RIPE NCC
RIPE NETWORK COORDINATION CENTRE

# A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists

Quirin Scheitle, Technical University of Munich
Oliver Hohlfeld, RWTH Aachen University
Julien Gamba, IMDEA Networks Institute/Universidad Carlos III de Madrid
Jonas Jelten, Technical University of Munich
Torsten Zimmermann, RWTH Aachen University
**Stephen D. Strowes**, RIPE NCC
Narseo Vallina-Rodriguez, IMDEA Networks Institute/ICSI

Internet Measurement Conference 2018

Top Lists are commonly used by our community

Top Lists are commonly used by our community

We wanted to enumerate how common this actually is,
and to understand what people are measuring

So we surveyed 687 papers from top conferences in 2017

So we surveyed 687 papers from top conferences in 2017

IMC, PAM, TMA, USENIX Security,
IEEE S&P (Oakland), ACM CCS, NDSS,
CoNEXT, SIGCOMM, WWW

| | Accepted | using list | | subset? | |
|---|---|---|---|---|---|
| Venue | papers | # | % | 1M | <1M |
| IMC | 42 | | | | |
| TMA | 19 | | | | |
| PAM | 20 | | | | |

|  | Accepted | using list |  | subset? |  |
|---|---|---|---|---|---|
| Venue | papers | # | % | 1M | <1M |
| IMC | 42 | 11 | **26.2%** |  |  |
| TMA | 19 | 4 | 21.1% |  |  |
| PAM | 20 | 4 | 20.0% |  |  |

|  | Accepted | using list |  | subset? |  |
|---|---|---|---|---|---|
| Venue | papers | # | % | 1M | <1M |
| IMC | 42 | 11 | **26.2%** | 7 | 5 |
| TMA | 19 | 4 | 21.1% | 2 | 2 |
| PAM | 20 | 4 | 20.0% | 0 | 4 |

**In all, 10% of papers we surveyed used a top list**

# Significance, Structure, and Stability of Internet Top Lists

# Significance, Structure, and Stability of Internet Top Lists

**0: How are top lists being used? (significance)**

# Significance, Structure, and Stability of Internet Top Lists

**0: How are top lists being used? (significance)**

**1: What goes into top lists? (structure)**

# Significance, Structure, and Stability of Internet Top Lists

**0: How are top lists being used? (significance)**

**1: What goes into top lists? (structure)**

**2: How stable are top lists? Can they be influenced? (stability)**

# Significance, Structure, and Stability of Internet Top Lists

**0: How are top lists being used? (significance)**

**1: What goes into top lists? (structure)**

**2: How stable are top lists? Can they be influenced? (stability)**

**3: How do they impact our research results?**

To answer these questions, we took a systematic look at three lists:

- Alexa top 1M (“temporarily available” in 2017[1]; now semi-public)
- Cisco Umbrella 1 Million
- the Majestic Million

[1] https://twitter.com/Alexa_Support/status/801167423726489600

# TOP-LIST STRUCTURE

## a.k.a., what gets into these lists?

# Top-list Structure

We wanted to understand:

- subdomain depth
- base-domain coverage
- TLD coverage
- and, how the lists intersect

# Top-list Structure: subdomain name depth

Alexa and Majestic: primarily base domains, with exceptions

- *.blogspot.com
- *.global.ssl.fastly.com

Umbrella doesn’t truncate; examples:

- 2.tlu.dl.delivery.mp.microsoft.com.edgesuite.net.globalredir.akadns.net
- a.a.a.a.a.a.a.a.a.a.a.a.a.a.a.a.a.a.a.a.a.a.a.3.id.ctoid.net

# Top-list Structure: base domains and TLDs

| List | # Base domains | # TLDs |
|---:|:---:|:---:|
| Alexa | ≈ 972k | ≈ 760 |
| Umbrella | ≈ 273k | ≈ 580 |
| Majestic | ≈ 994k | ≈ 698 |

