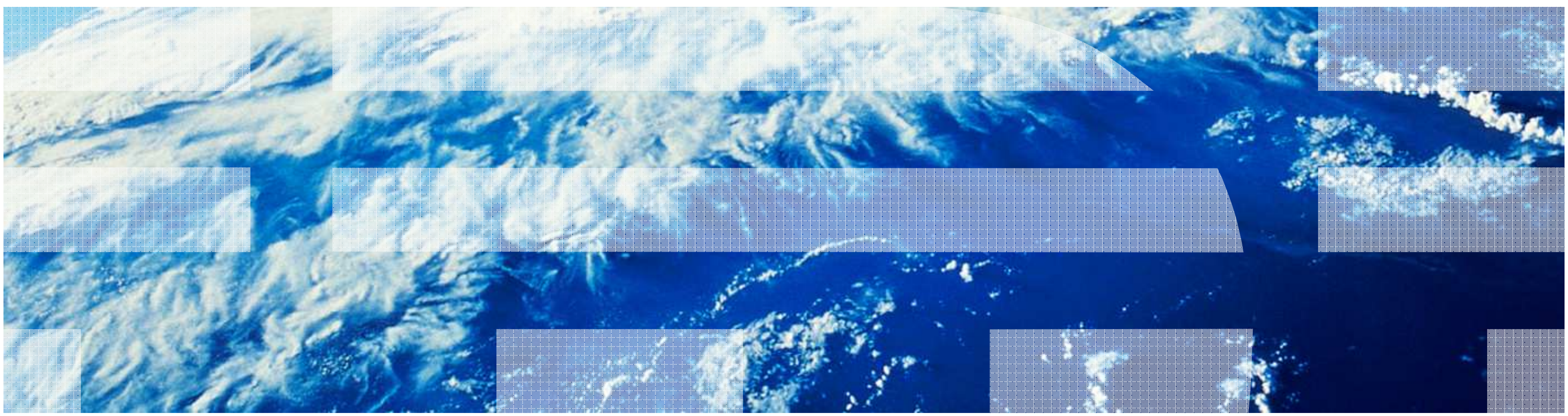
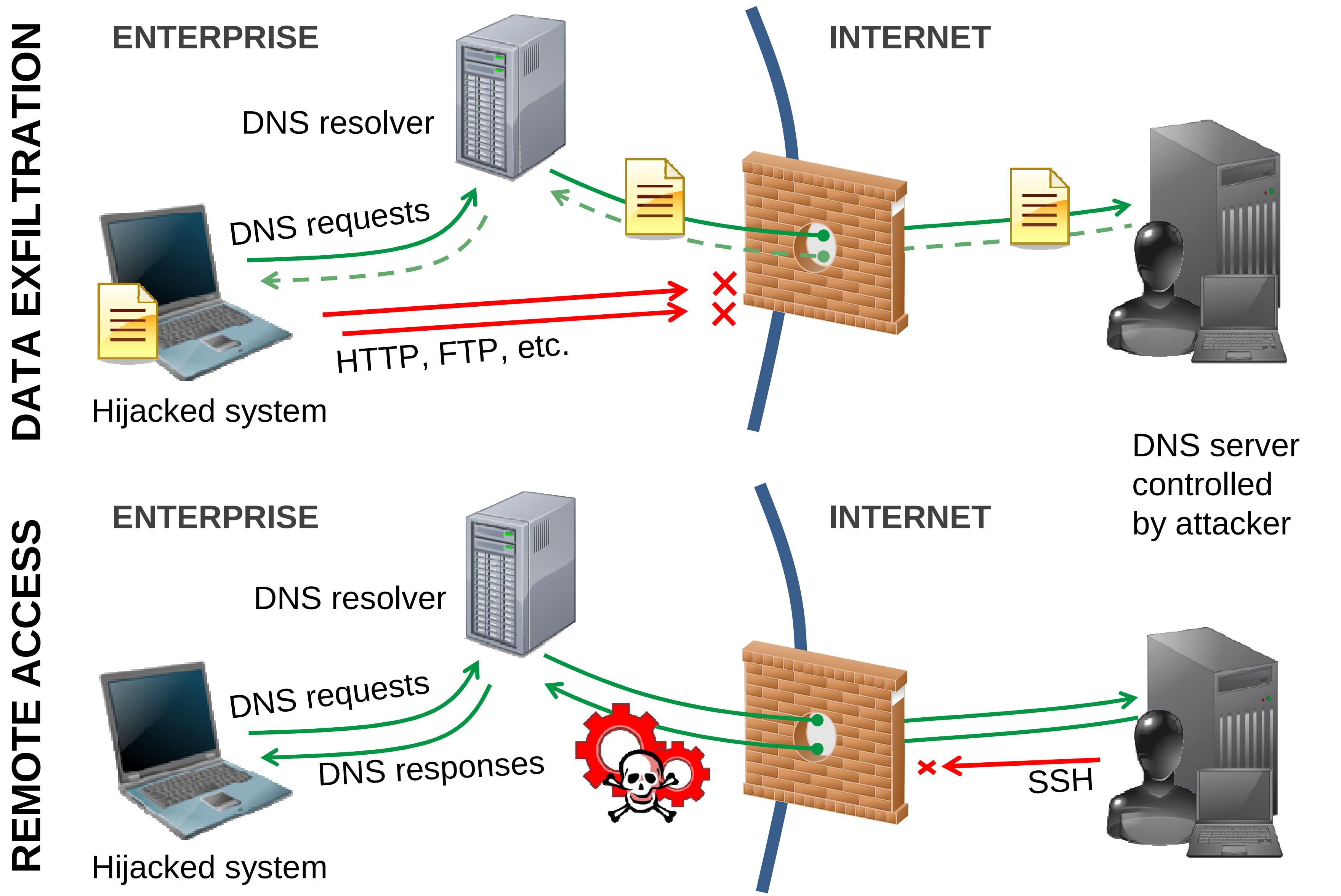
Presenting: Wietse Venema – IBM Research, Yorktown Heights, USA

Practical Comprehensive Bounds on Surreptitious Communication Over DNS

Vern Paxson[1,2] Mihai Christodorescu[3] Mobin Javed[1] Josyula Rao[4] Reiner Sailer[4]
Douglas Schales[4] Marc Ph Stoecklin[4] Kurt Thomas[1] Wietse Venema[4] Nicholas Weaver[2,5]
[1]UC Berkeley [2]ICSI [3]Qualcomm Research [4]IBM research [5]UC San Diego



