Crawling the USENET for DMCA

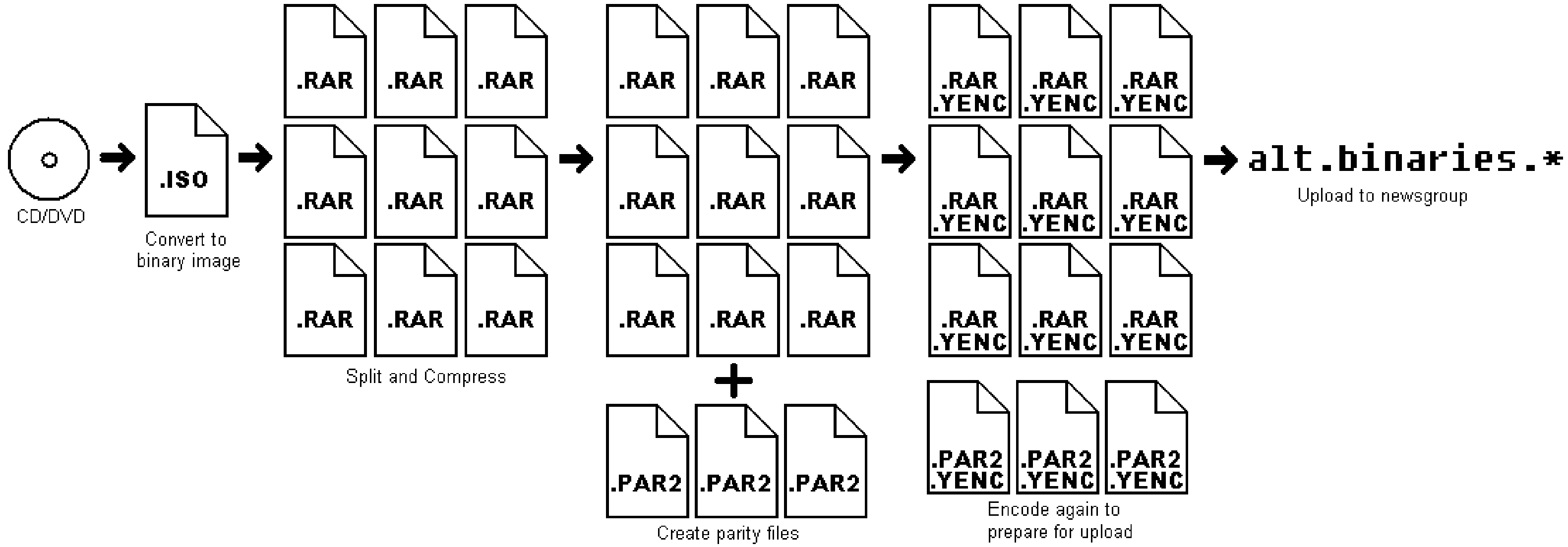
Project: NZB Ninja by Eddie Bijnen

29-05-2015

Digital Millennium Copyright Act

- DMCA
  - Take down notice
  - Electronic Commerce Directive
- DCMA must contain
  - Clear identication of the person or entity submitting the DMCA Notice.
  - Clearly stated relationship to the copyright holder (self or authorizedagent).
  - Message-IDs for all articles the DMCA Notice is requesting to take down.
  - Clear statement, that the information in the notication is accurate and that you are copyright holder, or authorized to act on behalf of the copy-right holder.
  - A "physical or electronic signature" of an authorized person to act on behalf of the owner.

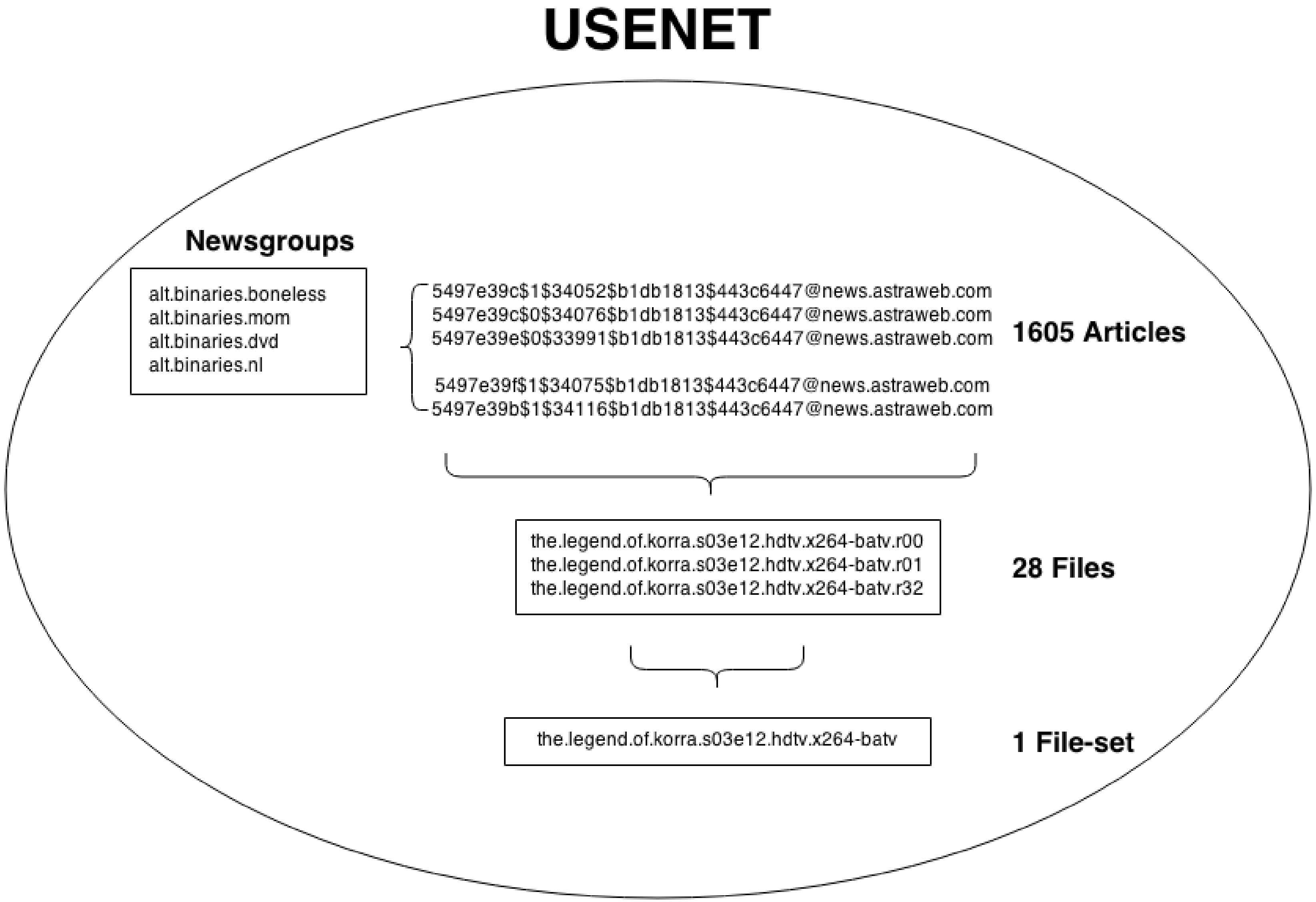
File Structure



Source: https://en.wikipedia.org/wiki/Usenet

File Type	Number of files
Orginal File	1
Compressed files	21
Compressed files & parity	28
Number of articles on USENET	1605

