NICTA

Managing security-relevant data from measurements on Internet scale

(Tales from the road)

Ralph Holz

9 June 2015

NICTA

About the speaker

- PhD from Technische Universität München, 2014
- Dissertation on measurement and analysis of X.509/TLS, SSH, OpenPGP (PKI)
- Now a Researcher at NICTA
- Working on data-driven security mechanisms
    - Understand real-world security problems by measurement
    - Develop defences that make good use of measurement data

Large-scale scans of security protocols

NICTA

- Large-scale Internet scans to determine state of security deployments
- Often results in quite large data sets to collect or handle
- Examples:
  - TLS/X.509 PKI: $\approx$ 40-50 GB collected
  - DNS: $\approx$ 800GB for .COM zone *per scan*
  - SSH: on the order of 10 GB collected
  - WHOIS: on the order of 50-100 GB
  - BGP: on the order of hundreds of GB (we do not store)

What do we do with these data?

- Document weaknesses
- Detect deployment issues (human problem, not a technical one)
- Derive new security mechanisms from insights
- Use measurement data to establish situational awareness

Let's have a look at some examples.

Acknowledgement

The following presents joint work with Lothar Braun and Nils Kammenhuber and Georg Carle (all TUM).

The X.509 Public Key Infrastructure (PKI)

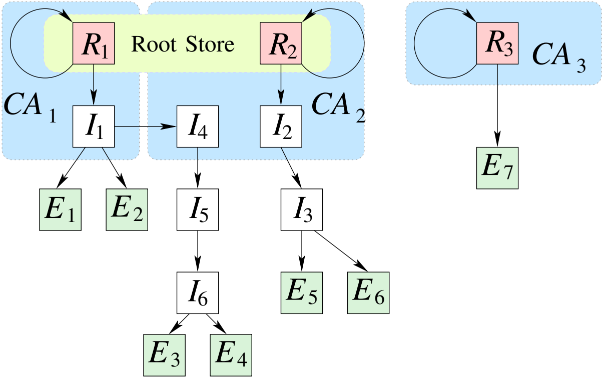
Much of our Internet security is built on X.509

- Every TLS-secured protocol uses X.509
- Further use cases: email, code-signing, . . .

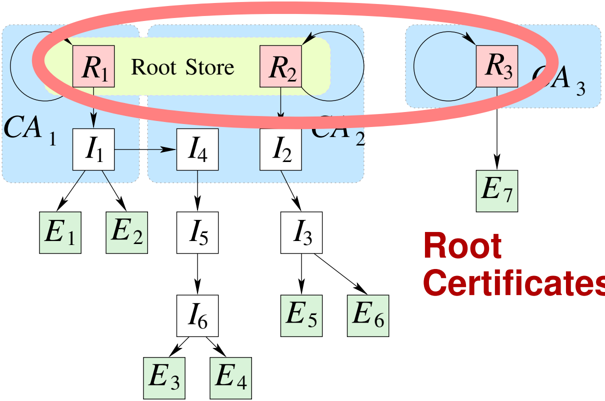
All X.509 PKIs share the same principle

- Certificates bind an entity name to a public key
- Certification Authorities (CAs) act as certificate issuers
- Browsers/OSes preconfigured with CAs' 'root' certificates

