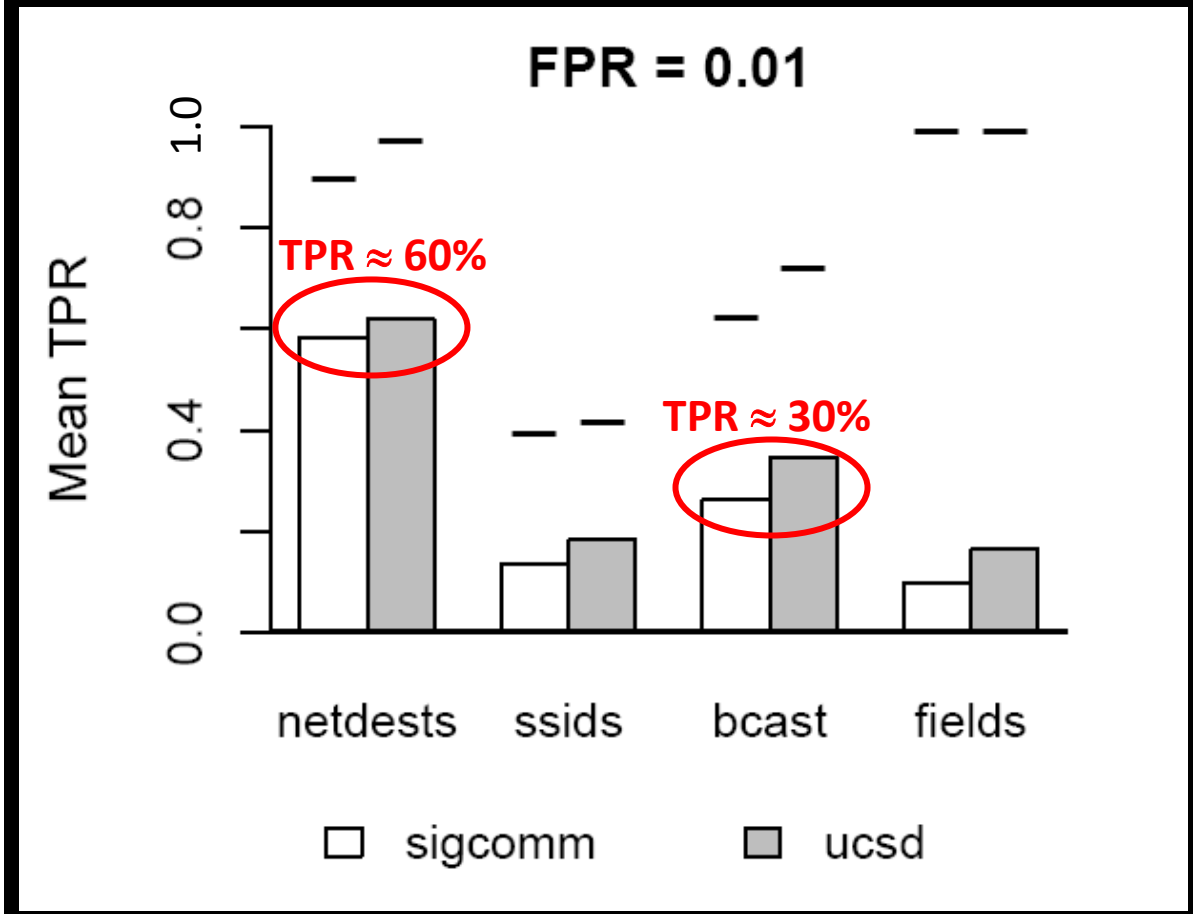


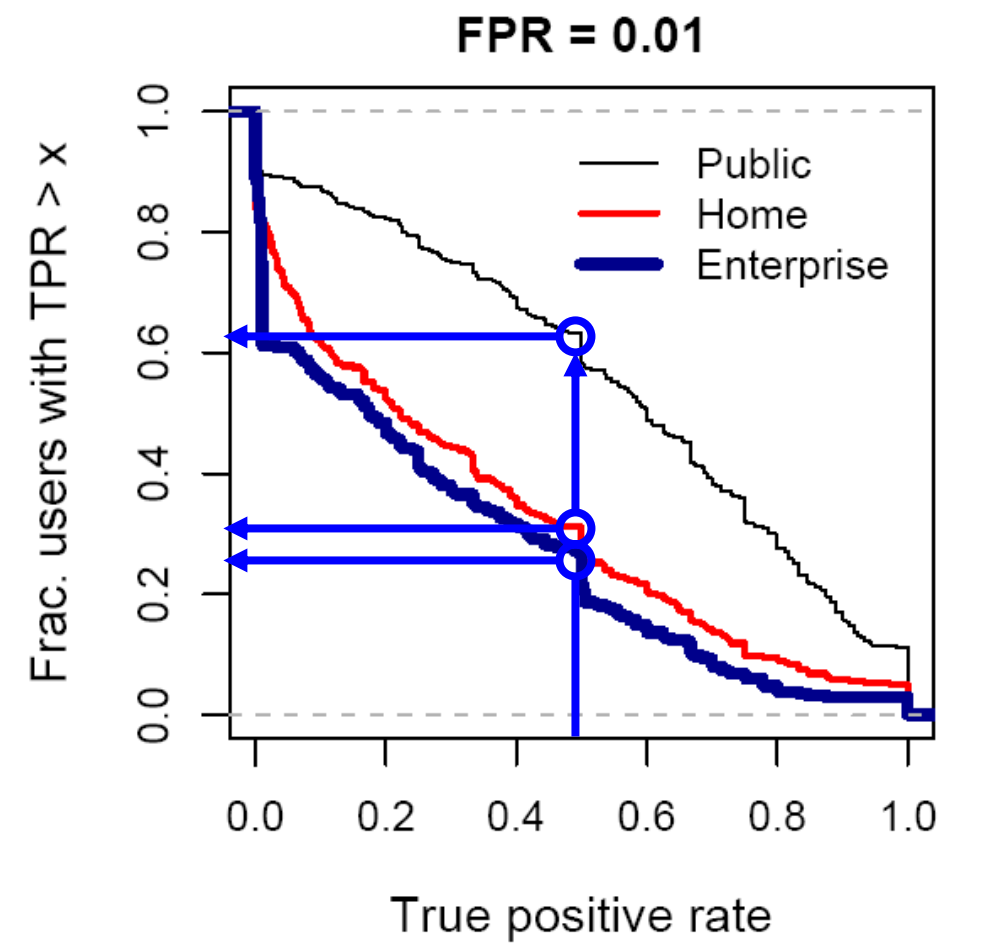
Figure 7—Classification accuracy for Question 1 if `sigcomm` users where in typical public, home, and enterprise networks.

