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/

ARC Centre of Excellance for Mathematical and Statistical Frontiers
School of Mathematical Sciences,
University of Adelaide

April 7, 2016

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

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

"There are lies, damn lies, and statistics. We're looking for someone who can make all three of these work for us."

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
  - it's not a unique problem (e.g., see health sciences)
  - but it can cause some misinterpretations and other problems
- War stories
  - e.g., X is better than Y, and related rankings
  - e.g., The red board

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, TMA, 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

Some Little Examples

Let's look at a few illustrative examples

Case 1: the test

> Statistics means never having to say you're certain

- Common test: test for a problem
  - in medicine it might be a disease
  - in networks, often look for an "anomaly"
- Let me propose a test for disease X
  - there are two types of error

type I	false alarm or false positive
type II	failed to detect the problem (false negative)

Case 1: example

- Imagine a hypothetical test for cancer with the following properties
  - if you have the cancer, it will be detected 90% of the time
  - if you don't have the cancer, then 90% of the time, the test will tell you that you don't
  - 1/100 people have the disease
- You go to your doctor, and he tells you (in a serious voice) that your test has come back positive
- Should you be scared?
  - what is the chance that you actually have the disease?

Case 1: analysis

It's a conditional probability problem, but it's actually easier to just consider frequencies.

Consider 1000 people, on average

- 1 in 100 has cancer, so there are 10 with the disease
- The test will identify 9 in the 10
- 990 don't have cancer, but 1 in 10 of these will have a false alarm
- So the test tell us 108 people have the disease, but only 9 are correct: so the probability you have the disease, given the test is only

$$\frac{9}{108} \simeq 9\%$$

- Our "90% accurate" test has a less than 10% chance of being right

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

Case 1: the issues

- How many False Alarms are too many
  - often we report a "false-alarm probability"
  - but these test might be conducted many times
  - too many false alarms, and you are "crying wolf"
  - the number depends
    - ★ how critical are alerts?
    - ★ how easy is it to fix alarms?
- False Discovery Rate is often what we really need
  - average number of false alarms per discovery
- Tests often have tradeoffs
  - often through choice of a threshold or similar parameter
  - by tuning this, we can exchange false alarms for failed detections
  - testing one without the other is pointless
  - comparisons must be of (ROC) curves of the tradeoff

Case 2: Simpson's Paradox

1. We commonly report results of experiments
   - often we group the data
   - often as percentages
   - and we think they are meaningful
     - e.g. we can see some causality in the data
   - we drawn conclusions from them
     - e.g., A is better than B
2. To do analysis properly
   - firstly we need to know whether our proportions are statistically significant
   - but even then beware Simpson's paradox

Case 2: Simpson's Paradox example

Berkeley gender bias case

- University was sued for bias against women
  - more men were accepted than women (of qualified applicants)

	applicants	admitted
Men	8442	44%
Women	4321	35%

- difference unlikely to be due to chance
  - looks like an obvious case of bias against women

Case 2: explanation

Examine individual departments

	Men		Women	
Department	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

- Larger proportion of female applicants to hard departments
- Not really any (provable) bias

Case 2: Simpson's Paradox

Other examples

The issue can often lead to reverses in conclusions

- Batting averages
  - player A has better average than B in 2012 and 2013
  - but player B's average over the two years is better
- Death penalty case
  - if you look uncritically, it looks like more white people than black are given the death penalty
  - if you control for the race of the victim, then the correlation goes the other way

Case 2: network measurement example

Cooked up example

We compare performance of two networks

- we conduct packet probe experiments
  - round-trip probes
  - assume we know how to do that correctly
  - assume we do enough to be statistically significant
- results

	loss rate
A	1%
B	5%

- Obviously A is better than B?

Case 2: network measurement example

Cooked up example

But really

- The networks carry 2 types of traffic
  - type X
    - is real-time, and unresponsive to congestion
    - both networks prioritise it and it has effectively 0% loss on both
  - type Y
    - is bulk data, and adapts to congestion
    - the two networks have the same "amount" of congestion, and a resulting loss rate of 10% for this type of traffic
- The two networks have different traffic mixes

	X	Y
A	90%	10%
B	50%	50%

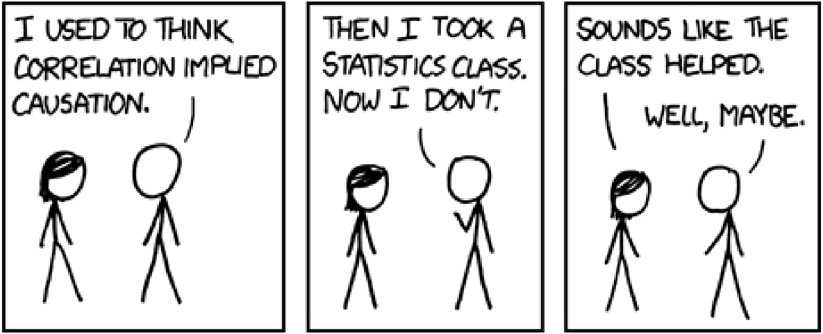
  - hence the loss measurements
  - but neither network is better than the other

