Domain and Website Attribution beyond WHOIS

Silvia Sebastián
IMDEA Software Institute
Universidad Politécnica de Madrid

Raluca-Georgia Diugan
IMDEA Software Institute

Juan Caballero
IMDEA Software Institute

Iskander Sanchez-Rola
Norton Research Group

Leyla Bilge
Norton Research Group

ABSTRACT

Currently, WHOIS is the main method for identifying which company or individual owns a domain or website. But, WHOIS usefulness is limited due to privacy protection services and data redaction. We present a novel automated approach for domain and website attribution. When WHOIS data does not reveal the owner, our approach leverages information from multiple other sources such as passive DNS, TLS certificates, and the analysis of website content. We propose a novel ranking technique to select the domain owner among multiple identified entities. Our approach identifies the domain owner with an F1 score of 0.94 compared to 0.54 for WHOIS. When applied on 3,001 tracker domains from the popular Disconnect list, it identifies needed updates to the list. It also attributes 84% of previously unattributed tracker domains.

CCS CONCEPTS

• **Security and privacy → Web application security**.

KEYWORDS

Attribution, Domain, Website, WHOIS, Trackers

ACM Reference Format:
Silvia Sebastián, Raluca-Georgia Diugan, Juan Caballero, Iskander Sanchez-Rola, and Leyla Bilge. 2023. Domain and Website Attribution beyond WHOIS. In *Annual Computer Security Applications Conference (ACSAC '23), December 04–08, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3627106.3627190

1 INTRODUCTION

Identifying which company or individual owns a domain or website is fundamental for many security applications including assisting law enforcement investigations [37], exposing intellectual property infringements [20], detecting phishing websites [17, 55, 56, 62, 80], reporting vulnerabilities in Internet-facing servers [27, 36, 52, 76], and measuring the coverage of Web tracking companies [32, 47, 70].

Domain and website attribution typically leverages the registration data (e.g., company name, person name, email, credit card) that domain registrars collect from registrants. Domain registration data can be accessed by law enforcement (e.g., by obtaining a warrant from a judge), but this can be a slow process especially if national boundaries are crossed (e.g., if a victim is in a country different from the ccTLD registrar managing the domain). Furthermore, some security-relevant scenarios like analyzing Web tracking domains may not necessarily violate the law (at least in some jurisdictions), making it difficult for law enforcement to perform the attribution. A subset of domain registration data (i.e., registrant name and email but not the payment data) can be publicly accessed through the WHOIS protocol. WHOIS is fundamental in scenarios where third parties other than law enforcement need to perform domain attribution including security companies investigating malicious domains, browser vendors that want to block tracking domains, and companies that believe their intellectual property has been abused. Unfortunately, WHOIS usefulness is limited due to two main challenges. First, registrars offer privacy protection (or proxy) WHOIS services, which replace the identity of the domain registrant with the registrar's identity, hiding the real domain owner [29]. Second, privacy regulations such as GDPR place limits on the WHOIS data that can be made available in some jurisdictions with 85% of large WHOIS providers redacting European Economic Area (EEA) records at scale, and over 60% also redacting non-EEA records [58].

One security application impacted by WHOIS limitations is the attribution of Web tracking domains [32, 47, 70], which is fundamental for measuring the coverage of Web tracking companies (and thus their impact on users' privacy) [32, 47], as well as for regulators to understand if some acquisitions may overly concentrate the market [47]. The challenge is that, as we will show, more than half of tracker domains lack useful WHOIS data. When WHOIS data is not useful, identifying the entity responsible for a domain becomes a tedious manual process. In particular, generating lists that link tracker domains to the entities behind them may require months of manual work [1, 3, 9]. Moreover, such attribution needs to be regularly repeated due to the dynamism of the targeted advertisement ecosystem, i.e., tracker domains changing ownership due to their companies being acquired, sold, or merged. For example, in 2017 Binns et al. manually created the X-Ray list [25], which in 2020, had to be manually updated to cope with the ecosystem changes [47].

Automating the domain attribution process is fundamental for scalability, and allows the attribution to be re-run periodically. A first step in this direction uses domain and IP WHOIS to automatically identify the owner of third-party tracker domains [70]. However, it suffers from the aforementioned problems of domain WHOIS (i.e., privacy protection services, data redaction). Moreover, IP WHOIS is noisy for attribution as most websites are hosted in cloud services where IP addresses may be shared or reused by different websites.

ACSAC '23, December 04–08, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0886-2/23/12…$15.00
https://doi.org/10.1145/3627106.3627190

This paper presents a novel automated approach for domain and website attribution. Given an input domain name, our approach outputs the most likely identity (i.e., organization or person name) for the domain owner. When WHOIS data is not useful, our approach leverages information from other sources such as passive DNS, TLS certificates, and the analysis of website content. In particular, it searches for Web servers on the input domain, its subdomains, and other domains from the same owner, examining their certificate and content including copyright strings, website metadata, privacy policies, terms of service (ToS) agreements, contact pages, and security.txt vulnerability disclosure policies [67].

There exists a tension between the right to privacy of domain and website owners and the need for accountability and transparency in the Web that attribution enables. In some cases, website users need attribution to protect their privacy rights, e.g., to identify the owner of tracker domains or websites that collect user data and whose privacy policies do not describe the entity responsible for the data collection. In other cases, attribution is undesirable, e.g., for websites with politically dissident content. We argue that novel attribution approaches are important, not only for security analysts, but also for privacy-sensitive domain owners, who can apply the proposed approaches to their websites to identify leaks that may lead to deanonymization. A similar argument applies to previous research on deanonymization in anonymity networks [44, 60].

A critical attribution challenge is that, while each source can provide useful attribution indicators, it can also lead to incorrect assessments due to, among others, privacy protection services in WHOIS, websites delegating TLS connections to hosting providers in certificate analysis, and third-parties mentioned in privacy policies and ToS agreements. Filtering misleading indicators is not enough because a blocklist can easily miss a previously unknown privacy protection service or hosting provider, and it is not possible to predict what third-parties may appear in a privacy policy or ToS agreement. To address this challenge, we propose a novel ranking technique, which identifies the domain owner identity among other third-party identities found in the different sources. The intuition behind our ranking is that the domain owner would be more prevalent than other entities. The ranking first clusters indicators by similarity and then ranks the clusters based on their indicator count, type, and source. The top-ranked cluster contains the domain owner indicators.

We have implemented our approach in a tool called WhoseDomain and have evaluated its attribution accuracy on a ground truth of 739 domains containing most popular domains, less popular domains, phishing targets, and tracker domains. Across all four datasets, WhoseDomain achieves a precision of 0.93, recall of 0.94, and F1 score of 0.94, compared with a F1 score of 0.59 when only using WHOIS. Tracker domains are hardest to attribute with WhoseDomain achieving a F1 score of 0.86 compared to 0.55 for WHOIS. We evaluate each data source, showing that content attribution performs best, but all sources contribute towards the attribution results. We apply WhoseDomain to 3,001 tracker domains in the Disconnect list [3], used by Firefox to block tracking, showing that it can identify needed updates to the list due to ownership changes. Finally, we apply WhoseDomain on 3,710 unattributed tracker domains, showing that it can automatically attribute 84%.

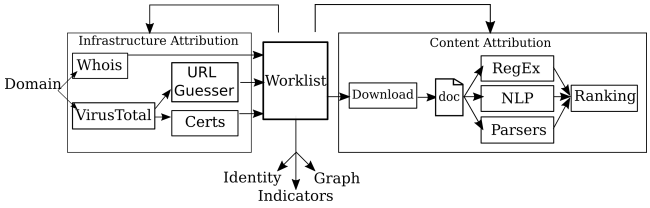

Figure 1: WhoseDomain's architecture.

This paper presents the following contributions:

- A novel automated approach for domain and website attribution that beyond performing WHOIS queries, also examines the content and configuration of Web servers hosted by the input domain, its subdomains, and other domains from the same owner.
- A novel ranking technique to identify the identity of the domain owner among multiple extracted identities. It groups indicators by similarity and ranks the clusters such that the top-ranked cluster corresponds to the domain owner.
- A novel NLP technique to extract the first-party (i.e., the domain owner) from definition sentences in legal documents, e.g., privacy policies and ToS agreements.
- An evaluation of our approach using a ground truth of 739 domains and 3,001 tracker domains. WhoseDomain achieves an F1 score of 0.94 compared to 0.54 using only WHOIS.

2 OVERVIEW

Given a domain, WhoseDomain outputs the identity of the domain owner together with a list of *indicators* from the owner (e.g., contact email and social network handles). An *indicator* comprises a type (e.g., *identity*), a string value (e.g., Adverts,Inc.), and an optional list of *sources* (e.g., *Whois*, *certificates*) from where the indicator was extracted. It also outputs an attribution graph that details the attribution steps, making the attribution process transparent. Nodes in the attribution graph are indicators. An edge from indicator A to indicator B captures that B was discovered from A and both indicators belong to the same owner.

Figure 1 shows the architecture of WhoseDomain, which performs an iterative process. At each iteration, it selects an indicator from the worklist (using a FIFO policy), applies a set of expansions specific to the indicator type to obtain new indicators, and adds them to the worklist. The process starts by placing the input domain in the worklist and finishes when an identity is found, the worklist is empty, or a maximum number of iterations is reached. The iterative process outputs the owner identity, other indicators belonging to the owner, and the attribution graph.

WhoseDomain expansions are split into two modules: infrastructure attribution (detailed in Section 3) and content attribution (detailed in Section 4). Infrastructure attribution uses WHOIS to identify the domain owner and passive DNS data to obtain its subdomains. For each domain, it looks for an associated Web server, examines its TLS certificate, and finds possible URLs to analyze. Content attribution downloads the document pointed by a URL, and extracts indicators from it using regular expressions, natural language processing (NLP), and document-specific parsers. Extracted

Indicator	Class	Extraction
identity	identity	NLP
organization	identity	NLP
personName	identity	NLP
email	contact	Regex
facebookHandle	social	Regex
githubHandle	social	Regex
instagramHandle	social	Regex
linkedinHandle	social	Regex
pinterestHandle	social	Regex
skypeHandle	social	Regex
telegramHandle	social	Regex
twitterHandle	social	Regex
whatsappHandle	social	Regex
youtubeHandle	social	Regex
copyright	ipr	Regex
trademark	ipr	Regex
aboutUrl	network	Regex
contactUrl	network	Regex
fqdn	network	Regex
esld	network	Regex
privacyUrl	network	Regex
securityUrl	network	Regex
tosUrl	network	Regex
url	network	Regex

Table 1: Indicators extracted by WhoseDomain, their class, and whether extracted using a regular expression or NLP.

indicators may belong to the domain owner or third parties. The ranking module (detailed in Section 5) selects the real owner among all identities identified. It clusters the indicators by similarity and selects the identity of the top-ranked cluster as the owner.

