The Modality of Mortality in Domain Names

Pawel Foremski, Paul Vixie
Farsight Security, Inc. - www.farsightsecurity.com
{pjf,vixie}@fsi.io

Abstract. Domain names are normatively registered for one or more years, and faithfully renewed thereafter. Domains slated for abusive uses, however, are effectively disposable: they are registered, quickly abused for cybercrime, and abandoned. In this study, we monitor an ongoing data feed of Newly Observed Domains (NOD) to generate a cohort data set, and periodically probe those names to determine what fraction of new domains get suspended by their registrar, deleted by the DNS provider, or effectively "killed" by several well-known blocklists. We then analyze how fast this happens, the most likely cause of domain "death," and how this may vary depending on the TLD involved. The study provides the first systematic study of domain life times, unraveling their complexities and showing the impact of blocklists on the new gTLDs. The results can be used to deploy more-secure DNS policy rules in a computer network.

1. Introduction.

The DNS industry lacks a complete, systematic study on what happens with domains once they have been registered, delegated, and used for the first time. One popular assertion is that most new domain names are malicious [10]. Several commercial and academic studies have addressed similar problems [13-15], but in this paper we combine a real-time stream of passive DNS observations with active DNS measurements, which we believe provides a much more comprehensive view on the life cycle of DNS names.

In this study, we uncover the first 7 days of life of 23.8M domains under 936 TLDs, measured over the course of 6 months from a global Internet perspective. We analyze differences among TLDs and various causes of new domain deaths.

We begin our paper by explaining our measurement methodology in Section 2. Then, Section 3 analyzes and visualizes obtained results. We conclude in Section 4 by summarizing our research and providing the industry with a recommendation for improving DNS security.

2. Measurement methodology.

Our measurement system consists of 4 stages: A) NOD, a data channel that notifies us in real-time about domain names used for the first time on the Internet, B) a scheduler that listens to NOD, looks up the nameservers responsible for each new domain, and schedules periodic measurements, C) a distributed pool of workers that run DNS lookups, and D) a database that collects and aggregates the results. The general method is to repeatedly check, at increasing time intervals, for each domain: Does it still exist in the DNS? Is it still *not* listed in the blocklists we were checking? We describe this process in more detail in the following.

Stage A. NOD is a streaming data service provided by Farsight Security, which operates a global network of passive DNS sensors [1]. The sensors collect DNS cache miss traffic from above recursive resolvers, i.e. they record communication with authoritative nameservers. The collected data is processed and stored in DNSDB, a historical DNS database, as detailed in [2].

When a domain not yet seen in DNSDB is observed, a message is emitted on NOD, notifying listeners of the detection of the first use of a new domain on the Internet. Note that this does not imply detection of newly *registered* domains, as a domain can stay dormant for months after registration - instead, NOD detects the moment of first query for a domain on the Internet. Also, note that NOD detects *effective second-level domains* (SLDs) - i.e. new labels directly under the IANA-accredited TLDs, e.g. *example.com* - and under the *de facto* TLDs - tracked in the Public Suffix List by Mozilla [3], e.g. *example.co.uk*.

As of Q1 2018, on average, more than 2 new SLDs are detected on the Internet each second. For comparison, if we consider Fully Qualified Domain Names (FQDNs, e.g. *ns1.example.co.uk*), more than 150 new FQDNs are detected on the Internet each second using the Newly Observed Hosts (NOH), another channel by Farsight Security.

Finally, note there are wildcard TLDs in the DNS - for instance, each query for any domain under **.pw* or **.ws* will succeed and return the same result. We ignore such TLDs to avoid tracking names in essentially random queries that should normally return an error.

Stage B. The next stage listens to NOD and schedules series of active probes for each new domain. We track each domain directly at 3 locations of the DNS hierarchy:

1. the *delegator* (the TLD): the top-level party that delegates authority over a domain by providing relevant *NS* records in the Authority Section of DNS replies,
2. the *authoritative nameserver* (the DNS hosting provider): the nameserver delegated at 1,
3. popular *URI DNSBLs* (the blocklist providers): zones maintained by Spamhaus [4], SURBL [5], and Swinog [6], which will include a domain if (and only if) it was blacklisted.

We repeatedly query these 3 locations over the course of 7 days, using 20 repetitions with increasing time pauses:

1. the first set of queries is executed immediately after the domain is detected,
2. the second set of queries is executed 1024 seconds after 1 (i.e. after roughly 17 min),
3. the third set of queries is executed 2048 seconds after 2 (i.e. after roughly 34 min),
4. the fourth and each subsequent set of queries is delayed linearly by a constant of 4096 seconds (i.e. by roughly 68 min).

The final 20th set of queries is executed exactly at 629,760 seconds, which equals 7d, 6h, and 56m since the domain was detected.

Stage C. The scheduled DNS queries are run via a pool of workers distributed globally in North America, Europe, and Asia. For each query repetition for a particular domain a worker is chosen at random, i.e. the queries are run around the world, each time from a different location.

For tracking delegators and authoritative nameservers, we send non-recursive DNS queries from workers directly to respective server IP addresses found in the previous stage. For tracking blocklists, we issue recursive DNS queries using a local instance of the *unbound* resolver [7]. In every case, we repeat the query 3 times in case of no reply. All DNS replies are encoded in the *dnstap* format [8], and sent to a centralized location for further processing.

Stage D. The last step is parsing a stream of real-time *dnstap* data: we extract the fields *identity, extra,* and from *message* we extract *response_time_sec* and *query_zone.* Finally, we parse *response_message* as a wire-format DNS response message, and extract the DNS *rcode*.

Thus, our basic data point is a tuple of (*identity*, *extra, response_time_sec, query_zone, rcode*), where *identity* is the worker that run the query, *extra* encodes one of the 3 locations described in "stage B", *response_time_sec* gives us the time when the response arrived, *query_zone* is the domain name, and *rcode* tells us the DNS response code.

