Statistical analysis of HTTPS reachability

ggm@apnic.net
{matthew.roughan, simon.tuke}@adelaide.edu.au
matt.wand@uts.edu.au
randy@psg.com

Why are we here? HTTPS Everywhere.

- Richard Barnes (Mozilla) asked informally if the APNIC Labs measurement system might be able to detect people who can't be taken to TLS from unsecure HTTP
  - 'HTTPS Everywhere' is a thing now.
  - If this is to become the norm, we have to understand who might be left behind
- We decided to have a look by active experiment:
  - Select people on port 80, ask them to fetch a URL on port 443, see what happens
- We got a lot of data (3.3m) but finding signals in noisy data is hard
- We definitely found some. But "what does it mean" ?

Cut to the chase

- Not everyone is going to be able to be ‘uplifted’ to TLS in the browser
  - It’s a tiny cohort, but it exists. These are real people. Real users who risk being excluded from secure communications only services.
- It’s possible to characterize this by a number of criteria:
  - Region, Economy, Provider, Browser, Device
- It’s not possible from this measurement to quantify the scale of the problem, but we are quite confident its above the noise threshold.
  - It would be good to test this by re-examination. Can you reproduce the same results?
- Getting there explored some interesting questions about how we might want to look at this kind of public policy question.
  - People should get used to using stronger statistical tools in the toolchest as we get into noisy data
  - Public policy expects blinded data, reproducible results. Avoid bias!
- Read the paper. Its in Arxiv. So is the data. Why don’t you see what you see?

Experiment

- 25 days of sample, one extra experiment added to the non-https clients
  - Cannot downgrade an existing port 443 service, can upgrade from port 80
  - Valid certificate only, no test of funny crypto or bad certs: this should work
- 3.3million samples worldwide (from larger sample set of 130m)
  - Not evenly distributed by either economy or ASN
  - No control over distribution by device
  - No control over ratio of HTTP/HTTPS clients given by Google (how we get to 3.3m)

- So a large sample but unable to apply this to determine worldwide scale, or % of world unable to be upgraded.
  - The experiment can’t directly give a quantifiable % of clients worldwide who are at risk.
  - However can indicate an underlying % of non-HTTP clients from this sample set, so can be applied to better figures from other measurements to scale problem appropriately.

Simple Analysis

- How might we characterize the problem?
  - We know Region, Economy, origin-AS, Browser, OS.
- Tabulate, compare. But sample counts vary wildly because of the imbalances in the measurement source. How can we meaningfully compare these things?

- Initial simple statistical analysis used ‘the usual tools’
  - This is a *categorical* question: Can the user use 443 or not?
  - Outcome? To hard to say from these tests alone
    - Strong indications there is a problem, and its got signs in our categories which do not appear strongly linked: thus we might have multiple independent underlying causes.
  - Too easy to dive into inferences about the data. THIS economy THAT browser.. A lot of ad-hoc AHA thinking going on…
- Ok. Lets drive harder: Whats in the toolchest?
  - Time to get competent statisticians onto the problem.

Statisticians to the rescue

- Statisticians like new problems. Can they have some fun here? Sure!
  - This is a problem about teasing out a signal from a high level of background noise, with several categories, and determining if its distinct from a control of random loss and if the categories are independent or linked
  - The actual subject is (almost) ephemeral interest. The statistics is the fun bit.
- First level statistics for dummies (me)
  - Avoid ad-hoc conclusions: blind the data, rely on the maths, not intuition
  - Construct proper null hypotheses against the data
  - Consider questions which can be tested using well known simple methods
  - Based on what the simple methods tell you, dive deeper.

First level outcomes: HTTPS is harder than HTTP, but there is some linkage in the test

- We have two fundamental questions regarding HTTP/HTTPS

1. Are the two tests actually independent or not?
   - We asked each client to fetch an http object, and an https object. It's a matched-pair problem.
   - This is validated by *Fishers exact test* for independence of matched pairs.
     - Null hypothesis: the tests are independent
2. Are the two tests of equal "hardness"?
   - We can use *McNemars test* because one test (HTTP) is logically a 'control'
     - Null hypothesis: the tests are equally "hard"

- This approach gets us to 'if' but not 'how much'
- Outcome: We reject the null hypothesis in both cases
  - we can't show independence, but we can show varying difficulty
  - Formally, we had to reject the null hypothesis in each case, because of the way we pose the question under test and the results of the stats.

