

# Uses and Challenges for Network Datasets

John Heidemann\* and Christos Papadopoulos

USC/ISI and CSU



4 March 2009



*funding agencies*

*policy makers*

# Why collect network data?

*network operators*

*security researchers*



# Some Reasons...

- how stable is the Internet topology?
- how widespread is IPv6? DNSsec? IPsec?
- how prevalent is peer-to-peer traffic?
- are ISPs blocking certain traffic?
- how many hosts have Conflictor worm?
- how many hosts are on botnets? control botnets?
- how many hosts are there at all?
- what does the net look like?



# Some Reasons...

- how stable is the Internet topology?

*reliability* IPv6? DNSsec? IPsec?

- how prevalent is peer-to-peer traffic?

- are ISPs blocking certain traffic?

*security* how many hosts have Conficker worm?

- how many hosts are on botnets? control botnets?

how many hosts are there at all? *trends*

- what does the net look like?



# Not “Why Collect Network Data”, but *What...*

*what...*

- problems does the Internet face?
- technical and policy questions follow?
- data can answer these questions?
- challenges in providing data?



# Not “Why Collect Network Data”, but *What...*

*what...*

- **problems does the Internet face?**
- **technical and policy questions follow?**
- **data can answer these questions?**
- **challenges in providing data?**



# *Problems -> Questions -> Data*

I want both  
*old and new*  
services

I want a  
*safer* Internet

I want  
to know how  
these *come*  
*together*

I want *business policies*  
*to promote network*  
improvements

I want to know  
how *technology*  
*changes* the net

I want to know how  
*people* use the net



# Problems -> *Questions* -> Data

what is typical traffic?  
at ISP? at home?  
on mobile devices?

what is atypical traffic?  
malware? botnets?  
spam? misconfigured devices?  
hostile attacks?

how do traffic  
and topology  
interact?  
congestion?  
loss?  
rate limits?

what is provider-level topology?  
how much competition?  
effects of regional policies?

what is router-level topology?  
how robust is routing?  
broadband deployment?

what is host-level topology?  
how big is the net?  
running out of addresses? trends?



# Problems -> *Questions* -> Data

**what is typical traffic?**

at ISP? at home?

on mobile devices?

**how do traffic  
and topology**

**interact?**

congestion?

loss?

rate limits?

**what is atypical traffic?**

malware? botnets?

spam? misconfigured devices?

hostile attacks?

**what is provider-level topology?**

how much competition?

effects of regional policies?

**what is router-level topology?**

how robust is routing?

broadband deployment?