The tuples are stored in an sqlite database [9], with one record per domain, in a table with the SQL schema presented in Listing 2.1.

```
CREATE TABLE nod(
  `id` integer primary key autoincrement,
  `domain` text,       -- the domain
  `observed` int,      -- when the domain was observed
  `scheduled` int,     -- when the domain was scheduled for measurements
  `delegator` text,    -- delegator: ip/name/bailiwick
  `deleg_count` int,   -- delegator: number of DNS responses so far
  `deleg_last` int,    -- delegator: timestamp of last DNS response
  `deleg_rcode` int,   -- delegator: rcode of last DNS response
  `deleg_nxd` int,     -- delegator: timestamp of first rcode=NXDOMAIN
  `authority` text,    -- auth NS: ip/name/bailiwick
  `auth_count` int,    -- auth NS: number of DNS responses so far
  `auth_last` int,     -- auth NS: timestamp of last DNS response
  `auth_rcode` int,    -- auth NS: rcode of last DNS response
  `auth_nxd` int,      -- auth NS: timestamp of first rcode=NXDOMAIN
  `dnsbl_count` int,   -- DNSBL: number of DNS responses so far
  `dnsbl_last` int,    -- DNSBL: timestamp of last DNS response
  `dnsbl_rcode` int,   -- DNSBL: rcode of last DNS response
  `dnsbl_listed` int,  -- DNSBL: timestamp of first rcode=SUCCESS
  `dnsbl_detail` text  -- DNSBL: the provider behind dnsbl_listed
);
```

Listing 2.1. SQL schema used for data storage.

We consider a domain effectively dead if any of the following happens:

- The delegator replies to a query for the domain with NXDOMAIN (rcode 3). In such a case, the SQL column *deleg_nxd* will contain the timestamp of the first such event.
- The authority nameserver replies with NXDOMAIN. As in the above case, the column *auth_nxd* will be updated.
- Any blocklist replies to a relevant query with SUCCESS (rcode 0). In such a case, the column *dnsbl_listed* will contain the timestamp of the first such event, and *dnsbl_detail* will identify the particular blocklist.

If more than one of the above happens, we treat the first event as the cause of death. If a domain dies, we do not cancel the scheduled measurements for it, but it cannot go back to the "alive" status, e.g. if the delegator starts replying with SUCCESS rcodes again. Also, note that we always retry DNS queries that fail with SERVFAIL or other rcode, except for REFUSED.

For performance reasons, the records are periodically removed from the database and archived in daily CSV files, after the measurements are finished for a particular domain. The archival process interprets the records by prepending 3 columns to each record: 1) *TLD*, which gives the effective TLD of the domain, 2) *status*, which is either "alive" or "dead-<cause>", and 3) *lifetime*, which for dead domains gives the time since appearance in NOD to the cause of death event. These additional columns make data aggregations easier.

3. Results.

3.1. Collected dataset.

We run our measurements for 6 months, from December 2017 till end of May 2018, which yielded raw data on 24.9M new domains. Out of these, we had to drop incomplete data on 1.1M domains, due to our data collection system being temporarily unavailable during measurements. Thus, in total we used a dataset on 23.8M new domains for the study.

Rank	TLD	Count	% of NOD
1	`com`	6,930,000	29.1%
2	`net`	955,000	4.0%
3	`cn`	900,000	3.8%
4	`tk`	785,000	3.3%
5	`top`	616,000	2.6%
6	`loan`	559,000	2.3%
7	`org`	556,000	2.3%
8	`ga`	531,000	2.2%
9	`ml`	498,000	2.1%
10	`cf`	495,000	2.1%
		12,825,000	53.9%

Tab. 3.1. Top 10 TLDs of Newly Observed Domains

Considering TLD types, 43.7% of the domains were registered under one of 226 ccTLDs (country code TLDs, e.g. *.de*), 35.5% belong to one of 6 legacy gTLDs (generic TLDs, e.g. *.com*), and 20.4% were under one of 607 new gTLDs (recently introduced gTLDs like *.xyz*). The remaining <1% belong to one of 85 internationalized TLDs or 12 sponsored TLDs.

If we consider specific TLDs instead of their types, almost one-third of all domains were registered under *.com*. All other TLDs were roughly at least an order of magnitude less popular, as visible in Table 3.1. Thus, note that TLDs differ considerably in their rates of new domains, and thus in their impact on the DNS.

3.2. General observations.

In Fig. 3.2a, we present the number of new domains observed each day. By the time we plot the data, we already know the ultimate state of each domain, so the blue vs. red color is used to show the proportion of domains that “survived” vs. “died”. Out of 23.8M evaluated domains, 21.6M (90.7%) survived, while 2.2M (9.3%) did not. This rate of "bad" domains was lower than we anticipated, expecting to find that "most new domains are malicious” [10].

The death rate was rather constant and roughly followed a normal distribution, with standard deviation of 1.9%. The number of domains evaluated each day varied due to NOD, but on average we saw 155K per day.

