MitM Attack by Name Collision: Cause Analysis and Vulnerability Assessment in the New gTLD Era

Qi Alfred Chen, Eric Osterweil†, Matthew Thomas†, Z. Morley Mao
University of Michigan, †Verisign Labs
alfchen@umich.edu, {eosterweil, mthomas}@verisign.com, zmao@umich.edu

***Abstract*—Recently, Man in the Middle (MitM) attacks on web browsing have become easier than they have ever been before because of a problem called "Name Collision" and a protocol called the Web Proxy Auto-Discovery (WPAD) protocol. This name collision attack can cause all web traffic of an Internet user to be redirected to a MitM proxy automatically right after the launching of a standard browser. The underlying problem of this attack is internal namespace WPAD query leakage, which itself is a known problem for years. However, it remains understudied since it was not easily exploitable before the recent new gTLD (generic Top-Level Domains) delegation.**

In this paper, we focus on this newly-exposed MitM attack vector and perform the first systematic study of the underlying problem causes and its vulnerability status in the wild. First, we show the severity of the problem by characterizing leaked WPAD query traffic to the DNS root servers, and find that a major cause of the leakage problem is actually a result of settings on the end user devices. More specifically, we find that under common settings, devices can mistakenly generate internal queries when used outside an internal network (e.g., used at home). Second, we define and quantify a candidate measure of attack surface by defining "highly-vulnerable domains", which are domains routinely exposing a large number of potential victims, and use it to perform a systematic assessment of the vulnerability status. We find that almost all leaked queries are for new gTLD domains we define to be highly-vulnerable, indirectly validating our attack surface definition. We further find that 10% of these highly-vulnerable domains have already been registered, making the corresponding users immediately vulnerable to the exploit at any time. Our results provide a strong and urgent message to deploy proactive protection. We discuss promising directions for remediation at the new gTLD registry, Autonomous System (AS), and end user levels, and use empirical data analysis to estimate and compare their effectiveness and deployment difficulties.

I. Introduction

Recently, Man in the Middle (MitM) attacks on web browsing have become easier than they have ever been before — the attacker only needs to register one of certain domain names, and web traffic of Internet users from all over the world can be automatically redirected to the attacker's MitM proxy. The underlying vulnerability comes from a problem called "Name Collision" [31]. Name collisions occur when administrators configure their internal systems to use names from local/internal namespaces that are also used in other namespaces (such as the global Domain Name System, DNS), and a collision happens when a query for a name is resolved in an unexpected namespace.

The MitM attack focused upon in this paper is a name collision based attack that arises from leakage of internal namespace Web Proxy Auto-Discovery (WPAD) queries. These WPAD queries are designed to automatically configure proxies for end systems only from within an administrative domain such as a corporate internal DNS namespace, but only in two of 13 DNS root servers, roughly 20 million such queries are observed to be leaking to the public DNS namespace every day. This has been a known problem for years but remains understudied, mainly because these queries typically use undelegated TLDs as internal Top-Level Domains (iTLDs) [5], [13], [21], and thus were not exploitable previously. However, in the recently-launched New gTLD (generic Top-Level Domains) Program [12], many of these popular iTLD strings have begun to be delegated and are open for public domain name registration, allowing attackers to exploit these leaked WPAD queries by setting up MitM proxies from anywhere on the Internet with only a domain name registration. Note that this is not a limitation or weakness of new gTLDs per se, but instead a manifestation of a name configuration problem leading to name collisions which we argue should be fully mitigated.

To characterize the magnitude of this newly-exposed MitM threat, we perform the first systematic study of the underlying problem causes and the vulnerability status in the wild. First, we investigate the fundamental underlying cause of WPAD query leaks from internal networks. Using a local testbed and traffic analysis, we find that a major cause that accounts for a significant proportion of the leakage traffic is actually a result of settings on the end user devices. More specifically, we find that under common settings, devices can mistakenly generate internal queries when used outside an internal network (e.g., used at home). From this finding, we identify a set of highly-vulnerable Autonomous Systems (ASes) with both high volume of leaked WPAD queries and high diversity of vulnerable query domain names, which is found to be dominated by home access network ASes.

Second, for these highly-vulnerable ASes, we perform a systematic assessment of the vulnerability status in the wild. Leveraging the insights that most domain names in leaked WPAD queries are transient and low-volume, we propose that a more useful characterization of attack surface should focus on domain names that persistently expose many victims. We call such domain names *highly-vulnerable domains* (HVDs), because an adversary could gain more value from operating them. From this definition, we then design an attack surface quantification method which systematically balances the trade-off between query persistence and high query volume. This

allows us to focus on the most exploitable domain names. For example, for the delegated new gTLD `.network`, only 4% of the domain names in the leaked WPAD queries match the HVD definition.

By applying our attack surface quantification method to the victim ASes, we find that almost all of the leaked queries are for new gTLD domain names defined to have high vulnerability, which indirectly validates our attack surface definition. If these domain names are registered by an attacker, she becomes authoritative to answer all the vulnerable queries, and actual exploits can start at any time. Fortunately, as of September 2015, the registration of these HVDs just started, and our registration status analysis (detailed in §VI-B) does not find statistical evidence showing that these domains are being maliciously targeted for registration. Nevertheless, we did find seemingly naïve attack registration patterns in the wild, showing potential attack attempts. These results illustrate real MitM threat for Internet users in the wild, and provide a strong and urgent message to deploy proactive protection.

To effectively defend against this attack, remediation strategies can be deployed at the new gTLD registry level to scrutinize the registration of HVDs, and also at the AS level and end user level to prevent the vulnerable queries from being leaked to the public DNS namespace. Based on the insights from the problem cause and vulnerability characterization, we discuss feasible defense methods for each of these three levels, and use empirical data analysis to estimate and compare their effectiveness and deployment difficulties.

We summarize the key contributions as follows:

- Targeting the new MitM attack vector exposed by name collisions, we perform a characterization of the problem and its severity, and an in-depth analysis on the fundamental internal namespace WPAD query leakage problem. From the analysis, we are able to uncover the major leak sources and the underlying device-side causes using both local testbed and DNS root server traffic analysis.
- We present a candidate definition and quantification method for the attack surface of this MitM threat, and use it to systematically study the vulnerability status in the wild. With this, we are able to find a set of highly-vulnerable domains (HVDs) which persistently expose many victims in the wild. We find that over 97% of the leaked WPAD queries are for these HVDs, and at this point, the HVDs for 10% of the new gTLDs have already been fully registered. These results show a real threat for Internet users in the wild.
- To prevent users from being exploited by this newly-exposed attack vector, based on the insights in our cause analysis and vulnerability quantification, we discuss a set of remediation strategies at the new gTLD registry, AS, and end user levels, and use empirical data analysis to evaluate their effectiveness and deployment challenges.

II. Background

In this section, we cover the necessary background of the public and internal DNS namespaces, and the focus of this paper, WPAD proxy discovery protocol.