Case 2: conclusion

- Obviously, the example is cooked
  - in reality, we might use two different types of probes to assess the different performance
  - but the problem is generic, not specific
- But the point remains
  - danger's of averages
  - correlation doesn't imply causality
  - beware hidden "confounding" variables

lurking variables

Obligatory xkcd cartoon



http://xkcd.com/552/

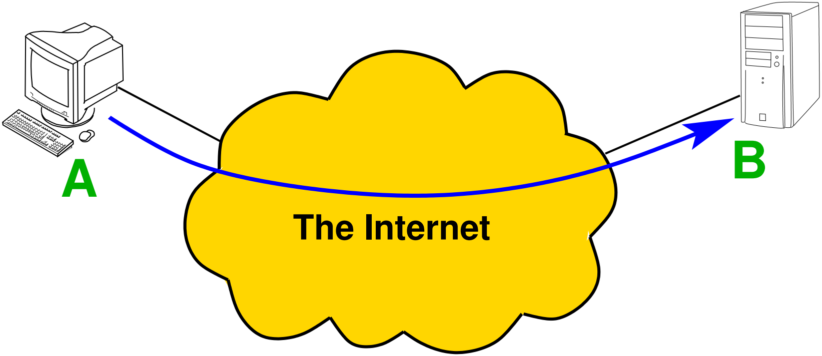
Case 3: estimating loss

- Estimating loss probability
  - packets are dropped in queues
  - want to measure end-to-end loss probability
  - it's a useful measure of how well the network is working
  - high loss rate indicates congestion, or other problems
  - SLAs (Service Level Agreements)
- Strategies
  - active: send probe packets
  - passive: measure traffic at two points
- Metric

$$Prob\{packet\ loss\} = \frac{N_{\text{lost packets}}}{N_{\text{measured packets}}}$$

Examples: Performance

- Active performance measurements
- Send probe packets from $A \rightarrow B$ across the network
- Measure the performance experienced by packets



Case 3: estimating loss

Questions?

- How many probe packets should I send?
- How accurate is a particular measurement?
- My measurement of network A $>$ network B, what does that mean?

These questions are all really asking the same question!

Case 3: estimating loss

Real question

- If we repeated a set of measurements under the same exact circumstances how much could the result vary?

  or the other way around

- Given a desired maximum variability in the estimates, how many measurements do I need?

We often wrap these ideas up in **confidence intervals**, though this isn't the only way to approach the problem.

Case 3: estimating confidence intervals for loss

Naive approach using Gaussian Confidence Intervals (CIs)

- For N measurements, with n losses

  $$\hat{p} = \frac{n}{N}$$

  and this estimate $\hat{p}$ is unbiased (its mean is correct) and its variance is

  $$\sigma_p^2 = p(1-p)/N$$

  and so we choose confidence intervals

  $$\hat{p} \pm z_\alpha \sigma_{\hat{p}}/\sqrt{N}$$

  where for 95% CIs (the typical case) $z_\alpha = 1.96$.
- Stats intuition: you need enough measurements for the Gaussian approximation to be correct, so make sure N is big enough that

  $$N\hat{p}(1-\hat{p}) > 10$$

Case 3: estimating confidence intervals for loss

What's wrong with this?

- The result is widely cited, but WRONG!
- Why?
  - The estimate $\hat{p}$ is used also to estimate CIs
  - The CIs are symmetric, which means you can have negative values!
  - The measure is continuous, but the experimental results are discrete
  - **The measure assumes that loss measurements are not correlated!**

