A Long Way to the Top

Significance, Structure, and Stability of Internet Top Lists
by Quirin Scheitle; Oliver Hohlfeld; Julien Gamba; Jonas Jelten; Torsten Zimmermann; D. Stephen Strowes; Narseo Vallina-Rodriguez

Presented by: Christian Wahl, David Hasselquist, Otto Bergdal

Why are these top lists important?

Build base for broad research

- Internet Measurements
- Privacy
- Network Security
- And many more

Seldomly questioned by community

- Might lead to bias in research
- Might not be stable
- Might not reflect the real usage
- Affects the reproducibility

The lists – Alexa 1M

- Data from over 25,000 browser extensions (including Alexa Toolbar)
  - Which are “used by ’millions of people’”
- Often used in network research as a ground for research
- License: Commercial

The lists – Majestic Million

- Uses a web crawler
- Ranks pages by IPv4 /24 subnets linking to this site
- License: Creative Commons

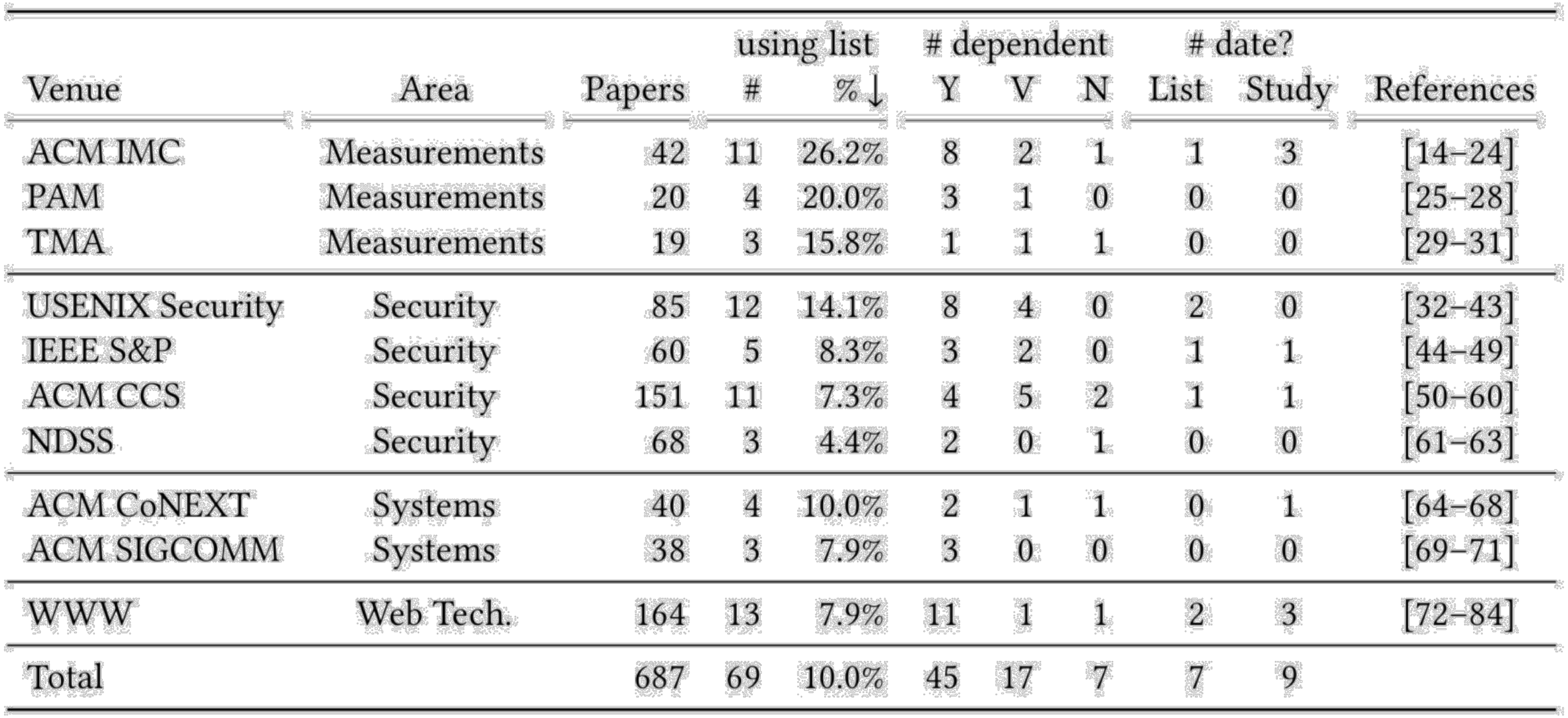
The lists – Cisco Umbrella 1M

- OpenDNS lookups per domain
- Does not check the validity of domain names
- Subdomains included (up to level 33)
- License: Not specified
  (but free of charge at the moment)

