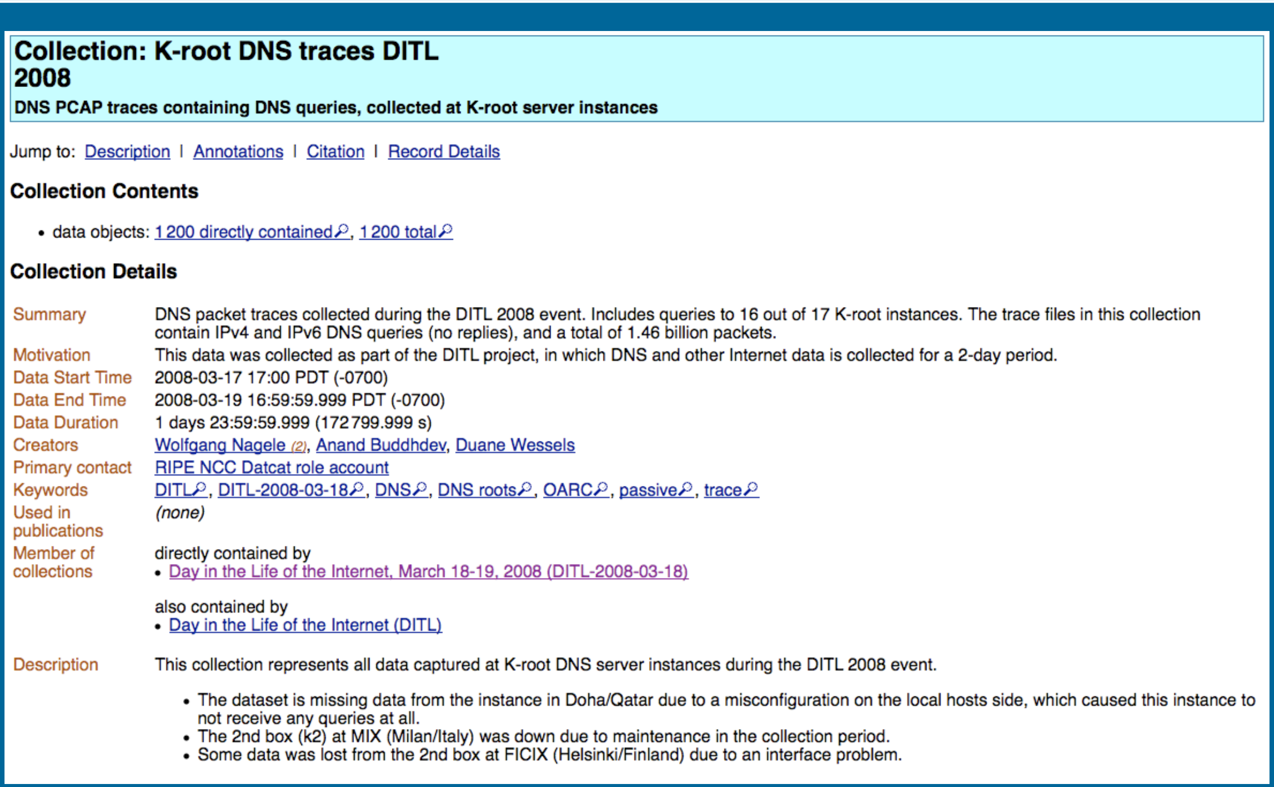
Measurable progress on empirical Internet science: (meta-)data repositories and privacy-sensitive data-sharing frameworks

kc claffy CAIDA
kc@caida.org

DatCat

- Internet Measurement Data Catalog (DatCat): http://imdc.datcat.org
  - Facilitate searching for and sharing of data among researchers
  - Enhance documentation of datasets via a public annotation system
  - Advance network science by promoting reproducible research
- No data storage, only meta-data:
  - How many IPv6 pkts in this pcap file?
  - How was this pcap file created?
  - How to get access to data