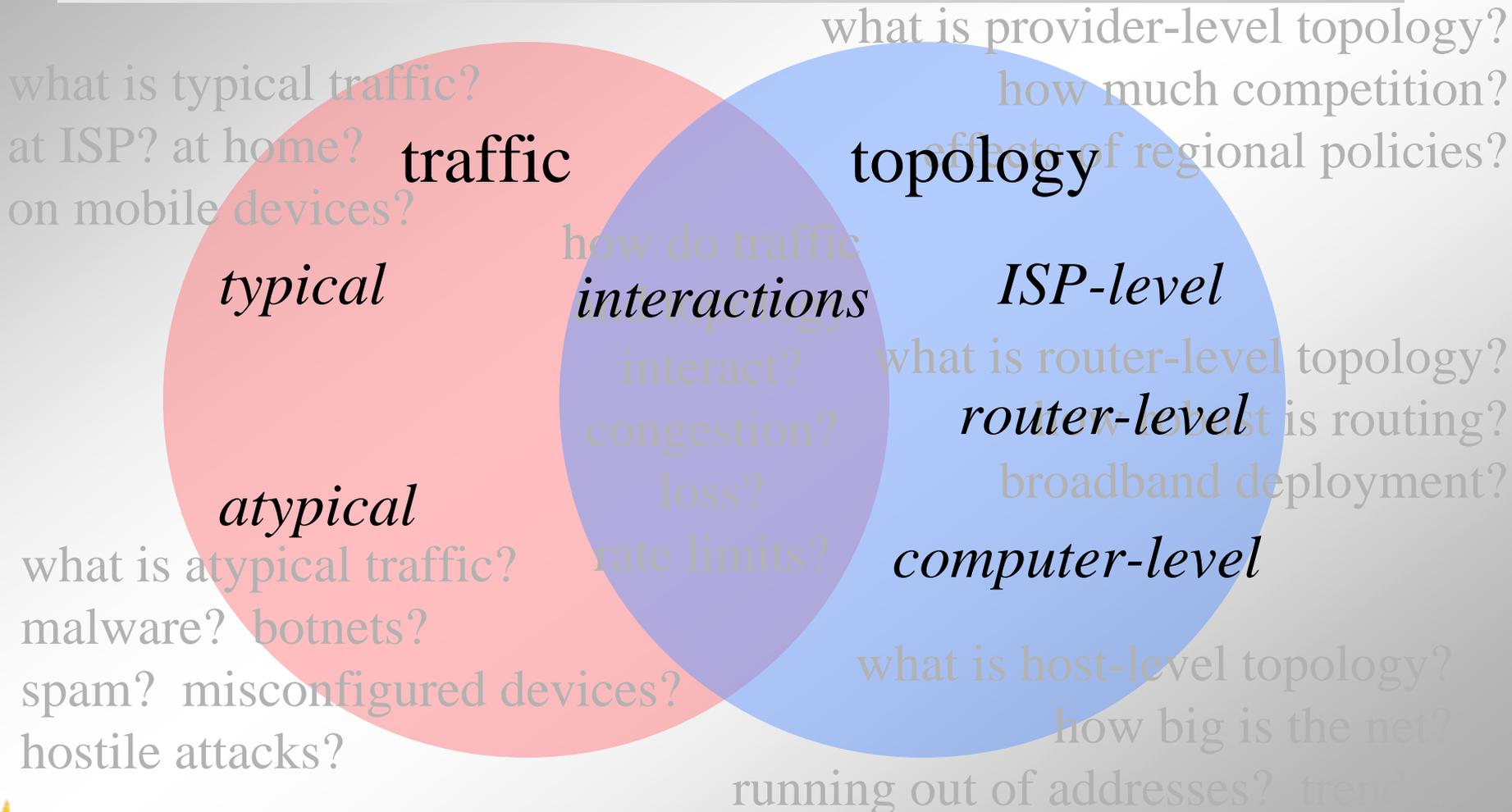
**what is host-level topology?**

how big is the net?

running out of addresses? trends?



# Problems -> Questions -> *Data*



# *Problems -> Questions -> Data*

For specific research problems and questions, see the paper Table 1 and Sections 2

**Table 1. General topics in network research and applications in those topics.**

<b>topic</b>	<b>applications</b>
<b>network traffic</b>	
typical traffic	protocol design, congestion control, router buffer sizing, traffic modeling, new traffic types
atypical traffic	malware detection: denial-of-service attacks, worm, virus spread, malware; spyware; unusual traffic types; protocol verification
<b>network topology</b>	
AS-level	understanding business relationships
router- or link-level	evaluation of network robustness, cross-section throughput, network coordinate systems
address-level	evaluation of network size
<b>topology and traffic</b>	localizing attack sources, mapping network to geography, physical cross-section throughput



# Not “Why Collect Network Data”, but *What...*

*what...*

- problems does the Internet face?
- technical and policy questions follow?
- **data can answer these questions?**
- challenges in providing data?



# Problems -> Questions -> *Data*

For specific data types,  
see the paper Table 2 and Section 3

**Table 2. List of data classes, instances of that class, and providers of that data (partially derived from data assembled by Jody Westby).**

class	examples (formats)	Providers	Privacy Concerns
local observations	packet headers for general links (pcap or ERF)	CAIDA [9], LANDER [72], LBNL [54]	addresses
	packet headers for events such as attacks or worm spread (pcap or ERF)	CAIDA, LANDER, MERIT [71], LBNL	addresses
	full packet contents (pcap or ERF)	<i>unavailable</i>	user data and addresses
	flow-level traces (netflow)	MERIT	addresses
	router statistics (SNMP)	-	-
local inferences	intrusion detection alerts (Snort, Bro, etc.)	<i>unavailable</i>	addresses and system data
	logs (syslog, firewall, spam)	LogAnalysis.org [5]	addresses and system data
network-wide observations	active IP addresses	<i>unavailable</i>	general addresses
	DNS requests	<i>unavailable</i>	user data
	BGP tables	MERIT, RouteViews [58]	-
	end-host scans (ping or nmap)	LANDER	addresses
	topology scans (traceroute)	CAIDA	general addresses
network-wide inferences	VOIP call records	PCH [48]	addresses
	BGP hijackings (PHAS, bgpmon)	<i>unavailable</i>	-
	darknet/telescope packet headers (pcap)	CAIDA, MERIT	addresses
	darknet/telescope full packets (pcap)	CAIDA, MERIT	addresses and user data
	IP reputations	Spamhaus [1]	addresses



# Digression: PREDICT

- one of the first HSARPA programs from DHS
- goal: collect network data for Internet security research
  - prior available traces are short and local
    - impossible to study long-term trends
  - prior anonymization schemes obscure all topology information
    - impossible to study worm propagation
  - prior data is mostly from backbones or local networks
    - missing data from regional networks, missing specialized data
- PREDICT aims to address these problems
- end result: stronger Internet Infrastructure
  - better firewalls, understanding of attackers and the Internet