A. DNS Ecosystem

DNS (Domain Name System) [27] is a distributed system which translates domain names to network service identifiers (such as IP addresses for computers in the Internet or a private network). Domain names are a set of labels separated by dots, for example `www.example.com`, and are organized in hierarchical subdomains of the DNS root domain. The first level of domain name labels under the root domain are the TLDs [9], including gTLDs such as `.com`, and country code Top-Level Domains such as `.us`. Directly below TLDs are Second-Level Domains (SLD) [7], e.g., `example` in `www.example.com`. In this paper, the term *domain* is defined to be any DNS name, and TLDs and SLDs are specific types of domains.

Domain name management and delegation. In DNS, a DNS zone is defined as the set of DNS domain names that are contiguous in the DNS tree hierarchy, and which are administered by the same authority. The DNS root zone is the canonical top of the DNS tree. It is the authoritative zone for all of DNS' TLDs. The structure and contents of the DNS root zone are determined by an organizational role called the Internet Assigned Numbers Authority (IANA), which is performed by the Internet Corporation for Assigned Names and Numbers (ICANN). The DNS root zone's actual operational and authoritative maintainer is a role called the Root Zone Maintainer (RZM), which is currently performed by Verisign. ICANN delegates the management of its subdomains, the TLDs, to TLD registry operators. Under TLDs, SLDs are registered in the process of domain name registration.

Domain name registration. A domain name registration is the delegation of the administration of an SLD and its subdomains under a TLD, which usually involves 3 parties: TLD registry operators, registrars, and registrants [32]. At a high level, registry operators manage TLDs, registrars conduct the daily business of transacting with clients for SLDs, and registrants pay to receive administrative authority to run SLDs. Once a domain is registered by a registrant, the registrar submits certain information to the corresponding TLD registry operators, and the WHOIS database [10] then maps the registered domain name to the registrant details.

Domain name resolution. In the domain name resolution process, end hosts rely on recursive DNS resolvers, usually configured by network providers, e.g., corporate network administrators and home network providers. Using the cached results whenever possible, the resolvers query the name servers following the DNS domain label hierarchy, getting either the corresponding IP address, or an NXDomain response (`rcode 3` in RFC1035 [28], NXD for short), indicating that no such domain name exists.

The New gTLD Program. In the history of DNS, the set of TLDs has remained relatively small and stable, with only 66 new TLDs added in 14 years before 2013 [31]. In 2011, with the goal of enhancing competition and consumer choice, ICANN approved the launch of the New gTLD Program [12], which in less than 2 years has added over 700 new gTLDs as of 2015/08/25. To differentiate these new gTLDs from the legacy

ones such as .com, in this paper they are also referred to as nTLDs. This enormous wave of new gTLD delegation raised name collision concern in the domain name industry [31], and in this paper, we perform the first systematic study of one of the consequences of this problem in the wild.

B. Internal DNS Namespace and iTLD Usage

The DNS ecosystem described above is the public DNS namespace for domain names visible to the Internet. Similarly, a local area network, e.g., a corporate network, can also set up an internal DNS namespace with private domain names. This helps control the access to internal confidential information, and can operate despite any external network connectivity disruption, making it a common practice for companies.

To create an internal DNS namespace, internal name servers are used to serve the zone files for a customized internal domain, and the resolvers are configured to query these servers instead of the DNS servers in public namespace. To make the internal domain name easy to reference and also to prevent confusion between internal and public namespaces, some administrators in the past used TLD strings that have not been delegated (in the public DNS namespace) as iTLDs.

The use of iTLDs implicitly assumes that these TLD strings will not be delegated in the public namespace; however, with the launching of the New gTLD Program, many of the popular iTLD strings have already been delegated today and are open for public registeration [16]. This breaks the implied assumption that previously undelegated iTLDs will never be delegated. As a side effect, the leaked internal queries to these iTLD strings that were previously benign now expose issuers to the MitM attacks studied in this paper.

C. WPAD: Automatic Proxy Discovery

WPAD (Web Proxy Auto-Discovery) is a protocol designed for browsers or operating systems (OSes) to automatically locate a web proxy configuration file. It is primarily used in internal networks where clients are restricted from communicating to the public HTTP network, e.g., in some corporate networks. The proxy configuration file is by default named `wpad.dat`, which is written in proxy auto-config (PAC) format, and specifies the proxy IP and port using code `PROXY` ⟨`IP`⟩`:`⟨`port`⟩.

To find the proxy configuration file, WPAD supports two methods: DHCP WPAD and DNS WPAD. In the implementation, usually DHCP WPAD is attempted first by issuing a `DHCPINFORM` message to the local DHCP server. If the local infrastructure supports this proxy configuration, the PAC file location is included in option 252 in the response.

If no such configuration is found in DHCP, DNS WPAD is performed. Without an explicit configuration like that in DHCP WPAD, DNS WPAD infers the location of the proxy file based on the device domain name. For example, in a company's internal network, a corporate device can be configured with internal domain `company.ntld` in the OS. In DNS WPAD proxy discovery, the proxy file location is inferred from this name and fetched using HTTP request `http://wpad.company.ntld/wpad.dat`, involving a DNS request for `wpad.company.ntld`. To serve this proxy discovery, a company can simply set up a web server with `wpad.dat` under its root directory, and point a DNS record for `wpad.company.ntld` in its local DNS zone file to this server. In this process, all the WPAD DNS queries should be served only by the local DNS resolvers, but as we show later, millions of such queries are leaked to the public DNS namespace every day, causing the name collision problem.

Supported OSes and browsers		Verified versions for DNS WPAD	Enabled by default
Browser	Internet Explorer	6–11	Yes
	Chrome	43	No
	Firefox	12, 33	No
	Safari	8	No
OS	Windows OS	XP, Vista, 7, 8, 8.1, 10	Yes
	Ubuntu	12.04, 14.04	No
	Mac OS X	10.10	No

TABLE I: Popular OSes and browsers that support WPAD.

Browser and OS support. WPAD service discovery can be supported in both OS and browser levels. The configuration is typically named "Automatically detect setting" in the LAN proxy setting [11]. Table I summarizes the popular browsers and OSes supporting WPAD, along with their versions which we have verified using a local testbed. As shown, DNS WPAD is supported by all popular browsers and OSes, and some of them even use it by default, e.g., Windows OSes and Internet Explorer (IE) browsers. Note that for the browsers and OSes that do not enable it by default, the local network administrator, e.g., the IT department in a company, may enable it during the device setup process so that end devices can use its convenient proxy discovery feature. For the browsers tested in our experiments, the discovery process starts right after the browser is launched. With a valid PAC file fetched, all subsequent web traffic is redirected to the configured proxy.

III. Threat Model and Attack Surface

In this section, we describe the threat model and attack surface definition of the newly-exposed MitM attack vector, which we call *WPAD name collision attack*.

A. Threat Model

As introduced in the previous section, the WPAD protocol is designed to only configure proxies for end systems from within an administrative domain such as a corporate internal DNS namespace. Ideally, for a device belonging to a corporate domain, it performs discovery to configure a WPAD proxy only inside that domain. While these queries may have always been vulnerable to DNS spoofing attacks, the adversaries would need to be on-path or be able to spoof DNS responses in a narrow attack window. The intended local scope of queries, the on-path requirement, and the narrow attack window have kept WPAD deceptively safe.