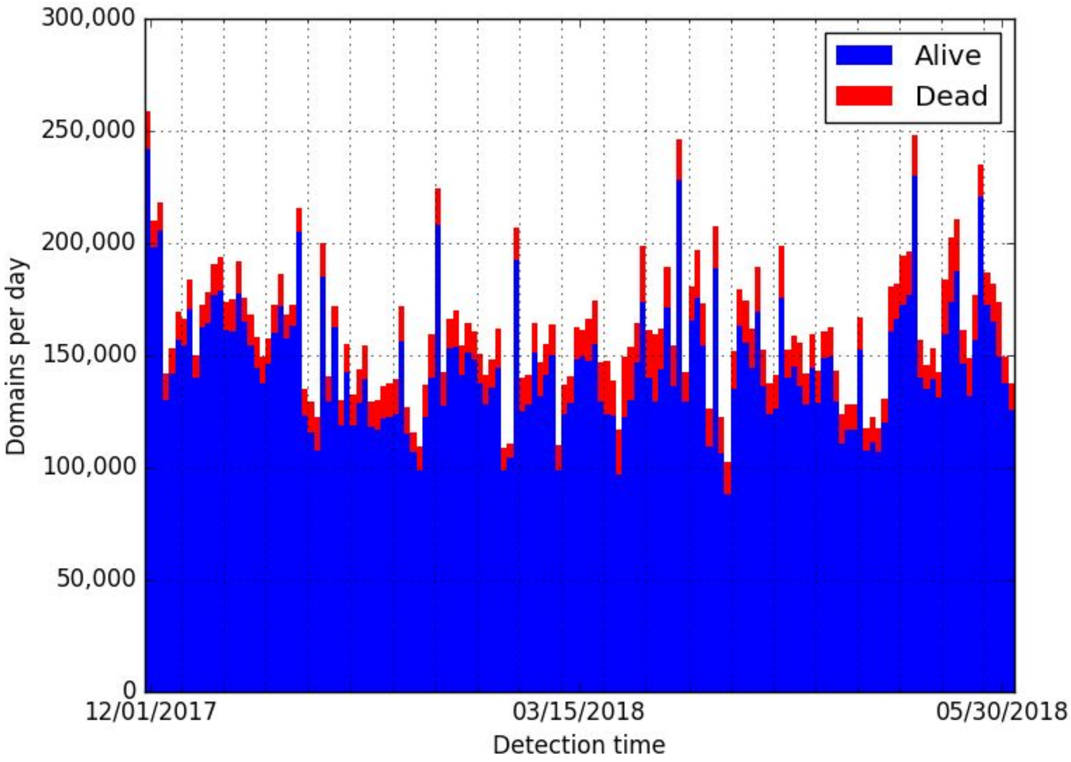

Fig. 3.2a. Number of new domains detected per day, with their state after 7 days.

In Fig. 3.2b, we analyze the 2.2M dead domains to see how fast they died. The plot is a histogram, were each bin represents one of the 20 DNS probes we send for each domain. Note that while the time gaps between the DNS probes are not constant (they grow as described in section 2), we draw the bars equidistantly to make interpretation easier. The bar labels give the *upper* time limits, e.g. "17m" below the first bar means that it represents the queries that finished in under 17 minutes since a domain was observed. The next bar labelled "51m" represents queries that finished between 17th and 51st minute, etc. On the vertical axis we give the death likelihood, i.e. the plot shows empirical probability distribution of the death event among new domains that will eventually die in under their first 7 days of life.

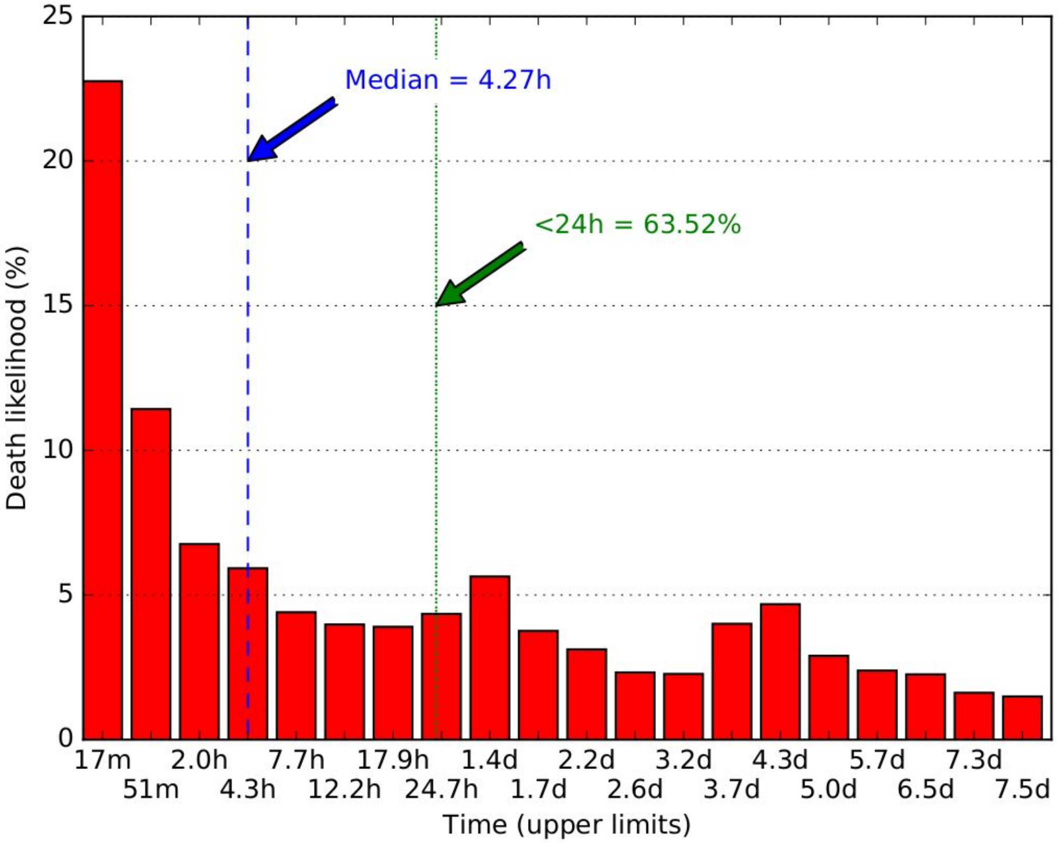

Fig. 3.2b. Death likelihood versus domain age.

Surprisingly, the median longevity marked on the plot with a dashed blue line is just 4.27h, i.e. 4 hours and 16 minutes. In other words, if a new domain is going to die in less than a week, most likely it will die really quickly: in just a few hours, or even just a dozen minutes. Apart from that observation, we identified 3 peaks in the data - roughly 1h, 1.5d, and 4d - which seem the 3 modes in mortality of new domains. Finally, note that although ~63% of domain deaths happen in less than 24h, a non-trivial amount of deaths (~37%) will happen in the following days after a domain is used for the first time.

3.3. Causes of death.

Fig. 3.3a presents the causes of new domain deaths. The horizontal axis gives time, and the vertical axis gives cumulative number of deaths, as percentage of *all* new domains, in order to better explain the net effect of 9.3% dead domains we reported in the previous subsection. Note that blacklisting is clear dominant, being responsible for 6.7% deaths of new domains, which is especially visible for the first 2 days of domain life. Next, we see 2.5% deaths due to action by the registrars: action that is important, but which however needs time to take place. Finally, only 0.2% of new domains deaths were due to the party responsible for DNS hosting, which was expected. It is the easiest for the bad actor to control this risk by running running their own dedicated nameservers, although this may increase their risk of being readily identified *en masse* via passive DNS methods.