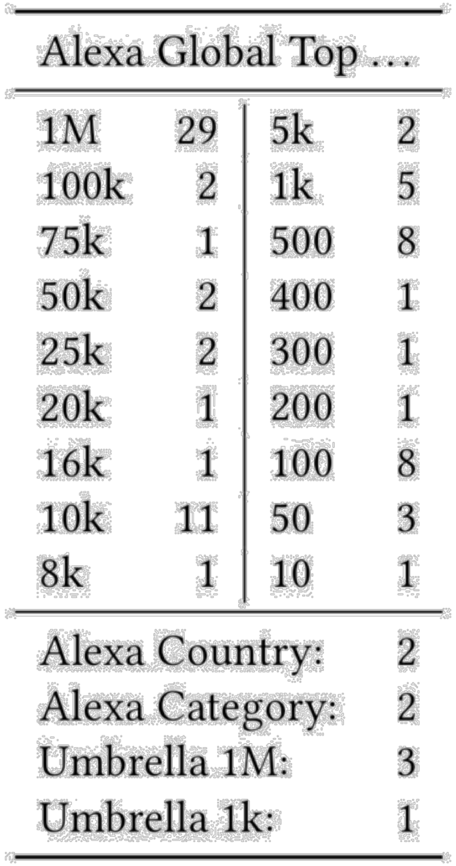
Published papers using top lists

- Survey of 687 networking-related papers published in 2017
- 69 papers (10.0%) make use of at least one top list
- Field of Internet measurement 22.2%
- Security (8.5%)
- Systems (6.4%)
- Web technology (7.9%)

