DGArchive

A Comprehensive Measurement Study of Domain Generating Malware

Daniel Plohmann, Khaled Yakdan, Michael Klatt, Johannes Bader, Elmar Padilla

Fraunhofer FKIE universitätbonn DOMAINTOOLS Fraunhofer FKIE

2016-08-10 | USENIX | Austin, Texas

Agenda

- Intro: Domain Generation Algorithms
- Summary of Findings

- Data Set
- DGA Taxonomy
- DGA Registrations & Collisions
- Conclusion

Intro

Domain Generation Algorithms

Domain Generation Algorithms

Definitions

- Concept first described ~2008: Domain Flux
- Domain Generation Algorithm (DGA)
  - An algorithm producing Command & Control rendezvous points dynamically
  - Shared secret between malware running on compromised host and botmaster
- Seeds
  - Collection of parameters influencing the output of the algorithm
- Algorithmically-Generated Domain (AGD)
  - Domains resulting from a DGA

Domain Generation Algorithms

Motivation for Usage

- Aggravation of Analysis
  - No hardcoded domains / dumping -> code analysis needed
- Evasion
  - Many DGAs have short-lived domains -> avoid blacklisting
- Backup
  - Registration only when needed
- Asymmetry
  - Attacker needs one domain, defender needs to prohibit access to all
- Feasability of DGAs
  - Domains are cheap

TL;DR

Key Findings

Key Findings

What we did

- Recovered 43 DGA implementations through reverse engineering
  - Average effort: 1 day per DGA, 1 hour per seed
- We used 253 seeds to generate 159,712,234 unique AGDs
  - Derived characteristics and a taxonomy
  - Analyzed collisions between DGAs
- Checked WHOIS records for 18,446,125 domains of 38 families
  - Analyzed registration status and behavior over time
  - Analyzed collisions with benign domains

Key Findings

What we found out

- 3/43 DGAs compose their domains based on wordlists
  - Previous detection methodology for DGA identification remains relevant, e.g. „Detecting the Rise of DGA based malware“, Antonakakis et al., 2012.
- DGAs increase in relevance:
  - 25/43 DGAs appeared 2013 and later (first DGA spotted in 2008)
- From a botmaster perspective, DGAs are effective
  - AGDs rarely / barely mitigated in time
  - Previous takedowns had significant financial efforts for domain purchase
- DGA domains barely collide (with other DGAs or benign domains)
  - DGA domains are a great feature to identify the malware family & campaign
  - Lists of DGA domains can be used for blocking with basically zero FPs

