• Jeffery Pang
  ➢ Carnegie Mellon
• Ben Greenstein
  ➢ Intel Research
• Ramakrishna Gummadi
  ➢ U of Southern California
• Srinivasan Seshan
  ➢ Carnegie Mellon
• David Wetherall
  ➢ U of Washington
The best practices for securing 802.11 networks, embodied in the 802.11i standard, provide user authentication, service authentication, data confidentiality, and data integrity. However, they do not provide anonymity, a property essential to prevent location tracking.

Problem Statement
- Demonstrate that existing user identification and tracking countermeasures are ineffective
- Highlight four previously unrecognized unique traffic identifiers (*implicit identifiers*)
- Present an automated procedure to uniquely identify wireless users

**Objectives**
The Implicit Identifier Problem

- SSID Broadcast with identifiable names
  - “MIT” or “UniversityofWashington”
- Traffic Patterns
  - Periodic IMAP or SMTP connections to an identifiable mail server
  - Unique repetitive packet/frame sizes
- Design Flaws/Implementation Variances
  - Fingerprinting higher layers in the stack (nmap/p0f)
  - Timing Characteristics

Problem In Detail
• Implicit Identifiers map to physical locations

• Map SSID’s from captures to a 1 block radius
Location Privacy
- RFID devices
- GPS enabled devices

Identity Hiding
- Using pseudonyms to mask MAC addresses (Gruteser, Jiang, Stajano)

Implicit Identifiers
- Fingerprinting 802.11 driver timings (Franklin, Kohno)
- Clickprints (Padmanabhan and Yang)

Related Work
• The Adversary
  ◦ Passive monitoring (weak adversary)
  ◦ Using TCPDUMP only

• The Environment
  ◦ Large and small wireless networks evaluated
    • 2004 SIGCOMM Conference (4 days)
    • U.C. San Diego CS Building (1 day)
    • Apartment Building (19 days)
  ◦ Encrypted (WEP/WPA) and Unencrypted

• Monitoring Scenario
  ◦ Assume pseudonyms are randomly chosen every hour

The Test Bed
Training
sigcomm – 1 day
uscd – 4 hours
apt – 5 days

NOTE: Profiled users are those users who were present in both the training set and the validation data.
• Q1) Did this traffic sample come from user U?
  ◦ Measuring the effectiveness of the classification model
  ◦ User distinctiveness

• Q2) Was user U here today?
  ◦ Can the presence of user U be detected during a particular 8 hour period?

Evaluation Criteria
<table>
<thead>
<tr>
<th>Traffic Characteristics/Identifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Network Destinations (netdests)</strong></td>
</tr>
<tr>
<td>◦ Set of IP &lt;address, port&gt; pairs</td>
</tr>
<tr>
<td><strong>SSID Probes (ssids)</strong></td>
</tr>
<tr>
<td>◦ SSID discovery probes typical of Windows XP</td>
</tr>
<tr>
<td>◦ Preferred networks list</td>
</tr>
<tr>
<td><strong>Broadcast Packet Sizes (bcast)</strong></td>
</tr>
<tr>
<td>◦ Set of &lt;application, size&gt; pairs (or just size if encrypted)</td>
</tr>
<tr>
<td><strong>MAC Protocol Fields (fields)</strong></td>
</tr>
<tr>
<td>◦ More fragments, retry, power management, order, authentication algorithm offered, supported transmission rates</td>
</tr>
</tbody>
</table>
• Naïve Bayes Classifier

From Bayes’ Theorem:

\[
p(C|F_1, \ldots, F_n) = \frac{p(C) p(F_1, \ldots, F_n|C)}{p(F_1, \ldots, F_n)}.
\]

\[
\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.
\]

\[
\text{classify}(f_1, \ldots, f_n) = \arg\max_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i|C = c)
\]

C = class (or U user)
F_i = implicit identifiers
n = 4 in this case
• Feature Generation

To compute probabilities implicit identifiers must be converted to real valued feature

- [fields] – each field combination represents a different value
- [ssids, bcast, netdests] – set of discrete elements
  - Weighted version of Jaccard similarity index to determine real-valued feature

\[
J(A, B) = \frac{|A \cap B|}{|A \cup B|}.
\]

\[
\text{feature}_U(s) = \frac{\sum_{e \in \text{Profile}_U \cap \text{Set}_s} w(e)}{\sum_{e \in \text{Profile}_U \cup \text{Set}_s} w(e)}
\]

Classification Model
• Accuracy measured by two components
  ◦ True Positive Rate (TPR)
    • Fraction of validation samples that user U generates that are correctly classified
  ◦ False Positive Rate (FPR)
    • Fraction of validation samples that user U does not generate that are incorrectly classified

Classifier Accuracy
Classifier Accuracy Metrics

Mean True Positive Rate for a failure rate of 1/100 and 1/10 respectively

Max expected TPR

Complementary cumulative distribution function (CCDF) on sigcomm users
(c) FPR = .01
(d) FPR = .1

Mean achieved TPR and FPR for sigcomm users
x=y line -> random guessing
• How accurately can the evaluation criteria be answered (Q1, Q2)?

Constraints:

◦ Public Network: [netdest, ssids, fields, bcast]
◦ Home Network: [ssids, fields, bcast]
◦ Enterprise Network: [ssids, bcast]
Tracking

Testing the Classifier

In all scenarios the classifier is able to identify unique users with 90%+ accuracy.

Classification accuracy using ‘Public, Home, Enterprise’ constraints

Complementary cumulative distribution function (CCDF)

FPR = .01

% of users that FPR errors that are less than .01 away from target FPR of 0.01

<table>
<thead>
<tr>
<th></th>
<th>% users with FPR error &lt; 0.01</th>
<th>median error</th>
<th>90th percentile error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public</td>
<td>97%</td>
<td>82%</td>
<td></td>
</tr>
<tr>
<td>Home</td>
<td>80%</td>
<td>64%</td>
<td></td>
</tr>
<tr>
<td>Enterprise</td>
<td>79%</td>
<td>68%</td>
<td></td>
</tr>
</tbody>
</table>
The number of users that can be accurately identified is between 50% and 83%.

Results
90% Accuracy

Adversary Detected Users

Min/Max sampled needed

Median active/hours needed to be detected

Active Users in a single hour
The number of users that can be accurately identified is still significant.
• Majority of users can be detected with 90% accuracy in public networks with less than 100 concurrent users
  ◦ 27% are detectable in all networks with less than 25 concurrent users.
  ◦ Even in large networks 12-52% are detectable
• Some users can be detected with 99% accuracy
  ◦ 12-37% in all networks with 25 users or less.

Conclusions
- Ability to identify user’s is not uniform
  - Some users do not display any characteristics that distinguish themselves
  - Majority of users can be tracked with 90% accuracy even when unique names/addresses are removed
- Any one implicit identifier can be highly discriminating
  - An adversary may only 1-3 samples of user’s traffic to track them on average
- Research assumptions serve to place a lower bound on the findings
  - Advanced adversary may have a significantly higher percentage of accuracy
- Applying existing best practices will fail to protect the anonymity of a non-trivial fraction of users
  - Pseudonyms alone are not enough to provide location privacy

**Summary of Findings**
• Similar Research

• Transmission Power Fluctuation

• Pseudonyms

• RSS

• Angle of Arrival

• Time of Arrival

Next Steps