Associated Domain Analysis Using Public Data

ICANN Office of the Chief Technology Officer

Sam Cheadle
Siôn Lloyd
Carlos Hernández-Gañán
Samaneh Tajalizadehkhoob
OCTO-044
18 September 2026

TABLE OF CONTENTS

EXECUTIVE SUMMARY 3

1 INTRODUCTION 3

2 METHODOLOGY 5

2.1 Data Sources 5
2.2 Detection Methods 6
2.2.1 Stage-1 Grouping 7
2.2.2 Stage-2 Lexical Filtering Analysis 8
2.3 Metric Calculation 9

3 RESULTS 10

3.1 Overall Analysis 10
3.2 Threat-Type Analysis 13
3.3 RBL-Type Analysis 15
3.4 Exclusion of a DGA Campaign Subset 16

4 DISCUSSION 16

5 CONCLUSION 19

6 REFERENCES 20

APPENDIX A 21

A.1 Batch Detection 21
A.1.1 Algorithm 21
A.1.2 Filtering Criteria 21

APPENDIX B 22

B.1 Random Forest Lexical Similarity Model 22
B.1.1 RF Training Data 22
B.1.2 Model Features 23
B.1.3 Model Training 23
B.1.4 Model Testing (Temporal Holdout) 24

This document is part of ICANN's Office of the Chief Technical Officer (OCTO) document series. Please see the OCTO publication page for a list of titles in the series. If you have questions or suggestions on any of these documents, please send them to octo@icann.org.

This document supports ICANN's strategic goal to evolve the assessment of, and responsiveness to, technological developments. It is part of ICANN's strategic objective to collaborate with relevant stakeholders to evolve the Internet's unique identifier systems.

Executive Summary

Large-scale abuse campaigns are often linked to bulk domain name registrations — the practice of rapidly registering high volumes of domains — often through a single registrar. Recent community efforts have highlighted the value of associated domain (AD) analysis for abuse mitigation purposes.

This report describes preliminary analysis of AD methods that rely on public data only. Results indicate that 1) over half of all reported maliciously registered domains in generic top-level domains (gTLDs) were registered (or used) in combination with other domains, consistent with the widespread practice of bulk registration for abuse campaigns; 2) a subset of reported abuse can be detected earlier using some AD methods, compared to reliance on standard security feeds alone; and 3) previously unknown suspected abuse can often be identified via 'expansion analysis' based on AD methods.

In summary, this report highlights the potential for AD analysis that uses public registration data only for monitoring bulk registration patterns (historical analysis) and the proactive detection of domains at high risk of abuse. A set of multiple, overlapping AD techniques were developed for gathering strong evidence of association between recently registered domains. These techniques have been documented to enable transparency and to help the community extend the work.

1 Introduction

Background. While the Domain Name System (DNS) is a foundational component of Internet infrastructure, the openness and scalability of the DNS has also made it a persistent target for abuse. Malicious actors routinely exploit the DNS by registering large numbers of domains in rapid succession—a practice referred to as *bulk registration*—to facilitate phishing campaigns, malware distribution, and other forms of cybercrime. This practice is often advanced through the use of automation, for example via application programming interfaces (APIs) that enable attackers to scale their operations with minimal overhead [1].

The identification and characterization of bulk domain registrations is of high importance for abuse mitigation. Detecting such patterns can provide early warning of malicious campaigns and, in some cases, enable investigation before domains become weaponized. A wide body of abuse-detection research has focused on the potential value of features available at, or close to, the time of domain registration. In early work, Felegyhazi et al. [2] explored the value of clustering newly registered domains, leveraging information contained in DNS zone files and WHOIS registration data. Other work highlighted the importance of detailed registration records, including personal registrant-level information, such as name, address, email, and phone number, for the detection of large-scale ongoing abuse campaigns [3]. A number of promising approaches have been developed by European country-code top-level domain (ccTLD) registries in this area [3], [4], [5], [6].

Motivation. Changes in data protection frameworks and registration-data availability have influenced how defenders conduct large-scale domain abuse analysis. Accordingly, the analysis described in this report focuses on publicly available technical registration and DNS

infrastructure data only, and does not rely on registrant contact information, payment data, or account-holder identifiers.

Publicly available technical registration and DNS infrastructure data includes sponsoring registrar, authoritative nameservers, and creation timestamps, to infer relationships among domains. Recently published work from OCTO-SSR highlighted the potential value of these *public* features [7], exploring the necessary data processing and filtering stages required to generate high-confidence outputs.

Analysis of public registration data is, in many cases, sufficient for associated domain discovery. Due to the operational reality facing attackers, overwhelming preference is given to automated, high-volume registration methods [1] in order to rapidly scale up and maintain operations. This behavioral preference creates observable patterns — notably, time-bound spikes in registration activity sharing registrar (or reseller), and the DNS infrastructure (such as authoritative nameservers) — which can be leveraged by defenders to identify coordinated malicious activity.

Public vs Private data. During 2025, ICANN gathered community feedback on a proposed Policy Development Process (PDP) on DNS Abuse Mitigation [8], [9]. One central proposal was that associated domain “pivot” analysis should be conducted internally by registrars. The proposal involves strengthening requirements for registrars to rapidly pivot on data such as account holder ID once a confirmed report of abuse is received. Theoretically, this can aid proactive abuse detection and investigation of all associated domains, rather than just the single (seed) domain with evidence of abuse.

Analysis of internal registrar data is valuable because registrars hold detailed customer and payment ID records, which can be used to rapidly identify associated domains. However, because this information is privately held and accessible only to the owning organization, there are certain advantages to the public registration features listed above, namely:

- Reproducibility - important for future auditing and transparency
- Wide-scale analysis - across multiple registrars and TLDs

Privacy considerations. The analysis described in this report uses publicly available technical registration and DNS infrastructure data only and does not rely on registrant contact information. Associated-domain matches are intended to support investigative analysis and should be interpreted in context. Depending on context, some technical data elements may nevertheless constitute or relate to personal data under applicable law where they can be linked to an identifiable natural person.

Current approach. In this study, we build on past work [10] using an approach to feature selection and grouping guided by past evidence, namely that bulk registrations are typically associated with identical registrar-nameserver or IP address-nameserver groupings. This *a priori* assumption enables more efficient clustering analysis by decreasing the search space, as well as improving the understandability and transparency of the algorithm.

AD analysis incorporates a series of methods for rapidly pivoting from reported abused domains, with the goal of identifying previously unknown malicious domains. These high-risk domains can then be reviewed by registrars and suspended if the evidence or probability of abuse is sufficiently strong.

The overall goal of the study is to explore the feasibility and value of AD analysis for

- Facilitating proactive abuse detection
- Improving wide-scale abuse campaign monitoring and reporting

While early identification of suspicious groups of domains can enable defenders to flag, monitor, or even preemptively block malicious domains before they are weaponized, improved reporting of the prevalence of wide-scale bulk registration campaigns can provide the community with valuable insight into when and where abuse is happening.

Summary of findings. Based on our analysis, we find that:

- Over half of all reported maliciously registered domains in gTLDs were registered (or used) in combination with other domains, consistent with the widespread practice of bulk registration for abuse campaigns.

- Some abused domains can be caught earlier (4.6% of the sample) under current data processing constraints. Modeling a hypothetical scenario with ideal data collection conditions indicates that early detection of abused domains could be improved.

- A large number of previously unreported high-risk or suspicious domains are identified via 'expansion analysis' wherein more than two new associated domains are found for each reputation block list (RBL) seed domain. These high-risk domains were not flagged by any RBL provider during the analysis period. Examination of the characteristics of this expanded set of high-risk domains revealed a large cluster of algorithmically generated domains within a single TLD.

Structure of the report. Following the introduction, the methodology section provides data sources and detection details. This is followed by results summarizing performance across multiple metrics (RBL coverage, expansion rate, and detection speed). The discussion section explores the potential value of associated domain analysis, and clarifies limitations of the current approach.

