# NETWORK TRAFFIC TRACING AT OSDI 2006

We are a team of researchers who are tracing network activity at OSDI 2006 with a view to making the data available to the research community. We are only recording information that is pertinent to networking research, in a suitably anonymized form. We are *not* recording sensitive information such as the user or client identities or the content of user communication. This handout details what we are tracing and why, how the traces are being processed to protect sensitive information, and whom to contact if you have further questions. Thank you.

## Frequently Asked Questions:

**Q1. What are the goals of this tracing project?**

Our goal is to gather a detailed trace of network activity at OSDI 2006 to enable analysis of the behavior of a wireless LAN that is (presumably) heavily used. In particular, we plan to examine questions pertaining to the nature and causes of wireless network failures as part of the WiFiProfiler project at Microsoft Research (http://research.microsoft.com/projects/NetProfiler/). Besides using this data for our research, we also plan to make the traces available to the research community.

**Q2. Who is gathering the traces?**

The traces are being gathered by a team of researchers from Microsoft Research _ Ranveer Chandra, Ratul Mahajan, and Venkat Padmanabhan _ in coordination with Tony Del Porto (tony@usenix.org) from USENIX.

**Q3. Who has approved this tracing project?**

The tracing plan, including this document, have been reviewed and approved by the OSDI 2006 co-chairs, Brian Bershad and Jeff Mogul, and by the USENIX staff.

**Q4. What is being traced?**

We are recording network protocol information from all packets sent on the air as well as on a wired switch that the access points connect to. The information being recorded for each packet includes physical layer information such as the wireless signal strength as well as the 802.11, IP, TCP, UDP, and ICMP headers, depending on the packet type. We are not recording payloads (packet bodies) except for DHCP and DNS payloads. However, we are anonymizing or deleting potentially sensitive information such as MAC and IP addresses, and DNS names.

**Q5. How is the trace being anonymized?**

We are using a keyed one-way hashing procedure (HMAC-SHA1, in particular) to create a one-way mapping between a potentially sensitive piece of information such as a DNS name and the anonymized version of the information. The key used in this computation will be thrown away once the trace collection is concluded, making it difficult to mount a dictionary attack on the one-way hash. Please see the appendix below for more information on the anonymization process.

**Q6. Will the packet payload be captured or stored?**

Packet payload will be recorded for DHCP and DNS requests and responses. However, information such as DNS names and IP addresses contained in the payload will be anonymized before being stored.

**Q7. Will my activities be identifiable?**

Given that the traces are being anonymized, we believe that it would be extremely difficult for anyone to identify users or learn which Internet services or hosts they have communicated with. That said, we are not in a position to prove that no such information can be gleaned from the anonymized traces.

**Q8. What will be done with the anonymized data? Who will have access?**

The anonymized traces will be made available to the research community, for example, through a repository such as CRAWDAD (http://crawdad.cs.dartmouth.edu/). We plan to make the data available within a month or so after OSDI 2006.

**Q9. Will any non-anonymized data be stored?**

The traces will be anonymized on-the-fly before they are stored on disk. However, certain information, such as the first 3 bytes of the MAC address, may turn out to violate the principle of *k*-anonymity (explained in Q11 below). If so, we will further anonymize the trace offline before anyone else sees it; this kind of anonymization cannot be done online.

**Q10. Who will have access to the non-anonymized data, and for how long?**

As stated in Q9, much of the anonymization will be performed on-the-fly, so no one should have access to the non-anonymized data, given that we intend to keep the tracing system as secure as possible. However, as noted in Q9, some of the anonymization can only be done offline, so the Microsoft researchers listed in Q2 will have access to the partially anonymized data during the time it takes to perform the offline anonymization (no more than a few days after the trace collection is concluded).

**Q11. What identifiable information could still be extracted from the final anonymized trace?**

It may be possible to identify users using a side-channel attack, for instance, by exploiting information such packet sizes and packet timing; we do not plan to protect the data against such attacks.

