LogoTrust: Leveraging BIMI to Build a Validated Dataset of Brands, Domain Names, and Logos

Youssef Abyaa*, Olivier Hureau*, Andrzej Duda*¶, and Maciej Korczyński*¶
*Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France
¶KOR Labs Cybersecurity, F-38000 Grenoble, France
firstname.lastname@univ-grenoble-alpes.fr

***Abstract*—Most brand logo datasets are static, manually labeled, and solely rely on the trust in the dataset creator. We leverage Brand Indicators for Message Identification (BIMI) to build a dynamic, automatically generated logo dataset with trust anchored in Certificate Authorities (CAs). The BIMI DNS TXT records, in which the domain owners publish logos with Mark Certificates (MCs), link brand logos to domain names, streamlining the collection and the verification of the mapping while potentially enhancing logo-based phishing detection methods. Using global-scale DNS measurements, we collect BIMI records, retrieve and deduplicate logos, and validate Mark Certificates to obtain the mapping of 1,680 brands, 1,811 logos, and 2,821 domain names. Our dataset, available at https://github.com/josef0x/LogoTrust, demonstrates the BIMI potential as a reliable means for constructing logo datasets and potentially provides a valuable resource for applications such as phishing detection.**

***Index Terms*—Logo Dataset, Brand Indicators for Message Identification (BIMI), Phishing**

1. Introduction

Phishing is a widespread cyber threat that involves tricking users into revealing sensitive information such as login credentials or financial details, by impersonating trusted entities. Attackers achieve this deception through various means, including emails, fake websites, and social engineering tactics. A key component of phishing schemes is visual impersonation in which fraudulent websites and messages mimic legitimate organizations to appear trustworthy. The crucial aspect of this impersonation is the use of brand logos [1] —companies and organizations heavily rely on visual elements in their communication with logos being among the most recognizable and widely used. Logos serve as concise identifiers that instantly convey the brand identity and authenticity. Cybercriminals exploit this trust by integrating brand logos in phishing emails and web pages, making fraudulent content appear legitimate. Based on this observation, anti-phishing solutions increasingly focus on logo detection, since logos are frequently exploited in phishing campaigns to impersonate targeted brands [2]–[5].

However, detecting a logo alone may not be sufficient to determine whether a webpage or an email is malicious or not. To assess a potential phishing attempt accurately, researchers often combine logo recognition with additional verification methods such as analyzing the domain name hosting the webpage (e.g., Phishpedia [4]). Since legitimate brands typically use well-established domains, whereas phishing sites often leverage compromised websites, newly registered domains, or free subdomain providers [6], domain name verification serves as a critical additional layer of defense in phishing detection.

In recent years, logo datasets have become increasingly popular, as they serve as essential resources for building and evaluating logo recognition solutions [7]–[12]. Brand Indicators for Message Identification (BIMI) [13], currently an active Internet draft, specifies how domain owners can publish their logos along with evidence documents that serve as the proof of ownership, enabling email receivers to verify the authenticity of the brand. By associating brand logos with domain names, BIMI helps prevent phishing attacks by ensuring that only legitimate emails from trusted brands display their logos. It aims to enhance trust in email communication and reduce the effectiveness of visual impersonation attempts.

This paper is the first to explore the potential of BIMI as a reliable means for constructing a logo image dataset and examines the advantages and limitations of this approach. Through global-scale DNS active measurements, we collect BIMI `TXT` records from our extensive domain dataset, retrieve and deduplicate logos, and validate Mark Certificates to create a mapping that can be applied in various contexts, including phishing detection. Our final dataset comprises 1,680 brands, 1,811 logos, and 2,821 corresponding domain names, and is available to the community at https://github.com/josef0x/LogoTrust.

2. Related Work

In practical scenarios, logos appear in unconstrained, real-world images at various scales, or orientations. This variability has spurred extensive research into constructing and evaluating logo datasets for advanced recognition algorithms. Early efforts, such as BelgaLogos [7], were developed to assess logo retrieval from real-world images. BelgaLogos comprises 10,000 manually annotated images featuring 26 logos, laying the groundwork for subsequent studies. FlickrLogos-32 [8] expanded this scope to 32 logos and brands, albeit with a lower total image count. Although these early datasets offered high-quality annotations, their limitations in scale and brand diversity became apparent with the rise of deep learning methods [9].

To overcome these challenges, later work leveraged automated data collection techniques. For instance, LogoNET [9] gathered data by crawling online retail platforms,

TABLE 1: A comparison between existing logo datasets.