Results: Individual Feature Accuracy



Individual implicit identifiers give evidence of identity

Results: Multiple Feature Accuracy



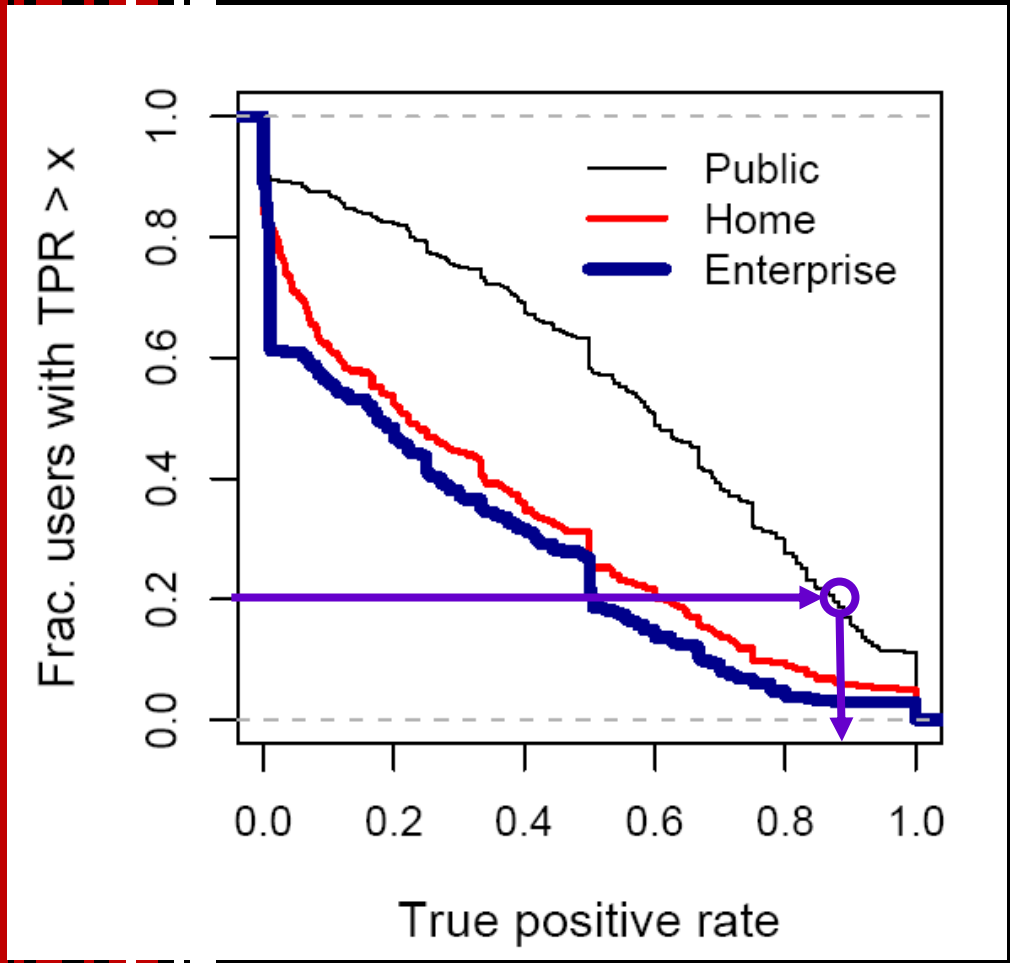
Users with TPR >50%:

