Ctrl+Alt+Deceive: Quantifying User Exposure to Online Scams

Platon Kotzias*§, Michalis Pachilakis*‡, Javier Aldana Iuit *,
Juan Caballero†, Iskander Sanchez-Rola* and Leyla Bilge*
*Norton Research Group
†IMDEA Software Institute
‡Computer Science Department, University of Crete
§BforeAI

***Abstract*—Online scams have become a top threat for Internet users, inflicting $10 billion in losses in 2023 only in the US. Prior work has studied specific scam types, but no work has compared different scam types. In this work, we perform what we believe is the first study of the exposure of end users to different types of online scams. We examine seven popular scam types: shopping, financial, cryptocurrency, gambling, dating, funds recovery, and employment scams. To quantify end-user exposure, we search for observations of 607K scam domains over a period of several months by millions of desktop and mobile devices belonging to customers of a large cybersecurity vendor. We classify the scam domains into the seven scam types and measure for each scam type the exposure of end users, geographical variations, scam domain lifetime, and the promotion of scam websites through online advertisements.**

We examine 25.1M IP addresses accessing over 414K scam domains. On a daily basis, 149K devices are exposed to online scams, with an average of 101K (0.8%) of desktop devices being exposed compared to 48K (0.3%) of mobile devices. Shopping scams are the most prevalent scam type, being observed by a total of 10.2M IPs, followed by cryptocurrency scams, observed by 653K IPs. After being observed in the telemetry, the scam domains remain alive for a median of 11 days. In at least 9.2M (13.3%) of all scam observations users followed an advertisement. These ads are largely (59%) hosted on social media, with Facebook being the preferred source.

I. Introduction

Online scams have become a top threat for Internet users. Only in 2023, the Federal Trade Commission (FTC) received over 2.6M reports of online fraud from US citizens, with total monetary losses of $10 billion [22]. Similarly, the Australian Competition and Consumer Commission (ACCC) received over 280K scam reports from Australian citizens with total losses of $455M [63]. Beyond financial losses, online scams often inflict profound emotional distress, including feelings of betrayal, embarrassment, and vulnerability[26], [29], which can even lead victims to commit suicide [88], [70].

Prior work has studied, and proposed defenses against, specific scam types including technical support scams [40], [46], [57], [84], shopping scams [15], [45], [101], [19], [100], [43], [84], [10]. romance scams [86], [6], pet scams [68], cryptocurrency and non-fungible token (NFT) scams [9], [28], tax-related scams [14], survey scams [43], and gaming scams [10]. However, it is not clear how those scam types compare. Furthermore, those works focus on the scam websites, without considering how often those scam websites are visited by users. Understanding the exposure of users to scams is an important, but challenging, problem that can lead to improvements in defenses and help prioritizing security investments.

Reports from consumer protection agencies such as FTC, ACCC, or the Better Business Bureau (BBB) cover multiple scam types, but they focus on one country [22], [63] or a few nearby countries [33]. Furthermore, these reports heavily rely on individuals reporting scams affecting them. Unfortunately, only a small fraction of victims are willing to report scams [92], [4].

In this work, we perform what we believe is the first quantitative study of user exposure to different types of online scams. For this, we leverage two data sources. First, we obtain 607K scam fully qualified domain names (FQDNs) from two feeds: 341K from the commercial ScamAdviser scam detection service [75] and 289K from a machine learning (ML) shopping scam detector used internally by a large cybersecurity vendor. These 607K scam FQDNs belong to 501K second level domains (SLDs). Our scam domain dataset is two orders of magnitude larger than those in previous scam studies [47], [15], [43], [10] and 1.5 times larger than phishing domain datasets [62]. Second, we leverage the telemetry of a large cybersecurity vendor covering millions of desktop and mobile devices located in over 230 countries.

Our approach scans 10 months of telemetry data using the 607K scam domains to identify observations of scam domains by real devices. To compare different scams, we classify the 607K scam domains into seven types: shopping, cryptocurrency, financial, gambling, dating, recovery services, and employment. Shopping, financial, and cryptocurrency scams consistently rank as the top three scam types in terms of victim reports or monetary losses, while gambling, dating, and employment scams are typically ranked in the top 10 [22], [33], [63]. For the classification, we leverage industry tags available in ScamAdviser, whose accuracy we evaluate, as well as shopping scam labels provided by the ML detector.

[§] The lead author was affiliated with Norton Research Group for the majority of this work, which was subsequently completed under affiliation with BforeAI

Network and Distributed System Security (NDSS) Symposium 2025
24-28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2025.241816
www.ndss-symposium.org

This approach assigns a type to 330K (54.5%) scam domains. We also analyze the lifetime of scam domains by measuring the *wait time* from domain registration to first observation, the *activity time* from first to last observations, and the *listing delay* from first observation to appearance in the scam feeds. Additionally, we analyze what fraction of scam domains are reached by victims by following online advertisements, and which advertisement platforms (e.g., social networks) are preferred by the scammers. For this, we analyze the usage in scam URLs of Urchin Tracking Module (UTM) parameters, used by popular analytics libraries such as Google Analytics [37]. Finally, we estimate the potential impact of scams on users by examining what fraction of users visit checkout or payment pages that may indicate they were scammed.

Our analysis answers the following research questions:

What is the overall user exposure to scams? Of the 501K SLDs in the scam feeds, 415K (82.7%) are observed in the desktop and mobile telemetry receiving visits by 25.1M IP addresses. Each day, over 149K devices of the vendor's customers are exposed to online scams. Desktop devices are more exposed (101K) than mobile devices (48K). After accounting for population differences, more than twice the fraction of desktop devices (0.8% daily IPs observed in the telemetry) is exposed compared to mobile devices (0.3%).

Are there geographical differences among scams? Exposure to scams varies significantly among countries. Top exposed countries are mostly from the European Union and have up to 10 times higher ratio of exposed IPs compared to the least exposed countries, most of them from Asia and America.

What are the most prevalent scam types? Users are most exposed to shopping scams with 10.2M affected IPs, followed by cryptocurrency (653K) and financial (443K) scams. Exposure to dating, gambling, employment, and funds recovery scams is significantly smaller with 3x-10x less exposed IPs.

What is the lifetime of scams? We examine the lifetime of the 242K scam domains registered since January 1, 2023. Of those, 223K (92%) are first seen in the telemetry, appearing in the scam feeds a median of 1 day later. After first observation, the scam domains are active for a median of 11 days, with dating and shopping scams having the longest activity of 59 and 15 days, respectively, while funds recovery, financial, and cryptocurrency scams are only active for 1 day. Compared to phishing domains, scam domains remain active 12–15 times longer.

What fraction of scam sites are reached through online advertisements? In 9.2M (13.3%) scam observations users followed an advertisement leading to a scam domain, 38.9K (9.7%) of scam domains are promoted through advertisements with 3.1M (15.3%) of IPs observing a scam advertisement. We observe 5.4M (59%) scam advertisements placed on social media. From all the advertisements with a UTM source parameter, 4.9M (75%) are placed on Facebook. Shopping and cryptocurrency scams attract the most users via advertisements.

What fraction of users are scammed? Of the 10.1M IPs visiting a shopping scam website, 411K (4%) reach the checkout page, indicating that they intend to complete a purchase.

II. Background & Related Work

A wealth of related work has analyzed scams. Most works have focused on studying a specific scam type including technical support scams [40], [46], [57], [84], romance scams [86], [6], pet scams [68], cryptocurrency and non-fungible token (NFT) scams [9], [28], tax-related scams [14], survey scams [43], and gaming scams [10]. Other works have focused on studying the distribution of scams campaigns through social media [16], [39], [40], [87] and the social engineering attacks enabling such scams [44], [65], [97], [27].

Our analysis does not focus on a single scam type, but compares seven prevalent scam types the victims observe, described below.

Shopping scams. Victims of this scam type place an order in an online store, but receive nothing, or receive a product that does not match the advertised one, e.g., a counterfeit product. Beyond the financial loss, victims may also expose their personal and payment information to the scammers. Shopping scams are also known as fake e-commerce websites, e-commerce scams, online purchase scams, and non-delivery scams. Shopping scams often attract victims by claiming to offer products at unusually low prices [71]. The stores look deceptively similar to legitimate online stores and may sell a wide range of products such as electronic equipment [30], clothes [19], furniture [3], and pharmaceuticals [53]. Shopping scams have been ranked by BBB as the top consumer risk for three years in a row until 2022 [32] and third riskiest in 2023 [33]. Similarly, the FTC ranks shopping scams as the second-highest by number of victims, with $392M reported losses in 2023 [22]. Previous work has proposed approaches for detecting shopping scams [15], [45], [101], [19], [100], [43], [84], [10].

Financial scams. This scam type involves websites promoting various investment opportunities, for example, on foreign currency (Forex), real estate, and high-yield investment programs (HYIP) [21]. It might also include blogs and sites giving iffy investment advice and promoting frauds like penny stock deals [52], [51]. They are also called investment scams and attract investors by promising unrealistically high returns. Investors do not obtain the promised returns and are not allowed to withdraw their original investment [77], [76]. Instead, scammers try to convince victims to invest more by requesting additional fees to allow a withdrawal that never happens. The FTC ranked investment scams as the most financially damaging scam in 2023 with reported total losses of $4,462 million [22]. The BBB ranked it first in 2023, with victims experiencing median losses of $3,800 per incident [33].

Cryptocurrency scams. These scams involve a payment in cryptocurrencies. We only consider scams that use a webpage to attract victims. These include cryptocurrency-focused investment scams such as token scams [103], mining investment scams [82], giveaway scams [98], [47], and ponzi scams [12], as well as exchange impersonation scams [104]. Out of scope are email-based scams such as sextortion [64] and malware-enabled attacks such as ransomware [48], [42], clippers [35], and cryptojacking [91]. While most scams in this category could be considered financial scams, keeping both categories

Dataset	Start	End	Data
Desktop telemetry	2023-01-01	2023-11-10	196.9B URL visits
Mobile telemetry	2023-01-01	2023-06-16	112.2B Domain visits
ScamAdviser	2023-03-20	2023-11-10	341K Scam FQDN (236K SLD)
Shopping scams	2023-01-01	2023-11-10	289K Scam FQDN (289K SLD)
All scams	2023-01-01	2023-11-10	607K Scam FQDN (501K SLD)

Table I: Datasets summary.

separate provides a finer-grained classification and allows merging results for both categories when desired.

Gambling scams. These websites allow visitors to engage in a variety of gambling activities such as sports betting and online casinos. They also include blogs and sites that push users to join gambling scams and give false reviews on lottery systems [74]. Similar to investment scams, victims of gambling scams are often unable to withdraw their earnings from the platform [93]. Some countries like China forbid online gambling, which has led to the proliferation of illegal gambling sites [105], [34]. However, illegal gambling sites may not be scams.

Dating scams. These websites offer deceptive adult contact subscription services. They create a false appearance of authenticity to attract individuals seeking romantic or sexual relationships. Once subscribed, victims discover the service is filled with fake profiles using stolen photos and made-up personal information [86], [6]. Additionally, victims report interacting only with automated bots instead of real humans [94]. Unsubscribing from these services is challenging and victims may keep receiving subscription charges after unsubscribing [95].

Funds recovery. These websites promise victims of other scams to recover their lost funds in exchange for an upfront fee, but the funds are never recovered and the fee is not returned to the victim [11]. They are particularly nasty as they offer false hope of restitution to individuals that have already been victimized. They are also known as recovery room fraud [80] and can be considered a sub-type of tech support scams [2]. Oftentimes, funds recovery scams focus on victims of cryptocurrency scams, but also include the recovery of non-crypto funds.