Dataset	# of logos	# of brands	# of annotated images	Release Year	Dynamic	Publicly Available
BelgaLogos	26	24	10,000	2009	✗	✓
FlickrLogos-32	32	32	8,240	2011	✗	upon request
Logo-NET (logos-160)	160	100	73,414	2015	✗	✗
WebLogo-2M	194	194	1,867,177	2017	✗	✗
Logo-2K+	2,341	2,341	167,140	2019	✗	✓
LogoDet-3K	3,000	2,864	158,652	2020	✗	✓
Our method	1,811	1,680	N/A	2025	✓	✓

resulting in two variants: i) logos-18 for small-scale applications and ii) logos-160 for larger-scale studies, with the latter offering nearly nine times more annotated images than FlickrLogos-32. WebLogo-2M [10] further pushed the boundaries by amassing nearly 2 million images across 194 brands. Additional datasets, such as Logo-2K+ [11] and LogoDet-3K [12], also contributed to increasing the diversity and volume of available logo data (see Table 1).

Unlike traditional logo datasets gathered via web scraping or manual annotation, our approach leverages BIMI to build a verifiable, high-integrity logo dataset with a novel perspective on dataset creation. The focus is on robust and reliable data, and with the growing deployment of BIMI [14], the coverage is continuously expanding. The next section provides essential background on BIMI, detailing its operation and role in our approach.

3. Background on BIMI

Introduced in 2021, Brand Indicators for Message Identification (BIMI) [13] leverages the Domain-based Message Authentication, Reporting, and Conformance (DMARC) [15] framework to help recipients verify the authenticity of senders by displaying a visual indicator (i.e., a logo) in supported email clients—provided that strict authentication requirements are met.

BIMI operates along the following key axes:

- **Authentication Prerequisite:** BIMI requires strict DMARC enforcement (i.e., a policy set to `p=quarantine` or `p=reject`) [15], ensuring that only authenticated emails qualify for logo display.
- **Logo Validation:** Domain owners must submit their logo to a Mark Verifying Authority (MVA) to obtain a Mark Certificate (MC). This digital certificate cryptographically binds the logo to the domain, preventing unauthorized use.
- **DNS Configuration:** A BIMI assertion is published as a `TXT` record containing URLs that point to both the logo image and the MC, enabling email clients to retrieve and display the logo with authenticated messages.

For a given domain—for example, `exa.com`—a query for its default BIMI DNS `TXT` record (typically located at `default._bimi.exa.com`) returns the data structured as shown in Figure 1. In this record, the `v` tag specifies the BIMI version (e.g., BIMI1), the `l` tag indicates the URL of the logo image, and the `a` tag provides the URL of the Mark Certificate that confirms the logo authenticity. This output supplies email clients with the necessary information to display a verified brand logo alongside authenticated messages.

```
"v=BIMI1; l=https://exa.com/logo.svg;
a=https://exa.com/document.pem"
```

Figure 1: Format of the BIMI assertion record.

3.1. Logos

To ensure a proper display across various email clients, the logos used for BIMI must comply with specific requirements. These requirements, as detailed in the draft RFC [13], include:

- **Format.** The logo must be in SVG (Scalable Vector Graphics) format, specifically SVG Tiny Portable/Secure (SVG P/S).
- **Dimensions.** It should have a square aspect ratio, i.e., equal width and height.
- **Size.** The SVG file must not exceed 32 KB in size.
- **Background.** A solid color background is required, as transparent backgrounds may not render consistently across all email clients.
- **Centering.** The logo should be centered within the SVG file to ensure consistent display.

Additionally, the draft mandates the use of HTTPS for transport.

3.2. Mark Certificate (MC)

A Mark Certificate (MC) issued by a trusted Mark Verifying Authority (MVA) asserts a cryptographically verifiable and auditable binding between an identity, a logo, and a domain [16]. There are two types of MCs: **Verified Mark Certificates (VMCs)** and **Common Mark Certificates (CMCs)**, both of which are published in CT logs [17].

VMCs are for entities formally holding **Registered Trademarks** with government intellectual property offices or possessing official **Government Marks**, whereas **CMCs** extend to a broader range of brand identifiers including **Prior Use Marks** and **Modified Registered Trademarks**. A **Prior Use Mark** refers to a brand name, logo, or other identifying symbol that has been consistently used in commerce to identify and distinguish the goods or services of a particular party before it was formally registered as a trademark by that party or another entity. A **Modified Registered Trademark** refers to a registered trademark that has undergone some form of alteration or variation from its originally registered form (minor stylistic changes or more substantial alterations of the logo, design, or even the wording of the mark). VMCs can be distinguished from CMCs based

on the Mark Type in the certificate (field with OID `1.3.6.1.4.1.53087.1.13`).