Public: 63%
Home: 31%
Enterprise: 27%

	Public	Home	Enterprise
netdests	✓		
ssids	✓	✓	✓
bcast	✓	✓	✓
fields	✓	✓	

We can identify many users in all environments

Results: Multiple Feature Accuracy



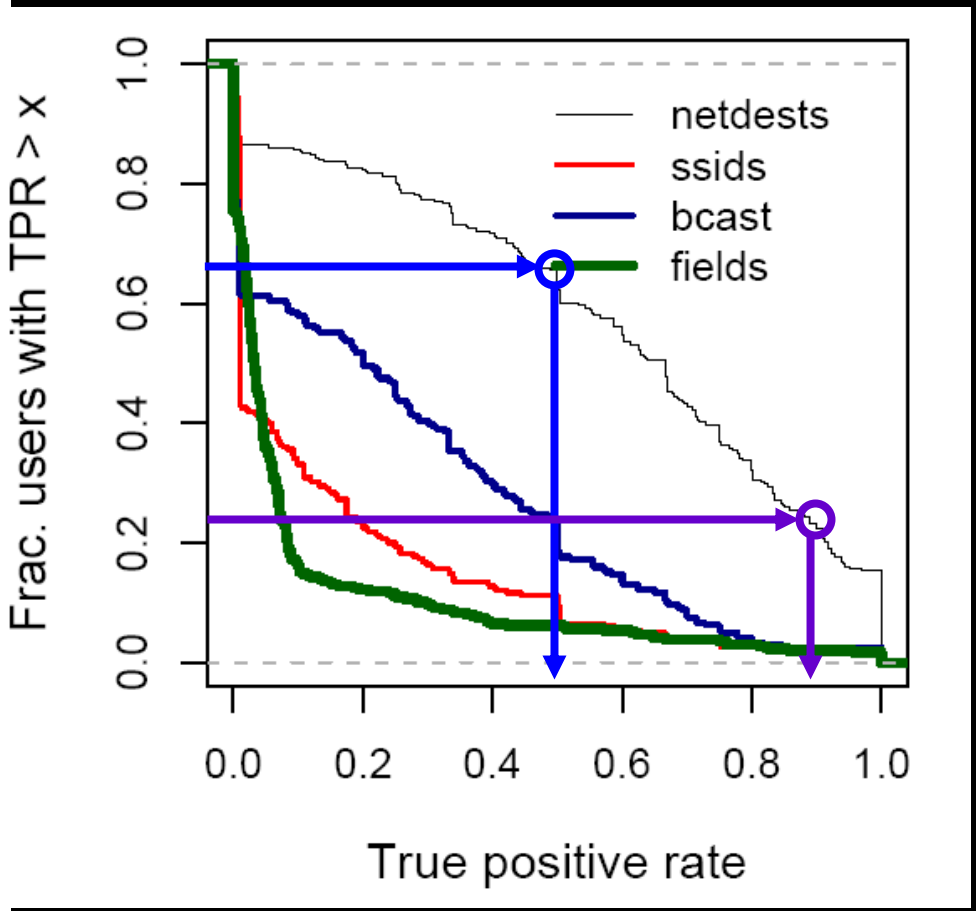
Public networks:
~20% users identified
>90% of the time