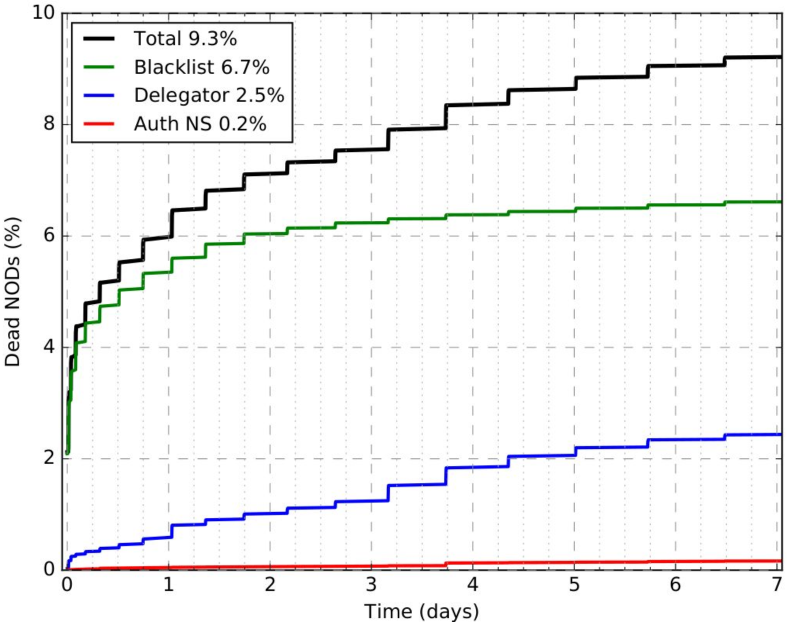

Fig. 3.3a. Causes of new domain deaths and their timing.

Note that we consider only the first cause of death for our study, which means the percentages we show in Fig. 3.3a are influenced by how fast a certain party can effectively kill a domain. For that reason, in Fig. 3.3b we analyze the intersections between various death events when considered independently of each other. That is, we check how often a new domain will experience 1, 2, or all of the 3 death events we described in section 2. Surprisingly, these events seem complementary, with the biggest overlap of 4.0% for a new domain being both blacklisted and deleted at the TLD level.

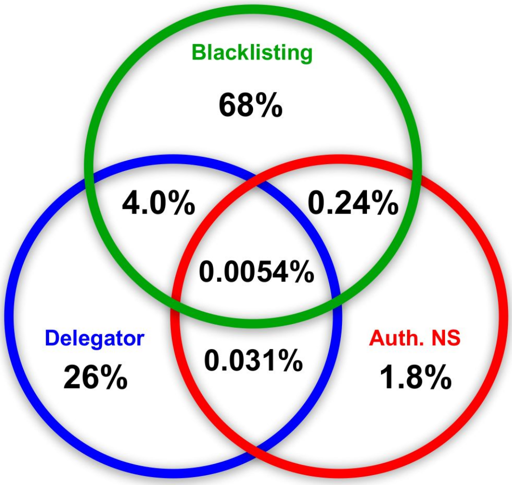

Fig. 3.3b. Intersection of independent causes of death (the diagram is not to scale).

Let us analyze timing of each “cause of death” in more detail. Figures 3.3c-3.3e present the death likelihood separately for each cause, which clearly shows considerable differences, and explain the sources of peaks in data we saw in Fig. 3.2b.

From Fig. 3.3c, it is apparent that blacklists kill new domains fast, in under an hour, and that the histogram resembles an exponential distribution. The histogram for delegators presented in Fig. 3.3d has a completely different probability distribution, with peaks around 1h, 1.5d, and 4d, which we speculate is an artifact of automated procedures for domain deletion. Also, note that the median time of delegator action is 3 days and 4 hours (76h).

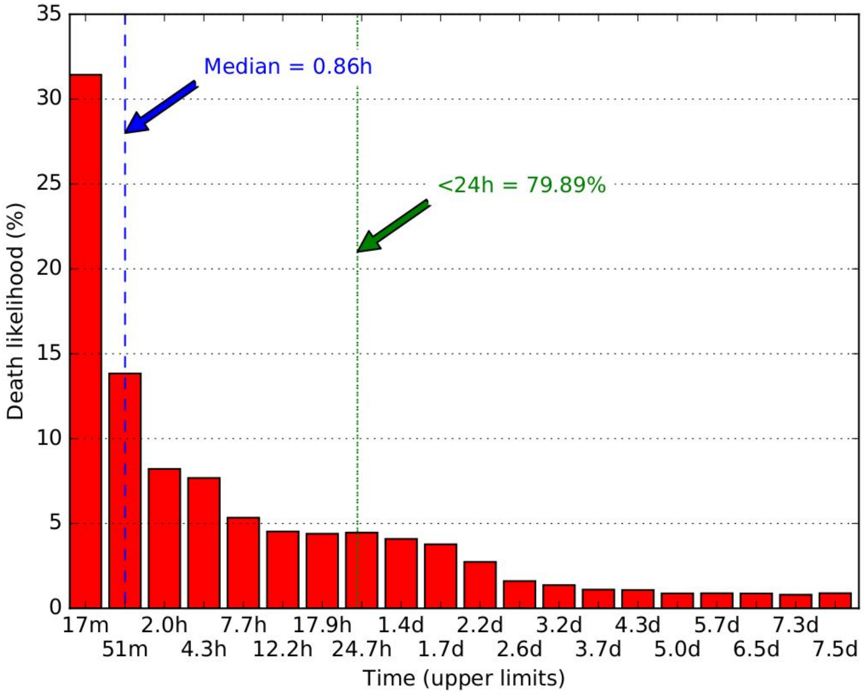

Fig. 3.3c. Deaths due to DNS blacklists: likelihood vs. domain age.