Indicators. The goal of WhoseDomain is to discover an *identity* indicator that identifies the domain owner. To achieve that goal, WhoseDomain needs to support a variety of other indicator types. In particular, WhoseDomain uses other indicator types to *pivot* using external datasets, i.e., to discover other indicators that belong to the same owner. For example, given a domain, passive DNS can be used to obtain subdomains that may host Web servers with certificates or content. In another example, the certificate of a website hosted on the input domain (e.g., www.amazon.com) may contain another domain (e.g., www.amzn.com) belonging to the same owner. Thus, WhoseDomain can jump from attributing the input domain to attributing other domains identified as belonging to the same owner. In addition, other indicator types serve as a context in the ranking module, detailed in Section 5.

Currently, WhoseDomain extracts the 24 indicators in Table 1. These indicators have been selected because of the following reasons. First, some indicators already capture an identity (i.e., *identity*, *organization*, *personName*). Second, other indicators may directly contain an identity (e.g., *copyright*). Third, some indicators intuitively lead to other indicators. For example, URLs and email addresses contain a domain name. Fourth, the resources captured by some indicators may require the owner to register its identity with a third party that can be leveraged by law enforcement as an attribution point (e.g., other domains, social network handles, email accounts). Finally, WhoseDomain supports five *url* subtypes linking to documents that may provide valuable attribution information: privacy policies (*privacyUrl*), ToS agreements (*tosUrl*), contact information webpages (*contactUrl*), descriptions of the website owner (*aboutUrl*), and security.txt policies describing how security issues in the website should be disclosed (*securityUrl*). The set of indicators and expansions potentially useful for attribution is quite broad. For that reason, WhoseDomain has a flexible and modular architecture that easily allows adding new indicators, expansions, and sources.

WhoseDomain supports two domain name indicators: fully qualified domain names (*fqdn*) and effective second-level domains (*esld*), also known as apex or base domains. The *esld* is the part of a *fqdn* that captures the owner. While in general the 2LD corresponds to the owner, if the 2LD assigns subdomains to third parties, the *esld* is the 3LD. For example, for www.amazon.com the *esld* is amazon.com, but for www.amazon.co.uk, it is amazon.co.uk. To extract an *esld* from a *fqdn*, WhoseDomain leverages the Public Suffix List (PSL) [63].

3 INFRASTRUCTURE ATTRIBUTION

These expansions examine WHOIS and Web servers on the input domain, its subdomains, and other domains of the same owner.

WHOIS.. The WHOIS protocol provides current registration data about domains and IP addresses. The WHOIS protocol is arguably the principal mechanism used by analysts to obtain information about registered domain names including the identity of the registrant, the registrar, and contact information for administrative and technical issues. WhoseDomain queries WHOIS with a given *esld* and parses the different response formats [57] using a popular library [10]. The WHOIS protocol is rate-limited, but WhoseDomain is not affected because it only performs an average of less than 3 WHOIS queries in each exploration (i.e., one for the input *esld* and another for each additional *esld* from the same owner that the exploration discovers). If higher rates are needed (e.g., to attribute many domains in parallel), commercial WHOIS services could be added to WhoseDomain. Protected domains may return empty registrant data, a generic string (e.g., *REDACTED FOR PRIVACY*), or the identity of a privacy protection service (e.g., *Domains By Proxy, LLC*). WhoseDomain uses a *WHOIS blocklist* to filter out such responses. To build this blocklist, we leverage a list of the top 100 WHOIS registrant strings obtained by querying 285M domains [11]. We generalize some list entries as regular expressions to cover cases such as "Jewella Privacy LLC Privacy ID# 14730082" and "Jewella Privacy LLC Privacy ID# 876917".

Passive DNS.. Given an *esld*, WhoseDomain uses the VirusTotal (VT) API [8] to obtain the list of subdomains in VT's passive DNS data. In general, subdomains should belong to the domain owner. However, some domains lease subdomains to third parties such as those of dynamic DNS providers (e.g., ddns.net), blog platforms (e.g., blogspot.com), and web hosting services (e.g., googlepages.com). WhoseDomain avoids exploring such subdomains by using a *user-Subdomain* blocklist with 209 *esld* that lease subdomains to third parties. To build the blocklist we use Mozilla's Public Suffix List [63] and also query the top 10K domains in the Tranco list [15][1] to VT, checking if the number of subdomains VT observed is larger than a

[1]We use the Tranco list from May 1, 2023 throughout the paper

threshold. We choose the threshold to be 50 following the procedure detailed in Appendix B.

Certificates. An HTTPS certificate can attribute a domain if it contains an organization name and is part of a valid certificate chain (i.e., certificates are not expired, chain ends in trusted CA). While free certificates (e.g., those issued by Let's Encrypt [50]) do not contain an organization, those issued by commercial CAs sometimes do. Furthermore, certificates may be valid for multiple domains listed in the Subject Alternative Name (SAN) extension. Those additional domains should belong to the same entity unless the entity is a hosting provider. WhoseDomain uses a *hostingProvider* blocklist to avoid analyzing hosting provider certificates. The blocklist contains 1,472 domains from a public catalog of top hosting providers [4]. Any *fqdn* (i.e., subdomain of the input *esld* or of a different *esld*) left after filtering will be examined in the following iterations since their attribution also attributes the input domain.

URL guesser. Checks if a given domain hosts Web servers. If either HTTP (tcp/80) or HTTPS (tcp/443) connections to the domain succeed, four URLs are returned for each successful connection: http(s)://DOMAIN/, http(s)://www.DOMAIN/, http(s)://DOMAIN/.well-known/security.txt, and http(s)://DOMAIN/security.txt. The last two are the possible locations defined by the security.txt standard [67]. URLs are analyzed by the content attribution module.

Other expansions. WhoseDomain also extracts the *fqdn* from a *url* and the *esld* from a *fqdn* using Mozilla's Public Suffix List [63].

4 CONTENT ATTRIBUTION

The content attribution module downloads and analyzes the document pointed by a URL. Only selected URLs in a discovered website are analyzed by the content attribution module, i.e., WhoseDomain does not crawl the discovered websites. In particular, the Guesser module generates URLs for the root page and the security.txt resource. In addition, URLs in the root page matching the regular expressions for privacy policies, terms of service, about us, and contact us resources are also passed to the content attribution module. These URLs are selected because they often contain useful attribution information. Crawling is not perfomed because websites can contain arbitrary content that can introduce false leads. For example, an online newspaper contains many articles mentioning identities related to the article's content, but unrelated to the owner of the newspaper. Those articles may contain links to previously related articles published by the newspaper as well as external references, which would further introduce false leads.

The content attribution module supports four document types common in websites: HTML, PDF, security.txt, and plain text. Indicators are extracted from the document's text: *readable text*, *other visible text*, and *nonvisible text*. Other types of content (e.g., images) are not analyzed as they rarely convey attribution information. Readable text corresponds to the main text of the document, e.g., the policy text for a privacy policy. In plain text files, the whole content can be considered readable. For PDF files, it is the concatenation of all the text objects. In webpages, it corresponds to the Reader View in browsers like Firefox, i.e., the subset of visible text after removing areas with a large density of links such as headers and footers [46]. Other visible text corresponds to parts of a webpage that are not the main text but are visible to a user such as headers and footers. Those areas may include useful attribution indicators such as social network accounts and copyright strings. Nonvisible text is text not shown to the user. It includes PDF and HTML metadata, which capture document properties (e.g., author), and scripts in webpages, which may contain Schema.org data [6]

To extract indicators from a document, content attribution uses regular expressions, document parsers, and two NLP techniques: Named Entity Recognition (NER) models and a novel synonym definition extraction technique.

Regular expressions. WhoseDomain uses the regular expressions in the *iocsearcher* tool [26] to extract 21 indicators with intrinsic structure such as email addresses, URLs, and copyright strings. Those regular expressions are also used to validate the correctness of the fields extracted by the document parsers (e.g., that a *mailto* link indeed contains an email address). A limitation of regular expressions is that they cannot accurately extract indicators without a well-defined structure such as organization and person names.

Document parsers. WhoseDomain parses HTML, PDF, and security.txt files to examine specific fields, which are grouped into four sources. (1) A dictionary of (field, value) pairs is built from the *metadata* object in PDF documents, and the title and meta tags in HTML webpages. (2) Regular expressions are applied on the contact field of *security.txt* files to extract *email* and *contactUrl* indicators. (3) Regular expressions are applied to the URL of HTML *link* tags to extract email addresses (*mailto:* scheme), Skype handles (*skype:*), and handles for a variety of social networks and IM services. The link's textual description is part of the readable text and is also searched for keywords to classify the URL into the 5 subcategories introduced in Section 2. (4) WhoseDomain examines HTML script tags for the presence of Schema.org data [6]. *Schema.org* maintains a large set of schemas for structured data that can be embedded in webpages. The schemas are used by over 10M sites [6]. The JSON data is parsed and selected fields such as Organization:email are examined with regular expressions to extract indicators.

NER.. NER models identify entities in natural text using neural networks. They do not require a pre-defined list of terms to identify. This is fundamental in our *open-world* scenario where any entity can be the domain owner. WhoseDomain identifies organization and person names in the readable text using Stanford's CoreNLP models for English, Chinese, French, German, and Spanish [7]. Off-the-shelf NER models are not specifically trained for our particular use case. Unfortunately, training a NER model from scratch requires very large annotated training data to achieve reasonable accuracy. Andow et al. improve NER accuracy through domain adaptation [19], i.e., updating an off-the-shelf NER model using additional training data from privacy policies. However, any NER model will still introduce noise in the form of incorrect entities (false positives). Thus, we opt to use the existing aforementioned models and address the false positives by introducing a novel ranking technique to identify the domain owner among multiple entities. We leave improving the NER models through domain adaptation or full re-training as future work.

