Identifying Subdomain Doppelgänger Attacks against Companies

Geoffrey Simpson
Tandy School of Computer Science
The University of Tulsa
geoffrey@utulsa.edu

Tyler Moore
School of Cyber Studies and
Tandy School of Computer Science
The University of Tulsa
tyler-moore@utulsa.edu

Abstract

Cybercriminals regularly impersonate organizations when carrying out attacks. This paper investigates a tactic that has not been studied previously. In so-called doppelgänger attacks, miscreants register domains similar to legitimate subdomains used by organizations. Investigation of domain registration data from 2009–2022 uncovers $84,952$ *1st-party doppelgänger attacks that mimic valid subdomains of organization websites, plus a further* $5,448$ *3rd-party doppelgängers in which service providers used by organizations are impersonated. By analyzing patterns of the gathered data, the paper studies how victims are affected and attackers organize their activities. It is hoped that by raising awareness to this attack technique, future malicious activities may be curtailed.*

1. Introduction

Organizations today face a myriad of cybercrime threats, from phishing to ransomware to business email compromise. In 2022 alone, the FBI's Internet Crime Complaint Center (IC3) fielded over 800,000 complaints totaling $10.7 Billion U.S. Federal Bureau of Investigation, 2022, likely a significant undercount.

Often, the method of attack involves impersonating the organization or its associated IT infrastructure. In this paper, we study two particular forms of impersonation in which attackers register lookalike web domains that closely resemble legitimate names already in use by the organization. For example, in November 2019 Russian intelligence officers launched a phishing campaign against the Ukrainian gas company Burisma (Area 1, 2020). They impersonated the Microsoft Sharepoint site of one of Burisma's subsidiaries, CUB Energy. The hackers registered the fake domain `cubenergy-my-sharepoint.com`, which closely resembled the legitimate site `cubenergy.my-sharepoint.com` operated by Microsoft.

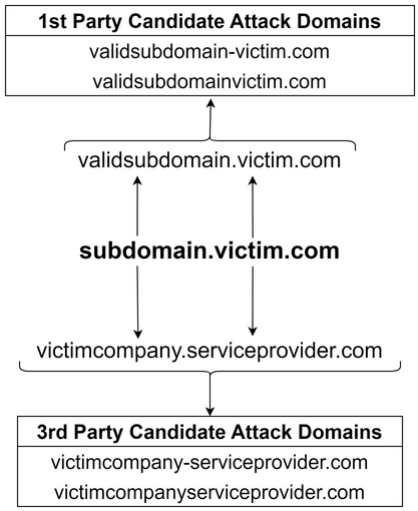

Figure 1: Method to construct candidate doppelgänger domains.

While cybercriminals have used such tactics for years, no prior work has systematically investigated such attacks targeting organizations at scale. In this paper, we investigate two classes of impersonation involving subdomains associated with organizations. Because these techniques have not been studied previously, we introduce a new term, *subdomain doppelgänger*, to describe the attack.

Subdomain doppelgaängers come in two forms. A *1st-party doppelgänger* is a domain that combines a legitimate organization's domain name and one of its own legitimate subdomain names, with or without the

conjunction character −. For example, the legitimate domain mail.victim.com can be impersonated as mail-victim.com. A *3rd-party doppelgänger* is a domain that combines a legitimate organization's name as a subdomain with a legitimate 3rd-party provider's name. The Burisma example above illustrates a 3rd-party doppelganger, where cubenergy is the victim name and my-sharepoint.com is the impersonated provider. The process for composing both types of doppelgängers is shown in Figure 1.

We now briefly summarize the paper's key contributions. First, we describe a method for identifying 1st- and 3rd-party doppelgängers. We construct very large historical datasets of domain name registrations and subdomain activity and locate historical registration of thousands of doppelgänger domains impersonating organizations between 2009 and 2022. We analyze the gathered data to document the attack prevalence and shed light on how attackers operate and how victims are affected.

2. Related work

The present work is motivated by the security economics literature (Anderson and Moore, 2006). This community has shed much light on how attackers operate (Moore et al., 2009; Levchenko et al., 2011; Liu et al., 2015, August) and the effectiveness of defenses (Liu et al., 2011; Metcalf and Spring, 2013).

Researchers have investigated many strategies for maliciously registering domain names to impersonate others. Table 1 illustrates attack categories for williams.com, with references to relevant work.

Typosquatting has been around for decades (Wang et al., 2006). Here, attackers register domains that result from input errors on the keyboard, such as swapping characters or "fat-fingering" nearby keys. Other studies have focused on tricking other senses. For example, Simpson et al., 2020 studies domains that impersonate other company domains by registering visually similar names (e.g., substituting the letter 'l' for the numeral 1). Soundsquatting attacks target common transcribing errors that occur when end users utilize text-to-speech software (Nikiforakis et al., 2014). Levelsquating attacks embed a legitimate-looking URL within the subdomain of a different domain Du et al., 2019. Meanwhile, combo-squatting attacks add ransom plausible words before or after the impersonated domain (Kintis et al., 2017; Tian et al., 2018). The final type is domain-squatting, in which different top-level domains are registered for a matching name (Pouryousef et al., 2020). This particular threat rose to the fore when available top-level domains expanded.