Statisticians push harder.

- Lets get beyond ‘IF’ into ’How much’
  - How can we begin to quantify the problem beyond the simple categorical questions? It’s a weak signal at best.
- Whats a good tool to do this kind of analysis? **Modelling**.
  - Find qualities in the data which conform to a model, understand the problem by understanding the model and its behaviors
  - Apply models to the covariates: the qualities of distinction we have identified in the test subjects
  - Region/Economy, Origin-AS, Browser, OS
  - Can we show a single cause, or understand the interrelationships of these qualities?

Statisticians push harder.

- Rasch Modelling
  - Item-Response Theory
  - Takes categorical method like fisher/pearson which is a simple yes/no, and projects it into an ordering. So can distinguish fine-grained variances in a signal.
  - Model requires a start condition/state, which is fed from the simple stats information
  - Comes from un-related disciplines, typically used to rank question/student exams to understand their appropriateness for easy/hard questions against student ability
  - Each category forms its own measure. So unrelated, and harder to identify covariates
- Generalized Linear Mixed Models
  - A more complex method of looking at several distinct groupings at once to determine linkage, independence and relative weight amongst covariates through randomized methods
  - Applicable to non-normal distribution data (which is good, since we probably are)
- I very quickly got out of my depth. But this is an interesting conversation to have with people who know the field. Working with professionals is always good. You need to read the paper to get this stuff. I'm still learning.

Rasch Model outcomes (*from the paper*)

- *The measurements show that there is (statistically) significantly more difficulty in performing HTTPS than HTTP measurements.*
  - The difference is often small, necessitating some extra care in order to determine whether the difference is significant.
- *There are country, OS, and browser differences, mainly important through a small set that exhibits more extreme variations from the norm.*
  - We decided to keep this blinded. We didn't want to get distracted by value judgements about the qualities

GLMM outcomes

- ASN and Browser rank higher than Country and OS as a covariate for HTTPS difficulty
- We would predict that this could be shown in a similar experiment, including one not subject to the unknown quantity/ratios of covariates (eg, a fully randomized sample method)
- The GLMM strongly suggested the APNIC collection methodology did not account for the region/country signals (we localize experiments to one of four nodes depending on the major continental location)

What did we find from pushing harder.

1. The APNIC experimental method with four collection points did not distort the underlying measurement. (from GLMM)
   - There is no strong regional bias, economy is a better predictor of likely failure to be taken to HTTPS (not surprising)
2. ASN is a strong predictor of failure (it often couples to economy)
3. Browser and OS are weaker but significant predictors of failure
4. There is therefore probably no one single root cause of the problem
   - Its wide-spread, and has a range of inputs which are potentially causing it
   - Economy/ASN goes to network. OS/Browser goes to on-host problems

Public policy deserves reproducible results

- The level of statistical rigor being applied in network measurement could be better overall
  - We should expect to work on blinded data
  - We should be prepared to share blinded data so methodology can be reproduced
  - We should invite critique of method/methodology and explore new ones and compare
- This is comparable to the application of statistics to public policy issues in public health, transport, economic planning.

So lets get statisticing.

- We've got a completely blinded dataset, arbitrary random uniques for each of region/economy/origin-as/browser/os
  - Its not that you couldn't work out which is which: the idea is not to **try**
  - No individual IP addresses are being revealed: the dataset doesn't have addresses in it any more. It's a set of matched-pair results with covariates.
- Can you reproduce the outcome?
  - Same methods, can you get the same result?
  - Different methods, do you converge to the same result?
  - Different data, do the conclusions hold up? Is the Rasch/GLMM model providing useful predictive indications?
- Isn't this the higher threshold we'd want in public policy?
  - Reproducible results, which can be tested independently

Public policy in HTTPS Everywhere

- “HTTPS-Everywhere” is a high goal, and a huge outcome for individual privacy and end-to-end security. We definitely want to go there.
- A small cohort of users are not going to be able to get there reliably
  - There isn’t a single root cause driving this. Its got several drivers which we haven’t explored fully (yet)
  - Its distributed in the network at large. There is no single smoking-gun
  - It includes Economy/ASN, and Browser/OS as indicators of risk
- Even though its small (order 1/1000 of http users seen, at a rate of 2.5% of overall clients seen) if you get to global internet scale, this is going to be a problem for some people. We’d need to understand the actual ratio of http/https to quantify it properly