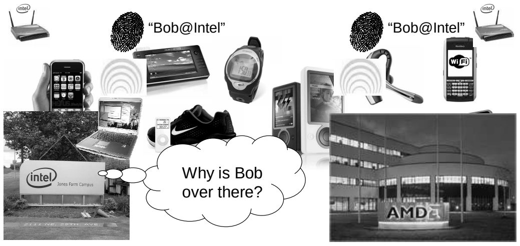
# 802.11 User Fingerprinting

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

## Motivation: The Mobile Wireless Landscape

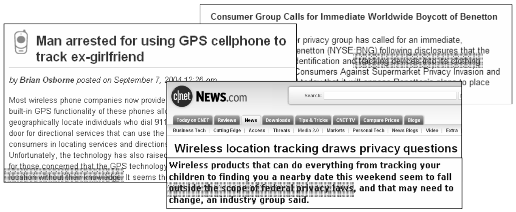


- They leave "digital fingerprints" that reveal who we are
  - And thus where we've been

## Motivation: The Mobile Wireless Landscape

➔ *Location privacy* is growing concern

- Wireless Privacy Protection Act [U.S. House Bill '05]
- Geographic location/privacy working group [IETF]



Consumer Group Calls for Immediate Worldwide Boycott of Benetton

r privacy group has called for an immediate,
enetton (NYSE:BNG) following disclosures that the
dentification and tracking devices into its clothing
Consumers Against Supermarket Privacy Invasion and
o place

Man arrested for using GPS cellphone to track ex-girlfriend

*by* ***Brian Osborne*** *posted on September 7, 2004 12:26 pm*

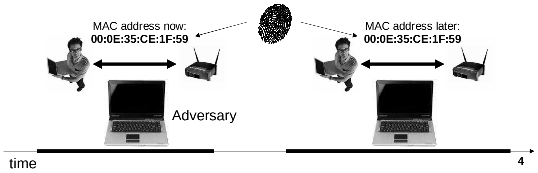
Most wireless phone companies now provide
built-in GPS functionality of these phones allo
geographically locate individuals who dial 911
door for directional services that can use the
consumers in locating services and directions
Unfortunately, the technology has also raised
for those concerned that the GPS technology
location without their knowledge. It seems th

Wireless location tracking draws privacy questions

Wireless products that can do everything from tracking your children to finding you a nearby date this weekend seem to fall outside the scope of federal privacy laws, and that may need to change, an industry group said.

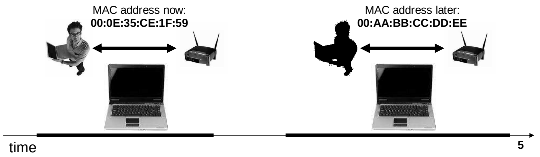
## Motivation: The Mobile Wireless Landscape

- A well known technical problem
  - Devices have unique and consistent addresses
  - e.g., 802.11 devices have MAC addresses
  - ➔ fingerprinting them is trivial!



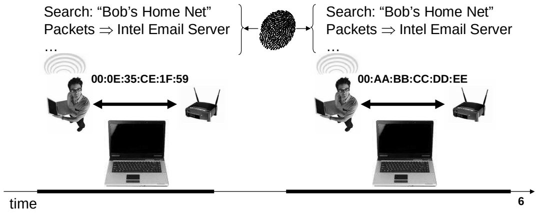
## Motivation: The Mobile Wireless Landscape

- 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

- Our work shows: **Pseudonyms are not enough**
  - *Implicit identifiers*: identifying characteristics of traffic
  - E.G., most users identified with 90% accuracy in hotspots



# Contributions

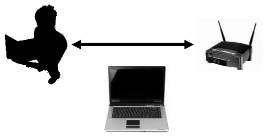
- Four Novel 802.11 Implicit Identifiers
- Automated Identification Procedure
- Evaluating Implicit Identifier Accuracy

# Contributions

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

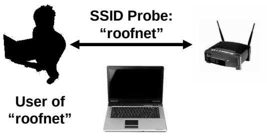
# Implicit Identifiers by Example

- Consider one user at SIGCOMM 2004
  - Transferred 512MB via BitTorrent
    - Poor network etiquette?
  - Seen in a “anonymized” wireless trace
    - MAC addresses hashed, effectively a pseudonym
- Can we identify the culprit using implicit identifiers?