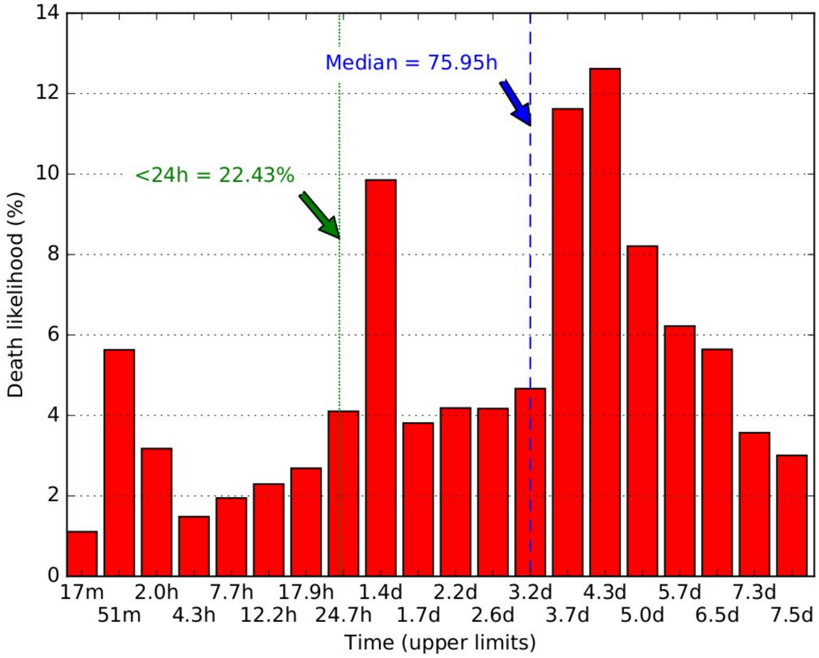

Fig. 3.3d. Deaths due to delegators: likelihood vs. domain age.

Fig. 3.3e for authoritative nameservers basically show a huge peak around 4 days, which we found to be caused by seemingly random domains under the *.to* TLD. If we ignore this TLD, the death likelihood is much more uniform.

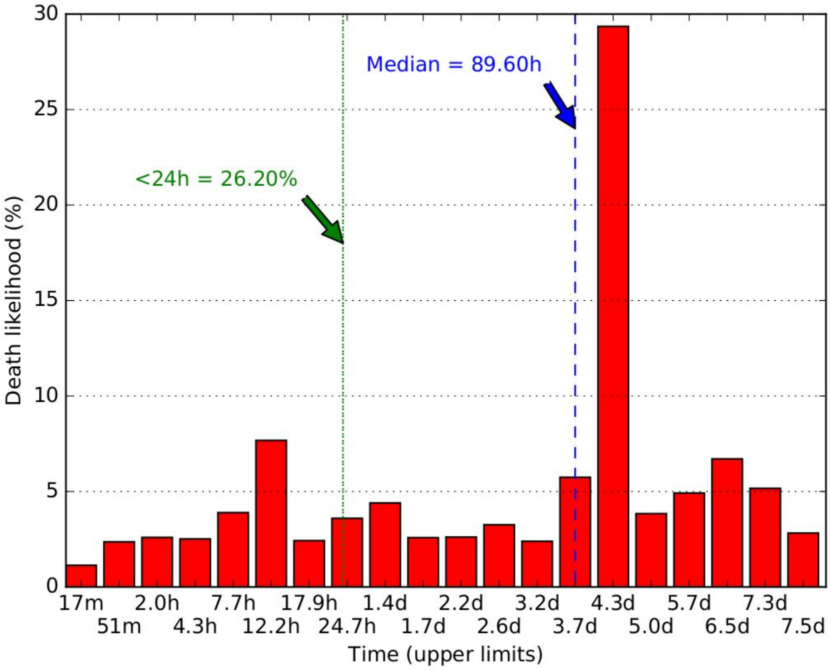

Fig. 3.3e. Deaths due to authoritative nameservers: likelihood vs. domain age.

3.4. Impact of TLD type.

Finally, let us see if the probability that a new domain will die quickly depends on the TLD. In Fig. 3.4a, we see that domains under the new gTLDs are on average roughly thrice as likely to experience early deletion as domains under the legacy gTLDs. There is a considerable difference in the causes of death: for the new gTLDs, the major cause of death is blacklisting, whereas for the legacy TLDs it is the registrar action. Note that we are considering the averages for all TLDs in a group, which means specific TLDs can exhibit more extreme characteristics, as will be presented below. Also, note that the IDN and sponsored TLDs experience relatively low rates of early deletion compared with the more popular TLD types.

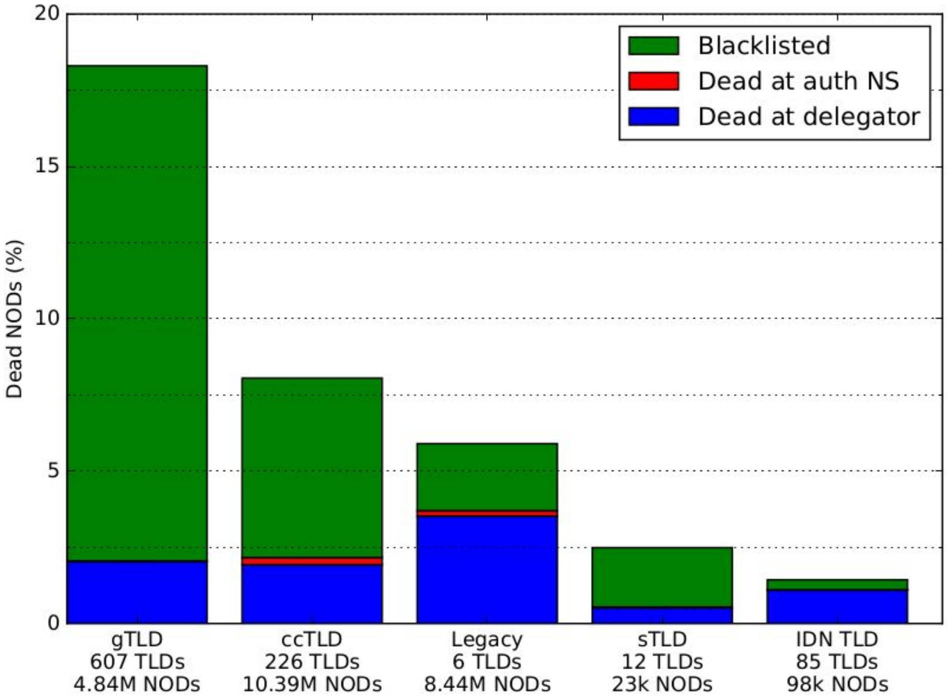

Fig. 3.4a. Impact of TLD group on new domain death rate.