2 Methodology

2.1 Data Sources

Multiple data sources were used for the study, including domain RBLs, registration data, and active DNS lookup data for newly observed domains.

RBLs. To assess reported domain abuse, we aggregate multiple threat intelligence and RBLs, including commercial and public feeds. The full set included Spamhaus, SURBL, PhishTank, Urlscan, the Anti-Phishing Working Group (APWG) and Abuse.ch's Urlhaus. These sources provide labels for domains associated with phishing, malware distribution, and botnet command-and-control (C&C) infrastructure.

RBLs were used to generate the seed domain list for further analysis based on recently registered gTLD domains appearing in any RBL within the main seven-day analysis period (24-30 September 2025, inclusive). AD output was cross-checked using a broader analysis period: domains first listed in the RBLs between 1 September and 31 October 2025.

Thin gTLD registration data. Analysis is based on the following publicly available domain registration features:

- Registrar name/IANA ID
- Authoritative nameserver(s)
- Creation timestamp

All listed features can be acquired via standard RDAP or WHOIS queries. For efficiency reasons, and in order to prevent possible rate limiting due to high-volume RDAP lookups, we utilized an internal ICANN-compiled gTLD registration dataset called bulk registration data access (BRDA) [11]. The BRDA dataset used for this analysis contains thin registration data elements supplied pursuant to applicable agreements. The analysis described in this report is limited to publicly available technical registration and DNS infrastructure data and does not utilize registrant contact information, payment details, or account-holder identifiers. Depending on context and applicable law, certain technical identifiers may nevertheless constitute or relate to personal data. The methodology therefore aims to minimize reliance on data elements that could directly identify natural persons.

DNS lookup data. Resolved IP addresses (v4) for newly registered domains were captured based on active DNS lookups through the urlscan.io platform.

Analysis periods. Data was compiled for two purposes: historical analysis and proactive detection. For historical analysis, we were interested in the proportion of reported malicious domains ever to be tagged as associated. Therefore, the main seven-day analysis period of new RBL entries was cross-checked using a longer, overlapping, two-month window of AD output domains (1 September-31 October inclusive). For proactive detection, the focus was on reported malicious domains for which detection could be accelerated. Therefore, we included only the associated entries that were first flagged within the main seven-day analysis period.

Processing pipeline. Analysis of new RBL entries was run on a ‘one-shot’ basis, at approximately 12 pm UTC on the day following the earliest RBL entry. The pipeline did not attempt to rerun AD analysis at later timepoints.

2.2 Detection Methods

The following sections describe a methodology for identifying bulk registered domains at high risk of abuse, based on RBL seed analysis. Three related techniques were used to search for associated domains. All three methods take as initial input a reported malicious RBL seed domain, pivoting from this initial seed to find related domains, based on either registration or infrastructure similarity.

Starting from an initial pool of reported malicious gTLD seed domains, lexically similar ADs are identified from the following candidate lists:

- Batch registration detection - capturing domains that share common registrar-nameserver details, registered closely in time to the seed (within seconds)
- Bulk registration detection - capturing domains that share common registrar-nameserver details, registered moderately closely in time (within hours)

- Bulk resolution detection - capturing domains that share common IP(v4) address-nameserver details, first seen resolving moderately closely in time (within hours)

Domain string filtering is based on a range of lexical characteristics, generating high-confidence evidence of association for all AD output domains.

For all three methods, the primary stages of analysis included:

- **Stage 1**. Preliminary grouping of domains based on shared registration or infrastructure data. In the first stage, new RBL additions are used as seed domains. Neighboring domains are identified either from registration (BRDA) or infrastructure (urlscan.io) data. These neighboring domain lists are referred to as 'candidate' associated domain lists.

- **Stage 2**. Lexical 'similar string' filtering. In the second stage, the candidate list is filtered, selecting only domains with shared lexical properties, providing additional evidence of association. This step involves either measuring the lexical similarity between the seed and candidate domains using a random forest machine learning model or using a measure of lexical homogeneity covering all domains within a candidate batch (See section 2.2.2 for details).

2.2.1 Stage-1 Grouping

The three different stage-1 grouping techniques for candidate domain list creation are summarized here. Full details for each method are provided in Appendix A.

Batch-registration clustering. Clustering analysis (i.e., density-based spatial clustering of applications with noise, or DBSCAN) is performed on all new gTLD registration timestamps, separately for each registrar-nameserver combination. (See Appendix A for details.) This method captures tightly timebound domain registrations, provisioned on shared infrastructure. The maximum interval between consecutive domain registrations is five seconds, the minimum cluster size is two.

Bulk-registration grouping. Identification of candidate neighboring bulk registrations over a longer time window. For each RBL seed domain, neighboring registrations that share an identical registrar-nameserver combination are captured, with creation dates falling within a two-hour window of the seed domain (± one hour). Candidate lists are excluded from further analysis if they contain fewer than 5% of domains flagged as malicious (reported in RBLs for phishing, malware, or botnet C&C).

Bulk-resolution grouping. Identification of candidate neighboring bulk-use domains. For each RBL seed domain, we capture neighboring domains first seen resolving to a common IP address (v4) nameserver pairing on the same day as the seed domain. IP address (v4) nameserver pairings with greater than 100 observations on the target date are excluded from further analysis, in order to help prevent the inclusion of false positives (that is, domains that are linked to shared web-hosting infrastructure).

Figure 1 presents a schematic of the Stage 1 analysis pipeline. Starting with an initial seed domain (filtered to malicious registered gTLD domain, recently added to RBLs), registration data

is collected (for gTLD domains created within the last six months only), followed by grouping and clustering analysis.

