Other Peoples' Data:
Pitfalls and Requirements

John McHugh

Canada Research Chair in Privacy and Security
Dalhousie University, Halifax, NS, Canada
3 July 2009

Premise

In many scientific disciplines, primary data collection and cleansing consumes the vast majority of the project budget. In many areas of computer science, especially network research, primary data collection is viewed as infeasible and research is attempted using third party datasets.

- As a result many studies are badly flawed.

The purpose of this talk is to discuss the some of the technical issues involved in using and providing third party data.

- The talk will be illustrated with several anecdotal case studies

Reuse

At one time code reuse was supposed to be the “silver bullet” of software engineering. Code reuse implies use in a context that differs substantially from the original.

- Developing for reuse requires additional effort in specification, development, testing, and documentation.
- The original developer is often unable to pay the cost.

Similar considerations apply to data reuse. In addition,

- Data may be collected but not used by the collector.
- There may have been no intention to reuse the data.
- Data may be badly flawed due to lack of attention to detail or other collection / collector failures.

Case 1 - Lincoln Lab IDS data

The 1998-2000 DARPA IDS data (-> KDD Cup 99) is still being used to evaluate IDS. It suffers from:

- Lack of variability compared to real data
- Freedom from noise
- Poor design and documentation
  - Said to be similar to real AF base traffic, but criteria (except word frequencies) not known.
  - Injected attacks have unique but irrelevant features making for trivial attack recognition.

We now know (suspect we know?) how hard it is to create good artificial data and to inject attacks.

Lincoln data is still used!

Lincoln data now a decade old. The nature of the internet and attack traffic has changed. Why is it used?

- Because it is there.
- Because attacks are labeled.
  - Axelsson spent 6 months manually labeling logs.
- Because researchers are incapable of collection.
  - Technical, organizational, legal barriers.
  - Many machine learning folks don't know much about machines or networks.
- Because they don't know any better.
  - Most CS education fails to include experimental design, analysis, and statistical training.

Artificial *vs.* Real data

The Lincoln data is artificial. The producers apparently thought that word frequencies were the key to IDS.

- Artificial data sets can be useful if they can be controlled for the features that are relevant for the task and
- Include features from the real world that are likely to confound the results

Lincoln failed on both counts. In general, these goals are hard to achieve. Ultimately, a field deployment is necessary - preferably before results are published.

Many of the problems associated with the Lincoln data might have been avoided if Lincoln had produced a paper describing the methods used to design and produce the data before releasing it.

Predict - a possible solution

In response to the Lincoln problem, DHS set out to build a repository of network data to support research.

- Legal and compliance issues underestimated. ***No other computer science data repository has directly confronted these issues!***
- Available only in US (to anyone there) with MOA
- Currently, account (requires sponsoring organization) necessary to see catalog; MOA necessary for data.
- Minimal data available in catalog.
  - This is in contrast to available clinical trial and other health related data sets.
  - See the next 3 slides plus discussion

Legal issues (an aside)

In medical experiments and many CS experiments with direct human participation, subjects are recruited into studies and give informed consent for participation.

- This usually allows the data to be shared among researchers. Subject identity is usually not a key variable.
- For computer science studies involving network observations, recruitment is difficult or impossible and there is often a question as to who the subject is.
  - Privacy laws may apply and there is a possibility of criminal sanctions. Strange exceptions apply but may limit data sharing, retention, etc.
  - Predict is grappling with these issues, painfully.

Predict catalog entry

Category: Enterprise Data from Internal LBNL networks

Hosted By: LBNL

Short Description: Anonymized Enterprise Packet Traces

Long Description: These datasets contain anonymized packet header traces from inside a medium-sized enterprise network. The main goal of this collection is to provide an example of benign background traffic as seen inside an enterprise. While the traffic includes some scanning and probing activity (much from the site's own internal scanning), it is believed to be free of traffic from internally compromised hosts.

Size: 2.0 MBytes

Formats: pcap (Packet Capture library)

Anonymization: Hashing, Prefix Preserving

BACH

https://www.niddkrepository.org/

The Boston Area Community Health (BACH) study is a longitudinal epidemiologic investigation of urogynecologic symptomatology and related risk factors. The study provides data on prevalence and risk factors for urogynecolgic symptoms, including urinary incontinence, benign prostatic hyperplasia, interstitial cystitis, chronic pelvic pain of bladder origin, prostatitis, hypogonadism, erectile dysfunction, and female sexual dysfunction.

General Description - 1 Page

BACH Metadata - 1 Page

Publications - Full citations 27 published, 6 submitted

Forms - 5 Multi-page survey instruments as word files

Roadmap - 1 Page description of data archive

Integrity Check - 100 Pages with statistical model and analysis code plus a medication supplement.

Contrast

The LBL data is also available at the ICIR web site. Two papers and a slide deck describe it, but one has to read these closely to determine the nature of the data and some of its limitations.