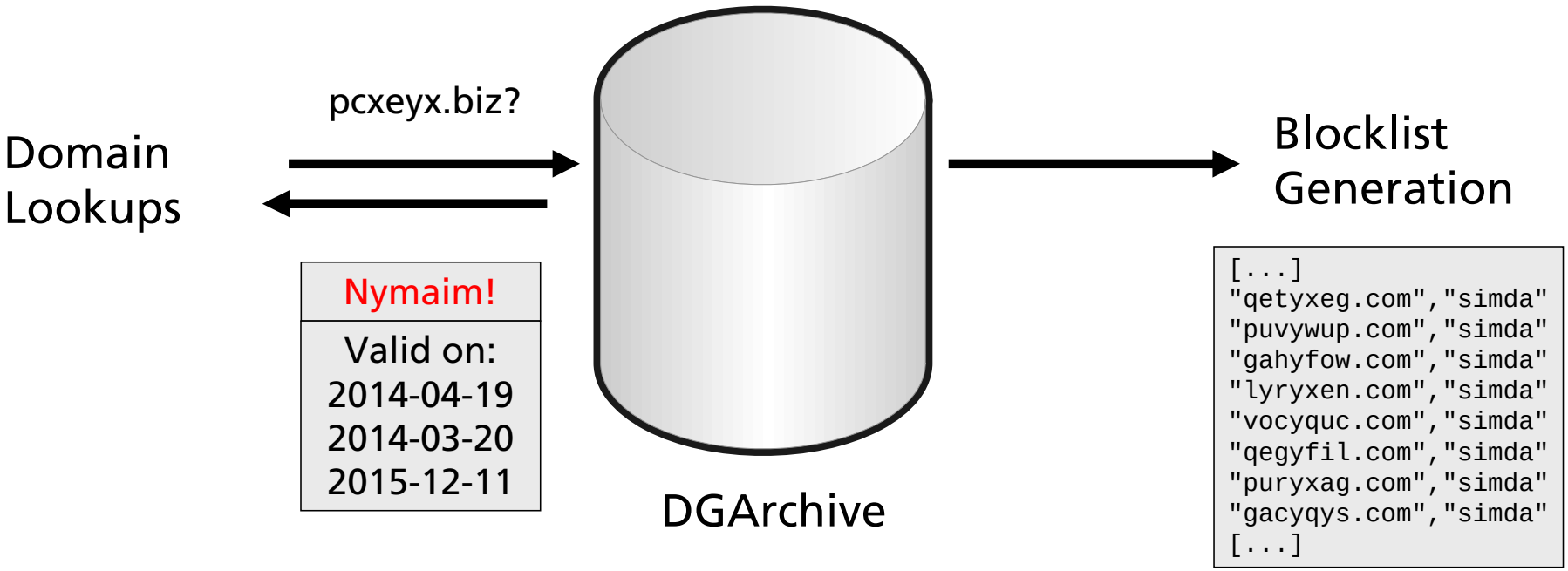
DGArchive

Data Set

DGArchive

The idea

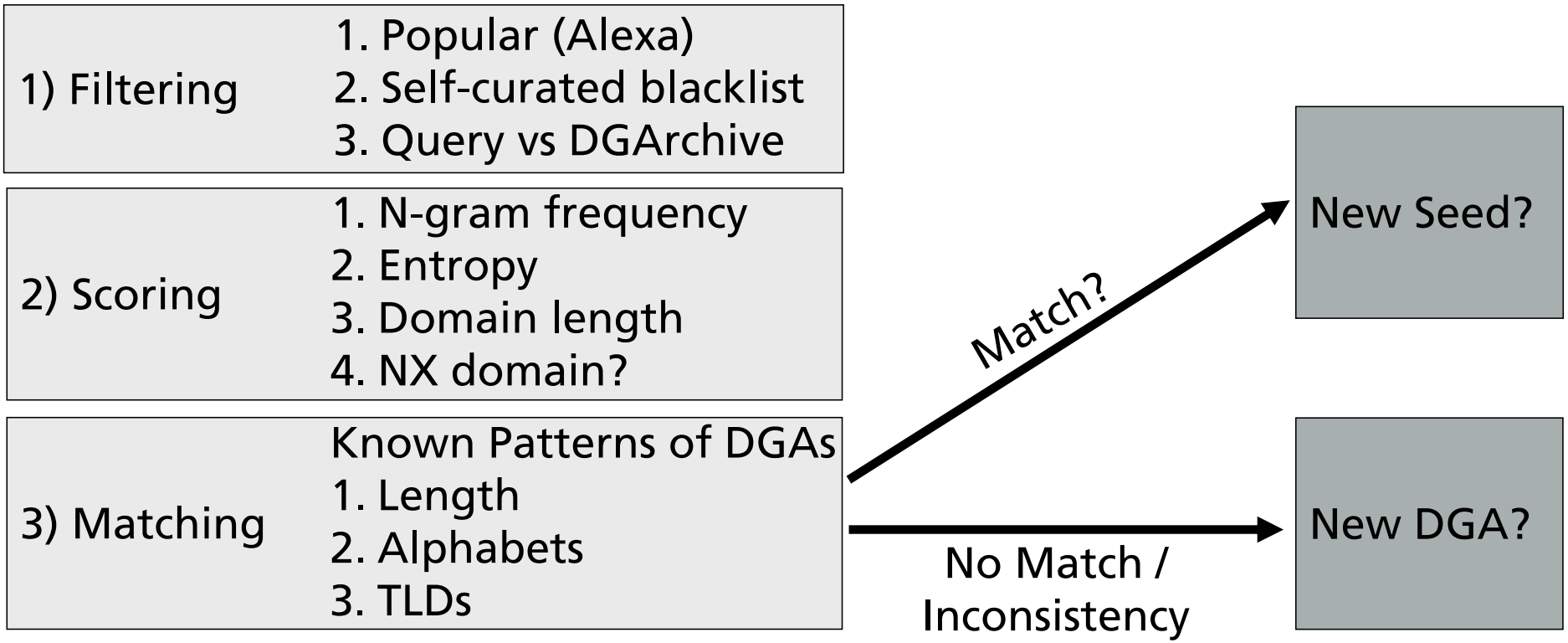
- Core idea:
  - **Reverse engineer** DGAs, then generate and archive all possible domains since first spotting of the respective malware family



Finding DGAs

Mining a Sandbox DNS feed

- Remix of academic approaches and common sense
  - Input: Lists of domains, as queried during a sandbox runs
  - DNS Feed by shadowserver ←THANK YOU!!!
  - Paper: 1,235,443 sandbox runs; 15,660,256 DNS queries (959,607 unique)
  - By now: 8,800,652 sandbox runs; 81,661,353 DNS queries (4,782,015 unique)



DGArchive

Status

- Data set used in this paper (September 2015)
  - 43 families/variants, 253 seeds, 20+ million domains
- DGArchive Today
  - 59 families/variants, 511 seeds, 49+ million domains
  - Web-service used by 300+ parties, including ISPs, financial institutions, CERTs, LEA, AVs, researchers

DGA Taxonomy

DGA Features

Taxonomy and Generation Schemes

- DGA Classes (cp. Barabosch et al. [1]):

Type	Time dependent	Deterministic	Example
TID	☒	☑	Kraken, TinyBanker
TDD	☑	☑	Conficker, Gameover Zeus
TDN	☑	☒	Torpig, Bedep
TIN	☒	☒	-

- Generation Schemes

Type	Example Family	Example Domain
Arithmetic (A)	DirCrypt	pibqzedhzwt.com
Wordlist (W)	Matsnu	termacceptyear.com
Hashing (H)	Bamital	b83ed4877eec1997fcc39b7ae590007a.info
Permutation (P)	VolatileCedar	dotnetexplorer.info

[1] https://net.cs.uni-bonn.de/fileadmin/user_upload/wichmann/Extraction_DNGA_Malware.pdf

DGA

Domain Usage

DGA Domain Space

as seen by DGArchive

- What we were interested in:
  - How many potential DGA domains are there?
  - How many of these domains are registered?
  - Are there collisions between DGAs / benign domains?
  - *Are there recognizable patterns when TD* DGA domains are registered?*
  - *How effective have been mitigations against DGA malware in the past?*

DGA Domain Space

as seen by DGArchive

- Data set fixed on 22nd September 2015
  - AGDs generated from malware's first spotting until 31.12.2015
- DomainTools provided historic WHOIS data for this study
- Evaluation of WHOIS features for majority of DGAs
  - Sinkholes
  - Mitigations (registration turned to sinkhole at later point)
  - Pre-registrations (registration before appearence of the family)
  - *Domain Parking*

Bamital 197,000	Fobber 2,000	Mewsei 1,984	Pykspa 2 775,342	Simda 11,528
Banjori 421,390	Geodo 90,232	Murofet 1 4,063,680	QakBot 385,000	Suppobox 98,304
Bedep 3,806	Gameover DGA 6,182,000	Murofet 2 262,000	Ramdo 3000	Szribi 2,949
Conficker 125,118,625	Gameover P2P 262,000	Necurs 3,551,232	Ramnit 18,000	Tempedreve 204
CoreBot 18,160	Gozi 16,963	Nymaim 65,040	Ranbyus 64,400	TinyBanker 81,930
Cryptolocker 1,108,000	Hesperbot 178	Pushdo 124,021	Redyms 34	Torpig 17,610
DirCrypt 420	Kraken 300	Pushdo TID 6,000	Rovnix 10,000	UrlZone 10,009
Dyre 592,000	Matsnu 3,346	Pykspa 1 22,764	Shifu 1,554	Virut 15,335,008

Generated Domains?

159,712,234 unique domains until end of 2015

Bamital 8,340 (4.22%) 197,000	Fobber 13 (0.65%) 2,000	Mewsei DDNS 1,984	Pykspa 2 1,927 (0.25%) 775,342	Simda 379 (3.29%) 11,528
Banjori 683 (0.16%) 421,390	Geodo 107 (0.12%) 90,232	Murofet 1 3,172 (0.08%) 4,063,680	QakBot 1,088 (0.28%) 385,000	Suppobox 11,338 (11.53%) 98,304
Bedep 654 (17.18%) 3,806	Gameover DGA 1,081 (0.02%) 6,182,000	Murofet 2 559 (0.21%) 262,000	Ramdo 47 (1.57%) 3000	Szribi 54 (1.83%) 2,949
Conficker - 125,118,625	Gameover P2P 74,755 (28.53%) 262,000	Necurs 295 (0.01%) 3,551,232	Ramnit 939 (5.22%) 18,000	Tempedreve 20 (9.80%) 204
CoreBot DDNS 18,160	Gozi 305 (1.80%) 16,963	Nymaim 656 (1.01%) 65,040	Ranbyus 98 (0.15%) 64,400	TinyBanker 1,733 (2.12%) 81,930
Cryptolocker 3,820 (0.34%) 1,108,000	Hesperbot 15 (8.43%) 178	Pushdo 453 (0.37%) 124,021	Redyms 11 (32.35%) 34	Torpig 139 (0.79%) 17,610
DirCrypt 86 (20.48%) 420	Kraken DDNS 300	Pushdo TID 245 (4.08%) 6,000	Rovnix 1 (0.01%) 10,000	UrlZone 127 (1.27%) 10,009
Dyre 850 (0.14%) 592,000	Matsnu 610 (18.23%) 3,346	Pykspa 1 455 (2.00%) 22,764	Shifu 11 (0.71%) 1,554	Virut - 15,355,008

Registrations? 115,079 (0.62%) of 18,465,427 unique domains we had data for.