The present work on doppelgängers differs from prior efforts in a few ways. First, we focus on impersonating business domains exclusively. Second, we are the first to study attacks that target subdomains in legitimate use by those organizations. Third, we investigate attacks over a longer period than any prior work has attempted.

3. Methodology for identifying doppelgängers

Consistent with most cybercrime measurement research, this paper employs an observational study method to investigate doppelgänger attacks. We discuss the data sources utilized in Section 3.1. We then describe how to identify 1st-party doppelgängers in Section 3.2 and 3rd-party doppelgängers in Section 3.3.

3.1. Data sources

We now describe three data sources for identifying attacks: legitimate company names, historical registration data, and subdomain utilization data.

Legitimate company names We elected to focus on companies as the targets of visual impersonation attacks because we know that business email compromise (BEC) attacks often employ look-alike email domains.

We use the Bureau van Dijk Orbis database, which holds data on over 375 million companies worldwide (Bureau van Dijk, 2023). We selected all active US-based companies with at least 35 employees (approximately 381K firms), as well as non-US companies with at least 350 employees (approximately 184K firms). In total, this gave us 565,269 records. These records provide the company name, website, NAICS Codes, and a unique identifier.

We extracted the second-level host name from the website URL. That is, from www.example.com/index.htm we selected example.com. We excluded any non-dedicated domains (e.g., companies reporting a Facebook page) and top-level domains other than .com, resulting in a list of 269 759 company domain names.

Some company names include common words (e.g., mail.com) or are very short (e.g., aa.com). In these cases, many candidate doppelgängers could be false positives. Hence, we take three approaches to filter our results. The first is comparing the list of domain names to the top 5,000 most commonly used English words according to the Corpus of Contemporary American English (COCA) (Davies, 2023) and removing those that are in the list. Secondly, we create a manual

Table 1: Domain name impersonation examples for target williams.com.

Category	Sample	Reference
Typosquatting	wililams.com	Wang et al., 2006; Moore and Edelman, 2010; Szurdi et al., 2014
Visual impersonation	williarns.com	Szurdi et al., 2014,Simpson et al., 2020
Soundsquatting	willyams.com	Nikiforakis et al., 2014
Levelsquatting	www.williams.com.anotherdomain.com	Du et al., 2019; Quinkert et al., 2021
Combo-Squatting	williams-login.com	Kintis et al., 2017; Tian et al., 2018
Domain-Squatting	williams.new	Pouryousef et al., 2020
1st-Party Doppelgänger	mail-williams.com	Present work
3rd-Party Doppelgänger	williams-sharepoint.com	Present work

filter list composed of common business phrases and geographic location names that are not contained in the top 5,000 English word list. Finally, we filter any names less than 3 characters in length.[1] This yielded a final list of 238,761 company .com domain names for analysis.

Historical registration data We obtained a dataset of .com zone file data from the Cambridge Cybercrime Centre, which provides a daily record of all domain name registrations and changes of name server from September 6, 2009 to September 16, 2022. Each record contains a domain name, name server name, and the start and finish dates that this entry was present in the zone file. Hence each domain can have many records, showing when it was registered (or re-registered), when it changed from one name server to another, and, by deduction, when it expired altogether. The entire data set comprises 2,972,097,397 records spanning 388,361,125 unique .com domain names. The availability of historical .com registration data is another reason we focused the investigation on for-profit companies.

Subdomain utilization data A second-level domain owner is not required to publish externally accessible subdomains in any central registry like a zone file. Hence, we sought out other sources of historical subdomain utilization data. Fortunately, Rapid7's Project Sonar makes available to researchers a DNS enumeration dataset called Forward DNS. It is comprised of DNS responses to the A, ANY, AAAA, CNAME, and TXT records. We used all available IPv4 data for the time-frames studied to maximize the potential for identifying the most subdomain names. To best represent longitudinal changes, we isolated our data reported in June for the five years spanning 2018 through 2022. In total, the selected data comprised 1,237 files, totaling approximately 100 Tb. We reduce that to a more manageable size by extracting only fully qualified domain names of interest.

[1]The full list of the manually created filter is available for inspection and download at https://www.dropbox.com/scl/fi/up1hzz80nuwaka7uj92g0/DomainFilter?rlkey=aolnuz6rbdapifgznd4p7pem1&dl=0.

3.2. 1st-party doppelgängers

The process of identifying candidate 1st-party doppelgängers is straightforward once the datasets are in place. First, we identify all subdomains associated with the Orbis company names in the Rapid7 dataset. Next, we search for each of the potential attack domains within the .com zone file data. As explained in Figure 1, 1st-party doppelgängers combine one of a company's legitimate subdomain names with its domain name, with or without the conjunction character –. The hyphen, or dash, character is included as a conjunction because it is the only non-alphanumeric character allowed in domain name values in the .com namespace according to RFC 1034 (Mockapetris, 1987). Compared to other forms of domain impersonation discussed in Table 1, 1st-party doppelgängers require additional effort from the malicious actors to tailor the attack to the victim's infrastructure.