Basic idea of X.509 PKI

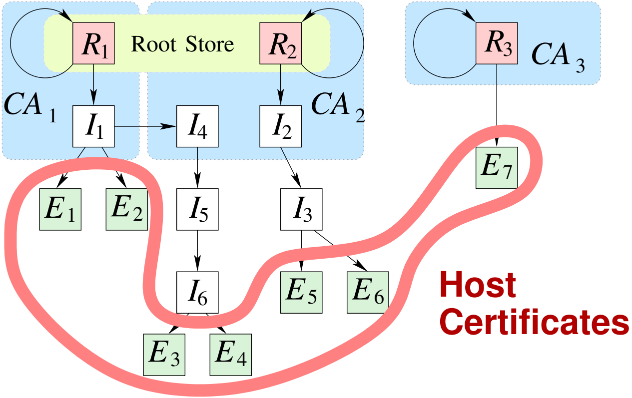


Basic idea of X.509 PKI

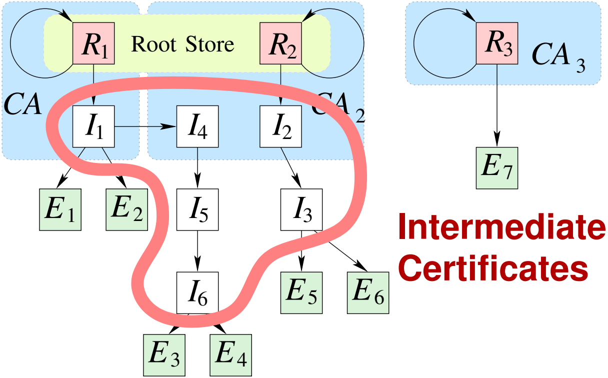


Root Certificates

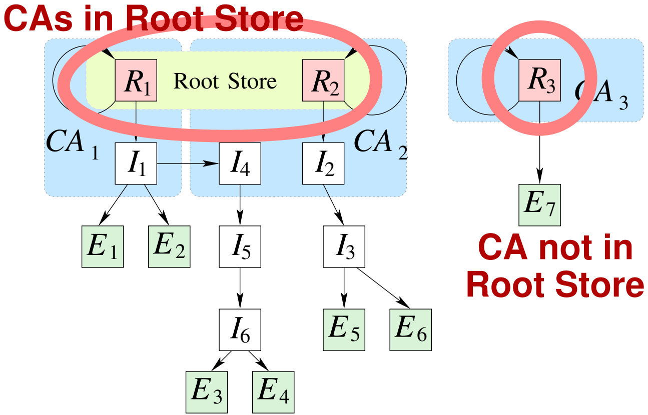
Basic idea of X.509 PKI



Basic idea of X.509 PKI

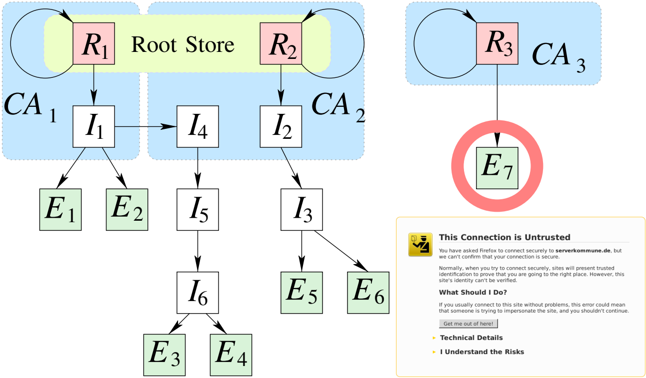


Basic idea of X.509 PKI



Basic idea of X.509 PKI

Root certificate not in Root Store



This Connection is Untrusted

You have asked Firefox to connect securely to **serverkommune.de**, but we can't confirm that your connection is secure.

Normally, when you try to connect securely, sites will present trusted identification to prove that you are going to the right place. However, this site's identity can't be verified.

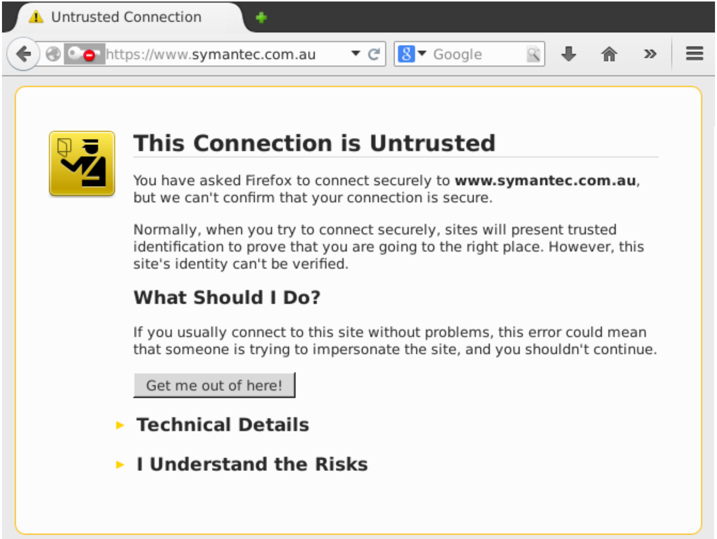
What Should I Do?