However, because internal queries leak to the DNS root servers and internal namespaces now collide with new gTLD domains, which are both happening in large scale today as characterized later in §IV-A, the inherent security weaknesses

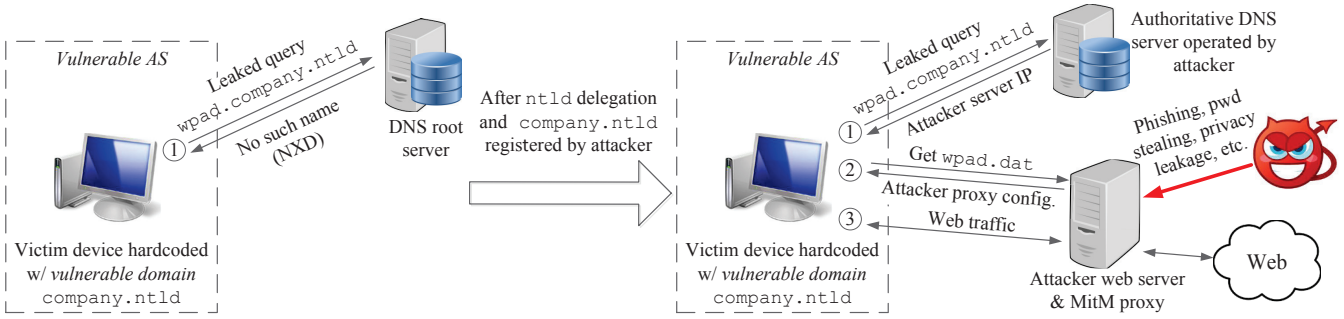

Fig. 1: Illustration of the WPAD name collision attack. If an internal namespace TLD is delegated as a new gTLD, internal namespace WPAD query leaks can be easily exploited using MitM attack from anywhere on the Internet.

in WPAD are significantly easier to exploit. Fig. 1 illustrates the WPAD name collision attack, in which a malicious domain registrant can exploit name collisions of leaked WPAD queries, and launch MitM attacks from anywhere on the Internet. In this attack, victim devices are assumed to be configured to use DNS WPAD for automatic proxy discovery by issuing WPAD queries in an internal DNS namespace, e.g., `company.ntld`. Here, `.ntld` is assumed to be used as iTLD but also delegated in the public DNS namespace. Under some common settings (uncovered in §IV), such queries are mistakenly leaked out. This allows an attacker to create name collisions for these queries by registering the domain name `company.ntld` in new gTLD `.ntld`. Thus, the leaked WPAD queries from affected systems, which may be anywhere on the Internet are sent to the attacker's authoritative name server and get resolved to fetch the attacker's proxy configuration file. This causes all the subsequent web traffic in the browser or traffic from the entire OS to be redirected to the proxy controlled by the attacker. The victim user may not even recognize the attack, since the WPAD proxy discovery is fully automated at the browser launch time, and some OSes and browsers enable it by default without explicit consent from users (shown in Table I). The attacker can leverage this MitM position to not only eavesdrop sensitive data such as confidential documents and user credentials, but also manipulate the traffic to inject malicious code, launch phishing attacks, or other malicious impacts to vulnerable systems.

In this attack, the adversaries only need to register new gTLD domains to direct potentially vulnerable WPAD queries to them. This means that if a potentially colliding internal domain is registered, the attacker can detect and respond authoritatively to WPAD queries without the need of spoofing. This frees the on-path requirement and eliminates the narrow attack window drawback of previous WPAD attacks. More importantly, the authoritative nature of the malicious responses makes this attack exploitable despite DNSSEC [18], [19].

This attack is also very stealthy, since once the domain name is registered, due to privacy protection it is difficult for both new gTLD registries and third parties to examine its subdomains for attack attempts. Note that we do not assume that the attacker is fully aware of the set of the vulnerable domains (i.e., domains with leaked queries), and thus deliberately exploits them. The attackers can be sophisticated registrants who know some vulnerable domains based on their own analysis, e.g., by sniffing local network queries or accessing DNS traffic collected by organizations such as DNS-OARC [8]. Meanwhile, the registrants can also be innocent at the domain registration time, but realize and start exploitation after observing a large number of misdirected WPAD queries. Another possibility is that the registrant is completely honest but the DNS servers are compromised by an attacker to exploit these vulnerable queries.

B. Attack Surface

In order to characterize the magnitude of this newly-exposed MitM threat, we propose a candidate methodology to quantify the WPAD attack surface exposed by registrations of new domain names under new gTLDs. With that, we describe a measure of how exposed (or open) the total attack surface is based on registration status.

Our threat model focuses on the fact that MitM attacks can be launched against any client who issues a WPAD query to a domain name that is controlled by an attacker. Thus, all domain names with leaked queries to the public namespace are vulnerable. However, we find that most of the domains in the leaked query traffic appears infrequently with low query volume, implying that they may not be easily exploited in practice. For example, we find that for the delegated new gTLD `.network`, 42.3% of the domains with leaked queries (e.g., `company.ntld` in Fig. 1) to two of 13 DNS root servers appeared in less than 14 days within a one-year period. Furthermore, less than 4% of these domains account for more than 98% of all leaked WPAD traffic observed at the two DNS root servers. Thus, using all the domains with leaked queries as the attack surface is an overestimate of the actual vulnerability status in practice. Therefore, we define a notion of "highly-vulnerable domains" based on a more accurate and useful attack surface characterization method described as follows.

Attack surface: highly-vulnerable domains (HVDs). In this paper, we define highly-vulnerable domains for a new gTLD to be those WPAD query domains persistently exposing a large number of victims. We denote these domains as the attack surface for this new gTLD. These attack surface domains

or HVDs need to have two properties: (1) high persistence, meaning that their queries are leaked to the public namespace frequently over a long time period, e.g., every day or days with regular periodicity, and (2) high query volume, indicating that once registered, many victims can be continuously exploited. From this definition, these domains are quantifiably attractive targets for adversaries, and are likely to keep exposing such vulnerability after the delegation of their TLD strings.

This methodology defines a measurably stable set of highly-vulnerable domain names. To quantify the attack surface based on this definition, we first concretely define the level of persistence using period length p and persistence duration n. We then balance the trade-off between persistence and high query volume by systematically exploring p and n, detailed later in §V-A. This quantification method allows us to estimate the size and composition of domains that, when registered, constitute the bulk of the WPAD name collision vulnerabilities.

C. Dataset

We describe the datasets used in our study as follows.

New gTLD list. We obtain the new gTLD list along with their delegation dates directly from ICANN website [16]. In this paper, we consider the new gTLDs delegated before 2015/08/25, consisting of 738 new gTLDs in total.

Root NXD WPAD. Due to the usage of non-delegated iTLDs, the leaked internal namespace queries are captured and replied with NXD by the DNS root servers. Thus, our vulnerability characterization and attack quantification mainly rely on NXD traffic collected at 2 of the 13 root servers — A root and J root, both managed by Verisign. Both root servers utilize IP anycasted services from a globally diverse set of locations [6], which should reduce any significant geographical biases in the data collection. The leaked queries become unobservable in this dataset after the delegation of their TLD strings. Thus, in the analysis of each new gTLD, we only use the data collected before its delegation date.