CMCs allow organizations with non-trademarked logos to obtain certificates if they can demonstrate at least 12 months of consistent logo usage. However, some email providers, such as Gmail, display a checkmark as an indicator of authenticity solely for VMCs while CMCs show the logo only [18].

4. Methodology

In this work, we focus on collecting the content of BIMI `TXT` assertion records, validating the Mark Certificates (MCs), and—when certificates are properly configured—analyzing the mapping between the brand names, their logos, and the associated domain names. Unlike prior work [19], we do not study BIMI deployments nor assess its misconfigurations.

4.1. Dataset Construction

We first compile a comprehensive list of domain names from multiple data sources. Specifically, we aggregate domain names from: i) the generic Top-Level Domain (gTLD) zone files provided by the ICANN Centralized Zone Data Service (CZDS) [20], ii) passive DNS data from SIE Europe [21], iii) Google certificate transparency logs [22], and iv) the Tranco 5M list [23] [24] generated on January 22, 2025. We then refine this dataset by using the Mozilla public suffix list [25] to extract only organizational domain names. This process results in a list of 513 million unique domain names (not all of which may be actively registered).

4.2. Measurements

Our measurement methodology focuses on mapping the domain names to their associated BIMI records. For each domain in our dataset, we query the BIMI `TXT` record using the default selector (i.e., `default._bimi.<domain>`). This record (if exists) contains the key information for the BIMI-based logo verification: the URL of the logo image and the URL of the Mark Certificate (MC) that validates the logo.

We adopt an approach similar to that used in previous work [19], extending its scope from the Tranco 1M list and a sample of potentially malicious domains to a comprehensive global scan of the domain name population. Our measurements are performed using `zdns` [26], which enables efficient querying of the BIMI `TXT` records across our large domain dataset.

Finally, for each BIMI record, we download the MC from the URL specified in the `a` tag and retrieve the corresponding logo from the URL provided in the `l` tag.

4.3. MC Validation

Unlike previous work [19], we validate certificates and discard misconfigured ones. Without proper verification, an adversary could register a domain and embed a brand logo within the BIMI assertion record, enabling impersonation. Our manual analysis has revealed such behavior in cases involving Apple, Microsoft, Meta and PayPal for which domains advertise logos without a valid legitimate certificate (see Figure 5 in Appendix). The robust certificate verification is therefore essential to protect the integrity of the final logo dataset.

To validate mark certificates, we implement the essential checks outlined in the draft RFC on VMC fetch and validation [27]. We consider a certificate valid if and only if it meets the following conditions:

- **Certificate Authenticity:** Verify the signatures and ensure that the end-entity certificate issuance chain leads to a BIMI root CA. Confirm that the root CA is included in the trusted BIMI roots, as defined by the path validation process in Section 6.1 of RFC5280 [28]. At the time of writing, only DigiCert, Entrust, and GlobalSign are authorized to issue MCs [29].
- **Chain Validity:** Check the validity of all certificates in the chain using the procedures outlined in Section 4.1.2.5 of RFC5280 [28].
- **CT Logging Proof:** Validate the presence of at least one Signed Certificate Timestamp (SCT) within the X.509 certificate.
- **MC Verification:** Confirm that the end-entity certificate is a Mark Certificate by ensuring that the `Extended-Key-Usage` extension includes the identifier `id-kp-BrandIndicator-forMessageIdentification`.
- **Domain Verification:** Verify that the domain name for which the BIMI `TXT` record was published is consistent with the domain names listed in the certificate Subject Alternative Name (SAN) field.
- **Logo Consistency:** Check for the presence of the logotype extension and compare the logo specified in the `l` tag of the BIMI record with the logo embedded in the certificate.

4.4. Data Aggregation

In the final step, we aggregate organization names from the common names in the validated MCs and compile all associated domain names. For each organization, we deduplicate logos by computing the SHA256 hash of each SVG file and retaining those with distinct hashes. Figure 3a in Appendix shows an example of a cluster made of duplicated images corresponding to the BIMI records of `babycare.de` and `baby-care.de`.

