Welcome to Maynooth!

Plato's Allegory of the Cave

Imagine prisoners in a cave, chained to the wall. They cannot see the real world, instead only shadows of objects. What is real? The shadows? The objects that cast them? The world above that inspired them?

The Internet and the Cave (discussion)

What is our cave? Is it right, flat, the cave. Think about some pre-conceived idea. Cave = traceroute; shadows = output; shadows = real world? The measurements we take. Objects? The ground truth real world? The devices we do not see the people holding the objects? The companies (or agencies) managing (or manipulating)

The Internet and the Cave

At one time the Internet fit on a napkin. Those days are long past...

- Many networks: >4M/24s
- Many computers: ~800M on public internet
- Many protocols

What to do?

[Map by CAIDA; data from Cheswick and Burch, 2000]

[Internet Census, USC/ISI, taken since 2006; this: 2017-02]
The Internet and the Cave

Imagine prisoners in a cave, chained to the wall. They cannot see the real world, instead only shadows of objects cast. What is real? The shadows? The objects that cast them? The world above that inspired them?

*by Markus Maurer, from Wikipedia.org*

researchers are “chained” limited in what we measure—but we see incomplete shadows. We should make inferences about the objects behind what we measure.

what is real? The shadows? the objects that cast them? the world above that inspired them?

Outline

- **Intro:** Platon’s Cave
- What Do We Want?
- 4 Case Studies and 5 Ground Truths
- Conclusions

What To Measure?

- Topology
  - Core (routers and links) and edges (hosts)
  - Relationships: AS-to-AS, AS-to-organizations
- Size and capacity
  - Numbers of end-systems, routers
  - Amount of traffic
  - Capacity of pipes and interconnection points
- Traffic and applications
  - QoE
- Reliability
  - Packet loss, outages, censorship

What To Measure? (My Take)

- Topology
- Size and capacity
- Traffic and applications
- Reliability

Established Research Topics

- What
  - Topology
    - Core (routers and links) and edges (hosts)
    - Relationships: AS-to-AS, AS-to-organizations
  - Size and capacity
    - Numbers of end-systems, routers
    - Amount of traffic, capacity of pipes and interconnection points
  - Traffic
    - Classification, trends
    - Quality of experience
  - Reliability
    - Packet loss, outages, censorship

- How
  - Traceroute
  - Ping
  - BGP peering (RouteViews)
  - Traffic analysis: HTTP, TCP, NTP
  - (Wireless stuff, also)
  - Platforms:
    - RIPE Atlas, CAIDA Ark, PlanetLab, private platforms
    - Simulators and simulators: DETER, Mininet, private
    - From users, Shadow, apps, Google ads
    - Simulators: ns-2, ns-3, OpNet, custom

Defining Ground Truth

- Goal: Is what we measure correct?
- Ground truth: Defines what is correct
  - But sometimes it is incomplete
  - Often unobtainable

But never forget that it exists; we must strive for it (There is an"outside the cave")
Elusive Ground Truth

- consider measuring height
  - ruler measured in cm: says h = 180cm
  - true height with ruler with infinite precision: h = 180.340cm
- is that true?
  - limitations on how accurately you can measure
  - you're taller in the morning
- (is meter well defined)

Elusive Ground Truth

- consider measuring height
  - ruler measured in cm: says h = 180cm
  - true height with ruler with infinite precision: h = 180.340cm
- is that true?
  - heights actually varies by around 1cm each day
  - even if true now, not true in 6 hours
- sometimes the truth varies; sometimes no single truth ever exists

Can we “Fix” Elusive Ground Truth?

- heights actually varies by around 1cm each day
- how to fix?
  - could define height more precisely
    - height must be measured at 9am
    - could define height as a range or distribution
    - 180 ± 1cm
    - an “envelope of truth”
  - why/when should height be measured? (non-stationary)
- both approaches have their place
  - range seems easier
  - why are you measuring?

Can we “Fix” Elusive Ground Truth?

- heights actually varies by around 1cm each day
- how to fix?
  - could define height more precisely
    - height must be measured at 9am
    - could define height as a range or distribution
    - 180 ± 1cm
    - an “envelope of truth”
    - why/when should height be measured? (non-stationary)
- both approaches have their place
  - range seems easier
  - why are you measuring?