This dataset was collected internally by Verisign for around 2 years, spanning from September 2013 to July 2015. Since the first new gTLD delegation in the New gTLD Program occurred in October 2013, this dataset covers leaked query traffic for all the new gTLDs delegated so far. To study leaked WPAD queries, we extract the query traffic with query names in the form of `wpad.`⟨`domain name`⟩. Considering that single label domains, e.g., `wpad.ntld` are more easily defended at the new gTLD registries, in this dataset we only include WPAD queries with at least 2 labels in ⟨`domain name`⟩, e.g., `wpad.sld.ntld`, `wpad.3ld.sld.ntld`, etc.

New gTLD zone files and WHOIS data. Once a domain is registered, it appears in the corresponding new gTLD's zone files. Meanwhile, mapping from registered domains to the domain registrants are included in the new gTLD's WHOIS data. To study the registration status and registration pattern of HVDs in our attack surface, we use new gTLDs' zone files from ICANN Centralized Zone Data Service (CZDS) [15] and WHOIS data from BestWhois service [1], which are both pulled daily from 2014/02 to 2015/09.

IV. WPAD Query Leakage Characterization

The WPAD name collision attack stems from the unintentional leakage of internal WPAD DNS queries into the public DNS namespace. This problem emerged soon after the popularization of the WPAD protocol [31], [33], however remains understudied since it was not easily exploitable until the expansion of the new gTLDs.

To systematically characterize this newly-exposed threat and help find effective solutions, we need to first have an in-depth understanding of this fundamental leakage problem. In this section, we first characterize its severity by quantitative measurements of leaked WPAD query traffic seen in the DNS root servers, and then elucidate the underlying causes of these leaks using query traffic analysis and controlled local testbed experiments.

A. Quantification of Leaked Queries

Fig. 2 shows the popular first labels ranked by their average daily query numbers in NXD traffic at DNS root server A and J from January to July in 2015. In DNS-based protocols, usually the protocol name is the first label. Thus, in the figure many labels belong to popular protocols such as WPAD, ISATAP, etc. The first label query number distribution exhibits a very long tail. As shown, WPAD protocol is ranked top 4 with more than 20 million leaked queries every day, showing high severity in terms of the query leakage problem. Using the number of distinct IP address and WPAD query domain pairs in our 2-year root NXD WPAD dataset, these queries are estimated to have at least 6.6 million potential victim users in the wild.

For these leaked WPAD queries to be exploitable in our attack, their TLD domains need to be delegated so that the attacker can register the SLD and create name collisions. We study the 738 new gTLDs that have already been delegated before 2015/08/25, and find that 65.7% (485) of them exhibited leaked WPAD queries to the 2 DNS root servers in our dataset before their delegation, revealing a significant attack surface. In §V, we use a more systematic approach to quantify the attack surface for these delegated new gTLDs based on the definition in §III-B.

To understand the vulnerability exposed by the new gTLDs that have already been delegated today, we measure the daily query percentage of these delegated new gTLD strings in the leaked queries using 1 month of root NXD WPAD data immediately prior to the delegation of the first new gTLD in the New gTLD Program on 2013/10/23. Fig. 3 shows the daily query volume and the overall query percentage in root NXD WPAD dataset for delegated new gTLD strings with leaked queries. As shown, even though the query percentage is not high, some top ones such as `.global` already have over 30,000 leaked WPAD queries every day. In total, 2.3% of the daily leaked WPAD queries, which are over 238,000 queries per day on average from only 2 DNS root servers, belong to the delegated new gTLD set. According to our threat model, these queries are already exploitable today. Note that these are query volumes from just 2 of the 13 DNS root

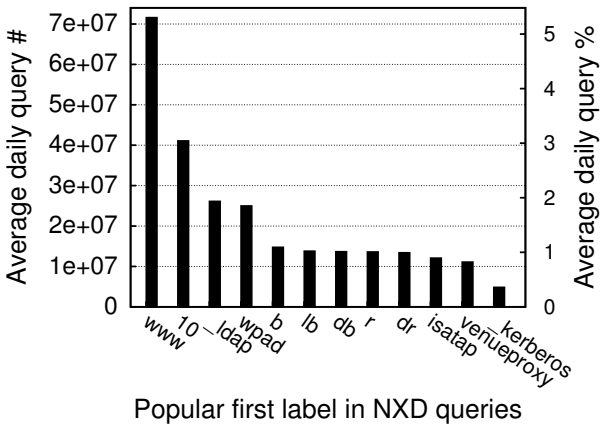

Fig. 2: The most popular first labels in root NXD traffic.

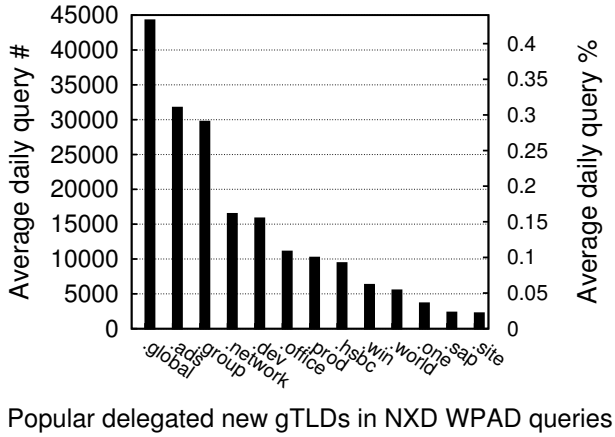

Fig. 3: The most popular delegated new gTLDs observed in root NXD WPAD queries.

servers. Furthermore, the number will only increase as more new gTLD strings continue to be delegated (as of 2016/03/20, 27.2% (201) more new gTLDs have been delegated since this study was conducted).

B. Leak Cause Analysis

1) Major Leak Source ASes: To identify the cause, we start by measuring where the leaks originate. We first break down the leaked WPAD traffic into country level according to their query IP addresses. Fig. 4 shows the country codes ranked by their average daily leak percentage in our root NXD WPAD dataset from January to July 2015. As shown, U.S. (United States) dominates the leaked traffic with nearly 70% worldwide, and its share is over 6× more than that of the country ranked the second. In the following analysis, our focus is mainly on the leaked query traffic from the U.S.

Within the U.S., we further characterize the query traffic according to ASes. Fig. 5 shows the ASes with top average daily WPAD query leaks from January to July, 2015. As shown, the overall distribution exhibits a long tail, in which nearly 2000 ASes have leaked queries, but the majority of these queries come from only a few top ASes. The top 12 ASes account for 85% of all the leaks, and their names are listed in Table II. In the table, we denote these ASes A1 to A12 to obfuscate the actual AS in our data. As shown, 10 out of the 12 ASes are home access network ASes. The remaining two ASes both operate open (publicly accessible) DNS resolvers, and we find that the queries come predominantly from source IP addresses within the IP address ranges listed as open DNS resolver servers on their websites. Thus, both ASes are associated with open resolver usage, which is also commonly configured by home access network users. These results suggest the major cause of WPAD query leaks is user behavior at home instead of in corporate networks.