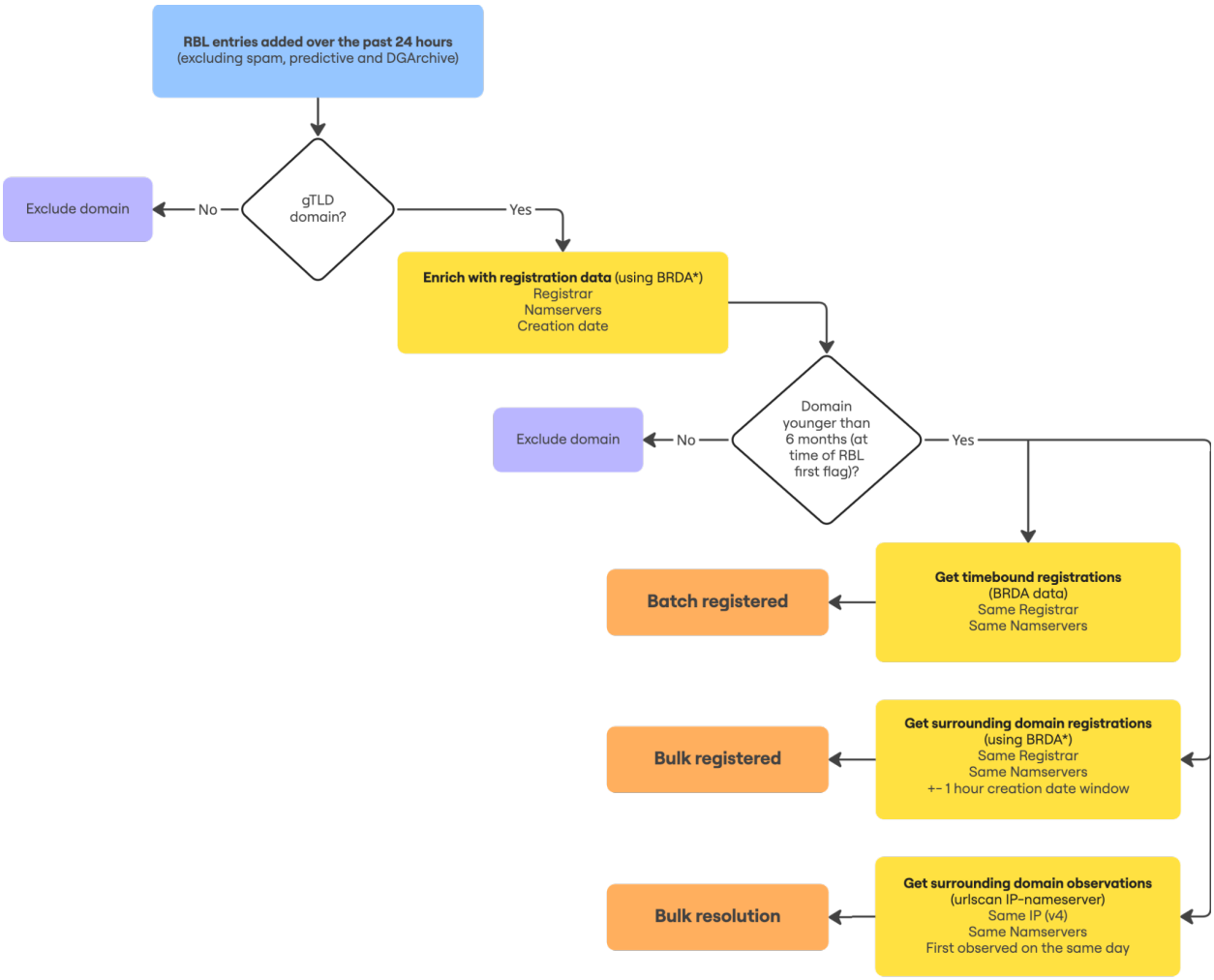

Figure 1. Schematic illustrating the Stage-1 data analysis processing flow, starting with a newly added RBL entry (top left, pale blue box) and ending with the AD method outputs (central, orange boxes).

2.2.2 Stage-2 Lexical Filtering Analysis

Although domains registered in bulk as part of a single campaign do not always share common lexical features, manual review of RBL lists suggests that this is often the case. Therefore, in order to increase confidence in the output of the AD methods, a strict layer of lexical filtering is applied, selecting only candidate domains that share clear lexical similarity to the seed domain. Lexical filtering is performed in two ways, depending on the AD method type.

Batch-registration string filtering. For batch-registered candidate domains, evidence of association is already strong. All domains are registered together closely in time, using shared infrastructure. For this class of candidate domain, therefore, it is sufficient to exclude candidate

clusters that contain high-variance domain strings (see Appendix A), resulting in a filtered list of batch-registered domains that exhibit strong first- or second-order lexical relationships.

This method is capable of highlighting domains that share common second-order features, sometimes linked to domain generation algorithms, because of the type of lexical filtering applied. See examples A and B (listing the first five domains from each group only) as well as clear first-order similarity based on common sub-strings (example C).

- **Example A** - kathryn******[.]shop, julie*******[.]shop, bridget****[.]shop, jessica********[.]shop, robert*******[.]shop (in this example, surnames have been replaced by asterisks of the same length)
- **Example B** - ox8b66[.]top, t0ov2d[.]top, py7yyv[.]top, g51rze[.]top, 50tlid[.]top
- **Example C** - govsxlfhj[.]vip, govhefh[.]vip, govsxlfpj[.]vip, govsxlhpx[.]vip, govsxlhty[.]vip

Bulk registration and bulk-resolution string filtering. Both of these methods loosen the time constraints on creation and first observation dates, leading to a lower-confidence candidate list. Therefore, stricter lexical filtering is applied to the candidate lists, using a custom trained random forest (RF) model, selecting only domains that share clear first- or second-order string similarity to the seed domain (see Appendix B for details).

Three examples of output-associated domain clusters, tagged as **bulk registered**, are shown here:

- **Example A** - uk-dwpbd[.]xyz, uk-dwpog[.]bond, uk-dwpmr[.]xyz, uk-dwprv[.]xyz, uk-dwpth[.]bond
- **Example B** - techformulazone[.]com, mediahivezone[.]com, designscopezone[.]com, techstartupzone[.]com, webservicexzone[.]com
- **Example C** - govstream[.]work, govstream[.]dev, boodstream[.]biz, boodstream[.]app, boodstream[.]club

Three examples of output-associated domain clusters, tagged as **bulk resolution**, are shown here:

- **Example A** - whats-bx[.]vip, whats-nf[.]vip, whats-ns[.]vip, whats-nt[.]vip, whats-ny[.]vip
- **Example B** - dmvfineoqr[.]vip, dmvfineruu[.]vip, dmvfinempj[.]vip, dmvfineguq[.]vip, dmvfinerfi[.]vip
- **Example C** - schwab-secure40ea[.]click, schwab-secure38ea[.]click, schwab-secure36ea[.]click, schwab-secure37ea[.]help, schwab-secure38ea[.]help

2.3 Metric Calculation

Four key metrics are used to evaluate the performance of the different AD methods:

- **Coverage.** The proportion of seed domains (RBL entries listed for the first time within the main seven-day analysis period) tagged by each AD method or aggregated (one or more AD) methods. To account for AD tagging outside of the main seven-day window, coverage of seed domains first seen within the main seven-day period (24 to 30 September 2025) was cross checked with AD output from an extended surrounding period (1 September to 31 October 2025).

- **Flag time deltas**. Time difference, measured in days, between the initial RBL entry and the earliest date of AD tagging.

- **Data delay modeling**. In order to compare detection times for overlapping entries (i.e., domains flagged by both RBLs and AD methods), we generate two measures of the flag time deltas. Firstly, the standard time delta based on the difference (measured in days) between the RBL earliest 'seen since' date and the timestamp of the daily AD batch job. This standard time delta accounts for real-world data acquisition and processing delays, which often can be in the order of days. A second measure is calculated based on an idealized scenario in which there are minimal data acquisition or processing delays. Although not currently feasible in a real-world setting, this idealized minimal delay condition is useful for predicting the benefits of rapid domain registration data sharing [12]. The minimal delay condition models a data processing delay of 30 minutes, post seed listing (i.e., the initial time of entry of the seed in RBLs).

- **Expansion.** The expansion rate metric captures the volume of previously unlisted associated domains flagged by AD methods. This includes domains seen in either BRDA or urlscan.io newly observed domains but not contained in our RBLs (during the analysis period) or flagged first by AD methods. The expansion metric is calculated as the ratio of the following domain volumes: AD flagged first or exclusively / RBL flagged first. For example, 1,000 RBL seed domains generating 100 previously unknown associated domains leads to an expansion metric of 0.1 (100/1000).

Lexical filtering thresholds for coverage and expansion analysis. Lexical similarity thresholds varied depending on the analysis type. Coverage analysis focused on the pool of reported malicious domains, employing a low lexical filtering threshold (RF lexical similarity >= 0.3) due to the implicit pre-filtering of the list by RBL providers. In contrast, expansion analysis required a higher threshold (RF lexical similarity >= 0.8) due to the more diverse candidate domain list, linked to a higher risk of false positives.

3 Results

3.1 Overall Analysis

The following results are compiled using phishing, malware, and botnet C&C domains first added to the RBLs between 24 and 30 September 2025, inclusive. Over the course of the seven-day analysis period, approximately 4,000 newly registered gTLD domains were added to the RBLs per day (with a domain age of six months or less at the time of RBL flagging). The dashed black line in Figure 2 illustrates fluctuations in daily RBL volumes over the one-week period, consisting of reported malicious domains (abused for purposes of phishing, malware, or botnet C&C) used as seeds for associated domains analysis.

Figure 2 also shows the volume of RBL seeds that were successfully tagged as associated or linked to one or more neighboring domains (dashed red line). Overall, across the seven-day period, 56.4% of RBL seeds were tagged as associated, based on a combination of the output of individual AD methods (solid lines). This means that over half of reported domains have at least one associated domain.