Bamital
7,891 / 8,340
197,000

Gameover P2P
72,713 / 74,755
262,000

Cryptolocker
2,899 / 3,820
1,108,000

	#domains	percent
Non-Takedown:	31,884	
Pre-registered:	9,646	30.25%
Mitigated:	1,199	3.76%
Sinkhole:	5,177	16.24%
Remainder:	15,862	49.75%

Takedowns actually account for 72,56% of all considered DGA registrations.
78,326 of 115,387 total registrations

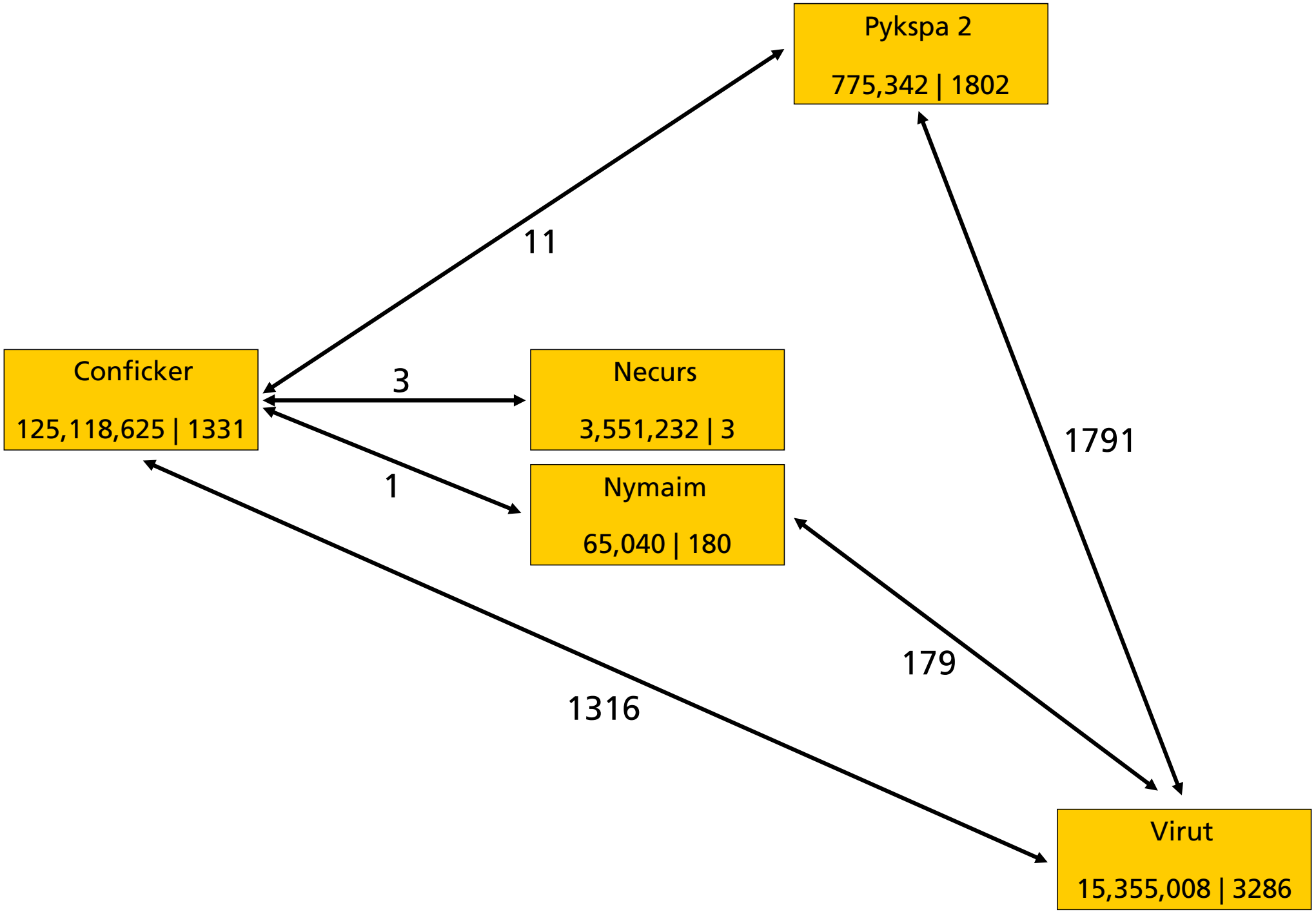
DGA Domain Space

Domain Collisions

- DGA domains may collide
  - With other DGAs
  - With benign domains

Bamital 197,000	Fobber 2,000	Mewsei 1,984	Pykspa 2 775,342	Simda 11,528
Banjori 421,390	Geodo 90,232	Murofet 1 4,063,680	QakBot 385,000	Suppobox 98,304
Bedep 3,806	Gameover DGA 6,182,000	Murofet 2 262,000	Ramdo 3000	Szribi 2,949
Conficker 125,118,625	Gameover P2P 262,000	Necurs 3,551,232	Ramnit 18,000	Tempedreve 204
CoreBot 18,160	Gozi 16,963	Nymaim 65,040	Ranbyus 64,400	TinyBanker 81,930
Cryptolocker 1,108,000	Hesperbot 178	Pushdo 124,021	Redyms 34	Torpig 17,610
DirCrypt 420	Kraken 300	Pushdo TID 6,000	Rovnix 10,000	UrlZone 10,009
Dyre 592,000	Matsnu 3,346	Pykspa 1 22,764	Shifu 1,554	Virut 15,355,008

How many domain collisions between Conficker+Virut and the other DGAs?



Considerably low (3,301 / 159,712,234).

Bamital 197,000	Fobber 2,000	Mewsei 1,984	Pykspa 2 775,342	Simda 11,528
Banjori 421,390	Geodo 90,232	Murofet 1 4,063,680	QakBot 385,000	Suppobox 98,304
Bedep 3,806	Gameover DGA 6,182,000	Murofet 2 262,000	Ramdo 3000	Szribi 2,949
Conficker 125,118,625	Gameover P2P 262,000	Necurs 3,551,232	Ramnit 18,000	Tempedreve 204
CoreBot 18,160	Gozi 16,963	Nymaim 65,040	Ranbyus 64,400	TinyBanker 81,930
Cryptolocker 1,108,000	Hesperbot 178	Pushdo 124,021	Redyms 34	Torpig 17,610
DirCrypt 420	Kraken 300	Pushdo TID 6,000	Rovnix 10,000	UrlZone 10,009
Dyre 592,000	Matsnu 3,346	Pykspa 1 22,764	Shifu 1,554	Virut 15,355,008

And now without Conficker & Virut?

Pykspa 2

775,342 | 1!

- One single collision between Nymaim and Pykspa2

„wttttf.net“!

Nymaim

65,040 | 1!

One single collision!

Conclusions:

- So if there are so few collisions between DGAs…
  - using pre-calculated AGDs to tag malware is extremely accurate!
  - For both family and campaign
  - Effectively **NO** False Positives for domain length 7+

Inter-DGA domain collisions: Basically non-existent!

