Flow Data Analysis in SWITCH / ETH Zurich Project DDoSVax

Arno Wagner

wagner@tik.ee.ethz.ch

Communication Systems Laboratory

Swiss Federal Institute of Technology Zurich (ETH Zurich)

Talk Outline

- The Dataset
- Flow Data Usage by SWITCH
- Offline Analysis Examples
- Traffic Amount vs. Unique Addresses
- Analysis Tools
- Performance questions

The DDoSVax Dataset

Project URL:
`http://www.tik.ee.ethz.ch/~ddosvax/`

- NetFlow v5 (converted from V7 by SWITCH)
- About 60.000.000 flows/hour
- Weekday: About 200k internal and 800k external IPs
- Unsampled
- Stored in full since March 2003

Flow Data Usage by SWITCH

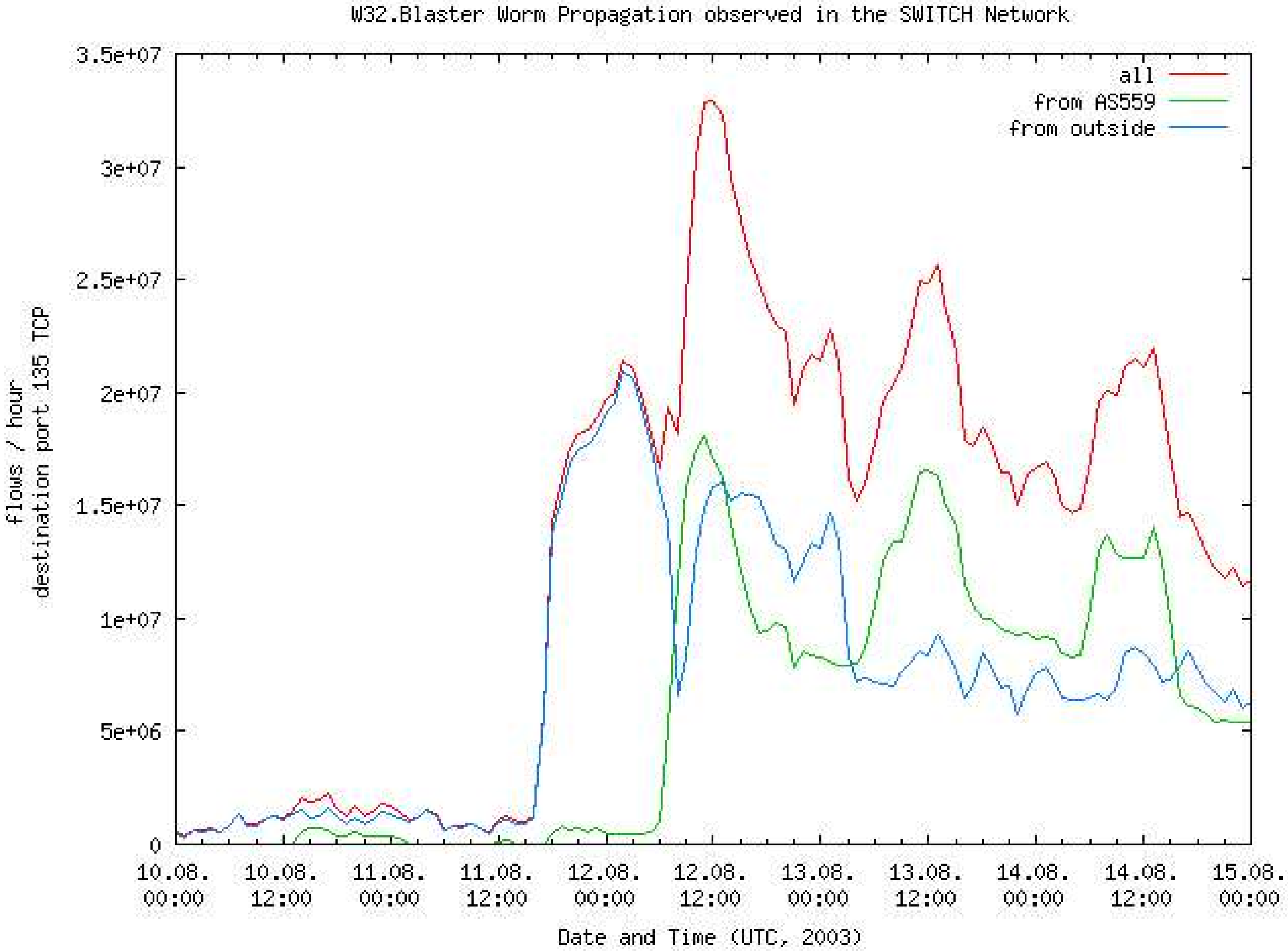
Independently done by SWITCH on NetFlow data