USENET Structure



USENET

Newsgroups

alt.binaries.boneless
alt.binaries.mom
alt.binaries.dvd
alt.binaries.nl

5497e39c$1$34052$b1db1813$443c6447@news.astraweb.com
5497e39c$0$34076$b1db1813$443c6447@news.astraweb.com
5497e39e$0$33991$b1db1813$443c6447@news.astraweb.com

5497e39f$1$34075$b1db1813$443c6447@news.astraweb.com
5497e39b$1$34116$b1db1813$443c6447@news.astraweb.com

1605 Articles

the.legend.of.korra.s03e12.hdtv.x264-batv.r00
the.legend.of.korra.s03e12.hdtv.x264-batv.r01
the.legend.of.korra.s03e12.hdtv.x264-batv.r32

28 Files

the.legend.of.korra.s03e12.hdtv.x264-batv

1 File-set

Removal Behavior

- All or nothing
- 100,000 articles over 200 filesets
- Zero deviation

Research Questions

- Can a comprehensive database, including DCMA take downs, be created?
- What kinds of method exist to keep article availability up-to-date?
- Is it feasible to keep the entire USENET article availability up-to-date?

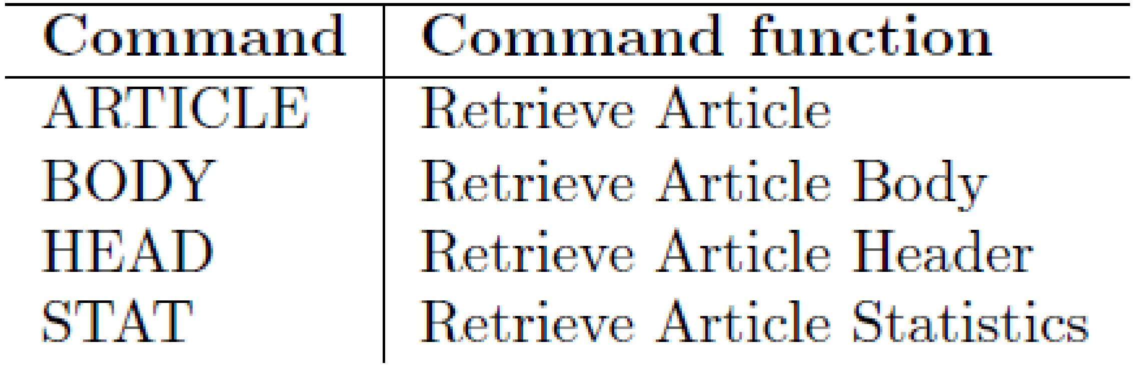
Can a comprehensive database, including DCMA take downs, be created?

- LIST command

- Files that should be available

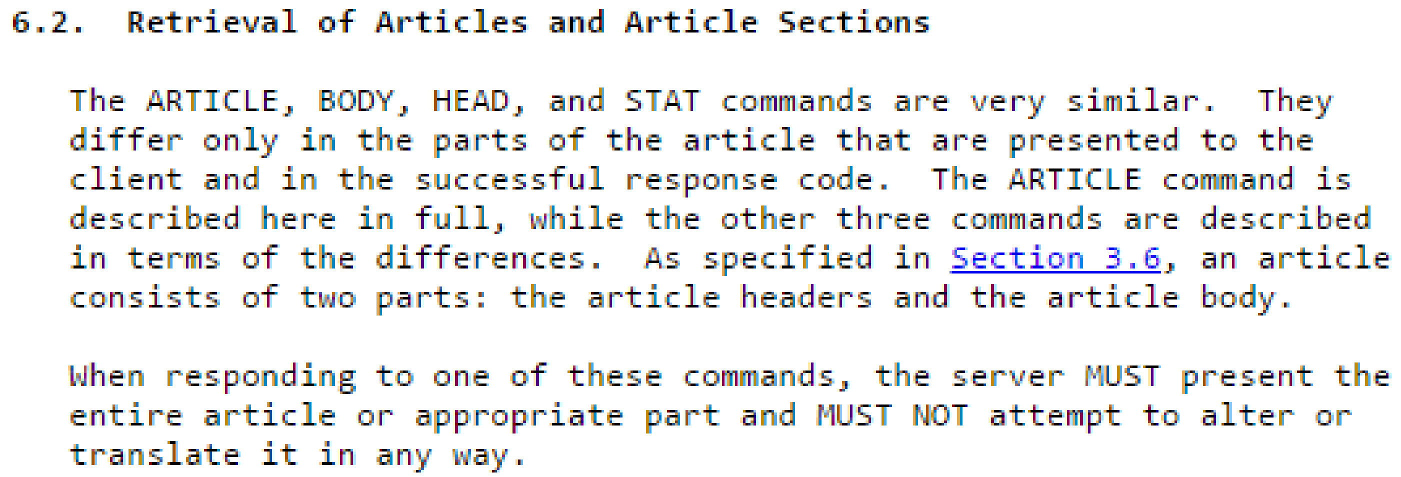
- Too unreliable for availability

What kinds of method exist to keep article availability up-to-date?



Command	Command function
ARTICLE	Retrieve Article
BODY	Retrieve Article Body
HEAD	Retrieve Article Header
STAT	Retrieve Article Statistics

What kinds of method exist to keep article availability up-to-date?



6.2. Retrieval of Articles and Article Sections

The ARTICLE, BODY, HEAD, and STAT commands are very similar. They differ only in the parts of the article that are presented to the client and in the successful response code. The ARTICLE command is described here in full, while the other three commands are described in terms of the differences. As specified in Section 3.6, an article consists of two parts: the article headers and the article body.

When responding to one of these commands, the server MUST present the entire article or appropriate part and MUST NOT attempt to alter or translate it in any way.

Is it feasible to keep the entire USENET article availability up-to-date?