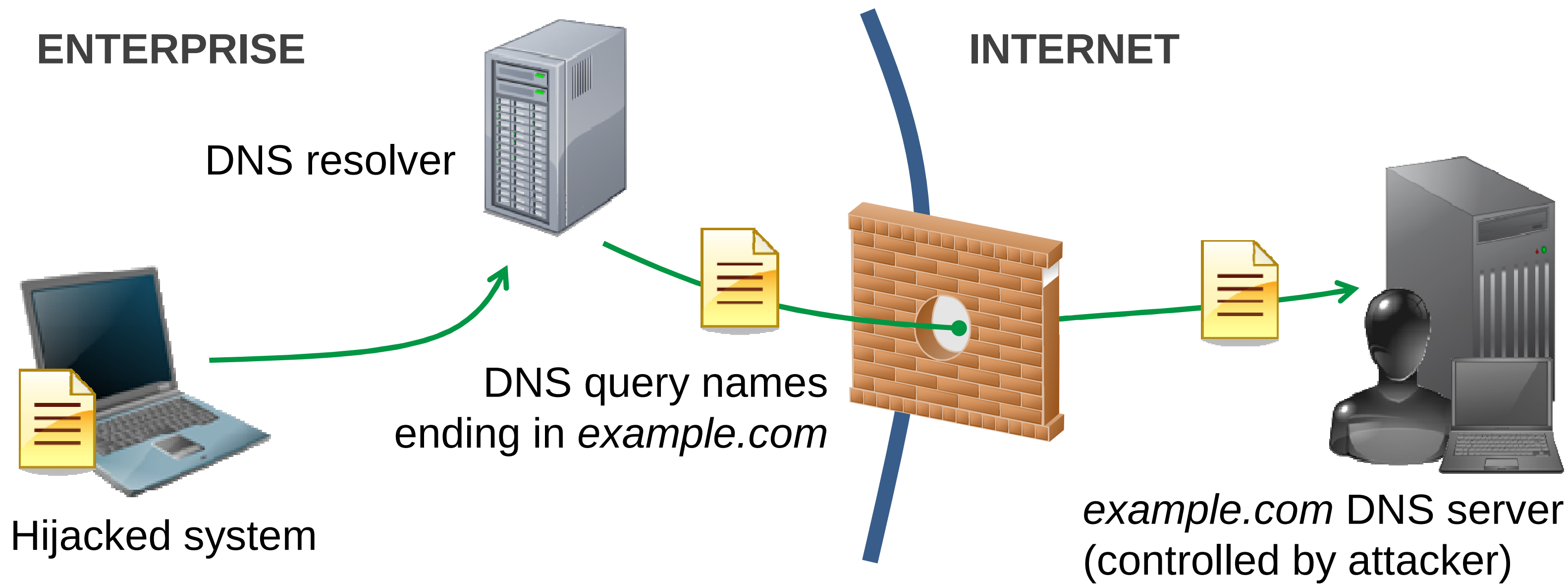


DATA EXFILTRATION
ENTERPRISE
INTERNET
DNS resolver
DNS requests
HTTP, FTP, etc.
Hijacked system
DNS server
controlled
by attacker
REMOTE ACCESS
ENTERPRISE
INTERNET
DNS resolver
DNS requests
DNS responses
SSH
Hijacked system

Our work in a nutshell

~~One query per day,~~
~~www.example.com to send “0”~~
~~mail.example.com to send “1”~~

- **Bound information content of DNS query sequences.**
- **4 kB/day per client and domain (*site*.com, *site*.co.uk).**
- **Lossless (reversible) compression: no false negatives.**
- **1-2 Alerts/week for enterprise-scale networks.**
- **59 Confirmed DNS tunnels in 230B lookups.**



- **Next: information embedding examples.**
  - Query content.
  - Query timing.

Information vector: DNS query name content

Actual queries, slightly altered for privacy

- **Tunnel: SSH over Iodine (TCP/IP over DNS).**

```
0ebba82??2db??Y?wl??bb??X?Ey0bdj?gZqH??4?lNM???0?aQ
l??????db??4.???Zz???4BJ?hLv????4a??i?G.t.porcupin
e.org
```

(? = non-ASCII or non-printable octet)

- **Non-tunnel: software installer.**

```
x--00453809-004d-0046-00523-004e-0051-0034004243-0
051-0055-.00583-0051-0053-0050-0056.val.linux.10-2
0-191-136.9_5-3532-6097.sn.msgserv.ZeroG.com
```

- **Capacity: up to 255 bytes/query.**
- **59 Confirmed name-content tunnel detections.**

Information vector: DNS query name codebook

Actual queries, slightly altered for privacy

query name	type	time (UTC)
a0.twimg.com	A	1286949054.503602
a3.twimg.com	A	1286949216.242019
a3.twimg.com	A	1286949251.387366
a1.twimg.com	A	1286949277.589322
a2.twimg.com	A	1286949295.694136
a3.twimg.com	A	1286949310.772878
a1.twimg.com	A	1286949310.816623
a3.twimg.com	A	1286949418.455759
a1.twimg.com	A	1286949418.627365
a3.twimg.com	A	1286949448.813207
a0.twimg.com	A	1286949461.172023

Is this a *name* tunnel? e.g.,
00 → a0
01 → a1
10 → a2
11 → a3

- **Capacity: up to $\log^2$(codebook size) bits/query.**
- **No confirmed tunnel detections.**

Information vector: DNS query type

Actual queries, slightly altered for privacy

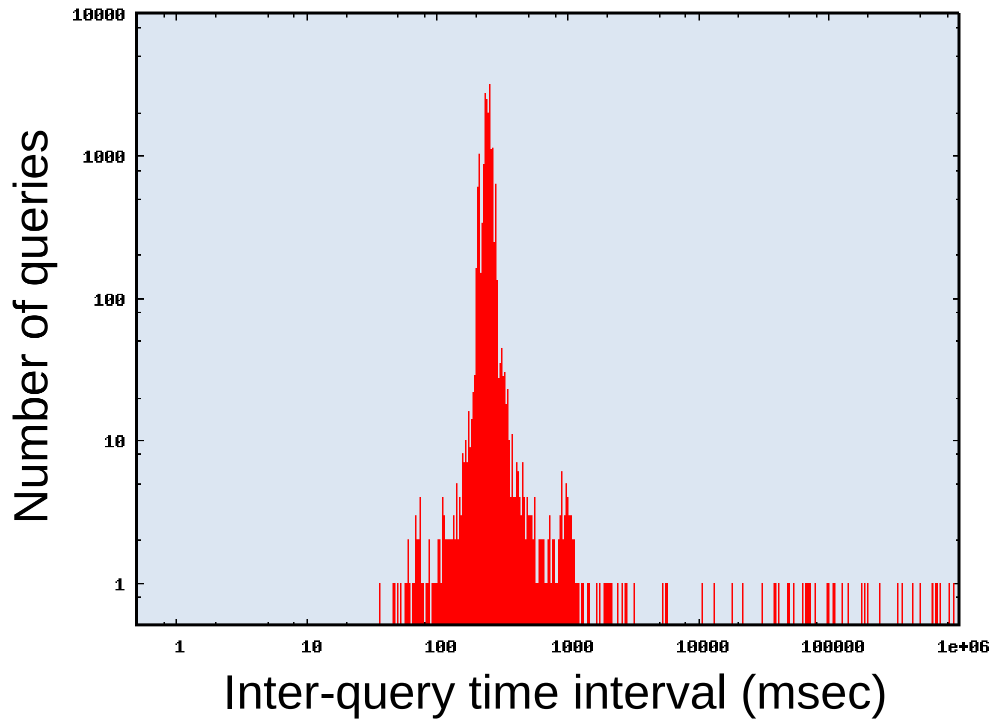
query name	**type**	time (UTC)
www.e-port.ru	**AAAA**	1363620228.803181
www.e-port.ru	**A**	1363620228.837213
www.e-port.ru	**AAAA**	1363620228.862057
www.e-port.ru	**A**	1363620228.878191
www.e-port.ru	**A**	1363620229.149720
www.e-port.ru	**AAAA**	1363620229.239968
www.e-port.ru	**A**	1363620229.269800
www.e-port.ru	**AAAA**	1363620229.319941
www.e-port.ru	**AAAA**	1363620229.377394
www.e-port.ru	**A**	1363620229.406241
www.e-port.ru	**AAAA**	1363620229.412821