Employment scams. Also known as job scams, these websites advertise fake job opportunities or services for assisting individuals in finding employment [5]. Victims of employment scams end up paying for starter kits or useless certifications, receive charges on their credit cards due to unexpected subscriptions, or are recruited to participate in shady tasks (e.g., writing fake reviews, generating fake traffic, reshipping scams [41]) without any financial return.

III. Datasets

For our analysis, we use telemetry data capturing URLs and domains visited by real users through desktop and mobile devices (detailed in Section III-A) and feeds of scam domains (detailed in Section III-B). Datasets are summarized in Table I.

A. Telemetry

We obtain telemetry data about URLs visited using desktop and mobile devices by customers of a large cybersecurity vendor that have opted in to the collection. The telemetry is collected by the vendor's security products and only includes data from users who install company's products, accept the company's privacy policy, and opt-in to share their data. To prevent user deanonymizaiton, device information is collected, processed, and stored in a privacy-preserving manner without device identifiers and in aggregate.

Desktop telemetry. This dataset contains URLs visited by 11M Windows desktop devices. The desktop telemetry is collected from a browser extension that supports the most popular browsers (i.e., Chrome, Firefox, Edge). When a user visits a URL, the extension sends a query to a backend server to obtain the URL reputation. The desktop telemetry covers 196.9 billion URL visits observed over 10 months (314 days) from January 1, 2023 until November 10, 2023. For each visit, the dataset contains a timestamp, the full URL visited, the hash of the client's IP address, and the user's country code obtained by geolocating the device's IP address.

Mobile telemetry. This dataset contains domains visited by 14M Android and iOS devices that have installed the vendor's VPN app. The dataset contains FQDNs in DNS requests and TLS Client Hello SNI headers. The domains may have been visited using different browsers and mobile apps installed on the devices. Content of the mobile telemetry is similar to the desktop telemetry but it captures visited domains rather than the visited URLs. The dataset contains 112.2 billion domain visits spanning 5 months (167 days) from January 1, 2023 until June 16, 2023.

The device IP addresses are geolocated in 234 (desktop) and 232 (mobile) country codes [1], thus covering nearly all countries in the world. However, not all countries are equally represented with North America, Europe, and Japan concentrating most devices. We will focus our geographical analysis of the 46 countries for which we observe at least 10K IP addresses daily.

B. Scam Domain Feeds

We use two feeds of scam domains to identify encounters of scam websites by devices in the telemetry. Overall, we collect 607K scam FQDNs: 341K from the ScamAdviser [75] commercial service and 289K from the vendor's shopping scam detector.

ScamAdviser. We obtain access to the commercial feed of ScamAdviser [75], a scam detection service to which users can submit a domain and receive a report about whether the domain hosts a scam website. The feed contains all reports for domains analyzed by ScamAdviser throughout nearly 8 months (234 days) from March 20, 2023 until November 10, 2023. Each report contains the date of the analysis, the analyzed domain, the first date the domain was analyzed by ScamAdviser (multiple users can request analysis of the same domain over time), a set of tags that categorize the content of the website hosted on the domain, the domain's WHOIS information, the number of reviews the domain received on third-party review platforms (e.g., TrustPilot [96], SiteJabber [81], Google Business [38]), and a trust score in the range [0,100] with low scores indicating the site is likely a scam and high scores indicating the site is trustworthy.

Over the 8 months, the ScamAdviser feed contained 21.1M reports for 14.7M fully qualified domain names (FQDNs) belonging to 11.8M SLDs. While the reports start on March 20, 2023, the domains were first analyzed by ScamAdviser as far back as September 2020. Figure 2 in the Appendix shows the trust score distribution for all FQDNs in the ScamAdviser feed. Of the 14.7M FQDNs, 6.8M (46.4%) have a trust score of at least 80 (indicating that they are potentially benign), 1.7M (11.4%) have a trust score less or equal to 10 (very likely scam), and 6.2M (42.2%) receive a score from 11 to 79.

We initially identify the 1.7M FQDNs with a trust score up to 10 as likely scam domains. Since ScamAdviser's detection process is proprietary, and thus its performance metrics are unknown, we err on the side of caution and perform two extra filtering steps that remove potential false positives (FPs) in the feed, at the expense of a smaller, but higher quality, dataset. First, we remove 67.3K (4%) domains appearing in the Tranco Top 1M [67] popularity list[§]. This step could remove popular scam domains if they manage to make it into the Tranco Top 1M. However, Tranco is designed to minimize those cases and we prefer to err on the side of caution, removing potential FPs. Second, during June 2024, we queried all ScamAdviser domains on VirusTotal [99]. We removed any scam domain that was unknown to VirusTotal or that received less than two detections. After the second filtering step, we retain 341K highly likely scam domains from ScamAdviser. Of those, 236K (69.2%) are SLDs and 105K (30.8%) contain a subdomain.

From October 11, 2023 and until the end of our analysis period on November 10, 2023, we crawled the likely scam domains that appeared in the ScamAdviser feed. The crawler visits the scam domains as soon as they appear in the feed, although small delays may be introduced due to queues, errors, and capacity limits. The crawler is built on top of Puppeteer [69], follows redirections, and is protected from cloaking using the *puppeteer-extra-plugin-stealth* add-on. Of the crawled domains, 130K returned a crawling error due to unsuccessful domain resolutions, timed-out connections, and connections refused by the server. For the remaining 211K domains, the crawler stores the HTML page with all its resources including HTTP requests and their return status, cookies, HTTP re-directions (if any), and a screenshot of the rendered page. We exclude a further 43K domains that return HTTP errors (4XX, 5XX, 6XX, 9XX), leaving 167K domains. Then, we use regular expressions to also exclude 21K domains returning non-meaningful content including default Web server pages, under construction messages, bot verification pages, and account suspended messages. We extract text from the remaining HTML pages using the *html2text* Python library [89]. We further filter webpages with 10 or less characters, to have enough meaningful content for us to analyze. After these filtering steps, 143,227 (42%) domains remain. We detect the language of the websites' text using *langid* [49]. It detect 96 languages with the most prevalent being English (63.4%), Chinese (7.7%), Japanese (3.9%), Indonesian (2.7%), and Russian (2.6%). We use the text and screenshot of the 143,227 domains as input to our scam domain classification in Section IV.

ScamAdviser Industry Tag	Domains	Prec.	Scam Type
Essay/Thesis/Dissertation Writers	323,060	0%	✗
Shopping	191,918	72%	Shopping
Cryptocurrency	90,873	80%	Cryptocurrency
Hacking - High Risk	72,264	0%	✗
Media - Games	65,658	12%	✗
Financial Service - Very High Risk	60,159	77%	Financial
Financial Services - High Risk	50,847	87%	Financial
Financial Services - HYIP	28,906	89%	Financial
Media - Movies	28,568	31%	✗
Gambling	18,453	90%	Gambling
Adult	17,209	26%	✗
Financial Services	13,825	100%	Financial
Media - Software	10,659	39%	✗
Sport Betting	10,229	58%	✗
Non-Profit Organization	9,924	24%	✗
Media - Books	8,225	27%	✗
Media Subscription Services	6,723	30%	✗
Adult - Dating	6,399	86%	Dating
File Sharing Service	5,943	27%	✗
Jobs	4,914	72%	Employment
Travel Services	3,290	65%	✗
NFTs	2,762	40%	✗
Recovery Services	1,331	84%	Funds recovery
Visa Services	1,121	66%	✗
Lending Service	432	62%	✗
Helpdesk - IT Support	406	40%	✗

Table II: ScamAdviser industry tags examined, precision measured, and scam type they are mapped to (if any).

Surprisingly, even though we crawled the scam domains soon after they appeared in the ScamAdviser feed, 179K (55%) were already dead by that time. The main reason is that only 1.3M (60%) of all 21.1M ScamAdviser reports are for fresh domains, the others are re-analysis of previously reported older domains, which may already be dead. As proposed by prior work [17], we also checked for domains blocked by their registrars by examining the WHOIS records for the presence of two Extensible Provisioning Protocol (EPP) status codes (*CLIENTHOLD* and *SERVERHOLD*). We observe that 4% of the unresolved domains were taken down by their domain registrars by the time they appeared in the feed.

Shopping scams. From the telemetry vendor, we obtain a dataset of domains identified by a ML-based detector specifically trained to identify shopping scams. The detector has been evaluated in a published work, achieving an F1-score of 0.973 with a precision of 0.988 and a recall of 0.959 [45]. The detections are weekly evaluated by analysts to ensure very low FPs. The shopping scams feed contains 289,434 FQDNs detected over 11 months from January 1, 2023 until November 10, 2023. Of those, 99.9% (289K) are SLDs and only 188 (less than 0.1%) contain a subdomain. Internally, the detector crawls the website content similar to what we implemented for the ScamAdviser domains.

IV. Scam Classification

This section explains our classification of scam domains into scam types. While the domains identified by the shopping scam detector have scam type information, the scam domains from ScamAdviser do not. ScamAdviser assigns industry tags

[§] Tranco permanent id: 828KV

to the domains in its feed. The advantage of these tags is that they are available for a significant fraction of domains in the ScamAdviser feed. The disadvantage is that ScamAdviser uses many tags and not all of them are accurate. To address this issue, we perform an accuracy evaluation that allows us to map a subset of ScamAdviser industry tags to 7 trustable types.

Industry tags provide information about the content of scam domains in the ScamAdviser feed. They are assigned based on the content of the main page of the scam domain. Among the 1.7M likely scam domains in ScamAdviser, i.e., prior to the two filtering steps detailed in Section III, 1.1M (63.7%) have no industry tags, 289K (17.0%) have one tag, and 330K (19.4%) have multiple tags. The 618K (36.4%) tagged domains use 42 distinct industry tags.

To evaluate the accuracy of the industry tags, four analysts performed a manual labeling exercise where they examined 26 industry tags: the 20 most prevalent (i.e., found in at least 1% of scam domains) and 6 additional ones whose names indicated scam types not covered in the top 20 (e.g., Help desk - IT Support). For each of those 26 tags, we randomly selected 100 domains to label. Labelers were tasked with assigning each domain a category based on their content (screenshot and text) and metadata. The labelers were provided the list of 26 industry tags and two generic categories (Other and Not enough information), but were allowed to refine the names of the categories and to add new ones. The labelers were not provided the industry tag ScamAdviser assigned to each domain. To establish consensus on the labeling process, the analysts performed two rounds of labeling in which each analyst separately labeled the same set of domains. After each round, the participants met, discussed, homogenized the categories, and revised the scam codebook. After the second round, we measured the inter-coder agreement using Fleiss' kappa statistic to be 0.756, indicating high agreement [31]. In the generated codebook, the labelers added a new category for SEO services and refined the names of 3 others: investment (instead of ScamAdviser's Financial Services), free downloads (Media - Software), and charity scams (Non-Profit Organization). Then, each labeler was randomly assigned a subset of the domains to label using the established codebook.

We compare the labeler-assigned tags to the ones assigned by ScamAdviser to determine the per-tag precision. Table II shows the 26 ScamAdviser industry tags, the number of ScamAdviser domains with the tag, the precision measured using the manual labeling, and the final type assigned to the tag. We select as trustable the 10 tags with at least 70% precision because there is an elbow in the distribution at that threshold, with most tags having lower precision. We also group the four tags ScamAdviser uses for financial services into a single type. This process outputs 7 scam types: financial, gambling, cryptocurrency, dating, shopping, employment, and funds recovery.

Table III summarizes the scam domain classification using these 7 types. It shows that 64K (19%) of ScamAdviser domains can be labeled, The other 276K (81%) domains remain *Unclassified* because they have no tags or they have low accuracy tags that are ignored. The most common scam types in ScamAdviser are shopping (13.7%), cryptocurrency (1.5%) and financial (1.2%). The classified domains increase to 330,898 (54.5%) when adding the domains from the shopping scam ML detector. We use these 330K classified domains throughout the paper to compare different scam types.