Bamital 0 / 8,340 197,000	Fobber 0 / 13 2,000	Mewsei	Pykspa 2 757 / 1,927 (39.28%)	Simda 66 / 379 (17.41%) 11,528
Banjori 0 / 683 421,390	Geodo 0 / 107 90,232	Murofet 1 0 / 3,172 4,063,680	QakBot 0 / 1,088 385,000	Suppobox 8.434 / 11,338 (74.39%)
Bedep 0 / 654 3,806	Gameover DGA 0 / 1,081 6,182,000	Murofet 2 0 / 559 262,000	Ramdo 0 / 47 3000	Szribi 0 / 54 2,949
Conficker	Gameover P2P 0 / 74,755 262,000	Necurs 10 / 295 (3.34%) 3,551,232	Ramnit 0 / 939 18,000	Tempedreve 0 / 20 204
CoreBot	Gozi 48 / 305 (15.74%) 16,963	Nymaim 70 / 656 (10.67%) 65,040	Ranbyus 0 / 98 64,400	TinyBanker 0 / 1,733 81,930
Cryptolocker 0 / 3,820 1,108,000	Hesperbot 0 / 15 178	Pushdo 3 / 453 (0.66%) 124,021	Redyms 0 / 11 34	Torpig 2 / 139 (1.44%) 17,610
DirCrypt 0 / 86 420	Kraken	Pushdo TID 0 / 245 6,000	Rovnix 0 / 1 10,000	UrlZone 0 / 127 10,009
Dyre 0 / 850 592,000	Matsnu 244 / 610 (40.00%) 3,346	Pykspa 1 12 / 455 (2.64%) 22,764	Shifu 0 / 11 1,554	Virut

Pre-Registrations: Wordlist-DGAs and short domains cause the most collisions.

Conclusions:

- Breakdown of Pre-Registrations (9,646)
  - Wordlist-DGAs: 8,726 (90.46%)
  - Remainder: 920
    - „Short“ domains (length 5-6): 856 (93.04%)
    - Remainder: 64
      - Accidentally „real“ words: „veterans.kz“
      - Pronounceable „words“: „kankanana.com“
      - „kandilmed.com“
- Basically no collisions with non-Wordlist DGAs or 7char+ domains
  - Using DGArchive for blocking -> very low FP rate for blocking!

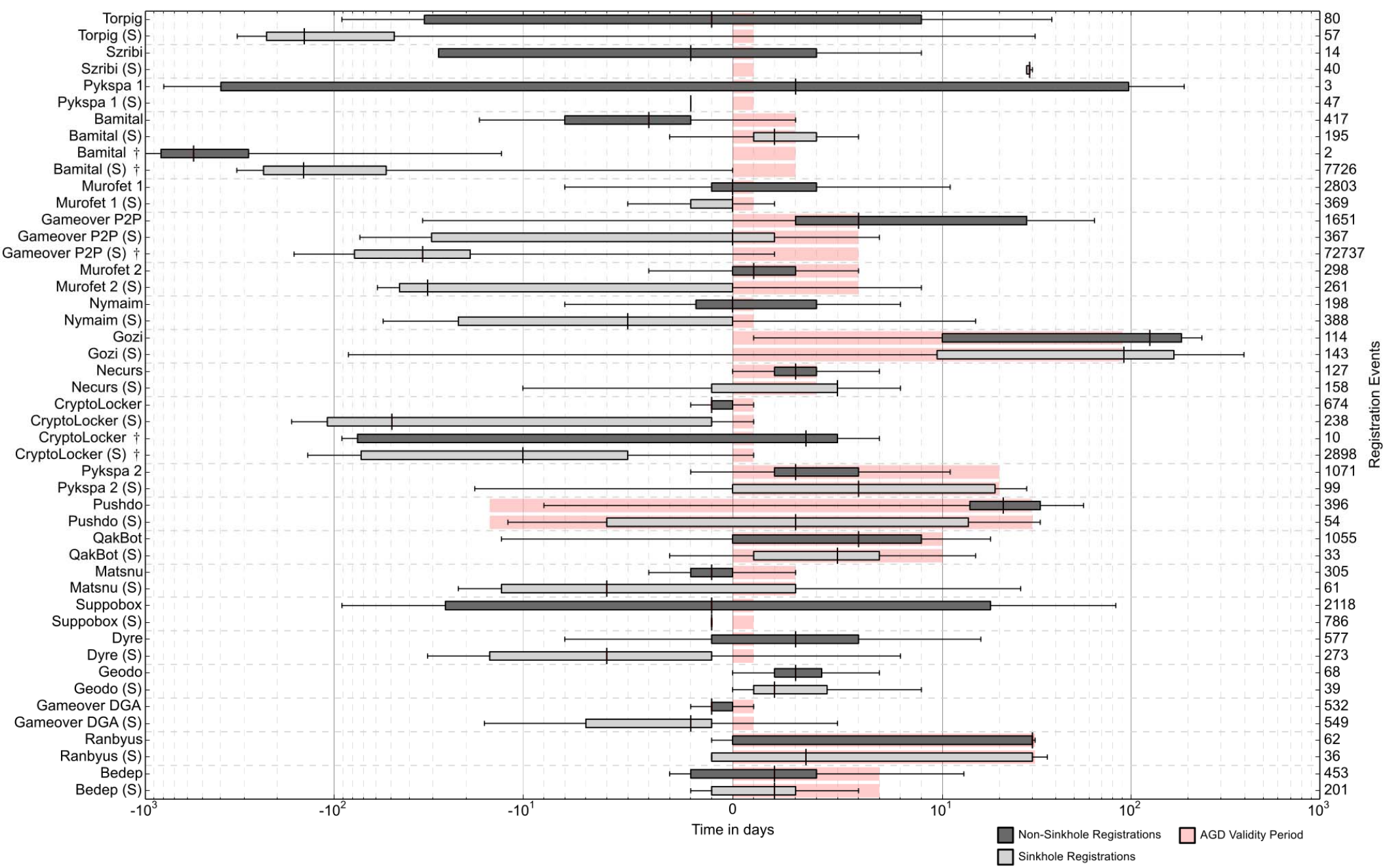
Pre-Registrations: With some exclusions basically non-existent.