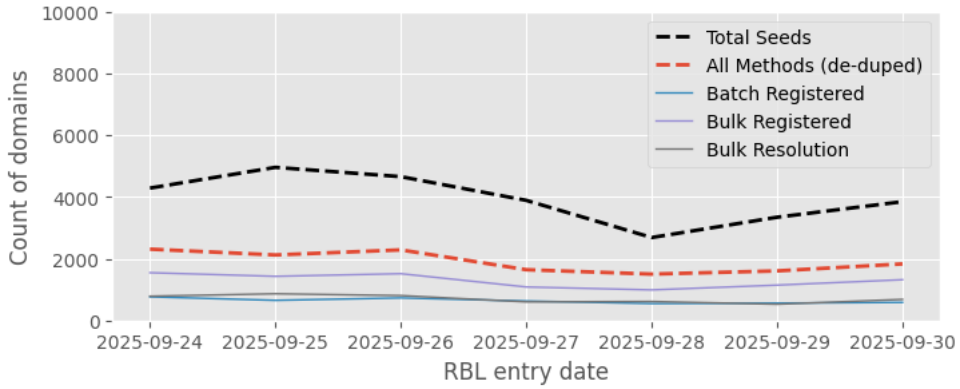

Figure 2. RBL coverage over the seven-day analysis period, displaying RBL daily addition volumes. Total gTLD newly registered counts (seed domains, dashed black), batch registered (blue), bulk registered (purple), bulk resolution (grey) and de-duplicated counts based on a combination of all three associated domain methods (dashed red).

Comparison of methods reveals that the highest proportion of RBL domains are tagged by the bulk registration method (shared registration details, loosely time-bound). The bulk registration method tags approximately 1,500 domains per day (32.6% coverage), while both the batch-registered and bulk-resolution methods tag a smaller proportion of the RBL seeds (between 500 and 1000 per day), corresponding to 16.1% and 17.6% coverage respectively.

Figure 3 illustrates the overlap between AD method output.

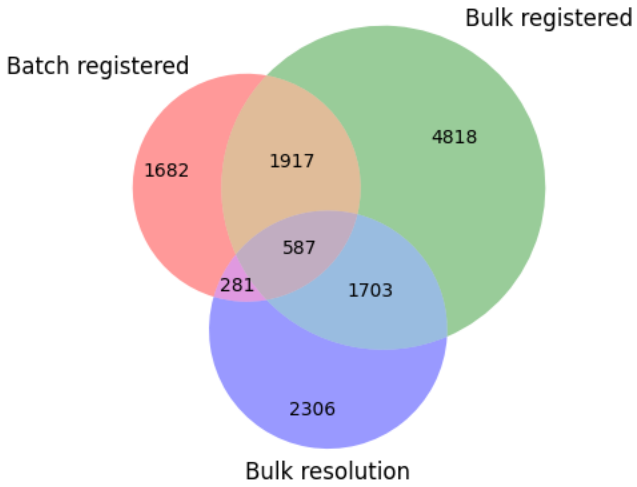

Figure 3. Venn diagram showing the overlap between individual AD method output.

The subset of RBL seed domains successfully tagged as associated by one or more methods are subsequently referred to as associated-RBL (a-RBL) domains.

While the majority of a-RBL domains are flagged first through standard RBLs, a minority (36.3%) can be detected earlier through associated domain analysis. To visualize the relative performance of AD method flag times, relative to standard RBLs, Figure 4 displays distribution in the form of violin plots, for each of the AD methods separately. Violin plots illustrate mirrored

(symmetrical) distributions, helping to highlight peaks where the majority of overlapping domains fall. Most domains fall within the positive part of the distribution, indicating initial detection by RBLs. The median flag delta (FD), the delay between RBL entry and a-RBL tagging, is less than 1 day for all methods, with the shortest delta associated with the batch registration method. Data collection delays, for both the registration and infrastructure data, contribute to these intervals. For instance, compilation of gTLD bulk registration data may take up to seven days for some gTLDs.

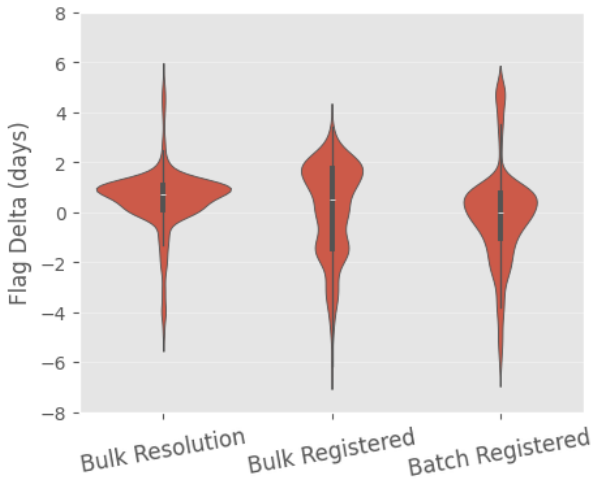

Figure 4. Violin plots showing flag time delta distributions (a-RBL flag time vs RBL flag time), split by AD method type. Negative values indicate initial flagging by AD methods.

Table 1 summarizes key performance metrics, split by AD method. In addition to the coverage and flag delta results described above, the table includes expansion rates. The expansion rate reflects the volume of previously unlisted associated domains (i.e., newly found), relative to the volume of original RBL entries. (See section 2.3 Expansion for details.) The results highlight that although the bulk registration method tags the highest proportion of existing RBL entries, batch registration analysis yields the highest expansion rate of 2.32 (for each RBL seed domain, more than two additional high-risk domains are identified).

	Coverage	Expansion	FD (Mean)	FD (Med)	FD (STD)
Batch registered	0.16	2.32	-0.15	0.01	1.97
Bulk registered	0.33	0.12	0.14	0.5	1.84
Bulk resolution	0.18	0.06	0.54	0.75	1.27

Table 1. Summary of key performance metrics. Coverage: the proportion of RBL seed domains tagged as associated. Expansion: ratio of expanded to seed domains, FD (Flag Delta) summary statistics including the mean, median (med) and standard deviation.

The data processing pipeline for AD analysis involves some data collection delays, mainly due to:

- Delays compiling the required gTLD registration data. Many gTLD operators only provide weekly updates, resulting in a maximum interval of seven days between domain creation and the registration data availability in BRDA, the centrally compiled data source.

- Fixed delays due to the daily batch scheduling of AD analysis. Run once per day, on the day following the initial RBL listing of each seed domain.

Figure 5 and Table 2 present the results under a minimal data delay assumption, in which all required registration data is available within 30 minutes of seed domain listing. This idealized scenario is useful for understanding the theoretical potential of AD methods for proactive detection.

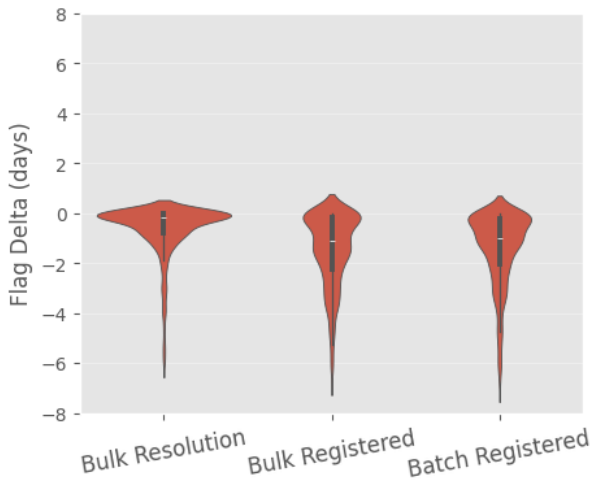

Figure 5. Violin plots showing flag time delta distributions (AD flag time vs RBL flag time), split by AD method type. Results are calculated assuming minimal data delay. Negative values indicate initial flagging by AD methods.