- Accounting and load monitoring (aggregated)
- SWITCH-CERT: Short-term forensics (reduced)
  - Single fast computer with hardware RAID-5
  - No compression
  - Sorted into minute (?) intervals
  - Fast search with regular expressions
  - Several weeks online
  - No (?) long term storage

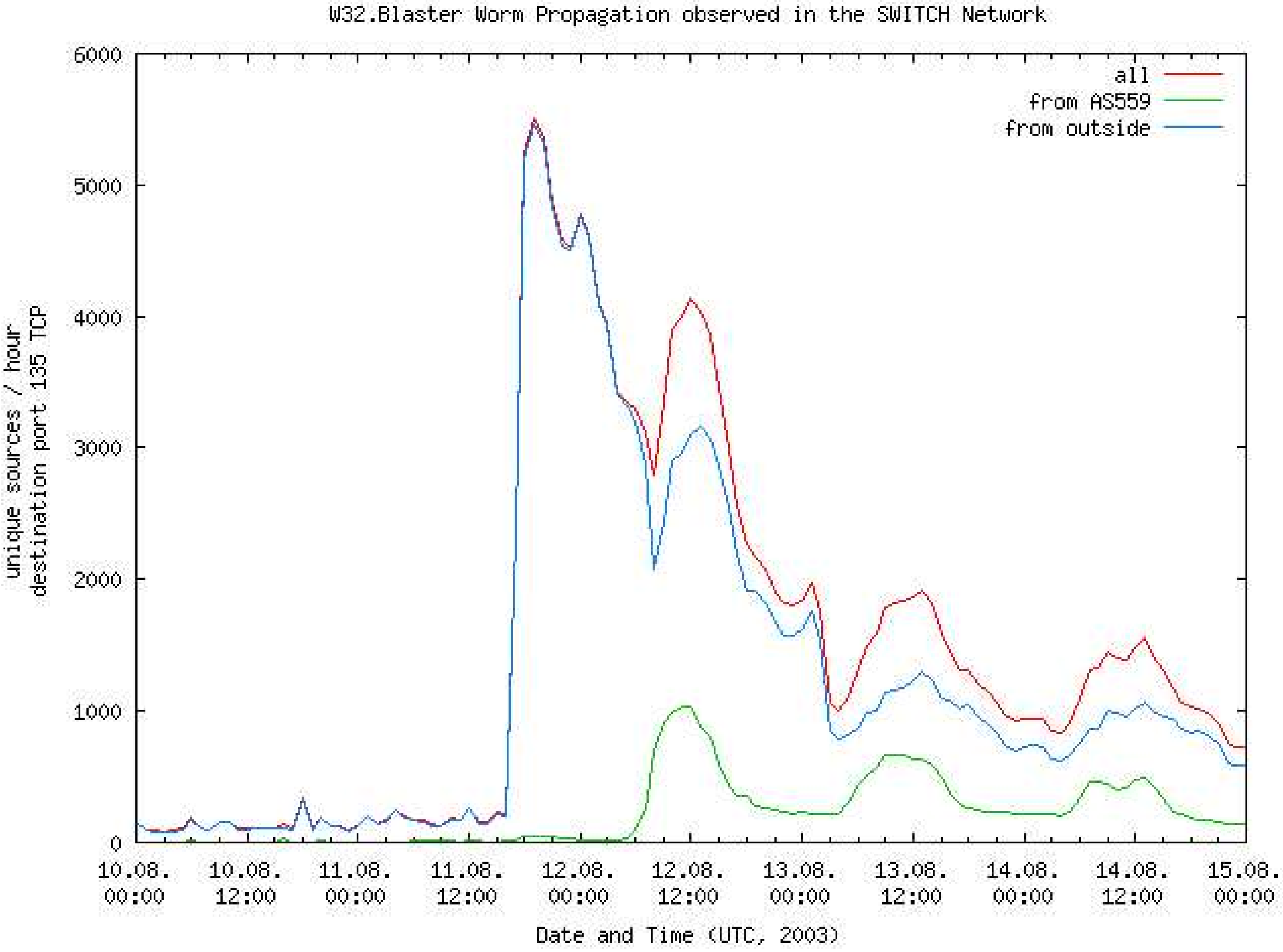
Offline Analysis

- E.g. for network/email worms
- Customised tools for some analyses
    - Single hour / prototyping: netflow_to_text and Perl
    - Days...weeks: From C-template
- Also other things: P2P, IRC, ...