DGA Domain Space

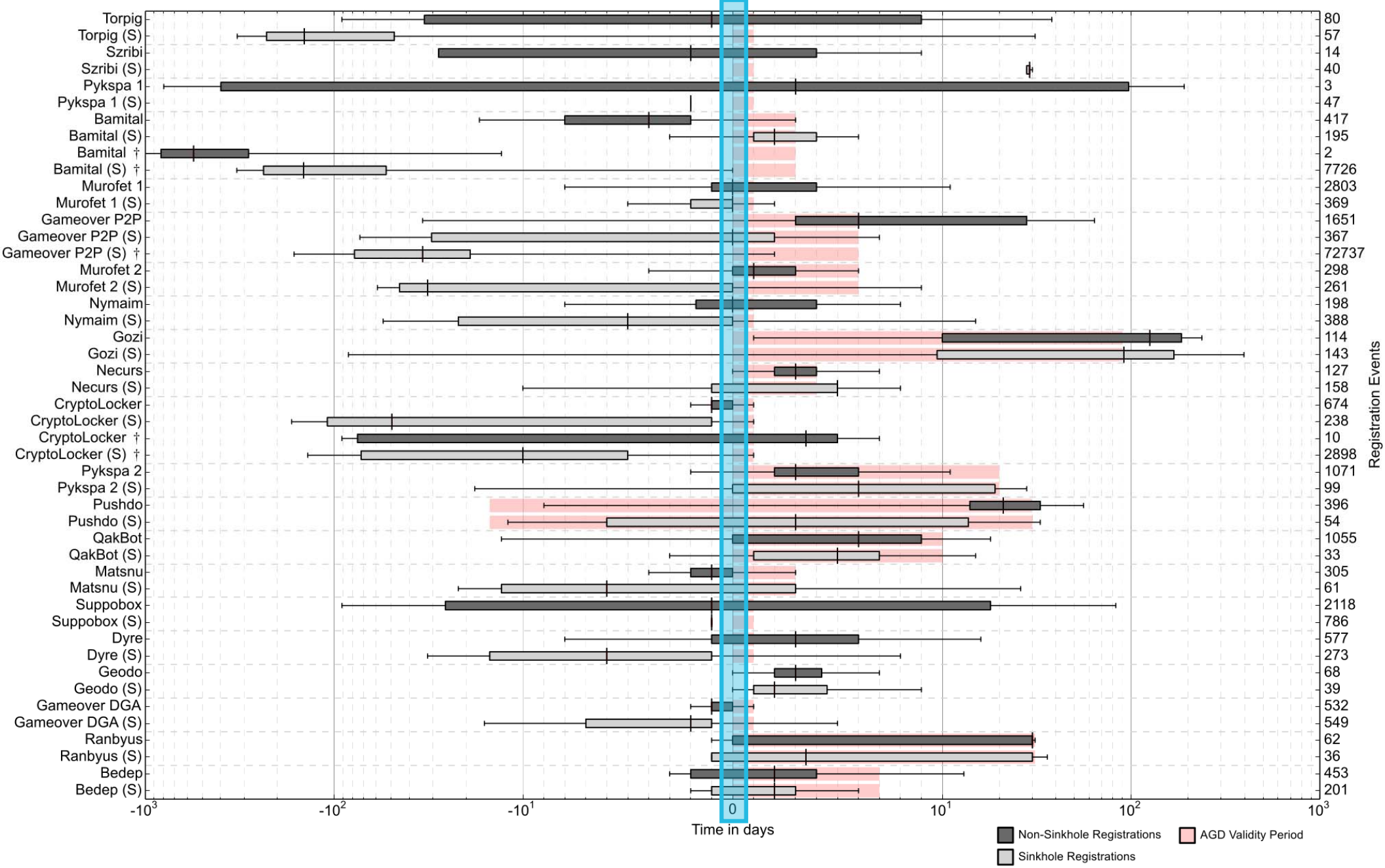
Registration Timings

- Consider time-dependent DGAs
  - Sets of domains have a window of validity!
- What is „registration lookahead“?
  - Relative „offset“ between start of validity and registration time

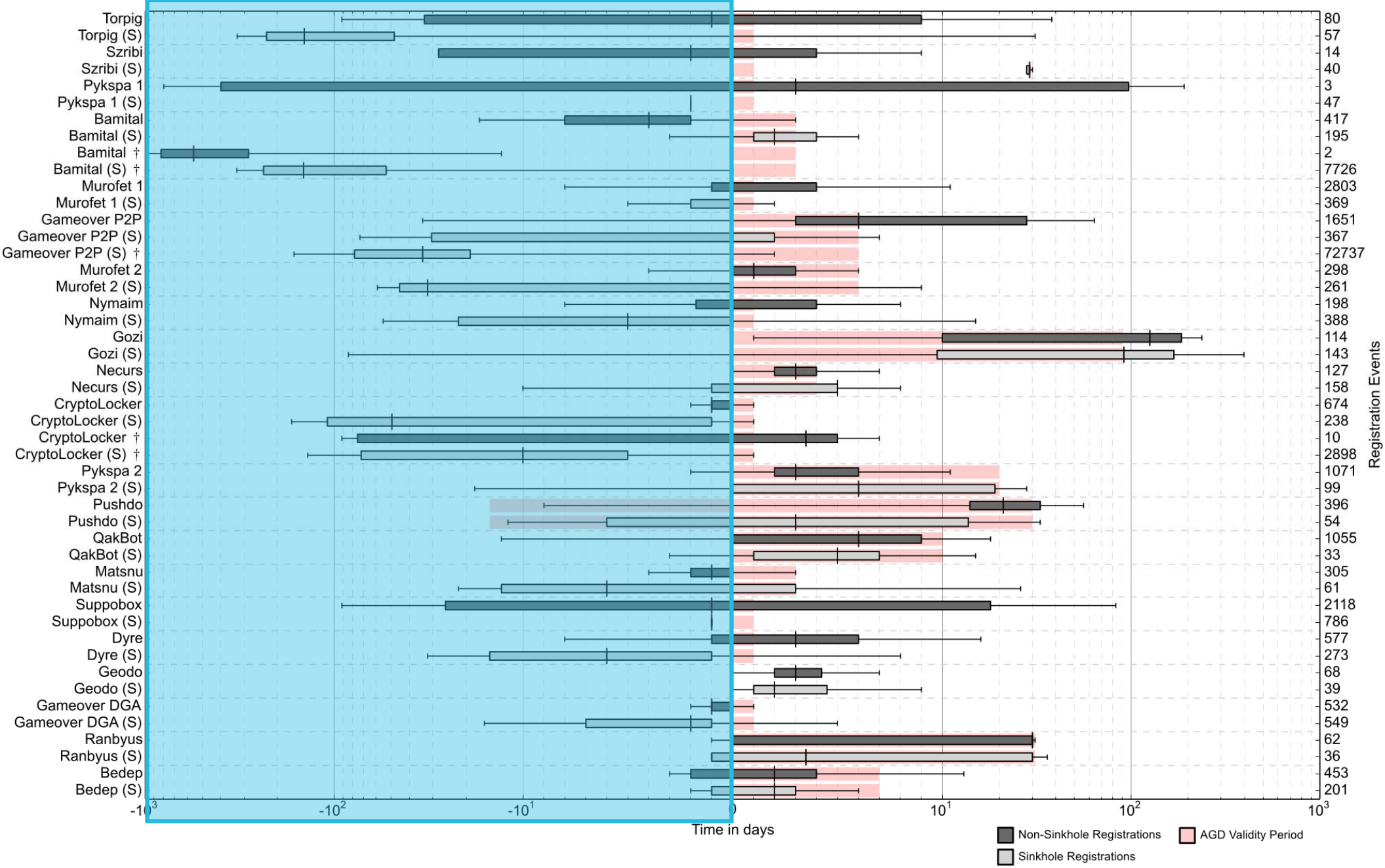
- In the following: Evaluation of registration lookaheads
  - For sinkholes
  - For „non-sinkholes“