Is this a *type* tunnel?
e.g.,
0 ⟶ A
1 ⟶ AAAA

- **Capacity: up to 16 bits/query (IANA defines 79 types).**
- **No confirmed tunnel detections.**

Information vector: DNS query timing
Actual queries

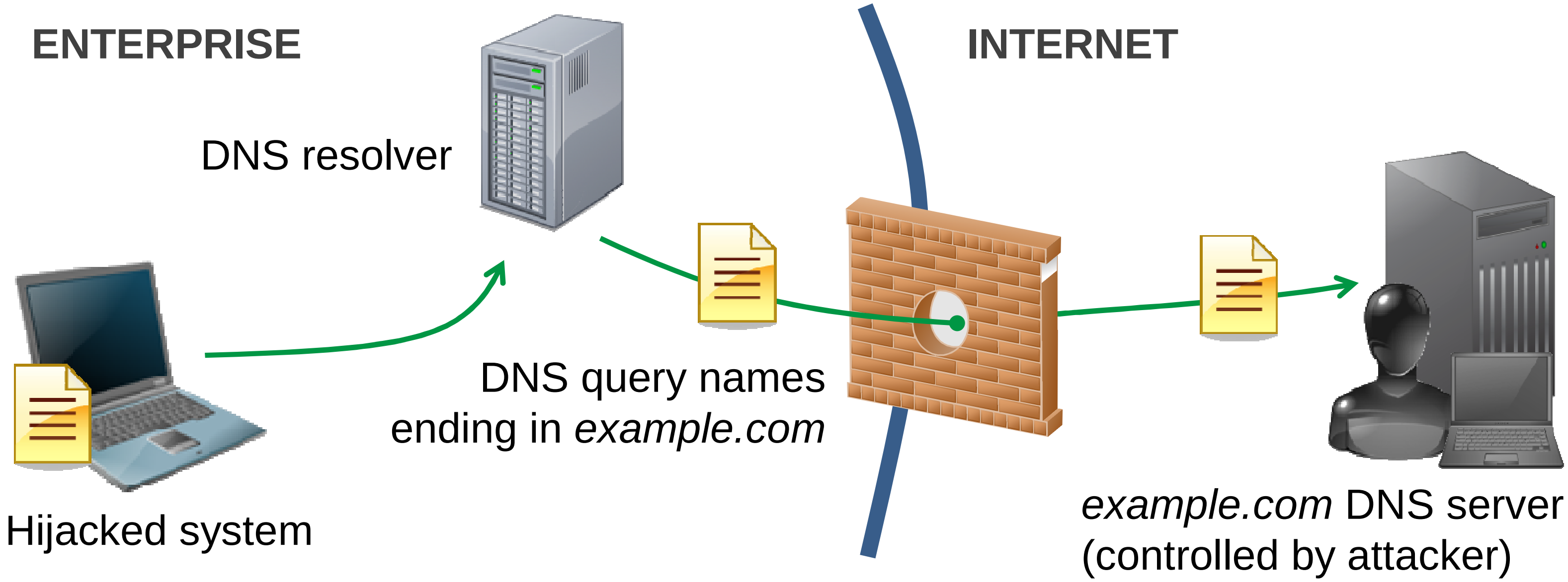
people.bu.edu query time-interval distribution



Is this a *timing* tunnel?

- **Capacity: O(100) bits/second at 10 msec resolution[1].**
- **No confirmed detections, but source of most alerts.**

[1]Conservative resolution based on median 23msec DNS timing variations observed with Netalyzr.



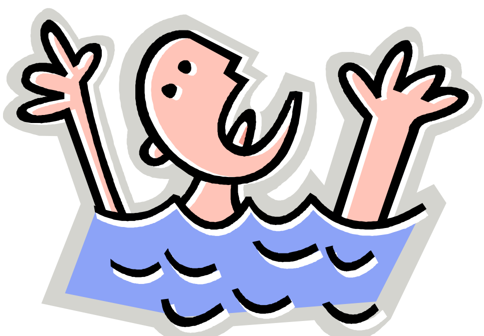
- **Next: measuring <u>all</u> information content in DNS queries.**
  - Regardless of encoding in names, types or timing.
  - First, focus on query names.

Measuring information in DNS query names, step 1

Result: 2174 alerts for IndLab dataset

“foo.example.com” + “bar.example.com” +…

- **Naive approach: concatenate all query names.**
  - Problem: too many alerts.