Published papers using top lists

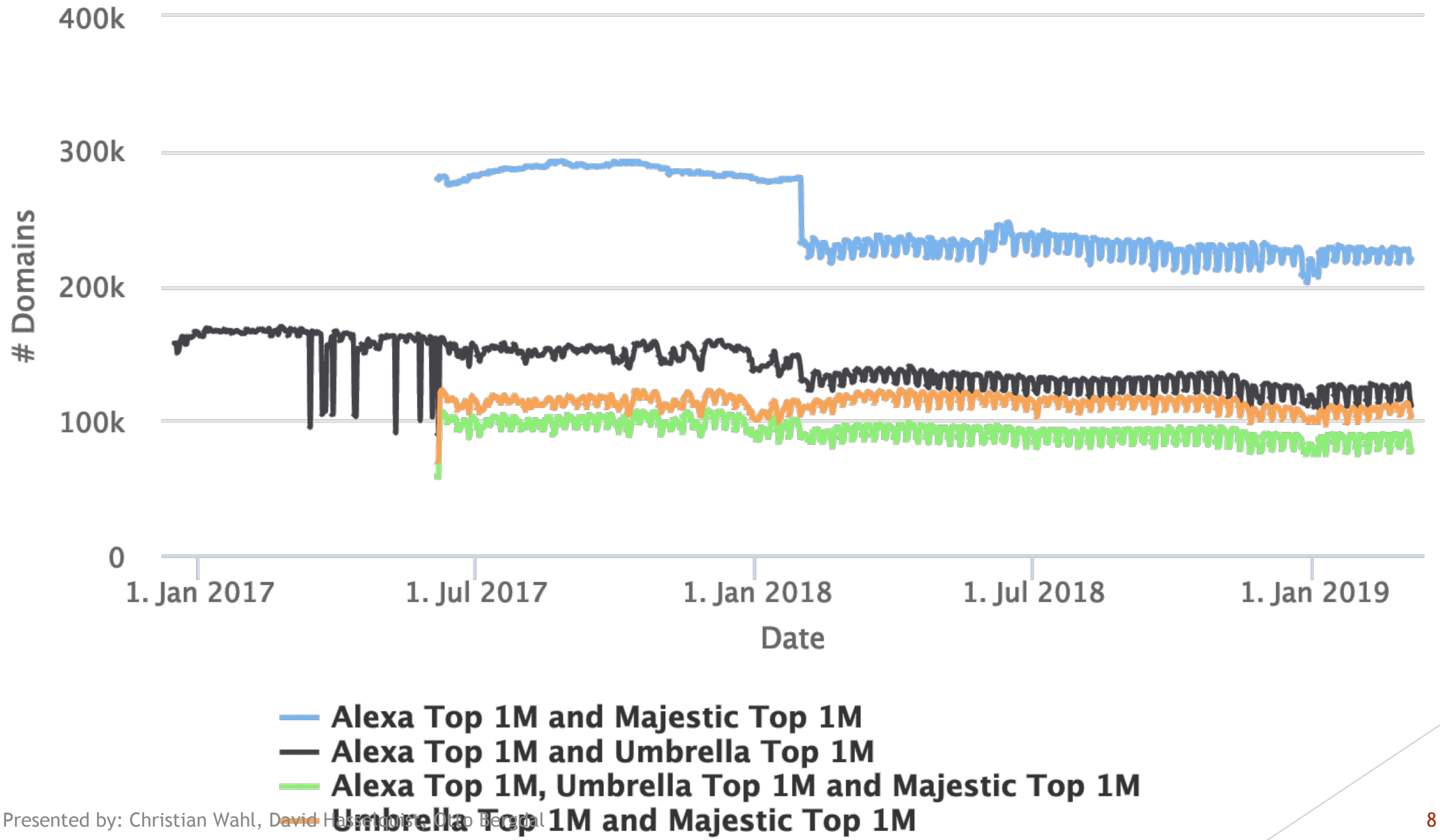


			using list		# dependent			# date?		
Venue	Area	Papers	#	% ↓	Y	V	N	List	Study	References
ACM IMC	Measurements	42	11	26.2%	8	2	1	1	3	[14–24]
PAM	Measurements	20	4	20.0%	3	1	0	0	0	[25–28]
TMA	Measurements	19	3	15.8%	1	1	1	0	0	[29–31]
USENIX Security	Security	85	12	14.1%	8	4	0	2	0	[32–43]
IEEE S&P	Security	60	5	8.3%	3	2	0	1	1	[44–49]
ACM CCS	Security	151	11	7.3%	4	5	2	1	1	[50–60]
NDSS	Security	68	3	4.4%	2	0	1	0	0	[61–63]
ACM CoNEXT	Systems	40	4	10.0%	2	1	1	0	1	[64–68]
ACM SIGCOMM	Systems	38	3	7.9%	3	0	0	0	0	[69–71]
WWW	Web Tech.	164	13	7.9%	11	1	1	2	3	[72–84]
Total		687	69	10.0%	45	17	7	7	9	