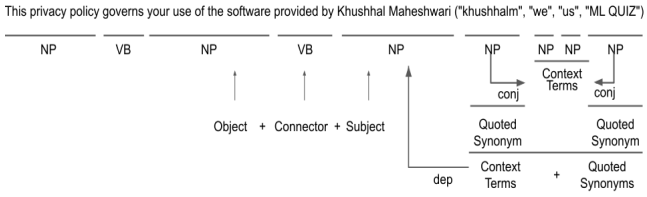

Figure 2: Synonym definition extraction example.

Synonym definition extraction. A common idiom in legal documents (e.g., privacy policies, ToS agreements) is to define the responsible first party, together with its synonyms, in the preamble of the document. Figure 2 shows an example definition sentence, which captures that the first party is "Khushhal Maheshwari" and that "khushhalm" and "ML QUIZ" are synonyms for it. We have developed a novel NLP technique to extract first-party entities and synonyms from such sentences in documents written in English and Spanish, languages for which we are fluent.

The technique first segments the readable text into sentences and discards those that do not contain language-specific *context terms*. In English, context terms appear quoted (e.g., *"we"*, *"us"*, *"provider"*). While context terms are widely used in documents, they only appear quoted in such definitions. In Spanish, context terms are not quoted (e.g., *a partir de ahora*, *con domicilio en*). Each sentence with context terms is analyzed using an NLTK [5] pipeline that focuses on nominal groups. A nominal group must contain a noun, but it may include other modifiers such as adjectives (e.g., *Mad Dog Studios*), other nouns (e.g., *Fox News*), or prepositions (e.g., *Moody Bible Institute of Chicago*). Intuitively, identities should appear in nominal groups. To identify nominal groups (NP) we define a grammar to split the sentence into parts of speech.

```
NP: {<JJ>?<NN.*>+<PP|IN>?<DT>?<JJ>?<NN.*>+<,>?<NN.*>}
    {<JJ>+<NN.*>+} {<NN.*><CD>} {<NN.*>+}
```

A sentence could include multiple nominal groups. For example, in Figure 2, the English grammar identifies *Khushhal Maheshwari*, *This privacy policy*, and *your use of the software*. We apply dependency parsing to identify nominal groups that depend on context terms, illustrated as arrows in Figure 2. Synonyms in definition sentences often appear quoted in both languages. Quoted terms that the dependency parser identifies as conjunctions (*conj* in Figure 2) of a context term are considered *quoted synonyms*, i.e., *"khushhalm"* and *"ML QUIZ"*. Then, all nominal groups that depend on a context term or quoted synonym are considered identities. In Figure 2, *Khushhal Maheshwari* is output as an identity with both *"khushhalm"* and *"ML QUIZ"* as synonyms for it.

5 RANKING

The goal of our novel ranking technique is to choose the subset of indicators that belong to the domain owner, among all indicators extracted by the content and infrastructure attribution modules. The ranking needs to differentiate the domain owner identity (and its aliases) from third-party identities (e.g., hosting providers in infrastructure attribution, regulators and advertisers in privacy policies, and website designers in copyright strings) and identities that do not correspond to any real entity (e.g., NER false positives). At a high level, the ranking first groups indicators by similarity such that each cluster roughly corresponds to the profile of an entity with its indicators. Then, clusters are ranked based on their number of indicators, indicator type, and sources they come from. The top-ranked cluster captures the domain owner identity and its indicators. The key insight is that the domain owner would be mentioned more often and in more ways (i.e., aliases) than other entities. For example, the domain owner may be mentioned in the visible text and in a copyright string in the footer, and its social network handles and domain may appear in links. In contrast, third parties appear in fewer sources, have less indicators, and are mentioned fewer times and use less aliases.

The ranking module comprises three steps: indicator *expansion*, indicator *clustering*, and cluster *ranking*. We describe them next using Figure 3 as an example where the 12 extracted indicators in Figure 3a are the input to the ranking.

Expansion. In addition to the expansions in Sections 3 and 4, WhoseDomain supports some *framework expansions* that do not require external requests. For *copyright* strings, it extracts the *identity* inside by removing all other components including copyright symbols, years, and common strings such as "All Rights Reserved". For *fqdn*, it extracts the *esld* using the PSL [63]. For *email*, it extracts the *fqdn* and filters it using an *email provider* blocklist that contains 3,791 email providers from a public list [18]. For *url*, if it belongs to a social network (Facebook, GitHub, Instagram, LinkedIn, Pinterest, Twitter, YouTube) or IM application (Telegram, WhatsApp) it extracts the handle using a regular expression. Otherwise, it extracts the *fqdn* and filters it out using the *userSubdomain* blocklist presented in Section 3 and a *userContent* blocklist with 98 *esld* for Web services that host user content (e.g., Facebook, Dropbox, Google Docs). In Figure 3b, the copyright has been expanded into *identity Cox Media Group* and the email address into *esld cmg.com*.

Clustering. Indicators are clustered based on their value similarity and alias information. First, each indicator is decomposed into a list of lowercase tokens. For *email*, one token with the username is generated since the domain was already extracted in the expansion step. For *esld*, one token is generated with the value minus the TLD since the same TLD may appear in multiple domains. For other indicators, a single token is produced with its value. Indicators are grouped using an agglomerative clustering that considers two indicators similar if at least one of the following conditions holds: if one is an alias of the other according to the synonym definition extraction, if one is the acronym of the other (e.g., *CMG* and *Cox Media Group*), if they share the same prefix (e.g., *CMG* and *CMG Affiliate*), or if their longest contiguous matching subsequence (LCS) ratio is higher than a predefined threshold (e.g., *countryLegends971* and *Legends971*). We selected 0.7 as the LCS threshold following the procedure in Appendix A. Figure 3c shows the 5 produced clusters.

Ranking. The ranking assigns a weight to each cluster and outputs the top-ranked cluster as corresponding to the domain owner. Each clustered indicator is assigned a weight based on the sources from where they were extracted. Higher confidence sources are set to a weight of 5, the remaining are set to 1. There are two high-confidence sources: (1) indicators extracted from nonvisible Schema.org data are trustworthy because they have been explicitly added by the website administrator and their extraction is very

```
email           cmgcopyright@cmg.com    ['link','regexp']
esld            daytondailynews.com     ['regexp']
facebookHandle  countrylegends971       ['link']
identity        Google Analytics        ['ner']
identity        American Arbitration Association  ['ner']
identity        American Stock Exchange ['ner']
identity        CMG                     ['dependency','ner']
identity        CMG Affiliate           ['dependency','ner']
identity        New York Stock Exchange ['ner']
organization    Cox Media Group, Inc.   ['ner']
twitterHandle   Legends971              ['link']
```

(a) Extracted indicators provided as input to the ranking

```
esld            cmg.com                 ['regexp']
esld            daytondailynews.com     ['regexp']
facebookHandle  countrylegends971       ['link']
identity        Google Analytics        ['ner']['generic']
identity        American Stock Exchange ['ner']
identity        CMG                     ['dependency','ner']
identity        CMG Affiliate           ['dependency','ner']
identity        New York Stock Exchange ['ner']
organization    Cox Media Group         ['regexp']
organization    Cox Media Group, Inc.   ['ner']
twitterHandle   Legends971              ['link']
```

(b) Expanded and filtered indicators

```
identity        CMG                     ['dependency','ner']
identity        CMG Affiliate           ['dependency','ner']
organization    Cox Media Group         ['regexp','ner']
organization    Cox Media Group, Inc.   ['ner']
email           cmgcopyright@cmg.com    ['link','regexp']
esld            cmg.com                 ['regexp']
CLUSTER 1:
facebookHandle  countrylegends971       ['link']
twitterHandle   Legends971              ['link']
CLUSTER 2:
identity        American Stock Exchange ['ner']
identity        New York Stock Exchange ['ner']
CLUSTER 3:
esld            daytondailynews.com     ['regexp']
CLUSTER 4:
identity        Google Analytics        ['ner']['generic']
```

(c) Clustered indicators

```
organization    Cox Media Group         ['regexp','ner']     7
organization    Cox Media Group, Inc.   ['ner']               6
identity        CMG                     ['dependency','ner'] 6
identity        CMG Affiliate           ['dependency','ner'] 6
email           cmgcopyright@cmg.com    ['link','regexp']    2
esld            cmg.com                 ['regexp']           1
CLUSTER 1: 2
facebookHandle  countrylegends971       ['link']             1
twitterHandle   Legends971              ['link']             1
CLUSTER 2: 2
identity        American Stock Exchange ['ner']              1
identity        New York Stock Exchange ['ner']              1
CLUSTER 3: 1
esld            daytondailynews.com     ['regexp']           1
CLUSTER 4: 1
identity        Google Analytics        ['ner']['generic']   1
```

(d) Ranked clusters

```
organization    Cox Media Group         ['regexp','ner']     7
organization    Cox Media Group, Inc.   ['ner']               6
identity        CMG                     ['dependency','ner'] 6
identity        CMG Affiliate           ['dependency','ner'] 6
email           cmgcopyright@cmg.com    ['link','regexp']    2
esld            cmg.com                 ['regexp']           1
```

(e) Winner cluster

Figure 3: Ranking example.

accurate and (2) identities extracted by the synonym definition extraction since that technique has higher precision (i.e., less false positives) than the NER. The indicator weight is the sum of the weights of its sources. For example, in Figure 3d *email cmgcopyright@cmg.com* weights two as it was extracted from the *link* and *regexp* sources, while the *identity CMG Affiliate* has a weight of six due to its *dependency* and *NER sources*. In addition, *organization* indicators are given a bonus weight of five to favor full organization names over their acronyms. For example, *Cox Media Group, Inc.* has a weight of 6, one for its *NER* source and five for being an *organization*. The weight of a cluster is the sum of the weights of its indicators. The cluster with the largest weight that contains at least one identity is chosen to be the domain owner, i.e, *Cluster 0* with weight 28 in Figure 3d. In the example, the winner cluster contains six indicators for the domain owner: the full identity (*Cox Media Group, Inc.*), three aliases also used to identify it (*Cox Media Group*, *CMG*, *CMG Affiliate*), the contact email for privacy policy questions, and the company's domain name.

6 EVALUATION

This section evaluates WhoseDomain. We first present the datasets used in Section 6.1. Then, we measure WhoseDomain attribution accuracy in Section 6.2. Next, we study the contribution of the different sources to the attribution accuracy in Section 6.3. and the impact of the ranking in Section 6.4. We then apply WhoseDomain to identify needed updates to the Disconnect list in Section 6.5. Finally, we evaluate WhoseDomain for attributing previously unattributed tracker domains in Section 6.6 and provide a case study on impersonation in Section 6.7.