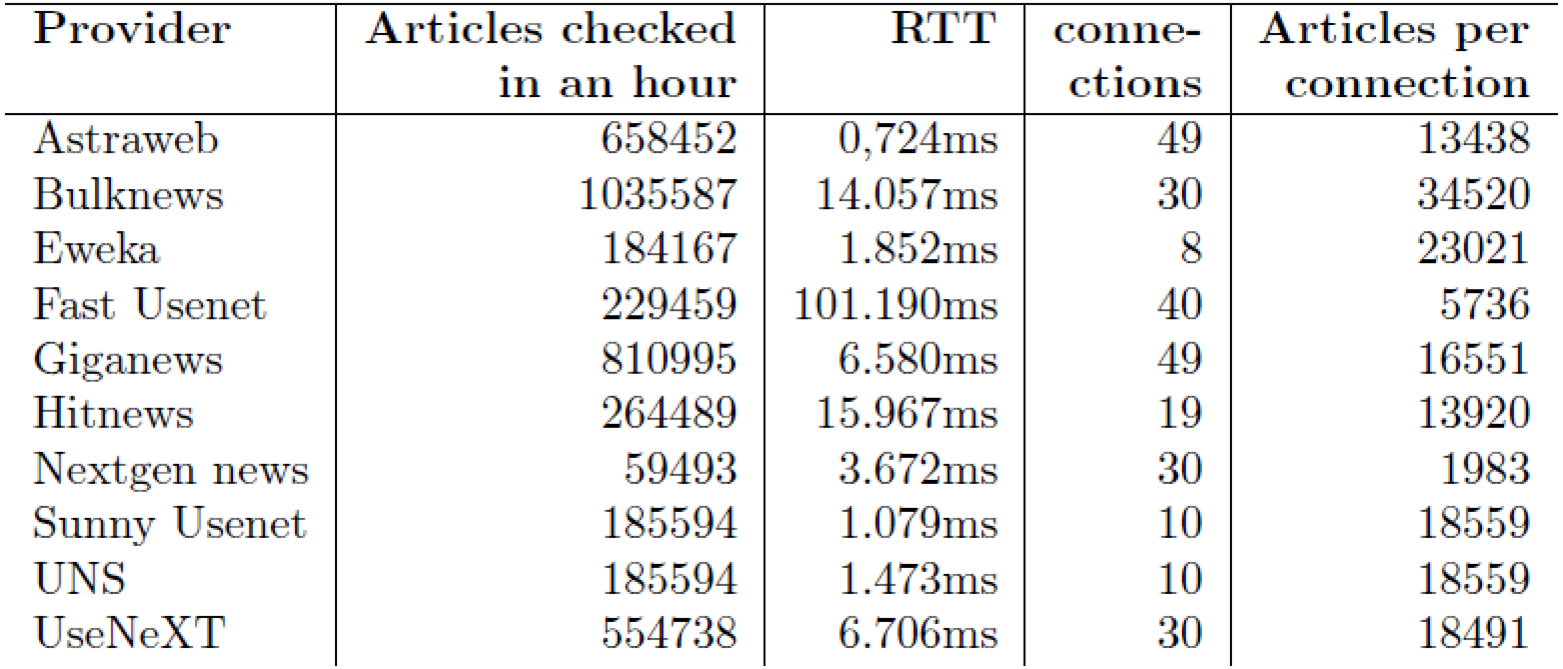
Is it feasible to keep the entire USENET article availability up-to-date?

- It all depends

Is it feasible to keep the entire USENET article availability up-to-date?

- Speed articles can be checked within an hour
  - Network latency
  - Server latency
  - Number of connections
- Growth of the USENET
- Size of the USENET
- Number of articles in a file set

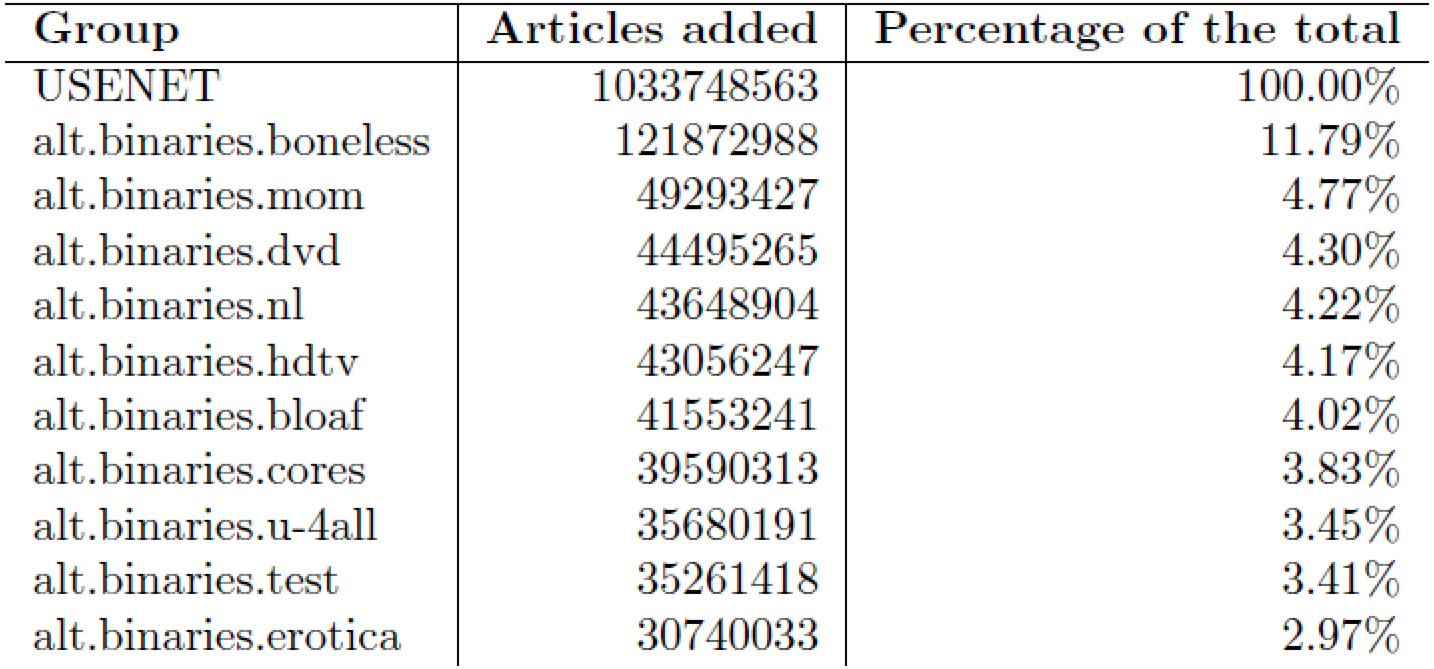
Speed articles can be checked within an hour



Provider	Articles checked in an hour	RTT	connections	Articles per connection
Astraweb	658452	0,724ms	49	13438
Bulknews	1035587	14.057ms	30	34520
Eweka	184167	1.852ms	8	23021
Fast Usenet	229459	101.190ms	40	5736
Giganews	810995	6.580ms	49	16551
Hitnews	264489	15.967ms	19	13920
Nextgen news	59493	3.672ms	30	1983
Sunny Usenet	185594	1.079ms	10	18559
UNS	185594	1.473ms	10	18559
UseNeXT	554738	6.706ms	30	18491

Growth of the USENET

- USENET growth in two weeks time



Group	Articles added	Percentage of the total
USENET	1033748563	100.00%
alt.binaries.boneless	121872988	11.79%
alt.binaries.mom	49293427	4.77%
alt.binaries.dvd	44495265	4.30%
alt.binaries.nl	43648904	4.22%
alt.binaries.hdtv	43056247	4.17%
alt.binaries.bloaf	41553241	4.02%
alt.binaries.cores	39590313	3.83%
alt.binaries.u-4all	35680191	3.45%
alt.binaries.test	35261418	3.41%
alt.binaries.erotica	30740033	2.97%