Aside: Truth is Often an Envelope

- TCP performance as a function of loss (p) and RTT?
  - bitrate = \( \frac{RTT^2}{2bp} \)
  - but there are many, different implementations
    - BSD, Linux, Windows
    - Vegas, FAST, CUBIC, BRR
- where does this matter?
  - validating TCP in ns-2
  - TCP friendliness: contention control that ties to be “like” TCP
  - future TCPs (CUBIC, BBR, etc.)
  - future other protocols (QUIC, etc.)

Outline

- intro: Plato’s cave
- what do we want?
- 4 case studies and 5 ground truths
- conclusions
Where to Get Ground Truth?
(discussion)

- DPI for traffic classification
  - modulo encryption
- SNMP to get data on congestion
- friendly network operators
- there are tradeoffs in privacy and proprietariness
- testbeds
  - complete control: good: you have control, bad: you set it up, so you have know the parameters and assumptions

Where to Get Ground Truth?
(my take)

- from the network operator
- from testbed experiments
- from simulations
- as seen in prior results

Case Study 1: Network Topology Mapping

- question: can we map ISPs, or the whole Internet?
- early work
  - “Heuristics for Internet Mapping” Govindan and Tangmunarunkit, INFOCOM 2000
  - “Mapping ISP Topologies with Rocketfuel”, Spring, Mahajan, Wetherall, SIGCOMM 2002
  - “Macroscopic analyses of the infrastructure: measurement and visualization of Internet connectivity and performance”, Huffaker, Fomenkov, Moore, Claffy, PAM 2001
- recent work
  - ITDK-2016 from CAIDA

Ground Truth 1: from the Operator

- ground truth: use a few research networks
  - “Heuristics for Internet Mapping” Govindan and Tangmunarunkit, INFOCOM 2000
  - 2 regional networks: Los Nettos and Calren2
  - “Mapping ISP Topologies with Rocketfuel”, Spring, Mahajan, Wetherall, SIGCOMM 2002
  - 3 (private) ISPs gave qualitative results
  - (the Huffaker et al 2001 paper did not evaluate correctness)

Ground Truth 1: from the Operator

- ground truth: defines what is correct
- but what does “correct” mean?
- unambiguous
- something that fits the purpose of this experiment
- optimum… algorithms? unproven? the best possible?
- scalable
- has high probability of being reproduced
- we can compare the

Ground Truth 1: from the Operator

- ground truth: defines what is correct
- but what does “correct” mean?
- from info theory
  - precision: is what you claim always true?
  - recall: is what you claim the complete truth?
  - accuracy: is what you claim and reject both correct
the Confusion Matrix: Formalizing Correct

<table>
<thead>
<tr>
<th></th>
<th>prediction positive</th>
<th>prediction negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>actual positive</td>
<td>true positive (TP)</td>
<td>false negative (FN)</td>
</tr>
<tr>
<td>actual negative</td>
<td>false positive (FP)</td>
<td>true negative (TN)</td>
</tr>
</tbody>
</table>

precision = TP / (TP+FP)  
recall = TP / (TP+FN)

accuracy = (TP+TN) / Population

Beware paper that talk about “correctness” without defining what metric of correctness.

They often focus only on precision and ignore recall—what they say is true, but they may miss a lot (and not know it).

Back to Ground Truth from the Operator

- “H outilities for internet mapping” Govindan and Tangmunarunkit, INFOCOM 2000
- “Regional networks: Los Nettos and CalREN” Bannister, ACM IMC 2008
- “we found all routers and all but 1 link in each network”
- “they don’t give counts, so we can’t estimate...”
- “measured SLAs via Rockfuel” Spring, Mubin, Wetherall, SIGCOMM 2002
- “surveyed ISP post-qualification stage”
- “did routers say ‘I don’t know’?”
- “did routers say ‘I don’t know’?”
- “they don’t give counts, so we can’t...”
- “how many customers did the router?”
- “how many customers did the router?”
- “we will only say what they showed
- “they don’t always know the truth!