6.1 Datasets

We use four datasets in our evaluation: a manually generated ground truth of 739 domains with their owner identity, the Disconnect list with 3,001 tracker domains, a manually generated list of 100 privacy policies labeled with the first-party identity in the policy, and 3,710 unattributed tracker domains. We detail these datasets next.

Domain owner ground truth. We manually build a ground truth (GT) with the owner entities for 739 domains, split into four datasets. The *tranco_top* dataset has the top 250 domains in the Tranco domain popularity list [66] while *tranco_100K* has 250 domains starting at position 100K in the Tranco list. We use these datasets to analyze the impact of domain popularity on attribution accuracy. The *brands* dataset was provided to us by a large security vendor. It contains the main domain for 100 large companies often targeted by phishing attacks (e.g., large banks, social networks). Finally, the *trackers* dataset contains 139 tracker domains randomly sampled from the Disconnect list, which is described below. To build the GT, two analysts examined the same sources that WhoseDomain uses (i.e., WHOIS, domain certificates, VT, website content), as well as external sources that WhoseDomain does not currently use (e.g., Google searches, company databases [31]). Each analyst independently attributed the domains and discussion followed until an agreement was reached. Thus, the GT allows us to evaluate how well WhoseDomain performs compared to human analysts. By using the GT, we can show that WhoseDomain can *automatically* attribute domains that are known to be attributable. We compute the string similarity between domain names and their owners using the ranking module. Across the 739 domains, 30% are not similar to any of their owners in the GT. This highlights the need for attribution, as the domain name does not always allow to identify the owner entity.

Disconnect. We use the version of the Disconnect list from May 11, 2022, which contains 3,001 tracker domains associated with 1,425 entities. While the Disconnect list associates domains to the entities that own them, we cannot use the list entries as GT because, oftentimes, the stated owner does not correspond to the latest owner, e.g., due to a company acquisition. That is the reason why we only included 139 Disconnect domains in our domain GT. For those 139 tracker domains, we had to perform the same manual

analysis done on the remaining GT domains. In particular, the analysts attributed 37% of the 139 Disconnect domains to owners different from those in Disconnect: for 22% of domains they updated the owner and for another 15% they added an additional owner (e.g., a parent company). This illustrates that domain attribution is often not a once-and-done process, but needs to be repeated over time. Even if tracker domains in Disconnect were correctly attributed when first added to the list, the entries often become stale. Periodically re-attributing all list entries is not feasible with the current manual approach. For example, it would have taken weeks for our analysts to manually label all 3,001 domains in Disconnect, so they restricted the analysis to 139 randomly selected domains. An automated approach such as the one we propose is needed for such periodic re-attribution.

Privacy policy ground truth. We also build a GT dataset of 100 privacy policies with the first-party identity in the policy, i.e., the owner of the domain the policy was collected from. We use this dataset to evaluate the content attribution and ranking modules. We focus on privacy policies because privacy regulations (e.g., GDPR, CCPA) require them when websites serve users from their regions.

Unattributed tracker domains. We obtain 3,710 unattributed tracker domains from a large security vendor. These are tracker domains that the vendor has identified using an in-house tracker detection system and that were not in public tracker lists at the end of May 2022. Thus, they were unattributed. We use this dataset in Section 6.6 to evaluate WhoseDomain for attributing previously unattributed tracker domains.

6.2 WhoseDomain Attribution Accuracy

We evaluate the attribution accuracy of WhoseDomain by comparing the domain owners it identifies with those in the GT. The evaluation needs to handle similar, but not identical, identities, as well as domains with multiple owners where only one owner is in the GT. For this, the indicators output by WhoseDomain are clustered with the GT identities using the clustering step of the ranking module. If the GT identity is in the same cluster as an identity output by WhoseDomain, the result is a true positive (TP). i.e., WhoseDomain correctly attributed the domain to its owner in the GT. If WhoseDomain did not output any identity, the result is a false negative (FN). i.e., WhoseDomain could not attribute the domain. And, if the GT identity is in a cluster without other identity indicators extracted by WhoseDomain, the result is a false positive (FP). i.e., WhoseDomain attributed the domain to an entity different from one in the GT.

The *WhoseDomain* part in Table 2 summarizes the attribution accuracy of WhoseDomain with a maximum of 150 iterations per input domain and using all sources. Across all GT datasets, WhoseDomain achieves a precision of 0.93, recall of 0.94, and F1 score of 0.94. The accuracy on the *tranco_top* and *brands* datasets is very high with F1 scores of 0.98 and 0.96, respectively. The F1 score reduces to 0.92 on *tranco_100K* indicating that lower domain popularity makes attribution harder. Attributing tracker domains is even more challenging with an F1 score of 0.86. This was expected since advertisement companies may have an interest in staying under the radar and not being associated with their tracker domains. For instance, the *Web* column in Table 2 captures the number of GT domains with a Web server. Only 69% of the tracker domains have a Web server, compared to 95%–99% in the other datasets. WhoseDomain reaches the maximum number of iterations for only 9 (1.2%) domains. Among the other FNs, a significant number are due to abandoned tracker domains, an issue we explore in Section 6.5. A common reason for FPs are website design services. Typically the owner predominates in the content and thus is selected by the ranking. However, in a few cases, the website designer appears so often that it is ranked above the owner. To address this issue, we plan to investigate changes to the ranking, e.g., lowering the score of identities containing keywords related to website design.

6.3 Attribution Sources Impact

This section examines the impact of the attribution sources. First, Table 2 shows an ablation study using only a single source at a time. We focus on the sources that can return identities (i.e., WHOIS, Certificates, Content). Then, Table 3 evaluates how sources complement each other by incrementally adding attribution sources.

WHOIS.. An FN corresponds to the domain having no WHOIS entry, a redacted entry, or a WHOIS identity in the blocklist of privacy protection services. An FP is an identity that does not match the GT. Across all four GT datasets, WHOIS achieves a precision of 0.95, recall of 0.43, and F1 score of 0.59. FPs are rare (2.1%), but FNs abound (55.9%) due to privacy protection services and data redaction. However, note that if we were to use the command line WHOIS, FNs in Table 2 may be FPs instead, as without the WHOIS blocklist privacy protection services may be identified as incorrect identities. Perhaps surprisingly, tracker domains can be attributed more often than the less popular domains and the phishing targets due to the high prevalence of protection services in those datasets. Thus, not all tracker domains hide their identity in WHOIS, although many do. In summary, WHOIS attribution is not enough as it only attributes 42% of the GT domains with an F1 score of 0.59. compared to 88% and 0.94 for WhoseDomain. The increased attribution by WhoseDomain is due to the use of additional sources, which we examine next.

Certificates. Using only certificates, attribution happens if the input domain hosts an HTTPS server, the server has a valid certificate, and the certificate contains the domain owner identity in the Subject's Organization attribute. Domains in the brands dataset can be attributed with a surprisingly high F1 score of 0.75 because they tend to provide a valid certificate with their company name. This may be due to phishing targets trying to help the user identify the proper domain owner. In the past, phishing targets used extended validation (EV) certificates for this purpose, although such certificates are no longer considered useful [23, 43]. On the other side, certificates only attribute 4% of tracker domains, largely due to only 38% of tracker domains having an HTTPS Web server with a valid certificate. Limited HTTPS support by trackers has been identified in prior work as a barrier to full HTTPS adoption [34]. Note that certificates without an Organization are still useful if they mention other domains or subdomains from the same owner, allowing WhoseDomain to pivot to those domains. We explore this effect when combining sources.

			WhoseDomain			WHOIS			Certificates			Content		
Dataset	**GT**	**Web**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
tranco_top	250	247	0.96	0.99	0.98	0.96	0.64	0.77	0.94	0.41	0.58	0.91	0.79	0.85
tranco_100K	250	237	0.89	0.95	0.92	0.93	0.29	0.44	0.94	0.19	0.31	0.93	0.80	0.86
brands	100	98	0.93	1.00	0.96	0.91	0.31	0.46	0.98	0.65	0.75	0.93	0.67	0.78
trackers	139	96	0.94	0.80	0.86	0.96	0.39	0.55	1.00	0.04	0.08	0.91	0.46	0.61
All	739	678	0.93	0.94	0.94	0.95	0.43	0.59	0.96	0.30	0.45	0.92	0.71	0.80

Table 2: Attribution accuracy for WhoseDomain with all sources and ablation study using only one source at a time. *GT* is the number of domains in the ground truth and *Web* the number of those domains with a Web server on ports 80/tcp or 443/tcp.

Content attribution. We use the Guesser to generate URLs on the input domain and then apply content attribution to those URLs. Table 2 shows that Web content attributes more (F1 score of 0.80) than WHOIS (0.59) and certificates (0.45). False negatives are dominated by domains without a Web server and a few websites with no real content (e.g., offering a Web service), both of which are especially common among tracker domains. For the interested reader, Appendix C evaluates the contribution of each content attribution technique to the attribution results.

In summary, content attribution performs best among individual sources on all datasets. But, WhoseDomain outperforms content attribution on all datasets, highlighting how the combination of sources in WhoseDomain improves domain and website attribution.

Combining sources. Next, we check whether all sources contribute towards the attribution. Since there are many possible combinations of sources, we focus on what we believe is the most logical one starting with WHOIS as the most popular source, then adding content attribution which the ablation study shows is most powerful, next adding VT to discover subdomains, and finally analyzing the certificates on all discovered domains. Table 3 summarizes the results. It shows that as more expansions are added, the F1 score increases. This highlights that all expansions positively contribute to the final attribution results. The largest increase is achieved by adding content attribution to WHOIS, as already hinted in the ablation study. But, adding VT and certificates still produces significant gains, e.g., 0.14 F1 score increase in the brands dataset. For example, tracker domain acuityplatform.com uses a WHOIS privacy protection service and has no website. A query to VirusTotal reveals 44 subdomains, which are all analyzed by WhoseDomain for websites. Most subdomains have a Web server with a certificate for CN=*.acuityplatform.com. However, origin.acuityplatform.com has a different certificate with an organization name of *AcuityAds Inc.*, which is the correct owner. Such configuration differences are often due to manual work, and identifying them typically requires an automated approach like ours.

6.4 Ranking Evaluation

We evaluate the accuracy of the ranking using the GT of 100 privacy policies. For each policy, we first apply the content attribution to extract indicators and then apply the ranking on the returned indicators. Finally, we compare if the identities returned by the ranking match the domain owner in the policy GT using the same methodology in Section 6.2. If the GT identity is in the winner cluster together with an identity output by the content attribution, the policy was attributed (TP). If the GT identity is in the winner cluster without any identity output by the content attribution, the policy was not attributed (FN). Otherwise, if the GT identity is not in the winner cluster, the policy was incorrectly attributed to a third party (FP).