2) Leak Domain Suffixes: To investigate why WPAD queries are leaked from home, we closely examine the domains of leaked WPAD queries in these home access network ASes. Surprisingly, instead of being dominated by a few popular home device domain names as we expected, we found that

AS code name	Home access network related
A1	Yes
A2	Yes
A3	Yes
A4	Likely
A5	Yes
A6	Yes
A7	Likely
A8	Yes
A9	Yes
A10	Yes
A11	Yes
A12	Yes

TABLE II: AS code names (used in Fig. 5 and Fig. 6) of the top 12 WPAD query leak ASes in the U.S., accounting for 85% of total leak queries. We anonymize the AS names for privacy consideration.

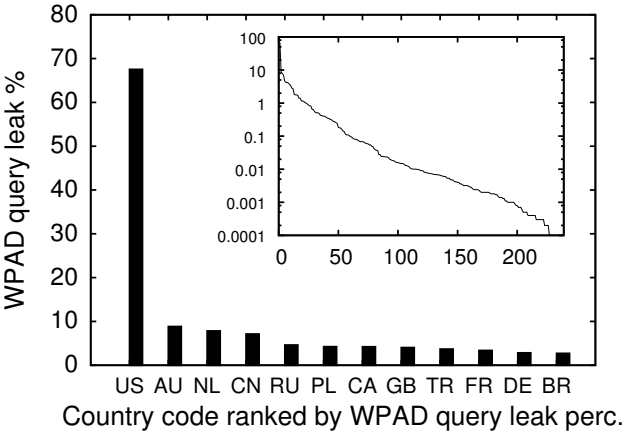

Fig. 4: Countries ranked by WPAD query leak percentage. The figure inset shows the complete probability distribution, illustrating the long tail.

the leaked queries have on average more than 10,000 different domain suffixes in these 12 ASes. For example, home access network AS A1 originated WPAD queries with more than 70,000 different domain suffixes, with the most popular one accounting only for 0.28% of all leaked queries. Moreover, we

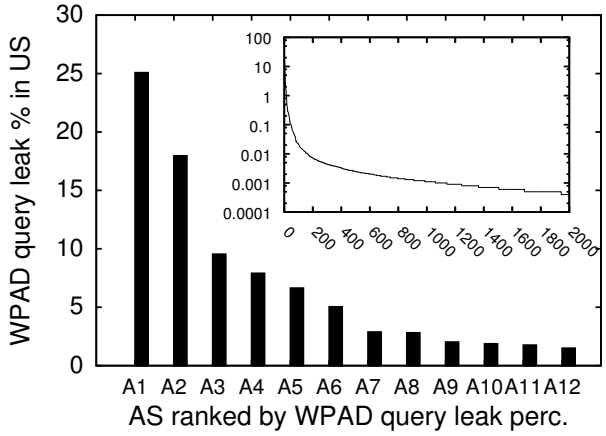

Fig. 5: ASes ranked by WPAD query leak percentage in US.

Domain suffix string	Query %	Home network related	Corporate network related
⟨defense contractor⟩ .master.	0.28	No	Unclear
corp.local.	0.26	No	Yes
⟨marketing⟩.local.	0.22	No	Yes
root.local.	0.21	Unclear	Unclear
⟨manufacture⟩.inc.	0.15	No	Yes
⟨town name⟩.local.	0.14	No	Yes
prod.dca.	0.13	No	Yes
⟨consulting⟩.local.	0.13	No	Yes
us.local.	0.13	Unclear	Unclear
⟨real estate⟩.local.	0.12	No	Yes
⟨computer⟩.lan.	0.11	No	Yes
⟨bank⟩.ubc.	0.11	No	Yes
datacenters.ww.	0.11	No	Yes
⟨marketing⟩.intraxa.	0.10	No	Yes
root.corp.	0.09	No	Yes

TABLE III: Top domain suffixes of the leaked WPAD queries in home access network AS A1. For privacy consideration, we anonymize some company or institution names with their business types in brackets.

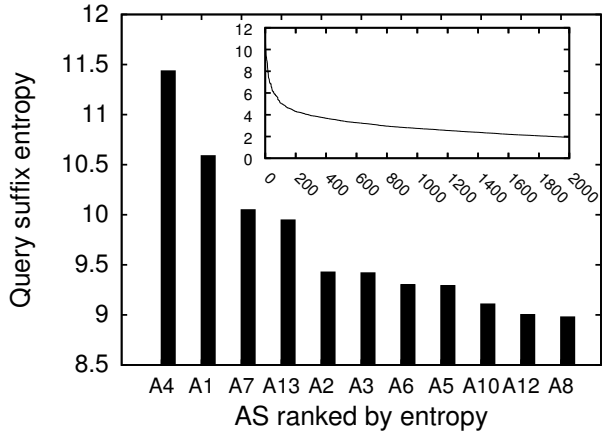

Fig. 6: ASes ranked by their domain suffix entropy scores. Home access networks with top leak query volume (Table II) are also high-entropy ASes. A13 is the only exception that did not appear in the top 12 WPAD query leak ASes.

manually classify the top domain suffixes and find that they are almost all corporate internal network suffixes instead of home device domains. Table III lists the top 15 leaked query domain suffixes from A1. As before, we obfuscate the details of the domain names for discretion. As shown, none are domains for home devices such as routers. Based on the labels, e.g., "corp", "inc", 12 of them are related to corporate internal networks that are unlikely to be hosted in home networks. This suggests that a potential cause of WPAD query leaks can be attributed to individuals using corporate devices on their home networks.

To further validate this cause, we measure the average daily domain query entropy of each leak source AS. The intuition is that home access networks with end-user machines using different internal network domain suffixes should have higher entropy due to the suffix diversity. In this analysis, we measure the daily query domain suffix entropy using equation $entropy(AS_i) = -\sum_{suf \in S} p_{suf} \ln p_{suf}$, where S is the set of distinct 2-level domain suffixes (e.g., `company.ntld` in Fig. 1) appearing in AS AS_i in a day, and p_{suf} is the query percentage of 2-level domain suffix $suf \in S$.

Fig. 6 shows the leak source ASes ranked by their average daily domain suffix entropy scores from January to July, 2015. As shown, the home access network ASes with top leak query volume are also high-entropy ASes. Moreover, the top 12 high leak volume ASes are all ranked top 15 in entropy scores out of over 2000 ASes in total, which supports our hypothesis. Thus, the major cause of the WPAD leaks is very likely using devices configured with internal domain names outside of internal networks, e.g., using corporate laptops at home.

3) Device-side Causes: From this cause, the major problem is on the device side: why does a corporate device still issue corporate internal WPAD queries when the device is actually not in the corporate network? In fact, with the support of DHCP, a device should be able to automatically update domain suffixes when the network changes. To find out the causes, we set up a local testbed to perform controlled experiments.