- Partial view - 2/20 link interfaces at a time. No net map.
- Some Scans blocked at border (LBNL uses Bro/TRW)
- Other scans in separate file. RFC 3330 IPs not anon.

The BACH data contains details of survey instruments and protocols.

- It seems typical of data at the NIDDK site.
- Why the contrast?

Why Indeed?

Experimental medicine has a long history. Stringent controls on human experimentation arose after abuses of WWII and the cold war. Where human (and sometimes animal) subjects are involved, protocols must be approved by Institutional Review Boards.

- If nothing else,this creates a need for careful planning and written protocols.
- Human experimentation is expensive. Even observational studies may cost a lot per subject.
- Experiments are complex and significant results hard.
  - Data reuse and Meta-analysis are commonplace.

Not just medicine

- Experimentation in the social sciences has similar constraints and shared datasets are also common.
- High energy physics makes extensive use of very expensive shared facilities and sharing of data sets is common. Experiments often reveal phenomena other than those hypothesized and detailed experiment / data collection documentation is necessary to make sense of the results.
- Careful documentation of experimental setups and operational conditions seems to be the rule in other fields. This facilitates data sharing.
- *Computer Science has yet to build a culture of sharing.*

Case 2 - CRAWDAD wireless

The CRAWDAD site at Dartmouth contains data sets that support a number of papers in wireless. The initial data set was packet header data from the Dartmouth campus. We have looked at Nov. 2003 - Feb. 2004.

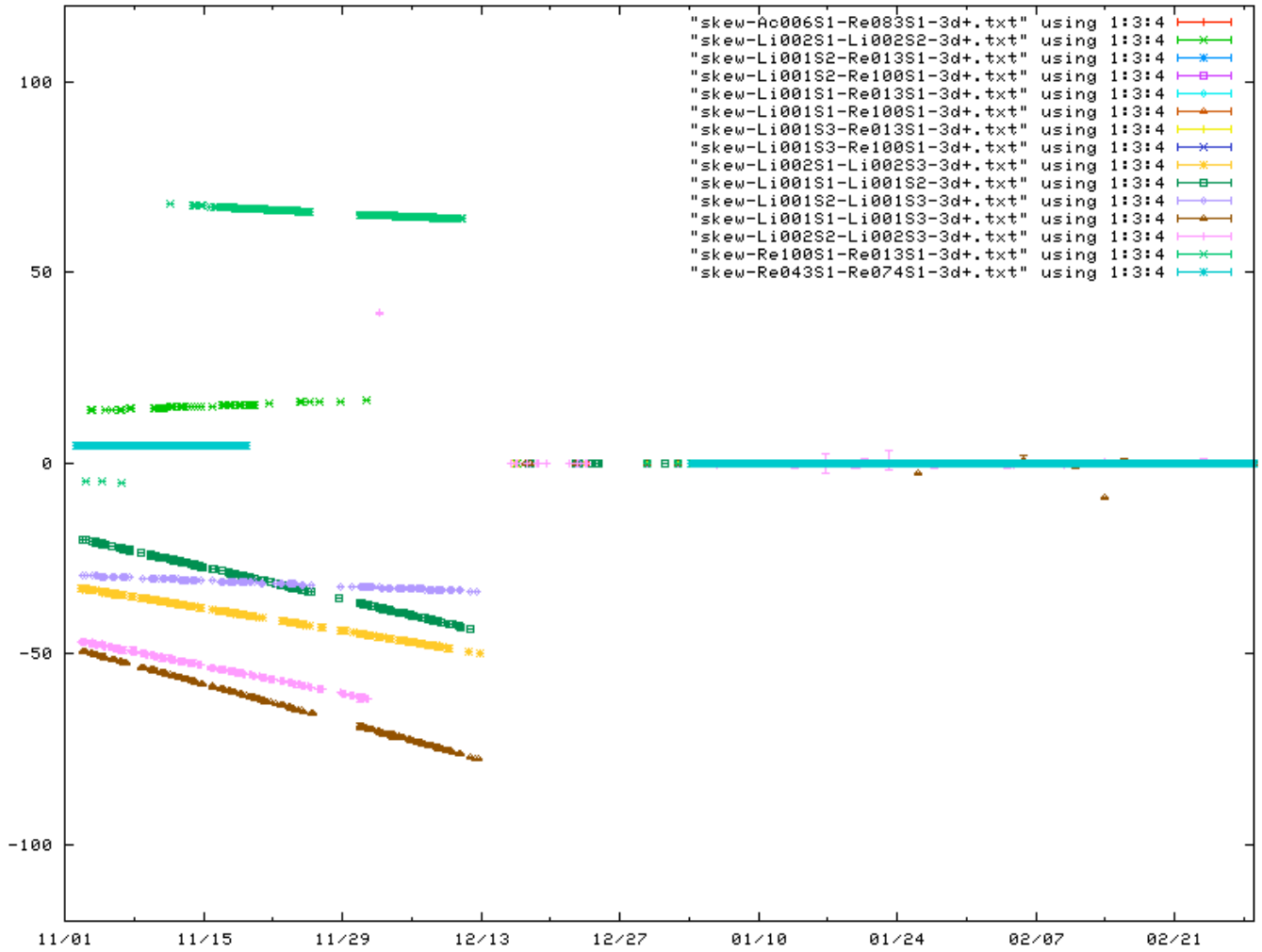
- Motivation was to provide a student training set
- 160+ GB compressed tcpdump headers.
- IPs prefix preserving anonymization ; MAC anon.
- 18 sniffers - Academic, Library, Residence
- Fair use does not allow attacking anonymization.
- Collectors not forthcoming about collection details
- Many problems discovered by trial and error