	Public	Home	Enterprise
netdests	✓		
ssids	✓	✓	✓
bcast	✓	✓	✓
fields	✓	✓	

Some users much more distinguishable than others

- **Question**: Was user X here today?
- More difficult to answer:
  - Suppose N users present each hour
  - Over an 8 hour day, $8N$ opportunities to misclassify
  - ➔Decide user X is here only if *multiple* samples are classified as his
- **Revised**: Was user X here today for a few hours?

Results: Individual Feature Accuracy



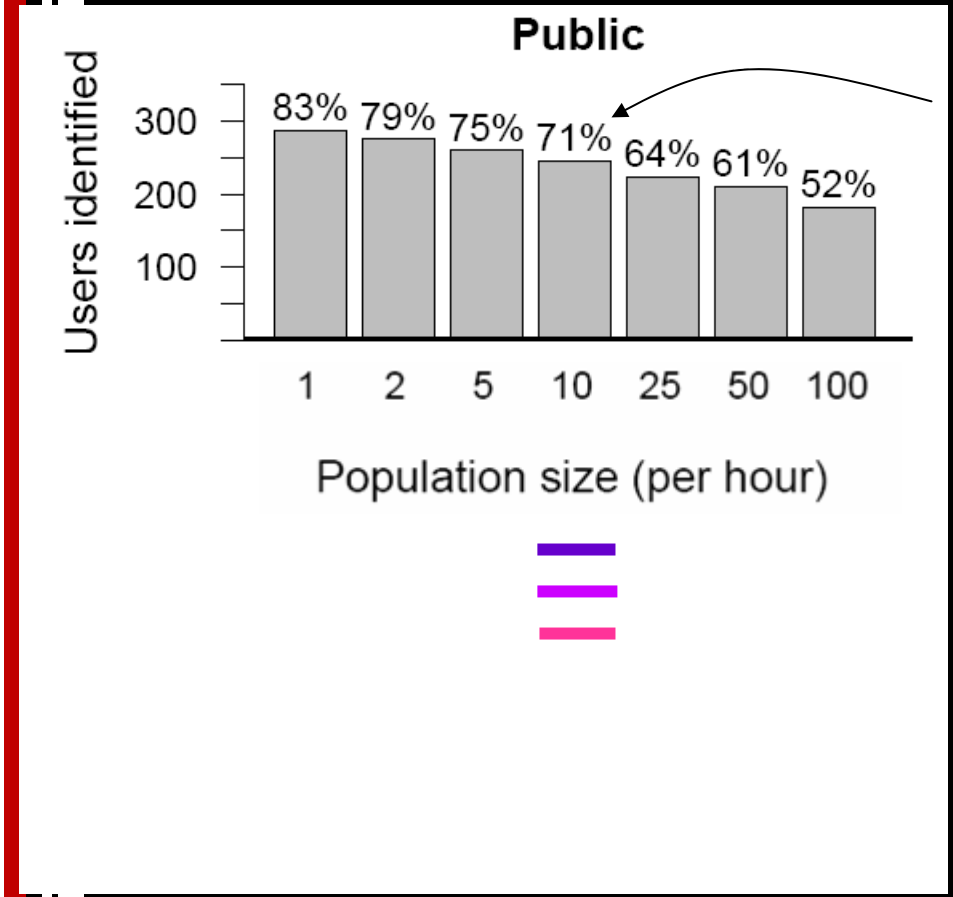
netdests:

~60% users identified >50% of the time

~20% users identified >90% of the time

Some users more distinguishable than others

Results: Tracking with 90% Accuracy



Of 268 users (71%):