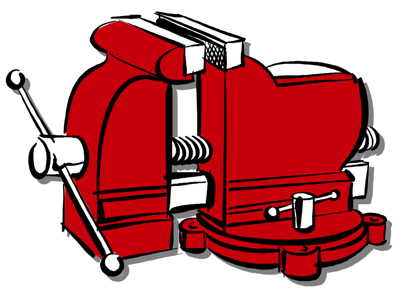


Measuring information in DNS query names, step 2

Result: 2174→145 alerts for IndLab dataset

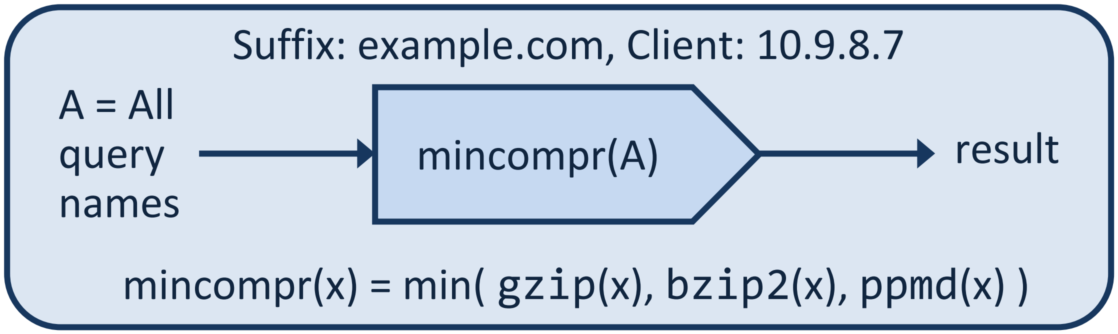
gzip(“foo.example.com..” + “bar.example.com..” +…)

- **Use lossless (reversible) data compression.**
  - Output length ≥ information content. No false negatives.
    - Insensitive to encoding details (8-bit, base64, etc.).
      - Append “**..**” to names, for reversible compression.

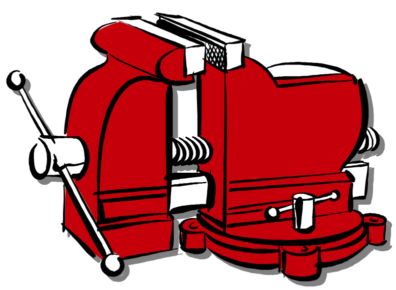


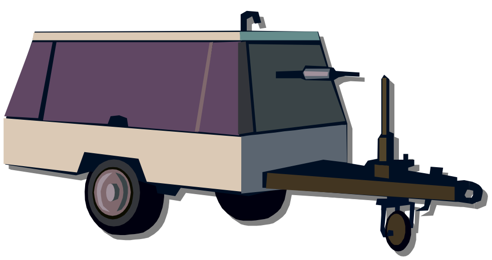
Measuring information in DNS query names, step 3

Result: 145→106 alerts for IndLab dataset



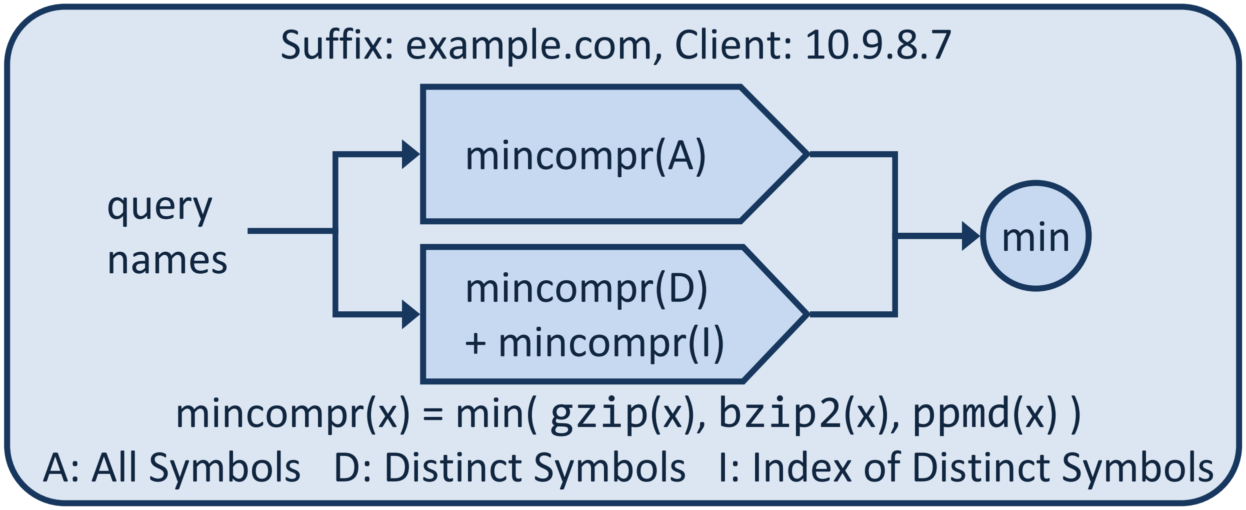
- **Use smallest result from different lossless compressors.**
  - Different compressors, different worst cases.





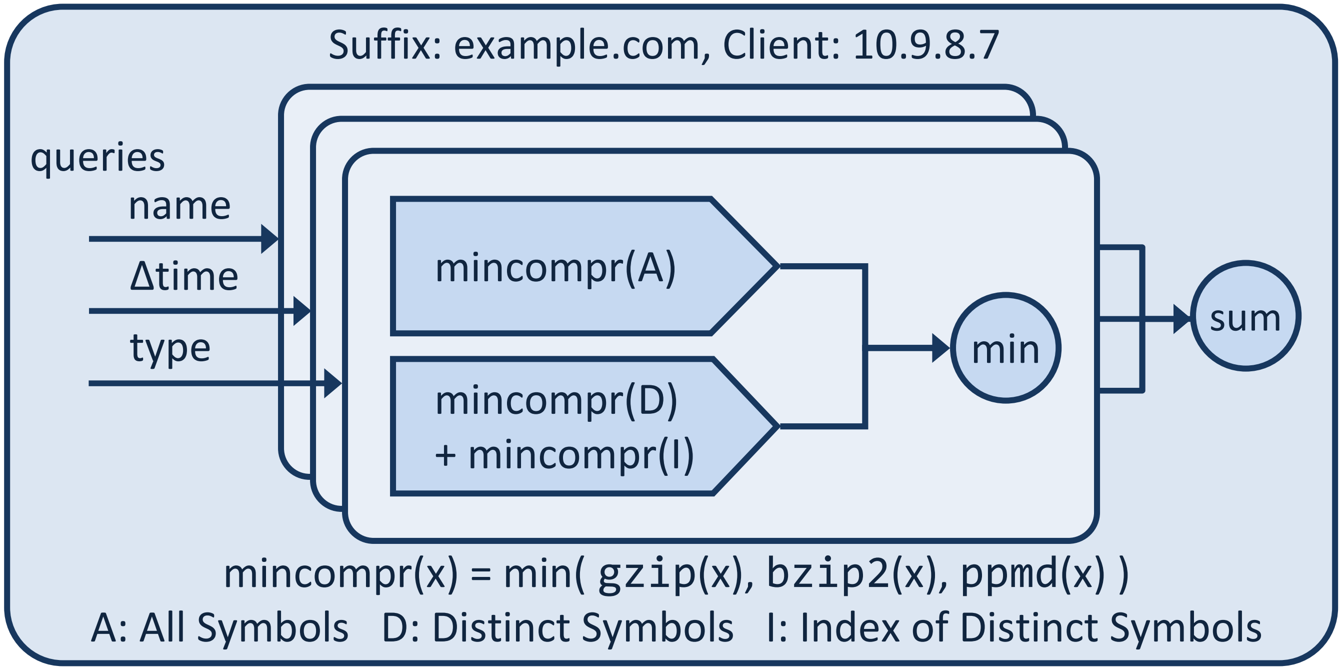
Measuring information in DNS query names, step 4

Result: 106→99 alerts for IndLab dataset



- **Use codepoints (besides straight compression):**
  - Transform names ***A*** → table of distinct names ***D*** + sequence of table indices ***I*** (codepoints). Then compress ***D*** and ***I***.
    - Exploit repetition at the granularity of entire query names.
      - Minor benefit for time-interval and query-type results (small symbols).

