Bare-Bones Measurement Data Archiving

Dave Plonka
University of Wisconsin – Madison
DoIT & WAIL

ISMA @ SDSC, June 3, 2004

Overview

- Our Data
- Archiving
- Namespaces
- Annotations
- Encoding / Anonymization / Obfuscation
- Access & Usage Policy
- Thoughts
- Tools

Our Data

- Passive:
  - Exported flow data
  - SNMP-gathered measurement data
- Active:
  - Some traceroute and ping-like text output
  - “show ip bgp” (from routeviews, campus routers)
- Flow data:
  - Packet-sampled flow records from Juniper
    - Varying sample rates, varying regularity
  - Non-sampled flow-data from Ciscos
    - Sometimes lossy, always voluminous

Archiving

- Short-term:
  - “raw” (binary) flow files, sometimes compressed
  - Random access to five-minute interval, sequential access to (unpredictably) ordered flows there-in
  - Usually retain for only 5-14 days (why? It's for operational use, storage space limited, open records law.. hmm.)
- Long-term:
  - Round-Robin Database (RRD) files
  - Occasionally copy raw flow files to tape for specific studies

Namespace

- We have used a directory hierarchy with “reversed” DNS of hostnames of the exporters or observation points:
  - edu/wisc/net/r-peer/...
- Complication: names in this space must change when anonymization is performed. One method is to create a script of shell commands (that is anonymized with the data) that will rename them
- Afterward, eg.:
  - mv 10\.42\.69\.10_log.txt 10.42.60.10_log.txt

Annotations

- We (ok, I) create detailed README files (!) in each directory containing the data.
- We maintain a journal / log of events, as “events.txt”:
  - eg. 2004/06/03 1600 something happened thru 1730
  - These events are web browsable using RRGrapher
- Flow file naming convention:
  - {collector}.{date}.{time}{TZ}[_{encoding}.{fmt}]
  - ft-v05.20040603.160000+0500_tcpdpriv-A50.cflow
  - ft-v05.20040603.160000+0500

Encoding / Anonymization / Obfuscation

- ip2anonip: simple filter for CSV files
- Pros:
  - People (and flow-{export,import}) grok CSV
  - Easy to add arbitrary field rewrites (such as aut-num, ifIndex, etc.)
- Cons:
  - Performance: hours to prep a day-long flow data set
  - Tedious:
    - one way to get it right, lots of ways to get it wrong
    - encode, examine, correct, repeat
  - Result depends on order of IPv4 addresses in input
  - Known attacks... better to use CryptoPAN?

Access and Usage Policy

- Tried NLANR/CAIDA? model c. years ago:
  - Usage agreement document, recipient signs-off
  - Data (and therefore analysis) resides on central server
- In theory: release as little as possible, but no less
  - Ask researcher to “apply” for access by describing the project
- In practice: increased levels of access with improved (trust) relationships between researcher and practitioner (creator/archiver).
  - The older the data the better (safer to release)?
- Result (IMO): minimally successful, time-consuming, not scalable

Thoughts

- Useful to store multiple encodings of same data:
  - Anonymized version more accessible than original
  - Follow-up questions can be asked of privileged users
- Canonicalize network element names (data set names?) in parallel with encoding:
    - r-peer.net.wisc.edu => border.our.domain
    - r-cssc-b280c-1-core.net.wisc.edu => core.our.domain
- We often find an anomaly in sampled data then drill-down into the non-sampled data based on point in time. Can this be accommodated in UI?

Tools

- Flow-tools: flow-import, flow-export, flow-stat
- perl: Cflow.pm (mnemonic: “See flow [data]”)
    - http://net.doit.wisc.edu/~plonka/Cflow/
  - flowdumper
- Visualization (browse by annotations):
  - RRGrapher (browser for RRDs)
    - http://net.doit.wisc.edu/~plonka/RRGrapher/
- Anonymization:
  - ip2hostname: 10.42.69.10 => host1.our.domain
    - http://net.doit.wisc.edu/~plonka/ip2hostname/
  - Ip2anonip -A50: 10.42.69.10 => n.x.y.z
    - http://net.doit.wisc.edu/~plonka/ip2anonip/