Scam Type	ScamAdviser	All Scams
Shopping	46,648 (13.7%)	312,783 (51.5%)
Cryptocurrency	8,919 (1.5%)	8,919 (2.6%)
Financial	7,230 (1.2%)	7,230 (2.1%)
Gambling	1,101 (0.2%)	1,101 (0.3%)
Employment	669 (0.1%)	669 (0.2%)
Dating	168 (<0.1%)	168 (<0.1%)
Funds recovery	28 (<0.1%)	28 (<0.1%)
Classified	64,763 (19.0%)	330,898 (54.5%)
Unclassified	276,329 (81.0%)	276,329 (45.5%)
Total	341,092 (100%)	607,227 (100%)

Table III: Scam domain classification results for ScamAdviser domains and all scam domains including the shopping scams from the ML detector.

For the interested reader, Appendix A applies a clustering approach to the content of the 143,227 successfully crawled scam domains, to determine whether unclassified domains (e.g., those without industry tags or with low accuracy tags) are instances of the 7 analyzed scam types or belong to previously unknown types.

V. Scam Observations

This section examines *observations* of the 1.6M scam domains in the desktop and mobile telemetry. An observation is a query in the telemetry to the reputation backend server that contains a domain appearing in the scam domain feeds (i.e., ScamAdviser or shopping scams) or any of its subdomains. For example, if the SLD example.com appears in a scam domain feed, then queries in the mobile telemetry for example.com, bad.example.com, or really.bad.example.com are all considered observations of the scam domain. In contrast, if bad.example.com appears in a scam domain feed, then queries in the mobile telemetry for bad.example.com and really.bad.example.com are considered observations, but queries for the 2LD example.com are not considered observations. This avoids flagging as malicious the parent of a reported subdomain since the subdomain may have been leased to a third party. Since the desktop telemetry contains URLs rather than domains, observations in the desktop telemetry correspond to appearances of the scam domain, or its subdomains, in the queried URLs.

Observations in the desktop and mobile telemetry are not directly comparable due to the different granularity (URLs vs domains) and collection methodology. For example, to limit frequent queries for the same domain, the desktop and mobile security products cache received reputation scores, but the caching policy differs for both clients. In general, the number of observations in the telemetry is a lower bound for the number of visits from users to the scam domains since not every visit may trigger a reputation query. To measure user exposure to scams we will focus on the number of IP addresses in the observations. One challenge is that an IP address may actually correspond to multiple devices if they are

Data	All Desktop	All Mobile	Overlapping Desktop
Start date	2023-01-01	2023-01-01	2023-01-01
End date	2023-11-10	2023-06-16	2023-06-16
Days	314	167	167
IP Hashes	20,396,359	4,680,260	10,132,210
Countries	237	229	234
URLs	34,642,929	N/A	14,699,697
FQDNs	1,040,251	153,998	610,069
SLDs	360,935	97,717	221,576
TLDs	680	461	623

Table IV: Observations of the 607K scam domains in the desktop and mobile telemetry. To facilitate the comparison, the rightmost column has the observations in the desktop telemetry over the same period covered by the mobile telemetry.

behind a NAT gateway. Thus, the number of IP addresses may underestimate the number of devices exposed to scams. On the other hand, the same device may be observed using different IP addresses over time. Due to this effect we will avoid computing user exposure through the whole analysis period. Instead, we will measure user exposure on a daily basis since the shorter the time frame, the less likely a device changes the IP address.

The scam observations are computed regardless of when the scam domains appeared in the feeds. For example, if a scam domain was reported on March 20 by ScamAdviser (the first day the ScamAdviser feed is available), we also include observations of that domain between January 1 and March 19 (if any). In Section V-B we examine whether scam domains are first seen by the telemetry or the scam feeds.

Note that we define scam domain observations and scam domain lifetime in terms of scam SLDs. Thus, throughout Section V when we refer to scam domains, we mean scam domain SLDs.

A. User Exposure to Scams

We measure user exposure in terms of IP addresses observing scam domains (i.e., SLDs) in the desktop and mobile telemetry. Of the 501K scam SLDs, 86,673 (17.3%) have not been observed in the desktop and mobile telemetry; they may be short-lived or fail to attract visitors. There are 99,608 (19.8%) that have been observed by a single IP, 200,501 (39.9%) observed by more than one and less than 10 IPs, and 114,934 (22.9%) observed by at least 10 IPs. Only 27,524 (5.5%) scam SLDs are observed by at least 100 IPs and 4,466 (0.9%) by at least one thousand IPs. Figure 3 in the Appendix shows the ECDF of the number of IPs accessing each scam FQDN and SLD, for domains with at least one observation. These results show that the majority of scam domains fails to attract a significant number of visitors with 77.1% of scam domains being observed by less than 10 IPs. However, some scams achieve significant user exposure with 4,466 scam domains being observed by at least 1K IPs. Furthermore, overall exposure across all 415K observed scam SLDs is high with 25.1M desktop and mobile IPs observing them.

Next, we examine observations in each telemetry separately. Of the 415K scam SLDs observed in the telemetry, 303K (73.2%) are observed only on desktop devices, 12K (3.1%) only on mobile devices, and 98K (23.7%) on both. Table IV summarizes the scam domain observations in each telemetry. The first column captures the observations in the desktop telemetry across the 314 days it is available. The rightmost two columns capture the mobile and desktop observations on the 167 days when we have telemetry data for both. Over the 167 days where we have both desktop and mobile telemetry, the volume of scam observation on desktop devices is larger than on mobile devices across all metrics: 3.9 times larger FQDNs, 1.9 times larger SLDs, and 2.1 times larger IP addresses. One could think this may be due to a larger number of devices in the desktop telemetry, but this is not so. The median number of daily IP addresses in the mobile telemetry is 24% larger than in the desktop telemetry. Thus, despite more potential mobile device targets, scam observations are larger on desktop devices, hinting that users are more exposed to scams on desktop devices compared to mobile devices. We further examine this result next.

	Median	Mean	Stdev	Min	Max
Desktop	101K	101.7K	20K	61.9K	145.5K
Mobile	48K	37.3K	23.8K	9.1K	71.6K

Table V: Daily stats of IP addresses observing scam domains.

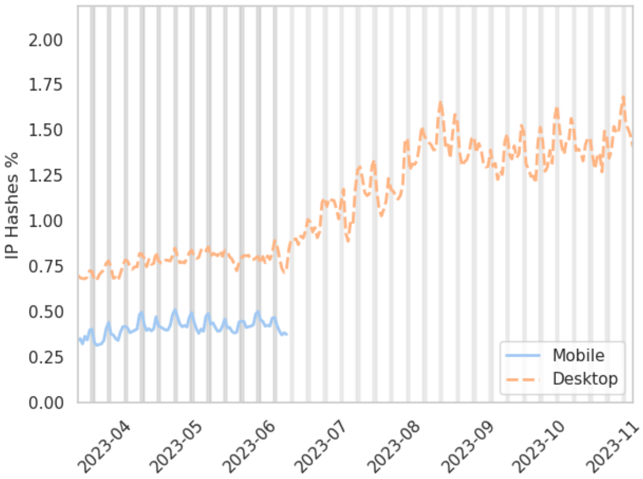

Figure 1: Fraction of active daily IPs observing scam domains. Plot starts on March 20 when we start collecting ScamAdviser feed. Mobile telemetry is only available until June 16, 2023. Weekends are marked with vertical gray lines.

Daily exposure. Table V shows the daily statistics for IP addresses observing scam domains. On a daily basis, a median of 101K desktop and 48K mobile IP addresses observe a scam domain. Thus, over 149K devices are exposed to scam domains each day. To account for the different number of IP addresses in the telemetry each day, we normalize the exposed IP addresses by the total IP addresses in Figure 1. More than twice the fraction of desktop devices (daily median of 0.8%) are exposed to scam domains compared to mobile devices (daily median of 0.3%). Multiple reasons may cause this phenomenon. First, desktop devices may be used more frequently than mobile devices (e.g., during work-time) or may be used more frequently for riskier tasks such as online shopping. For example, reports indicate that users are more intent on purchasing products when using a desktop [78] and that they prefer large devices for important tasks [59]. Second, mobile users could be less exposed to scams because of different interactions on both platforms. For example, mobile users often interact with services through apps. If a user shops on Amazon or Ali Shopping through their apps, he may be less exposed to scams impersonating those sites. Moreover, mobile browsers often have limitations on certain scripts, pop-ups, and other potentially malicious activities, which can contribute to a more secure browsing experience. Because of these factors mobile device activity is more controlled than on desktop, possibly exposing users to less risk. Finally, we could also have a selection bias if the scam domain feeds are biased towards desktop scams making us miss mobile-specific scam domains.

Figure 1 also shows some temporal variations. First, the number of exposed IPs follows a weekly pattern, with a higher fraction of exposed IPs on weekends (gray vertical lines). While the telemetry shows lower overall traffic on weekends, users have more time those days for personal surfing, which may lead to a higher encounter of malicious pages, as mentioned in prior work [18]. The increase of 0.5% in the exposed desktop IPs starting on July 2023 is related to a decrease in the total number of daily desktop IPs due to a client-side update of the vendor's software.

Takeaway 1

Of the 501K SLDs in the scam feeds, 415K (82.7%) are observed in the desktop and mobile telemetry receiving visits by a total of 25.1M IP addresses. Each day, over 149K devices are exposed to online scams. After accounting for population differences, more than twice the fraction of desktop devices (0.8% daily IPs) is exposed compared to mobile devices (0.3%).

Exposure to scam types. Table VI summarizes the scam observations over the whole analysis period, split by scam type. The relative ranking of scam types according to the number of exposed IP addresses is the same for both desktop and mobile. Users are most exposed to shopping scams being observed by 10.1M desktop IPs and 2.8M mobile IPs. Even when removing all shopping scam domains flagged by the ML detector (second data row in Table VI) users are still most exposed to shopping scams in ScamAdviser (2.8M desktop and 1.5M mobile IPs). Cryptocurrency scams rank second being observed by 652K desktop IPs and 50K mobile IPs, with financial scams closely behind with 442K desktop IPs and 20K mobile IPs. If we combined cryptocurrency and financial scams in the same category, the ranking would not change. Users are much less exposed to the other scam types with dating scams having one third the exposed IPs compared to financial scams, and gambling, employment, and funds recovery having an order of magnitude less exposed IPs than the top 3 scam types.

While the user exposure ranking resembles the ranking of scam types by number of scam domains in Table III, some scam types have larger user exposure per domain. For example, the scam domain feeds have 6.5 times more gambling domains than dating domains, but dating scams are visited by three times more IPs compared to gambling scams.

Country exposure. Next, we compare the exposure to scams of users in different countries. While the telemetry has devices geolocated in 239 (desktop) and 230 (mobile) country codes, we restrict this analysis to the 50 countries with a median of at least 10K daily active IPs. The countries with most overall scam observations are those where the vendor has a larger user base, namely the US, Japan, and European countries. However, once we normalize by the per-country user base, the picture changes. Table VII shows the top 10 and bottom 10 countries with the highest and lowest ratio of daily IP addresses (desktop and mobile) observing scams. The highest exposure happens in Philippines with 1.2% of daily devices in that country encountering scams, followed by Turkey (1.2%), Vietnam (1.1%), Norway (1.1%), and Hungary (1.0%). The lowest exposure is for Japan (0.1%), China (0.2%), Russia (0.3%), and South Korea (0.4%) Differences between countries can be significant, with the top exposed country (Philippines) having 10 times higher ratio of exposed IPs than the least exposed country (Japan). Of the top 10 exposed countries, seven are European (Romania, Denmark, Greece, Portugal, Hungary, Norway). In contrast, the least exposed countries are predominantly in Asia (Japan, China, South Korea, Thailand, Indonesia, Russia) and America (Chile, Brazil, Mexico), with only two European countries (Germany, Switzerland). These may point to cultural differences. For example, Asian countries often have higher volumes of mobile traffic, where we have measured a lower exposure to scams.