DatCat



Collection: K-root DNS traces DITL 2008

DNS PCAP traces containing DNS queries, collected at K-root server instances

Jump to: Description | Annotations | Citation | Record Details

Collection Contents

- data objects: 1 200 directly contained, 1 200 total

Collection Details

Summary	DNS packet traces collected during the DITL 2008 event. Includes queries to 16 out of 17 K-root instances. The trace files in this collection contain IPv4 and IPv6 DNS queries (no replies), and a total of 1.46 billion packets.
Motivation	This data was collected as part of the DITL project, in which DNS and other Internet data is collected for a 2-day period.
Data Start Time	2008-03-17 17:00 PDT (-0700)
Data End Time	2008-03-19 16:59:59.999 PDT (-0700)
Data Duration	1 days 23:59:59.999 (172 799.999 s)
Creators	Wolfgang Nagele (2), Anand Buddhdev, Duane Wessels
Primary contact	RIPE NCC Datcat role account
Keywords	DITL, DITL-2008-03-18, DNS, DNS roots, OARC, passive, trace
Used in publications	*(none)*
Member of collections	directly contained by • Day in the Life of the Internet, March 18-19, 2008 (DITL-2008-03-18) also contained by • Day in the Life of the Internet (DITL)
Description	This collection represents all data captured at K-root DNS server instances during the DITL 2008 event. • The dataset is missing data from the instance in Doha/Qatar due to a misconfiguration on the local hosts side, which caused this instance to not receive any queries at all. • The 2nd box (k2) at MIX (Milan/Italy) was down due to maintenance in the collection period. • Some data was lost from the 2nd box at FICIX (Helsinki/Finland) due to an interface problem.

- Current stats:
  - 121 datasets
  - Representing 26 TB of data

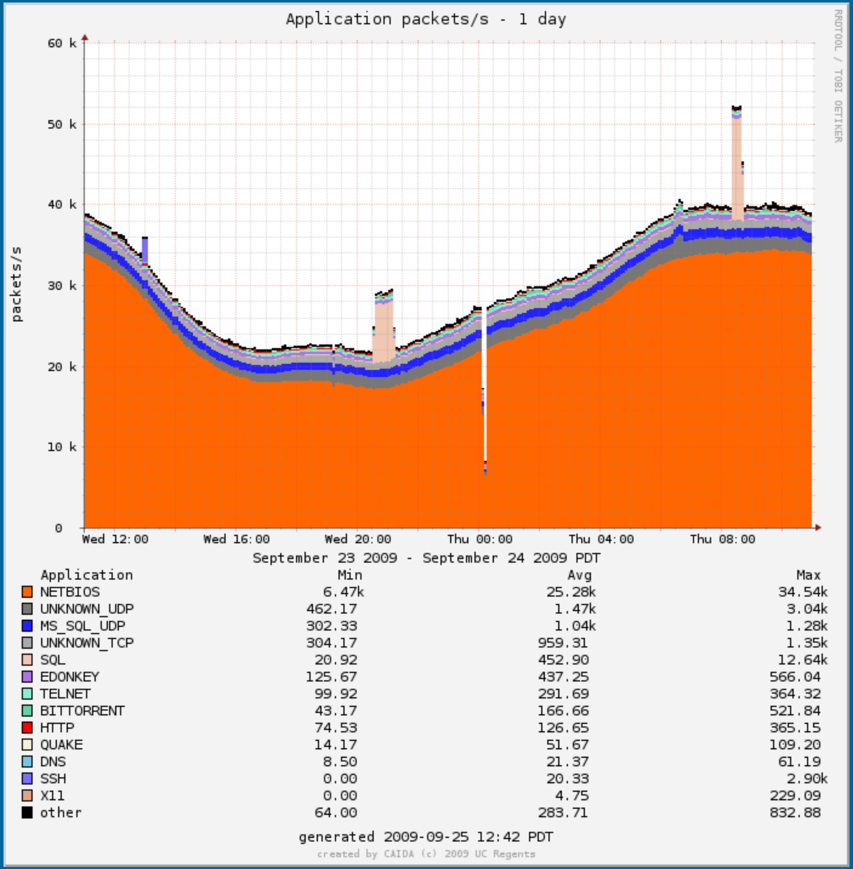
PREDICT

- Dept. of Homeland Security (DHS) attempt to formalize data-sharing relationships
- Lots of legal documentation
- Also some data being collected, not necessarily what's needed
  - Mostly topology data
  - Traffic data challenging
- most recent PI meeting:
  - http://www.caida.org/publications/presentations/2009/predict_pi_aug/predict_pi_aug.pdf