Experiment setup. We use VirtualBox to launch a virtual machine running different testing OSes on a host machine using NAT (Network Address Translation) configuration. In our experiments, we choose Mac OS X, Ubuntu 12.04, Windows XP, Vista, 7, 8, 8.1, and 10 as testing OSes.

The host machine can be connected to 3 different network environments. Two of them have 2 different domain suffixes configured in DHCP, which are automatically propagated to the host. The third environment does not have a domain suffix, which is created using the tethering feature of a smartphone on a cellular network. In our experiment, we switch the network of the host machine among the 3 environments to simulate network condition changes on the testing OSes, e.g., from corporate to home network.

Results. As summarized in Table IV, we find several common OS settings under which internal query leaks can happen even with automatic domain configuration from DHCP. The first case is setting the domain of a computer, which can be found in the control panel of Windows OSes. This configuration is recommended for businesses and schools, since it can remotely manage laptops they provide to their employees and students with their domain controller over VPN

OS configuration	Affected OSes
Set Windows PC domain	Windows XP/Vista/7/8/8.1/10
Hardcode domain search list	Windows XP/Vista/7/8/8.1/10, Mac OS X, Ubuntu 12.04
Change from a network with domain to a network without domain	Windows Vista

TABLE IV: Common OS configurations that can cause a device to mistakenly issue internal queries when the device is used outside internal network.

or Internet connection [14]. However, we find that once this is set, the OS keeps this domain name regardless of the DHCP domain configuration, and thus still issues internal WPAD queries even after the network has already changed.

The second case is about domain search list configuration, which can be accessed in the network setting panels of all OSes we tested. When a queried name is not considered fully-qualified [28], e.g., a dotless single label like `wpad`, the OS appends the domains in this search list one by one until obtaining a valid response. This search list can enable the OS to support both home network and corporate network by including both of their network domain suffixes. But if the corporate network domain suffixes are listed first, internal queries are tried first and thus leaked when outside the internal network. This cause has been discussed before in the web browsing context [20]; in contrast, in our experiment we study it for the WPAD proxy discovery process.

The third case is specific to Windows Vista, where we find that the domain is not unset when changing from a network with a configured domain to a network without a configured domain. This is likely a specific implementation flaw in Windows Vista, as all other OSes quickly change the domain setting to an empty string under the same condition. Due to this problem, corporate computers with Windows Vista leak internal queries when connected to a network without a configured domain, which can happen both at home and at public networks such as a café.

These results show that there exist common configurations in popular OSes that can mistakenly issue internal WPAD queries when the device is used outside corporate networks, causing internal query leaks. Note that these experiments are not intended to be exhaustive in finding all possible device-side causes, which is a rather difficult task. In fact, these identified causes might just be the tip of the iceberg, and merely patching them may only fix a small portion of the problem.

C. Result Summary and Highly-vulnerable ASes

Concluding from the characterization results above, we find that millions of vulnerable queries are leaked from internal networks every day, and the cause for the majority of the leaks is on the device side. Under common OS configurations, devices with popular OSes mistakenly keep internal domains even outside internal networks, and thus issue internal namespace WPAD queries. Once these queries are issued outside an internal network, the DNS resolvers have no idea where the local name servers are for these internal domains. Thus, they end up querying the DNS servers in the public namespace.

From our analysis above, we are also able to find 10 ASes with both highest query leak volume and query domain suffix entropy score in the U.S. as shown in Fig. 5 and Fig. 6. These ASes account for 81.2% of total WPAD query leaks in the U.S., and at the same time expose the largest variety of different victims. Thus, we consider them as the most vulnerable leak sources in our study. In the following sections, we will focus on these 10 ASes, especially the one with highest query leak volume, A1, to perform systematic assessment of the vulnerability status in the wild.

V. Attack Surface Quantification

Shown in the previous section, a large number of vulnerable WPAD queries are found in the public DNS namespace, many of which are already exploitable today. In this section we propose a candidate attack surface quantification method derived from the definition in §III-B, and evaluate its effectiveness.

A. Quantification Method

As defined in §III-B, the attack surface for a new gTLD is highly-vulnerable SLDs with two properties: (1) high persistence, and (2) high query volume. Because "high" query volume is a relative measure, we use query ratio, qr, as the metric for the high query volume property. For an SLD set S under a new gTLD $ntld$, we represent query ratio as $qr_{ntld}(S) = \frac{\sum_{sld \in S} Q_{sld.ntld}}{Q_{ntld}}$, where $Q_{sld.ntld}$ and Q_{ntld} are the number of leaked queries with domain $sld\,.\,ntld$, and with new gTLD $ntld$ respectively.

To find highly-vulnerable domains, our method is to first identify domains with high persistence. This is because a domain can be exploited as long as it is queried again for WPAD proxy discovery after domain registration. To quantify the level of persistence for a domain $sld_i\,.\,ntld$, we use period length p and persistence duration n to identify domains with leaked WPAD queries to the DNS root server in every p-day period for at least n days until the delegation of $ntld$. High persistence is reflected by a small p and large n, e.g., the domain has leaked queries every day for at least 1 year before the delegation of $ntld$. We use this as evidence indicating that the leakage may likely occur with some degree of frequency even after the delegation because of high persistence.

For a new gTLD $ntld$, given a certain p and n, we can find a set of SLDs under $ntld$, $S^{p,n}$, that meet this level of persistence in root WPAD NXD dataset, with a corresponding average query ratio value $qr_{ntld}(S^{p,n}) = \frac{\sum_{i=1}^{D} qr^i_{ntld}(S^{p,n})}{D}$. Here, $D = \lfloor \frac{n}{p} \rfloor$ is the number of p-day periods during which WPAD query leaks with domains in $S^{p,n}$ are observed, which we call *persistence period*. In this equation, $qr^i_{ntld}(S^{p,n})$ is the query ratio for the i-th period.

To meet the high query ratio property, we need to find the set $S^{p,n}$ with the highest $qr_{ntld}(S^{p,n})$ under a satisfiable persistence level defined by p and n. This is non-trivial as there are trade-offs between the choices of p, n and the query ratio value. For the period length, the smaller, the more persistent, but with a small p we may lose high query ratio domains with longer appearing periods. And for the persistence duration,

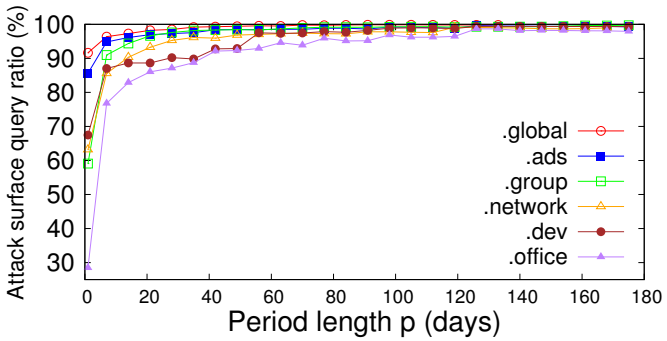

Fig. 7: Relationship of attack surface query ratio and period length p.