We also explored an alternative approach for near-duplicate identification using Perceptual Hashing (pHash), where the similarity between a pair of images is quantified by the Hamming distance between their corresponding hashes —smaller distances indicate greater visual similarity. This approach can potentially cluster logos that, while not identical, share significant visual characteristics. Figure 3b in Appendix illustrates the extent to which similar logos are grouped despite differences in color. However, due to the introduction of false positives (as shown in Figure 4 in Appendix) and our emphasis on reliability, we ultimately opted to use only SHA256 hashing for deduplication.

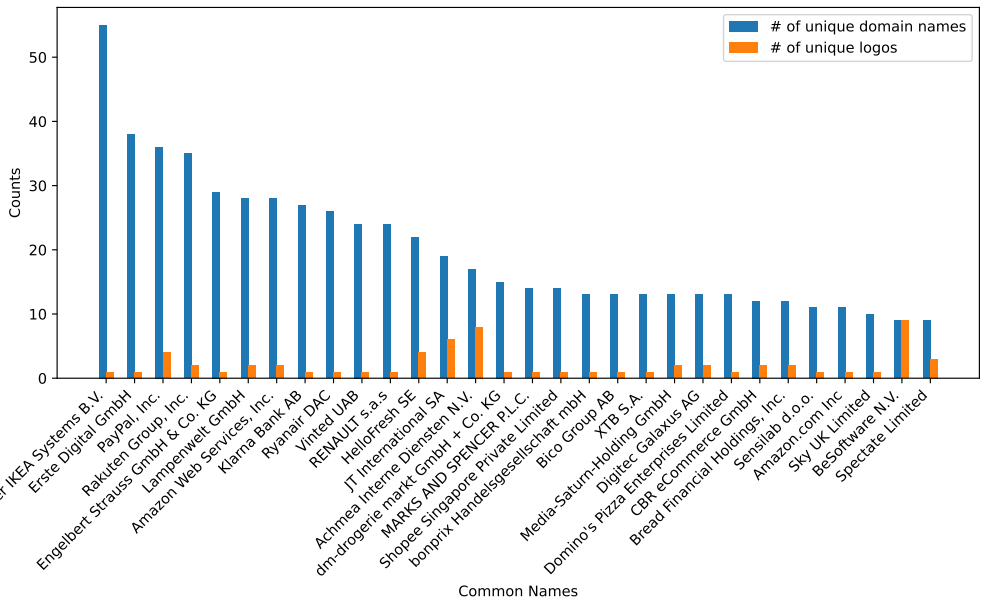


Figure 2: Number of unique domain names and logos per common name

5. Results

Upon completion of our measurements, we collected a total of 55,650 logos each corresponding to exactly one domain name publishing a BIMI **`TXT`** record, and 5,430 certificates. We began by removing duplicate certificates, which accounted for 1,771. Among the remaining 3,659 certificates, only 2,821 met the criteria for valid MCs as described in Section 4.3. Looking at the **`subject:markType`** field reveals that 95.4% of the valid certificates were VMCs with 2682 **Registered Marks** and 9 **Governement Marks** while CMCs only represent 4.6% with 118 **Prior Use Marks** and 12 **Modified Registered Marks**. Our logo deduplication approach yielded 42,845 clusters indicating that approximately 23% of the collected logo images were redundant.

Figure 2 shows the distribution of domain names and BIMI logos for the 30 leading common names, sorted by domain name count. The majority of brands (94.22%) use a single BIMI logo. For example, IKEA, a multinational company, employs a single logo across more than 50 domains, while the average domain count per common name is 1.6.

The results reveal that the proposed methodology not only achieves broad brand coverage despite the early deployments of BIMI, but also captures the corresponding domain names (e.g., PayPal and Rakuten each cover more than 30 domain names). In the context of phishing, this approach may help ensure a low false positive rate: when ccTLDs or even defensively registered domain names (provided the BIMI certificates cover them) are encountered, they will not be mistakenly classified as phishing solely based on logo detection.

Finally, when comparing the raw counts of covered brands and logos to prior methods (see Figure 1), the proposed approach shows the same orders of magnitude as the methods with the best coverage (i.e., Logo-2K+ and LogoDet-3K). However, further work is needed to compare the coverage of brands and logos quantitatively.

6. Limitations

Despite its advantages for logo collection and brand verification, BIMI has several inherent limitations. First, the resulting dataset is intrinsically tied to the current BIMI adoption rate. Thus, the repository represents only the brands participating in BIMI, potentially excluding many others. Nevertheless, as BIMI adoption is expected to increase, the method should yield more comprehensive data over time.

