


Internet Populations (Good and Bad): Measurement, Estimation, and Correlation

John Heidemann
USC/ISI

joint work with Genevieve Bartlett, Joseph Bannister, Xue Cai, Yuri Pradkin, Ramesh Govindan, Christos Papadopoulos, Lin Qian

Institute for Computational and Experimental Research in Mathematics:
Mathematics of Data Analysis in Cyber Security
20 October 2014

This work is classified as non-human-subjects research by USC's IRB (IRB00001648).



This research is sponsored by the Department of Electrical and Information Systems (EIS) and the Department of Information Systems (IS) at the University of Southern California (USC) and the Information Systems Research Laboratory (ISRL) at the University of California, San Diego (UCSD). This research is also supported by the National Science Foundation (NSF) under grant number 0946542 and grant number 0946542. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation thereon. This notice does not constitute an endorsement or approval of the views or opinions of the authors or those of USC or the U.S. Gov't.

Internet Censuses

- since 2003 we've taken *Internet Censuses*
- probe all 4 billion IPv4 addresses in ~2 months
 - pings (ICMP echo request)
 - 0.0.0.0 to 255.255.255.255 (except not private or multicast or reserved)
- one census is at right:
 - each pixel: average from 65k addrs
 - brightness (red & green): how many respond
 - plotted as Hilbert Curve

1D: 0 1 2 3 4 5 6 7 ...

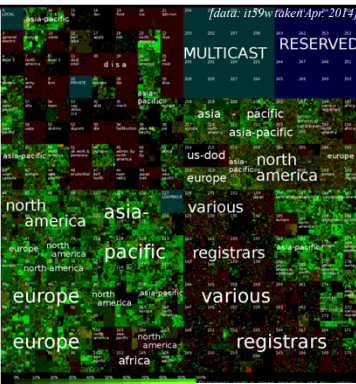
etc.

2D: 0 14 15

3 2 13 12

4 8 11


5 6 9 10



Heidemann et al., "Census and Survey of the Visible Internet", ACM IMC, Oct. 2008. doi.acm.org/10.1145/1452520.1452542

Can we quantify:

- how big is the Internet?
- how reliable?
- how does it change?



Internet Population Measurement / 2014-10

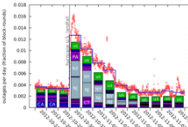
From Census to Outages


can pings detect network outages?
yes...with proper interpretation

responding and stopping... says something

net up

down






Quan et al., "Trinocular: Understanding Internet Reliability Through Adaptive Probing", ACM SIGCOMM, Aug. 2013. doi.acm.org/10.1145/2486001.248017

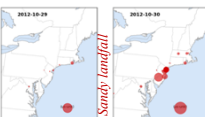
Measuring the Internet

Internet censuses



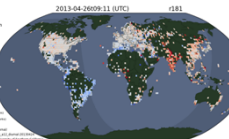
about 650-780M active IPv4 addresses (Apr. 2014)

network outages




about 3x more U.S. outages day after Sandy landfall

diurnal networks



11% of blocks are diurnal (25% relaxed diurnal); more in some areas than others



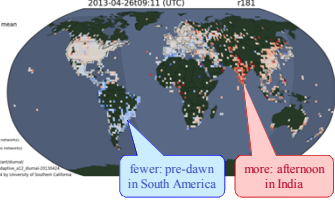
From Outages to Sleep (Diurnal Networks)

pinging the Internet for outages shows active addresses


red: more than typical
white: typical
blue: fewer

parts of the Internet sleep: they are more active during the day

fewer: pre-dawn in South America
more: afternoon in India




paper: Quan et al., "When the Internet Sleeps: Correlating Diurnal Networks With External Factors", ACM IMC, Nov. 2014. doi.org/10.1145/2663716.2663721
animation: <http://www.isi.edu/ant/diurnal/>




Why Measure the Internet?

- baseline for cybersecurity
 - must know *how many* to find if *disproportionately bad*
- network resilience is a form of security
- data helps correlation studies
 - what policies make things better?