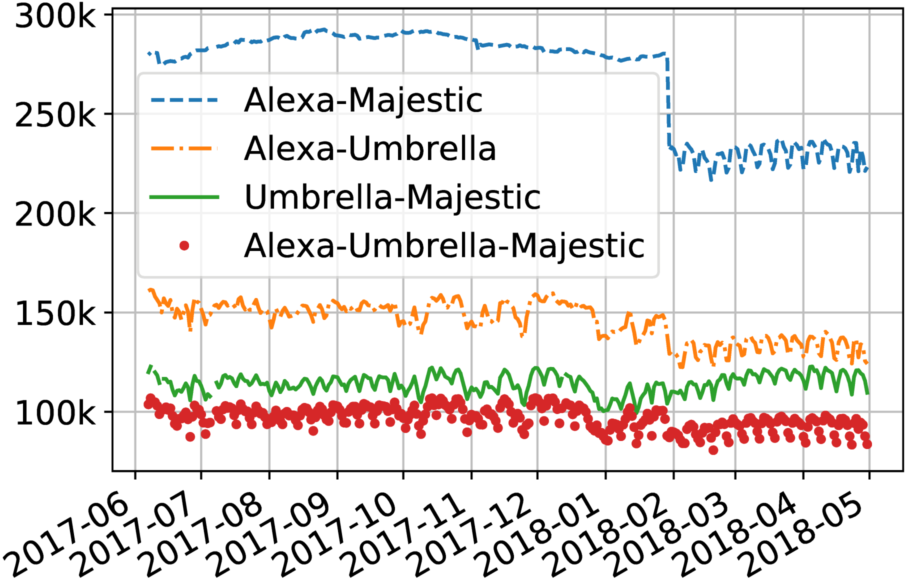
# Top-list Structure: base domains and TLDs

| List | # Base domains | # TLDs |
|---:|:---:|:---:|
| Alexa | ≈ 972k | ≈ 760 |
| Umbrella | ≈ 273k | ≈ 580 |
| Majestic | ≈ 994k | ≈ 698 |

**Cisco Umbrella emphasises depth, the others, breadth.**

**Top lists miss >50% of the active set of TLDs.**

# Top-list Structure: intersection



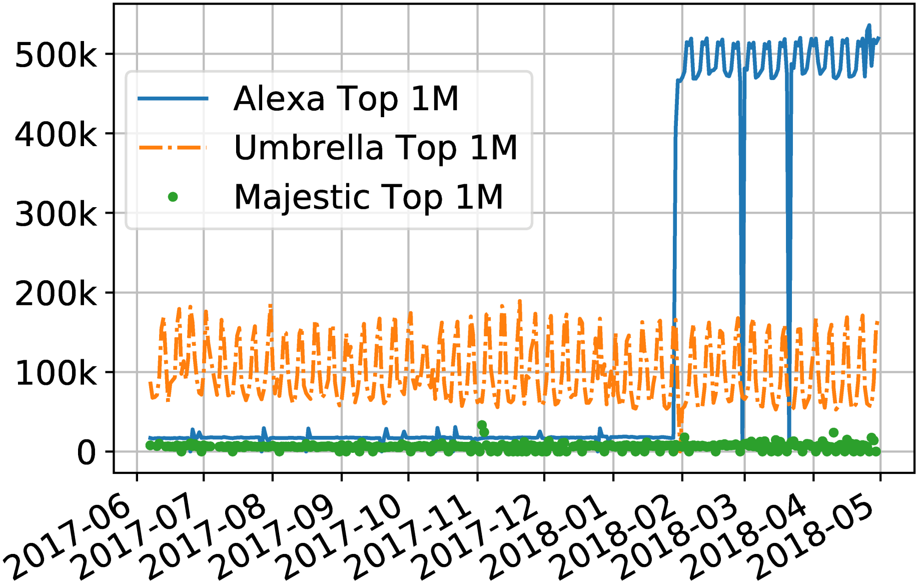
# Top-list Structure

- Domain/subdomain/TLD coverage is not consistent
- The intersection of base domains between these lists is remarkably low

# TOP-LIST STABILITY

a.k.a., how stable are the lists; can they be influenced?