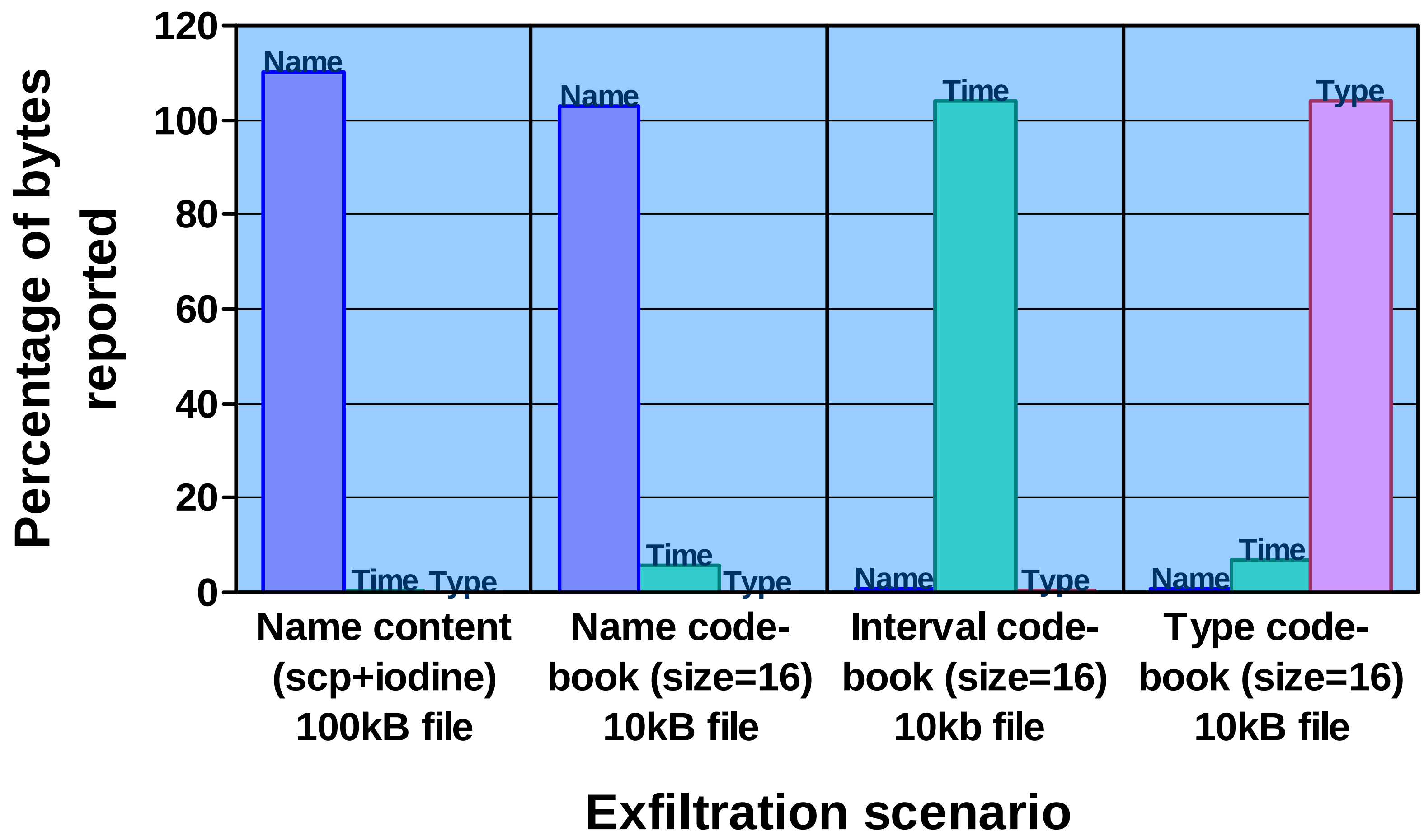
Combined DNS query information measurement