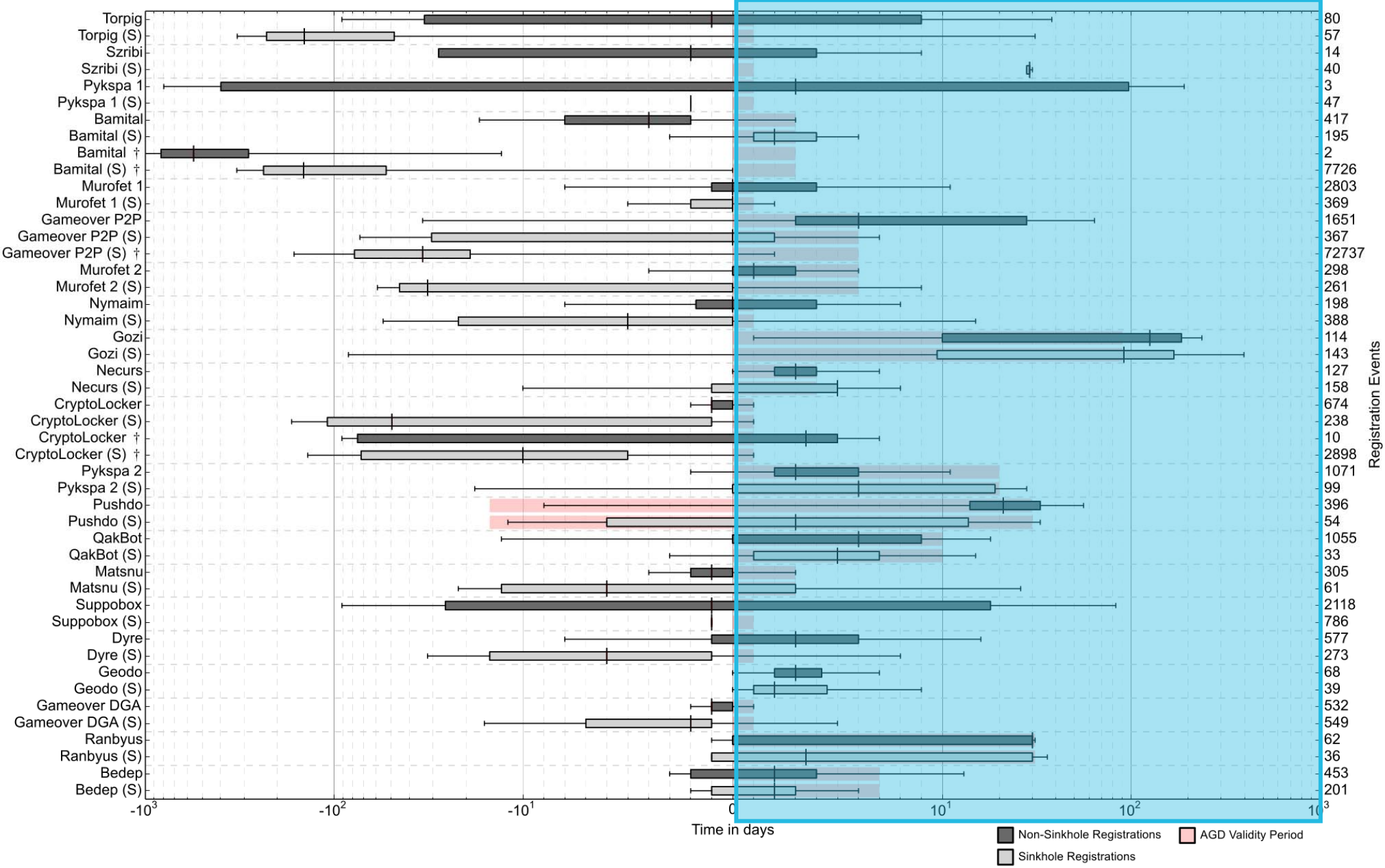
Overview of registration lookaheads



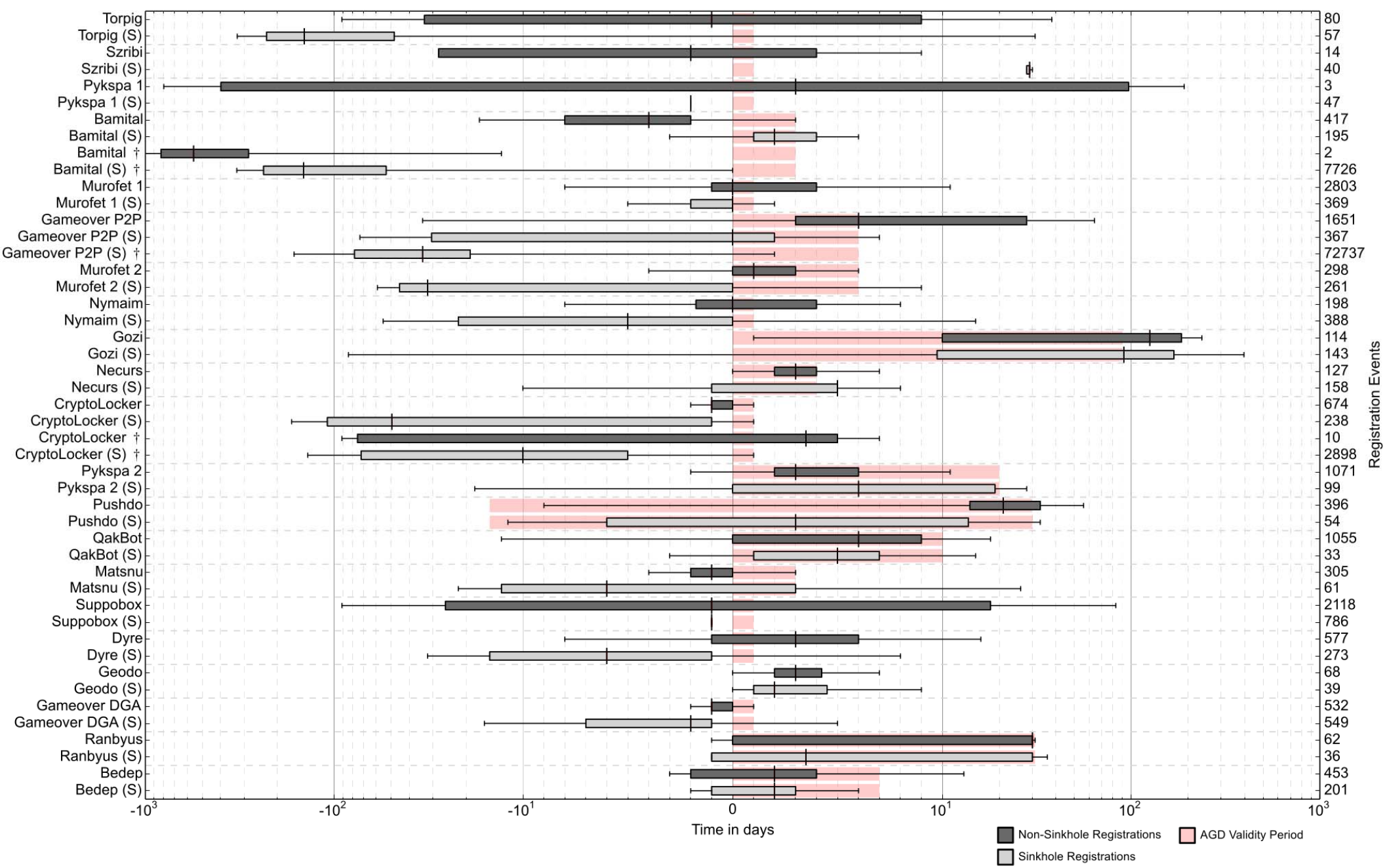
Domains registered ON the first day they became valid in the DGA



