SIMR

Collecting useful metadata

http://www.icir.org/mallman/papers/simr-pam2002.ps

http://www.cs.purdue.edu/homes/eblanton/slides/isma-elb-0406.pdf

Ethan Blanton
eblanton@cs.purdue.edu

SIMR Overview

- Hand-waving forerunner of the IMDC
  - Mark Allman, Ethan Blanton, Wesley Eddy. A Scalable System for Sharing Internet Measurements. Proceedings of the Passive and Active Measurement Workshop, March 2002.
  - http://www.icir.org/mallman/papers/simr-pam2002.ps
  - Big thanks to CAIDA for turning some vague text into a product!
- Stores only metadata
- **Datatypes have administratively defined schema**

Schema definitions

- Schema definition seems to be the crux of the project
- Determining what is ‘‘useful’’ turns out to be tricky
- Getting this right is Really Important

Administrative definition

- Maximizes consistency
  - Intended to make searching more effective
  - We’ve all seen what happens with, e.g., unrestricted ‘keyword’ fields in databases
- Loses flexibility
  - This is why Getting it Right is so critical

Why it’s so hard

- Details of measurement collection or manipulation may be both invisible and critical to the task at hand

- Examples:
  - Anonymization/sanitization
  - Capture network or machine’s purpose and conditions
  - Large measurements broken up in some fashion
  - Selective packet sampling

Example: anonymization

- May be irrelevant
  - Studying the behavior of individual TCP transfers
- May be ‘‘sort of’’ relevant
  - Perhaps prefix-preserving transformations are OK
- May be critical
  - Topology studies
  - Eliminating local traffic

Example: anonymization (cont.)

- Annotating the specific anonymization method is hard
- Even harder when multiple measurements are involved
  - Multiple measurements using the same mapping
  - Using different mappings but having overlapping hosts
- Different studies are likely to care about different facets of the transformation

Example: bizarre conditions

- Host is behind a satellite phone
- Network is behind a mobile router
- Host is on Mars

Example: selective sampling

- “Simple” filters
  - tcp port 80
- Time-based sampling
  - The first 5 minutes of every hour
- Other types of slices
  - Every nth packet
  - The first packet of every TCP connection
  - ...

Other dangers

- We want to store metadata about data
  - This puts metadata about results explicitly out of scope
  - Where is the line between data and results?
- Database pollution
  - Can schema definitions be used to reduce this?
  - What about ‘‘meta-pollution’’?
- User interaction for individual data items doesn’t scale
  - Or, as Mark says, "reading cruddy READMEs doesn’t scale"

Solutions

- Careful enumeration of interesting characteristics
  - Future-proofing is hard
  - If we knew all of the interesting characteristics, we'd be doing the study ourselves
  - Searches become easy
    - ''Prefix-preserving anonymized traffic with identified local links''
- Free-form comment structure
  - Future-proof by definition
  - You say ''sanitize,'' I say ''anonymize''
- A middle ground

Solutions (cont.)

- Insert your ideas here

Comments?

(It doesn’t say questions because I don’t have the answers)

http://www.icir.org/mallman/papers/simr-pam2002.ps

http://www.cs.purdue.edu/homes/eblanton/slides/isma-elb-0406.pdf