Ground Truth 2: from Our Enterprise

- ground truth from Our Enterprise
  - University of Southern California
  - disadvantage
  - can talk to the operators (we know them)
  - can apply multiple measurement methods
    - test active probing (ICMP, the new method being considered)
    - against other kinds of active probing (TCP SYNs)
    - and against passive traffic analysis
  
other ground truth about topology

- academic network
  - Los Nettos: LA regional consortium
  - CalREN: California academic net
- pro:
  - GRANT
  - 
  - open topologies
  - cost
  - not profit based maybe not optimized the same way
  - The Griffin had a population-centric research net to commercial net: The Internet Consortium of Southern California 2000
- con:
  - where you have to fill in all the details

Sources of Error for Edge Address Activity

- variance
  - measurement location: downstream: normal error
  - sampling error
    - random due to probe order;
    - function of probe frequency
    - survey within 5-9% with 95% confidence
  - birth/deaths during survey: estimate in paper
  - probe type: (ICMP vs. TCP): ICMP consistently more complex
- overcounting
  - routers and multi-homed hosts: estimated at <5% in paper
- undercounting
  - probes: random due to probe order;
  - use a square purview to recover single insertion in survey
  - firewall hosts: coming up

⇒ warning: if error was always pro or con, can set a bound
⇒ (but no: it can go either way)
Enterprises are Not Perfect

- USC has ~89k IPv4 addresses
- management is partially decentralized
  - no one has complete, current status of all addrs
- current status is sensitive
  - anti-file sharing requests: who was using IP x and time t?
  - will not share DHCP information with researchers
- operator knowledge ages
  - address use changes over time; tracking is incomplete
- the network operator does not know the ground truth
  - big is hard! (even where big == one enterprise)

Advantages at Your Enterprise

- getting all the local traffic
- combining passive and active to get bigger view
- still not perfect
  - passive at edge misses hosts with local-only traffic
  - printers, internal telephones, etc.
  - hard to get all traffic at the edge
  - modems? internal caches? direct peering?
  - and... how do we know USC is representative of the Internet whole?

Evaluating at USC (Our Enterprise)

<table>
<thead>
<tr>
<th>category</th>
<th>any active</th>
</tr>
</thead>
<tbody>
<tr>
<td>addresses probed</td>
<td>81,064</td>
</tr>
<tr>
<td>non-probed</td>
<td>54,078</td>
</tr>
<tr>
<td>responding any</td>
<td>27,596</td>
</tr>
<tr>
<td>responding any TCP</td>
<td>19,806</td>
</tr>
<tr>
<td>responding any ICMP</td>
<td>17,045</td>
</tr>
<tr>
<td>false</td>
<td>14,794</td>
</tr>
<tr>
<td>Passive only</td>
<td>25,706</td>
</tr>
<tr>
<td>TCP only</td>
<td>1,081</td>
</tr>
<tr>
<td>ICMP only</td>
<td>7,720</td>
</tr>
</tbody>
</table>

Census is incomplete, but can estimate error => recall is 62%

Ground Truth 3: Random Sampling

(for Case Study: Edge Address Activity)

- take a random sample of all Internet addresses
- pro:
  - could do it repeatedly
  - there is no bias
- con:
  - use of IP address space is not equally distributed
    - so many of what we pick might not be used
    - and some parts are reserved for private use
    - don't know if they're in use
- fix: probe with TCP and ICMP

Ground Truth 3: Random Sampling

(for Case Study: Edge Address Activity) (my take)

- take a random sample of all Internet addresses
- pro:
  - should be unbiased (by definition)
- con:
  - what is their truth?
  - what about rare parts of the Internet?
  - 1M addresses might only get 10 servers (?), or 10 users in developing world...
Random Sampling for Active Addresses

<table>
<thead>
<tr>
<th>category</th>
<th>active</th>
<th>non-active</th>
</tr>
</thead>
<tbody>
<tr>
<td>addresses probed</td>
<td>1,200,000</td>
<td>945,703</td>
</tr>
<tr>
<td>responding active</td>
<td>280,477</td>
<td>62,003</td>
</tr>
<tr>
<td>TCP only</td>
<td>1,088,092</td>
<td>765,530</td>
</tr>
<tr>
<td>ICMP only</td>
<td>20,145</td>
<td>14,264</td>
</tr>
<tr>
<td>both ICMP and TCP</td>
<td>382,108</td>
<td>42,983</td>
</tr>
</tbody>
</table>

