

M2C* Deliverable D1.5

M2C Measurement Data Repository

Remco van de Meent

r.vandemeent@utwente.nl

22nd December 2003

University of Twente, Enschede
The Netherlands

Abstract

This report describes a repository of network traffic measurement data, collected at various institutes in the Netherlands. The measurement data consists of *anonymized packet header traces*, and is available for download. The repository can be found at this location: <http://m2c-a.cs.utwente.nl/repository/>.

**Measuring, Modelling and Cost Allocation (M2C)* is a project sponsored by the Dutch Telematica Instituut. Partners in this project are the University of Twente (through its CTIT research institute) in Enschede, and the National Research Institute for Mathematics and Computer Science (CWI) in Amsterdam.

1 Introduction

One of the objectives of the *Measuring, Modelling and Cost Allocation (M2C)* project is to get an understanding of the characteristics (e.g., throughput on small time scales) of network traffic on the Internet. This objective is (partly) achieved by performing detailed traffic measurements on various institutes. Other researchers may have other questions about Internet traffic they want to answer.

Network traffic on the Internet looks, in detail, different on various measurement points. These differences are caused by local circumstances, such as the number of users, link speeds, applications that are used, etc. Researchers studying Internet traffic want to (in)validate hypotheses in various scenarios, but not all have access to as much different scenarios as they would like to have. Thus there is a need for data sharing, achieved by (semi-)public repositories of network traffic, measured on various places on the Internet.

1.1 Structure

The remainder of this report is organized as follows: 1.2 points to some related work, both within and outside the M2C project. Section 2 presents the measurement procedure, and points out how the privacy of users whose network traffic is used, is protected in the repository. Section 3 provides information about the different scenarios of which data currently is in the repository.

1.2 Related Work

Within the M2C project, the traffic repository is used for all research involving real network traffic (as opposed to simulation or purely theoretical models). For instance, in the visualization tools [1], the traffic repository is used as a source for live (packet-level) data, which is converted into flow-level information that is stored in a database that the visualization tools use.

Around the world, some other research group also share measurement data, in various formats.

There are the Packet traces from WIDE backbone in Japan [2], the Internet Traffic Archive [3] containing several TCP traces (some covering special things like the web-server of a large sporting event), traces from NLANR [4] and a list of different traffic traces put together by Schulzrinne [5].

2 Measurements

The M2C traffic repository is currently filled with measurement data from 3 institutions in The Netherlands. In this section we describe how the data is collected (2.1), and how privacy is protected (2.2).

2.1 Data Collection

The measurements are performed by capturing the headers of all packets that are transmitted over the (Ethernet) “uplink” of an access network to the Internet, as outlined in Figure 1. The switch (can also be a router) copies all traffic flowing in to and out of the access network to the measurement PC. The configuration of the measurement PC is listed in Table 1. Even with such a (moderate) PC we are able to handle a load on the uplink of several hundreds Mbit/s [6]. The tool that has been used on the measurement PC to capture packets is the standard tcpdump [7] utility.

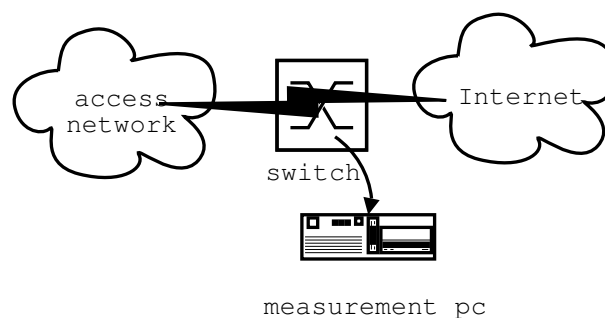


Figure 1: Measurement Setup

Tcpdump is run for 15 minutes, generating a binary file that is stored on disk, containing a *packet trace*: a dump of the headers of all packets that have been transmitted over the up-

Component	Specification
CPU	Pentium-III 1 GHz
Mainboard	Asus CUR-DLS (64 bit 66 MHz PCI)
Hard disk	60 + 160 Gigabyte, UDMA/66
Operating system	Debian Linux, 2.4.19-rc1 kernel
Network interface	1 x Gbit/s Intel Pro/1000T
Main memory	512 MB reg. SDRAM

Table 1: Measurement PC Configuration

link in that period. As we want to capture everything up to the transport (e.g., TCP) header, whilst ignoring the payload of the packet, we have capture the first 64 octets of each Ethernet frame.

The resulting packet trace is a file of possibly several gigabytes, depending on the load of uplink. In order to save resources, the traces are compressed. This saves on average some 60% in disk space.

