Lies, Damn Lies, and Internet Measurements

Statistics and Network Measurements

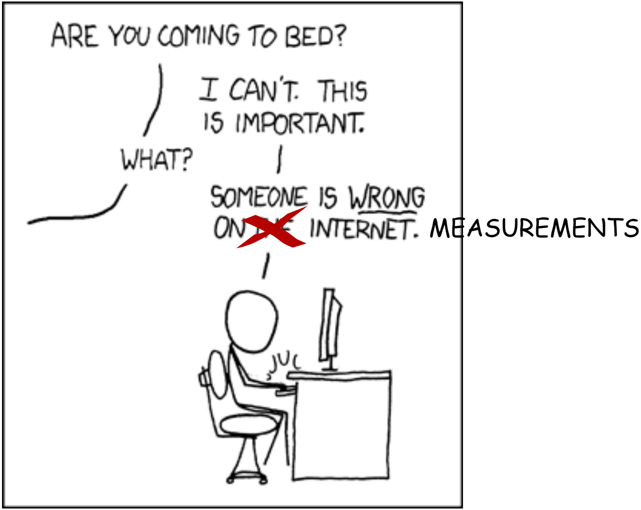
Matthew Roughan
<matthew.roughan@adelaide.edu.au>
http://www.maths.adelaide.edu.au/matthew.roughan/

School of Mathematical Sciences,
University of Adelaide

July 11, 2014

There are three kinds of lies: lies, damned lies, and statistics.

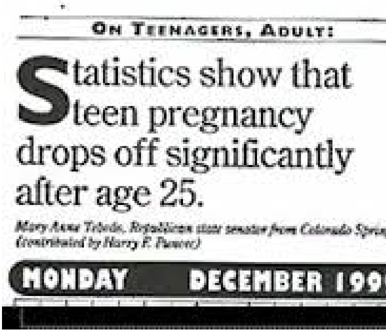
Mark Twain



http://xkcd.com/386/

Statistics and Network Measurements

- Everyone here understands the value of network measurements
- However, not wanting to be too controversial, the NM community is hopeless at statistics
  - its not a unique problem
  - but it can cause some misinterpretations
- War stories
  - e.g., X is better than Y, and related rankings
  - e.g., The red board



A little history of Stats

- **1560s** Cardan, calculate dice probabilities
- **1654** Pascal and Fermat, theory of probability
- **1713** Bernoulli, Law of large numbers
- **1756** Simpson, Theory of Errors
- **1761** Bayes' Theorem
- **1801** Gauss, line of best fit
- **1814** Laplace, lots of contributions

A little history of Stats

1560s Cardan, calculate dice probabilities

1654 Pascal and Fermat, theory of probability

1713 Bernoulli, Law of large numbers

1756 Simpson, Theory of Errors

1761 Bayes' Theorem

1801 Gauss, line of best fit

1814 Laplace, lots of contributions

1854 Jon Snow, Broad Street

A little history of Stats

- 1560s Cardan, calculate dice probabilities
- 1654 Pascal and Fermat, theory of probability
- 1713 Bernoulli, Law of large numbers
- 1756 Simpson, Theory of Errors
- 1761 Bayes' Theorem
- 1801 Gauss, line of best fit
- 1814 Laplace, lots of contributions
- 1854 John Snow, Broad Street

A little history of Stats

- 1560s Cardan, calculate dice probabilities
- 1654 Pascal and Fermat, theory of probability
- 1713 Bernoulli, Law of large numbers
- 1756 Simpson, Theory of Errors
- 1761 Bayes' Theorem
- 1801 Gauss, line of best fit
- 1814 Laplace, lots of contributions
- 1854 John Snow, Broad Street
- 1854+ a little other stuff happened!

A little history of Network Measurements

1969- ARPANET and all that ...

- measurements are part of it, but not much is published (as far as I know)
- stochastic simulation is the norm
- lots of stochastic models proposed and used for data traffic – few measurements used

c1992-97 Beran, Erramilli, Leland, Taqqu, Sherman, Willinger, Wilson, and a few others publish a series of papers about self-similar traffic

c1992-97 Vern Paxson does his PhD at Berkeley on "Measurement and Analysis of End-to-End Internet Dynamics"

c1995-97 Cunha, Bestavros, and Crovella look at web traces

2000+ Network measurements exploded

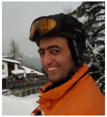
- **2000** First PAM
- **2001** First IMW (becomes IMC in 2003)
- **2001** Endace founded

A little history of Network Measurements

- This is hardly a fair history
  - much is missing
  - focus on what I see as seminal (because it influenced me)
  - apologies to those I left out (CAIDA, Neville Brownlee, and many others)
- I'm trying to make a point though
  - around 92-97 the Internet was growing and changing very rapidly
  - and we went from being data poor to data rich very quickly
  - initial studies were motivated and supported by stochastic models
  - their impact derived from data
- We took the last bit on board
  - data is now seen as key
  - huge efforts to make this data "good"
  - we seem to have forgotten some of the original modelling and statistics that also made those early result so valuable