# PREDICT Challenges

- matching data and researchers
    - web site with metadata: <http://www.predict.org>
  - getting interesting data
    - five data providers: USC, MERIT, CAIDA, LBL, PCH
    - packet traces, VoIP logs, network topology, darknets, etc.
    - getting data to researchers: RTI and SRI
  - balancing privacy and research needs
    - technical: anonymization
    - policy: process for matching data to researcher
    - legal: researcher MOAs and vetted process
- substantial progress to making data sharing possible



# Not “Why Collect Network Data”, but *What...*

*what...*

- problems does the Internet face?
- technical and policy questions follow?
- data can answer these questions?
- **challenges in providing data?**



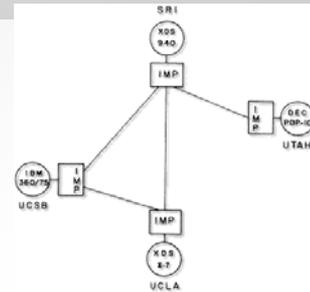
# Challenge: Managing Privacy

- anonymization is central
    - if you can't manage privacy, you can't give out the data
  - tension with researchers' needs
    - can't anonymize information needed for the answer
  - my opinion: need both *technical* and *policy* components
    - unlikely (to me) that there is just a technical solution
- need additional study and improvements



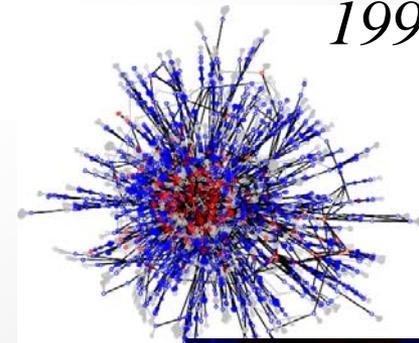
# Challenge: New Datatypes

- the network keeps changing
  - new traffic types
  - new attacks and vulnerabilities
  - evolving topology
- data collection must keep changing
  - new kinds of collection
  - ongoing collection
  - multiple points of view