If you usually connect to this site without problems, this error could mean that someone is trying to impersonate the site, and you shouldn't continue.

Get me out of here!

- **Technical Details**
- **I Understand the Risks**

A typical Internet experience



Reason (not a UX fail)

NICTA

Technical Details

www.symantec.com.au uses an invalid security certificate.

The certificate is only valid for the following names:
symantec.com, norton.com, careers.symantec.com, customercare.symantec.com, jobs.symantec.com, www.account.norton.com, account.norton.com, mynortonaccount.com, www.nortonaccount.com, nortonaccount.com, downloads.guardianedge.com, www.pgp.com, store.pgp.com, na.store.pgp.com, eu.store.pgp.com, uk.store.pgp.com, row.store.pgp.com, nukona.com, www.nukona.com

(Error code: ssl_error_bad_cert_domain)

Our research goal: assess the quality of X.509

X.509 should:

- ...allow HTTPS on all WWW hosts
- ...contain only valid certificates
- ...offer good cryptographic security

And there should be:

- Long keys, only strong hash algorithms, ...
- Correctly deployed certs

Does it?

Data sets: 25m certificates

Active scans to measure *deployed* PKI

- Scan hosts on Alexa Top 1 million Web sites
- Nov 2009 – Apr 2011: 8 scans from Germany
- April 2011: 8 scans from around the globe

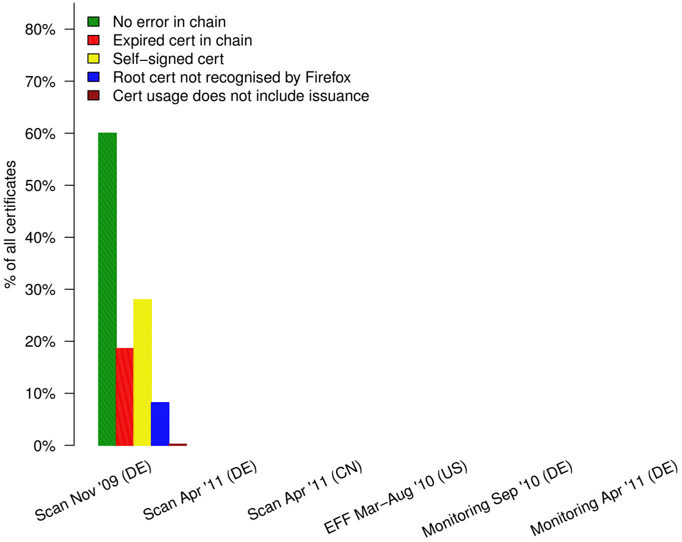
Passive monitoring to measure *user-encountered* PKI