3.5. The new gTLDs.

In Fig. 3.5a, we present the top 25 new gTLDs that have the highest new domain death rates. We skip the gTLDs with low impact by considering only those with more than 5K domains in our dataset, i.e. roughly more than 1 new domain detected per hour.

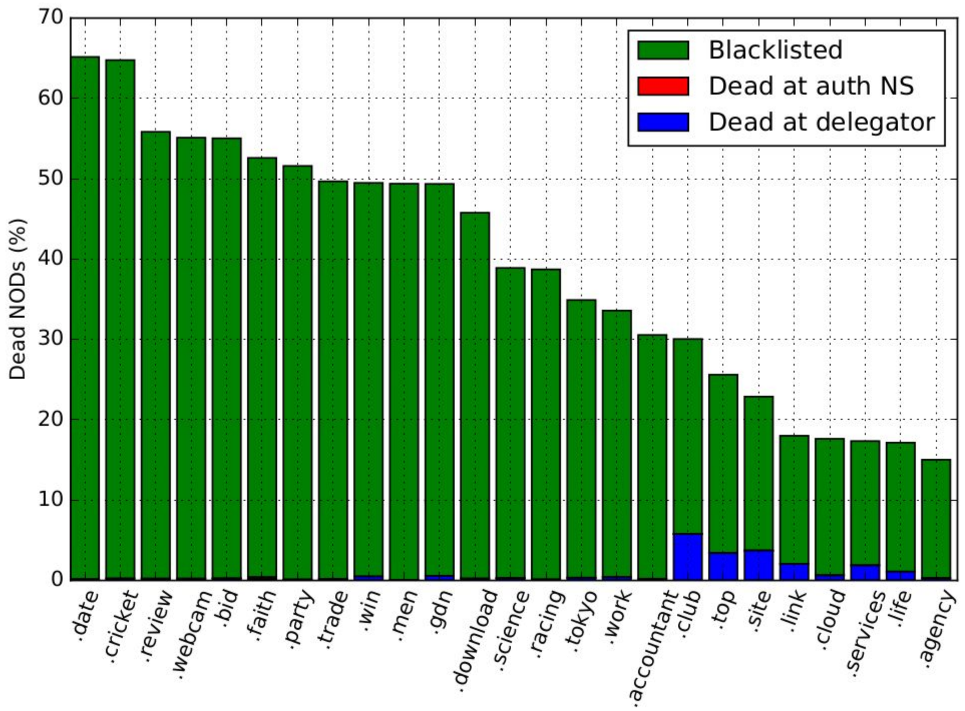

Fig. 3.5a. New gTLDs with the highest new domain death rates.

We found many gTLDs with high new domain death rates, for example *.date* and *.cricket*, with >65% death rates. Almost all of those domain deaths were due to blacklisting, with little impact of the delegator.

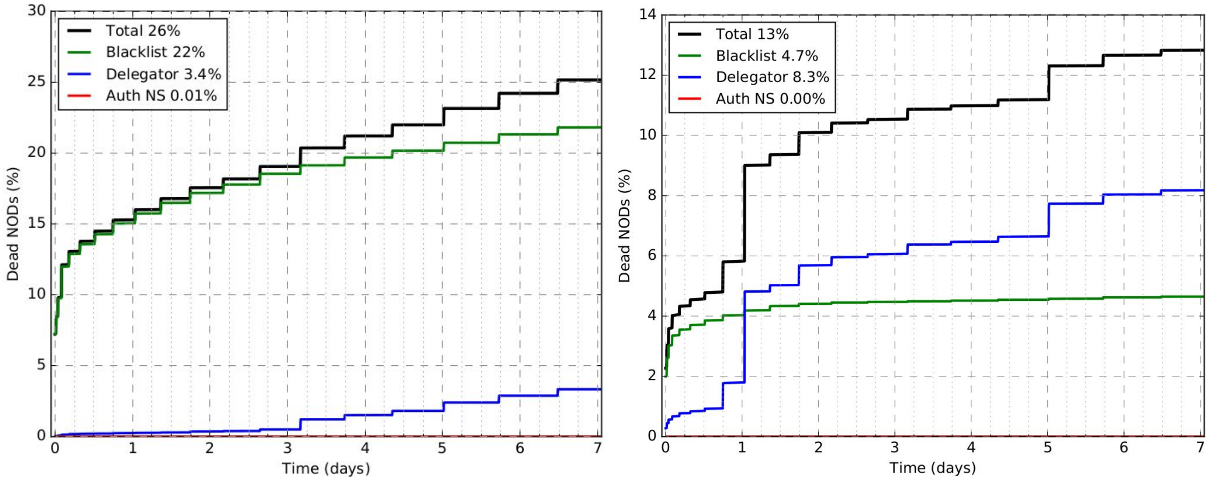

Fig. 3.5b. Causes of death for 2 gTLDs: *.top* (left) and *.xyz* (right).

In Fig 3.5b, we analyze the causes of death for 2 illustrative gTLDs, *.top* and *.xyz*. We see that one in four *.top* domains die quickly, mainly due to blacklisting, and that the registrar does not take much action until the third day since the domain is observed. On the other hand, for *.xyz* we see that the registrar deletes almost twice as many of its domains as blacklists do, and that majority of these deletions happen on either the 1st or 5th day of the domain life, which suggests automated procedures.

We conclude that the new gTLDs are highly susceptible to abuse. Many gTLDs cost less than $1 to register, which makes them easily available for cybercrime [11]. Such a low price means low profits to the registrar, and in consequence little budget for mitigating domain abuse [12].

3.6. The country-code TLDs.

Fig. 3.6a demonstrates that some ccTLDs are abused and experience high new domain death rates. The TLDs managed by Freenom - *.gq, .ml, .cf, .ga,* and *.tk* - experience the highest death rate of 18.7% on average. Note that if considered together, they are the world's second most popular source of new domains (total 11.6%, see Table 3.1). Those TLDs offer free domains, but unfortunately their abuse may damage the reputation of all domains under such a TLD. Surprisingly, one of the top TLDs in this plot is *.eu*, connected with the European Union. We speculate the reason is that it is cheap, people generally trust it, and for a long time it had no domain whois. Note there is a clear gap in the death rates among ccTLDs. The ones being used for their original purpose, as a TLD for use by citizens or nationals of that country, experience new domain death rates below 5%. The 2 notable exceptions are *.cn* and *.us*, with death rates of 18.5% and 7.7%, respectively.