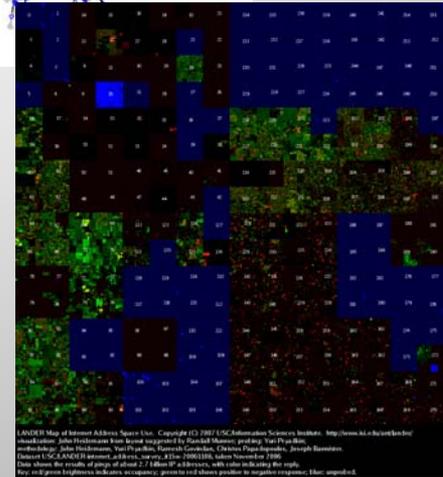


1969

1999



2006



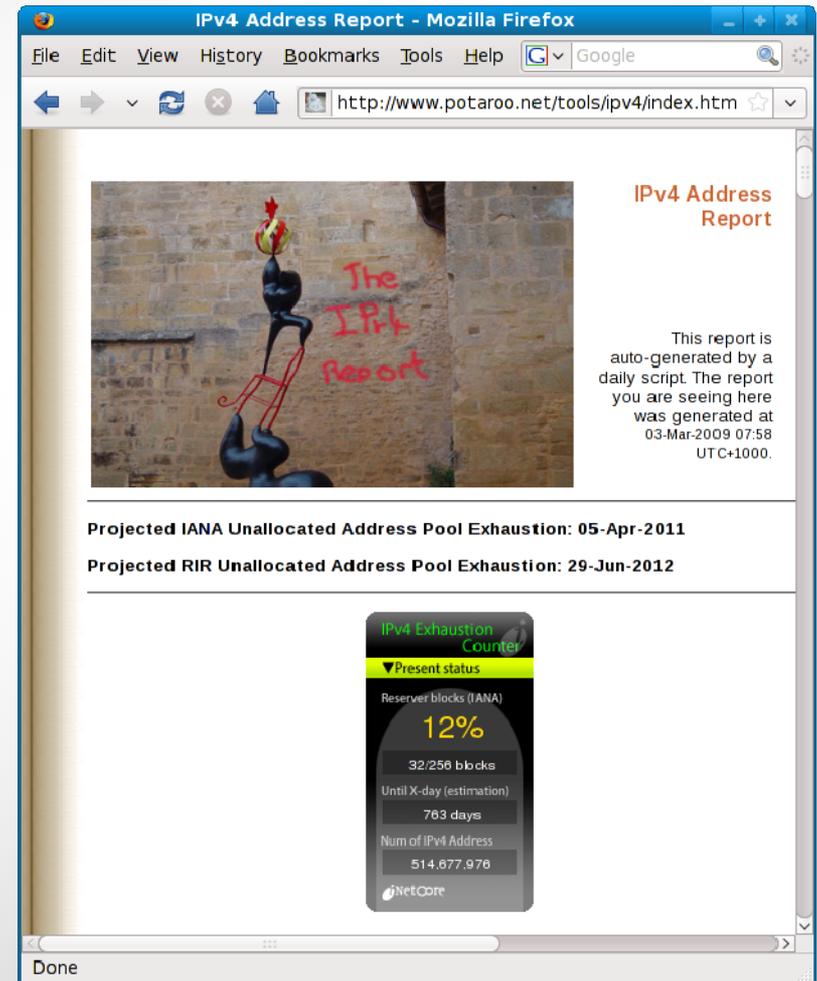
# Challenge: Validating New Approaches

- establishing a new methodology is difficult
  - what is ground truth?
- need
  - small, controlled studies to establish ground truth
    - often with relaxed anonymization from volunteers
  - iterate: data collection  $\Leftrightarrow$  research
    - ensure data answers the question
    - ensure answer is correct



# Example Problem-Question-Data: IPv4 Address Use

- problem: exhaustion of the IPv4 address space
- question: how effective is current address use?
- data: take a *census* of all addresses
  - probe each address to record use



# Example Ground Truth: Is the Census Accurate?

the raw data: 2.8 billion probed,  
only ~4% reply

validation:

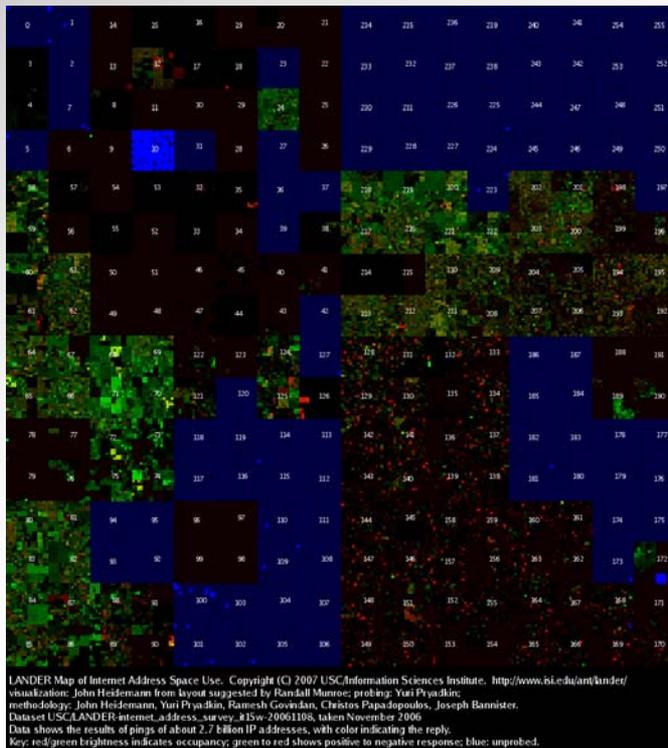
compare to *ground truth* at USC

- no actual list of address use
- so compare to addresses observed in network traffic

conclusion:

estimate is 40-45% low  
(but useful with correction!)

*result required multiple datasets and  
controlled anonymization*



# Example Iteration: Observing IPv4 Addresses

- in collecting and refining this data, we went through *five* iterations
  - started simple
  - got much more accurate as we learned
- iteration and use of data is essential

**Table 3. Evolution of information saved in address scans.**

version	year	information
0	2003	bit per responding addresses, for ICMP echo reply only
0.1	2004	adds TTL, RTT
1	2005	new format: encoded ICMP type and reply code (not all saved), TTL, RTT, for three ICMP message types only
2	2007	new format: full ICMP type and reply code, TTL, RTT, for all valid ICMP message types
2.1	2008	adds pcap capture of all invalid ICMP message types



# Open Challenges

- improved anonymization techniques
  - need iteration: attacks  $\Leftrightarrow$  improvements
- new data collection
  - shift from getting “*the dataset*”  
to *process for continuous measurement*
- need broader data collection ecosystem
  - technical
  - *and* policy and legal



# Conclusions

- need network *data to improve* the Internet
- real progress so far
  - *new types of data* available
  - *better process* for getting and using data
- working on *building ecosystem*

<http://www.isi.edu/ant/>

<http://www.predict.org/>