Table 4 presents the number of indicators and identities extracted by the content attribution, the number of total clusters output by the ranking, the number of winner clusters and the indicators and identities they contain, and finally the attribution accuracy metrics. On average, content attribution extracts 63.9 indicators from each privacy policy, grouped by the ranking into 2.1 clusters. Each winner cluster has on average 19 indicators. Overall, the ranking removes 70% of identities, as well as 70% of all indicators, corresponding to third parties. Prior to the ranking, 97% (2,255) identities come from the NER module. The ranking reduces the identities in the winner clusters that come from the NER to 86% (609). This shows that the NER introduces many third parties that, if not removed by the ranking, would lead to false attributions, highlighting the ranking importance. However, identities extracted by the NER are fundamental to attribute the 50% of policies without definition sentences.

Overall, 92 policies are correctly attributed, 5 are misattributed, and 3 are not attributed. We manually analyze the FPs and FNs. The root cause of the three FNs is the NER failing to identify the following entities: "Preisvergleich Internet Services AG" (German name in English policy), "LiveStreaming" (two common words joined), and "ZZB, LLC" (short). Three FPs are also due to NER FNs, causing the ranking to select a third party as the domain owner. The two other FPs are due to the text extraction step.

In summary, this experiment demonstrates the importance of the ranking to eliminate third parties introduced by the content attribution (most often by the NER), and thus limit false attributions. Furthermore, none of the attribution errors are due to the ranking, but rather to limitations in the NER and text extraction.

6.5 Updating Tracker Domain Lists

We apply WhoseDomain to the 3,001 tracker domains in the Disconnect list and compare the owners output by WhoseDomain with the ones in Disconnect to identify domains whose ownership should be updated. Overall, WhoseDomain outputs an identity for 77% of the tracker domains, a similar rate to the trackers GT dataset. In 60% of the attributed domains, WhoseDomain outputs the same owner in Disconnect, measured using the string similarity of the ranking module to account for minor differences like *Google Inc.* vs *Google*.

We evaluate the accuracy of WhoseDomain on this dataset by randomly sampling 100 tracker domains from the Disconnect list.

	Whois				+ Content Attribution				+ VirusTotal				+ Certificates			
Datasets	**TP**	**FP**	**FN**	**F1**	**TP**	**FP**	**FN**	**F1**	**TP**	**FP**	**FN**	**F1**	**TP**	**FP**	**FN**	**F1**
tranco_top	156	6	88	0.77	227	13	10	0.95	229	16	5	0.96	239	10	1	0.98
tranco_100k	71	5	174	0.44	210	15	25	0.91	212	21	17	0.92	212	26	12	0.92
brands	33	6	61	0.50	69	8	23	0.82	76	10	14	0.86	93	7	0	0.96
trackers	53	2	84	0.55	93	7	39	0.80	99	7	33	0.83	105	7	27	0.86
All	313	19	407	0.59	599	43	97	0.89	616	54	69	0.91	649	50	40	0.94

Table 3: Attribution accuracy increase as expansions are incrementally added from left (WHOIS only) to right (all expansions).

Ranking All			Ranking Winners			Accuracy		
Ind.	**Iden.**	**Clust.**	**Ind.**	**Iden.**	**Clust.**	**TP**	**FP**	**FN**
6,386	2,325	2,102	1,907	707	100	92	5	3

Table 4: Document ranking results on 100 GT privacy policies. Ind. are all indicators, Iden. identities, and Clust. clusters.

If the domain owner output by WhoseDomain matches the owner in Disconnect, we consider it a TP. This happens for 52 domains. If WhoseDomain did not attribute the domain, we consider it an FN since Disconnect has an identity assigned to each domain. This happens for 19 domains. However, we later show that this may underestimate the attribution rate of WhoseDomain, as Disconnect may contain dead domains that arguably should no longer be on the list. Finally, if the identities differ, we manually check them to determine which one is right. This happens for 29 domains.

For 16 disagreeing domains, WhoseDomain outputs the parent company while Disconnect contains a subsidiary, or vice versa. In these cases, we consider that both WhoseDomain and Disconnect are correct. Thus, we assign a TP to WhoseDomain and there is nothing to report to Disconnect since keeping a single owner per domain is a design decision. In contrast, other tracker domain lists include both parent and subsidiary companies for each tracker domain [1]. For another 7 disagreeing domains, we determine that WhoseDomain misattributed the domain. Thus, we assign an FP to WhoseDomain and there is nothing to report to Disconnect.

For 6 disagreeing domains, we determine that WhoseDomain correctly attributed the domain. Thus, we assign a TP to WhoseDomain and consider that the entity in Disconnect should be updated. One of these domains is futureads.com, a privacy-protected domain attributed to *Future Ads* in Disconnect and to *Propel Media LLC* by WhoseDomain. The LinkedIn page of *Future Ads* states that they are now *Propel Media LLC* [39]. Three domains are assigned to *VerizonMedia* in Disconnect (aolcloud.com, my.yahoo.com, huffingpost.com). Yahoo and AOL were sold by VerizonMedia on September 2021 and became Yahoo again [2]. WhoseDomain correctly attributes both domains to Yahoo Assets LLC. Similarly, WhoseDomain attributes HuffingPost.com to BuzzFeed, which acquired it in November 2020 [40]. In the remaining two domains, WhoseDomain's output matches the non-generic WHOIS identity.

In summary, on the 100 analyzed tracker domains, WhoseDomain achieves 74 TPs, 7 FPs, and 19 FNs, for a precision of 0.91, recall of 0.79, and an F1 score of 0.85, only slightly worse than the 0.86 reported in Table 2. Thus, WhoseDomain results are consistent on different sets of tracker domains. Note that, as shown in Section 6.2, tracker domains are the hardest to attribute as they might not have an associated Web server. Thus, these attribution rates should increase on other types of domains (e.g., popular, phishing targets).

We also explore the reasons behind the 19 FNs. Of those, 3 are real FNs. All others correspond to domains that no longer appear to be used for tracking. Among those, 4 domains are currently not registered or are available for sale (addlvr.com, adonnetwork.net, csm-secure.com, tmnetads.com). Since these 4 domains do not currently have an owner, they could arguably be considered WhoseDomain TNs instead of FNs. We believe that these domains should be removed from Disconnect as they no longer belong to Web tracking companies. The other 12 domains have a privacy-protected owner in WHOIS. Of those, 5 do not resolve, 2 are parked, and for the others, there is no website associated and no activity we can identify. These 12 domains could still belong to a tracking company, although in that case the tracking likely has moved to different domains. Still, it is safer to keep them in Disconnect, as they could resurrect as trackers. Among these, we observe cases where Disconnect assigns the domains to different owners, but our manual analysis determines the owner is the same. For example, adfunkyserver.com and batanganetwork.com are assigned to different entities in Disconnect, but we track both to *VIX*, which was recently acquired by Univision Communications [64]. Thus, they should be merged in Disconnect under the same entity.

These results highlight how entries in tracker domain lists can quickly become stale due to the dynamic ecosystem. Tracker domains often change ownership and they may be removed from operation and eventually become available to buy. Automated tools like WhoseDomain can be used to assist managers to keep their lists updated. In particular, WhoseDomain identified 6 domains that should be updated in Disconnect, as well as four unregistered domains that should be removed. We also observe instances of domains that should be merged under the same entity. We have contributed the identified updates to the Disconnect repository and they have been accepted and applied to the tracker list [12].

6.6 Attributing Unattributed Tracker Domains

This section evaluates the ability of WhoseDomain to attribute the 3,710 unattributed tracker domains provided to us by a large security vendor. We set a maximum of 150 iterations for the exploration of each domain. Of the 3,710 domains, WhoseDomain attributes 3,113 (84%) and it does not find an identity for 597 (16%). The accuracy evaluation in Section 6.2 measured a precision of 0.86 on the *tracker* domains dataset. Assuming a similar precision holds over these unattributed tracker domains, we can estimate that 2,677 previously unattributed tracker domains would be correctly

attributed and 436 would be falsely attributed. Of the 597 FNs, there are 82 (2.2% of all domains) for which the exploration reaches the maximum number of iterations. For the other 515, WhoseDomain explores all leads but fails to identify an identity.

Figure 4 at the appendix shows the attribution path WhoseDomain followed to attribute the *bs.ad-stir.com* domain to *UNITED, inc.* The path comprises 9 expansions and captures how WhoseDomain explored the webpage at *https://ad-stir.com/* identifying in its content the link to a privacy policy hosted at *mt.united.jp*. That new domain had a WHOIS entry identifying *UNITED, inc.* as the owner. This is an example of the attribution pivoting through a domain different than the input domain.

Attribution points. Among the 16% tracker domains WhoseDomain could not attribute, we observe that for 3% WhoseDomain discovered other attribution points i.e., social handles and other domains. For malicious domains, such attribution points could be leveraged by law enforcement for attribution by requesting registration data from social networks and domain registrars. For example, the exploration does not find an identity for *analytics.trovit.com*. But, WhoseDomain finds the Twitter account *Trovit* and the domain *lifullconnect.com*. This saves time for the analyst as *Lifull* acquired Trovit in 2014.

6.7 Impersonation Case Study

Our evaluation has so far focused on attributing benign domains that do not attempt to hide their identity (e.g., popular, brands) and gray domains that may prefer not to be attributed and thus avoid disclosing their identity (e.g., some tracker domains). A different class are malicious domains that not only hide their real identity but impersonate a third-party. To examine the impact of impersonation, we apply WhoseDomain for attributing apesorigami.net, a phishing domain submitted to PhishTank on September 13, 2023 [14]. This phishing domain impersonates app.1inch.io, a domain from the *1inch* decentralized finance (DeFi) crypocurrency exchange.

We provide apesorigami.net as input to WhoseDomain. WhoseDomain first queries WHOIS, but the input domain is privacy-protected. Then, it identifies a Web server listening on ports 80/tcp and 443/tcp. The HTTP server on 80/tcp redirects to the HTTPS server on 443/tcp. The HTTPS web server has a valid certificate issued by "Google Trust Service LLC" that does not contain any identity. The security.txt URL returns the 9-byte string "Not found". From the downloaded content, WhoseDomain extracts 3 indicators: an iOS app ID (1546049391), a Google Tag Manager ID (GTM-TFJV2F3), and identity "1inch" extracted from the copyright string. All 3 indicators belong to the impersonated entity and have been copied from the original website. Since an identity has been found, the exploration stops and outputs *1inch* as the owner's identity.