Data-Sharing framework

- DHS S&T interested in “Belmont report” as a model
- blog.caida.org: “what's Belmont got to do with it?”
- Proposing privacy-respecting Internet data sharing framework
  - – http://www.caida.org/publications/papers/2009/engaging_data/
  - – Will experiment with on UCSD Network telescope data

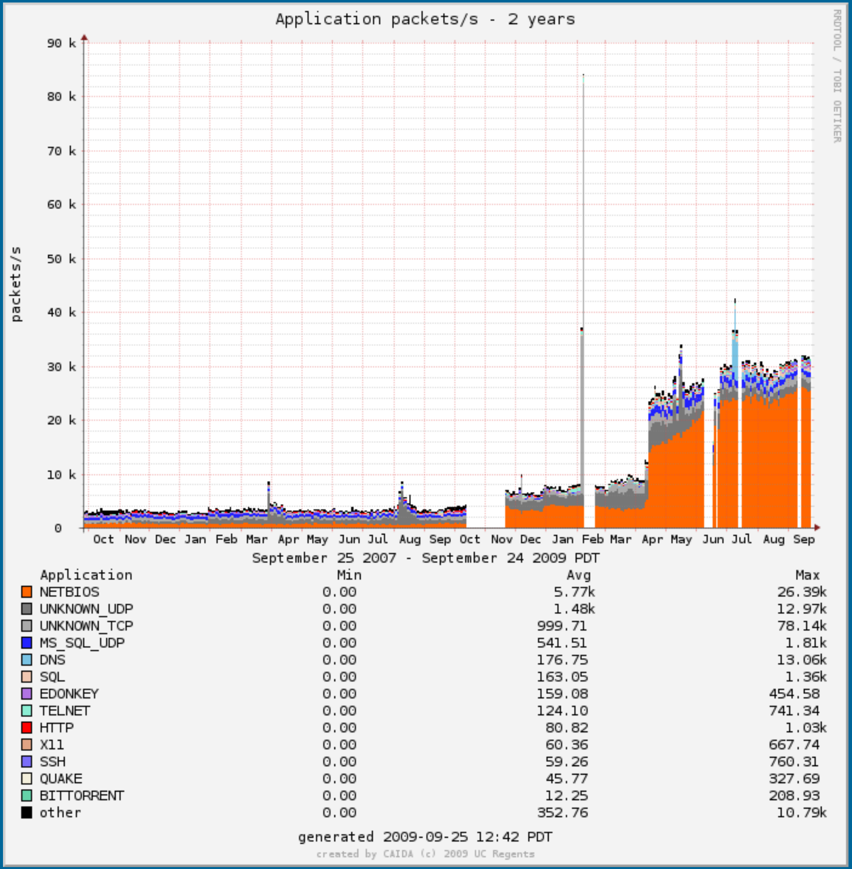
Data-Sharing framework



	Application	Min	Avg	Max
🟧	NETBIOS	6.47k	25.28k	34.54k
⬛	UNKNOWN_UDP	462.17	1.47k	3.04k
🟦	MS_SQL_UDP	302.33	1.04k	1.28k
⬜	UNKNOWN_TCP	304.17	959.31	1.35k
🟫	SQL	20.92	452.90	12.64k
🟪	EDONKEY	125.67	437.25	566.04
🟩	TELNET	99.92	291.69	364.32
🟩	BITTORRENT	43.17	166.66	521.84
🟥	HTTP	74.53	126.65	365.15
⬜	QUAKE	14.17	51.67	109.20
🟦	DNS	8.50	21.37	61.19
🟪	SSH	0.00	20.33	2.90k
🟫	X11	0.00	4.75	229.09
⬛	other	64.00	283.71	832.88