Internet Population Measurement / 2014-10
8

Big Measurements: Probing the Whole Internet


- actually, pretty easy
- in 2003
 - 4 machines in parallel over 2-3 months
- today
 - 1 machine does most of net every 11 minutes
 - 19k probes/s per core
- our goal: careful and complete and correct
 - but others optimize for speed: ZMap scans in 15 minutes (with high parallelism)


Internet Population Measurement / 2014-10
11

What About Math?


an essential tool under all of this

- statistics: evaluating correctness
- census taking and sampling
- inference: Bayesian analysis
- correlation (with imperfect data)


Internet Population Measurement / 2014-10
9


Polite Measurement

- a *big deal* if you want to keep going
 - people *do* complain
 - and they cc the president of Colorado State U. to help
 - often naively
 - ex: Belkin routers report “ping of death” (fixed since 2000)
- our approach
 - have an opt-out policy and blacklist
 - broad measurements are slow
 - frequent measurements are careful


Internet Population Measurement / 2014-10
12

Challenges

- big measurements
- polite measurement
- calibrated measurement

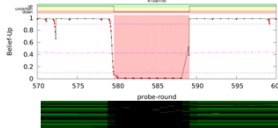

Internet Population Measurement / 2014-10
10


Politeness by Modeling

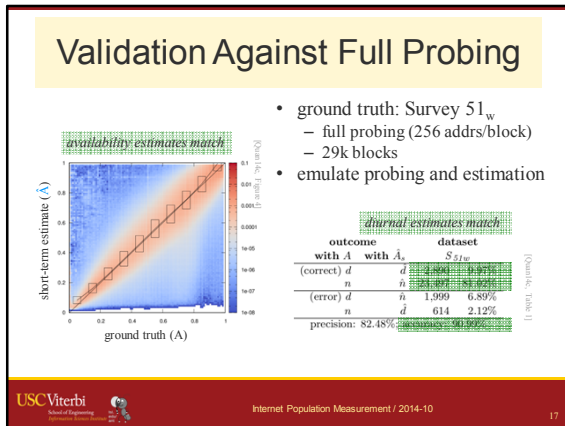
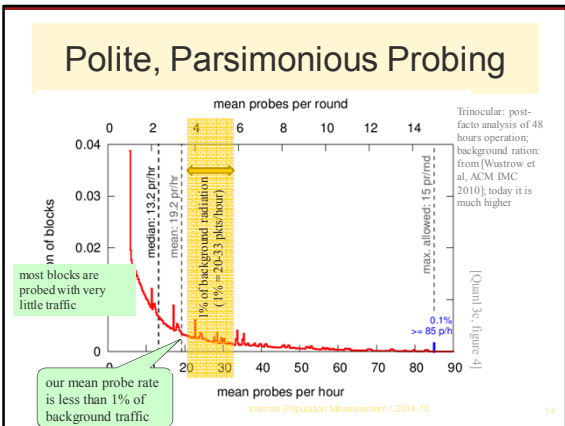
- model each block b
 - $E(b)$: addresses that ever respond
 - $A(E(b))$: Pr[those addresses will respond]
- build model from Internet census
- refine $A(E(b))$ online
 - concurrent with outage detection
 - exponential weighted moving average

• Bayesian model of block state

probe result	prior U^b	$P(\text{probe}(U^b))$	response
a	U	$1 - A(E(b))$	inactive addr.
p	U	$A(E(b))$	active addr.
n	\bar{U}	$1 - (1 - \epsilon)^{ b }$	no-response to block
p	\bar{U}	$(1 - \epsilon)^{ b }$	lose router?

$$B^b(U) = \frac{P(p|U)B(U)}{P(p|U)B(U) + P(p|\bar{U})B(\bar{U})}$$


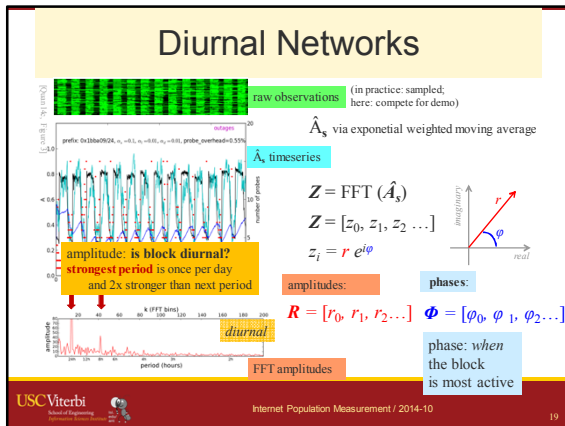

Internet Population Measurement / 2014-10
13



- ### Calibrated Measurement
- what did you *actually* measure?
 - how precise?
 - is there systematic bias?
 - general approach to calibration:
 - ask operators for ground truth
 - but they don't know
 - approximate ground truth by overprobing
 - compare a narrow slice (for us: USC)
 - compare a random sample and second location
- USC Viterbi, Internet Population Measurement / 2014-10.