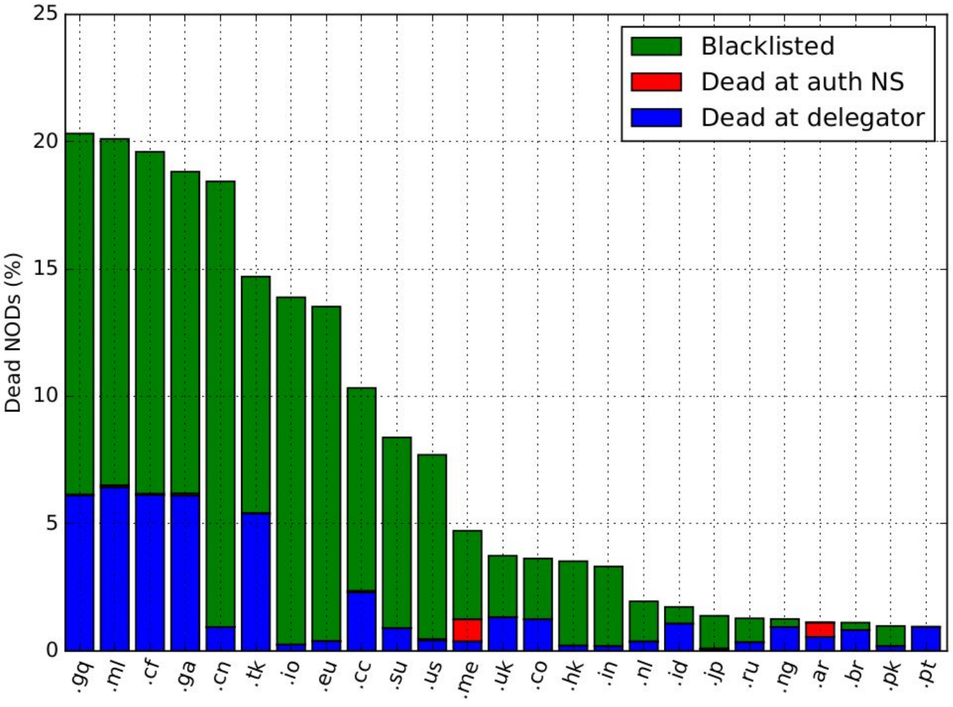

Fig. 3.6a. Country-code TLDs with the highest new domain death rates.

In Fig. 3.6b, we analyze the causes of death for 2 misused ccTLDs *.tk* and *.cc*. In both cases, we see the leading role of blacklists in preventing domain abuse, with some action of the registrar. For *.tk*, we see a difference in the death rates before and after the 1st day.

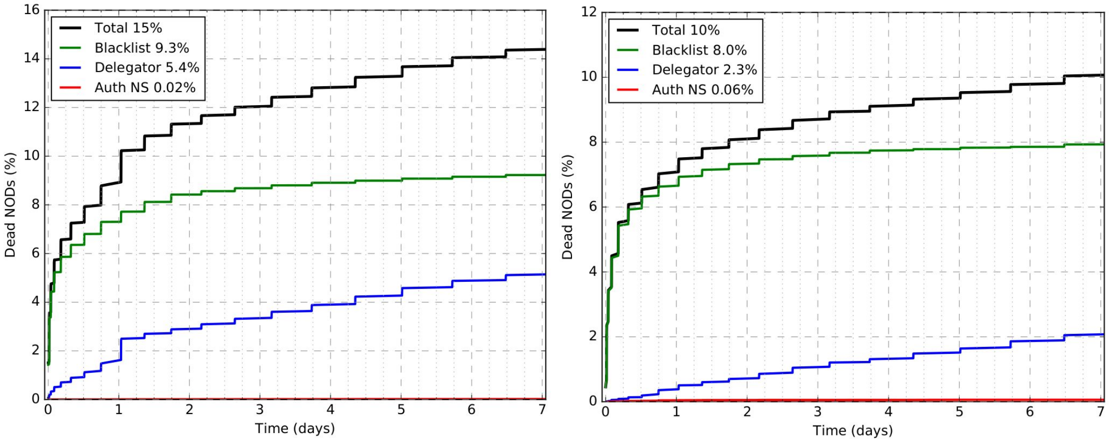

Fig. 3.6b. Causes of death for 2 ccTLDs: *.tk* (left) and *.cc* (right).

In summary, we think ccTLDs are in general well protected from abuse, with a very important exception of some misused ccTLDs, which are no longer associated with a specific country.

3.7. The legacy gTLDs.

Last but not least, we analyze the legacy gTLDs. In Fig. 3.7a, we present the TLDs with more than 5K new domains. Note that *.net* is twice as likely to experience an early domain death compared with *.org*, which seems much better protected from abuse. The largest source of new domains on the Internet, the *.com* TLD, experienced just 5.8% of new domain deaths.

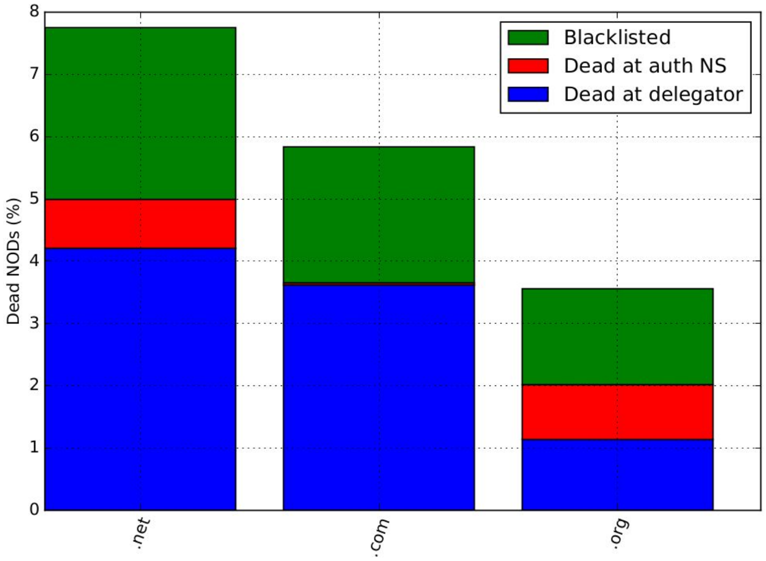

Fig. 3.7a. Legacy gTLDs and their death rates.