Delays acquiring the necessary BRDA data for both the batch and bulk registration analysis strongly influence FD results (Table 2). Under these new minimal data delay conditions, median flag time deltas are negative for all AD methods (between -1.41 and -0.64 days). The majority (87.0%) of overlapping entries are now detected earlier by AD analysis, prior to inclusion in RBLs.

	Coverage	Expansion	FD (Mean)	FD (Med)	FD (STD)
Batch registered	0.16	2.34	-1.41	-0.98	1.48
Bulk registered	0.33	0.17	-1.46	-1.1	1.47
Bulk resolution	0.18	0.09	-0.64	-0.17	1.07

Table 2. Summary of key performance metrics. Results are calculated assuming minimal data delay. Coverage: the proportion of RBL seed domains tagged as associated. Expansion: ratio of expanded to seed domains, FD summary statistics including the mean, median, and standard deviation.

3.2 Threat-Type Analysis

The following section illustrates differences in AD performance across threat types: phishing, malware, and botnet C&C. It should be noted that the majority of domains included in the sample were tagged as phishing (n=27,206), with much smaller sample sizes for malware (n=344) and botnet C&C (n=143). Therefore, the results for malware and botnet C&C threat

types should be interpreted with caution. For the following tables and plots, botnet C&C and malware categories have been combined to increase the sample size.

Table 3 presents key metrics split by threat type, highlighting significantly higher coverage for phishing (57%) relative to botnet C&C-malware (43%) domains. Expansion rates, however, are significantly higher for botnet C&C-malware (5.88) relative to phishing (2.36) domains.

	Coverage	Expansion	FD (Mean)	FD (Med)	FD (STD)
Botnet C&C-malware	**0.43**	**5.88**	**-1.65**	**-2.38**	**2.37**
Phishing	**0.57**	**2.36**	**0.14**	**0.47**	**1.71**

Table 3. Summary of key performance metrics, split by threat type. See Table 1 caption for details.

For domains contained in RBLs *and* tagged by an AD method, flag delta results demonstrate a delay of approximately half a day for phishing domains (FD Median= 0.47 days), while botnet C&C-malware domains are frequently flagged first by AD methods (FD Median= -2.38 days). This difference is statistically significant based on an independent-samples t-test, $t(3,229) = 4.05$, $p < 0.001$, suggesting that AD methods are more effective at early detection of botnet C&C-malware than phishing domains, although it is necessary to replicate on a larger sample in future.

With the *minimal data delay* assumption (Table 4), median flag deltas decrease for both groups (FD Median= -0.69 and -4.40 for phishing and botnet C&C-malware respectively), implying that a large number of domains could be tagged as associated more quickly if the necessary registration data was ingested and processed rapidly.

	Coverage	Expansion	FD (Mean)	FD (Med)	FD (STD)
Botnet C&C-malware	0.43	5.89	-3.59	-4.4	2.43
Phishing	0.57	2.44	-1.14	-0.69	1.39

Table 4. Summary of key performance metrics, split by threat type. Results are calculated assuming minimal data delay. See Table 1 caption for details.

Figure 6 presents flag delta distributions split by threat type and data delay assumptions.

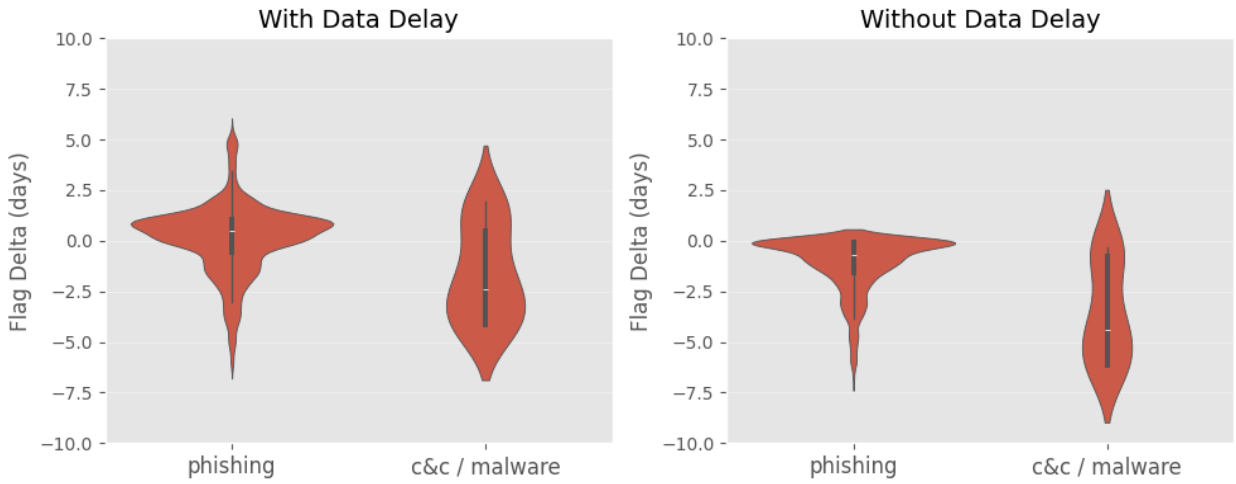

Figure 6. Violin plots showing flag time delta distributions (AD flag time vs RBL flag time), split by threat type and data delay assumptions. Negative values indicate initial flagging by AD methods.

3.3 RBL-Type Analysis

In this section, RBLs are split into two groups in order to explore AD performance for public (free) RBLs versus commercial RBLs. RBLs contained in the public group included APWG, PhishTank, and OpenPhish. RBLs contained in the commercial group included Spamhaus, SURBL, and Urlscan.

The sample consisted of 27,492 commercial RBL and 1,905 open RBL domains. Table 5 shows that coverage is significantly higher for the public RBLs group (68% of entries were tagged as associated vs. 56% for commercial RBLs). Expansion rates are significantly higher for the public RBL group (4.39 previously unknown domains detected per seed for public seeds vs. 2.53 for commercial seed).

	Coverage	Expansion	FD (Mean)	FD (Med)	FD (STD)
Commercial RBLs	0.56	2.53	0.13	0.41	1.71
Public RBLs	0.68	4.39	0.63	0.91	1.57

Table 5. Summary table listing key performance metrics. See Table 1 caption for details.

Commercial RBL listings tagged as associated are typically detected earlier than those from public RBLs, $t(4,144) = 6.34$, $p < 0.001$. Recalculation of metrics under the *minimal data delay* assumption yields the results in Table 6. As expected, median flag deltas are decreased for both groups (-0.69 days for commercial and -0.06 days for public RBLs).

	Coverage	Expansion	FD (Mean)	FD (Med)	FD (STD)
Commercial RBLs	0.56	2.64	-1.15	-0.69	1.39
Public RBLs	0.68	4.58	-0.59	-0.06	1.03

Table 6. Summary of key performance metrics, split by RBL type. Results are calculated assuming minimal data delay. See Table 1 caption for details.

Figure 7 presents flag delta distributions split by RBL type and data delay assumptions.