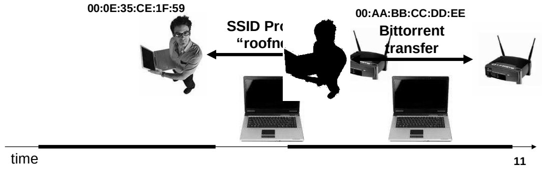
# Implicit Identifiers by Example

- Implicit identifier: SSIDs in probes
  - Set of SSIDs in 802.11 probe requests
  - Many 802.11 drivers search for preferred networks
  - Usually networks you have associated with before



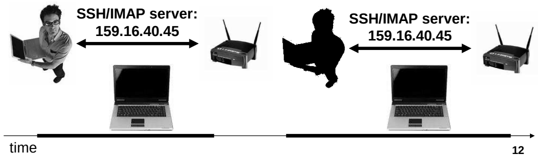
# Implicit Identifiers by Example

- Implicit identifier: SSIDs in probes
  - Set of SSIDs in 802.11 probe requests
  - Many 802.11 drivers search for preferred networks
  - Usually networks you have associated with before



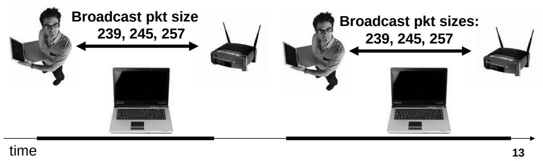
# Implicit Identifiers by Example

- Implicit identifier: network destinations
  - IP <address, port> pairs in network traffic
  - At SIGCOMM, each visited by 1.15 users on average
  - Some nearly-unique destinations repeatedly visited (e.g., email server)



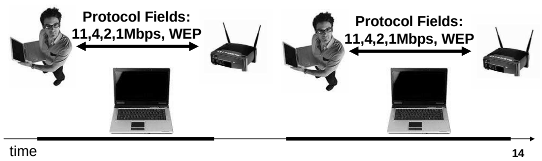
# Implicit Identifiers by Example

- Implicit identifier: broadcast packet sizes
  - Set of 802.11 broadcast packet sizes in network traffic
  - E.g., Windows machines NetBIOS naming advertisements; FileMaker and Microsoft Office advertise themselves

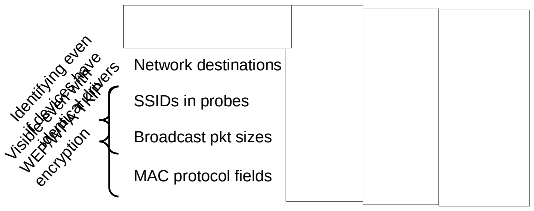


# Implicit Identifiers by Example

- Implicit identifier: MAC protocol fields
  - Header bits (e.g., power management., order)
  - Supported rates
  - Offered authentication algorithms



# Implicit Identifier Summary


Identifying even if devices have identical drivers
Visible even with WEP/WPA encryption
Network destinations
SSIDs in probes
Broadcast pkt sizes
MAC protocol fields

- More implicit identifiers exist
  - ➔ Results we present establish a lower bound

# Fixing Implicit Identifiers is not Simple

- Encryption does not prevent traffic analysis
  - Cover traffic?
  - **Challenge**: Shared medium ⇒ large performance hit
- Service discovery is done in the clear
  - Don't probe?
  - **Challenge**: Beaconing is often undesirable also
- Implementation and configuration variation
  - Standardize?
  - **Challenge**: Ambiguity of specifications

# Contributions

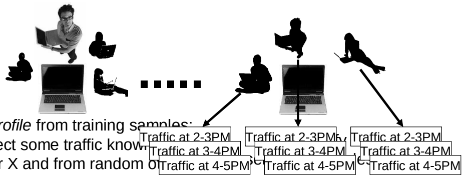
- Four Novel 802.11 Implicit Identifiers
- Automated Identification Procedure
- Evaluating Implicit Identifier Accuracy

# Tracking 802.11 Users

- 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.