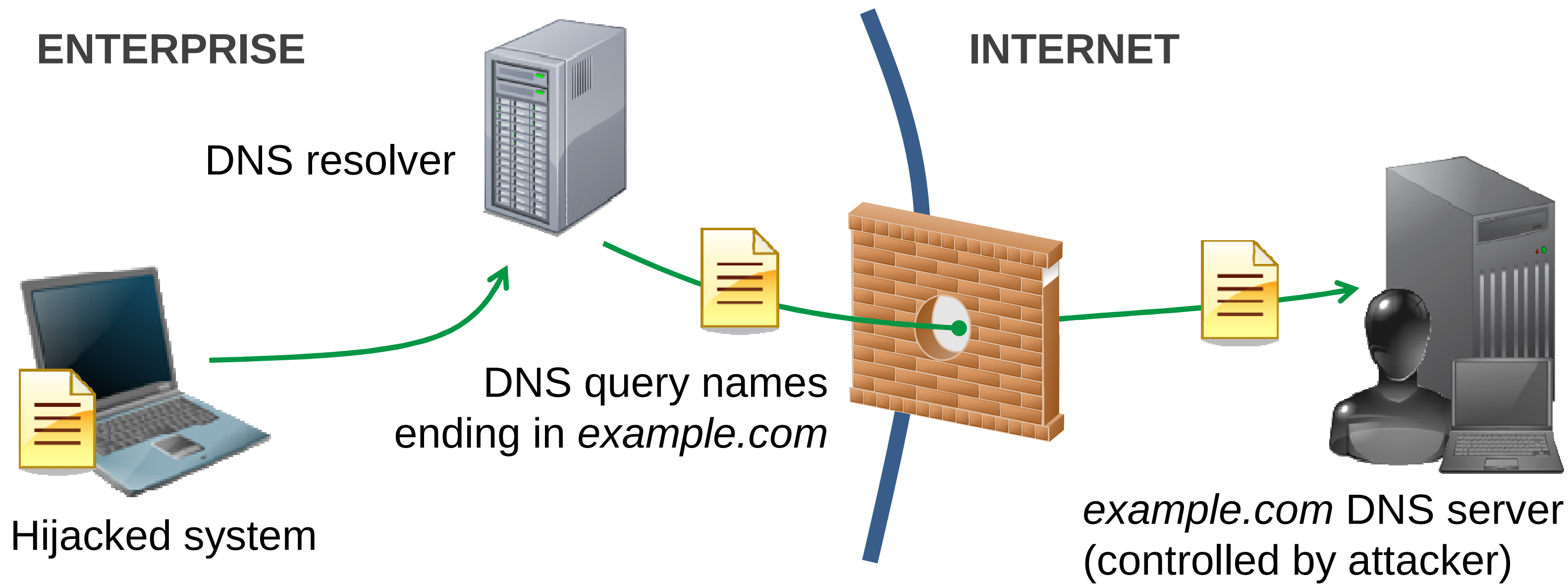


- **Use same procedure to separately measure:**
  - Sequence of query names; sequence of query types;
  - sequence of inter-query time intervals (10 msec units[1]).

[1]Conservative resolution based on median 23msec DNS timing variations observed with Netalyzr.

Validation with synthetic traffic





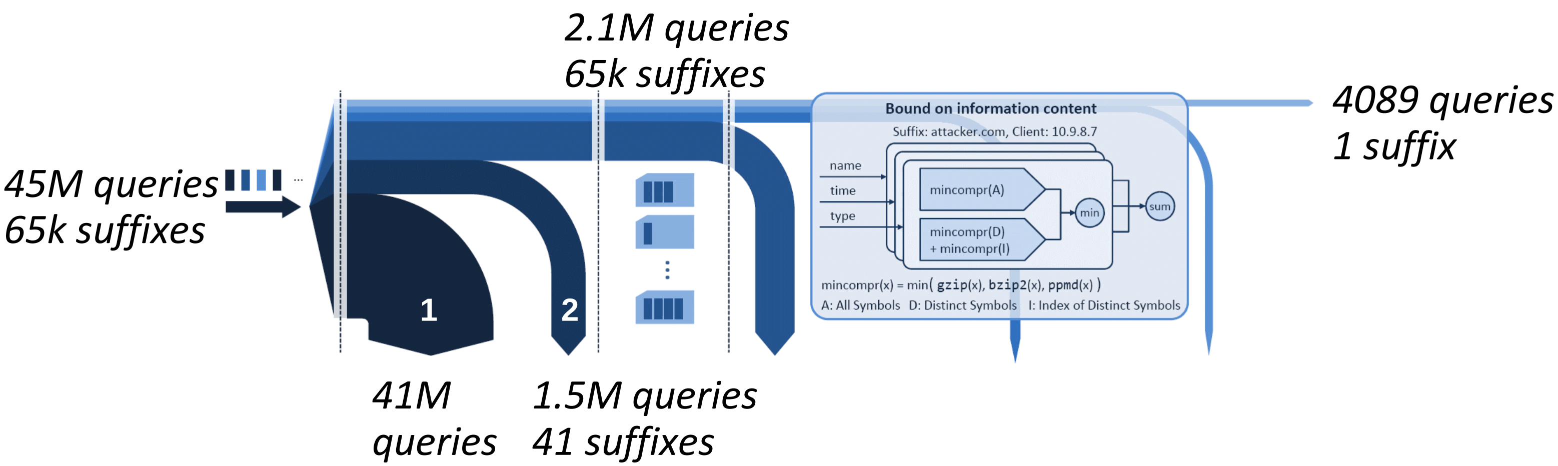
- **Next: detecting DNS tunnels in mostly-benign traffic.**
  - From 45M→4089 queries without introducing false negatives.

Searching a haystack of 230B lookups

Site	Vantage point	Clients	Days	Lookups (daily)
IndLab[1]	Internal DNS server	10k	1212	57B (47M)
LBL[2]	Internal DNS server	6.8k	2776	79B (28M)
NERSC[3]	Internal DNS server	1.3k	1642	14B (9M)
UCB campus	Network perimeter	2.1k	45	1.7B (38M)
China campus	Caching resolver	61k	5	69M (14M)
SIE[4]	Reply mirrors	123	53	77B (1.5B)

[1]Undisclosed Industrial Research Laboratory, USA.
[2]Lawrence Berkeley National Laboratory, USA.
[3]National Energy Research Supercomputing Center, USA.
[4]ISC Security Information Exchange, contributions mainly from USA and Europe.

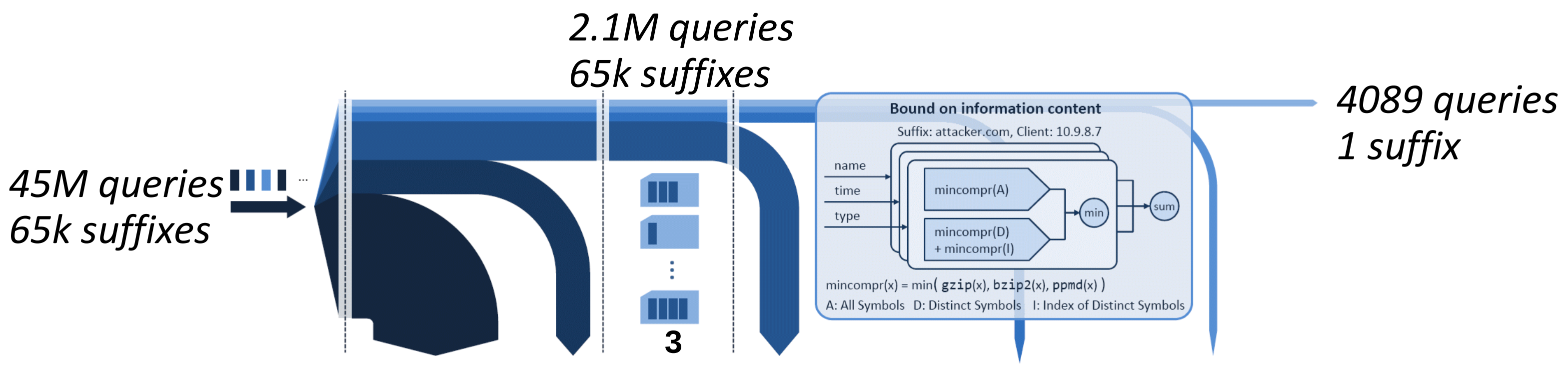
Input filters
Numbers for 1 day of IndLab traffic



- **1: Eliminate queries that hit the local DNS resolver cache.**
  - Model local DNS resolver cache (requires reply TTLs).
- **2: Eliminate “uninteresting” queries.**
  - Non-existent top-level domains (Mozilla “effective TLD” list).
  - Local/sister/reserved domains and (PTR) address ranges.