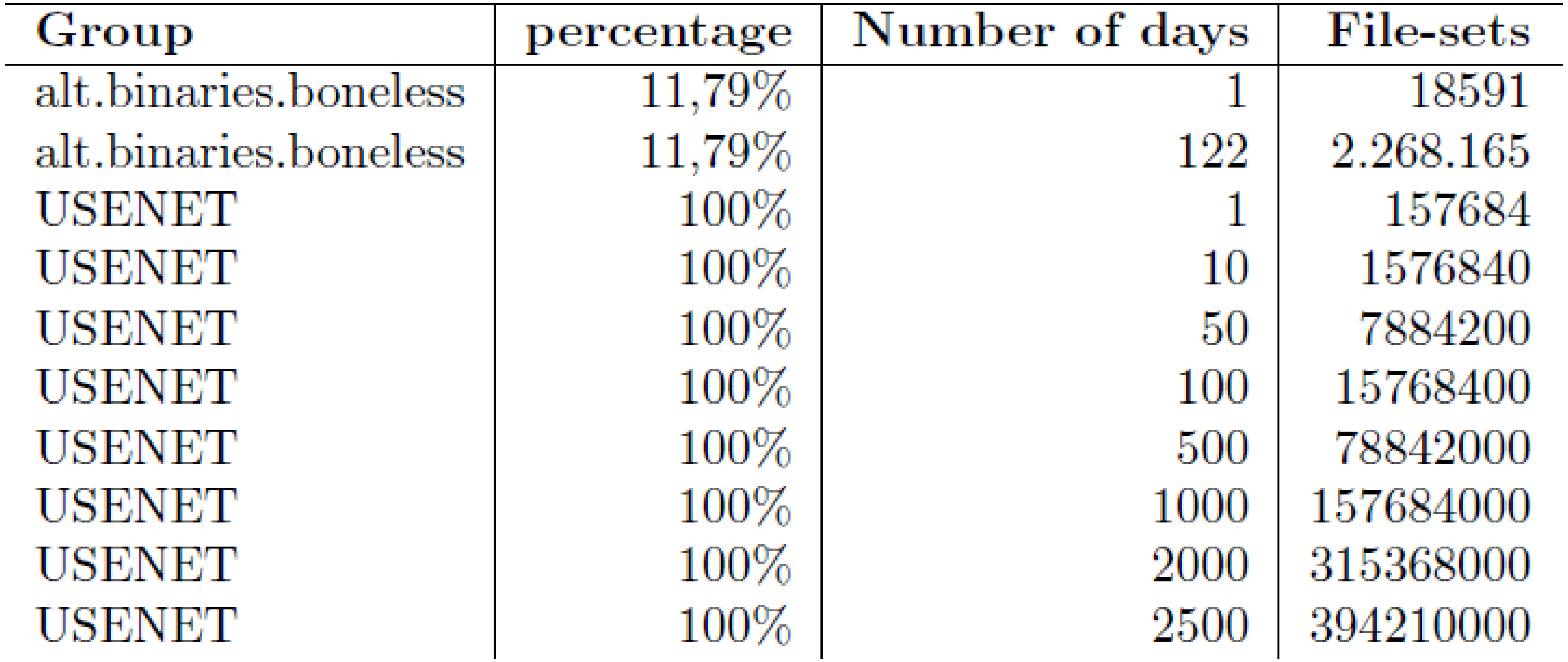
Size of the USENET

- 5 quadrillion
- Number of file sets?

Number of articles in a file set

- The following assumptions had to be made to make this calculation
  - The average number of articles in a file set is the same for all newsgroups;
  - The number of articles posted between 1-04-2015 and 14-04-2015 is representative for an average day;
  - The growth of the amount of articles has been constant.

Number of articles in a file set



Group	percentage	Number of days	File-sets
alt.binaries.boneless	11,79%	1	18591
alt.binaries.boneless	11,79%	122	2.268.165
USENET	100%	1	157684
USENET	100%	10	1576840
USENET	100%	50	7884200
USENET	100%	100	15768400
USENET	100%	500	78842000
USENET	100%	1000	157684000
USENET	100%	2000	315368000
USENET	100%	2500	394210000

Is it feasible to keep the entire USENET article availability up-to-date?

- Show hands if you think we can check 10 days worth of USENET in a day.

Is it feasible to keep the entire USENET article availability up-to-date?

- Show hands if you think we can check 10 days worth of USENET in a day.

- 50 days worth of USENET

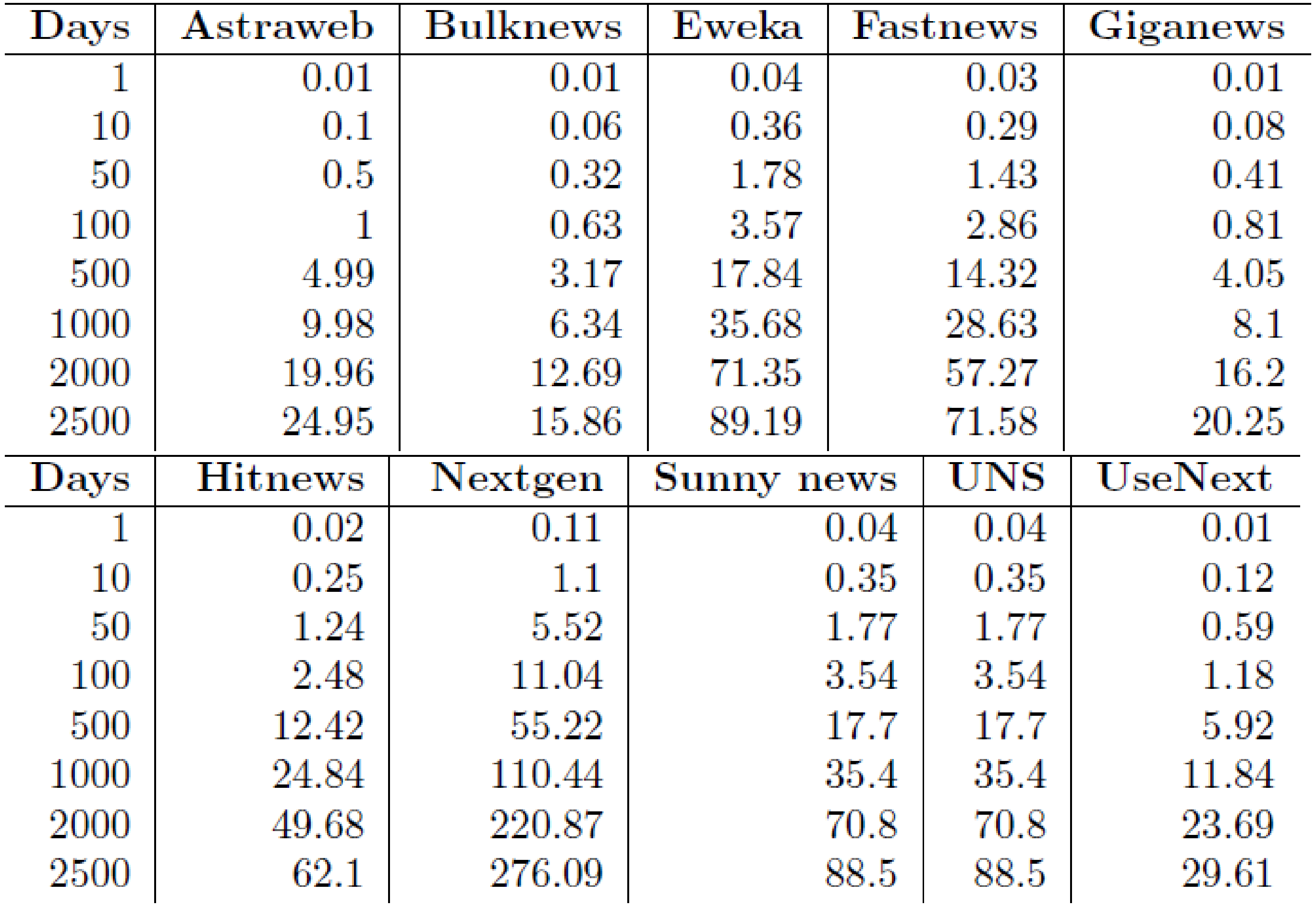
Is it feasible to keep the entire USENET article availability up-to-date?