Takeaway 2

Users are most exposed to shopping scams with 10.2M affected IPs, followed by cryptocurrency scams (653K), and financial scams (443K). Exposure to dating, gambling, employment, and funds recovery scams is significantly smaller with 3x-10x less exposed IPs. User exposure to scams varies significantly among countries. Top exposed countries are mostly from the European Union and have 10 times higher ratio of exposed IPs compared to the least exposed countries, most of them from Asia.

B. Scam Domain Lifetime

Analyzing the scam domain lifetime is important for understanding how well existing defenses are working and for designing new defenses against scams. For this analysis we consider observations on both the desktop and mobile telemetry and focus on SLDs, i.e., observations of subdomains count as observations of the SLD. We analyze three aspects of the lifetime of scam domains, summarized in Table VIII. We first measure the *active time* as the time difference between the first and last observations of the SLD in the telemetry. Then, we measure the *listing delay* as the time difference between the first observation of the SLD in the telemetry and the first appearance of the domain in a scam domain feed. Finally, we measure the *wait time* as the time difference between the WHOIS domain registration date and the first observation of the domain in the telemetry. We measure all times in days.

This analysis includes 242,625 (58%) of the 415K scam SLDs observed in the desktop and mobile telemetry. We

	Desktop					Mobile			
Scam Type	URLs	FQDNs	SLDs	IP Hashes	CC	FQDNs	SLDs	IP Hashes	CC
Shopping (All)	32,072,843	362,942 (34.9%)	288,224 (71.7%)	10,194,912	236	85,923 (55.8%)	77,794 (69.9%)	2,885,616	221
Shopping (SA)	7,338,132	41,488 (4.0%)	30,022 (7.5%)	2,850,994	234	23,104 (15.0%)	20,023 (18.0%)	1,496,865	214
Cryptocurrency	2,691,870	10,222 (1.0%)	7,182 (1.8%)	652,869	220	1,440 (0.9%)	1,389 (1.2%)	50,951	140
Financial	1,742,407	8,174 (0.8%)	5,638 (1.4%)	442,826	219	1,094 (0.7%)	957 (0.9%)	20,457	110
Dating	2,992,402	618 (0.1%)	348 (0.1%)	177,837	153	208 (0.1%)	176 (0.2%)	13,278	67
Gambling	171,985	1,272 (0.1%)	858 (0.2%)	54,875	158	175 (0.1%)	170 (0.2%)	3,361	54
Employment	100,249	801 (0.1%)	651 (0.2%)	49,248	202	139 (0.1%)	134 (0.1%)	2,146	74
Funds recovery	48,484	76 (<0.1%)	51 (<0.1%)	25,663	84	14 (<0.1%)	12 (<0.1%)	438	14
Unclassified	29,304,371	655,721 (63.0%)	98,926 (24.6%)	10,720,908	236	65,005 (42.2%)	32,951 (29.6%)	2,039,492	219
All	69,137,946	1,040,251 (100%)	402,216 (100%)	20,396,359	237	153,998 (100%)	111,353 (100%)	4,680,260	229

Table VI: Scam domain observations per type sorted by exposed IPs. URLs are only available in desktop telemetry. The second data row captures only the ScamAdviser (SA) shopping scams, showing that shopping scams dominate even when excluding the internal detector.

Rank	Country	FQDN	SLD	IP Ratio
1	Philippines	243	209	1.250%
2	Turkey	272	256	1.167%
3	Vietnam	100	97	1.131%
4	Norway	391	372	1.129%
5	Hungary	180	175	1.088%
6	Portugal	436	428	1.083%
7	Denmark	405	390	1.052%
8	Romania	147	143	1.033%
9	Greece	211	207	0.997%
10	Finland	242	233	0.991%
41	Switzerland	435	412	0.497%
42	Brazil	1,155	1,062	0.493%
43	Chile	126	120	0.473%
44	Germany	1,665	1,448	0.423%
45	Thailand	80	79	0.410%
46	Indonesia	47	46	0.390%
47	South Korea	66	64	0.387%
48	Russia	37	35	0.314%
49	China	22	21	0.187%
50	Japan	2,491	1,684	0.120%

Table VII: Daily desktop and mobile scam observations for the top 10 and bottom 10 countries by ratio of daily IPs observing a scam domain. Only the 50 countries with at least a median number of 10K daily IPs are considered.

	Active Time		Listing Delay		Wait Time	
Scam Type	Med	Mean	Med	Mean	Med	Mean
Dating	59	75.2	8	25.3	19	20.9
Shopping	15	42.6	1	11.3	24	43.2
Employment	4	31.7	14	34	9	21.7
Gambling	3	37.9	4	24.7	10	11.6
Funds recovery	1	42.6	2	29.5	29	29.6
Financial	1	32.0	1	-0.6	22	41.2
Cryptocurrency	1	25.9	1	-0.3	11	30.8
Unclassified	3	27.2	5	22.5	5	16.4
All	11	38.7	1	13.8	18	36.5

Table VIII: Median and mean (1) active time between first and last observations, (2) listing delay from first observation to first appearance in scam feeds, and (3) wait time from domain registration to first observation. All measured in days.

exclude 48K (11.6%) scam SLDs for which no WHOIS registration date is available. To ensure the telemetry covers the birth of a scam domain, we also exclude 124,163 (29.9%) scam SLDs registered prior to January 1, 2023. By doing that, we may remove long-lived scam domains, thus our active time estimation becomes a lower bound. However, we will additionally estimate the active time when including these 124K scam SLDs.

Active time. Overall, the median active time for scams domains is 11 days, although the mean is 38.7 days as some scam domains are active for much longer. This a lower bound because we removed domains registered prior to January 1, 2023. If we include those domains, the median active time doubles reaching 21 days (mean 59.4 days). We observe notable differences among scam types with *dating*, *shopping*, and *employment* scams being active the longest, with a median of 59, 15, and 4 days, respectively. On the other hand, *funds recovery*, *financial*, and *cryptorcurrency* scams last 1 day. We compare our estimates with those of prior work. Li et al. recently measured the median lifetime of cryptocurrency giveaway scams to be 26.18 hours [47]. Although, the cryptocurrency scams we measure offer general investment opportunities in cryptocurrencies (i.e., not only giveaways), they have very similar active times. Prior works have quantified the lifetime of phishing websites to range from 17 hours [61] up to 21 hours [62]. Based on those estimations, scam domains are active 12–15 times longer than phishing domains, possibly indicating scams are harder to identify and take down for current defenses such as blocklists and interventions by domain registrars and hosting providers.

Listing delay. We examine whether scam domains are first observed by the telemetry or the scam feeds. We measure the delay between the first telemetry observation and the earliest appearance of the SLD in the scam domain feeds. A positive listing delay means that the telemetry observed the scam domain first, while a negative delay means that the feeds listed the scam domain before any user in the telemetry observed it. Overall, the median listing delay is 1 day. During this 1 day the scam domains attract the majority of their total traffic, i.e., on average a scam domain receives 58.6%

of its visits within the first active day. On the positive side, 19.1K (7.9%) scam domains are listed in the feeds before any user observes them. These scam domains may have been identified by crawling recently registered domains or domains recently listed in Certificate Transparency logs [47]. By scam type, employment scams are harder to identify with a median of 14 days of activity before they are listed, followed by dating scams (8 days). Other scam types are listed much faster within 1–3 days. Identifying employment scams and fake dating platforms often requires registration which may not always be free. This complicates automated detection efforts and increases their costs.

Since the median active time is 11 days and it takes 1 day for the feeds to list scam domains, the median time from listing to removal is 10 days. The removal could be due to blocking by the domain registrar, intervention by the hosting provider, or the scammers moving to another domain.

Wait time. We also measure the delta between the domain registration and the first observation, which captures how fast the domains are utilized by scammers for hosting scams. Overall, scammers wait a median of 18 days before utilizing their domains, with the largest wait time being 29 days for funds recovery scams and the shortest 9 days for employment scams. It is worth noting that we removed domains registered prior to January 1, 2023 for the lifetime analysis. That filtering should have removed most compromised domains that are not registered by the attackers. Li et al. measured a median wait time of 14.14 hours for giveaway scams [47] and we measure a wait time of 11 days for cryptocurrency scams. However, they count from domain registration until the website is up and we count until the first victim arrives. The difference may indicate that scam sites stay up for days before they start being advertised.

We find 2.8K (0.7%) scam domains observed in the telemetry prior to registration. These domains were previously registered by other entities and have been re-registered by the scammers to benefit from their residual trust [83]. Re-registrations happen for 12.7% of gambling domains and 2% of dating domains, while the percentage is up to 5% for other types.

Takeaway 3

Scam domains are observed in the telemetry a median of 1 day before they are listed in a scam feed. On the positive side, 19.1K (7.9%) scam domains are listed on the feeds before any device observes them. After first observation, the scam domains are active for a median of 10 days, with dating and shopping scams having the longest activity of 59 and 15 days, respectively, while financial,cryptocurrency, and funds recovery scams are only active for 1 day. Compared to phishing domains, scam domains remain active 12–15 times longer.

VI. Scam Advertising

This section examines what fraction of victims are following advertisements to arrive at scam pages, and from which sources the advertisement traffic comes from, e.g., social networks. To identify advertised scam pages, we examine Urchin Tracking Module (UTM) parameters in the scam URLs. Advertisers can define UTM parameter values for each source they use to promote a website. The UTM parameters are then added to the advertised URLs before posting them in the advertising sources (e.g., social networks). When a user clicks on a promoted link with UTM parameters, those become available to tracking and analytics platforms (e.g., Google Analytics), which use them to segment the received traffic. The segmentation is used by the libraries to produce reports on how well advertisement sources and campaigns work.

Scam Type	Observations	SLD	IP
Shopping	6,615,332 (20.6%)	32,901 (11.4%)	2,337,930 (22.9%)
Cryptocurrency	349,168 (13.0%)	340 (4.7%)	131,611 (20.1%)
Financial	46,367 (2.6%)	302 (5.3%)	22,498 (5.1%)
Funds recovery	250 (0.5%)	5 (9.8%)	184 (0.7%)
Gambling	659 (0.4%)	32 (3.7%)	276 (0.5%)
Employment	310 (0.3%)	32 (4.9%)	208 (0.4%)
Dating	7,335 (0.2%)	108 (31.0%)	138 (<0.1%)
Unclassified	2,086,766 (7.1%)	5,179 (5.2%)	826,973 (7.7%)
All	9,231,334 (13.3%)	38,982 (9.7%)	3,127,873 (15.3%)

Table IX: Desktop observations of advertised scam URLs, SLDs in those URLs, and IPs observing the advertised URLs. Percentages are computed over all desktop scam observations in Table VI.