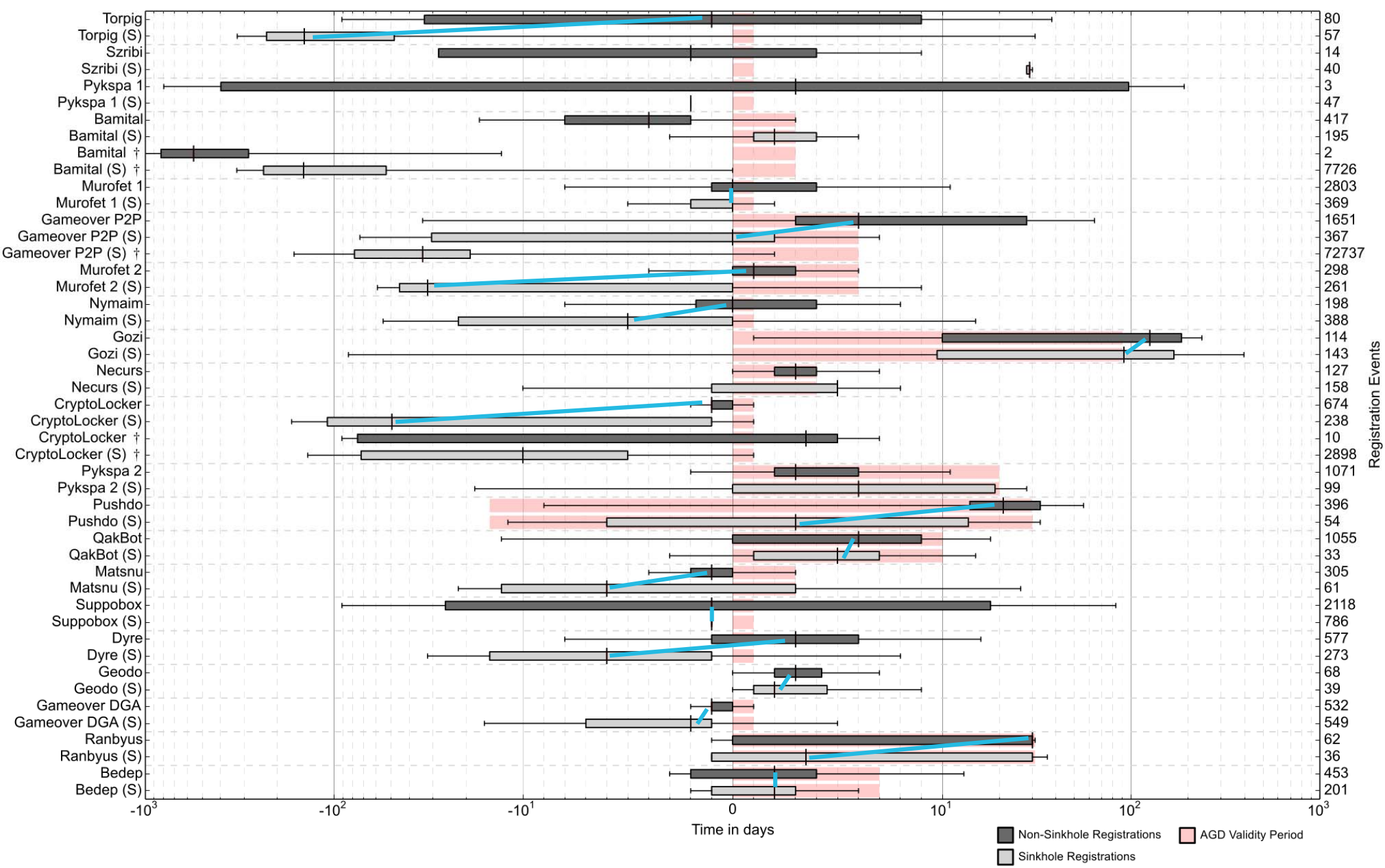
Domains registered BEFORE they become valid in the DGA



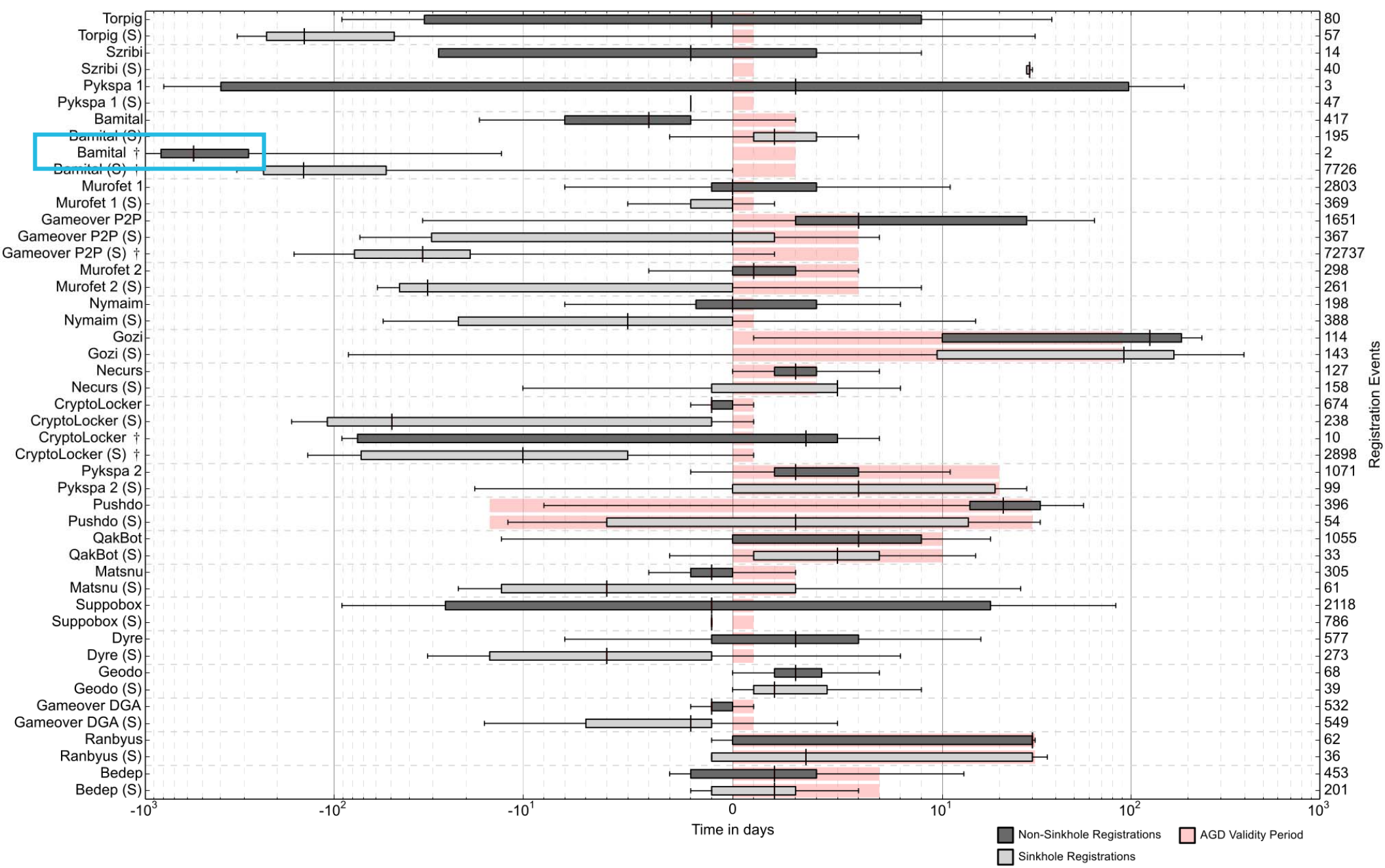
Domains registered AFTER they become valid in the DGA