- Show hands if you think we can check 10 days worth of USENET in a day.
- 50 days worth of USENET
- 100 days worth of USENET

Is it feasible to keep the entire USENET article availability up-to-date?

- Show hands if you think we can check 10 days worth of USENET in a day.
- 50 days worth of USENET
- 100 days worth of USENET
- 500 days worth of USENET

Is it feasible to keep the entire USENET article availability up-to-date?



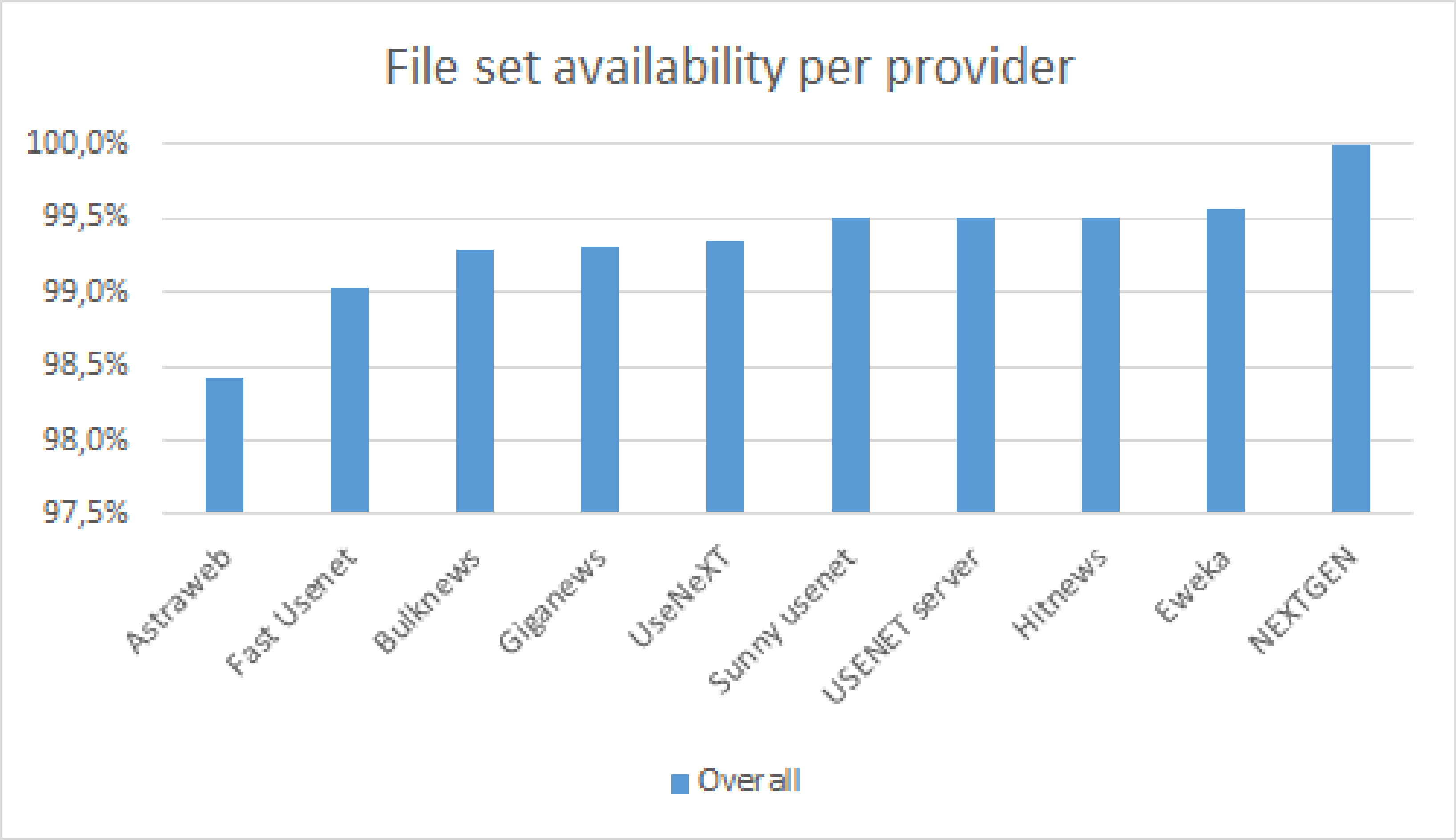
Days	Astraweb	Bulknews	Eweka	Fastnews	Giganews
1	0.01	0.01	0.04	0.03	0.01
10	0.1	0.06	0.36	0.29	0.08
50	0.5	0.32	1.78	1.43	0.41
100	1	0.63	3.57	2.86	0.81
500	4.99	3.17	17.84	14.32	4.05
1000	9.98	6.34	35.68	28.63	8.1
2000	19.96	12.69	71.35	57.27	16.2
2500	24.95	15.86	89.19	71.58	20.25

Days	Hitnews	Nextgen	Sunny news	UNS	UseNext
1	0.02	0.11	0.04	0.04	0.01
10	0.25	1.1	0.35	0.35	0.12
50	1.24	5.52	1.77	1.77	0.59
100	2.48	11.04	3.54	3.54	1.18
500	12.42	55.22	17.7	17.7	5.92
1000	24.84	110.44	35.4	35.4	11.84
2000	49.68	220.87	70.8	70.8	23.69
2500	62.1	276.09	88.5	88.5	29.61