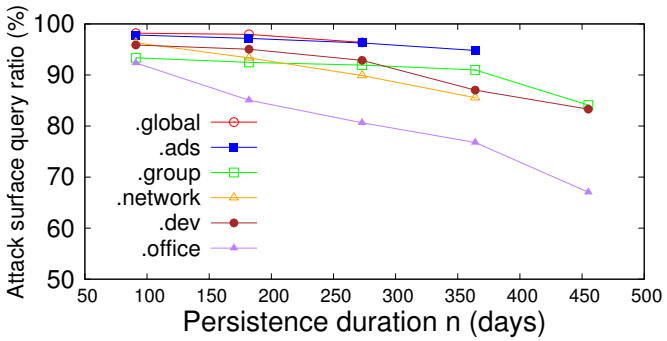

Fig. 8: Relationship of attack surface query ratio and persistence duration n. Since the 6 new gTLDs have different delegation dates, the data range for the curves are different.

the larger, the more persistent, but with a large n we may lose some recent high query ratio domains.

Fig. 7 and Fig. 8 show examples of these trade-offs using 6 delegated new gTLD strings with the highest leaked query percentage (according to Fig. 3). As shown, when p increases, the increase of query ratio slows down, and when n increases, the decrease of query ratio starts to drop more sharply. Thus, to balance the trade off, for a period length p, we stop increasing it to avoid sacrificing the persistence level, once the increase rate of qr reaches a limit, thr_p, indicating that we have already included enough high query ratio domains. For the persistence duration n, we also set such a limit, thr_n, and stop increasing persistence level once the decrease rate of qr exceeds this limit, indicating more sacrifice in the high query ratio property.

Algorithm 1 shows the pseudocode of our quantification method. For p, our method first tries $p = 1$, and then tries multiples of 7 days considering the weekly pattern of DNS queries, i.e., $p = 7(j-1)$ where $j = 2, 3, ...$ This process stops when $\triangle qr_{ntld}(S^{p,n})$ is less than thr_p, or $\lfloor \frac{n}{p} \rfloor$ is less than 2, which reaches the point of no periodicity. For n, our method tries multiples of 91 days, i.e., $N = 91i$ where $i = 1, 2, 3, ...$, until $\triangle qr_{ntld}(S^{p,n})$ is larger than $thres_n$, or the n is so large that it exhausts our 2-year dataset. We choose 91 days because it is roughly 3 months, which is considered the least persistence duration in this paper to avoid short-term domain query phenomena.

Algorithm 1 Attack surface quantification method

Input: Q_{ntld} (the set of daily leaked WPAD query domains for new gTLD $ntld$ in a victim AS), thr_p, thr_n
Output: Attack surface domain set S for new gTLD $ntld$

```
n_i = 91i, where i = 1, 2, 3, ...
p_1 = 1
p_j = 7(j − 1), where j = 2, 3, 4, ...
for i = 1.2, 3, ... do
    for j = 1, 2, 3, ... do
        Find domain set S^{p_j,n_i} from Q_ntld
        d^P_qr = qr_ntld(S^{p_j,n_i}) − qr_ntld(S^{p_{j−1},n_i})
        if d^P_qr ⩽ thr_p or ⌊n_i / p_{j+1}⌋ < 2 then
            break
        end if
    end for
    q_i = qr_ntld(S^{p_j,n_i})
    d^N_qr = q_{i−1} − q_i
    if d^N_qr > thr_n or n_{i+1} > |Q_ntld| then
        break
    end if
end for
return S^{p_j,n_i}
```

B. Evaluation

We implemented our attack surface quantification method, and applied to the 10 highly-vulnerable ASes using the root NXD WPAD dataset. In this section we use A1 as an example to show our results, because it was the top AS in both query leak volume and domain suffix entropy score, and the findings below also apply to the other 9 highly-vulnerable ASes.

In total, A1 presented queries in 255 out of the 738 new gTLDs delegated as of 2015/08/25. Among them, 19 new gTLDs only have leaked query data for 1 day, which are not enough to conclude their attack surface according to our definition of persistence. For the remaining 236 new gTLDs, our method is able to find attack surface domains for 204 (86.4%) of them, which are the ones accounting for 99.99% of total new gTLD WPAD query leaks in this AS.

Fig. 9 shows CDF of attack surface query ratio qr_{ntld}, for the 204 new gTLDs in TLD percentage and leaked WPAD query traffic percentage. As shown, for 185 (90.7%) of them, the attack surface query ratio qr output by our method are over 92.1%. These 185 new gTLDs account for 98.4% of total new gTLD WPAD query leaks in A1, showing that we are able to find domains meeting high query ratio property for new gTLDs that expose most vulnerabilities in a victim AS.

We also evaluate how well the attack surface output by our method can meet the high persistence property. As shown in Fig. 10, for 148 (72.5%) out of the 204 new gTLDs, which account for 98.8% of total new gTLD WPAD query leaks in this AS, their attack surface domains have periodical appearance for more than 4 periods ($D >= 4$). Thus, our method is also able to find domains meeting high persistence property for new gTLDs exposing most vulnerabilities.

AS code name	Attack surface domain characterization						Registration status (as of 2015/09/26)		
	Domain #	Domain query %	Distinct TLD #	# of TLDs have only 1 SLD	Distinct SLD #	# of SLD strings unique to 1 TLD	Reg. #	# of TLDs w/ reg.	# of TLDs w/ full reg.
A1	1185	97.4	204	109 (53.4%)	1122	1080 (96.3%)	129 (10.9%)	56 (27.5%)	18 (8.8%)
A2	486	97.0	122	75 (61.5%)	463	447 (96.5%)	49 (10.1%)	28 (23.0%)	10 (8.2%)
A3	747	97.7	154	91 (59.1%)	714	694 (91.2%)	68 (9.1%)	34 (22.1%)	16 (10.4%)
A4	3621	96.2	331	130 (39.3%)	3324	3145 (94.6%)	284 (7.8%)	79 (23.9%)	8 (2.4%)
A5	704	96.1	146	80 (54.8%)	673	653 (97.0%)	67 (9.5%)	35 (24.0%)	15 (10.3%)
A6	701	97.2	144	75 (52.1%)	668	646 (96.7%)	66 (9.4%)	31 (21.5%)	9 (6.3%)
A7	1751	95.7	230	117 (50.9%)	1633	1566 (95.9%)	123 (7.0%)	55 (23.9%)	17 (7.4%)
A8	457	97.6	113	74 (65.1%)	439	426 (97.0%)	43 (9.4%)	27 (23.9%)	12 (10.7%)
A10	254	96.8	73	42 (57.5%)	235	224 (95.3%)	28 (11.0%)	17 (23.3%)	8 (11.0%)
A12	255	95.5	70	44 (62.9%)	239	227 (95.0%)	33 (12.9%)	19 (27.1%)	14 (20.0%)
Union	8918	97.0	406	92 (22.7%)	7966	7447 (93.5%)	589 (6.6%)	123 (30.3%)	16 (3.9%)
Intersection	90	58.2	33	21 (63.6%)	80	73 (91.3%)	14 (15.6%)	9 (27.3%)	7 (21.2%)

TABLE V: Attack surface domain characteristics and registration status (as of 2015/09/26).