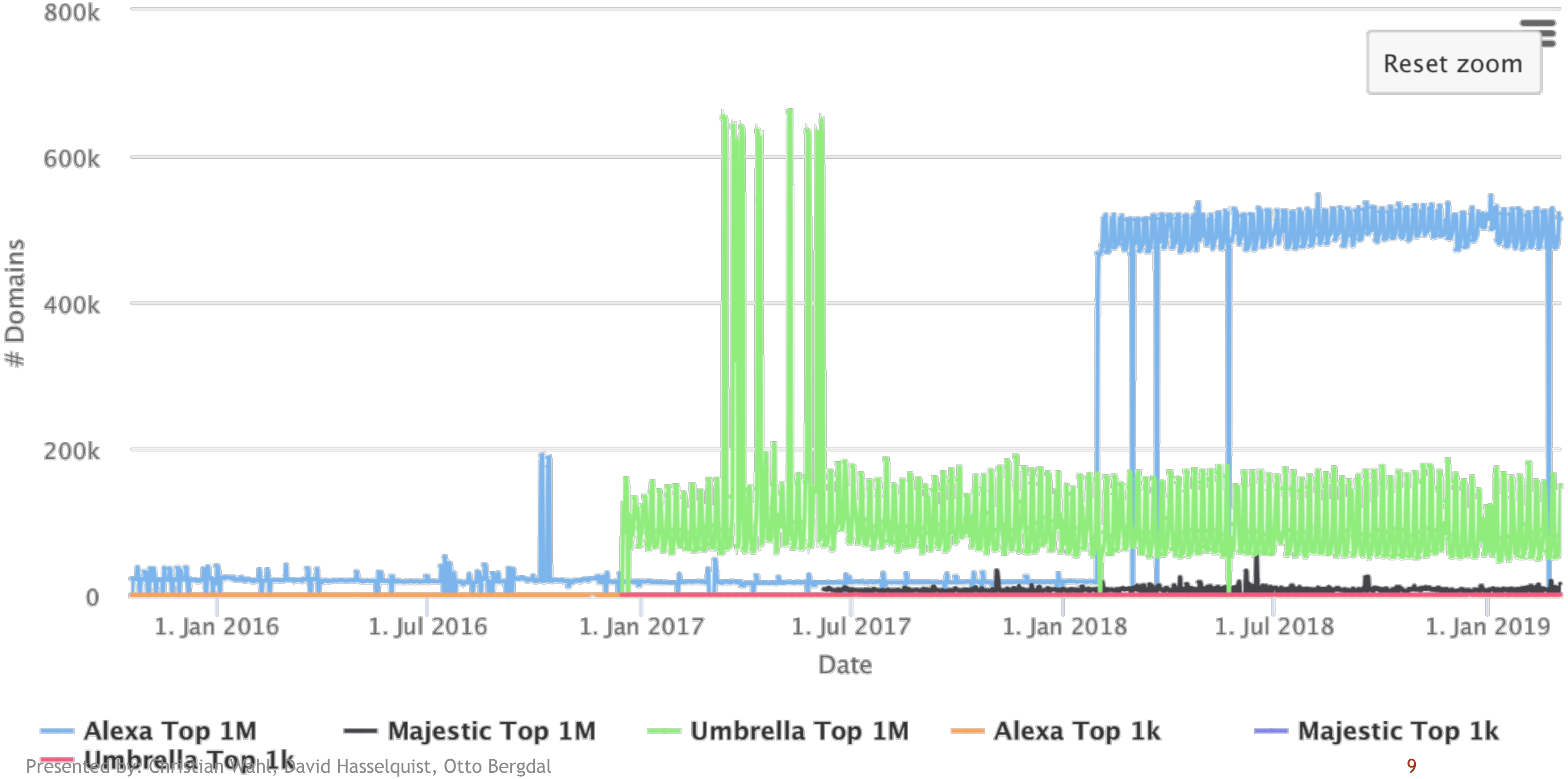


Alexa Global Top …			
1M	29	5k	2
100k	2	1k	5
75k	1	500	8
50k	2	400	1
25k	2	300	1
20k	1	200	1
16k	1	100	8
10k	11	50	3
8k	1	10	1
Alexa Country:			2
Alexa Category:			2
Umbrella 1M:			3
Umbrella 1k:			1