generated 2009-09-25 12:42 PDT

created by CAIDA (c) 2009 UC Regents



	Application	Min	Avg	Max
🟧	NETBIOS	0.00	5.77k	26.39k
⬛	UNKNOWN_UDP	0.00	1.48k	12.97k
⬜	UNKNOWN_TCP	0.00	999.71	78.14k
🟦	MS_SQL_UDP	0.00	541.51	1.81k
🟦	DNS	0.00	176.75	13.06k
🟫	SQL	0.00	163.05	1.36k
🟪	EDONKEY	0.00	159.08	454.58
🟩	TELNET	0.00	124.10	741.34
🟥	HTTP	0.00	80.82	1.03k
🟫	X11	0.00	60.36	667.74
🟪	SSH	0.00	59.26	760.31
⬜	QUAKE	0.00	45.77	327.69
🟩	BITTORRENT	0.00	12.25	208.93
⬛	other	0.00	352.76	10.79k

generated 2009-09-25 12:42 PDT

created by CAIDA (c) 2009 UC Regents

Attack traffic to UCSD Network telescope

QuickTime™ and a
decompressor
are needed to see this picture.

INRDB, the Internet Number Resources Database

Emile Aben RIPE NCC / CAIDA
emile.aben@ripe.net

RIPE (NCC)

- RIPE : Europe++ community, focus on administrative and technical coordination about the Internet
- RIPE NCC :
  - Supporting organization for RIPE
  - Driven by RIPE community
- RIPE NCC Science Group :
  - Making science work for the RIPE (NCC)

Data problem

- RIPE NCC has large volumes of data on Internet number resources (IPv4, IPv6, ASN) used for:
  - policy impact analysis
  - measuring quality of registration data
  - research (Mediterranean cable cut / Youtube hijack incident)
- Problem: heterogeneous data: different file formats, databases
- Solution: fast / uniform access to this data

INRDB

- Internet Number Resource DataBase
- A system to store and retrieve large amounts of Internet Number Resource related data
- Efficiently stores “blobs”, ie. observations on Internet resources and the time intervals related to these resources
- Efficiently query for Internet resources
- For now: a prototype service

Our ‘public’ INRDB instance

- Runs on 6 servers , 64 G memory total
- Data contained:
  - BGP table dumps from RIPE NCC RIS service
  - RIPE database (whois + IRR)
  - RIR stats (when was a resource allocated)
  - CAIDA AS relationships
  - CAIDA Reverse DNS lookups
  - IANA assignments IPv4 and IPv6
  - Blacklists (spamlists)
  - GeoIP (maxmind)
- Also other instances: Internal, test@CAIDA
- Over 1 G blobs and 7.5 G time intervals (‘94-’09)

Querying INRDB

- Syntax tailored to Internet Number resources and time dimension
- More/less specific IP prefixes
  - – 'show me all information you have on prefixes contained in 193.0.0.0/21' : -M 193.0.0.0/21
- Restrict by time
  - – 'show me all information you have between 2002 and 2008 : +sT 2002-01-01 +eT 2008-01-01
- Full example, from data class RIPE_DB
  - – +dc RIPE_DB +sT 2002-01-01 +eT 2008-01-01 -M 193.0.0.0/21
- And more …

When not to use INRDB

- Not meant as general data store, focus on information on number resources and time dimension
  - Example: Internet users per country (not a number resource)
- Less efficient on high entropy data
  - Real benefits show when blobs can be aggregated over time and/or are valid for multiple time intervals

INRDB / MOMENT mediator

- INRDB: specialized
  - High query speed on large volumes of data
- MOMENT mediator: generalized
  - Rich query syntax
  - Rich functionality (middleware)

INRDB + MOMENT mediator ?

- INRDB as data source (of data sources) for MOMENT seems possible, reverse also
  - But you are the experts :)
  - Not currently standard SQL, XML or RDF
- Open question
  - What control does a data service have over who uses data and under what policies?
- Benefits
  - For MOMENT : All TBs of INRDB data become available through mediator architecture
  - For RIPE : Access to various mediator services. Anonymization, workflow etc.

Questions?