- Munich Research Network
- Real SSL/TLS as caused by *users*

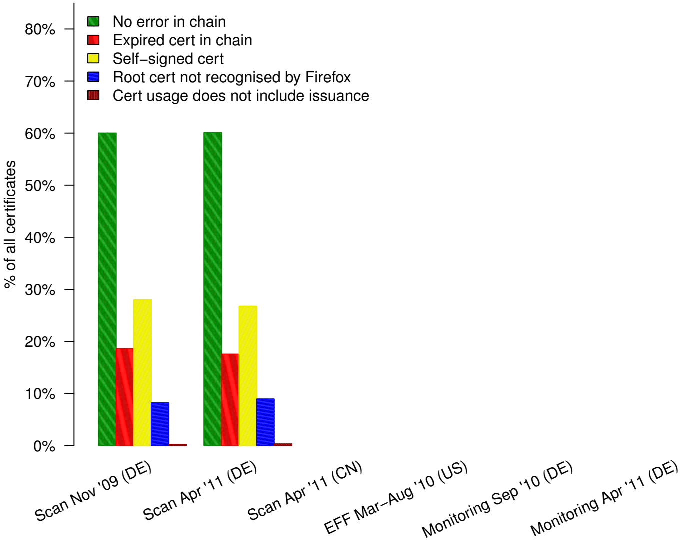
EFF scan of IPv4 space in 2010

- Different kind of scan—months-long, no domain information

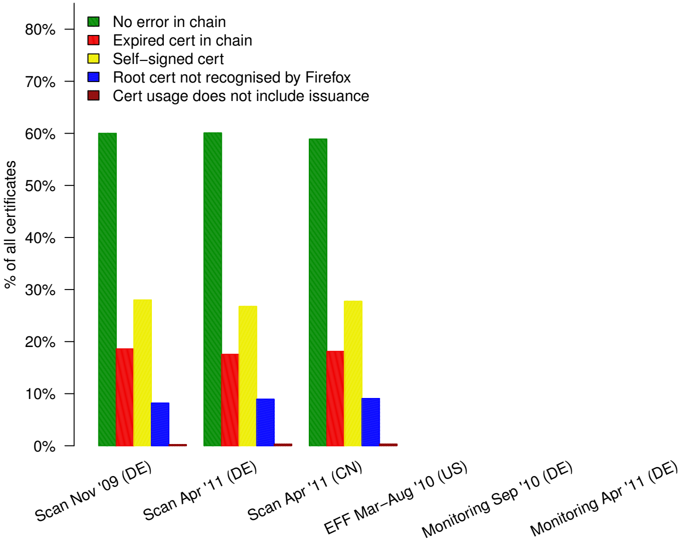
Correctness of certificate chains



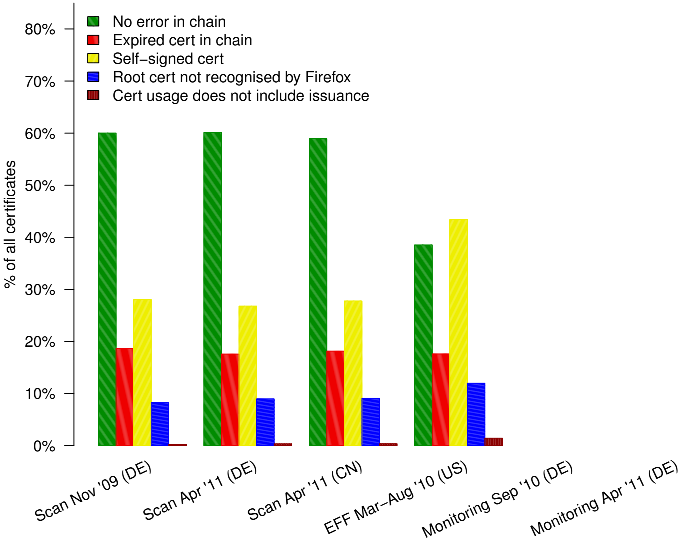
Correctness of certificate chains



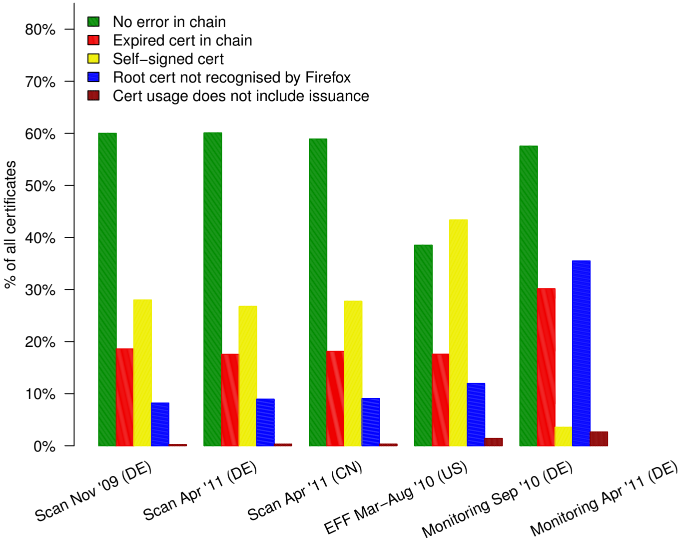
Correctness of certificate chains



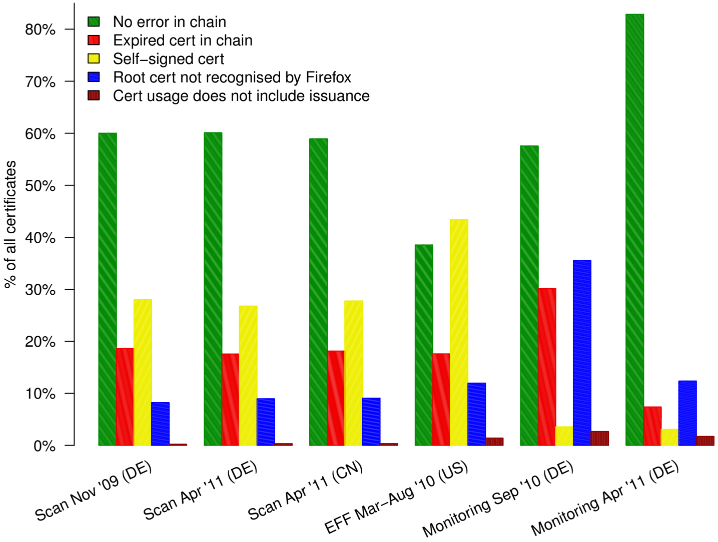
Correctness of certificate chains



Correctness of certificate chains



Correctness of certificate chains



Domain names in certificates

NICTA

Are certificates issued for the right domain name?

- Tested for scans of Alexa Top 1m
- Compare name in certificate against domain name, incl. wildcard matching
- Only **18%** of certificates are fully verifiable
- **More than 80%** of the deployed certificates show errors

Continued work

- Our scans were the first long-term, large-scale, and globally distributed scans of a popular Internet security protocol
- Important insight: X.509 needs deployment mechanisms—not more crypto
- Indeed, letsencrypt.org has in the meantime taken up this idea

- In 2012, we switched to Internet-wide scans for both SSH and TLS