Because the Rapid7 data dates to 2018, we identify all legitimate subdomains in use by companies between 2018–2022. We then check for possible doppelgängers in the zone file dating back to 2009. Hence, we will miss any doppelgängers impersonating subdomains that were in use before 2018 but were no longer used by then.

3.3. 3rd-party doppelgängers

As organizations host more services online, they regularly outsourcce to 3rd-party providers. While some firms choose to self-host on their own infrastructure, they increasingly rely on the service providers to host in "the cloud". These cloud-hosted services often offer a personalized 3rd-level subdomain named after the customer. For example, Microsoft's SharePoint online collaboration platform creates two dedicated subdomains. For example, a fictional company, WidgetsXYZ would be assigned widgetsxyz.sharepoint.com and widgetsxyz-my.sharepoint.com. Once again, we expect that targeted attacks may involve registering lookalike domains that mimic closely the names used in the legitimate services. In 3rd-party doppelgänger

attacks, the attacker's intended victim is the company name that is being hosted as a legitimate subdomain of a legitimate service provider, not the service provider itself.

In order to find attack registrations, we must find domain that are composed of the victim subdomain name followed by the service provider domain name. We first describe two approaches that we considered but ultimately rejected. The naïve approach is to search for all registrations in the zone file that use the company name as part of a potential attack domain name. For example, for the fictitious `widgetsxyz.com`, we could search for any domains in the zone file that use `widgetsxyz` at the beginning of their name. Once identified, the next step would be to determine if they are related to a service provider whose second-level domain is impersonated. Identifying that what follows is in fact a service provider and not something else is computationally infeasible, and a cursory inspection of this approach reveals unacceptably high false positives.

Another option, to compile a list of providers, is doomed to be forever incomplete. While we could identify large providers such as SharePoint, there may be smaller operators that we would miss if we relied on a list of known providers.

We now describe the utilized method for identifying 3rd-party doppelgängers. First, we search the subdomains of *all* second-level domain names reported by Rapid7. Second, we check to see if any company names appear as valid subdomains of these second-level domains. When they do, we conclude that there are two possible 3rd-party doppelgängers that must be checked for in the zone file. For example, if `widgetsxyz.provider.com` appears in the Rapid7 data, we then check to see if `widgetsxyzprovider.com` or `widgetsxyz-provider.com` was ever registered in the historical zone file data. Any matches are deemed 3rd-party doppelgängers.

4. Empirical analysis

We now describe the findings of our large-scale examination of doppelgänger attacks spanning over 12 years. Section 4.1 discusses 1st-party attacks, followed by 3rd-party attacks in Section 4.2. In Section 4.3 we examine the cybercriminal infrastructure utilized in both attacks.

4.1. 1st-party doppelgängers

Of the $238,761$ Orbis second level domains searched for, we observed $222,328$ of them in the Rapid7 data. From there, we extracted a total of $270,978,567$ unique subdomain and domain name combinations from Rapid7 for the Orbis companies identified, giving an average of $1,218$ subdomains per domain. In total, we observed $22,223,197$ unique subdomain names, which highlights how large the potential attack surface is.

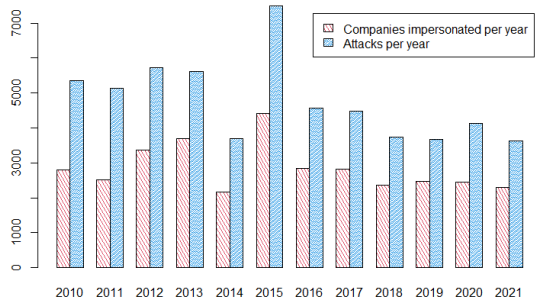

Figure 2: 1st-party doppelgänger attacks created over time (blue) and number of companies attacked (red).

We next search the historical `.com` zone file for each of the $541,957,134$ potential attack domains. In total, we observed registrations for $84,952$ of them. On the one hand, this confirms that 1st-party doppelgängers have been a common attack vector to impersonate companies through the years. $24,343$ companies, 10.2% of the total, are targeted by at least one 1st-party doppelgänger attack. On the other hand, it also reveals that attackers have not taken anywhere near full advantage of the opportunity, registering just 0.016% of the candidate attack domains.

In Figure 2 we see the total 1st-party doppelgänger attacks per year along with the total number of companies attacked per year. The number of companies affected is similar in magnitude to the total number of attacks, suggesting a small number of attacks per company.

Differences in subdomain popularity In order to differentiate the most commonly used subdomain names, such as `www` and `mail` from more domain-specific names, we evaluate how companies utilize subdomains. We observe huge variations in how frequently a particular subdomain is used. At one extreme, 91% of companies utilize the `www` subdomain, along with 41% utilizing `mail`. Popularity falls off greatly from there. Only seven subdomains were utilized by more than 5% of companies and were included in doppelgänger attacks. We classify such subdomains as *common*. For the subdomain names that are observed for less than 5% and more than 0.001%,

Table 2: 1st-party attack domains by frequency of occurrence.

Category	Subdomains	Attack Domains
Common	7	22,414 (26%)
Uncommon	9 713	53,629 (63%)
Rare	8 229	8,934 (11%)

Table 3: Common subdomain observation and 1st-party doppelgänger attack rates.