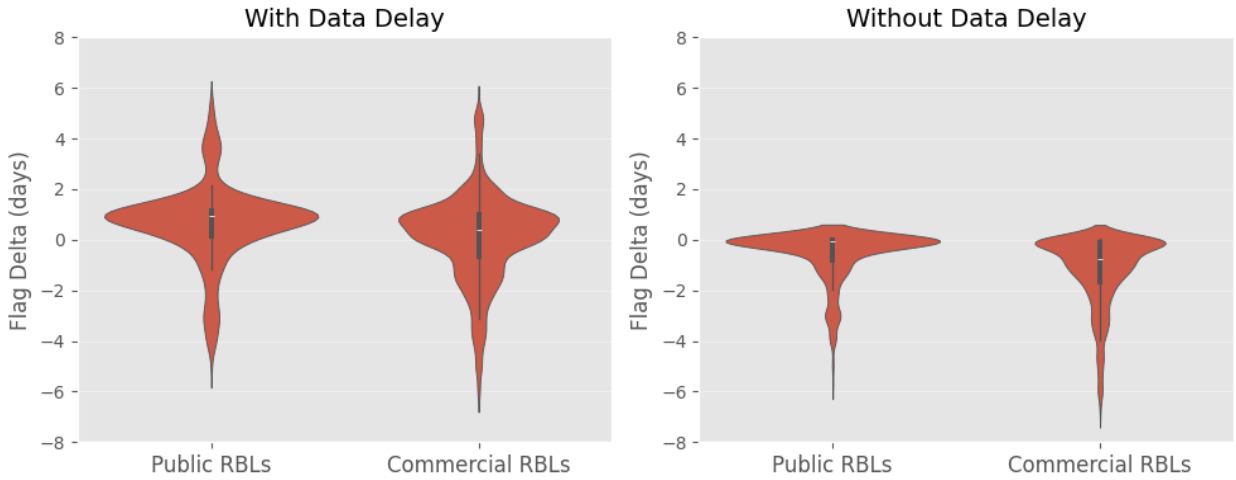

Figure 7. Violin plots showing flag time delta distributions (AD flag time vs RBL flag time) split by RBL-type and data-delay assumptions. Negative values indicate initial flagging by AD methods.

3.4 Exclusion of a DGA Campaign Subset

A review of the expanded set highlighted large clusters of batch registered domains under a single gTLD, which is consistent with there being a single campaign. Specifically, expansion analysis via the batch registration method yielded a large number of domains linked to a domain-generating algorithm (DGA) sharing consistent lexical properties (six characters, alphanumeric). The majority of these domains were not contained within existing RBLs.

Some of the batches linked to this DGA campaign are extremely large (containing over 6,000 domains) and, therefore, may skew the results shown in Table 5. Exclusion of this subset of domains from the analysis sample substantially decreases the expansion rate, as seen in Table 7. This is particularly true for the public RBL group, decreasing from 4.39 expanded domains per seed to 0.99. This result indicates that the majority of expanded public RBL output for the current sample period consists of domains linked to this single DGA campaign.

	Coverage	Expansion	FD (Mean)	FD (Med)	FD (STD)
Commercial RBLs	0.54	1.54	0.22	0.5	1.72
Public RBLs	0.64	0.99	0.84	0.96	1.44

Table 7. Summary of key performance metrics, split by RBL type. Excluding DGA campaign domains. See Table 1 caption for details.

4 Discussion

Applying AD detection to a sample of RBL domains from late 2025, the study led to a number of findings.

- Over half of all reported malicious gTLD domains in our sample were registered in combination with other domains or linked to common DNS infrastructure. In total, AD techniques tagged 56.4% of the RBL sample, consistent with the widespread practice of bulk registration.
- A subset of RBL entries could be detected earlier via associated domain analysis. Results indicate that just under 5% could be detected earlier based on the current data processing pipeline; this number is estimated to rise to at least 10% if data processing delays could be minimized.
- Expansion analysis revealed high-risk associated domains, often registered in batches with reported malicious RBL domains. These domains received no confirmed reports of abuse within the analysis period, but shared strong similarities to known abused domains.

Facilitating registrar checks. Results are consistent with the importance of policies and practices put in place by individual registrars (e.g., regarding API availability and restrictions) for limiting high-volume registrations. The AD methods described above have been designed to operate using publicly available data, in contrast to privately held registrar data, such as account holder ID.

The public data approach is beneficial for a transparent, wide-scale, centralized analysis of registration patterns across multiple TLDs and registrars, and does not require analysis of any personal registrant data. This is currently done for gTLDs only using centralized BRDA registration data, but is also possible for ccTLDs with bulk RDAP access. This permits comparable 'sector-wide' results to be generated, which are important for benchmarking and auditing purposes.

On the other hand, the use of public data for AD analysis requires a process of inference and filtering to achieve high-confidence results, which likely means that some associated domains are being missed by the current methods. Public data AD analysis, as described in this report, is viewed as complementary to registrar-directed analysis using privately held records.

Overall bulk rates. Our findings highlight the prevalence of bulk registrations within our sample pool of gTLD domains reported as maliciously registered. Over half of domains in our sample were tagged as associated (linked to one or more neighboring domains via registration or infrastructure patterns) at some point over the two-month review period (September to October 2025).

This supports the value of bulk-registration monitoring described in the DNS Abuse Mitigation PDP, initiated in 2025, and suggests that associated domain checks run by registrars are likely to be an effective strategy for proactive identification of domains at high risk of abuse. The speed of detection using the AD methods described in this report is constrained by data processing delays, as well as delays building up the required level of evidence (see Methods section for details).

Therefore, whereas many bulk registered domains reported as malicious are successfully identified using the current methods, data delays often mean this happens *after* active abuse, as measured by the initial date of RBL entry. In-house pivot analysis performed by registrars would bypass many of these constraints, leading to faster detection times.

In this study we have focused on the prevalence of bulk registrations linked to reported abuse. It is important to note that the results do not imply that *all* bulk registrations are intended for malicious use, and that the findings are consistent with the legitimate use of many bulk registered domains. There are many possible benign use cases for bulk-registered domains, including defensive registrations, advertising networks, or traffic distribution systems. Previous work has highlighted the large number of batch-registered gTLD domains not reported as abused in RBLs, and presumed legitimate [7]. Understanding the use cases for these gray-area bulk domain registrations is a potentially valuable line of future investigation.

Multiple AD methods. Multiple related but distinct bulk-detection methods are shown to add value. Previous analysis highlighted that close to 20% of gTLD domains reported as maliciously registered were also batch-registered [7], due to shared registration details and tightly time-bound creation dates. The current results support this finding and additionally highlight the added value of multiple overlapping AD methods. Using multiple different detection techniques, over 50% of the equivalent set of reported malicious domains are covered. The true rate is likely to be higher, due to the conservative lexical filtering employed in the current study, which excludes many valid batches.

Threat-type comparison. Sample sizes for botnet C&C and malware were small, preventing robust comparison of threat types. After combining the botnet C&C and malware data, we found reduced coverage for higher severity threat types (botnet C&C and malware) compared to phishing (43% vs 57% respectively). This may indicate that bulk registrations are less frequent for higher severity abuse types, such as botnet C&C, due to deliberate attacker strategies to evade detection and suspension of high-value domains.

Attacker strategies. Attackers may adapt to increased monitoring of registration data. For instance, by distributing registrations over longer time periods or across multiple registrars. The proposed AD techniques rely on the assumption that bulk registrations are associated with identical registrar-nameserver or IP address(v4)-nameserver groupings, which will not always be the case. While attackers may attempt to evade detection by bulk registration of domains across multiple registrars and nameservers, this process increases the time and complexity involved, and is unlikely to be adopted at scale. Emerging practices, such as the increasing use of agentic AI in the domain registration pipeline, may lower the barrier to distributing large-scale malicious campaigns across multiple registration and infrastructure providers. Importantly, the AD grouping techniques outlined in this document may be adapted to changing requirements and are not restricted to the set of assumptions outlined above.

Data delays. Access to timely bulk registration data has the potential to facilitate rapid abuse detection. The results of this study highlight that rapid access to domain registration data is important for effective AD based abuse detection; the percentage of reported abused domains caught earlier by AD methods more than doubles under conditions of rapid data collection — from 4.6% to 10.0%. Current analysis has been performed on a bulk dataset available for internal ICANN analysis, compiled using thin registration data supplied by gTLD operators (BRDA agreement [14]), negating the need to run multiple RDAP queries. At the time of analysis all gTLD operators were required to supply updates at a minimum frequency of once per week, although it should be noted that the latest Registry Agreement [12] decreases this minimum update frequency to once per day.