Query aggregation by client and organization

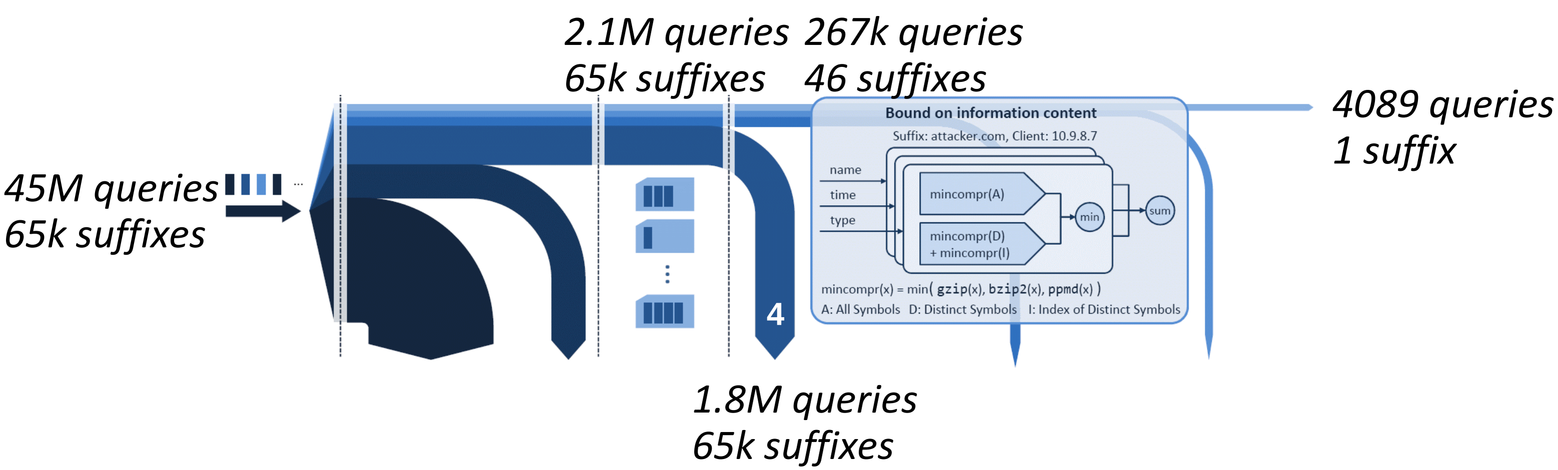
Numbers for 1 day of IndLab traffic



- **3: Aggregate queries by (client, query name suffix).**
  - 1 Query name suffix ≤ 1 organization.
    - *site*.com, *site*.co.uk (Mozilla “effective TLD” list).
    - *in-addr.arpa* at /16 and /24 boundaries, *ip6.arpa* at /48.

Quick information estimate before data compression

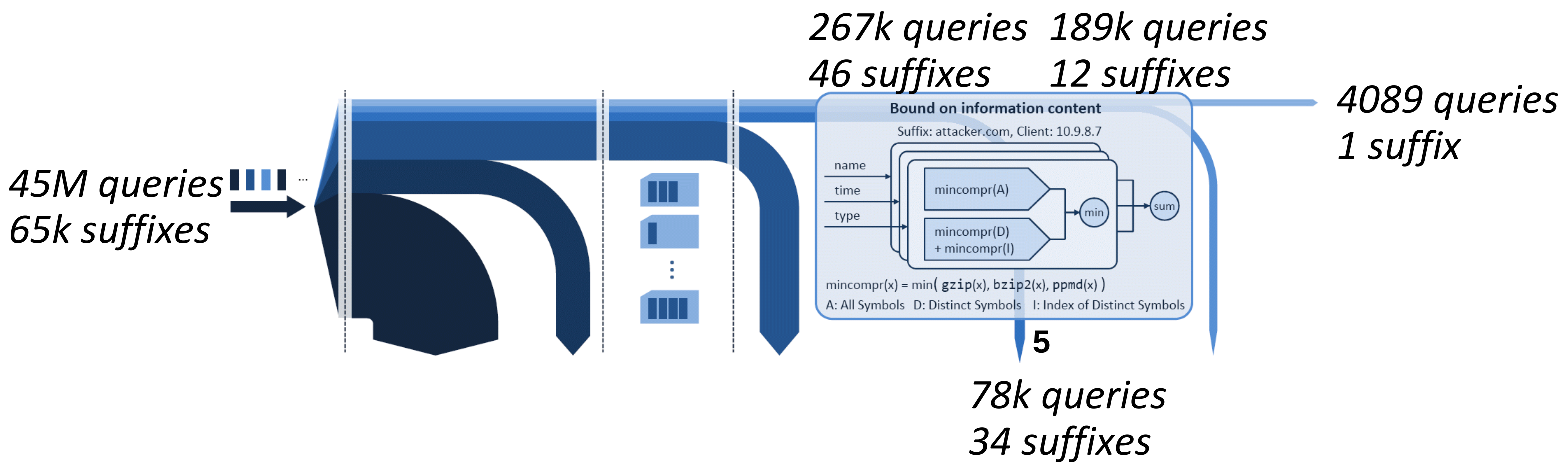
Numbers for 1 day of IndLab traffic



- **4: Eliminate (client, name suffix) based on per-day totals.**
  - Worst-case Shannon entropies: assume uniform distributions over distinct inter-query time intervals, names, and types.
  - Plus length of distinct query names and types.