In Fig. 3.7b, we compare *.com* with *.net.* Surprisingly, only 2% of new domains under *.com* are blacklisted, and eventually the registrar plays more important role, especially between the 3rd and 5th day. For *.net*, we also see the leading role of the domain delegator. The vast majority of the 0.8% quick domain deaths detected at the authoritative DNS level were due to the Microsoft Azure cloud platform.

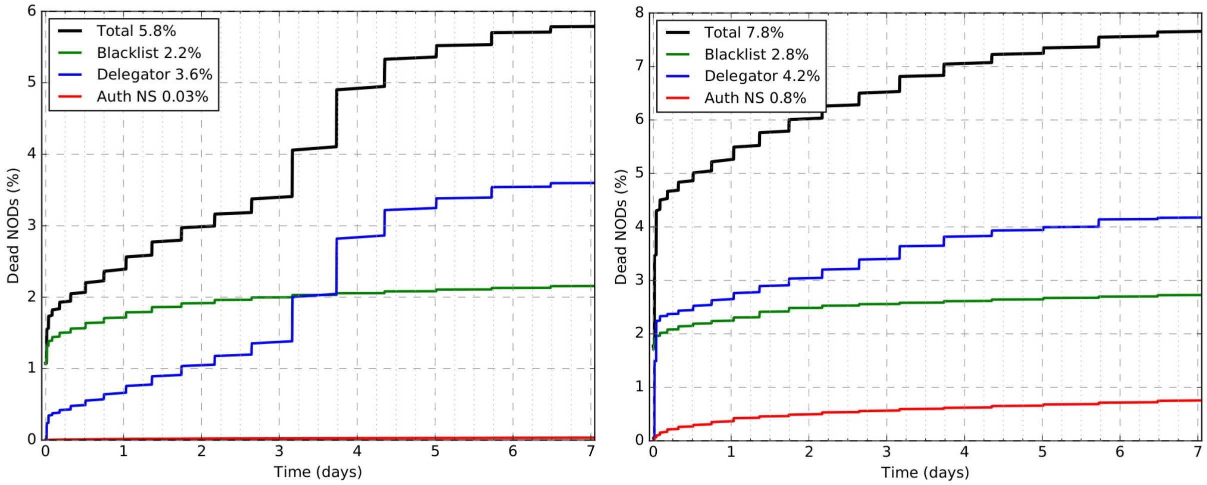

Fig. 3.7b. Causes of death for 2 legacy TLDs: *.com* (left) and *.net* (right).

4. Conclusions.

We presented a systematic study on the lifetime of newly observed domain names. We found that on average 9.3% of new domains effectively died in their first 7 days, with a median time of just 4 hours and 16 minutes. This, however, differs considerably for various causes of deaths and TLDs. Blacklisting, which was responsible for 6.7% deaths, in majority cases blocked domains in under 1 hour, whereas DNS registrars and hosting providers needed over 3 days. We also found great influence of the TLD. In general, the new gTLDs exhibited roughly thrice as many quick deaths as the legacy gTLDs, with a dozen cases where more new domains died than survived their first week. Surprisingly, the largest TLD on the Internet, *.com*, exhibited only 2% of new domains being blacklisted, with 3.6% of new domains being deleted by the registrar.

We conclude that TLDs can be roughly divided into 2 classes: those that experience a high percentage of quick deaths among new domains (i.e., well above the average 9.3%), and those that do not. In the former case, it is usually associated with: a) registration under a new gTLD or under a misused ccTLD, b) cheap price, and c) little domain deletions at the TLD level.

Our research highlights the need for a secure DNS policy in a computer network. On one hand side, presumably the vast majority of 9.3% new domains will be used for cybercrime. On another hand, even blacklists need time to stop a potential incident, and they did not contain 26% of new domains that experienced a quick deletion at the TLD level (see Fig. 3.3b). Thus, a sensible DNS policy should block access to new domains for a few hours since detection, a few days, or even a week for maximum protection.

References.

1. F. Weimer. "Passive DNS replication." *FIRST conference on computer security incident*. 2005.
2. R. Edmonds. "ISC passive DNS architecture." *Internet Systems Consortium,* 2012.
3. Mozilla Foundation. “Public Suffix List.” URL: https://publicsuffix.org/learn/, 2018.
4. The Spamhaus Project. “The Domain Block List.” URL: https://www.spamhaus.org/dbl/, 2018.
5. SURBL. “Combined SURBL list.” URL: http://www.surbl.org/lists#multi, 2018.
6. SWINOG. DNSBL at uribl.swinog.ch, 2018.
7. NLnet Labs. “Unbound.” URL: https://www.unbound.net/, 2018.
8. R. Edmonds, P. Vixie. “dnstap.” URL: http://dnstap.info/, 2018.
9. SQLite. URL: https://www.sqlite.org/, 2018.
10. P. Vixie. "Taking Back the DNS." URL: http://www.circleid.com/posts/20100728_taking_back_the_dns/
11. TLD List. URL: https://tld-list.com/, 2018.
12. M3AAWG. “Recommendations for Preserving Investments in New Generic Top-Level Domains (gTLDs).” URL: http://www.m3aawg.org/gTLDInvestments, 2018.
13. The Spamhaus Project. “The World’s Most Abused TLDs.” URL: https://www.spamhaus.org/statistics/tlds/, 2018.
14. S. Hao, N. Feamster, R. Pandrangi. “Monitoring the Initial DNS Behavior of Malicious Domains.” *ACM Internet Measurement Conference IMC’11.* 2011.
15. S. Hao, M. Thomas, V. Paxson, N. Feamster, C. Kreibich, C. Grier, S. Hollenbeck. “Understanding the Domain Registration Behavior of Spammers.” *ACM Internet Measurement Conference IMC’13.* 2013.