Subdomain	Attack Domains		Observations	
	#	Rate	#	Rate
www	20 029	9.2 %	216 704	90.8 %
mail	1 578	1.6 %	97 717	40.9 %
webmail	451	1.1 %	39 226	16.4 %
remote	224	1.5 %	15 297	6.4 %
vpn	122	0.9 %	13 782	5.8 %
autodiscover	7	0.0 %	52 992	22.2 %
cpanel	3	0.0 %	26 443	11.1 %

Table 4: Distribution of doppelgängers per company.

Group Size	1st Party		3rd Party	
	#	%	#	%
1	15,388	63.0 %	1,396	67.2 %
2	3,649	14.9 %	314	15.1 %
3	1,597	6.5 %	143	6.9 %
4	906	3.7 %	64	3.1 %
5	573	2.3 %	36	1.7 %
6	377	1.5 %	22	1.1 %
7	366	1.5 %	16	0.8 %
8	223	0.9 %	17	0.8 %
9	166	0.7 %	12	0.6 %
10	138	0.6 %	8	0.4 %
11-20	575	2.4 %	40	1.9 %
21-50	334	1.4 %	8	0.4 %
51-100	94	0.4 %	2	0.1 %
101-200	37	0.2 %		
201-1000	16	0.1 %		
>1000	2	0.0 %		

we classify them as *uncommon*. Finally, the subdomain names that were observed in less than 0.001% of the domains are classified as *rare*.

This differentiation is noted in Table 2. We can see that just seven subdomains account for 26% of all 1st-party doppelgängers. Uncommon domains account for 63% of the total. Subdomains that are rare, often unique to the victim company, account for nearly 9,000 observed 1st-party doppelgänger attacks. This indicates that many attackers are choosing to impersonate subdomains tailored specifically to the victim company.

Returning to the popular end of the spectrum, Table 3 shows the observation and attack count for the most common subdomains. The subdomains `webmail`, `autodiscover`, and `cpanel` are all common parts of the cPanel server management application. While these cPanel subdomains are widely used by companies, they have not been utilized in attacks very often. Hence, the popularity of a subdomain does not always lead to its use by attackers.

Victim companies The resulting group size distribution of 1st-party doppelgängers is shown in the left columns of Table 4. Most victim domains were only attacked once with this technique, and out of the 24,343 observed attacks, 63.2 % were against a single company. This still leaves multiple attacks a reality for more than one-third of the victim companies, with some experiencing hundreds of attacks.

Table 5 shows the 25 companies most targeted by 1st-party doppelgänger attacks. Attacks against `tmall`, `instructure`, `blackboard`, and `salesforce` account for a large portion of the 1st-party attacks in 2015, the top year overall for such attacks. Some companies in this list also feature prominently in the list of most-targeted providers in 3rd-party doppelgänger attacks, as we will show in the next subsection. To understand why, consider `tmall`, a leading Chinese e-commerce site. Tmall sets up brand-specific pages as subdomains (e.g., `nike.tmall.com`). Because each of these brands are themselves companies, when an attacker registers `nike-tmall.com`, it is both a 1st-party doppelgänger (impersonating `tmall.com`) and a 3rd-party doppelgänger (impersonating Nike).

While Figure 2 suggests a roughly steady if slightly declining occurrence of 1st-party attacks over time, we often see in Table 5 individual rises and falls in attack frequency by companies. For example, the top two targeted companies, `tmall` and `taobao`, the first half of the 2010 decade was much worse than the later half. By contrast, for `instructure` and `blackboard` (the two leading academic learning mangagement system providers), attacks peaked in 2015 but decline in later years. Encouragingly, only one company in the top 25 (`okta`) experienced its worst year for 1st-party doppelgängers in 2021, the most recent year with complete data. Overall, such variations suggest that for the most-targeted victim companies, attacker behavior evolves over time, potentially in response to defenders identifying and blocking doppelgänger domains.

Table 5: 1st-party doppelgänger attacks per year for 25 most targeted companies.

Domain	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
tmall	205	549	471	313	170	209	106	79	63	63	71	65
taobao	538	431	281	323	194	74	71	51	35	36	27	10
instructure		3	12	20	47	208	53	45	27	29	80	48
okta	7	7	5	11	16	28	41	59	60	64	102	125
yelp	50	65	54	44	39	46	42	36	23	21	20	27
blackboard	17	38	51	56	43	109	46	39	24	2	12	6
salesforce	8	6	3	5	5	232	2	70	1	5	4	7
facebook	80	89	48	15	16	15	15	12	8	8	7	6
yandex	17	13	61	41	14	19	16	17	20	13	16	21
microsoft	21	25	22	15	18	15	23	22	18	15	33	29
juiceplus	57	65	63	41			1			1	1	
webs	26	16	29	12	19	27	22	15	8	5	14	8
clickfunnels					8	24	30	24	29	35	32	10
alibaba	23	36	11	15	20	26	19	13	5	8	4	7
att	24	13	9	9	11	22	4	7	13	16	20	14
adp	12	5	12	17	17	28	26	5	8	9	11	7
squarespace	5	1	7	10	7	23	13	16	16	15	22	11
cisco	16	20	10	6	13	14	13	10	12	10	7	7
intuit	17	9	12	15	9	15	13	4	9	12	16	6
baixing	10	33	21	14	17	9	8	7	3	3	1	2
alipay	15	19	8	3	16	19	3	10	5	7	17	5
weibo	29	60	18	5	2	6	3	3		1		
uber	6	4	7	4	18	23	20	11	8	9	11	5
grubhub	7	10	8	11	8	15	12	15	6	7	12	8
adobe	9	11	12	3	9	16	9	4	17	5	10	12