This example shows that in face of impersonation, WhoseDomain may output the impersonated identity instead of the identity of the impersonator. Since the phishing website in this case is simply a copy of the original site on a different domain, identifying the attacker's identity is not possible, even for a human analyst. Furthermore, we believe there is still value to identify the impersonated entity for providing brand protection services. For example, if WhoseDomain is applied to a large number of domains and reports all domains attributed to 1inch back to the exchange, the exchange will know that app.1inch.io belongs to them, but apesorigami.net does not and thus is a phishing website that should be taken down. We furher discuss impersonation in Section 7.

7 DISCUSSION

Ethical considerations. WhoseDomain does not collect private user information. It only examines data publicly available on the Internet and VirusTotal. One concern is that WhoseDomain may de-anonymize domain owners that prefer not to be de-anonymized. This likely happens because the domain owner publicly leaked some data that it did not mean to. We argue that our approach can be used by the domain owner to identify and fix such leaks.

Misleading information. A key challenge in attribution is misleading information, which can introduce false leads and may lead to incorrect attribution. We differentiate the cases of unreliable and planted false information. Unreliable information is intrinsic to the attribution process (e.g., third parties in the privacy policy). The ranking of WhoseDomain is designed to address naturally occurring unreliable information by selecting the correct identity among all observed identities. Our evaluation covers the natural occurrence of such cases for popular, less popular, brands, and tracker domains.

Our case study on a phishing domain shows that if planted false information dominates the content, WhoseDomain may output the planted identity as the owner. Impersonating another entity is illegal in most jurisdictions and thus often avoided by companies behind attribution-worthy domains such as trackers [70], download portals [69], and commercial PPI services [48]. When impersonation happens (e.g., in phishing websites), we argue that there is still value to identify the impersonated entity. For example, companies know the websites they own and could search for other websites that WhoseDomain attributes to them in order to identify impersonators.

In attribution, an investigator needs to follow all leads (even false ones) because a priori it is not known which leads may be useful, unfruitful, or planted. WhoseDomain assists an investigator by automating repetitive manual tasks that are error-prone and can take long time. In the presence of misleading information, WhoseDomain would still save time for the analyst by investigating all leads automatically, even if false. The analyst would then examine the results to determine whether false leads were planted.

Attribution points. When WhoseDomain fails to identify the owner identity for a domain or website, it may still discover useful attribution points like other domains that belong to the same owners and certificates issued by commercial certificate authorities (CAs). Law enforcement can then continue the attribution process by requesting the registration information from the domain registrars and commercial CAs. The key advantage is that WhoseDomain automatically finds those attribution points, saving investigation time.

Other applications. WhoseDomain is best suited for attributing gray domains that belong to companies, which may not be able to completely hide their identity. Beyond trackers, WhoseDomain could be applied to scam websites (e.g., [45, 53, 79]), download portals [69], abusive affiliate programs [28, 61], and pay-per-install (PPI) commercial services [48]. Prior work by Starov et al. [75] uses

advertisement identifiers to cluster phishing webpages from the same owner. WhoseDomain could also be applied to that scenario by adding advertisement identifiers to the supported IOCs. The combination of such identifiers with the other sources WhoseDomain already supports could potentially attribute higher numbers of sites. WhoseDomain can also be applied to determine the company that owns a benign website. For example, when a Web server is identified to have a vulnerability, WhoseDomain can be used to identify the owner of the Web server and extract its contact indicators. On the other hand, WhoseDomain may not be well suited for attributing other malicious Web servers that hardly provide any info such as exploit servers used to deliver malware to their visitors [35] and C&C servers [33].

Blocklists. WhoseDomain uses a variety of blocklists. While building such blocklists requires significant work, it is largely a one-time effort. Furthermore, the design of WhoseDomain assumes that filtering is incomplete by nature and includes a novel ranking technique to address third-party identities introduced due to filtering limitations.

Related identities. The ranking currently only outputs the top-ranked cluster. It is possible that there exist multiple dissimilar, but related, identities, e.g., parent-child relationships due to company acquisitions. If so, WhoseDomain will only output one company. One possibility for future work would be to leverage company databases (e.g., [31]) to check if multiple identities in the ranking are related.

Other expansions. Adding more expansions could increase the attribution rate. For example, historical WHOIS information may help to identify the domain owner even if a domain currently uses a privacy protection service. And, advertisement identifiers could be used to identify other webpages belonging to the same owner [75].

8 RELATED WORK

Rid and Buchanan [68] define attribution as answering the question "Who did it?", which in this paper we convert into "Who owns it?". Automating the attribution of domains and websites is a little-studied problem. Sanchez et al. [70] take a first step by using domain and IP WHOIS to identify the owner of third-party tracker domains. However, WHOIS usefulness is hampered by the problems of privacy protection services and data redaction (e.g., to satisfy privacy regulations). Attribution as a resource ownership problem has also been recently addressed for IP addresses [77] and Autonomous Systems (ASes) [82]. WhoseDomain currently does not address the attribution of IP addresses, and their ASes, because due to IPv4 scarcity most websites use hosting services instead of their own IP ranges. In hosting services, IP addresses may be shared or reused by different websites over time, which can lead to incorrect attributions. Starov et al. [75] leverage advertisement identifiers to link websites belonging to the same owners, while Retriever [72] links developer accounts in mobile markets from the same owners. In contrast, WhoseDomain goes beyond linking resources of the same owner to automatically identify the identity of the owner. Attribution frameworks such as Maltego [59] and VT Graph [78] produce attribution graphs but are designed for manual analysis.

Other works propose infrastructure and content features to identify malicious domains and websites (e.g., [24, 41]). But, those features are not designed for attribution. For example, a common feature is whether the Whois entry is anonymized. There is also work on the impact of domain ownership changes [51]. Handling such temporal ownership changes would require adding other datasets to WhoseDomain like historical Whois registrations. Also related is the work by Squarcina et al. [74] that analyzes security issues introduced by subdomain leasing. Our work shows that subdomain leasing can also affect attribution.

WHOIS.. Early WHOIS works analyzed the accuracy of its data [21], the abuse of privacy protection and proxy services [29], and the data format inconsistency [57]. Recently, Lu et al. quantified the impact of GDPR on WHOIS [58], finding that 85% of large WHOIS providers redact European Economic Area (EEA) records at scale, and over 60% also redact non-EEA records. Our work proposes a novel attribution approach that can attribute domains and websites even when WHOIS data is not useful.

Tracking. Web tracking can be traced back at least 25 years [49] and has become widespread with over 90% of websites including at least one tracking script [34, 71]. Tracker companies may own multiple tracker domains [32, 70, 73] and may collaborate to increase their coverage [38, 65, 70]. Manually-curated lists associate tracker domains with tracking companies [3, 9, 30], but they can quickly become obsolete due to company acquisitions. Our novel attribution approach allows building such lists in an automated manner and keeping them up-to-date with the dynamics of the tracker ecosystem.

Indicator extraction. Other research extracts indicators of compromise (IOCs) from security articles [42, 54, 81]. Our approach is similar in using regular expressions and NLP to identify indicators but differs in the goal of attributing domains and websites.

9 CONCLUSION

We have presented a novel automated approach for domain and website attribution. When WHOIS data is not useful, our approach leverages other sources such as passive DNS, TLS certificates, and the analysis of website content. We have proposed a novel ranking technique to select the domain owner and its indicators among multiple entities. We have implemented our approach into WhoseDomain [16], which achieves an F1 score of 0.94 compared to 0.59 for WHOIS. We have applied WhoseDomain on 3,001 tracker domains in the Disconnect list showing that WhoseDomain can identify needed updates to the list.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments and VirusTotal for providing academic access to their API. This work was partially funded by the Spanish Government MCIN/AEI/10.13039/501100011033/ through grants TED2021-132464B-I00 (PRODIGY), PID2022-142290OB-I00 (ESPADA), and FPU18/06416. This work was also partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101019206 (TESTABLE). The above grants are co-funded by European Union ESF, EIE, and NextGeneration funds.

REFERENCES

[1] 2020. Xray2020 blacklist. https://github.com/TrackerControl/tracker-control-android/blob/master/app/src/main/assets/xray-blacklist.json.

[2] 2021. Yahoo is Yahoo once more after new owners complete acquisition. https://www.theverge.com/2021/9/2/22653652/yahoo-aol-acquired-apollo-global-management-private-equity.

[3] 2022. Disconnect, Inc. disconnect.me.

[4] 2022. Hosting provider catalogue. https://hostings.info/catalog.

[5] 2022. Natural Language Toolkit. https://www.nltk.org/.

[6] 2022. Schema.org. https://schema.org/.

[7] 2022. Stanford CoreNLP NER Model. https://stanfordnlp.github.io/CoreNLP/index.html.

[8] 2022. VirusTotal. https://www.virustotal.com/.

[9] 2022. webXray. https://github.com/timlib/webXray.

[10] 2022. whois: A Python package for retrieving WHOIS information of domains. https://github.com/DannyCork/python-whois.

[11] 2022. Zoxh. https://zoxh.com/.

[12] 2023. disconnectme issue #330 : Recategorizing the Entity which they belong to. https://github.com/disconnectme/disconnect-tracking-protection/issues/330.

[13] 2023. List of DynDNS Pro (Dynamic DNS) Domain Names. https://help.dyn.com/list-of-dyn-dns-pro-remote-access-domain-names/.

[14] 2023. PhishTank Submission 8296476. https://phishtank.org/phish_detail.php?phish_id=8296476.

[15] 2023. Tranco popular domain list. https://tranco-list.eu/list/K2XYW.

[16] 2023. WhoseDomain. https://hub.docker.com/r/dianecode/whosedomain.

[17] Sahar Abdelnabi, Katharina Krombholz, and Mario Fritz. 2019. WhiteNet: Phishing Website Detection by Visual Whitelists. *CoRR* abs/1909.00300 (2019).

[18] ammarshah. 2022. Email provider list. https://gist.github.com/ammarshah/f5c2624d767f91a7cbdc4e54db8dd0bf.

[19] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play. In *USENIX Security Symposium*.