# Stability: daily changes

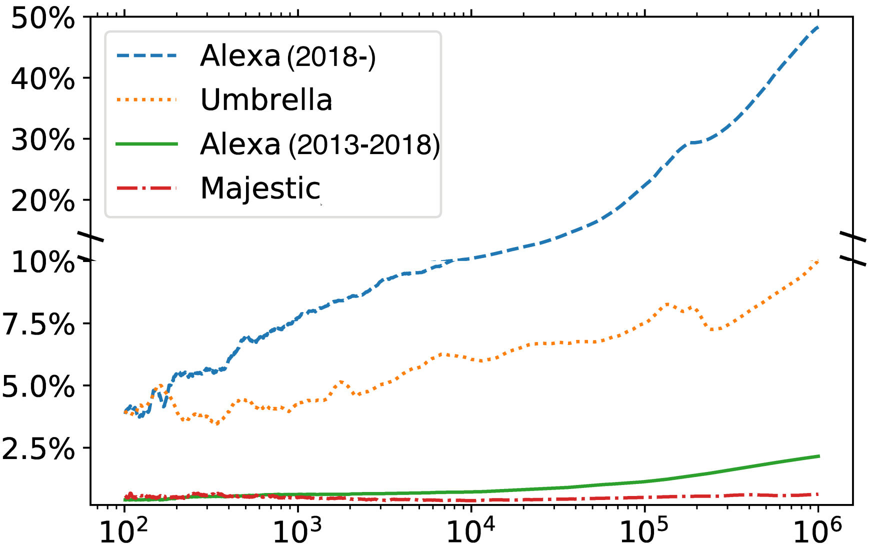


# Stability: daily changes

**Top lists can undergo rapid and unannounced changes**

**Alexa is now the most unstable list of these lists**

# Stability: rank volatility and the long tail



# Stability

Stability varies considerably

- lower ranks are more volatile than higher ranks, especially in Alexa, Umbrella
- enumerating this may affect your experimental design

In the paper, we also look at weekly patterns, and order stability

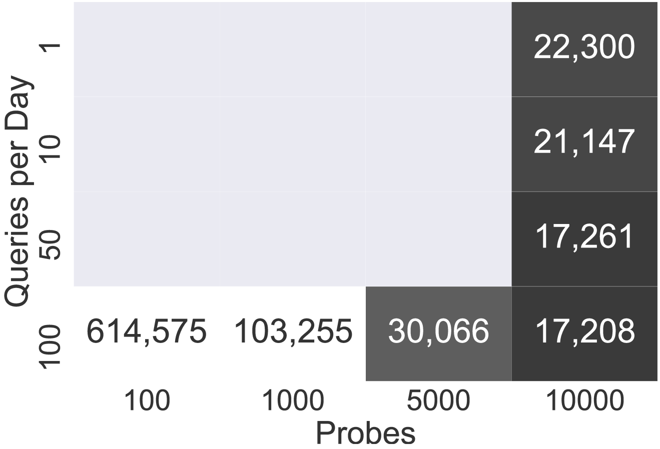
# INFLUENCING LISTS

## a.k.a., how easily can we add domains to Umbrella?

a.k.a., how to make lists and influence people

# Influencing lists: “hacking” Umbrella

It is remarkably easy to promote domains far up the Umbrella ranking:



# IMPACT ON RESEARCH

a.k.a., how do lists potentially affect measurement results?

# Understanding potential impact on research

We ran measurements against the names in these lists

For example: IPv6, CAA, TLS, HSTS, HTTP/2, CDN coverage

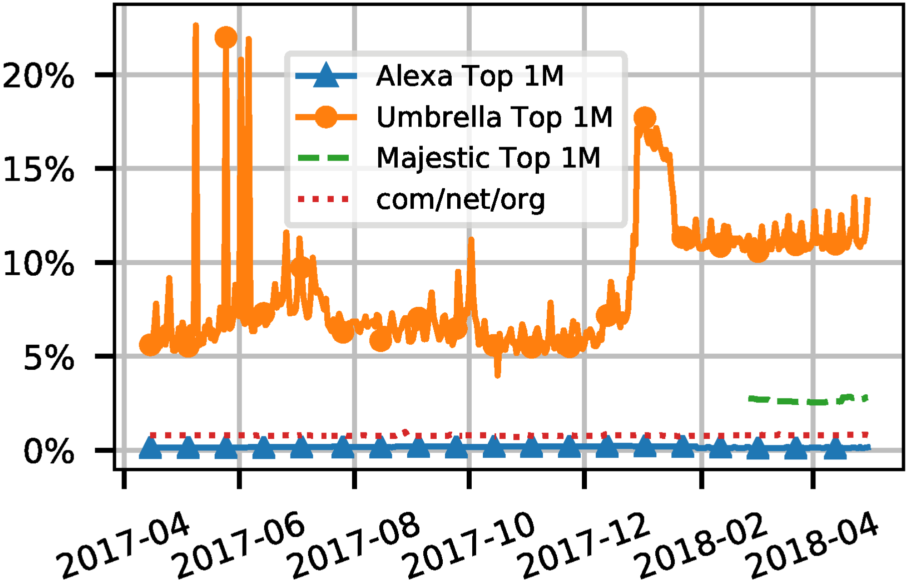
In each case, the list chosen gives different results