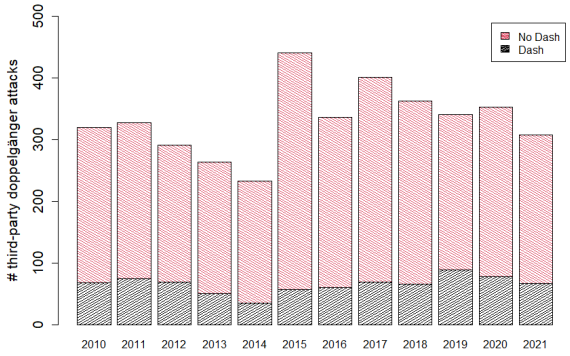

Figure 3: 3rd-party doppelgänger attacks per year split by conjunction type.

4.2. 3rd-party doppelgängers

In total, we observe 5,448 3rd-party doppelganger attacks. We distinguish between victim companies (whose names appear in the subdomain) and providers (whose second-level domain is utilized).

Figure 3 shows the total number of 3rd-party doppelgänger attacks over time and split by conjunction types. The peak observation was in 2015, and the relative proportion of dash to no-dash conjunction remained similar across all years. There appears to be a modest decline in total attacks since 2017.

We observed 2,343 unique subdomain names used across the attack domain registrations, meaning there were 2,343 unique companies targeted by these attacks.

Victim company target frequency We first consider how often companies are targeted by 3rd-party doppelgängers. The right-most columns in Table 4 report how the attacks are distributed amongst companies. Two-thirds of companies are attacked only once in this way, whereas just $2.4\,\%$ of companies experience more than 10 attacks.

The vast majority of companies experiencing many 3rd-party doppelgängers are consumer-facing product companies (e.g., Facebook, Samsung, Disney, Toyota, T-Mobile). As described in the previous section, we anticipate that many of these attacks are simultaneously 1st- and 3rd-party doppelgängers, since consumer-facing websites sometimes place brands as subdomains.

We have also found evidence of defensive registrations among potentially malicious domains, meaning one of the companies in the attacked domain

Table 6: Attack rate for the most-targed 3rd-party providers.

Provider	Observed Subdomains	Attack Domains #	Attack Domains %	Naïve Observed
benefitsnow	43	10	23.3%	542
klimaservis	184	17	9.2%	347
partswebsite	248	17	6.9%	52
corporateperks	303	20	6.6%	151
mysecurebill	232	13	5.6%	240
csod	1,073	34	3.2%	456
fbmta	2,028	43	2.1%	133
followmyhealth	1,187	20	1.7%	151
onelogin	893	14	1.6%	266
nfl	1,578	15	1.0%	17,445
ultipro	1,749	15	0.9%	489
recruiting	20,312	163	0.8%	21,658
blackboard	3,611	21	0.6%	1,758
custhelp	2,112	12	0.6%	170
okta	26,420	120	0.5%	2,586
juiceplus	2,846	12	0.4%	71,059
powerschool	3,740	13	0.3%	642
meetup	12,356	34	0.3%	7,291
tmall	59,280	149	0.3%	19,315
facebook	13,081	32	0.2%	25,088
blackberry	10,067	21	0.2%	3,648
quip	20,270	22	0.1%	15,984
nedir	14,712	15	0.1%	3,475
instructure	25,164	20	0.1%	892
fandom	22,706	18	0.1%	1,884
taobao	126,181	80	0.1%	10,911
teamwork	21,079	11	0.1%	1,904
visualstudio	20,927	10	0.0%	486
sharepoint	104,244	49	0.0%	4,274
webex	148,359	65	0.0%	1,201
fang	39,632	11	0.0%	58,996
homestead	45,519	12	0.0%	13,377
zendesk	475,831	13	0.0%	308
blogspot	9,929,781	29	0.0%	9,406

has registered it to prevent attackers from doing so. For example, the website adobefacebook.com uses name servers hosted by Adobe. Such defensive registrations are unfortunately quite rare.

Provider popularity for attackers We now investigate a related, but subtly different, question. We investigate whether certain 3rd-party providers are utilized more often in these attacks. In other words, is SharePoint the only provider impersonated by attackers to trick victim companies, or are there others?

In total, 3,177 distinct provider second-level .com domains are utilized in the 5,188 3rd-party doppelgänger attacks. Hence, most providers are used for such attacks once or twice. Hence, it will be difficult for most providers to recognize that they have been exploited in this manner.