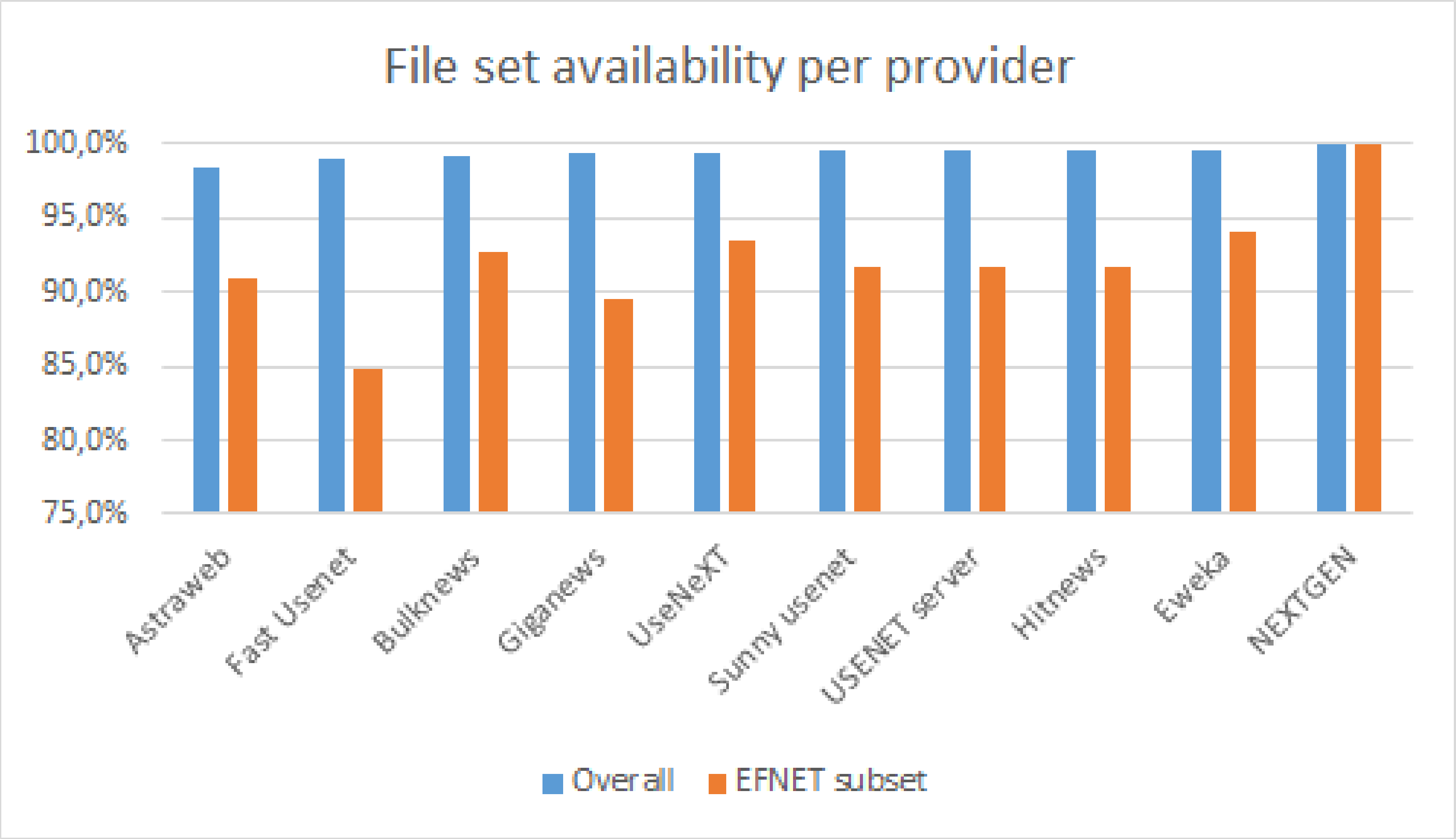
Interpreting gathered data

- 3,211,532 file sets
- 382,364,225 availability checks
- 1,980,947,402 articles
- Database of 692 GB