[20] International Trademark Association. 2020. WHOIS Challenges: A Toolkit for Intellectual Property Professionals. https://www.inta.org/wp-content/uploads/public-files/advocacy/committee-reports/WHOIS-Challenges-A-Toolkit-for-Intellectual-Property-Professionals-3.20.20.pdf.

[21] NORC at the University of Chicago. 2010. Draft Report for the Study of the Accuracy of WHOIS Registrant Contact Information. hhttps://www.icann.org/en/resources/compliance/reports/whois-accuracy-study-17jan10-en.pdf.

[22] Ulrich Bayer, Paolo Milani Comparetti, Clemens Hlauschek, Christopher Kruegel, and Engin Kirda. 2009. Scalable, Behavior-Based Malware Clustering. In *Network and Distributed System Security*.

[23] Robert Biddle, Paul C Van Oorschot, Andrew S Patrick, Jennifer Sobey, and Tara Whalen. 2009. Browser interfaces and extended validation SSL certificates: an empirical study. In *ACM Workshop on Cloud Computing Security*.

[24] Leyla Bilge, Sevil Sen, Davide Balzarotti, Engin Kirda, and Christopher Kruegel. 2014. EXPOSURE: a Passive DNS Analysis Service to Detect and Report Malicious Domains. *ACM Transactions on Information and System Security* 16, 4 (2014), 1–28.

[25] Reuben Binns, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2018. Measuring Third-party Tracker Power across Web and Mobile. *ACM Transactions on Internet Technology* 18, 4 (2018), 1–22.

[26] Juan Caballero, Gibran Gomez, Srdjan Matic, Gustavo Sánchez, Silvia Sebastián, and Arturo Villacañas. 2023. The Rise of GoodFATR: A Novel Accuracy Comparison Methodology for Indicator Extraction Tools. *Future Generation Computer Systems* 144 (July 2023), 74–89. https://doi.org/10.1016/j.future.2023.02.012

[27] Orcun Cetin, Carlos Ganan, Maciej Korczynski, and Michel Van Eeten. 2017. Make notifications great again: learning how to notify in the age of large-scale vulnerability scanning. In *Workshop on the Economics of Information Security*.

[28] Neha Chachra, Stefan Savage, and Geoffrey M Voelker. 2015. Affiliate Crookies: Characterizing Affiliate Marketing Abuse. In *Internet Measurement Conference*.

[29] Richard Clayton and Tony Mansfield. 2014. A Study of Whois Privacy and Proxy Service Abuse. In *Workshop on the Economics of Information Security*.

[30] Cliqz GmbH. 2019. WhoTracks.me: Bringing Transparency to Online Tracking. https://github.com/cliqz-oss/whotracks.me.

[31] crunchbase 2022. Crunchbase. https://www.crunchbase.com/.

[32] Savino Dambra, Iskander Sanchez-Rola, Leyla Bilge, and Davide Balzarotti. 2022. When Sally Met Trackers: Web Tracking From the Users' Perspective. In *USENIX Security Symposium*.

[33] David Dittrich and Sven Dietrich. 2007. Command and control structures in malware. *Usenix magazine* 32, 6 (2007).

[34] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *ACM SIGSAC Conference on Computer and Communications Security*.

[35] Chris Grier et al. 2012. Manufacturing Compromise: The Emergence of Exploit-as-a-Service. In *ACM Conference on Computer and Communication Security*.

[36] Durumeric et al. 2014. The Matter of Heartbleed. In *Internet Measurement Conference*.

[37] Europol. 2015. Use of WHOIS for cyber investigations. https://gac.icann.org/briefing-materials/public/gregory-mounier-ec3-lea-use-case-examples-of-whois-icann-54-publish-2015-10-19.pdf.

[38] Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, and Richard Mortier. 2016. Tracking personal identifiers across the web. In *International Conference on Passive and Active Network Measurement*.

[39] fuelads 2022. Future Ads LLC (now Propel Media). https://www.linkedin.com/company/future-ads-llc/about/.

[40] Anthony Ha. 2020. BuzzFeed acquires HuffPost. https://techcrunch.com/2020/11/19/buzzfeed-acquires-huffpost/.

[41] Austin Hounsel, Jordan Holland, Ben Kaiser, Kevin Borgolte, Nick Feamster, and Jonathan Mayer. 2020. Identifying Disinformation Websites Using Infrastructure Features. In *USENIX Workshop on Free and Open Communications on the Internet*.

[42] Ghaith Husari, Ehab Al-Shaer, Mohiuddin Ahmed, Bill Chu, and Xi Niu. 2017. TTPDrill: Automatic and Accurate Extraction of Threat Actionsfrom Unstructured Text of CTI Sources. In *Annual Computer Security Applications Conference*.

[43] Collin Jackson, Daniel R Simon, Desney S Tan, and Adam Barth. 2007. An Evaluation of Extended Validation and Picture-in-Picture Phishing Attacks. In *International Conference on Financial Cryptography and Data Security*.

[44] Ishan Karunanayake, Nadeem Ahmed, Robert Malaney, Rafiqul Islam, and Sanjay K. Jha. 2021. De-Anonymisation Attacks on Tor: A Survey. *IEEE Communications Surveys and Tutorials* 23, 4 (2021), 2324–2350. https://doi.org/10.1109/COMST.2021.3093615

[45] Amin Kharraz, William Robertson, and Engin Kirda. 2018. Surveylance: Automatically Detecting Online Survey Scams. In *IEEE Symposium on Security and Privacy*.

[46] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate Detection Using Shallow Text Features. In *ACM International Conference on Web Search and Data Mining*.

[47] Konrad Kollnig, Anastasia Shuba, Reuben Binns, Max Van Kleek, and Nigel Shadbolt. 2022. Are iPhones Really Better for Privacy? Comparative Study of iOS and Android Apps. *Proceedings on Privacy Enhancing Technologies* 2022, 2 (2022), 6–24. https://doi.org/doi:10.2478/popets-2022-0033

[48] Platon Kotzias, Leyla Bilge, and Juan Caballero. 2016. Measuring PUP Prevalence and PUP Distribution through Pay-Per-Install Services. In *USENIX Security Symposium*.

[49] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. 2016. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *USENIX Security Symposium*.

[50] letsencrypt 2022. Let's Encrypt. https://letsencrypt.org/.

[51] Chaz Lever, Robert Walls, Yacin Nadji, David Dagon, Patrick McDaniel, and Manos Antonakakis. 2016. Domain-Z: 28 Registrations Later. Measuring the Exploitation of Residual Trust in Domains. In *IEEE Symposium on Security and Privacy*.

[52] Frank Li, Zakir Durumeric, Jakub Czyz, Mohammad Karami, Michael Bailey, Damon McCoy, Stefan Savage, and Vern Paxson. 2016. You've Got Vulnerability: Exploring Effective Vulnerability Notifications. In *USENIX Security Symposium*.

[53] Xigao Li, Anurag Yepuri, and Nick Nikiforakis. 2023. Double and Nothing: Understanding and Detecting Cryptocurrency Giveaway Scams. In *Network and Distributed Systems Security Symposium*.

[54] Xiaojing Liao, Kan Yuan, XiaoFeng Wang, Zhou Li, Luyi Xing, and Raheem Beyah. 2016. Acing the IOC Game: Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence. In *ACM SIGSAC Conference on Computer and Communications Security*.

[55] Yun Lin, Ruofan Liu, Dinil Mon Divakaran, Jun Yang Ng, Qing Zhou Chan, Yiwen Lu, Yuxuan Si, Fan Zhang, and Jin Song Dong. 2021. Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages. In *USENIX Security Symposium*.

[56] Ruofan Liu, Yun Lin, Xianglin Yang, Siang Hwee Ng, Dinil Mon Divakaran, and Jin Song Dong. 2022. Inferring Phishing Intention via Webpage Appearance and Dynamics: A Deep Vision Based Approach. In *USENIX Security Symposium*.

[57] Suqi Liu, Ian Foster, Stefan Savage, Geoffrey M Voelker, and Lawrence K Saul. 2015. Who is. com? Learning to parse WHOIS records. In *Internet Measurement Conference*.

[58] Chaoyi Lu, Baojun Liu, Yiming Zhang, Zhou Li, Fenglu Zhang, Haixin Duan, Ying Liu, Joann Qiongna Chen, Jinjin Liang, Zaifeng Zhang, et al. 2021. From WHOIS to WHOWAS: A Large-Scale Measurement Study of Domain Registration Privacy under the GDPR. In *Network and Distributed System Security Symposium–NDSS*.

[59] Maltego 2022. https://www.maltego.com/.

[60] Srdjan Matic, Platon Kotzias, and Juan Caballero. 2015. CARONTE: Detecting Location Leaks for Deanonymizing Tor Hidden Services. In *ACM Conference on Computer and Communication Security*.

[61] Damon McCoy, Andreas Pitsillidis, Grant Jordan, Nicholas Weaver, Christian Kreibich, Brian Krebs, Geoffrey M Voelker, Stefan Savage, and Kirill Levchenko. 2012. PharmaLeaks: Understanding the Business of Online Pharmaceutical Affiliate Programs. In *USENIX Security Symposium*.

