Issues and Etiquette Concerning Use of Shared Measurement Data

Mark Allman (*ICSI*)
Vern Paxson (*ICSI & LBNL)*

Internet Measurement Conference
October 2007

"Country roads, take me home
To the place I belong"

Introduction

- Collecting a substantive set of network measurements is tough
- Many reasons to share collected data (see [4,16,9])
- *We strongly encourage widespread sharing of measurement data*
- However, sharing data and using shared data must be done carefully

Scope

- This talk is not focused on specific incidents
  - We do not live up to our guidelines
  - See section 3.3.1 for a case study that shows
    - These are not theoretical issues
    - Community lacks a unified view of issues

- The goal of this conversation is to look to the future

Releasing Data

- Goal of data release is to *minimize exposure* while *maximizing research usefulness* within some context

- Ultimately a policy decision (see [13])

Releasing Data (cont.)

- Consideration #1: Providers should carefully understand the threat model and apply an appropriate anonymization policy
  - Anonymization depends on context
  - Lots of tools developed to help

Releasing Data (cont.)

- Consideration #2: Data providers should provide an explicit Acceptable Use policy
  - The term "scholarly use" means different things to different people, so be specific
    - E.g., "user of data should not attemp de-anonymization"
    - E.g., "the shared data is not a general resource, but given for task X"
  - A statement does not prevent inappropriate use, but gives the community a lever

Releasing Data (cont.)

- Consideration #3: Data providers should be explicit about the interactions they are willing to have with data users
  - Often a broader context helps solve puzzles
    - E.g., topology information
    - E.g., explaining site-specific traffic patterns
- Consideration #5: Data providers should indicate what sort of raw data is retained and for how long

Releasing Data (cont.)

- Consideration #5: Data providers should indicate what sort of notification or acknowledgment they desire regarding its appearance in a publication

Using Data

- Prudence is required when *using* shared data
  - Organizations already have an immense list of legitimate reasons to withhold data, we don't need to provide more
  - *Data is our gold*

Using Data (cont.)

- Consideration #6: Be careful in reporting sensitive phenomena found in shared data
  - E.g., developing new scanning detectors could show scanners how to evade current scheme
  - Balancing act: Learning something new vs. data sensitivity
- Coping strategies:
  - Aggregate
  - Further anonymize

Using Data (cont.)

- Privately shared data needs to be used as carefully as publicly released data (or, more so)

Using Data (cont.)

- Consideration #7: Researchers *must not* redistribute non-public data
- Consideration #8: Exercise great care in storing non-public data
- Consideration #9: Researchers should only use the data for the purpose it was given and not treat it as a general resource
- Consideration #10: Researchers should seek permission before using privately shared data for another purpose

Using Data (cont.)

- De-anonymizing data can have serious repercussions for the community and the data provider
- *De-anonymization should not be undertaken as sport*
- De-anonymization has its scholarly place
  - Better understanding of what can be inferred leads to better anonymization techniques

Using Data (cont.)

- Consideration #11: De-anonymization should only be undertaken with the explicit permission of the data provider
  - Takes data sensitivity into account
  - Leads to better science
- Coping strategy: Analyze the *anonymization techniques* not the *anonymized data*

Using Data (cont.)

- Consideration #12: Researchers should respect provider's acknowledgment and notification wishes
- Often provides the "ammo" necessary for data release

Interactions

- Communication between providers and users is important
- When in a grey area: *ask!*

A Plea

- Read the paper
- Have your students read the paper
- We do not claim to have all the answers, but the conversation is important

Questions?

Discussion?