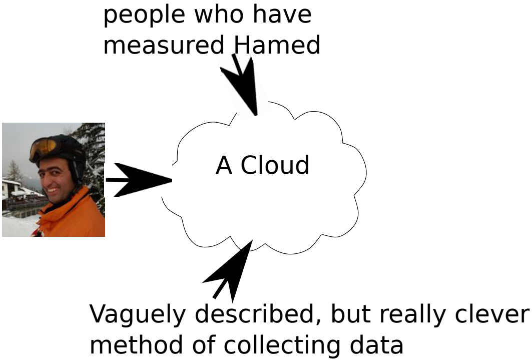
Obligatory Block Diagram

Lets use Internet Measurement to find out how tall Hamed is



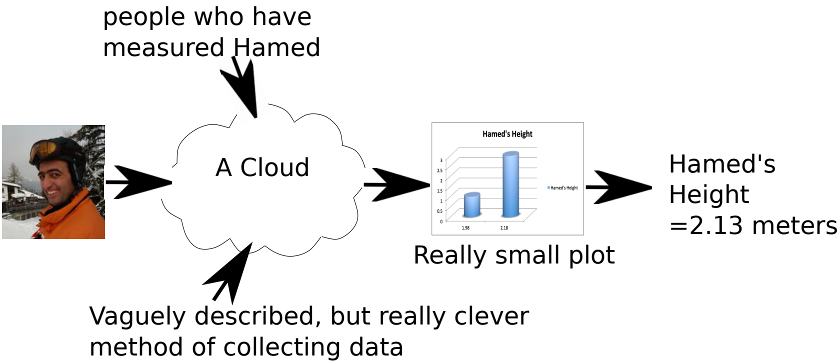
Obligatory Block Diagram

Lets use Internet Measurement to find out how tall Hamed is



Obligatory Block Diagram

Lets use Internet Measurement to find out how tall Hamed is



Case 1: the test

- Common test: test for a problem
  - in medicine it might be a disease
  - in networks, often look for an "anomaly"
- Consider the following
  - There's a chance you have a horrible disease
  - Your doctor comes to you with test results, and says "your test was positive", he also says "the test is 90% accurate".
  - How worried are you?

Case 1: example

- There are two types of error

  type I false alarm or false positive

  type II failed to detect the problem (false negative)

- Imagine a hypothetical test for disease with the following properties
  - if you have the disease, it will be detected 90% of the time
  - if you don't have the disease, then 90% of the time, the test will tell you that you don't

  It seems fair to call it 90% accurate

- Now suppose that 1 in 10 people have the disease
- You go to your doctor, and he tells you (in a serious voice) that your test has come back positive
  - what is the chance that you actually have the disease?

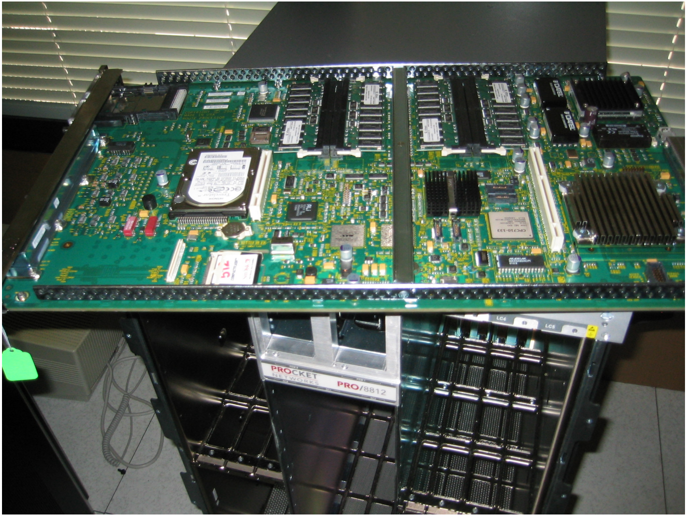
Case 1: analysis

Its a conditional probability problem, but its actually even easier: imagine 100 people:

1. One person in 10 has the disease, so 10 in total
2. If the blood test is 90% accurate, 9 of these will show up in the test
3. The other 90 do not have the disease, but 10% will still get a positive result, i.e., 9
4. So 9 people with a positive test have it, and 9 dont
5. Your chances are 50:50

Interlude: some hardware porn

A Naked Procket



Case 1: network measurement case

- Anomaly detection:
  - 99% detection probability
  - 1% false alarm probability
- Applied to network
  - SNMP link traffic: bytes and packets
  - collected every 5 minutes, on each link
  - 1000 links
  - average 10 real problems per day

$$\text{false alarms per day} \simeq 1000 \times 24 \times 12 \times 2 \times 2 \times 0.01 = 11,520$$

$$Pr(\text{alarm is genuine}) = 9.9/11,520 \simeq 0.0009$$

- Result: ops switch off the alarm system

It's not always easy

If you choose an answer to this question at random, what is the chance you will be correct?

A 25%

B 50%

C 66%

D 25%

What to do

- There's lots of research going on
  - some is on how to do this stuff better
- Be careful with statistics (obviously)
  - learn enough (to be dangerous)
  - consult with a statistician
    - ⋆ this seems to be becoming the norm for medical studies
- Consult your statistician early
  - preferably before experimental design
  - otherwise results may be usefulness, but at the very least you will waste resources, and your statisticians time
- All is not lost
  - results may be useful despite model failures
  - proof is in the pudding
  - but it better be good

- Sorry about the Stats 101 for those already initiated
- Any questions?

Further reading I

J.Beran, R.Sherman, M.Taqqu, and W.Willinger, *Variable-bit-rate video traffic and long range dependence*, Tech. Report TM-ARH-020766, Bellcore, 1992.

Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson, *On the self-similar nature of Ethernet traffic (extended version)*, IEEE/ACM Transactions on Networking **2** (1994), no. 1, 1–15.

V. Paxson, *Measurements and analysis of end-to-end internet dynamics*, Ph.D. thesis, U.C. Berkeley, 1997, ftp://ftp.ee.lbl.gov/papers/vp-thesis/dis.ps.gz.