However, red bars show how long domains REMAIN valid

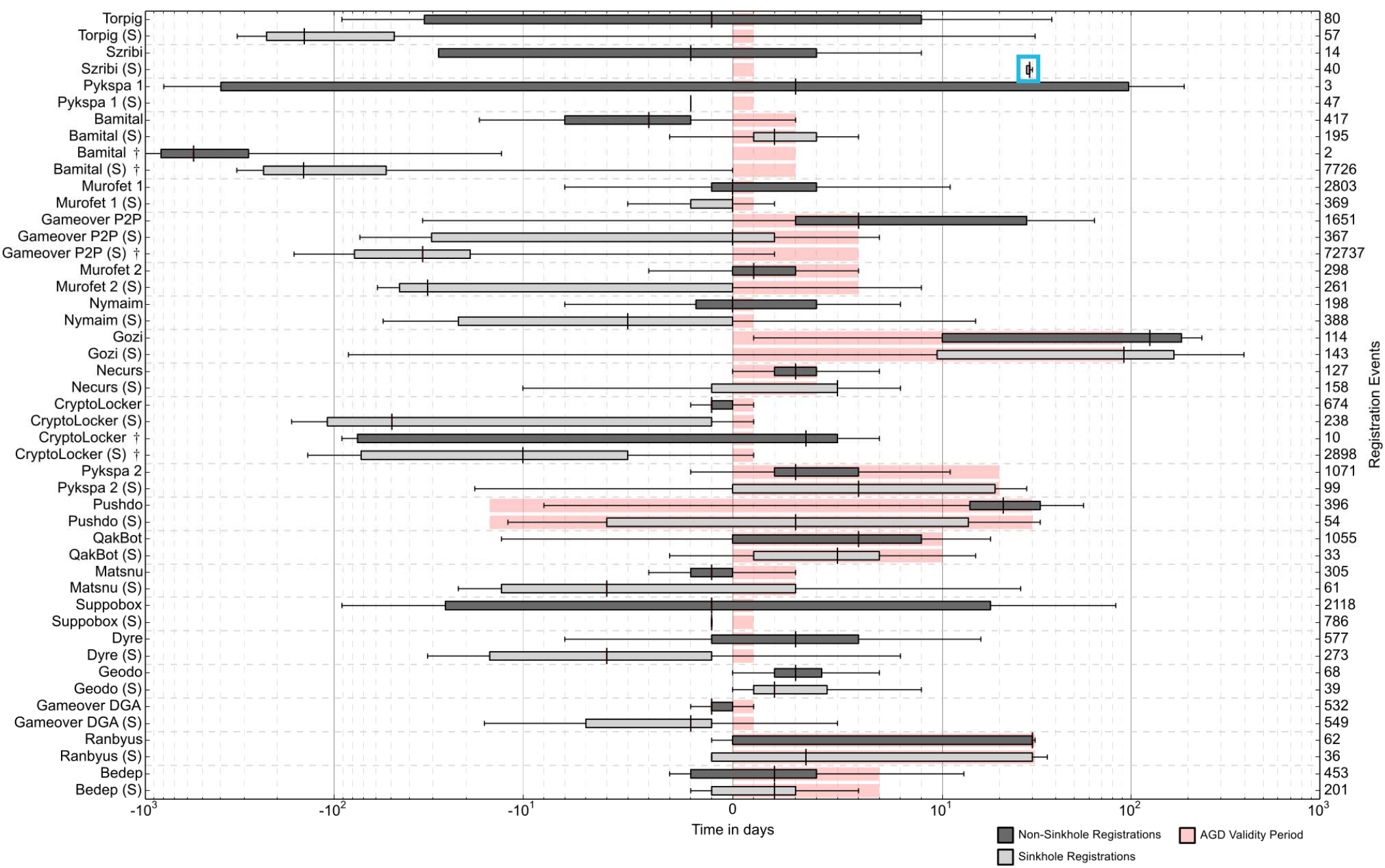


Observation: Sinkholes are often registered earlier than „non“-sinkholes



Observation: Some domains are registered far into the future

Bamital:	Same day registrations for domains 1, 2, 3 years in advance
Nymaim:	Same day registrations for domains 1, 2 years in advance
Murofet:	Same day registrations for domains 1, 2, 3, 4 years in advance



Observation: Reversing Fail by some sinkholer?

32 domains registered exactly one month after validity?

Wrapping up

Conclusion

Conclusion

Wrapping it up

- DGAs are a prevalent malware issue
  - Many active families have a DGA; 25/43 analyzed DGAs surfaced 2013+
- Low collisions between DGAs or with existing benign domains
  - Domains can be blocked with almost no FPs
  - AGDs serve well for identifying malware family based on domains
- With regard to botnet takedowns
  - Botnet operators can be caught off-guard as they register mostly short-termed or „on demand“
  - Rigorously blocking AGDs would have negliable impact on benign users, as there are close to no collisions with „desirable“ domains
  - We are in contact with ICANN, VeriSign, and Nominum to look into how our data can be put to good use at scale

Conclusion

Wrapping it up

- DGArchive
  - Looking for more users / contributors!
  - Request free access: daniel.plohmann@fkie.fraunhofer.de
  - Required: basic proof of identity (e.g. no freemailer) or vetting