There are 5 UTM parameters that can appear in any order. It is highly recommended to include the *utm_source* parameter, which identifies the origin of the traffic (e.g., facebook, email, google). Two other parameters are recommended: *utm_medium* identifies whether the traffic is organic or paid for (e.g., cpc for cost-per-click, cpm for cost-per-mile) and *utm_campaign* is the advertiser-selected name of the campaign. The other two UTM parameters are infrequent: *utm_term* identifies search terms the user typed when clicking the ad (e.g., travel+island) and *utm_content* identifies a specific ad when a campaign uses multiple ones (e.g., logolink, textlink). We call a scam URL with UTM parameters an *advertised scam URL.*

Since only the desktop telemetry has URLs, we focus on desktop devices. Table IX captures the number of advertised scam URLs in the desktop telemetry, the number of SLDs in those URLs, and the IPs observing those URLs. The percentages are computed over all desktop observations in Table VI. Overall, in 9.2M (13.3%) of all scam observations users followed an advertisement to reach the scam domain, 38.9K (9.7%) of scam SLDs are being advertised, and 3.1M (15.3%) IPs are exposed to the advertised scam URLs. The numbers are a lower bound since it is possible to advertise sites without using UTM parameters.

Promotion via advertisements is more prevalent for shopping scams with 2.3M (22.9%) of users having visited shopping scams by clicking an ad. Advertisements are also prevalent with cryptocurrency scams having attracted 131K (20.1%) of their victims via ads. For all scam types except dating, the number of IPs exposed to the advertised URLs is one to two orders of magnitude larger than the number of advertised SLDs. This is an indication of the effectiveness of the promotion of scams via online advertisement. Ads for some scam types seem to be more successful. For example, while the number of the advertised cryptocurrency SLDs (340) is very similar to that of financial scams (302), advertised cryptocurrency scams affect 5.8 times more users. We also observe that despite the

Source	Values	Observations
Facebook	780	4,948,800 (75.1%)
X (formerly Twitter)	5	503,259 (7.6%)
Newsletter	34	66,710 (1.0%)
Taboola	1	65,830 (1.0%)
Copernica	1	24,649 (0.4%)
Shopify	2	9,371 (0.1%)
All	28,605	6,592,206 (100%)

Table X: Top advertising sources for scam URLs and number of utm_source parameter values considered.

stricter policies of major social networks on cryptocurrency advertisements [56] over 131K users have visited such scam pages via online ads. Next, we analyze the role of social media platforms on scam advertisements.

Ad sources. We use the *utm_source* parameter values to analyze which advertisement platforms are preferred by the scammers. Since UTM parameter values are defined by the advertisers, the same advertisement platform (e.g., Facebook) may be referred under different names (e.g., fb, facebook, facebook_ads). We focus on the 6.5M advertised scam URLs that include a UTM source parameter. We also group tokens that identify the same platform: 1.1K tokens for Facebook, 8 for Twitter (now called X), and 42 for email newsletters. Table X presents the top sources (with a ratio higher than 0.1%) after aggregation. Facebook is the advertisement platform most used to promote scam websites, with 4.9M observations (75% of all advertised URL observations). In comparison, Twitter has 503K (7.6%) observations. The userbase of Facebook is 3 billion users [60], about 6 times bigger than Twitter [85], which explains why Facebook is the preferred advertisement platform also for scammers. Newsletters are another common advertisement platform with 66.7K (1.0%) observations. We also observe advertisement platforms like Taboola [90] (65K), the email-based Copernica [24] (24.6K), and Shopify (9.3K). Google also appears among the advertisement platforms, but with a low volume (2.7K) of observations. Our analysis shows that 13.3% of all scam domain observations on desktop devices are produced by following online advertisements, with 5.4M (59%) of those ads shown in social media, predominantly in Facebook. A similar estimation by FTC reported that 12% of scam victims in 2023 reached scams via social media. [22].

Takeaway 4

In 9.2M (13.3%) of all scam observations users followed an advertisement to reach the scam domain, 38.9K (9.7%) of scam SLDs are promoted through advertisements, and 3.1M (15.3%) of IPs are exposed to the advertised scam URLs. Scam advertisements are placed largely (59%) on social media, most often on Facebook (75%). Shopping and cryptocurrency scams are the most advertised and attract the highest number of users through advertisements. This may explain the widespread popularity of these two scam types in our dataset.

VII. User Impact

So far, we have analyzed the exposure of users to scams, i.e., the number of potential victims that visit the scam websites. However, not all visitors to the scam websites will end up being scammed, e.g., users may realize the scam and navigate away. In this section, we estimate the potential impact of scams on users by examining what fraction of users visit pages that may indicate they were scammed, e.g., checkout or payment pages. We largely focus on shopping scams because they are the most prevalent and tend to have a fully online experience where the user selects the products, adds them to a cart, and proceeds to checkout and payment. For other scams, user interactions and the payment process may not be handled online, e.g., the user may be provided a bank account number or cryptocurrency wallet and pays through his bank or a crypto exchange.

For this analysis, we examine the path and parameters of the scam URLs using keywords capturing different types of webpages such as the main page, product, and checkout pages. For example, a main page contains no path or a filename containing the *main* keyword (e.g., main.php), and a checkout page contains at least one of 18 checkout-related keywords including *order*, *checkout*, *trackingorder*, and *track-your-order*.

For each page type, Table XI summarizes the number of keywords used and the number and fraction of scam URLs, advertised scam URLs, shopping scam URLs, and advertised shopping scam URLs. We are able to categorize 37.0M (53.6%) of all scam URLs, with the largest categories being, 21.6M (31.3%) product pages, 14.3M (20.7%) main pages, and 891K (1.2%) checkout pages. The 32.1M (46.4%) uncategorized URLs do not match any keywords (e.g. the paths contained just numbers). A small fraction of scam domain URLs point to contact us, about us, policy (e.g., privacy policy, terms of use), and careers pages. These may be due to users trying to determine whether the site is legit. For shopping scams, we categorize a significantly higher 25.3M (78.8%) pages. Product pages (18.8M, 58%) and checkout pages (761K, 2.4%) are almost twice as prevalent in shopping scams compared to their prevalence among all scam URLs (31.3% and 1.2%, respectively).

Most advertised scam URLs correspond to product pages (79.3%) with the ratio being even higher for shopping scams (90.7%). Thus, scammers directly advertise specific products, rather than advertising the main page of the scam website.

The visit of a checkout page indicates the user is in the final stages of completing a purchase. While we only observe a handful of payment pages, this is likely due to payments requiring the user to first log in to the site. We assume that users visiting a checkout page are likely to complete a purchase, although a fraction of them may still navigate away. Fortunately, only 891K (1.2%) scam URLs are checkout pages although the ratio is higher (761K, 2.4%) in shopping scams. Thus, even when users are convinced to visit a scam page, they often identify it as a scam or are not interested in the offered products, thus avoiding a purchase. However, despite the small fraction of scam checkout pages, over 411K desktop IPs (4% of all IPs visiting a shopping scam) reach a checkout page. On a daily basis, a median of 1.8K IPs (mean: 1.8K, min: 486, max: 2.7K, std: 339) reach a checkout page.

		All Scams		Shopping Scams	
Page Type	**Keyw.**	**URLs**	**Advertised URLs**	**URLs**	**Advertised URLs**
Product	15	21,613,681 (31.3%)	6,889,232 (79.3%)	18,881,170 (58.0%)	5,862,484 (90.7%)
Main page	1	14,296,197 (20.7%)	41,024 (0.5%)	5,464,326 (17.0%)	27,268 (0.4%)
Checkout	18	891,132 (1.2%)	10,971 (<0.1%)	761,352 (2.4%)	8,089 (0.1%)
Contact us	5	105,005 (0.1%)	323 (<0.1%)	85,708 (0.3%)	221 (<0.1%)
About us	5	63,303 (<0.1%)	335 (<0.1%)	43,079 (0.1%)	317 (<0.1%)
Policy	9	25,753 (<0.1%)	2,545 (<0.1%)	11,122 (<0.1%)	2,156 (<0.1%)
Payment	1	17,558 (<0.1%)	5 (<0.1%)	10,437 (<0.1%)	0 (0.0%)
Careers	1	21,309 (<0.1%)	463 (<0.1%)	18,700 (<0.1%)	0 (0.0%)
Uncategorised	-	32,104,008 (46.4%)	1,741,010 (20.0%)	6,796,949 (21.2%)	562,788 (8.7%)
All	55	69,137,946 (100%)	8,685,908 (100%)	32,072,843 (100%)	6,463,323 (100%)

Table XI: For each page type, number of keywords used to identify the category, number of all scam URLs, number of advertised URLs (i.e., with UTM parameters), number of shopping scam URLS, and number of advertised shopping scam URLs.

Takeaway 5

We observe that 4% of all IPs visiting a shopping scam reach a checkout page. Thus, a small, but not negligible fraction of users exposed to scams may end up completing a purchase and becoming a victim. We also observe that most advertised scam URLs correspond to product pages (79.3% across all scams, 90.7% for shopping scams) indicating scammers prefer to advertise specific products to users.

VIII. Discussion

This section discusses our results, proposes recommendations, presents limitations, and details ethical considerations.

A. Results & Recommendations

Scam type differences. Our study investigates 7 popular scam types and identifies differences among them. Users are most exposed to shopping scams with 10.1M affected IPs, followed by cryptocurrency scams (652K), and financial scams (442K). Exposure to dating, gambling, employment, and funds recovery scams are significantly smaller with 3x-10x less exposed IPs. On the other hand, some smaller scam types attract a disproportionately large number of visitors per domain. For example, funds recovery services and dating attract 337.7 and 287,7 IPs per domain, respectively, much higher than shopping (28.1), cryptocurrency (63.9), and financial (54.2).

The observed differences highlight the need for specialized scam detection mechanisms that focus on specific scam types in order to complement general defenses against malicious websites. The development of tailored defenses can be costly given the large number of scam types and the frequent emergence of new ones. Our findings indicates that shopping, cryptocurrency, and financial scams affect the largest number of users. Thus, these scam types should be prioritized, as specialized detection systems may offer a larger return for the investment. Several specialized approaches have already been proposed for detecting shopping scams [15], [45], [101], [19], [100], [43], [84], [10] and for some cryptocurrency scams such as giveaway scams [47] and ponzi scams [12]. However, specialized detection approaches for financial scams are sorely needed. Other scam types affecting less users may also deserve specialized defenses. For example, dating domains are the second hardest to detect with a median listing time of 8 days, compared to one day across all scam types.

Defenses. Our analysis allows evaluating how well existing defenses work against scams. First, we identify that at least 149K customers of the vendor are being exposed to scams on a daily basis. We also measure a median of 11 active days for scam domains, much longer than the 17–21 hours reported for phishing domains [62], [61]. These results indicate that improvements are needed to scam defenses. We evaluate the potential of scam domain feeds for proactive blocking. We observe that 92% scam domains are first seen in the telemetry. Thus, users are exposed to those scams before the feeds allow blocking them. On the other hand, 7.9% scam domains are reported by ScamAdviser prior to devices being exposed to them, so there is some value in using the feeds for proactive blocking. It takes a median of one day for scam domains to be listed in the feeds We also find a modest 4% of unresolved domains that were already blocked by their domain registrars when they appeared on the feed. Scam domains can also be taken down by hosting providers but those interventions rarely leave a trace (i.e., oftentimes only a default error page is returned). Finally, we measure a median wait time from domain registration to first visit of 18 days indicating that defenses monitoring new domain registrations and HTTPS certificates have great potential to identify scam domains before users are exposed to them.

We also observe a lack of open scam feeds, beyond phishing domain feeds (e.g., PhishTank [66]), that may hamper research in the area. While scam domains likely appear in open threat intelligence platform such as AlienVault OTX [7], they are not marked as such so they have to be distinguished from other malicious domains. We leveraged the commercial ScamAdviser feed, but commercial feeds may not be available to many researchers and their accuracy cannot be easily evaluated, e.g., we had to apply conservative filtering to remove potential false positives. Developing crowd-sourced scam feeds would be important to foster scam research. Finally, our analysis indicates that user interactions with scam websites may be long (i.e., advertisement, account creation, product selection, checkout, payment), which may provide defenses with multiple opportunities for intervention, e.g., for extracting classification features from various vantage points.