Precise information measurement

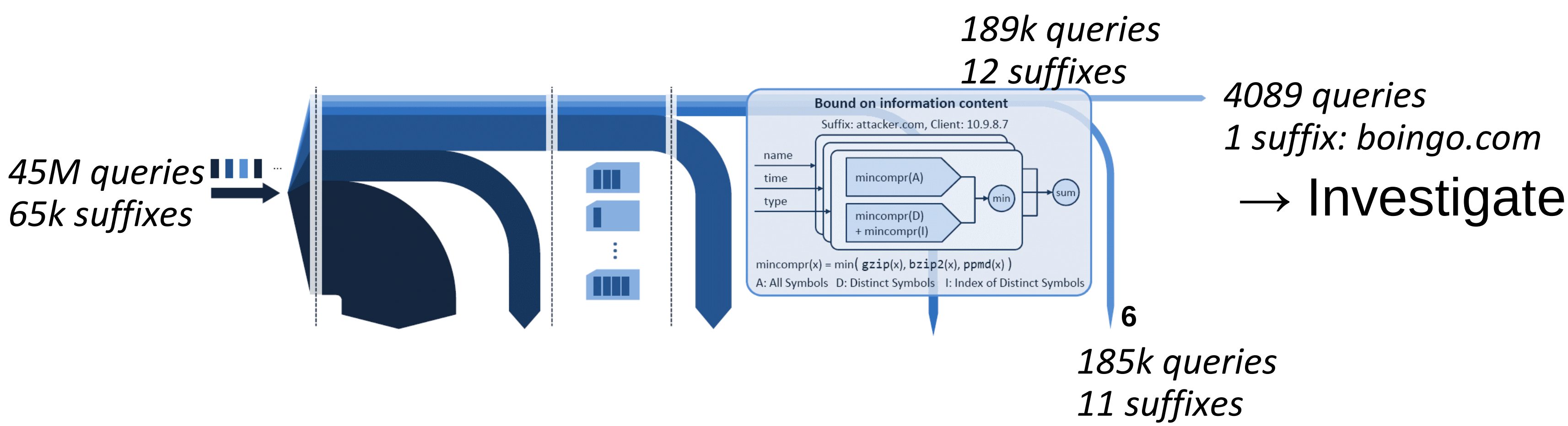
Numbers for 1 day of IndLab traffic



- **5: Eliminate (client, suffix) with too little information.**
  - Compressor and codepoint bakeoff.
  - 4 kB bound for targeted environments (individual clients).
    - 10 kB bound for aggregated query streams.

Inspected domains list

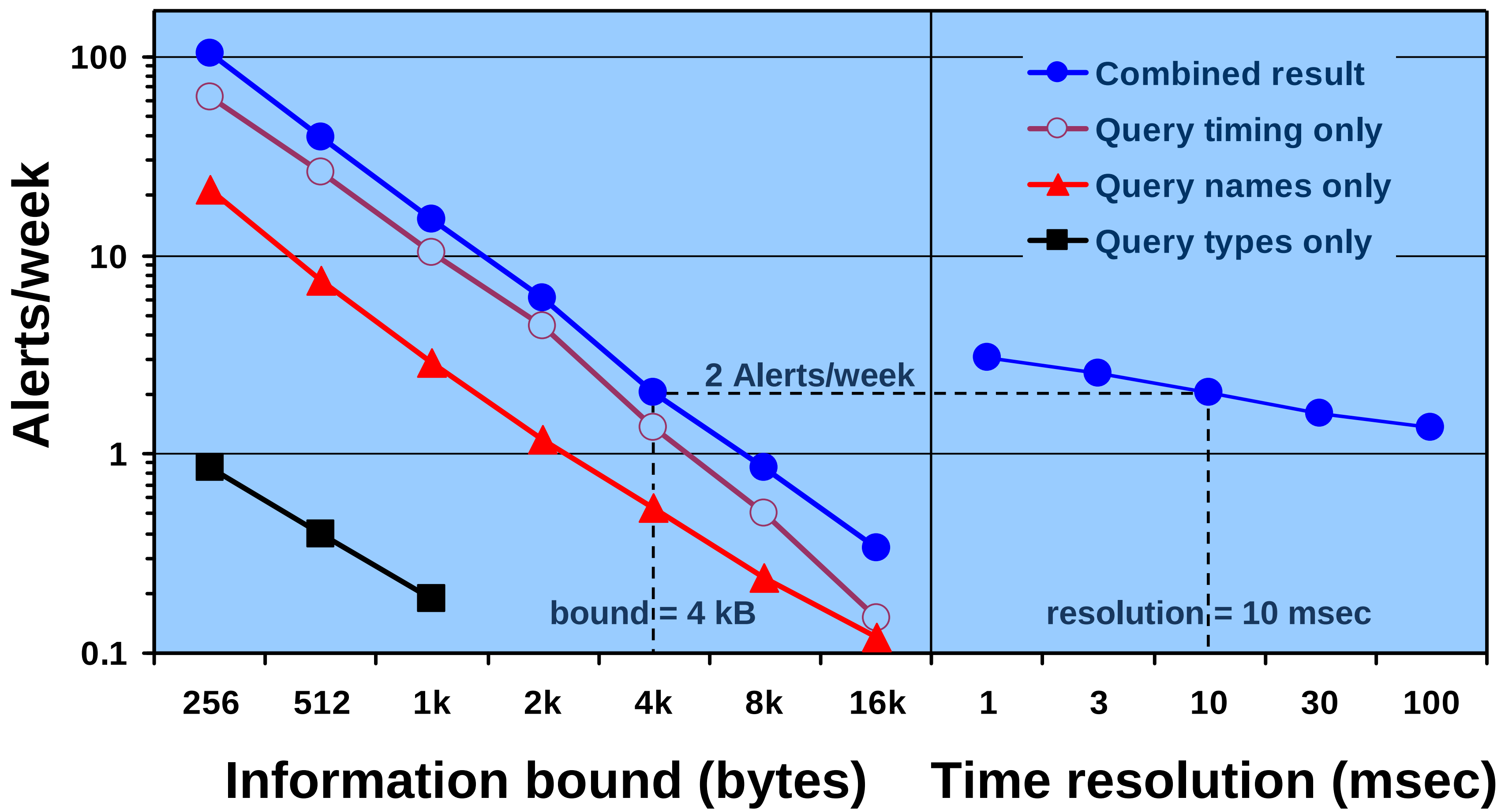
Numbers for 1 day of IndLab traffic



- **6: Eliminate already-inspected domains.**
  - Each flagged query name suffix is inspected only once.
    - If a benign domain becomes malicious after inspection:
      - It is a major site (Google, etc.) → we have worse problems.
      - It keeps mimicking benign behavior → it remains undetected.
      - It exposes itself to signature-based detection.

Alert rate sensitivity to parameter settings

IndLab data set, 1212 days



Detection breakdown

Dominant source	Individual clients			Aggregates		
Site	IndLab	LBL	NERSC	UCB	China	SIE
Lookups (days)	57B(1212)	79B(2776)	14B(1642)	1.7B(45)	69M(5)	77B(53)
Information bound	4kB	4kB	4kB	10kB	10kB	10kB
Confirmed tunnel	0	2	0	0	0	57
Benign	286	306	29	200	41	4815
Malware	2	2	0	5	2	74
Misconfiguration	49	62	5	126	8	310
IPv4 PTR	11	29	4	26	3	N/A
IPv6 PTR	0	5	0	1	0	N/A
Unknown	14	27	0	13	13	1
Total alerts	362	433	38	371	67	5257
Alerts/typ. week	2.0	1.1	0.15	32	N/A	358

Conclusion

- **Novel procedure to measure the information content of DNS query streams.**
- **1-2 Alerts/week for enterprise-scale networks.**
  - 4 kbytes/day threshold per local client and remote domain.
  - Lossless compression, no false negatives.
- **59 Confirmed DNS tunnels in 230B queries.**
  - All conventional name-content based.