Nonetheless, some providers are used repeatedly by attackers. Providers utilized in at least ten attacks are shown in Table 6. The total number of subdomains observed for each of the providers is shown, along with the percentage of subdomains that are attacked. *Naïve observed* reports the number of registered domains that start with the victim company and include the provider name (i.e., one of the approaches we described in Section 3.3 but did not pursue). While it is likely that some of these will be attacks, our more conservative approach does not include them. It also may explain why we observe an order of magnitude fewer attacks compared to 1st-party doppelgängers.

We observe several service providers in the list. For example, online HR services provider benefitsnow had the highest observed attack percentage of all 3rd-party doppelgänger victims. Of the 43 observed subdomains for benefitsnow in the Rapid7 data, 10, or 23.3%, also had 3rd-party doppelgängers registered. We identified 49 attacks on sharepoint, a tiny fraction of the 104,244 legitimate subdomains observed. Since companies do not have to choose the same name for their subdomain that they do for their own website, it is likely that quite a few of the 4,274 domains identified using the naïve method are in fact doppelgängers.

Other service providers in the list include HR providers (corporateperks, csod, recruiting, ultipro,), healthcare management services (followmyhealth), customer service providers (custhelp.com, zendesk), technology providers (okta, onelogin, webex), project management providers (teamwork), and learning management platforms (blackboard, instructure, powerschool). These all represent providers of outsourced services who delivered that service by creating subdomains with client names that proved attractive for attackers to impersonate.

Not all of the domains in the list are providers, though. For example, while nfl has a high attack percentage, the combination of the association of the domain with a popular sports league and the domain only having three letters in it leads to a very high naïve observation count. Also, websites such as tmall and taobao appear on both this list and the list of high 1st-party doppelgängers, reflecting the fact that these sites use brands as subdomains and they are popular enough to be impersonated. By contrast, most of the providers have high attack percentages coupled with a relatively lower number of naïve observed domains.

Looking at the top 30 most victimized 3rd-party provider domains across the years in Table 7, there is only one domain that was used across each of the years and it is the third most victimized domain, recruiting. recruiting had a little overall variance in the observed attacks across the years, unlike the top two victimized domains, tmall and okta, which both had large variations in the number of observations over the years, including several years that had no observations at all. The observations for

Table 7: 3rd-party doppelgänger attacks per year for 30 most targeted companies.

Row Labels	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	Total
tmall		21	42	38	13	8	7	3	4		2	1	6		145
okta						2	2	6	18	14	11	22	24	21	120
recruiting	3	11	6	8	12	4	6	3	11	3	9	9	10	9	104
taobao	32	14	10	5	2	1	1	1		1	1	1			69
webex		6	2	5	12		5	3	5	4	8	4	6		60
fbmta				1		2	6	1	17	10	3	2	1		43
sharepoint		2		1	1	2	2	3	3	7	3	8	2	6	40
csod				1	3	3	5	1	3	4	1	10	1	2	34
meetup		12	2		2		2	3	1	1			2		25
facebook	2	3	9	5		1		1	1		1	1			24
followmyhealth					1		5	1	1	1	3	4	3	1	20
instructure						2	8	2	1	1	1	1	4		20
blogspot	1	1	4	3	2		2		1	2	1		2		19
blackboard	1	1	1	1	2	1	6		1	1					15
fandom		1			1		3		1	2	2	1	4		15
ultipro						4	4	1	2	1		2		1	15
onelogin							1		3	1	3	2	2	2	14
quip	1	3			2	1	1				1	4		1	14
mysecurebill										1		11	1		13
zendesk							2		2	1	4	2	2		13
blackberry	4	4	2							1				1	12
custhelp		2			1		1	1	2	1	1	3			12
klimaservis	2	3		2	1		4								12
powerschool					1		4	1		1	1	3		1	12
benefitsnow		2		1		2	4	1							10
corporateperks		3			3	1	3								10
homestead				1		1		1	1		2	2	1	1	10
nedir			1	3			1		1	1	1		1	1	10
fang		1		3	1		2	2							9
turkiyeservisi			3		1							3		2	9

the rest of the domains in the top 30 are infrequent, with some like `mysecurebill` and `fbmta` having one or two years that were large deviations from the other years, suggesting a potential targeted malicious campaign utilizing those domains.

Case study: `fbmta.com` We conclude by discussing the details of one little-known provider, `fbmta.com`, which hosts a "restaurant technology marketing platform". Originally named Fishbowl, but now called Personica[2], this platform offers a way for restaurants to track customers. The `fbmta` domain was observed to have 1,674 unique subdomain names in June 2022, mostly names of restaurants and associated locations. The attack domains formed from these types of existing relationships prey on an expectation of legitimacy based on a previous interaction with the specific restaurant. We observed 43 3rd-party doppelgänger attacks on `fbmta`, including `arbys`, `peiwei`, `goldencorral`, `schlotzskys`, `mazzios` and `dennys`.

4.3. Attacker infrastructure

The prior analysis focused on the victims selected by attackers. We now briefly examine the infrastructure utilized to perpetrate the attacks. The best data sources we have to study attacker infrastructure are the nameservers used to host doppelgänger domains. Hosting two distinct doppelgänger domains with the same nameserver suggests that the same criminal may have registered both impersonating domains, particularly if it is unpopular. Researchers have also found a strong correlation between nameserver and domain name registrar (Simpson et al., 2020).