Take-down mechanisms. Our scam domain lifetime analysis highlights a significant gap between the scam detection time (1 day) and the scam active time (10 days). This disparity suggests that either scam sites are not being systematically reported or that current take-down procedures are not sufficiently responsive to the rapidly evolving tactics employed by scammers. One potential reason for this gap is the fact that scammers often abuse hosting platforms (e.g., Shopify), which may have unique take-down procedures that differ from those used by domain registrar or registries. We recommend that further research is done on how scam take-down procedures can be improved to reduce the active time of scams and thus their impact.

Advertising. We measure that in at least 13.3% of scam observations users followed an advertisement leading to a scam domain and at least 9.7% scam SLDs are being promoted through advertisement platforms. Most of those advertisements come from social networks, predominantly Facebook (75%) and Twitter/X (7.6%). But, we also observe email-based scam advertisement campaigns. Given the relatively short lifetime of scam domains, scammers abuse advertisement platforms to quickly drive potential victims to their sites.

Advertising platforms (including social media) should prioritize user safety over ad revenue. They should rigorously vet the advertised URLs using domain reputation websites (e.g., TrustPilot) and manually review those with low ratings. Additionally, platforms should offer clear reporting mechanisms for users to flag ads leading to scams and proactively investigate other URLs within the same advertising campaign as those flagged by reporters.

Scam type classification. Our evaluation of the ScamAdviser industry tags shows that they are often inaccurate with only 7 out of 26 categories having high precision. Thus, 98.9K (25%) scam SLDs remain unclassified (24.6% desktop, 29.6% mobile) as they lack trustable industry tags. We believe this low accuracy is not specific to ScamAdviser, but plagues most commercial website classification services. For example, we tried the services used by VirusTotal, but the accuracy did not seem better since most scam domains were detected either as malicious or as phishing. An alternative classification approach is leveraging machine learning. Previous work has designed one-class classifiers to identify a specific scam type, e.g., shopping scams [101], [19], [100], [45]. An interesting research question is whether we should build many highly specific one-class classifiers (i.e., one for each scam type) or to build an n-class classifier to classify a domain into n scam types at once. One-class classifiers are typically more accurate, but when combining them a domain may end up being assigned multiple scam types. Furthermore, supervised classifiers are limited by the number of scam types in the training dataset. To address this issue we leveraged unsupervised clustering, but 38.2% of the websites ended up as singletons. Novel scam domain and website classification approaches remain an important area for future work.

Scam taxonomy. Our work has analyzed 7 popular scam types. However, the accuracy evaluation of ScamAdviser tags and the clustering validation reveal multiple other scam types among the unclassified domains such as SEO services, charity scams, tech support scams, package delivery scams, membership cancellation services, and dream interpretation. Currently, our community lacks a unified taxonomy of scams, with each scam reviewing site and consumer protection organization having its own types, sometimes with conflicting definitions. Furthermore, current website classifiers focus on the industry a website belongs to, but scam types may need to be more fine-grained. For example, cryptocurrency scams can be exchange impersonation scams [104], giveaway scams [98], [47], mining investment scams [82], ponzi scams [12], and token scams [103]. Defining a comprehensive taxonomy of scams is an extremely challenging proposition given the many scam types and variations, combinations of scam types (e.g., romance and financial as used in pig butchering scams [25]), and the creation of new scam types over time. Similar to what has been proposed for malware [79], we believe future work should propose an open scam taxonomy that while not complete, covers the most popular scam types and can be easily extended.

Identifying scam websites. While our work does not focus on detection, we have examined a large number of scam websites and can offer some recommendations to users on how to detect them. As expected, if an offer (e.g., investment advice) sounds too good to be true, it probably is a scam. Users that are suspicious about a website should leverage reputation services (e.g., TrustPilot [96], SiteJabber [81], Google Business [38]) and avoid sites with low reputation. Furthermore, if the site has no reviews that should raise suspicions as many scam sites are short-lived. Many scams request up front fees so users should be especially suspicious if those are requested. Another property of many scam websites is that they avoid identifying the entity (e.g., the company) behind the website. If the privacy policy, terms of service, and contact us webpages do not explicitly list the entity that owns the site, that is typically a good sign that the website may be a scam.

B. Limitations

Selection bias. Our study is constrained by the used dataset, which introduces some selection bias. We only analyze the exposure to scams of users that have purchased a security product and have opted in to the telemetry collection. Other users that do not invest in security or decline the collection due to privacy concerns may have different scam exposure. We examine the telemetry from a single cybersecurity vendor, which introduces a geographical bias towards the regions where the vendor's customers are located. The telemetry covers millions of desktop and mobile devices, with devices in nearly all countries (239 country codes in desktop telemetry, 230 in mobile telemetry). However, the devices are not equally spread with higher income regions having more devices. In particular, 80% of the devices are located in the United States, the European Union, Japan, and the United Kingdom. User exposure in other regions with lower income (e.g., Africa) could differ. Despite the geographic imbalance, the telemetry still contains 50 countries with at least 10K daily active devices.

The scam domain feeds we use come mostly from desktop devices. Thus, we may have a negative selection bias towards mobile-specific scam sites. This could affect the fact that desktop devices are affected by scams twice more than mobile devices. However, the total number of detections the vendor observes (scams or other) is indeed significantly larger on desktop devices, despite the mobile user base being larger. Given the prevalence of mobile devices, it is unlikely the difference is only due to scammers targeting desktop devices more. We believe differences in user behavior on both platforms play an important role in the different exposure.

Shopping scams skew. Shopping scams are over-represented since they come from two datasets: ScamAdviser and the internal ML detector. To address this issue, Table VI includes a row with only observations of the ScamAdviser shopping scam domains, i.e., excluding the shopping scam domains only identified by the internal ML detector. The results show that, even if considering only ScamAdviser scam domains, users are still most exposed to shopping scams.

IPs vs users. The same device may appear under different IP addresses over time. To address this issue, we measure user exposure to scams daily, as shorter time frames reduce the chance of IP changes. Additionally, multiple devices may share a single IP, e.g., due to network address traversal (NAT) in home networks, potentially underestimating the number of users exposed to online scams.

Scam domain filtering. We apply two conservative filtering steps to remove benign domains in the ScamAdviser feed. These filtering steps may remove domains that are indeed scams if those domains made it into the Tranco Top 1M list, were not submitted to VirusTotal, or had less than two detections on June 2024 when we queried VirusTotal. Still, we prefer to err on the side of caution by removing any potentially benign domains in ScamAdviser to avoid polluting our measurements.

User impact. Measuring the impact of scams on users is challenging. We measured the user exposure to scams and that 4% of all IPs visiting a shopping scam reach a checkout page. However, not all users who visit a checkout page will complete a purchase. Thus, that percentage may overestimate scam success. Furthermore, our approach does not allow us to quantify the financial and emotional impact on victims [58], [102]. One approach for measuring the financial impact is to leverage victim reports. But, each abuse reporting service (e.g., BBB [13], FTC [23], ChainAbuse [20]) has its own scam report format and categories, making aggregation difficult. It is also hard to map the reports to specific scam domains, as these are hardly reported. For cryptocurrency scams, an alternative is collecting blockchain addresses used for scam payments and analyzing the deposits to those addresses in the public ledger (e.g., [12], [42], [47], [36]).

C. Ethical Considerations

During the installation of the vendor's products, customers may choose to opt in for sharing telemetry data. When users who agree to the data collection visit a URL, a query is made to a backend to obtain the URL's reputation. Queries are anonymized by removing unique device identifiers so they cannot be mapped to specific users. Device IP addresses are first geolocated at a country level, then hashed, and only the hash is stored. The telemetry data is stored in the vendor's data lake. Aggregate statistics are directly computed on the data lake, so no local copies of the telemetry data are kept. Only employees of the vendor have access to the data lake. The academic authors did not need IRB approval as they do not access user data, only aggregate statistics.

IX. Conclusions

Online scams have emerged as a significant threat to internet users worldwide, with substantial financial and emotional impacts. In this work, we perform what we believe is the first study that measures the user exposure to different types of online scams, geographical variations, scam domain lifetime, and the promotion of scam websites through online advertisements. We discover that hundreds of thousands of devices are exposed to scams every day, with shopping and cryptocurrency scams affecting the most devices. Most scam domains are observed in the telemetry 1 day before they appear in a scam feed. After first observation, the scam domains are active for 11 days. The longer activity period of scam domains compared to phishing domains highlights the need for quicker detection mechanisms. In addition, our study shows that an important portion of scams is propagated through online advertisements hosted largely on social media, especially Facebook. This shows that advertising platforms need to step up their defenses against scams. At last, we observe a small, but not negligible fraction of users exposed to scams may complete a purchase and become victims of shopping scams.

Acknowledgments

We thank the reviewers for their time and valuable comments. This work was partially funded by the Spanish Government MCIN/AEI/10.13039/501100011033/ through grants TED2021-132464B-I00 (PRODIGY) and PID2022-142290OB-I00 (ESPADA). The above grants are co-funded by European Union ESF, EIE, FEDER, and NextGeneration funds.

References

[1] Iso 3166-1 alpha-2. https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2.

[2] Bhupendra Acharya, Muhammad Saad, Antonio Emanuele Cinà, Lea Schönherr, Hoang Dai Nguyen, Adam Oest, Phani Vadrevu, and Thorsten Holz. Conning the Crypto Conman: End-to-End Analysis of Cryptocurrency-based Technical Support Scams. In *IEEE Symposium on Security and Privacy*, 2024.

[3] AdScams. Ricelazily.com review - are ricelazily reviews real or fake?, 2022. https://ad-scams.com/ricelazily-com.

[4] Citizens Advice. How to communicate about scams in an effective and engaging way. https://www.cas.org.uk/system/files/citizens_advice_scams_awareness_toolkit2018b.pdf.

[5] Federal Trade Commission (FTC) Consumer Advice. Job scams. https://consumer.ftc.gov/articles/job-scams.

[6] Suhaib Al-Rousan, Abdullah Abuhussein, Faisal Alsubaei, Ozkan Kahveci, Hazem Farra, and Sajjan Shiva. Social-guard: Detecting scammers in online dating. In *2020 IEEE International Conference on Electro Information Technology (EIT)*, pages 416–422. IEEE, 2020.

[7] AlienVault OTX.
https://otx.alienvault.com/.

[8] Dimo Angelov. Top2vec: Distributed representations of topics. 2020.

[9] Emad Badawi, Guy-Vincent Jourdan, Gregor Bochmann, and Iosif-Viorel Onut. An automatic detection and analysis of the bitcoin generator scam. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 407–416. IEEE, 2020.

[10] Emad Badawi, Guy-Vincent Jourdan, Gregor Bochmann, Iosif-Viorel Onut, and Jason Flood. The “game hack” scam. In *International Conference on Web Engineering*, pages 280–295. Springer, 2019.

[11] American Riviera Bank. The ultimate guide to recognizing and avoiding recovery scams. https://americanriviera.bank/blog/the-ultimate-guide-to-recognizing-and-avoiding-recovery-scams.

[12] Massimo Bartoletti, Barbara Pes, and Sergio Serusi. Data Mining for Detecting Bitcoin Ponzi Schemes. In *Crypto Valley Conference on Blockchain Technology*, June 2018.

[13] Scam Tracker, 2024. https://www.bbb.org/scamtracker/reportscam.

[14] Morvareed Bidgoli and Jens Grossklags. “hello. this is the irs calling.”: A case study on scams, extortion, impersonation, and phone spoofing. In *2017 APWG Symposium on Electronic Crime Research (eCrime)*, pages 57–69. IEEE, 2017.