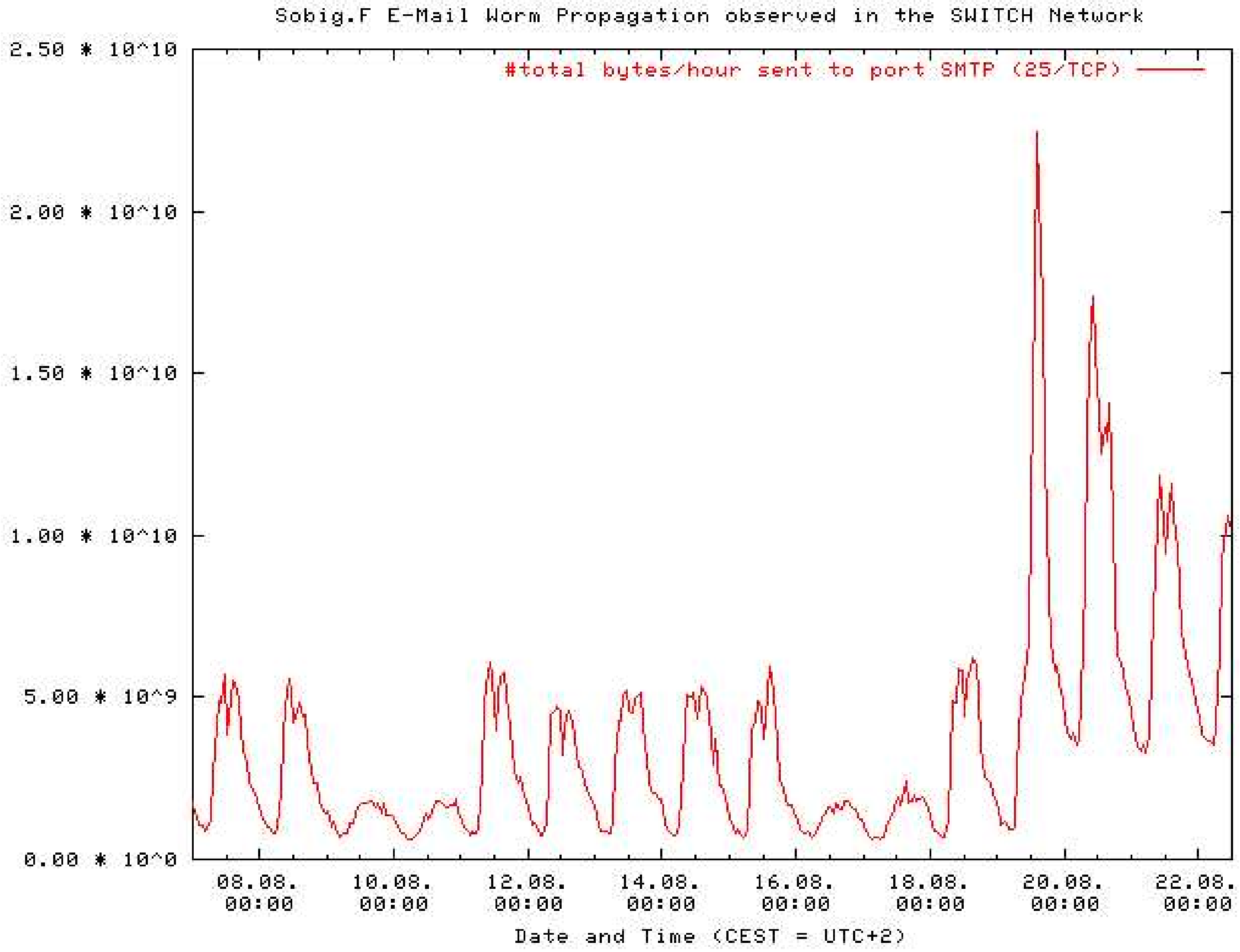
Example: Blaster - Flows



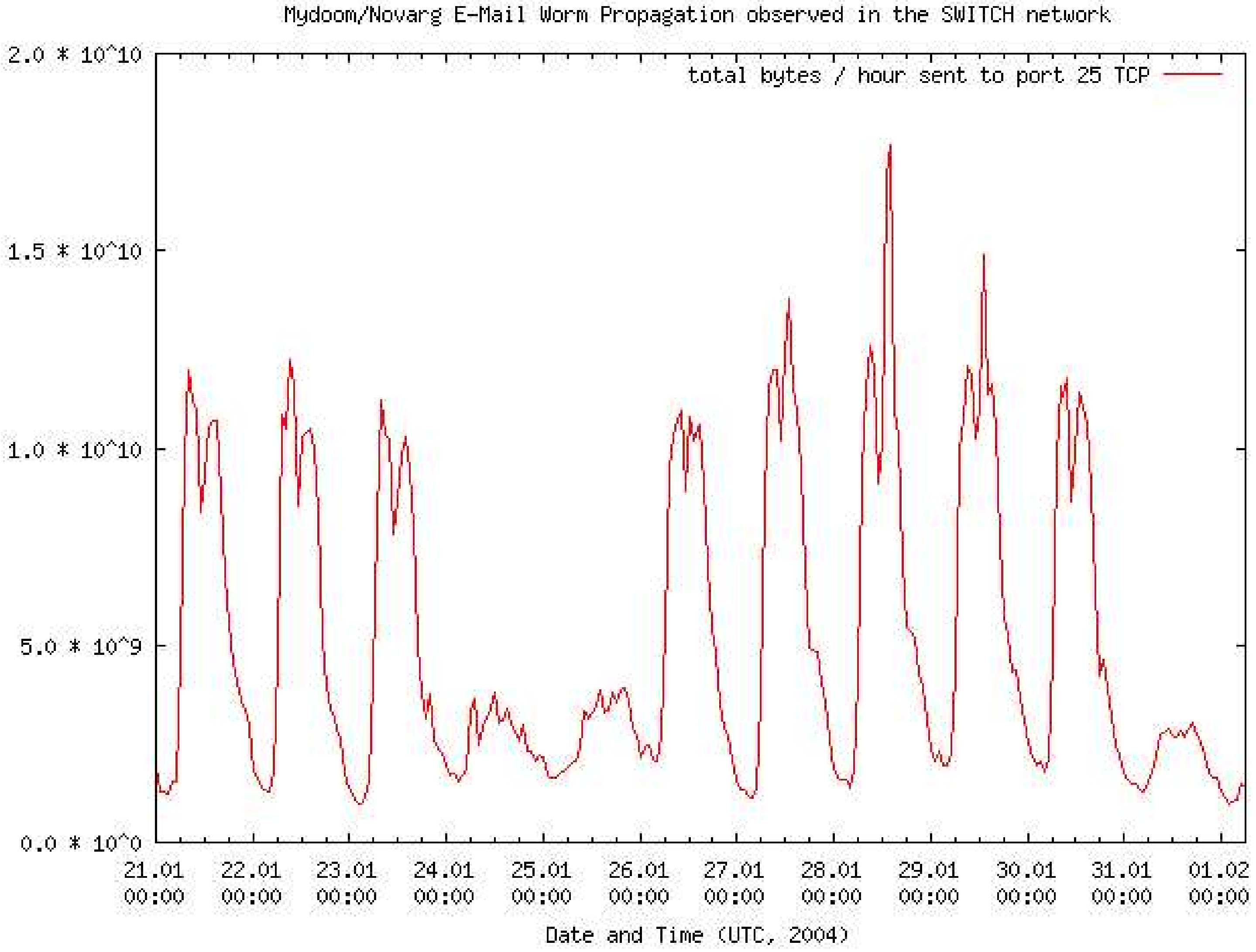
Example: Blaster - Unique Sources



Example: Sobig



Example: MyDoom



Traffic vs. Unique Sources

Traffic:

- Easy to do
- Works reasonably well
- Sensitive to data generation problems
- Sensitive to observed network

Unique Sources:

- More complicated, more robust
- Weakly dependent on observed network
- Allows to get global picture

Analysis-tools: Scripting

”netflow_to_text”

- Takes one data file, outputs one line
- Well suited as ”grep”/Perl input

Example:

```
TCP pr 111.131.210.8 si 1111.136.200.121
di 1264 sp 135 dp 48 le 1 pk
12:59:51.965 st 12:59:51.965 en 0.000 du
```

Analysis-tools: C

”Iterator template”

- Iterates over all records in a set of files
- Preprocesses timestamps, etc.
- Reading of input files encapsulated

Performance Issues

- 5-10 minutes / hour of data bunzip2
- I/O limit at 10 cluster nodes reading from one NFS partition
- Memory limitations