# Understanding potential impact on research

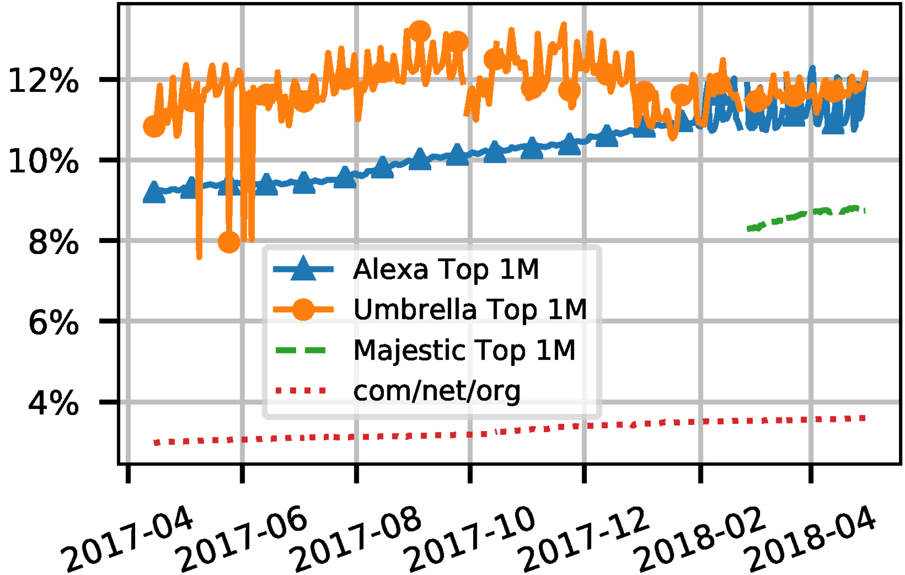
Here, I'll discuss: NXDOMAINs, IPv6, and HTTP/2.

All measurements conducted on the day the list was fetched.

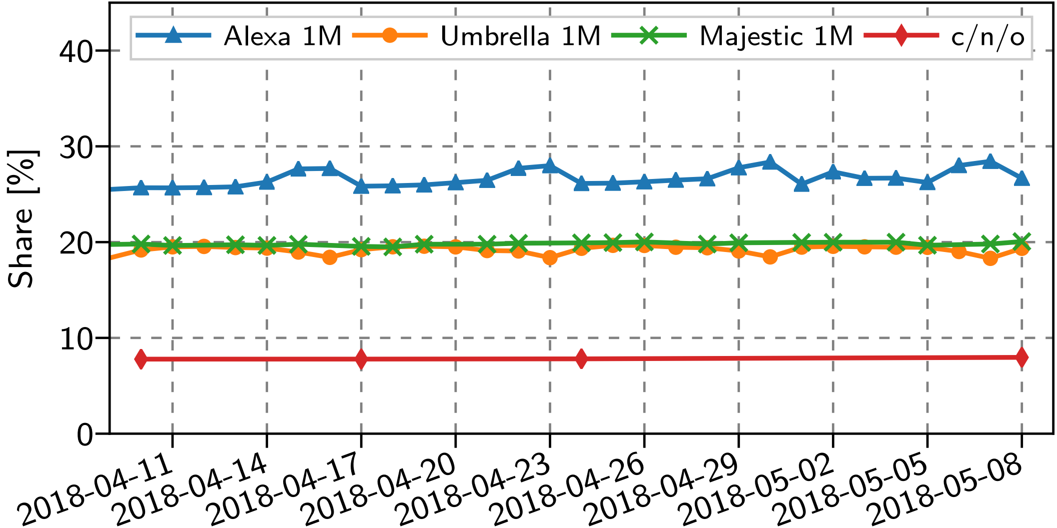
# Impact: example 1, NXDOMAINs



# Impact: example 2, IPv6 adoption



# Impact: example 3, HTTP/2



# Understanding potential impact on research

**Studies frequently do not mention *when* they retrieved a top list, and/or when they ran measurements against that list.**