Verification, Evaluation and False Positives. One method of assessing the validity of the AD clustering method is to evaluate the output against a registrant/account holder ground-truth

dataset. Domains grouped into distinct campaigns using the method described in this study should commonly map to a single owning registrant or account holder. Although this type of evaluation is informative, it has not been possible to include in this study because ICANN lacks visibility of registrant level personal data, which is typically held by registrars and some registries. To perform a comprehensive, privacy preserving, evaluation of the method, ongoing collaboration between ICANN and the registrar community is required. It should be noted that the high threshold for domain string similarity employed in this study may result in missing many true positives. Specifically, the method may fail to capture some bulk-registered domains linked to the same registrant or account ID, but with diverse domain string characteristics.

Although strong emphasis has been placed on designing an algorithm that minimizes the number of incorrect reports, it is still possible that the current AD output contains some false positives. We believe that this is most likely to be the case for the batch detection algorithm, for which, unlike previous analysis [7], batches containing over 100 domains were retained. This raises the possibility that the present results may contain a small number of datapoints mislabeled as associated domains, linked to registrars with non-standard portfolio domain string characteristics. Future work will seek to ensure the consistent removal of this class of false positives.

Types of evidence. It is important to distinguish evidence of *abuse* from evidence of *association*. Our approach differs fundamentally from many other predictive techniques for DNS and domain name abuse - rather than reputation scoring broad features based on historical abuse levels, our goal is to reliably group associated domains into distinct clusters or campaigns. The output is thus purely a signal of how likely it is that groups of domains will be used for the same purpose, with no insight into what that purpose is. It is only when this output is paired with evidence of abuse sourced from external feed providers that we start to find previously unknown domains at high risk of abuse. Our results are agnostic with respect to the question of what constitutes sufficient *evidence* of abuse. In this study we have relied on widely used RBLs, but the AD expansion method is equally applicable when a more restricted set of seed domains are employed; for instance, using only confirmed abused domains (e.g. with direct evidence of abuse, such as a phishing page screenshot). The AD grouping method is designed to only highlight domains sharing *strong* evidence of association. We suggest that associated domains identified as at high risk of abuse are suitable candidates for further investigation. However, it is important to highlight the prevalence of obfuscation techniques including cloaking [15], which can significantly impair the ability of investigators to gather *direct* evidence of abuse at scale. Therefore, an important consideration is whether strong evidence linking a domain to a known abuse campaign can ever be considered sufficient to take action (e.g. domain suspension), without additional *direct* evidence of abuse.

5 Conclusion

This report highlights the potential value of associated domain analysis using only publicly available domain registration and DNS infrastructure data, for monitoring both bulk-registration patterns (historical analysis) and proactive detection of high-risk domains. A set of multiple overlapping AD techniques have been developed for gathering a high threshold of associative evidence. These techniques have been documented to enable transparency and further the security, stability and resilience goals of ICANN and the larger community.

6 References

[1] Y. Nosyk et al., “Exposing the Roots of DNS Abuse: A Data-Driven Analysis of Key Factors Behind Phishing Domain Registrations,” in Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security, Taipei Taiwan: ACM, Nov. 2025, pp. 618–632. doi: 10.1145/3719027.3744869.

[2] M. Felegyhazi, C. Kreibich, and V. Paxson, “On the potential of proactive domain blacklisting,” in Proceedings of the 3rd USENIX Conference on Large-scale Exploits and Emergent Threats (LEET), USENIX Association, USA, 2010.

[3] J. Spooren, T. Vissers, P. Janssen, W. Joosen, and L. Desmet, “Premadoma: an operational solution for DNS registries to prevent malicious domain registrations”, in *Proceedings of the 35th Annual Computer Security Applications Conference*, San Juan Puerto Rico USA: ACM, Dec. 2019, pp. 557-567. doi: 10.1145/3359789.3359836.

[4] A. Del Soldato, D. Sartiano, and M. Martinelli, “Reads: Enhancing .it domains integrity with registrant’s anomalies detection system,” in 2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), 2024, pp. 1–7.

[5] J. Prins, “Proactive Recognition of Domain Abuse”, M.S. thesis, DACS, University of Twente, 2021.

[6] T. Wabeke and T. van den Hout, “Assessing the risk of new .nl registrations using RegCheck”. Accessed: July 01, 2025. [Online]. Available: https://www.sidnlabs.nl/en/news-and-blogs/assessing-the-risk-of-new-nl-registrations-using-regcheck

[7] S. Cheadle, C. Gañán, S. Lloyd, S. Tajalizadehkhoob, “Detecting Malicious Domain Registration Batches: Patterns, Prevalence, and Security Implications”, in *2025 APWG Symposium on Electronic Crime Research (eCrime)*. doi:10.1109/eCrime66972.2025.11327823

[8] Generic Names Supporting Organization, “Preliminary Issue Report on a Policy Development Process on DNS Abuse Mitigation”. Accessed: Nov. 25, 2025. [Online]. Available: https://www.icann.org/ru/public-comment/proceeding/preliminary-issue-report-on-a-policy-development-process-on-dns-abuse-mitigation-08-09-2025

[9] G. Bunton, “White Paper: Proposal for PDPs on DNS Abuse”, NetBeacon Institute. Accessed: July 30, 2025. [Online]. Available: https://netbeacon.org/white-paper-proposal-for-pdps-on-dns-abuse/

[10] S. Hao, A. Kantchelian, B. Miller, V. Paxson, and N. Feamster, “PREDATOR: Proactive Recognition and Elimination of Domain Abuse at Time-Of-Registration”, in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Vienna Austria: ACM, Oct. 2016, pp. 1568-1579. doi: 10.1145/2976749.2978317.

[11] ICANN, “Base Registry Agreement - Approved 21 January 2024”. Accessed: July 24, 2026. [Online]. Available: https://www.icann.org/en/contracted-parties/registry-operators/registry-agreements/base-agreement/2012

[12] R. Sommese, G. Akiwate, A. Affinito, M. Müller, M. Jonker, and K. C. Claffy, “DarkDNS: Revisiting the Value of Rapid Zone Update”, in *Proceedings of the 2024 ACM on Internet Measurement Conference*, Nov. 2024, pp. 454-461. doi: 10.1145/3646547.3689021.

[13] M. Ester, H.-P. Kriegel, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, in Proc. 2nd Int. Conf. Knowl. Discovery Data Mining (KDD), Portland, OR, USA, 1996, pp. 226–231.

[14] ICANN, “2026 Base Registry Agreement - Approved 12 March 2026”. Accessed: Sep. 7, 2026. [Online]. Available: https://www.icann.org/en/contracted-parties/registry-operators/registry-agreements/base-agreement/2026
[15] W. Li, S. Manickam, S. U. A. Laghari, and Y.-W. Chong, “Uncovering the Cloak: A Systematic Review of Techniques Used to Conceal Phishing Websites,” IEEE Access, vol. 11, pp. 71925–71939, 2023, doi: 10.1109/ACCESS.2023.3293063.

Appendix A

A.1 Batch Detection

A.1.1 Algorithm

We define a *batch* as a group of domains registered close in time, through the same registrar, and served by the same authoritative nameservers. The detection process involves:

- Grouping domains by unique combinations of registrar and authoritative nameservers
- Applying density-based clustering (DBSCAN) to the creation timestamps within each group to identify time-bound clusters
- Applying a set of filtering criteria to exclude low-reliability clusters

Density-based spatial clustering of applications with noise (DBSCAN [13]) is a popular clustering algorithm suitable for segregating data into regions of high and low density, well suited to the task of identifying dense clusters of registrations against a sparse background of unrelated activity. Unlike centroid-based or hierarchical-clustering methods, which typically do not consider the density of datapoint distribution, DBSCAN segregates regions of high density (clusters) from lower-density regions (noise). This quality makes the algorithm suitable for batch registration detection.