- We have continued scanning TLS ever since
- Several other protocols have been added to our scanning methodology

Acknowledgement

The following presents joint work with Johann Schlamp, Oliver Gasser, Andreas Korsten, Georg Carle (all TUM), and Quentin Jacquemart and Ernst Biersack (Eurecom).

The fragility of Internet routing

NICTA

Connect with us

c|net Search CNET Reviews News Video How To Games Download Log In / Join US Edition

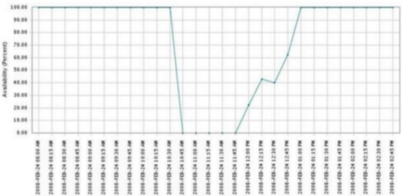
CNET › Tech Culture › How Pakistan knocked YouTube offline (and how to make sure it never happens again)

How Pakistan knocked YouTube offline (and how to make sure it never happens again)

YouTube becoming unreachable isn't the first time that Internet addresses were hijacked. But if it spurs interest in better security, it may be the last.

by Declan McCullagh @declanm / February 25, 2008 2:30 PM PST / Updated: February 25, 2008 4:28 PM PST

0 / / / / / more +



This graph that network-monitoring firm Keynote Systems provided to us shows the worldwide availability of YouTube.com dropping dramatically from 100 percent to 0 percent for over an hour. It didn't recover completely until two hours had elapsed.