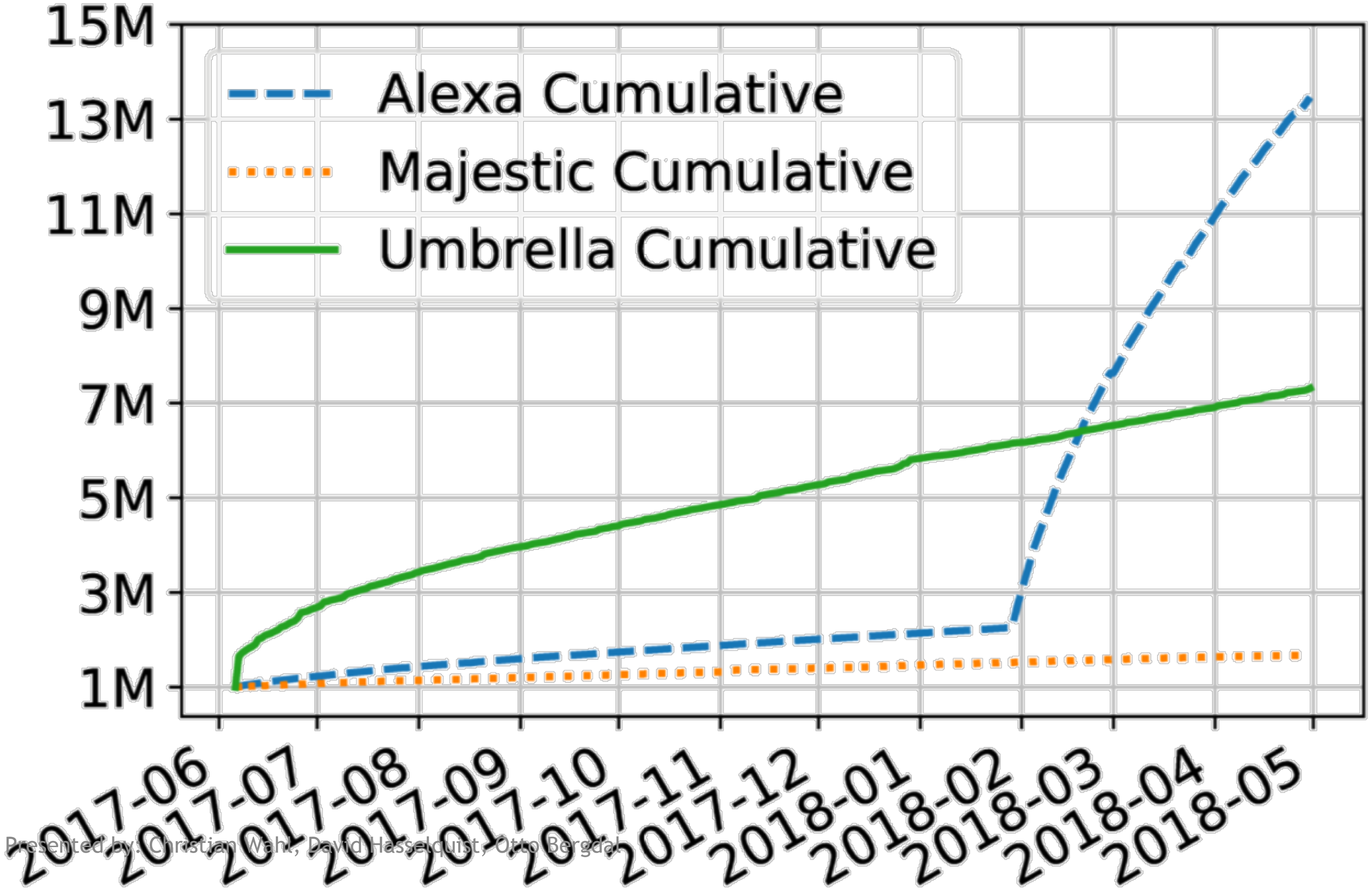
Intersection between Top 1M lists



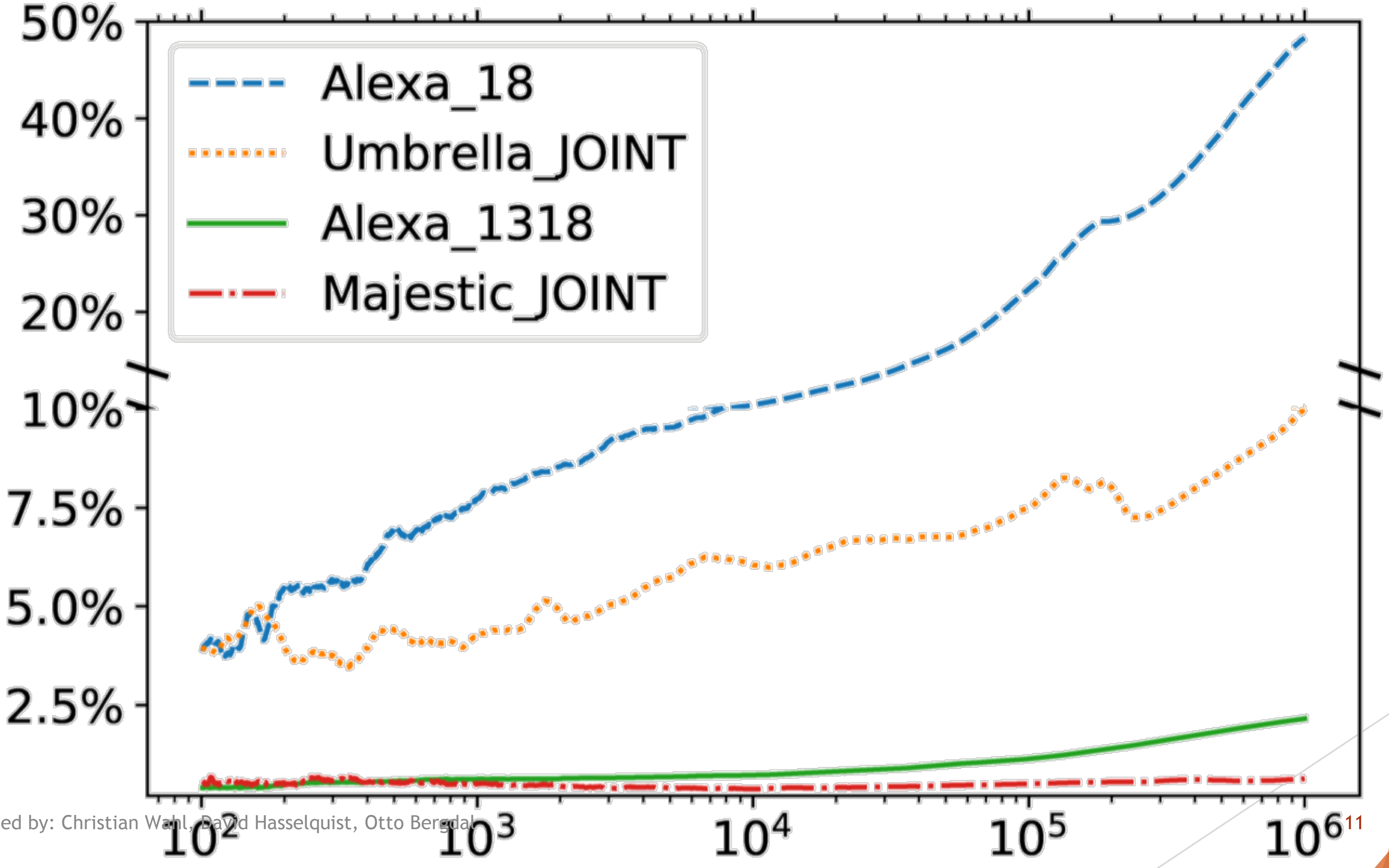
Daily changes of Top 1M entries