[15] Marzieh Bitaab, Haehyun Cho, Adam Oest, Zhuoer Lyu, Wei Wang, Jorij Abraham, Ruoyu Wang, Tiffany Bao, Yan Shoshitaishvili, and Adam Doupé. Beyond phish: Toward detecting fraudulent e-commerce websites at scale. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2566–2583. IEEE Computer Society, 2023.

[16] Elijah Bouma-Sims and Brad Reaves. A first look at scams on youtube. *arXiv preprint arXiv:2104.06515*, 2021.

[17] Xander Bouwman, Victor Le Pochat, Pawel Foremski, Tom Van Goethem, Carlos H. Ganan, Giovane C. M. Moura, Samaneh Tajalizadehkhoob, Wouter Joosen, and Michel van Eeten. Helping hands: Measuring the impact of a large threat intelligence sharing community. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1149–1165, Boston, MA, August 2022. USENIX Association.

[18] Davide Canali, Leyla Bilge, and Davide Balzarotti. On the effectiveness of risk prediction based on users browsing behavior. In *ACM Symposium on Information, Computer and Communications Security*, 2014.

[19] Claudio Carpineto and Giovanni Romano. Learning to detect and measure fake ecommerce websites in search-engine results. In *Proceedings of the international conference on web intelligence*, pages 403–410, 2017.

[20] ChainAbuse, 2024. https://www.chainabuse.com/.

[21] BC Securities Commission. Types of investment scams. https://www.investright.org/avoid-fraud/types-of-investment-scams/.

[22] Federal Trade Commission. Consumer sentinel network - data book 2023. February 2024. https://www.ftc.gov/system/files/ftc_gov/pdf/CSN-Annual-Data-Book-2023.pdf.

[23] Federal Trade Commission. ReportFraud.ftc.gov, 2024. https://reportfraud.ftc.gov/.

[24] Copernica.com: Powerful Email Marketing Solutions, 2024. https://www.copernica.com/.

[25] Cassandra Cross. Romance baiting, cryptorom and ‘pig butchering’: an evolutionary step in romance fraud. *Current Issues in Criminal Justice*, pages 1–13, 2023.

[26] Cassandra Cross, Kelly Richards, and Russell G Smith. The reporting experiences and support needs of victims of online fraud. *Trends and issues in crime and criminal justice*, (518):1–14, 2016.

[27] Tobias Dam, Lukas Daniel Klausner, Damjan Buhov, and Sebastian Schrittwieser. Large-scale analysis of pop-up scam on typosquatting urls. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pages 1–9, 2019.

[28] Dipanjan Das, Priyanka Bose, Nicola Ruaro, Christopher Kruegel, and Giovanni Vigna. Understanding security issues in the nft ecosystem. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022.

[29] Marguerite DeLiema and Paul Witt. Profiling consumers who reported mass marketing scams: demographic characteristics and emotional sentiments associated with victimization. *Security Journal*, pages 1–44, 2023.

[30] Jillian D’Onfro. Google broke up a vietnamese con scheme after an employee was scammed buying a bluetooth headset, May 2018. https://www.cnbc.com/2018/05/03/how-google-weeds-out-fraud-in-its-shopping-tool-.html.

[31] JL Fleiss. Measuring nominal scale agreement among many raters, psychological bulletin. 1971.

[32] Better Business Bureau Institute for Marketplace Trust. Bbb scam tracker risk report - employment scams make a resurgence. February 2023. https://bbbfoundation.images.worldnow.com/library/8923baa8-e503-45c9-9f2c-c57995ed4a2e.pdf.

[33] Better Business Bureau Institute for Marketplace Trust. Bbb scam tracker risk report. April 2024. https://bbbmarketplacetrust.org/wp-content/uploads/2024/04/2023-BBBScamTracker-RiskReport-US-040224.pdf.

[34] Yuhao Gao, Haoyu Wang, Li Li, Xiapu Luo, Guoai Xu, and Xuanzhe Liu. Demystifying illegal mobile gambling apps. In *Proceedings of the Web Conference 2021*, pages 1447–1458, 2021.

[35] Gibran Gomez, Pedro Moreno-Sanchez, and Juan Caballero. Watch your back: Identifying cybercrime financial relationships in bitcoin through back-and-forth exploration. In *ACM SIGSAC Conference on Computer and Communications Security*, 2022.

[36] Gibran Gomez, Kevin van Liebergen, and Juan Caballero. Cybercrime Bitcoin Revenue Estimations: Quantifying the Impact of Methodology and Coverage. In *ACM Conference on Computer and Communication Security*, 2023.

[37] Google Analytics, 2024. https://developers.google.com/analytics.

[38] Google Business.
https://www.google.com/business/.

[39] Payas Gupta, Roberto Perdisci, and Mustaque Ahamad. Towards measuring the role of phone numbers in twitter-advertised spam. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pages 285–296, 2018.

[40] Srishti Gupta, Gurpreet Singh Bhatia, Saksham Suri, Dhruv Kuchhal, Payas Gupta, Mustaque Ahamad, Manish Gupta, and Ponnurangam Kumaraguru. Angel or demon? characterizing variations across twitter timeline of technical support campaigners. *The Journal of Web Science*, 6, 2019.

[41] Shuang Hao, Kevin Borgolte, Nick Nikiforakis, Gianluca Stringhini, Manuel Egele, Michael Eubanks, Brian Krebs, and Giovanni Vigna. Drops for stuff: An analysis of reshipping mule scams. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1081–1092, 2015.

[42] Danny Yuxing Huang, Maxwell Matthaios Aliapoulios, Vector Guo Li, Luca Invernizzi, Kylie McRoberts, Elie Bursztein, Jonathan Levin, Kirill Levchenko, Alex C. Snoeren, and Damon McCoy. Tracking Ransomware End-to-end. In *IEEE Symposium on Security and Privacy*, May 2018.

[43] Amin Kharraz, William Robertson, and Engin Kirda. Surveylance: automatically detecting online survey scams. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 70–86. IEEE, 2018.

[44] Takashi Koide, Daiki Chiba, and Mitsuaki Akiyama. To get lost is to learn the way: Automatically collecting multi-step social engineering attacks on the web. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, pages 394–408, 2020.

[45] Platon Kotzias, Kevin Roundy, Michalis Pachilakis, Iskander Sanchez-Rola, and Leyla Bilge. Scamdog millionaire: Detecting e-commerce scams in the wild. In *In Proceedings of the 39th Annual Computer Security Applications Conference*, 2023.

[46] Jonathan Larson, Bryan Tower, Duane Hadfield, Darren Edge, and Christopher White. Using web-scale graph analytics to counter technical support scams. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3968–3971. IEEE, 2018.

[47] Xigao Li, Anurag Yepuri, and Nick Nikiforakis. Double and nothing: Understanding and detecting cryptocurrency giveaway scams. In *Network and Distributed Systems Security (NDSS) Symposium*, 2023.

[48] Kevin Liao, Ziming Zhao, Adam Doupé, and Gail-Joon Ahn. Behind Closed Doors: Measurement and Analysis of CryptoLocker Ransoms in Bitcoin. In *APWG Symposium on Electronic Crime Research*, June 2016.

[49] Marco Lui and Timothy Baldwin. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30, 2012.

[50] Claudia Malzer and Marcus Baum. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, September 2020.

[51] UNITED STATES DISTRICT COURT DISTRICT OF MASSACHUSETTS. Securities and exchange commission v. james p. anglim. https://www.sec.gov/files/litigation/complaints/2023/comp25780.pdf.

[52] UNITED STATES DISTRICT COURT DISTRICT OF MASSACHUSETTS. Securities and exchange commission v. joseph a. padilla and kevin c. dills. https://www.sec.gov/files/litigation/complaints/2023/comp25745.pdf.

[53] Damon McCoy, Andreas Pitsillidis, Jordan Grant, Nicholas Weaver, Christian Kreibich, Brian Krebs, Geoffrey Voelker, Stefan Savage, and Kirill Levchenko. {PharmaLeaks}: Understanding the business of online pharmaceutical affiliate programs. In *21st USENIX Security Symposium (USENIX Security 12)*, pages 1–16, 2012.

[54] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2, 03 2017.

[55] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[56] Meta. About Meta's advertising policy on Cryptocurrency Products and Services, 2024. https://www.facebook.com/business/help/438252513416690?id=595195347635322.

[57] Najmeh Miramirkhani, Oleksii Starov, and Nick Nikiforakis. Dial one for scam: Analyzing and detecting technical support scams. In *22nd Annual Network and Distributed System Security Symposium (NDSS)*, volume 16, 2016.

[58] David Modic and Ross Anderson. It's all over but the crying: The emotional and financial impact of internet fraud. *IEEE Security & Privacy*, 13(5):99–103, 2015.

[59] Kate Moran and Kim Salazar. Large devices preferred for important tasks. https://www.nngroup.com/articles/large-devices-important-tasks/, August 2019.

[60] Oberlo. How Many Users Does Facebook Have? (2013–2023), 2023. https://www.oberlo.com/statistics/how-many-users-does-facebook-have.

[61] Adam Oest, Yeganeh Safaei, Penghui Zhang, Brad Wardman, Kevin Tyers, Yan Shoshitaishvili, and Adam Doupé. {PhishTime}: Continuous longitudinal measurement of the effectiveness of anti-phishing blacklists. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 379–396, 2020.

[62] Adam Oest, Penghui Zhang, Brad Wardman, Eric Nunes, Jakub Burgis, Ali Zand, Kurt Thomas, Adam Doupé, and Gail-Joon Ahn. Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020.

[63] ScamWatch National Anti-Scam Centre of Australian Goverment. Scam statistics 2023. February 204. https://www.scamwatch.gov.au/research-and-resources/scam-statistics?scamid=all&date=2023.

[64] Masarah Paquet-Clouston, Matteo Romiti, Bernhard Haslhofer, and Thomas Charvat. Spams Meet Cryptocurrencies: Sextortion in the Bitcoin Ecosystem. In *ACM Conference on Advances in Financial Technologies*, 2019.

[65] Youngsam Park, Jackie Jones, Damon McCoy, Elaine Shi, and Markus Jakobsson. Scambaiter: Understanding targeted nigerian scams on craigslist. *system*, 1:2, 2014.

[66] PhisTank.
https://phishtank.org/.

[67] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. *arXiv preprint arXiv:1806.01156*, 2018.

[68] Benjamin Price and Matthew Edwards. Resource networks of pet scam websites. In *2020 Symposium on Electronic Crime Research*. Institute of Electrical and Electronics Engineers (IEEE), 2020.

[69] Puppeteer.
https://pptr.dev/.

[70] Andrea Ramey. Scam artists drive seniors to suicide in mobile county. https://mynbc15.com/news/local/scum-of-the-earth-scam-artists-drive-seniors-to-suicide-in-mobile-county, October 2021.

[71] Andrew R.Chow. Here's how shopping scams on facebook are ripping off thousands of customers, with the money flowing overseas, December 2020. https://time.com/5921820/facebook-shopping-scams-holidays-covid-19/.

[72] Nils Reimers. Sentence transformers. https://www.sbert.net/.

[73] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[74] ScamAdviser. Lottery-guy scamadviser reviews. https://www.scamadviser.com/check-website/lottery-guy.com.

[75] ScamAdviser.
https://www.scamadviser.com/.

[76] Fake Website Buster Exposing Scams & Scammers. Augur capital (augurcapital.com.au) — fake or real? https://fakewebsitebuster.com/augur-capital-limited-augurcapital-com/.

[77] Fake Website Buster Exposing Scams & Scammers. Bitzent.com — fake or real? https://fakewebsitebuster.com/bitzent-com/.