Figure 4 plots the count of doppelgängers each year grouped by the associated nameservers. It is common for cybercriminals to use an infrastructure provider until that provider cracks down, at which point they move onto a new target (Böhme and Moore, 2016). We can observe this "iterated weakest link" in the figure. The most striking example is `rookdns.com`. It is not used at all until 2014, when it rises in prominence. In 2015, it becomes the most commonly used nameserver, drops off in 2016 before disappearing completely in 2019. Not coincidentally, `namedynamics.net` appears in 2019 in its place. We observe similar patterns of introduction, increase, and decline with `mytrafficmanagement.com` and `worldnic.com`.

Of course, there are exceptions to this pattern, such as `domaincontrol.com`, which appears consistently each year. In these cases, the nameservers are associated

[2]https://personica.com/new-brand-faq/

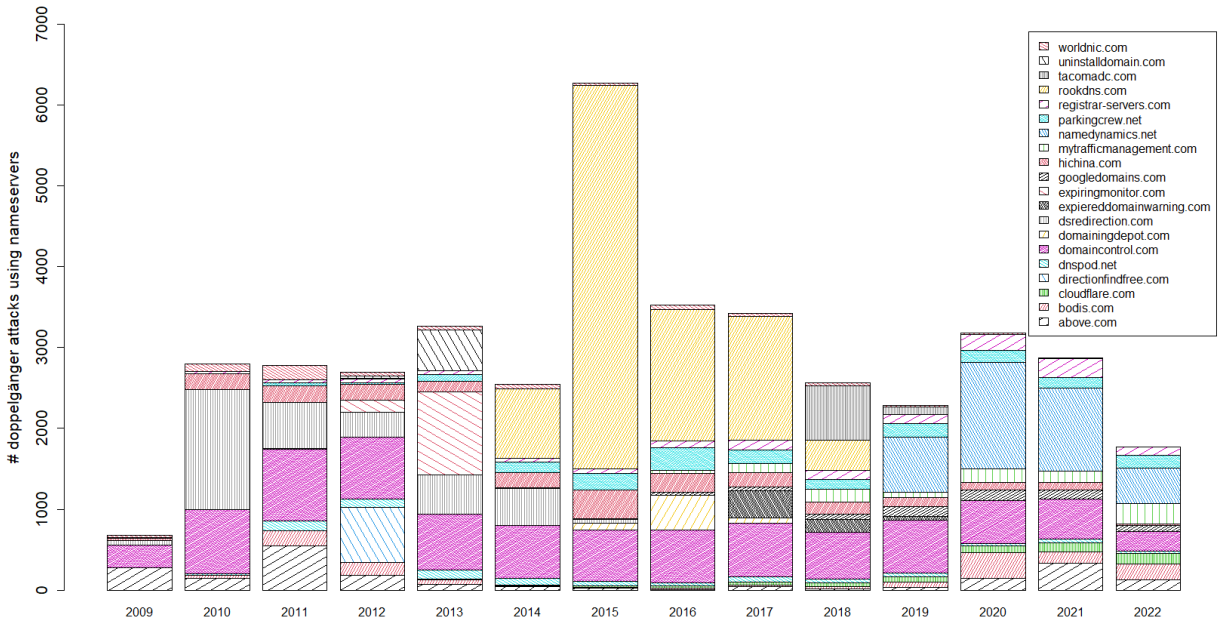

Figure 4: Combined top 20 nameserver usage across both doppelgänger attacks

with popular registrars, which some cybercriminals choose to utilize. Where the iterated weakest link pattern does hold, however, this suggests that a smaller number of criminal actors may be carrying out doppelgänger attacks en masse and with impunity.

5. Conclusion

Companies have experienced targeted cyber attacks for a long time. In this paper, we have studied an underappreciated mimicry tactic we call subdomain doppelgänger attacks at scale for the first time. For the period from 2009–2022, we present a method that identifies $84,952$ 1st-party doppelgänger attacks affecting $24,343$ companies in which legitimate company subdomaines are impersonated. As outsourced cloud providers have risen in popularity, so too have 3rd-party doppelgänger attacks. We identified $5,448$ such attacks in which 3rd-party service providers used by companies are impersonated.

While the main goal of the research is to improve our understanding of cybercriminal behavior and its impact, we can offer some advice to organizations on practical steps they should take to protect themselves.

Fortunately, it is much easier for a single organization to determine its own susceptibility to the attacks we presented than to track the global phenomenon. So what should defenders do? Dealing with 1st- and 3rd-party doppelgangers are not currently identified by existing tools that enumerate possible malicious domain names such as URLCrazy (Horton, 2023), though they could be augmented to do so. More broadly, as with any crime, reducing exposure to the attacks is good practice. Keeping an inventory of exposed subdomains, and then removing them when not needed helps minimize the external vulnerability to such attacks. Moreover, companies who use outsourced services should also keep an inventory of legitimate services and monitor for doppelgänger domain registrations.

Providers that host customers on named subdomains must be aware of the increased risk of attack. Such services should proactively check for impersonated domain name registrations and notify their customers when identified.