## Tracking 802.11 Users

- Tracking scenario:
  - Every users changes pseudonyms every hour
  - Adversary monitors some locations
  - ➔One hourly traffic sample from each user in each location



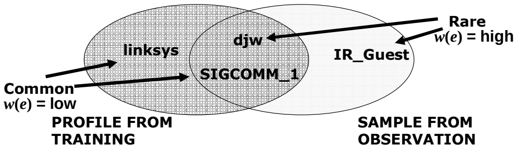
Build a *profile* from training samples:
First collect some traffic know
from user X and from random o

## 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



Set similarity (Jaccard Index), weighted by frequency:

$$feature_{\square}(s) = \frac{\sum_{e \in Profile_{\square} \cap Set_s} w(e)}{\sum_{e \in Profile_{\square} \cup Set_s} w(e)}$$

## Contributions

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

## Evaluating Classification Effectiveness

- Simulate tracking scenario with wireless traces:

| | Duration | Profiled Users | Total Users |
|---|---|---|---|
| SIGCOMM conf. (2004) | 4 days | 377 | 465 |
| UCSD office building (2006) | 1 day | 153 | 615 |
| Apartment building (2006) | 14 days | 39 | 196 |

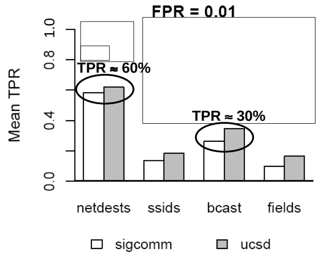
  - Split each trace into training and observation phases
  - Simulate pseudonym changes for each user X

## Evaluating Classification Effectiveness

- **Question**: Is observation sample ***s*** from user X?
- Evaluation metrics:
  - **True positive rate (TPR)** = **???**
    Fraction of user X’s samples classified correctly
    ↑ Measure TPR
  - **False positive rate (FPR)** = **0.01**
    Fraction of other samples classified incorrectly
    ↕ Fix ***T*** for FPR (See paper)

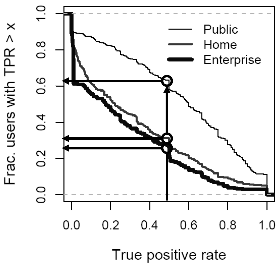
**Pr[ *s* from user X | *s* has features *F* ] > *T***

## Results: Individual Feature Accuracy



Individual implicit identifiers give evidence of identity

## Results: Multiple Feature Accuracy



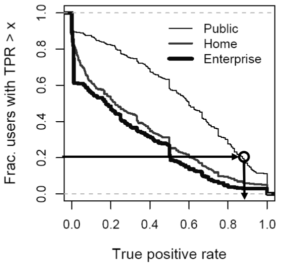
**Users with TPR >50%:**

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

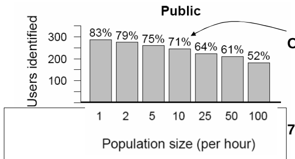
| | Public | Home | Enterprise |
|---|---|---|---|
| netdests | ✓ | | |
| ssids | ✓ | ✓ | ✓ |
| bcast | ✓ | ✓ | ✓ |
| fields | ✓ | ✓ | |

Some users much more distinguishable than others

## One Application

- **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: 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

## Conclusions

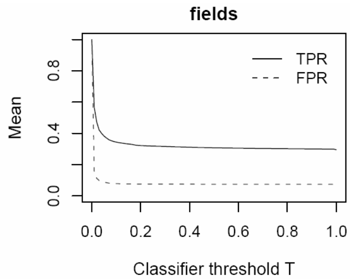
- Implicit identifiers can accurately identify users
  - Individual implicit identifiers give evidence of identity
  - We can identify many users in all environments
  - Some users much more distinguishable than others
- Understanding implicit identifiers is important
  - Pseudonyms are not enough
  - We establish a *lower bound* on their accuracy
- Eliminating them poses research challenges
  - Current work: Confidential service discovery
  - Current work: Traffic analysis resistant MAC