Second, our measurement methodology exclusively considers the default BIMI selector. Consequently, if a domain owner specifies its logo using a custom selector without providing one for **`default`**, our approach fails to detect the logo. For example, querying the DNS **`TXT`** record of **`default._bimi.facebookmail.com`** returns an empty response. However, using the **`fb2023q1v3`** selector instead of **`default`** returns an answer. Nevertheless, this limitation (and certificate verification process) could be overcome by fetching MCs directly from the CT logs.

Third, although MCs enhance verification, their optional publication within BIMI introduces ambiguity regarding brand ownership and logo authenticity. To ensure robustness, we exclude organizations that do not publish these certificates, thereby sacrificing coverage.

Fourth, the BIMI specification dictates the publication of a single logo per certificate, optimized for email display. This contrasts with typical logo datasets that often require multiple variations (e.g., different resolutions, color schemes, or styles) to support diverse applications. Consequently, the BIMI-derived dataset may lack the breadth and diversity needed for certain logo analysis tasks. Nevertheless, the present logo dataset could still serve as a resource for data augmentation.

7. Conclusion and Future Work

In this paper, we have presented an original rigorous methodology for collecting brand logos using BIMI. Our

final dataset consists of 1,680 brands, 1,811 logos, and 2,821 corresponding domain names. Even if the number of the collected logos may appear as relatively small, we can observe that most of phishing attacks target the considered brands. In fact, the incentive to impersonate a widely-known brand, able to afford an MC, is often higher than for little-known brands.

While BIMI is an adequate way for brand logo retrieval, we are aware of its several limitations, including its dependence on the degree of adoption, the optional nature of the mark certificate publication, and the single-logo constraint.

In future work, we plan to conduct a comprehensive evaluation of our findings, including a detailed analysis of the added value of our approach to phishing detection, which will involve enhancing existing datasets of logos and organization names (i.e., potential phishing targets) and assessing both the domain name coverage and the overall accuracy of logo-based phishing detection methods, such as Phishpedia.

References

[1] K. L. Chiew, E. H. Chang, S. N. Sze, and W. K. Tiong, "Utilisation of Website Logo for Phishing Detection," *Computers & Security*, vol. 54, pp. 16–26, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404815001145

[2] A. S. Bozkir and M. Aydos, "LogoSENSE: A Companion HOG Based Logo Detection Scheme for Phishing Web Page and E-mail Brand Recognition," *Computers & Security*, vol. 95, p. 101855, 2020.

[3] T. van den Hout, T. Wabeke, G. C. M. Moura, and C. Hesselman, "LogoMotive: Detecting Logos on Websites to Identify Online Scams - a TLD Case Ctudy," in *Proc. PAM*, 2022.

[4] Y. Lin, R. Liu, D. M. Divakaran, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si, F. Zhang, and J. S. Dong, "Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages," in *Proc. USENIX Security*, 2021.

[5] R. Liu, Y. Lin, X. Yang, S. H. Ng, D. M. Divakaran, and J. S. Dong, "Inferring Phishing Intention via Webpage Appearance and Dynamics: A Deep Vision Based Approach," in *Proc. USENIX Security*, 2022.

[6] S. Maroofi, M. Korczyński, C. Hesselman, B. Ampeau, and A. Duda, "COMAR: Classification of Compromised versus Maliciously Registered Domains," in *Proc. IEEE EuroS&P*, 2020, pp. 607–623.

[7] A. Joly and O. Buisson, "Logo Retrieval with a Contrario Visual Query Expansion," in *Proc. ACM Multimedia*, 2009.

[8] S. Romberg, L. G. Pueyo, R. Lienhart, and R. van Zwol, "Scalable Logo Recognition in Real-World Images," in *Proc. ACM Multimedia Retrieval*, 2011.

[9] S. C. H. Hoi, X. Wu, H. Liu, Y. Wu, H. Wang, H. Xue, and Q. Wu, "LOGO-Net: Large-scale Deep Logo Detection and Brand Recognition with Deep Region-based Convolutional Networks," *CoRR*, vol. abs/1511.02462, 2015.

[10] H. Su, S. Gong, and X. Zhu, "WebLogo-2M: Scalable Logo Detection by Deep Learning from the Web," in *Proc. IEEE ICCVW*, 2017.