DBSCAN relies on two primary hyperparameters: *epsilon* and *min-samples*. Epsilon defines the maximum distance between data points required for them to be considered neighbors - initial testing explored a range of epsilon values, from 0.1 to 10 seconds. Although higher precision is achieved for low epsilon values, capturing only batches of domains registered near simultaneously, we were interested in capturing expanded batch registrations taking place over the course of multiple seconds or even minutes. Therefore, an epsilon value of 5 seconds was selected. A min-samples value of 2 was selected, enabling analysis of all batch sizes.

A.1.2 Filtering Criteria

Filtering is applied to generate a high confidence sub-set of batches for detailed analysis, excluding all batches of domains with inconsistent lexical characteristics. The method represents a conservative filtering approach, highlighting batches linked to strong evidence of association only, likely to also exclude many real batch registrations. Specific filtering criteria are listed and justified below.

Many batches contain domains sharing common keywords (e.g. login123[.]com, 456login[.]com) or higher order lexical properties linked to algorithmic domain generation. In order to select batches with strong evidence of association, we used two forms of string similarity validation, similar to the techniques described in the PREDATOR study [10].

First order similarity. Trigrams of the domain strings (excluding TLD) were calculated, followed by Jaccard similarity scoring. Scores were calculated by comparing the initial domain within each cluster, to all other domains within the cluster. Only clusters with a mean Jaccard score >=0.3 were selected for inclusion based on first-order string similarity.

Second-order similarity. Some clusters lack first-order feature similarity, such as identical sub-strings, but nevertheless contain domains that share common second-order features (commonly linked to DGA). To capture similarity of this type, the following set of second-order metrics were generated for all domains within the sample:

- String length
- % highest frequency character
- % distinct characters
- % digits
- % hyphens

Following the calculation of these scores at the domain level, intra-batch statistics were generated, capturing the mean and variance for each of the five separate metrics. Anomalous batches were identified via comparison to a baseline population of domains (randomly sampled Tranco top-1M domains, n=500). Mann-Whitney U tests (non-parametric test of rank distribution differences) were performed on all metrics, on a per cluster basis, in order to highlight statistically significant second order string differences. Clusters were selected for inclusion if two or more metrics exhibited a statistically significant difference ($p<0.05$) from the baseline population.

Appendix B

B.1 Random Forest Lexical Similarity Model

The following sections describe the training and evaluation of a decision tree based machine learning model, referred to as a *random forest* (RF) model, used to analyse the candidate domain lists for two AD methods (bulk registration and bulk resolution). See section 2.2.2 for an overview.

B.1.1 RF Training Data

Training data comprised candidate lists generated from registration (BRDA) data analysis, using a two-hour window (± one hour) surrounding each RBL seed domain. A total of 674 RBL seed domains and 6,789 neighboring domains were included in the sampled dataset, added to the RBLs between 2025-08-04 and 2025-09-04.

'Associated' labels were generated by manual inspection of all domains in the sample, comparing the string properties of seed domains to neighboring domains. Domains were tagged as related only when sharing clear lexical characteristics with the seed domain, focusing on first-order lexical similarity (shared sub-strings). Table 8 illustrates a typical example of a candidate cluster, with manual inspection labels included.

Domain	Seed domain	First seen delta	Associated
tigersecurlties.vip	com-paykea.vip	-2084	✗
com-payjfe.vip	com-paykea.vip	0	✓
com-payumv.vip	com-paykea.vip	0	✓
com-paykea.vip	com-paykea.vip	0	✓
com-paytkn.vip	com-paykea.vip	1	✓
com-payaoj.vip	com-paykea.vip	3	✓
com-paycdo.vip	com-paykea.vip	3	✓
mo-suzuku.cyou	com-paykea.vip	253	✗
dscloth.cyou	com-paykea.vip	3037	✗
stodindiatraeef.vip	com-paykea.vip	3217	✗

Table 8. Example training data, showing neighboring domains or a single RBL seed domain (com-paykea[.]vip), together with first seen deltas (the seconds between registration dates), and the manually generated tag.

B.1.2 Model Features

The following feature types were calculated for all seed-neighbor pairs in the training and test data. Note that all lexical features were calculated after excluding the TLD from the domain string.

- First and second order lexical domain string features (see A1.2 for details)
- Word based features, including exact seed-neighbor matches and English language corpus prevalence
- Registration time features, including the time delta between seed-neighbor registrations, and registration density over multiple time windows.
- TLD features, including seed-neighbor exact matches and TLD prevalence

An extended technical report describing the features in greater detail is available on request: SSR-Research@icann.org.

B.1.3 Model Training

A random forest (RF) decision tree classifier (scikit-learn library) was trained to score the similarity between seed-neighbor domain pairs. Number of estimators=100, Max leaf nodes=100, all other parameters were set to default. Five-fold cross validation was performed on the training data. The results reported below represent the average from the five train-test splits.

The predicted class probabilities from the trained RF model were used to evaluate model performance. The predicted probability that the input seed domain is associated with the neighbor domain is denoted AP (associated probability). Figure 8 displays Receiver Operating Characteristic (ROC) and Detection Error Tradeoff (DET) curves for test data splits, based on the AP output scores.

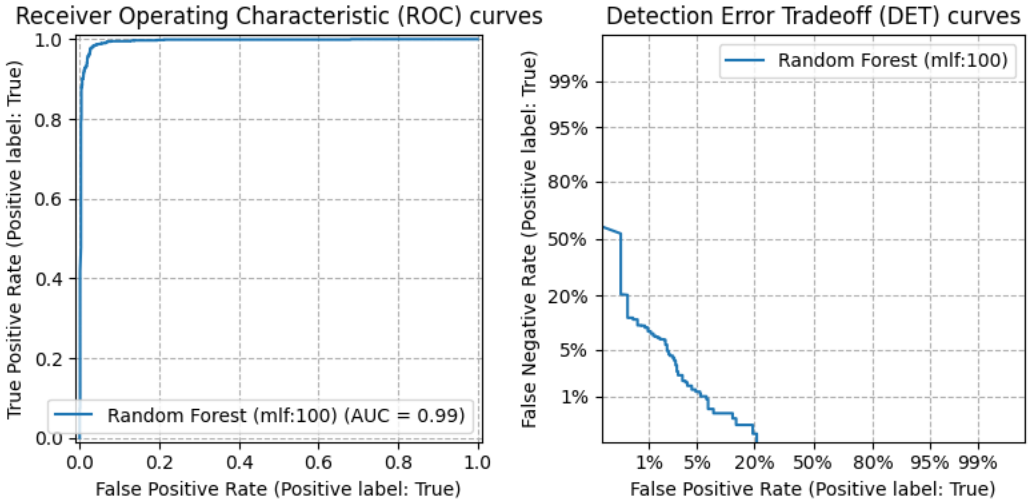

Figure 8. ROC and DET curves for test data split (20%, five-fold cross validation). AUC = Area Under the Curve.

Based on an AP threshold of 0.7 (reflecting a 70% or greater predicted probability of association), Table 9 summarizes the accuracy, precision, recall, and F1 score. Results demonstrate strong performance, with an F1 score of almost 0.97.

	Accuracy	Precision	Recall	F1_score
Test set	0.967	0.98	0.956	0.968

Table 9. Test results from the five-fold cross validation process (average of five runs). Model AP score threshold of 0.7 used for the predicted 'associated' class.

B.1.4 Model Testing (Temporal Holdout)

In order to account for the possibility that the model was overfitting to only lexical features linked to a limited group of malicious domains, a new test set (n=1001) was collected and manually labelled, based on RBL feed entries covering a later time period (between dates 2025-10-01 and 2025-10-07).

Results (Table 10) demonstrate that the model generalizes well to test data from this later period, with similar overall performance (F1 score =0.975).

Accuracy	Precision	Recall	F1_score
0.978	0.973	0.977	0.975

Table 10. Test results from the holdout test dataset. Model AP score threshold of 0.7 used for the predicted 'associated' class.