2.2 Anonymization

The headers in the packet trace include source and destination IP addresses and port numbers. Although the payload of the IP packets is discarded, careful analysis of the packet trace still may reveal possibly sensitive information, such as which websites are visited by who, which threatens users' privacy. On the other hand, removal of addresses etc. from the packet traces severely reduces their usefulness. Thus there is a trade-off to be made between protecting privacy and usability of the traces.

Hence, to protect users' privacy, the packet traces are made anonymous, by scrambling the source and destination IP addresses, using the `tcpdpriv` [8] utility. This process is called *anonymization*. Other information, such as transport port numbers and the timestamps at which packets arrive are left unchanged.

Tcpdpriv is called using the following options:

```
tcpdpriv -A50 -P99 -r original_packet_trace -w anonymized_packet_trace
```

The “-A50” option ensures, within a single packet trace, that if two of the original addresses are equal in the most significant n bits, then these two addresses will map to scrambled addresses that are similarly equal in the most significant n bits. For example, if source address $a.b.c.p$ is mapped to $x.y.z.k$, then $a.b.c.q$ is mapped to $x.y.z.l$. A possible drawback of this approach, however, is that some topological information might be revealed, whereas strict random mapping would not.

The effect of the “-P99” option is that transport port numbers are unchanged, e.g., if an original packet was sent from TCP port 1025 to port 80 (i.e., web-browsing), the same port numbers will be stored in the anonymized packet trace.

All the packet traces that are available from the M2C traffic repository are made anonymous following the procedure outlined above.

3 Repository

The M2C traffic repository available at this location:

```
http://m2c-a.cs.utwente.nl/repository/
```

It currently contains several hundred traces, measured at three different locations for 3 organizations, various times of the day, 7 days per week. Each trace contains 15 (real) minutes worth of packet headers.

Within the repository, the structure is as follows:

```
<organization>/<org>-<date>-<time>.gz
```

The date and time refer to the local time at which the measurement period of 15 minutes has started. After uncompression, the files can be processed using `tcpdump/libpcap`.

In the remainder of this section the three locations are briefly described, to give the context of the traces.

Location #1

On location #1 the 300 Mbit/s (a trunk of 3 x 100 Mbit/s) ethernet link has been measured, which connects a residential network of a university to the core network of this university. On the residential network, about 2000 students are connected, each having a 100 Mbit/s ethernet access link. The residential network itself consists of 100 and 300 Mbit/s links to the various switches, depending on the aggregation level. The measured link has an average load of about 60%.

Location #2

On location #2, the 1 Gbit/s ethernet link connecting a research institute to the Dutch academic and research network has been measured. There are about 200 researchers and support staff working at this institute. They all have a 100 Mbit/s access link, and the core network of the institute consists of 1 Gbit/s links. The measured link is only mildly loaded, usually around 1%.

Location #3

Location #3 is a large college. Their 1 Gbit/s link (i.e., the link that has been measured) to the Dutch academic and research network carries traffic for over 1000 students and staff concurrently, during busy hours. The access link speed on this network is, in general, 100 Mbit/s. The average load on the 1 Gbit/s link usually is around 10–15%.

References

- [1] R. van de Meent and A. Pras, “Visualization Tools,” tech. rep., University of Twente, November 2003. M2C Project Deliverable 1.3.
- [2] MAWI Working Group, “Packet traces from WIDE backbone,” 2003. <http://tracer.cs1.sony.co.jp/mawi/>.

- [3] P. Danzig, J. Mogul, V. Paxson, and M. Schwartz, “The Internet Traffic Archive,” 2000. <http://ita.ee.lbl.gov/index.html>.
- [4] NLANR Measurement & Operations Analysis Team, “NLANR network traffic packet header traces,” 2003. <http://pma.nlanr.net/Traces/>.
- [5] H. Schulzrinne, “Traffic Traces,” 2003. <http://www.cs.columbia.edu/~hgs/internet/traces.html>.
- [6] R. Poortinga, R. van de Meent, and A. Pras, “Analysing campus traffic using the meter-MIB,” in *Proceedings of the Passive and Active Measurement Workshop (PAM2002)*, (Fort Collins, Colorado, U.S.A.), pp. 192–201, March 2002.
- [7] Lawrence Berkeley National Laboratory Network Research, “TCPDump: the Protocol Packet Capture and Dumper Program,” 2003. <http://www.tcpdump.org/>.
- [8] Ipsilon Networks, “tcpdpriv,” 1997. <http://ita.ee.lbl.gov/html/contrib/tcpdpriv.html>.