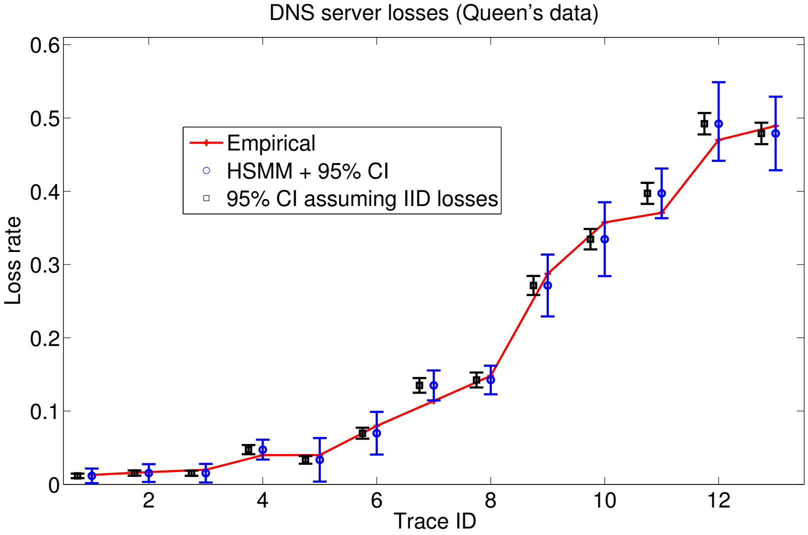
Case 3: what do we do about it?

- The actual variance of the estimate is

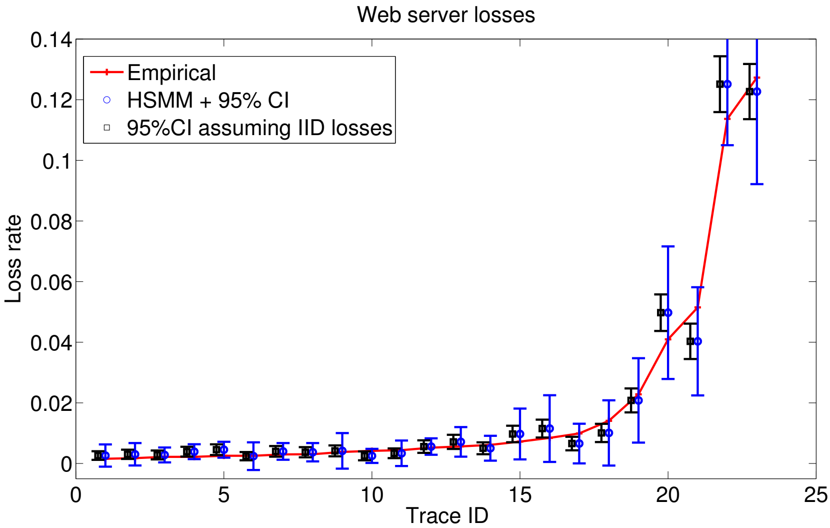
$$\mathrm{Var}(\hat{p}) \quad = \quad \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} R(\tau_{ij}),$$

  where $R(\cdot)$ is the autocovariance function and the τ_{ij} are the times between measurements
- Process
  - Estimate $R(\cdot)$
    - ★ have to be careful how to do this with limited measurements [NR13]
  - Use CIs, but with a better variance estimate

Case 3: cross-validation



Case 3: cross-validation



Case 3: conclusion

- CIs for loss-probabilities estimates need more care
  http://bandicoot.maths.adelaide.edu.au/SAIL/
  - reasonable CIs are usually MUCH wider than IID Gaussian CIs
  - more measurements are needed than you think
- Most Internet loss measurements studies and tools have ignored the problem
  - many research conclusions are WRONG!!!!
  - there may have been SLA violations reported that weren't supportable
  - network op.s decisions made on the basis of bad information, or network op.s stop listening to measurements
- And that doesn't even take into account the other problems which occur when probabilities are small [SBE+11, BCD01, Wil27]

Some other statistical problems

- Sampling
  - do I need to test everyone?
  - remember many experiments are just samples of some underlying phenomena
    - ★ e.g., packet probes sample a network's performance
- Comparisons
  - is A better than B?
  - this is a statistical question, whether you know it or not
  - there are aspects to the question not discussed above
  - ranked orderings are particularly dangerous
- Models
  - curve fitting is potentially misleading
  - but lots of people do even that part really badly
- Gnarly "little" issues
  - long-range correlations
  - infinite variance
  - PASTA

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

> To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.
> *Ronald Fisher*

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

Further reading I

- Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta, **Interval estimation for a binomial proportion**, Statistical Science **16** (2001), no. 2, 101–133.
- 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.
- H.X. Nguyen and M. Roughan, **Rigorous statistical analysis of internet loss measurements**, IEEE/ACM Transactions on Networking **21** (2013), no. 3, 734–745.
- 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.

Further reading II

Joel Sommers, Rhys A. Bowden, Brian Eriksson, Paul Barford, Matthew Roughan, and Nick G. Duffield, **Efficient network-wide flow record generation**, IEEE Infocom, 2011, pp. 2363–2371.

Edwin B. Wilson, **Probable inference, the law of succession, and statistical inference**, Journal of the American Statistical Association **22** (1927), no. 158, 209–212.