Future work could include developing a real-time monitoring service to check for doppelgängers in closer to real time. Further investigation of the attacker infrastructure may yield new insights on how best to disrupt attacks. Finally, our detection methodology could be expanded to identify more doppelgängers.

Acknowledgements

The authors acknowledge support from US National Science Foundation Award No. 1652610.

References

U.S. Federal Bureau of Investigation. (2022). 2022 Internet Crime Report [https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf].

Area 1. (2020). Phishing Burisma holdings [https://cdn.area1security.com/reports/Area-1-Security-PhishingBurismaHoldings.pdf].

Anderson, R., & Moore, T. (2006). The economics of information security. *Science*, *314*(5799), 610–613.

Moore, T., Clayton, R., & Anderson, R. (2009). The economics of online crime. *Journal of Economic Perspectives*, *23*(3), 3–20.

Levchenko, K., Pitsillidis, A., Chachra, N., Enright, B., Félegyházi, M., Grier, C., Halvorson, T., Kanich, C., Kreibich, C., Liu, H., et al. Click trajectories: End-to-end analysis of the spam value chain. In: In *IEEE Symposium on Security and Privacy*. IEEE. 2011, 431–446.

Liu, Y., Sarabi, A., Zhang, J., Naghizadeh, P., Karir, M., Bailey, M., & Liu, M. Cloudy with a chance of breach: Forecasting cyber security incidents. In: In *24th USENIX security symposium*. Washington, D.C., 2015, August, 1009–1024. ISBN: 978-1-931971-232.

Liu, H., Levchenko, K., Félegyházi, M., Kreibich, C., Maier, G., Voelker, G. M., & Savage, S. On the effects of registrar-level intervention. In: In *USENIX Workshop on Large-scale Exploits and Emergent Threats*. Boston, MA, 2011, 5–5.

Metcalf, L., & Spring, J. M. (2013). *Everything you wanted to know about blacklists but were afraid to ask* (tech. rep.). Software Engineering Institute – Carnegie Mellon University.

Wang, Y.-M., Beck, D., Wang, J., Verbowski, C., & Daniels, B. Strider typo-patrol: Discovery and analysis of systematic typo-squatting. In: In *2nd USENIX Workshop on Steps to Reducing Unwanted Traffic on the Internet*. San Jose, CA: USENIX Association, 2006, 5.

Simpson, G., Moore, T., & Clayton, R. Ten years of attacks on companies using visual impersonation of domain names. In: In *APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, 2020.

Nikiforakis, N., Balduzzi, M., Desmet, L., Piessens, F., & Joosen, W. Soundsquatting: Uncovering the use of homophones in domain squatting. In: In *International Conference on Information Security*. Springer. 2014, 291–308.

Du, K., Yang, H., Li, Z., Duan, H., Hao, S., Liu, B., Ye, Y., Liu, M., Su, X., Liu, G., et al. TL; DR hazard: A comprehensive study of levelsquatting scams. In: In *International Conference on Security and Privacy in Communication Systems*. Springer. 2019, 3–25.

Kintis, P., Miramirkhani, N., Lever, C., Chen, Y., Romero-Gómez, R., Pitropakis, N., Nikiforakis, N., & Antonakakis, M. Hiding in plain sight: A longitudinal study of combosquatting abuse. In: In *ACM Conference on Computer and Communications Security*. ACM, 2017, 569–586.

Tian, K., Jan, S. T., Hu, H., Yao, D., & Wang, G. Needle in a haystack: Tracking down elite phishing domains in the wild. In: In *ACM Internet Measurement Conference*. 2018, 429–442.

Pouryousef, S., Dar, M. D., Ahmad, S., Gill, P., & Nithyanand, R. Extortion or expansion? an investigation into the costs and consequences of ICANN's gTLD experiments. In: In *International Conference on Passive and Active Network Measurement*. Springer, 2020, 141–157.

Moore, T., & Edelman, B. Measuring the perpetrators and funders of typosquatting. In: In *Financial cryptography and data security*. *6052*. LNCS. Springer, 2010, 175–191.

Szurdi, J., Kocso, B., Cseh, G., Spring, J., Felegyhazi, M., & Kanich, C. The long "taile" of typosquatting domain names. In: In *23rd USENIX Security Symposium*. 2014, 191–206.

Quinkert, F., Tatang, D., & Holz, T. Digging deeper: An analysis of domain impersonation in the lower DNS hierarchy. In: In *International conference on detection of intrusions and malware, and vulnerability assessment*. Springer. 2021, 68–87.

Bureau van Dijk. (2023). Orbis: Company information across the globe [https://orbis.bvdinfo.com/].

Davies, M. (2023). Word frequency data [https://www.wordfrequency.info/].

Mockapetris, P. V. (1987). RFC1034: Domain names-concepts and facilities.

Tange, O. (2011). Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, *36*(1), 42–47. http://www.gnu.org/s/parallel

Böhme, R., & Moore, T. (2016). The "iterated weakest link" model of adaptive security investment. *Journal of Information Security*, *7*(2), 81–102.

Horton, A. (2023). Urlcrazy [https://morningstarsecurity.com/research/urlcrazy].