[11] J. Wang *et al.*, "Logo-2K+: a large-scale logo dataset for scalable logo classification," in *Proc. AAAI*, 2020.

[12] ——, "LogoDet-3K: A Large-scale Image Dataset for Logo Detection," *ACM Trans. Multimedia Comput. Commun. Appl.*, Jan. 2022.

[13] S. Blank, P. Goldstein, T. Loder, T. Zink, M. Bradshaw, and A. Brotman, "Brand Indicators for Message Identification (BIMI)," IETF Internet Draft draft-brand-indicators-for-message-identification-08, 2024.

[14] "Bimi radar," https://bimiradar.com/, accessed: 2025-04-04.

[15] M. Kucherawy and E. Zwicky, "Domain-based Message Authentication, Reporting, and Conformance (DMARC)," RFC 7489, 2015.

[16] BIMIGroup, "Minimum Security Requirements for Issuance of Mark Certificates," 2025, https://bimigroup.org/resources/VMC-_Requirements_latest.pdf.

[17] https://gorgon.ct.digicert.com/log, accessed: 2025-03-04.

[18] https://workspaceupdates.googleblog.com/2024/09/gmail-additional-bimi-protections.html, accessed: 2025-03-04.

[19] M. Yajima, D. Chiba, Y. Yoneya, and T. Mori, "A First Look at Brand Indicators for Message Identification (BIMI)," in *Proc. PAM*, A. Brunstrom, M. Flores, and M. Fiore, Eds., 2023, pp. 479–495.

[20] https://czds.icann.org/, accessed: 2025-03-04.

[21] https://www.sie-europe.net/, accessed: 2025-03-04.

[22] "Certificate transparency in chrome," https://googlechrome.github.io/CertificateTransparency, accessed: 2025-03-04.

[23] https://tranco-list.eu/download/4QJ3X/full, accessed: 2025-03-04.

[24] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation," in *Proc. NDSS 2019*, Feb. 2019.

[25] https://publicsuffix.org/, accessed: 2025-03-04.

[26] L. Izhikevich, G. Akiwate, B. Berger, S. Drakontaidis, A. Ascheman, P. Pearce, D. Adrian, and Z. Durumeric, "ZDNS: a Fast DNS Toolkit for Internet Measurement," in *Proc. ACM IMC*, 2022.

[27] W. Chuang, M. Bradshaw, T. Loder, and A. Brotman, "Fetch and Validation of Verified Mark Certificates," IETF Internet Draft draft-fetch-validation-vmc-wchuang-08, 2024.

[28] S. Boeyen, S. Santesson, T. Polk, R. Housley, S. Farrell, and D. Cooper, "Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile," RFC 5280, 2008.

[29] "Mark certificates issuers," https://bimigroup.org/vmc-issuers/, accessed: 2025-04-04.

[30] D. Dittrich and E. Kenneally, "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research," https://catalog.caida.org/paper/2012_menlo_report-_actual_formatted, 2012.

[31] C. Partridge and M. Allman, "Ethical Considerations in Network Measurement Papers," *Commun. ACM*, vol. 59, no. 10, p. 58–64, 2016.

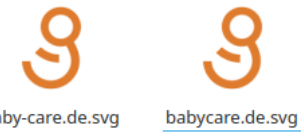
[32] Z. Durumeric, E. Wustrow, and J. A. Halderman, "ZMap: Fast Internet-Wide Scanning and Its Security Applications," in *Proc. USENIX Security*, 2013, pp. 605–620.

Appendix

Ethical Considerations and Reproducibility

Our research uses active network measurements while adhering to industry best practices [30]–[32]. We randomize our input list of domain names to distribute the load across various authoritative nameservers and over time. Additionally, we query the Google's public resolver that functions as a caching resolver. As a result, some of our requests are likely to be served from its internal cache.

We make the mapping between organization names, domain names, and their corresponding logos publicly available at https://github.com/josef0x/LogoTrust.

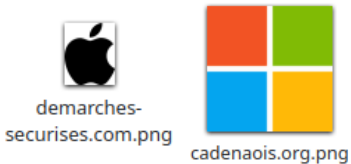


(a) Cluster constructed using SHA256 hashes

(b) Cluster built using pHash

Figure 3: Two examples of clusters

Figure 4: Example of a cluster with false positives



(a) Apple (b) Microsoft

(c) Meta (d) Paypal

Figure 5: Cases of impersonation targeting Apple, Microsoft, Meta, and Paypal using their logos