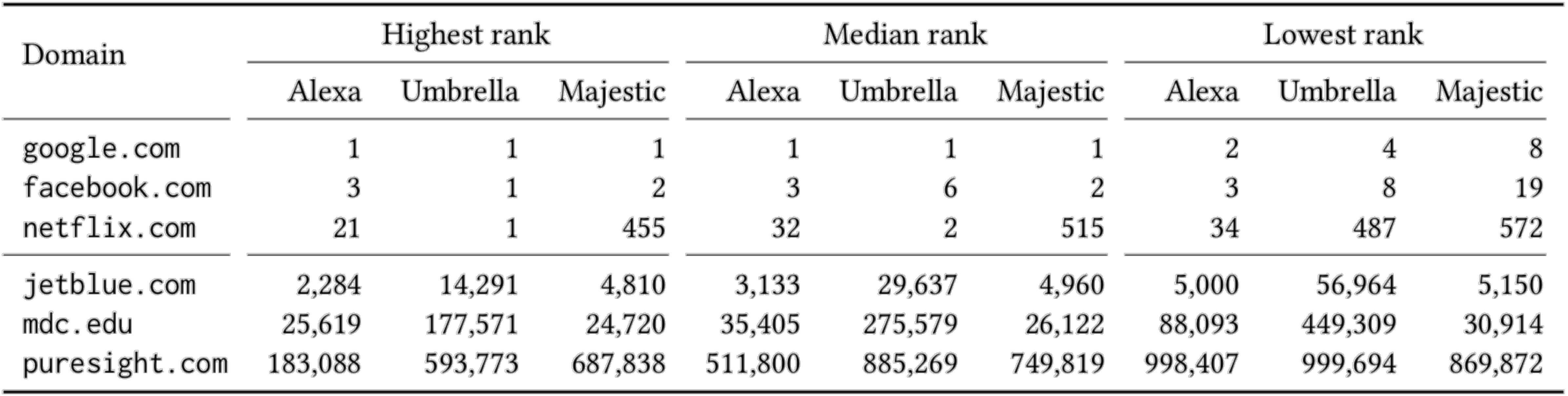
Cumulative sum of all domains ever included in Top 1M lists (Top 1k similar)