Time problems

For multi-sensor data, a common time base is necessary if events crossing sensors are to be analyzed (e.g. platform mobility studies). During reorganization of the data into hourly files, we checked for time reversals and large gaps.

- Time should always increase, but we found
  - several small reversals (10s of microseconds)
  - 7 reversals of slightly less than an hour
- Tried to resolve this by finding the same records in several sensors (it is wireless and there are several ways this can happen)
- Results on next slide.

Sensor pair time differences



What is going on?

Clock drift is apparent up through about Dec 13.

- The clocks in the sensors are drifting badly
- At some point, ntp was activated and clocks converge.
- Can we converge the drifting clocks?
  - During Nov / Dec there are many scanning worms.
  - Can we find external evidence of scans; correct time?
  - Must do this so as to preserve anonymization.

This should have been detected during the data collection setup, but no one seems to have checked. Must have been noticed in Dec, but not documented.

Not clear this will help with the big back jumps.

MAC addresses and DHCP

IP addresses are assigned by DHCP and the same platform may have many IP addresses. Questions involving worm infections and propagation involve platforms, not IP addresses. Since this is packet data, we have the anonymized MAC addresses, as well.

- Can we use the MAC address as a platform ID
- Assume MAC spoofing rare. (likely in this case)
- Need to examine constancy of MAC / IP relationship

Not told whether DHCP is local or global. IP is associated with both platform and gateway MACs. At boot time, MAC may adopt several IPs (0.0.0.0, 169.254.x.x, etc.), prior to getting IP from DHCP.

Working through this prior to reassigning platform IP.

Why bother?

- This data is now 5+ years old. Is it worth trying to understand what is happening and cleaning it up.
  - It is one of a very few large sets available to my students. They learn useful skills coping with its problems. Maybe they are more careful.
  - It is a time machine. We discover something in current data and may be able to ask “Was this present at Dartmouth in 2003?”
  - My own data collection practices are improved by solving problems with others data sets.
- All data should be pristine. If pigs had wings they could fly. ... Everyone should be able to vet data.
- *Again, a paper on collection methods would help.*

The sermon

Data collection is part of experimentation. Good data shares a lot with good experimental papers.

- A good experimental paper should start with a hypothesis. What are we trying to show (or refute) with the experiment. Directed data is closely tied to the hypothesis.
- A good experimental paper contains a detailed methods section. Data collection is part of of the methods.
- Data for reuse may be intended to support a range of hypotheses. This places additional burden on documenting the collection process

Hypotheses

Whether the hypotheses are specific or general, the collector must explicitly deal with the ability of the collection to support the analysis needed.

- The laws of physics - time precision and light speed
- Sensor bias and blindness - seeing relevant phenomena?
- Capacity and performance - data drops and other loss
- Ambiguity, redundancy, and consistency - knowing when data is defective. Sensors with differing transfer functions

Even (especially) when the collector is not the user of the data set, integrity tests and a sample analysis should be performed.

Methods

Documenting the collection process and setup is essential. No detail is unimportant. Things that seemed unimportant for the original use can be fatal for reuse.

- It helps the user to determine the utility of the data for specific research questions.
- It may explain observations that come to light during analysis.
- It is necessary if continuation data or data for replication are to be collected by others.

Integrity and Consistency

Most of the NIDDK data sets have explicit integrity components. These describe the steps that the experimenter took to ensure the quality of the data.

A large part of the preparation of data for analysis involves integrity and consistency checks. In most collections, invariants can be constructed from the laws of physics, common sense, and the experimental configuration. A preliminary analysis of the data should be made to ensure that the invariants hold.

Pilot studies should be run to “debug” the collection setup.

Deferring all analysis until all the data is in is usually a bad idea.

Parting Thoughts

Plans are nothing; planning is everything.
Dwight D. Eisenhower

Think first; code later.

Anonymous

Anything not worth doing is not worth doing well.

Dick Kemmerer (attributed to John McHugh)

Thank You