[78] Broadband Search. Mobile vs. desktop internet usage 2024. https://www.broadbandsearch.net/blog/mobile-desktop-internet-usage-statistics.

[79] Silvia Sebastián and Juan Caballero. AVClass2: Massive Malware Tag Extraction from AV Labels. In *Annual Computer Security Applications Conference*, 2020.

[80] Financial Services and Markets Authority (FSMA). The FSMA warns against recovery room fraud, September 2020. https://www.fsma.be/en/warnings/fsma-warns-against-recovery-room-fraud-0.

[81] Sitejabber.
https://www.sitejabber.com/.

[82] Gilberto Atondo Siu, Alice Hutchings, Marie Vasek, and Tyler Moore. "invest in crypto!": An analysis of investment scam advertisements found in bitcointalk. In *APWG Symposium on Electronic Crime Research*, 2022.

[83] Johnny So, Najmeh Miramirkhani, Michael Ferdman, and Nick Nikiforakis. Domains do change their spots: Quantifying potential abuse of residual trust. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2130–2144. IEEE, 2022.

[84] Bharat Srinivasan, Athanasios Kountouras, Najmeh Miramirkhani, Monjur Alam, Nick Nikiforakis, Manos Antonakakis, and Mustaque Ahamad. Exposing search and advertisement abuse tactics and infrastructure of technical support scammers. In *Proceedings of the 2018 World Wide Web Conference*, pages 319–328, 2018.

[85] Statista. Number of X (formerly Twitter) users worldwide from 2019 to 2024, 2023. https://www.statista.com/statistics/303681/twitter-users-worldwide/.

[86] Guillermo Suarez-Tangil, Matthew Edwards, Claudia Peersman, Gianluca Stringhini, Awais Rashid, and Monica Whitty. Automatically dismantling online dating fraud. *IEEE Transactions on Information Forensics and Security*, 15:1128–1137, 2019.

[87] Karthika Subramani, Xingzi Yuan, Omid Setayeshfar, Phani Vadrevu, Kyu Hyung Lee, and Roberto Perdisci. When push comes to ads: Measuring the rise of (malicious) push advertising. In *Proceedings of the ACM Internet Measurement Conference*, pages 724–737, 2020.

[88] Bob Sullivan. The darkest side of online scams – when victims attempt suicide. https://bobsullivan.net/cybercrime/the-darkest-side-of-online-scams-when-victims-attempt-suicide/, September 2023.

[89] Aaron Swartz. Html2text. https://pypi.org/project/html2text/.

[90] Taboola.com: Content Discovery & Native Advertisin, 2024. https://www.taboola.com/.

[91] Ege Tekiner, Abbas Acar, A Selcuk Uluagac, Engin Kirda, and Ali Aydin Selcuk. SoK: Cryptojacking Malware. In *IEEE European Symposium on Security and Privacy*, 2021.

[92] TMJ4. Victim blaming prevents consumers from coming forward and reporting fraud. https://www.tmj4.com/news/i-team/victim-blaming-prevents-consumers-from-coming-forward-and-reporting-fraud.

[93] Truspilot. Galacticwins trustpilot reviews. https://www.trustpilot.com/review/galacticwins.com?stars=1.

[94] Truspilot. Gaysgodating trustpilot reviews. https://www.trustpilot.com/review/gaysgodating.com?stars=1.

[95] Truspilot. Maturedating trustpilot reviews. https://www.trustpilot.com/review/maturedating.com.

[96] TrustPilot. https://www.trustpilot.com/.

[97] Phani Vadrevu and Roberto Perdisci. What you see is not what you get: Discovering and tracking social engineering attack campaigns. In *Proceedings of the Internet Measurement Conference*, pages 308–321, 2019.

[98] Iman Vakilinia. Cryptocurrency giveaway scam with youtube live stream. In *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0195–0200. IEEE, 2022.

[99] VirusTotal. http://www.virustotal.com/.

[100] Thymen Wabeke, Giovane Moura, Nanneke Franken, and Cristian Hesselman. Counterfighting counterfeit: detecting and taking down fraudulent webshops at a cctld. In *International Conference on Passive and Active Network Measurement*, pages 158–174. Springer, Cham, 2020.

[101] John Wadleigh, Jake Drew, and Tyler Moore. The e-commerce market for" lemons" identification and analysis of websites selling counterfeit goods. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1188–1197, 2015.

[102] Monica T Whitty and Tom Buchanan. The online dating romance scam: The psychological impact on victims–both financial and non-financial. *Criminology & Criminal Justice*, 16(2):176–194, 2016.

[103] Pengcheng Xia, Haoyu Wang, Bingyu Gao, Weihang Su, Zhou Yu, Xiapu Luo, Chao Zhang, Xusheng Xiao, and Guoai Xu. Trade or trick? detecting and characterizing scam tokens on uniswap decentralized exchange. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 5(3):1–26, 2021.

[104] Pengcheng Xia, Haoyu Wang, Bowen Zhang, Ru Ji, Bingyu Gao, Lei Wu, Xiapu Luo, and Guoai Xu. Characterizing Cryptocurrency Exchange Scams. *Computers & Security*, 98, 2020.

[105] Hao Yang, Kun Du, Yubao Zhang, Shuang Hao, Zhou Li, Mingxuan Liu, Haining Wang, Haixin Duan, Yazhou Shi, Xiaodong Su, et al. Casino royale: a deep exploration of illegal online gambling. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 500–513, 2019.

Appendix

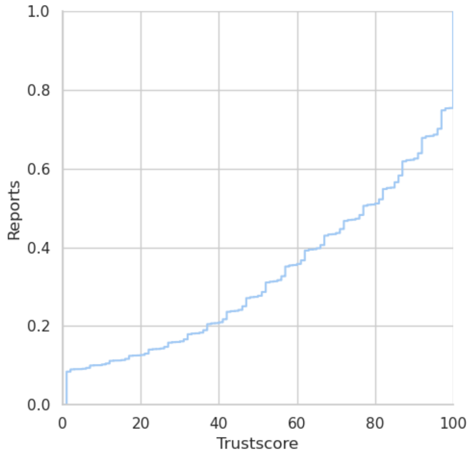

Figure 2: Distribution of ScamAdviser trust scores for all 21.1M reports in the feed.

A. Scam Clustering

This section clusters the 143,227 successfully crawled ScamAdviser domains by the similarity of their downloaded content. The goal is to examine to what degree unclassified ScamAdviser domains (i.e., with no industry tag or a low-quality tag) belong to the 7 selected scam types, and to identify additional scam types.

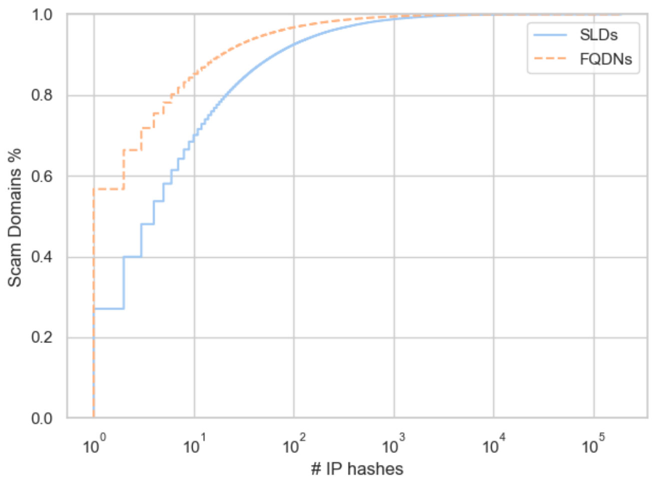

Figure 3: ECDF plot of the number of IPs observing each scam FQDN and SLD on both desktop and mobile telemetry.

Preprocessing. Prior to the clustering, we clean the text extracted from the 143K domains by removing adjacent blank spaces, newlines, and character tabulations. These do not provide semantic information and might push relevant descriptive sentences out of range of the feature extraction input.

Feature extraction. From each preprocessed text, we extract a feature vector using the Sentence-BERT embedding [73]. Specifically, we use the *all-MiniLM-L6-v2* pre-trained model provided by the SentenceTransformers Python library [72]. The language model takes only the first 256 tokens out of the full text contained in the main page to contribute to the final 384-dimensional embedding, which is the mean vector of the token embeddings after applying the attention mask. To alleviate the sparsity of the high-dimensional embedding space and to increase the density of local regions, following [8], we apply dimensionality reduction over the embeddings using UMAP [55] to get 5-dimensional projections using cosine distance and setting the size of the local neighborhood to 15 samples.

Clustering. We cluster the feature vectors using HDBSCAN [54], [50]. The minimum cluster size and minimum core point neighborhood size is set to 40, the distance threshold for merging clusters is zero, and the persistent clusters are selected by the excess of mass algorithm. The clustering produces 525 clusters containing 88,481 (61.8%) of the 143K scam webpages, with a median cluster size of 86 and the largest cluster having 4,866 webpages. The remaining 54,746 (38.2%) scam webpages are not similar to those in the clusters, and thus are placed by themselves in singleton clusters.

Label expansion. Among the 143K clustered scam domains, there are 58,703 (41%) that have been assigned one of the seven scam types. We apply a label expansion process to increase the labeling coverage. For each cluster, we first

Scam Type	Before Expansion Domains	After Expansion Domains
Shopping	26,134 (18.2%)	60,069 (41.9%)
Cryptocurrency	14,766 (10.3%)	22,684 (15.8%)
Financial	12,476 (8.7%)	15,420 (10.8%)
Gambling	3,489 (2.4%)	6,206 (4.3%)
Dating	878 (0.6%)	2,157 (1.5%)
Employment	838 (0.6%)	1,348 (0.9%)
Funds recovery	122 (<0.1%)	24 (<0.1%)
Unclassified	84,524 (59.0%)	35,319 (24.7%)
Total	143,227 (100%)	143,227 (100%)

Table XII: Classification of the crawled scam domains before and after applying label expansion on the clustering results.

compute the most common type for the classified domains in the cluster. We select that type as the cluster type, and apply it to all domains in the cluster. Clusters in which no domains had a type assigned are marked as unclassified. Table XII shows the number of domains assigned to each type before (left part) and after (right part) applying the label expansion on the clustering results. Using the expansion we reduce the unclassified domains more than half from 59.0% to 24.7%, showing that a significant portion of the unclassified domains belong to one of the seven selected types.

To validate the clustering results, we compare the labels from the human annotators. From the 1.2K domains that human annotators assigned to one of the seven scam types, 46% (552) remain as singletons after clustering and thus maintain their original ScamAdviser labels (if any). From the remaining 648 domains, 74% (479) of the domains are labeled the same both by the automated clustering and the human annotators. For the remaining 26% (169) domains both labels disagree. A closer look on the disagreements reveals that funds recovery scams are often mislabeled by the expansion as investment scams. This is likely due to those services often advertising the recovery of lost funds from cryptocurrency and other investment scams, making the webpage text resemble that of investment scams.

New scam types. The 35,319 unclassified domains after expansion belong to 94 clusters. We manually inspect these 94 clusters to check for new scam types. During this analysis, two authors inspected the screenshots, HTML content and text extracted for 20 domains from each cluster, 10 domains close to the cluster's centroid and 10 domains randomly selected from the remaining ones. We find that 51 clusters only contain custom error pages, 32 belong to the seven scam types but the expansion failed to tag them due to the lack of labels, and 11 clusters represent potentially new scam types. The new scam types include two clusters offering fake IT Help desk Services (177 domains), a cluster with US Postal Service (USPS) package delivery scams (141 domains), a cluster with membership cancellation services (52 domains), and clusters with the minimum 40 domains offering services like student assistance, personal care, house construction, and dream interpretation.