Also, we would like to permit the identification of the manufacturer of a wireless NIC (which could be useful when analyzing the traces), so the first 3 bytes of the MAC address will be left unanonymized. However, this could violate the principle of *k*-anonymity, i.e., that it should not be possible to identify any user as being a member of a group with fewer than *k* members. If a group size is smaller than 10, our offline anonymization will replace this MAC-address prefix with another value so as to create a group of at least 10 nodes (i.e., we set *k* to 10). So it would be possible to identify the 3-byte prefix of a node's MAC address provided that there are at least 10 nodes that share the same prefix.

**Q12. How should I protect my data and identifiable activities if I use the wireless network?**

As noted above, we are taking every care to obscure sensitive information while still leaving the traces useful for research. However, we have no control over who else might be sniffing on the network traffic, even though such sniffing is against the terms of use for the OSDI wireless network. Since this is an ever-present danger, especially in wireless networks, we strongly recommend that you use secure protocols and procedures for communication (e.g., SSL, SSH, VPN). That said, we are not in a position to provide definitive advice on how best to protect yourself when using a wireless network. You would have to consult your IT staff regarding this.

**Q13. Whom should I contact if I have further questions about this tracing project?**

Please contact Venkat Padmanabhan (padmanab@microsoft.com).

## Appendix: Technical Details

We are gathering traces of wireless traffic at several monitoring nodes distributed across the conference floor and breakout areas. In addition, we are gathering traces on the wired switch to which the wireless access points connect.

Here is a description of the traces we are gathering and the anonymization that is being performed (the anonymization will be performed on-the-fly except where noted otherwise). Our description here focuses on tracing on the wireless LAN. A subset of this (viz., everything above the PHY layer) also applies to the tracing on the wired LAN.

**What traffic is being monitored?**

Each monitor will capture all of the 802.11 frames it sees, including:

1. Data frames
2. Management frames (e.g., association, authentication)
3. Control frames (e.g., RTS, CTS, ACK)

**What information is being logged?**

For each wireless frame captured at a monitor, we record the following information:

1. Per-frame PHY information, including:
   a. Channel frequency
   b. RSSI
   c. Modulation rate

2. Entire MAC header, with only the source and destination MAC addresses being anonymized as follows:
   a. In real-time, the first 3 bytes of the MAC address will be copied over as is. The last 3 bytes are replaced with a one-way hash.
   b. Offline, we replace all the 3-byte MAC prefixes that occur fewer than 10 times with a common prefix. This ensures *k*-anonymity, for k=10.

3. The entire IPv4 and TCP/UDP header, with the source and destination IPv4 addresses anonymized as follows:
   a. The IP address is replaced with a one-way hash.
   b. In addition, we record whether the IP address belongs to the following categories:
      i. Auto conf (169.254/16).
      ii. Private address space (10/8, 172.16/12,192.168/16).

4. The entire DHCP payload, with the following anonymization:
   a. Client IP address (ciaddr) is anonymized as in 3.a.
   b. Client hardware address (chaddr) is anonymized as in 2.
   c. Your IP address (yiaddr) is anonymized as in 3.a.
   d. The "client identifier" option, if present, is replaced with a one-way hash.

5. The DNS request/response payload, with the following anonymization/deletion:
   a. The domain name in each RR is replaced with a one-way hash.
   b. The resource data contained in each RR is deleted.

**Security and privacy issues:**

1. We have taken reasonable measures to secure the machines used for tracing: kept them up-to-date on patches, turned off unnecessary services, protected access with a strong password, etc.

2. We will throw away the secret key used for the keyed one-way hash once the trace collection is concluded to make it difficult to perform a dictionary attack on the one-way hash.

3. Despite the anonymization, it may be possible for some information to leak. For example, it may be possible to infer which website was visited based on the size of the response received. We are unable to obfuscate such information without damaging the data significantly.