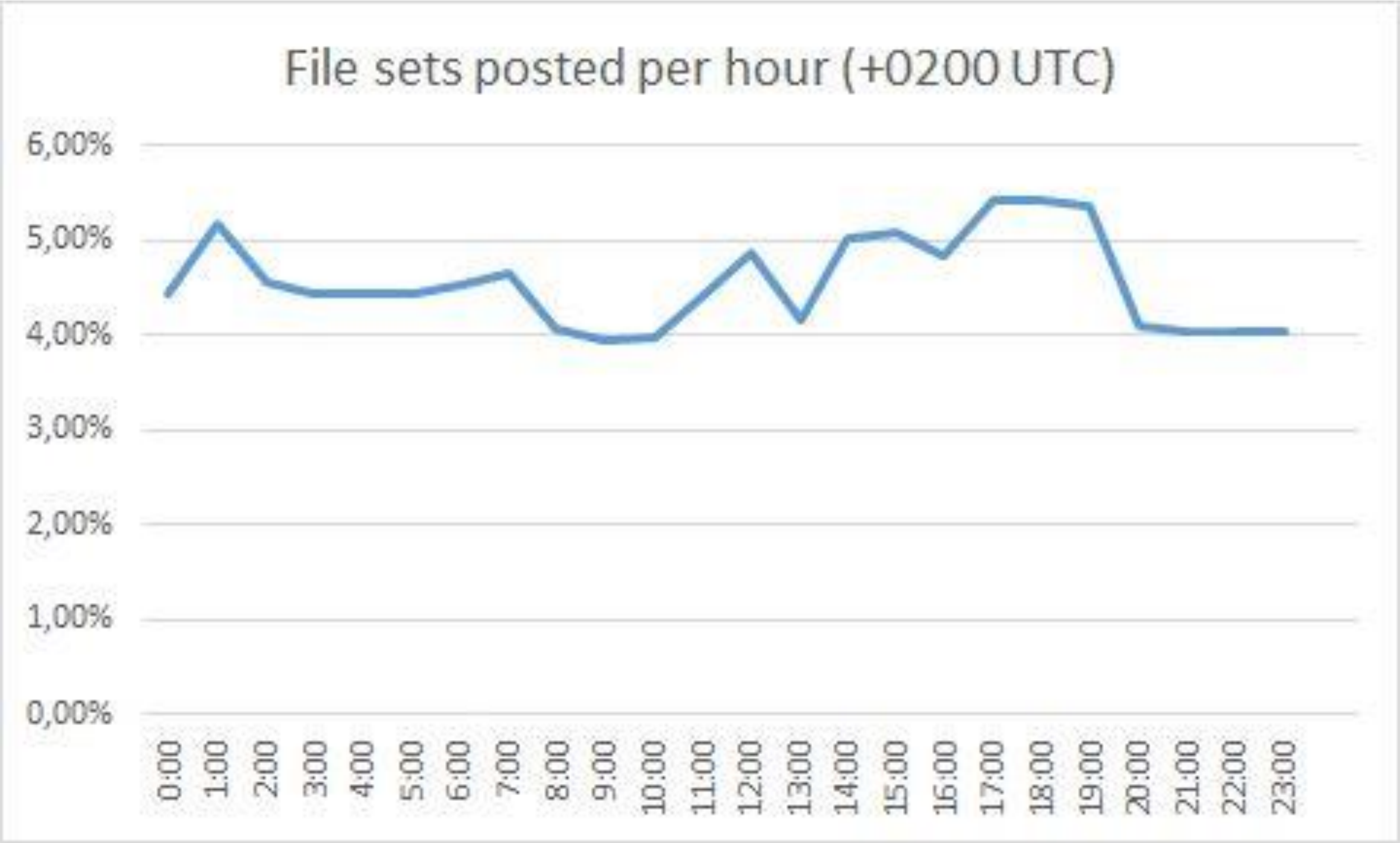
File set availability



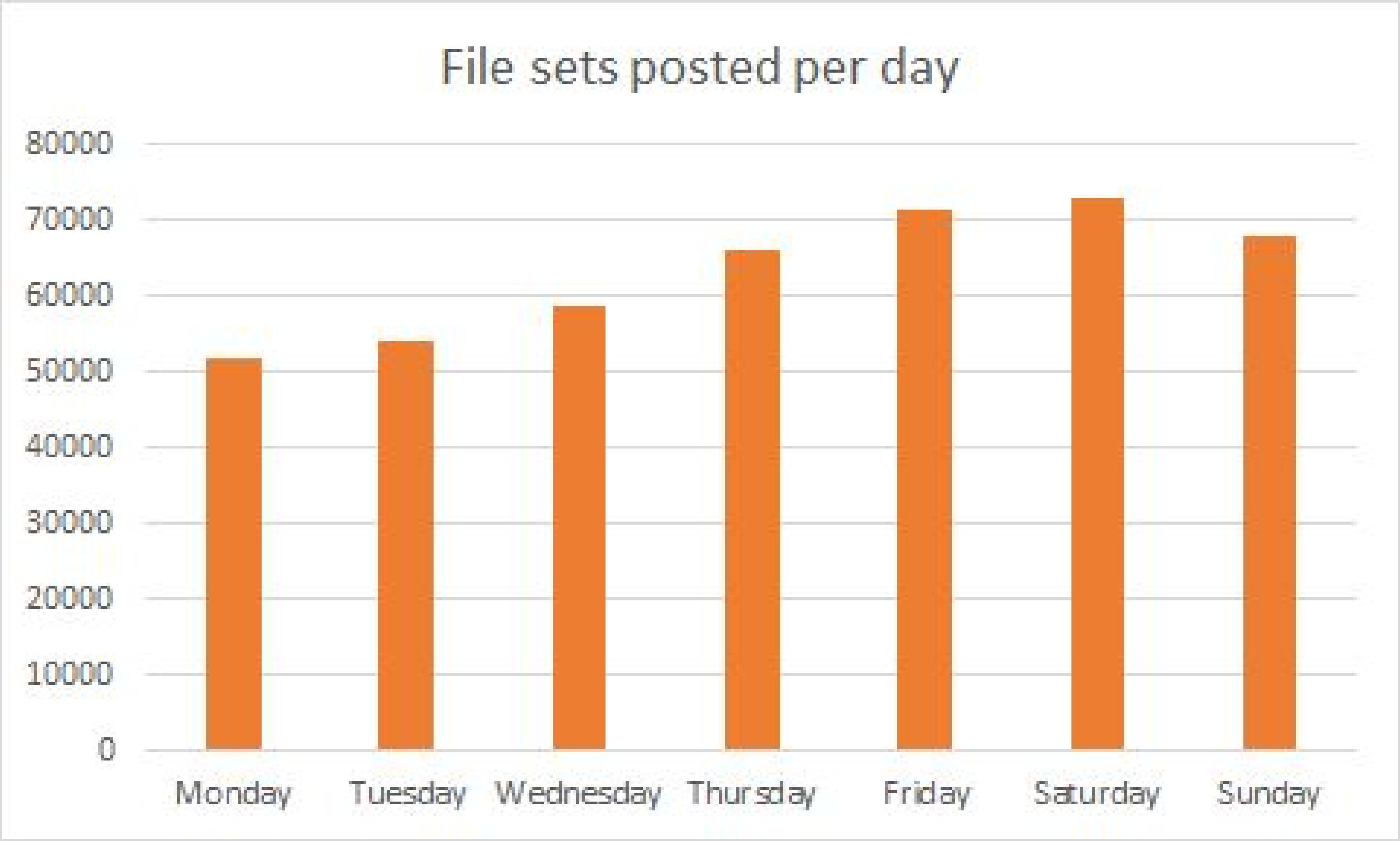
File set availability



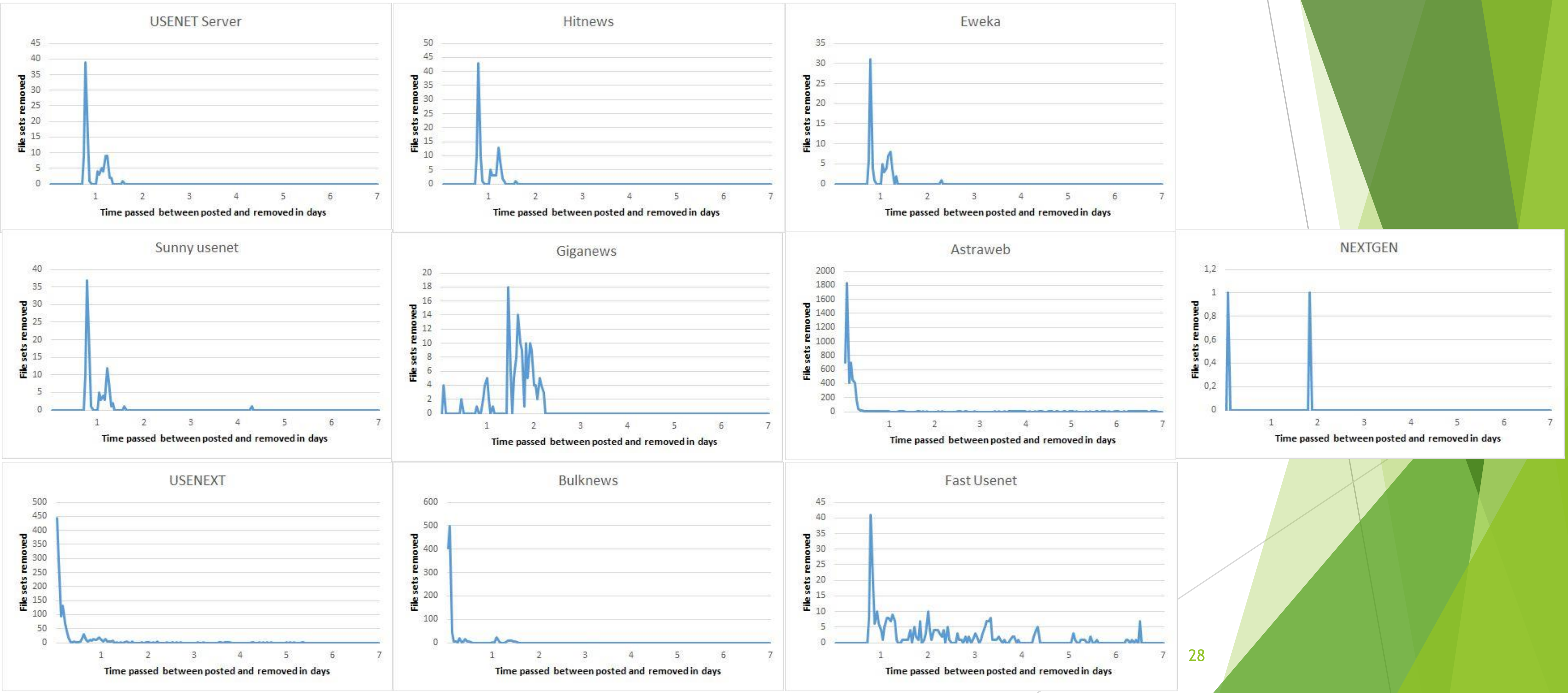
File sets posted to the USENET



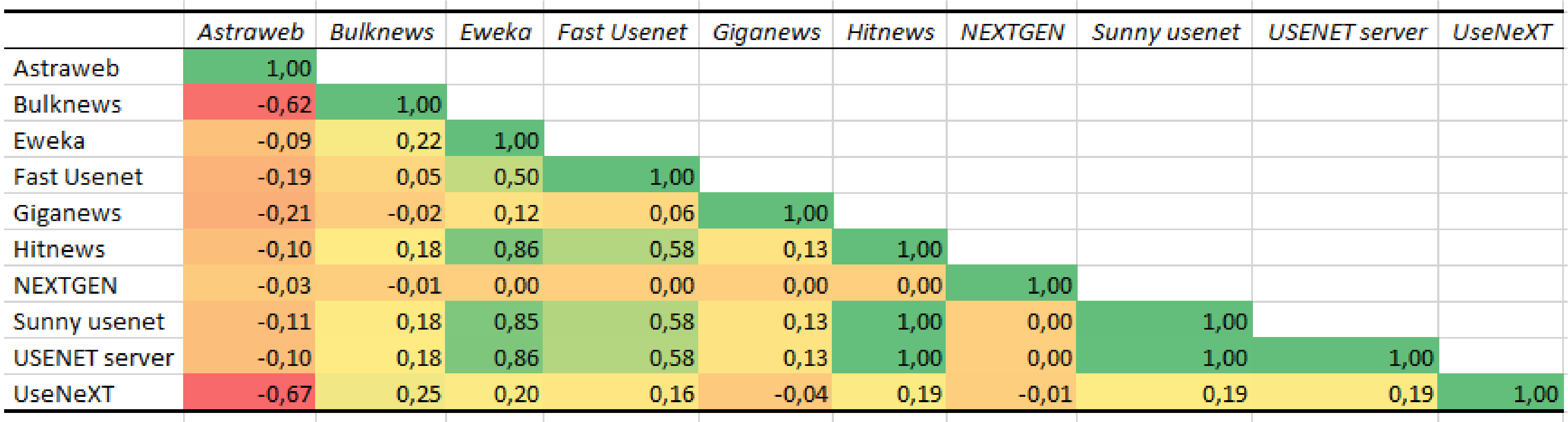
File sets posted to the USENET



File sets time until removed

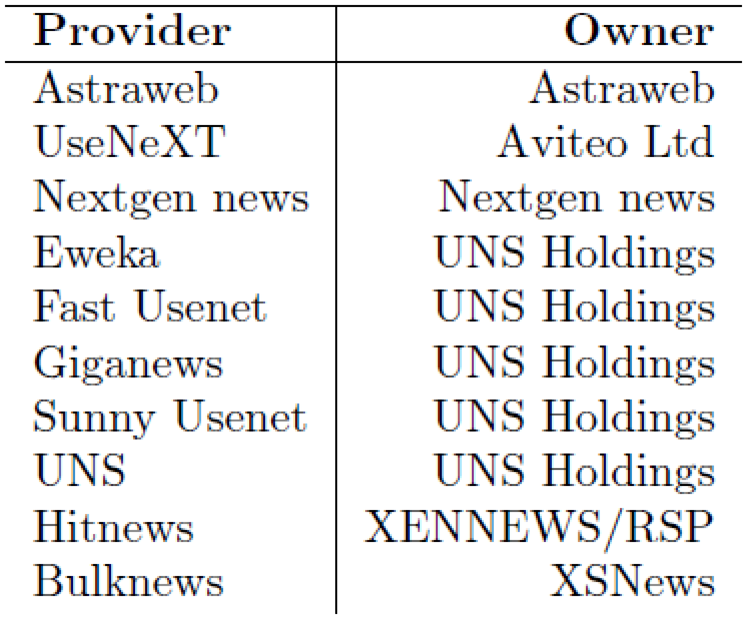


Correlation between providers



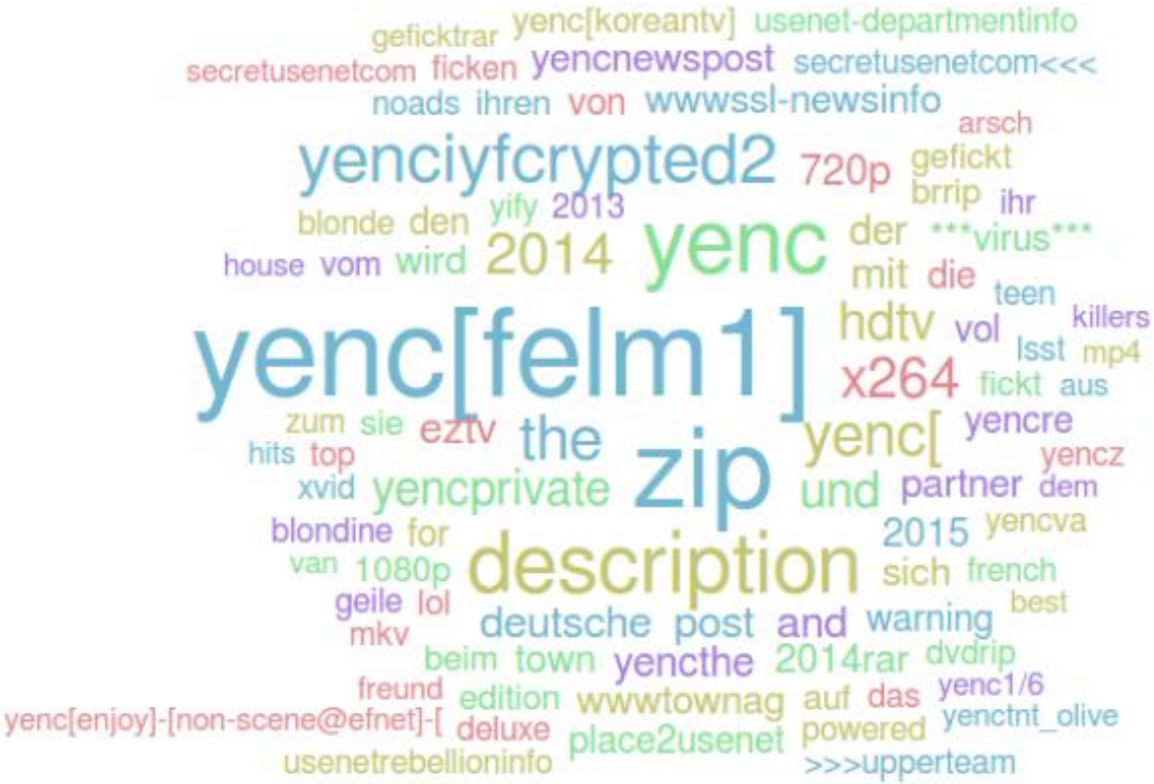
	Astraweb	*Bulknews*	*Eweka*	*Fast Usenet*	*Giganews*	*Hitnews*	*NEXTGEN*	*Sunny usenet*	*USENET server*	*UseNeXT*
Astraweb	1,00									
Bulknews	-0,62	1,00								
Eweka	-0,09	0,22	1,00							
Fast Usenet	-0,19	0,05	0,50	1,00						
Giganews	-0,21	-0,02	0,12	0,06	1,00					
Hitnews	-0,10	0,18	0,86	0,58	0,13	1,00				
NEXTGEN	-0,03	-0,01	0,00	0,00	0,00	0,00	1,00			
Sunny usenet	-0,11	0,18	0,85	0,58	0,13	1,00	0,00	1,00		
USENET server	-0,10	0,18	0,86	0,58	0,13	1,00	0,00	1,00	1,00	
UseNeXT	-0,67	0,25	0,20	0,16	-0,04	0,19	-0,01	0,19	0,19	1,00

Correlation between providers



Provider	Owner
Astraweb	Astraweb
UseNeXT	Aviteo Ltd
Nextgen news	Nextgen news
Eweka	UNS Holdings
Fast Usenet	UNS Holdings
Giganews	UNS Holdings
Sunny Usenet	UNS Holdings
UNS	UNS Holdings
Hitnews	XENNEWS/RSP
Bulknews	XSNews

Word cloud of removed articles



Got time for a quick demo?

Conclusion

- You can create an comprehensive database, including DCMA take downs.
- Stat command is the ideal command to check article availability.
- It is feasible to keep the entire USENET article availability up-to-date.
  - Depends on your definition of up-to-date and amount of USENET accounts.

Future Work

- Intelligent scheduler
- Larger retention
- Popularity

Special thanks to

- nZEDb creators
- The USENET providers that provided a free trial account
  - Guido for his Hitnews account
- Arno Bakker & Niels Sijm for supervising
- Tessa Wassenaar for spell checking

Questions?

- Paper available on:
  - http://nzb.ninja/RP.pdf
- Search engine URL
  - http://nzb.ninja/
- Contact info
  - eddie@edworks.info
  - PGP SIG: 64D1 C460 FB44 F5CE 8D31 9767 326F 9E8D B421 792B