- ### Powerful Tools
- correctness: statistics
 - census taking, understanding sampling error
 - calibration different observations
 - decision making: Bayesian Inference
 - interpretation
 - spectral analysis
 - correlations with other data
- USC Viterbi, Internet Population Measurement / 2014-10.

- ### Validating Diurnal
- "When the Internet Sleeps: Correlating Diurnal Networks with External Factors", Quan et al., ACM IMC 2014*
- availability
 - vs. full probing: strong correlation 0.957 (§3.2.2)
 - diurnal
 - vs. full probing: good precision (82%) and accuracy (90%) (§3.2.3)
 - sensitivity (simulation): works w/10% diurnal (§3.2.2)
 - vs. operators: few false positives (at most 3%) (§3.2.4)
 - vs. second location: good agreement (85% exact, 98.8% relaxed) (§3.3)
- USC Viterbi, Internet Population Measurement / 2014-10.



Phase Correlates with Longitude

- compare block phase
 - when usage peaks
- vs. geolocated longitude
 - when people wake up

(phase is modular, so "unroll" it based on longitude)

=> phase correlates with longitude (conf: 0.763)

USC Viterbi School of Engineering Internet Population Measurement / 2014-10 20

Better Powerful Tools?

- correctness: statistics
 - census taking, understanding sampling error
 - calibration different observations
- decision making: Bayesian inference
- interpretation
 - spectral analysis
 - correlations with other data

The Internet is not human: how is an Internet Census different?

From data to policy?

USC Viterbi School of Engineering Internet Population Measurement / 2014-10 23

Phase Predicts Longitude

phase predicts longitude but only roughly: worst case: ±90°, best: ±10°

USC Viterbi School of Engineering Internet Population Measurement / 2014-10 21

Thoughts about Internet Census Taking

- human census taking
 - census is labor intensive and relatively slow (months)
 - targets are big and slow
- Internet census taking
 - can be really fast
 - targets do move: DHCP
- when do these differences matter?
 - does fast help? what about repetition?
 - can we model movement?
 - when should we do adaptive probing?
 - can we exploit these to improve our estimates of error?

USC Viterbi School of Engineering Internet Population Measurement / 2014-10 25

Correlating Diurnal with Many Factors

- apply ANOVA (Analysis of Variance)
 - smaller values => greater correlation
- factors:
 - GDP, electricity consumption, number of Internet users per host, time of first block allocation, mean age of allocations
 - link type (inferred from DNS)

	per-capita gdp	p.cpt. elec. cons.	l-aet. users/host	age/first alloc.	mean age/alloc.
gdp	0.000111	0.306	0.822	0.789	0.595
elec.	0.000111	0.703	0.609	0.830	0.00114
users/host	0.0366	0.959	0.819	0.00114	0.00114
first alloc.				0.00114	
mean alloc.				0.00114	

dynamic => diurnal (as expected) but also *dsl*

diurnal & GDP: money => always on

electricity and age: more dynamic => more diurnal when "new" (IPs are tight?)

policy studies are early: approach seems promising

USC Viterbi School of Engineering Internet Population Measurement / 2014-10 22

Thoughts about Correlation and Policy

- how best to study correlation?
 - higher quality input?
 - correlation with different precision inputs?
 - ex: coarse grain estimates from surveys or highly aggregated data vs. precise, direct measurements
- how do we get to root causes?
 - natural experiments that compare policy A vs. policy B?
- cross-disciplinary work
 - with policy experts? social science and public policy
 - statistics

USC Viterbi School of Engineering Internet Population Measurement / 2014-10 26

Network Measurement Needs Math

- Internet measurement is fascinating
- analysis is essential
- can we do better?

- your take?
 - papers: <http://www.isi.edu/ant/pubs/>
 - our data is available at no cost;
<http://www.isi.edu/ant/traces/> or
<https://www.predict.org>