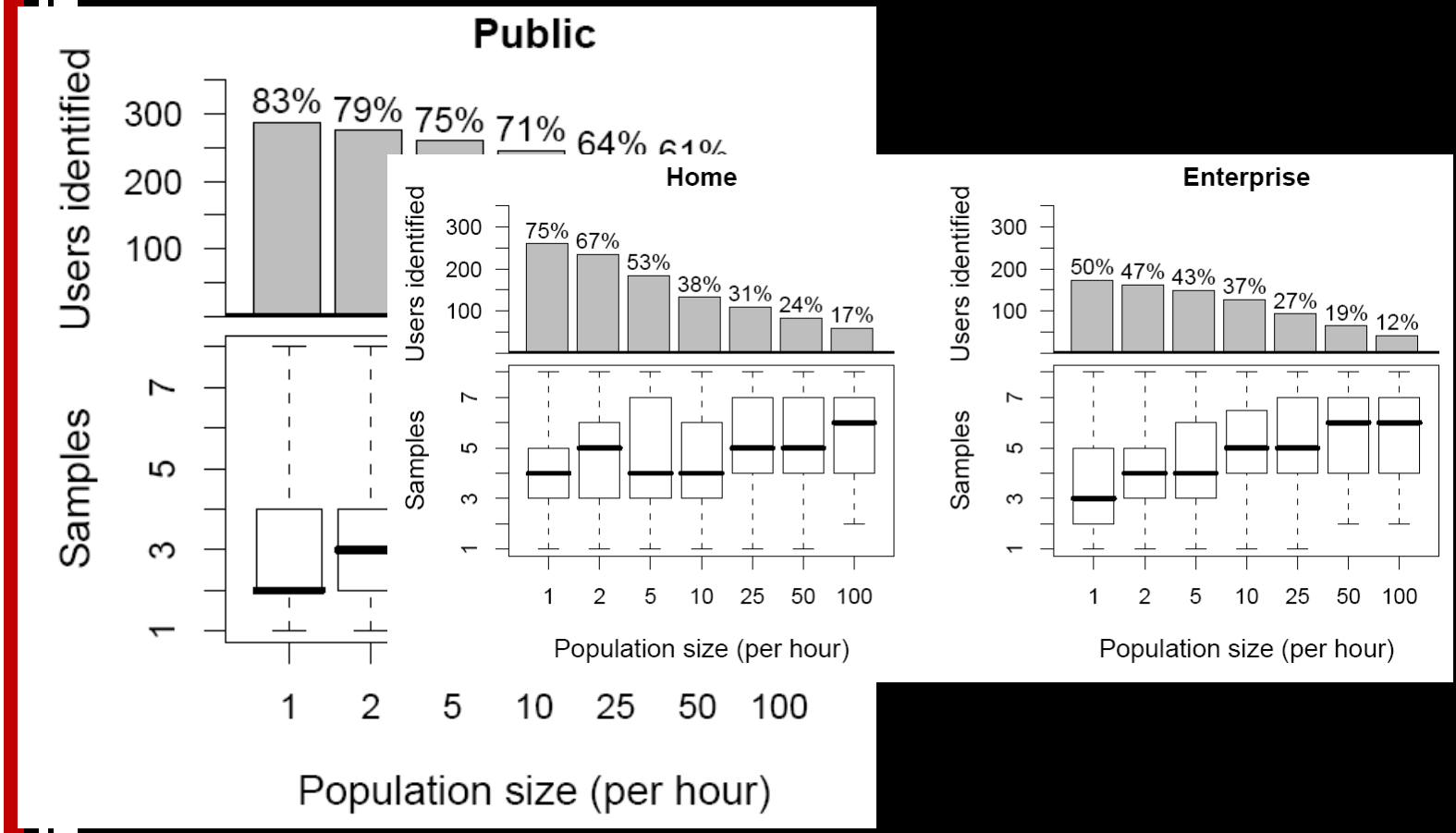
75% identified with ≤4 samples

50% identified with ≤3 samples

25% identified with ≤2 samples

Majority of users can be identified if active long enough

Results: Tracking with 90% Accuracy



Many users can be identified in all environments

Conclusions

- Implicit identifiers can accurately identify users
  - Individual implicit identifiers give evidence of identity
  - We can identify many users in all environments
  - Some users much more distinguishable than others
- Understanding implicit identifiers is important
  - Pseudonyms are not enough
  - a *lower bound* on their accuracy is established

- Future
  - Uncover more identifiers (timing, etc.)
  - Take measures to resolve the issues regarding the implicit identifier problem and build a better link layer and to prevent detection from these identifiers.

THANK YOU…