- only have weaker ground truth, active probing only
- Census is still incomplete, but can estimate error
  => recall now 74%
  => confirms prior results

Ground Truth in IP Alias Resolution

- early work:
  - "M e a s u r i n g I S P T o p o l o g i e s w i t h I n t e r n e t S c a l e I P A l i a s R e s o l u t i o n T e c h n i q u e s " , K e y s ,  C C R  2010
  - "M e a s u r i n g I S P T o p o l o g i e s w i t h I n t e r n e t S c a l e I P A l i a s R e s o l u t i o n T e c h n i q u e s " , K e y s ,  C C R  2010
- recent work:
  - "M e a s u r i n g I S P T o p o l o g i e s w i t h I n t e r n e t S c a l e I P A l i a s R e s o l u t i o n T e c h n i q u e s " , K e y s ,  C C R  2010
  - "M e a s u r i n g I S P T o p o l o g i e s w i t h I n t e r n e t S c a l e I P A l i a s R e s o l u t i o n T e c h n i q u e s " , K e y s ,  C C R  2010

Case Study 2: IP Alias Resolution

- question: when are IP addresses in traceroutes the same device?
- early work:
  - "M e a s u r i n g I S P T o p o l o g i e s w i t h I n t e r n e t S c a l e I P A l i a s R e s o l u t i o n T e c h n i q u e s " , K e y s ,  C C R  2010
  - "M e a s u r i n g I S P T o p o l o g i e s w i t h I n t e r n e t S c a l e I P A l i a s R e s o l u t i o n T e c h n i q u e s " , K e y s ,  C C R  2010
- recent work:
  - "M e a s u r i n g I S P T o p o l o g i e s w i t h I n t e r n e t S c a l e I P A l i a s R e s o l u t i o n T e c h n i q u e s " , K e y s ,  C C R  2010
  - "M e a s u r i n g I S P T o p o l o g i e s w i t h I n t e r n e t S c a l e I P A l i a s R e s o l u t i o n T e c h n i q u e s " , K e y s ,  C C R  2010

Case Study 4: Effects of Cable Cuts

- question: what are the effects
  of breaks in undersea cables
  on the countries they serve?
- work-in-progress (tech report)

Ground Truth 4: Prior Work

- can compare to prior published work
  - or get and run prior code
- but can compare to prior results
- challenge:
  - errors can propagate
  - "better than before" gives no clue about "good"
Ground Truth 5: Modeling (discussion)

- **idea:** let's model the network as best we can
- **pros:**
  - simplifies the problems
  - can compare your results to alternatives, based on your knowledge
- **cons:**
  - simplifies the problems
  - but maybe alternatives that you consider are not right or missing

Modeling What Ifs

- can evaluate likely outcome of cable cut (~20Gb/s capacity)
- for assumed traffic load (50k flows)

Ground Truth 5: Modeling All Options (my take)

- **idea:** let's model the network as best we can
  - look at all possible parameters
- **pros:**
  - can look at many parameters quickly
  - if all parameters give same result, have answer!
  - if most parameters give same result, answer is likely
  - worst case: provide possible outcomes, others (w/more info, or in future) can fill in
- **cons:**
  - can be lots of parameters!
  - each layer of model adds uncertainty
  - not ground truth, but all possible truths (many incorrect!)

Some Options for Ground Truth

- ask the operators
- your enterprise
- random sampling
- prior work
- model all the things!
- (your ideas here)

Outline

- **intro:** Plato's cave
- what do we want?
- 4 case studies and 5 ground truths
- **conclusions**
imagine prisoners in a cave, chained to the wall
they cannot see the real world, instead only shadows of objects
(outrage of objects, not even of the real things!)
what is real?
the shadows? the objects that cast them?
the world above that inspired them?

So What Is Real?
(the truth we cannot directly see)
the shadows?
the objects that cast them?
the world above that inspired them?
an ideal world that could exist?

Conclusions
• strive to search for the truth
  – don’t stop at what you see
  – “best available data” today… can you do better tomorrow?
  – not not just what exists, but what should be
• use strong correctness (from info theory)
  – precision and recall, not just “correctness”
• be creative about ground truth
  – you can often dig it out, if you work
  – explore all possibilities if that’s the best you can do