# Extra Slides

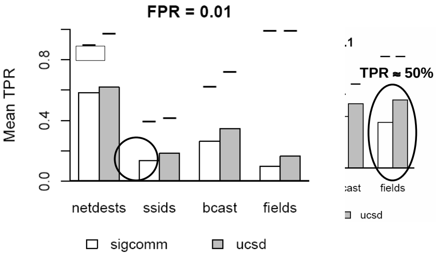
# Related Work

- Other Implicit Identifiers
  - Device driver fingerprints [Franklin ’06]
  - Clock-skew fingerprints [Kohno ’05]
  - Click-prints [Padmanabhan ’06]
  - RF antenna fingerprints [Hall ’04]
- Our work:
  - 802.11 fingerprints for *individual users*
  - Tracking with only commodity hardware/software
  - Better coverage than some previous work
  - Procedure to combine implicit identifiers

# Evaluating Classification Effectiveness

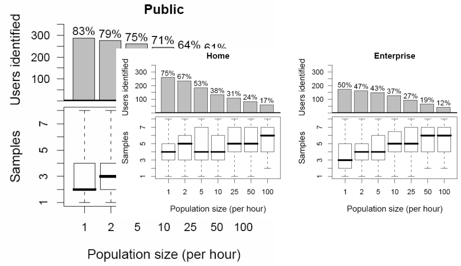


# Results: Individual Feature Accuracy



Other implicit identifiers distinguish groups of users

# Results: Tracking with 90% Accuracy



Many users can be identified in all environments

# Old Slides

# Answers to Common Questions

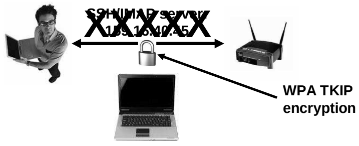
- Is tracking accuracy good enough to track an entire city/how does it vary with population size?
  - Our accuracy results aren't good enough to allow fine-grained tracking of most users in environments with more than 100s of users because the adversary will simply obtain too many samples and have too many opportunities for false positives. However, we note that by modifying the question a little, such as "was this person here for x hours, where x is reasonably large, such as a couple days" or "was this person here in this specific area with smaller population in the sub-100s", we can answer it fairly accurately.
- Your traces were only over a couple days, why would one expect fingerprints to remain identifying over longer periods of time?
  - You are correct and we can't say anything empirical for longer term tracking. However, we suspect that the implicit identifiers would remain identifying because they are mostly caused by automatic behavior of drivers and applications and thus wouldn't change due to human behavior or location. In the paper we show that ssids are stable over weeks at least for example.
- Do you think people really care about location privacy?
  - I think that people don't care until they realize that they have a problem and then they do (e.g., the article about a guy tracking his ex-girlfriend). One magnifying factor is that because of the low cost of 802.11 radios and the ubiquity of 802.11 in devices, almost anyone can easily monitor some locations and track users of interest.
- Don't you think people could modify their behavior in order not to be as identifying when they want to be private?
  - Sure; if they had the technical knowledge to do so, they could easily modify their behavior to "look different" and leave a different implicit fingerprint. However, I would argue that this is undesirable because it means that privacy concerns have a chilling effect on the types of act ivies we are willing to do thus limiting the wireless applications. I think the protocols themselves should protect us enough so that we don't have to worry about these things.
- What about wireless protocols other than 802.11?
  - I believe many of the implicit identifiers still exist, for the reason that the fundamental reasons behind their existence remain: you can still do traffic analysis even with encryption, all wireless protocols have some form of service discovery, and there is implementation variation.

# Outline

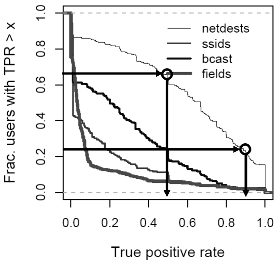
- Problem
  - People are worried about tracking, 802.11 is especially worrisome
  - Pseudonyms proposed, not enough
- Bittorrent Example
  - Use to explain each identifier
  - Summarize implicit identifiers
- How to train as an example
  - Points to bring up and jx:
    - 1 hour sample size
    - How to select the classifier threshold
    - How adversary could obtain training samples
  - Learning process
- Q1 results
- Q2 results
  - An attacker can use multiple features and multiple samples, so ask question…

# Implicit Identifiers by Example

- Implicit identifier: network destinations
  - IP <address, port> pairs in network traffic
  - At SIGCOMM, each visited by 1.15 users on average
  - Some nearly-unique destinations repeatedly visited (e.g., email server)



# 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