Average daily percentage change over rank



Rank variation for some more and less popular websites in the Top 1M lists



Domain	Highest rank			Median rank			Lowest rank		
	Alexa	Umbrella	Majestic	Alexa	Umbrella	Majestic	Alexa	Umbrella	Majestic
google.com	1	1	1	1	1	1	2	4	8
facebook.com	3	1	2	3	6	2	3	8	19
netflix.com	21	1	455	32	2	515	34	487	572
jetblue.com	2,284	14,291	4,810	3,133	29,637	4,960	5,000	56,964	5,150
mdc.edu	25,619	177,571	24,720	35,405	275,579	26,122	88,093	449,309	30,914
puresight.com	183,088	593,773	687,838	511,800	885,269	749,819	998,407	999,694	869,872

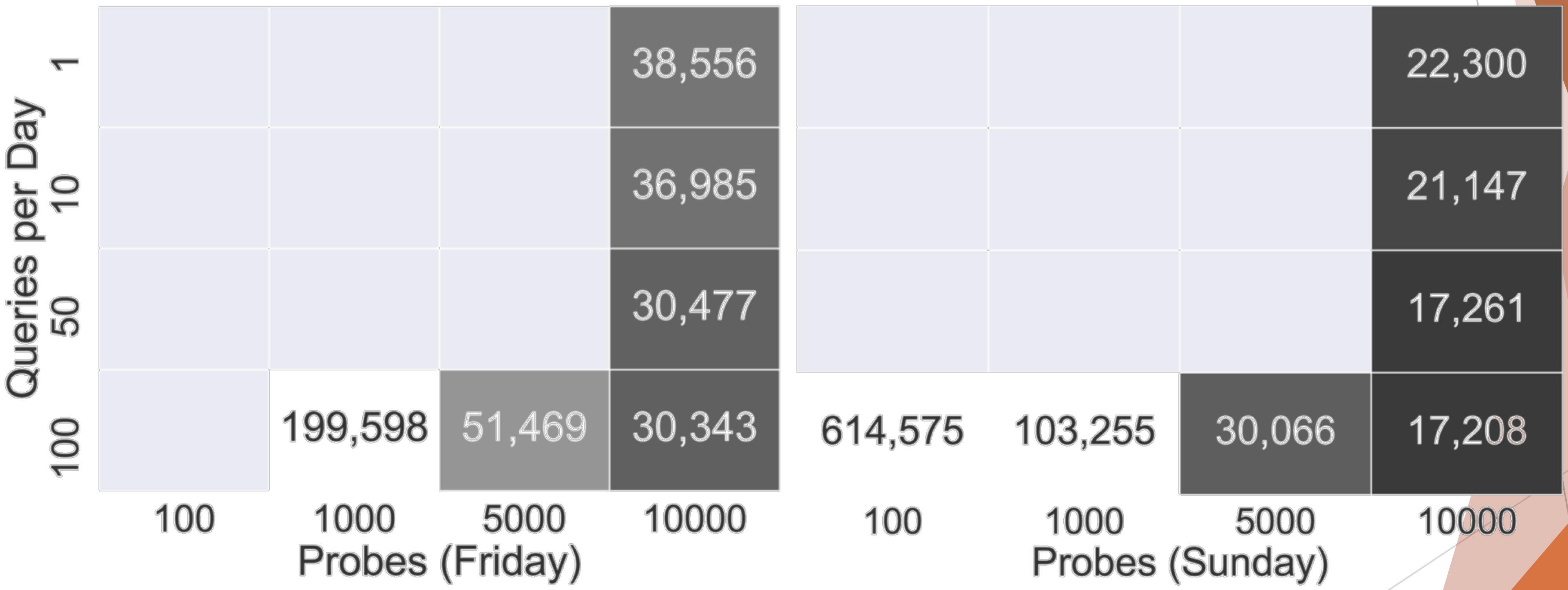
Ways to manipulate the lists

MAJESTIC LIST DIFFICULT TO MANIPULATE, DUE TO ITS RANKING ALGORITHM

ALEXA MANIPULATION WAS INVESTIGATED IN [2]

UMBRELLA TOP LIST CAN BE MANIPULATED AND DONE IN THE PAPER [1]

Umbrella Top 1M list manipulation



Conclusion

- Constantly changing
  - Date of update important
- The lists are not completely representative
- The lists can be manipulated
- Live data available at: toplists.github.io

Sources:
[1]: https://dl.acm.org/citation.cfm?id=3278574
[2]: https://arxiv.org/pdf/1806.01156.pdf

Presented by: Christian Wahl, David Hasselquist, Otto Bergdal