Keynote Systems

THIS WEEK'S MUST READS /

1. How Pakistan knocked YouTube offline (and how to make sure it never happens again) — Tech Culture
2. Google's Android Wear software will let you leave your phone at home (if there's Wi-Fi) — Mobile
3. In Adobe's new Lightroom, multiple photos can now meld into one — Photography
4. Yahoo's new search pact with Microsoft includes opt-out clause — Internet
5. Microsoft rolling out 34 unscheduled patches for Windows today — Operating Systems

The fragility of Internet routing

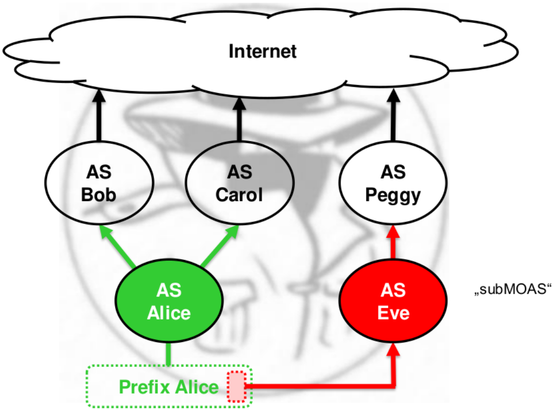


Figure: J. Schlamp, TUM

The subMOAS problem

- subMOAS: 'subprefix multiple origin AS'
- A more specific prefix belonging to AS *A* is announced by a different AS *B*
- Example of attack: blackholing attack
- subMOAS *might* also be absolutely legitimate—business relationship etc.
    - On just one day, we observed >75k subMOAS events
- subMOAS can be extremely transient (hours to days)
- How to distinguish the attack from the benign case?

- Our approach: turn the problem on its head—try to rule out an attack by analysing evidence of benign behaviour.

Data source 1: RIPE

RIPE is the Regional Internet Registry for Europe. It stores information about registered Autonomous Systems, prefixes, relationships between them.

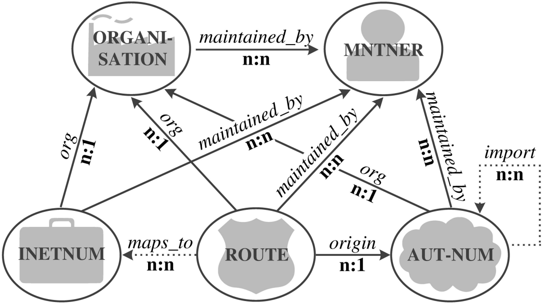


Figure: J. Schlamp, TUM

Data source 2: topology reasoning

- Build a graph of Autonomous Systems to affected subprefix
- Check if the ‘attacking’ AS is actually a downstream AS from the ‘attacked AS’
- In other words, if the ‘attacker’ is attacking her upstream AS
- This is very unlikely to be an attack—the upstream AS would simply shut down the attacker because they would be the first to notice

Data source 3: SSL/TLS scans

- Hosts may have unique keys—identify such stable hosts
- If they remain reachable during a subMOAS, there is no (blackholing) attack
- First, establish a ground truth (methodology in paper)
- Then, scan all hosts in affected subMOAS for their keys

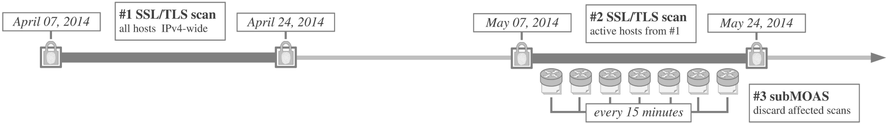


Figure: J. Schlamp, TUM

Results for subMOAS analysis

Coverage (subMOAS recognised as benign):