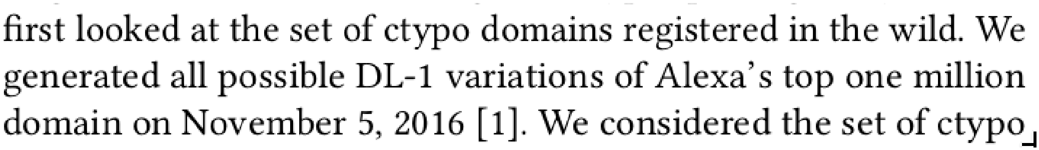
| | Accepted | using list | | dates? | |
|---|---|---|---|---|---|
| Venue | papers | # | % | fetch | study |
| IMC | 42 | 11 | 26.2% | | |
| TMA | 19 | 4 | 21.1% | | |
| PAM | 20 | 4 | 20.0% | | |

# Understanding potential impact on research

**Studies frequently do not mention *when* they retrieved a top list, and/or when they ran measurements against that list.**

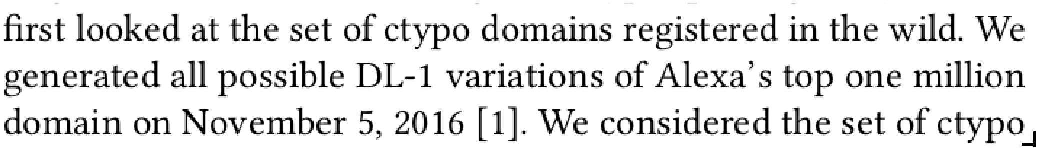
| | Accepted | using list | | dates? | |
|---|---|---|---|---|---|
| Venue | papers | # | % | fetch | study |
| IMC | 42 | 11 | 26.2% | 1 | 3 |
| TMA | 19 | 4 | 21.1% | 0 | 0 |
| PAM | 20 | 4 | 20.0% | 0 | 0 |

## More of this please?

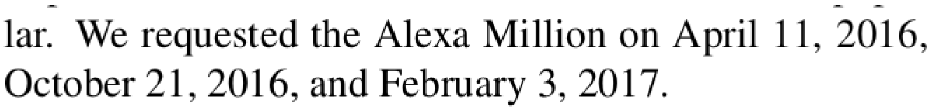


first looked at the set of ctypo domains registered in the wild. We generated all possible DL-1 variations of Alexa's top one million domain on November 5, 2016 [1]. We considered the set of ctypo

# More of this please?



first looked at the set of ctypo domains registered in the wild. We generated all possible DL-1 variations of Alexa's top one million domain on November 5, 2016 [1]. We considered the set of ctypo



lar. We requested the Alexa Million on April 11, 2016, October 21, 2016, and February 3, 2017.

# Understanding potential impact on research

In summary:

1. choice of top list will affect measurement results
2. day the list was fetched will affect measurement results
3. ... but authors frequently don’t tell us the important dates

# List/Study Considerations

# List/Study Considerations

# List/Study Considerations

**These are disparate sources with proprietary behaviours that use differing methods.**

We ought to be careful with how we use them.

- Consider the contents of the list: Alexa clearly not always best
- Consider temporal aspects: longitudinal study may be appropriate
- State the date that list was fetched, and when dates it was used for measurements

# Conclusions

# Conclusions

We have shown that:

- There is significant churn in Alexa, Umbrella
- The choice of top list, or even day of week, can clearly affect the result of measurement studies
- The Alexa list changed its behaviour significantly in January 2018
- Domains can be trivially inserted into the Cisco Umbrella list

And finally, we hope we encourage deeper rationale in top list usage in the future.

# Thank you!

Questions?

`sds@ripe.net` / `@sdstrowes`

You can find code and data at:

`https://toplists.github.io/`