[62] Rami M Mohammad, Fadi Thabtah, and Lee McCluskey. 2012. An Assessment of Features Related to Phishing Websites using an Automated Technique. In *IEEE International Conference for Internet Technology and Secured Transactions.*
[63] Mozilla. 2022. Public Suffix List. https://publicsuffix.org/.
[64] Ben Munson. 2021. Univision acquires Vix ahead of PrendeTV launch. https://www.fiercevideo.com/video/univision-acquires-vix-ahead-prendetv-launch.
[65] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. 2019. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *The World Wide Web Conference.*
[66] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened against Manipulation. In *Network and Distributed Systems Security.*
[67] Tara Poteat and Frank Li. 2021. Who you Gonna Call? An Empirical Evaluation of Website security.txt Deployment. In *ACM Internet Measurement Conference.*
[68] Thomas Rid and Ben Buchanan. 2015. Attributing cyber attacks. *Journal of Strategic Studies* 38, 1-2 (2015), 4–37.
[69] Richard Rivera, Platon Kotzias, Avinash Sudhodanan, and Juan Caballero. 2019. Costly Freeware: A Systematic Analysis of Abuse in Download Portals. *IET Information Security* 13, 1 (January 2019), 27–35.
[70] Iskander Sanchez-Rola, Matteo Dell'Amico, Davide Balzarotti, Pierre-Antoine Vervier, and Leyla Bilge. 2021. Journey to the center of the cookie ecosystem: Unraveling actors' roles and relationships. In *IEEE Symposium on Security and Privacy.*
[71] Iskander Sanchez-Rola and Igor Santos. 2018. Knockin' on trackers' door: Large-scale automatic analysis of web tracking. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment.*
[72] Silvia Sebastián and Juan Caballero. 2020. Towards Attribution in Mobile Markets: Identifying Developer Account Polymorphism. In *ACM Conference on Computer and Communication Security.*
[73] Konstantinos Solomos, Panagiotis Ilia, Sotiris Ioannidis, and Nicolas Kourtellis. 2020. Clash of the Trackers: Measuring the Evolution of the Online Tracking Ecosystem. In *Network Traffic Measurement and Analysis Conference.*
[74] Marco Squarcina, Mauro Tempesta, Lorenzo Veronese, Stefano Calzavara, and Matteo Maffei. 2021. Can I Take Your Subdomain? Exploring Same-Site Attacks in the Modern Web. In *USENIX Security Symposium.*
[75] Oleksii Starov, Yuchen Zhou, Xiao Zhang, Najmeh Miramirkhani, and Nick Nikiforakis. 2018. Betrayed by Your Dashboard: Discovering Malicious Campaigns via Web Analytics. In *World Wide Web Conference.*
[76] Ben Stock, Giancarlo Pellegrino, Christian Rossow, Martin Johns, and Michael Backes. 2016. Hey, You Have a Problem: On the Feasibility of Large-Scale Web Vulnerability Notification. In *USENIX Security Symposium.*
[77] Florian Streibelt, Martina Lindorfer, Seda Gürses, Carlos H Gañán, and Tobias Fiebig. 2023. Back-to-the-Future Whois: An IP Address Attribution Service for Working with Historic Datasets. In *International Conference on Passive and Active Network Measurement.*
[78] vtgraph 2022. VirusTotal Graph overview. https://support.virustotal.com/hc/en-us/articles/360004679937-VirusTotal-Graph-overview.
[79] Pengcheng Xia, Haoyu Wang, Bowen Zhang, Ru Ji, Bingyu Gao, Lei Wu, Xiapu Luo, and Guoai Xu. 2020. Characterizing Cryptocurrency Exchange Scams. *Computers & Security* 98 (2020).
[80] Yue Zhang, Jason I Hong, and Lorrie F Cranor. 2007. CANTINA: A Content-Based Approach to Detecting Phishing Web Sites. In *International Conference on World Wide Web.*
[81] Ziyun Zhu and Tudor Dumitras. 2018. ChainSmith: Automatically Learning the Semantics of Malicious Campaigns by Mining Threat Intelligence Reports. In *IEEE European Symposium on Security and Privacy.*
[82] Maya Ziv, Liz Izhikevich, Kimberly Ruth, Katherine Izhikevich, and Zakir Durumeric. 2021. ASdb: A System for Classifying Owners of Autonomous Systems. In *ACM Internet Measurement Conference.*

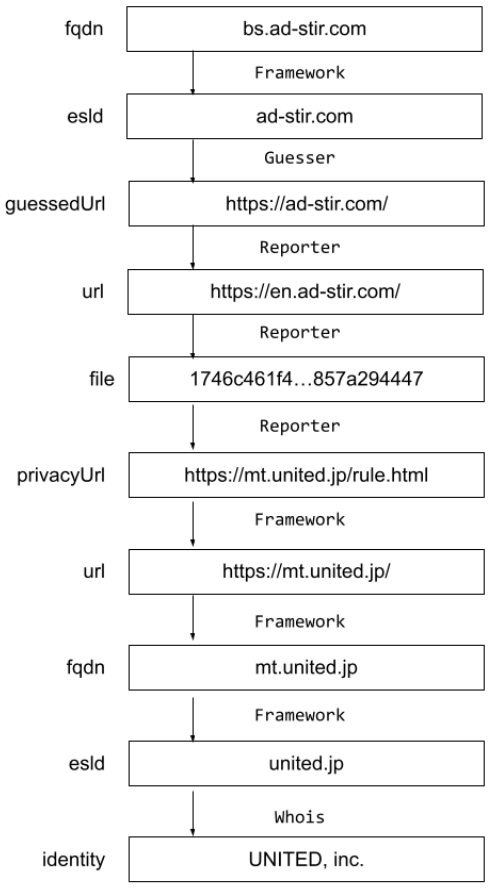

Figure 4: Attribution path followed for a tracker domain.

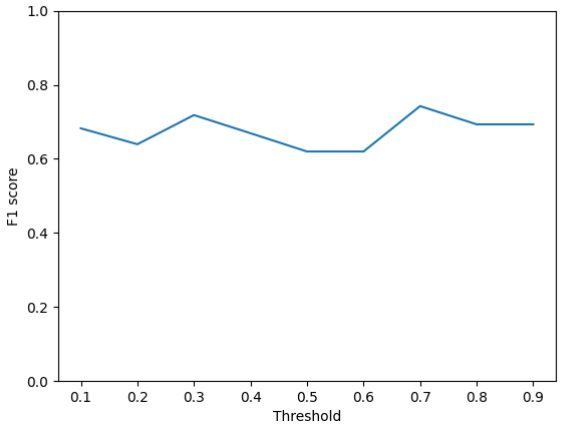

Figure 5: F1 score for different LCS thresholds in the ranking module. The best F1 score is achieved by threshold 0.7.

A LCS THRESHOLD SELECTION

The clustering step in the ranking module groups indicators with similar values (e.g., countryLegends971 and Legends971). To determine if two indicator values are similar, it computes the longest contiguous matching subsequence (LCS) for their values and compares the LCS value (in the range [0,1]) with a threshold. Strings with LCS larger than the threshold are considered similar.

To select the optimal threshold value we use a dataset of 200 strings corresponding to identities the NER extracted from 50 privacy policies that do not belong to any dataset used in Section 6. We produced a reference clustering by manually grouping the 200 strings such that similar strings are in the same cluster and dissimilar strings are in different clusters.

	Regexp	NLP-NER	NLP-Synonym	DP-Metadata	DP-Schema	DP-Link
Before Ranking	1,069 (71)	2,377 (2,255)	110 (110)	117 (10)	87 (17)	2,494 (0)
After Ranking	326 (65)	625 (609)	104 (104)	83 (7)	47 (12)	1,138 (0)

Table 5: Number of indicators (identity indicators in parenthesis) extracted by the different content attribution techniques on the dataset of 100 GT privacy policies.

	GT	Content	Regexp	NLP-NER	NLP-Synonym	DP-Metadata	DP-Schema	DP-Link
tranco_top	250	42	26	8	0	25	10	41
tranco_100k	250	151	118	14	1	54	43	134
brands	100	80	74	19	3	24	10	73
trackers	139	82	69	40	11	29	24	73
All	739	355	287	81	15	132	87	321

Table 6: Number of indicators extracted by the different content attribution techniques that end up in the final attribution results for each GT dataset.

We run the clustering step of the ranking module on the 200 strings multiple times, increasing the LCS threshold in intervals of 0.1. In this experiment, the clustering only uses the LCS feature, other features (e.g., if the two strings share the same prefix) are disabled. To determine the clustering accuracy at each threshold, we compare the clustering results with the manually generated reference clustering. For this, we use precision, recall, and F1 score, common metrics for evaluating malware clustering results [22]. These metrics do not require or use cluster labels. They measure structural similarity between the obtained clusters and the reference clusters. The results are shown in Figure 5. The F1 score is maximized with threshold 0.7, which achieves 1.00 precision, 0.83 recall, and 0.74 F1 score. We use 0.7 as threshold in our evaluation.

B SUBDOMAIN THRESHOLD SELECTION

The passive DNS expansion in the infrastructure attribution module ignores the subdomains of a given *esld* if they are larger than a threshold to avoid expanding domains that lease subdomains to third parties. To select the threshold value, we sample 200 domains belonging to the DynDNS Pro dynamic DNS provider [13] that leases subdomains on those domains to third parties. We query VT to obtain their number of subdomains. The lowest number of subdomains VT reports is 52. To account for some variability, we conservatively select 50 as threshold value. To check the impact of this threshold, we randomly select 200 domains among the top 100K domains in the Tranco list. As far we know, none of these 200 domains lease subdomains. We query VT for their subdomains, observing that 70 have more than 50 subdomains. Thus, while our selected threshold of 50 does not miss leasing domains, it may not expand some non-leasing domains with many subdomains, which could potentially introduce false negatives. However, other expansions may still attribute those domains, as illustrated by the 0.94 F1 score WhoseDomain achieves across the GT datasets. In future work, we would like to explore if this simple threshold-based filter could be replaced with a machine learning classifier.

C CONTENT ATTRIBUTION TECHNIQUES

We evaluate how each content attribution technique contributes to the attribution of a document and to the final domain attribution.

Document attribution. We first evaluate how much each content attribution technique contributes towards attributing a document. We focus on privacy policies as those are the most prevalent documents. In particular, we apply the content attribution module on the dataset of 100 privacy policies used for the ranking evaluation in Section 6.4. Each indicator has an attribute with the list of sources (module and technique) that extracted it. Table 5 shows the number of indicators in the content attribution results, before and after the ranking, that contain a technique among their sources. Numbers in parenthesis correspond to *identity* indicators.

The technique that extracts most indicators is the document parsing link analysis (DP-Link), followed by the NER, and the regular expressions. Indicators extracted from links are largely domains and social network handles; this technique does not extract any identity indicators. The smallest contributor is the parsing of the Schema.org information (DP-Schema) because this data is only available in a small fraction of documents. The results also show that many identities extracted by the NER are false positives that the ranking drops.

Domain attribution. Next, we evaluate how much each content attribution technique contributes to the final domain attribution results. For each technique, we count the number of indicators in the domain attribution results of the 4 GT datasets that contain the technique among their sources. Table 6 first shows the number of domains attributed (GT) and the number of domains where the content attribution module was used (Content), which is smaller since some domains may not have websites or may be attributed before content is downloaded (e.g., by WHOIS). Then, it shows the number of indicators in the results contributed by each technique.

All techniques contribute indicators to the final domain attribution results, highlighting their usefulness. The technique that contributes most indicators is the document parsing link analysis, followed by regular expressions, and the document parsing metadata analysis (DP-Metadata). The latter extracts indicators from the non-visible content, e.g., HTML meta tags. Compared to the document attribution, the NER importance decreases as the final ranking further removes false positives.