- Investigated > 8,000 events over ten days—automatically
  - RIPE: 88% of prefixes covered
  - Global IP space: 60% of prefixes covered
- Globally, we could prove 46% of all subMOAS events were benign

Conclusions:

- Method has great potential, especially for limited IP spaces
- For a better global coverage, we need the other RIRs and more scans of security protocols

Data management

Data has to be managed to achieve two very desirable properties:

- **Reproducibility**—measurements and results are (near) useless if no-one can reproduce them for their own purposes
- **Reuse**—data sets can be combined and reanalysed to investigate new research questions

There are two urgent requirements:

- Very **precise understanding of the methodology** used during data collection
- Data must be **meticulously annotated** to ensure it can be linked to another data set.

We had the advantage that we collected all data ourselves. But what if others wish to combine two data sets?

Best practices

NICTA

We argue that the following are *minimal* best practices:

- Precise documentation of methodology
  - Software developed: code release
  - Libraries/tools used: version (better: compile flags)
  - Setting: IP, time, upstream, bit rate etc.
  - Document resolution of measurement
- Use precise definitions of key terms
- Document research questions you answered
- Document all known limitations

It would be great to have machine support for such documentation.

Example: methodology

A mistake we made in our early scans

- We forgot to document the openssl version we used
- In SSL/TLS the client sends a list of cipher suites, from which the server chooses
- This list may change between releases, and even Linux distributions
- While this did not affect the data we released and wrote about...
- ...we now use custom-built openssl from vanilla sources

Example: unclear definition

NICTA

'Certification Authority' (CA)

Claim by EFF in 2010: > 1,500 CAs are trusted. This is wrong.

- 60-80 organisations in Mozilla root store; ca. 180 root certificates.
- Intermediate certs $\neq$ CA. It is not possible to identify a CA by technical means alone. You need to read their policy documents.
- Estimates for the correct number vary; general opinion: about 500.

Example: research questions

NICTA

Research questions

These are often easy to specify and document. . .

- 'What is the distribution of public key strength over the IPv4 space?'
- 'What percentage of certificates with validating chains?'

. . . however:

- What is the definition of 'validating chain'?
  - Which root store?
  - Which tool?
  - Which version?
- It's about precise documentation of your tools again

Example: unclear definition and/or limitation

What is a host?

- An IP address is not a host, and a domain is not a host.
- Hosts may have many interfaces, or virtual interfaces, and thus many IPs.
- A host may serve many WWW domains, have many DNS names.
- There is no generally recognised way to identify a host.
- To complicate things: maybe 30% of IP addresses are assigned dynamically.

Consequence: the duration of a scan has a direct effect on the accuracy of your results.

Some more examples

Effects that should be documented

- Experiment resolution: to which degree are short-lived effects captured?
- What mappings exist between your variables (e.g. IP to domain)
- Which confounding effects could technically not be avoided?

This is really just good scientific practice. The question is not whether it should be done—but how it can be done to enable other researchers to combine their data sets. Can it be done (almost) M2M?

Thanks!

I would love to discuss:

- Ways to protect data and yet release it
    - (Semi-)automatically to friendly researchers?
- Ways to make experiment data reproducible and useful for others
    - Clearly, there is no one-size-fits-all solution
    - But can we make a step towards automation?
- What data is similar in nature to the data we have experience with?
- Experiences from which we can learn—e.g. sensitivity, privacy,. . .

Scanners, data sources, collaborations

NICTA

- Fully-resolving DNS scanner
  - Joint work with J. Naab, TUM
- Internet-wide scans of the SSH protocol
  - Joint work with with O. Gasser, TUM
- MP-TCP: active and passive collection of data
  - Joint work with with O. Mehani, NICTA; J. Amann, ICSI
- Further data sources:
  - BGP (with TUM)
  - Internet registries (with TUM)
  - WHOIS (Autonomous Systems)
  - Geo-location

Next: joint work with J. Schlamp from TUM