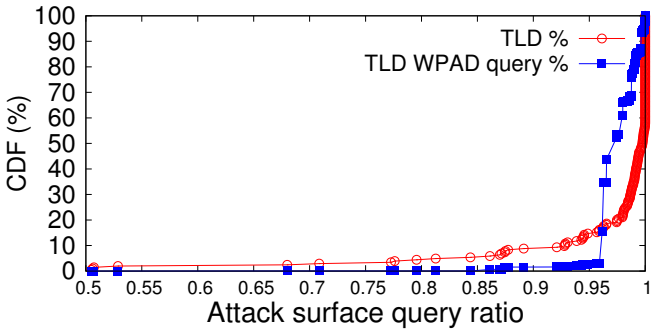

Fig. 9: CDF of attack surface query ratio in TLD percentage and TLD leaked WPAD query traffic percentage.

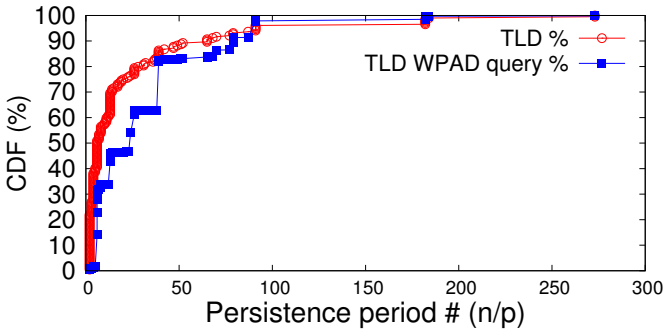

Fig. 10: CDF of attack surface period number in TLD percentage and TLD leaked WPAD query traffic percentage.

VI. Attack Surface and Exploit Status Characterization

With attack surface successfully computed, in this section we characterize their properties in the victim ASes, and also study their registration and exploit status in the wild.

A. Attack Surface Characterization

Finding 1. Among the 10 top vulnerable victim ASes, ASes operating open resolvers expose the largest attack surfaces. Column 2, 4, and 6 in Table V show the number of attack surface domains, distinct attack surface TLDs and SLDs for the 10 highly-vulnerable ASes discussed in §IV. As shown, A4 and A7, which both run open resolvers as discussed in §IV-B1, have significantly more attack surface domains, TLDs and SLDs than other victim ASes, even though their leaked WPAD query traffic is much less than some home access network ASes such as A1. This is likely because these popular open resolvers are used in all kinds of network environments and the exposed suffixes are more diverse compared to a single home access network AS (also shown in Fig. 6). This suggests that ASes running popular open resolvers should be the first priority for deploying AS-level defense.

Finding 2. In victim ASes, large fractions of leaked WPAD queries are for new gTLD domains defined to have high vulnerability (using our attack surface definition). Column 3 of Table V lists the percentage of leaked WPAD queries for the attack surface domains during their persistence periods in the 10 highly-vulnerable victim ASes. As shown, for all of these ASes, on average 96.7% of the leaked queries are in the HVDs, showing a high ratio of exploitability in the wild if these domains are registered.

Finding 3. For most of the new gTLDs, only very few SLDs are highly vulnerable. Fig. 11 shows the attack surface size distribution for new gTLDs with leaked queries from A1. In the figure, even though some new gTLDs can have very large attack surface, e.g., over 250 for `.office`, 184 (90.2%) of the 204 new gTLDs have fewer than 10 domains in their attack surface. This uneven distribution also holds for other highly-vulnerable victim ASes. As shown in column 5 of Table V, for 9 of the 10 ASes, more than half of the new gTLDs only have one domain in their attack surface. This indicates that for most new gTLD strings, the attack surface size is actually very small, and thus only a few domains need to be treated more carefully in registration.

Finding 4. Most SLD strings only appear in one new gTLD's attack surface. We then measure the popular SLD strings shown across the new gTLD attack surface in A1. From the result, the 5 most popular SLD strings are `us`, `corp`, `local`, `home`, and `net`, which are mostly generic ones. Out of the 204 distinct new gTLD string in A1, we find that the most popular SLD string, `.us`, is only shared by 7 new gTLDs' attack surface. As shown in column 7 of Table V, for all the 10 highly-vulnerable victim ASes, more than 90% SLD strings only appear in one new gTLD's attack surface in the victim AS. This suggests that if applying SLD reservation as

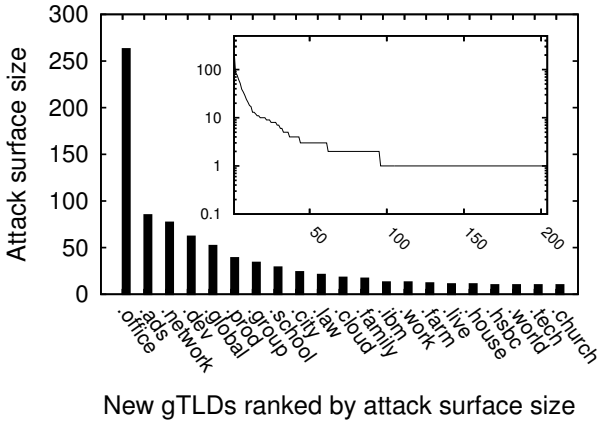

Fig. 11: Attack surface size distribution for new gTLDs delegated as of 2015/08/25.

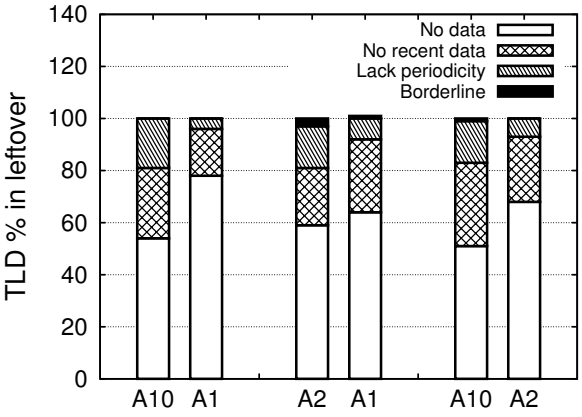

Fig. 12: Breakdown of new gTLDs in the leftover part in cross AS attack surface comparison.

a defense strategy, each new gTLD registry needs to identify its own SLD reservation list based on its WPAD query traffic patterns.

Finding 5. A large portion of the attack surface domains are victim AS unique. As shown in Table V, 8918 domains across 406 new gTLDs are in the union set of the attack surface of the 10 highly-vulnerable victim ASes, but only 90 (1%) of these domains are in the intersection set. Thus, very few attack surface domains are in common among different victim ASes. Moreover, 3689 (41.4%) of these domains are included in only 1 AS's attack surface. These results indicate that most attack surface domains are actually victim AS unique.

To understand why large numbers of domains are AS unique, we pick 3 home access network ASes in the highly-vulnerable AS set, and pair-wisely compare their attack surface domains. More specifically, for comparison between AS A_x's attack surface, S_{A_x}, and AS A_y's attack surface, S_{A_y}, we study the leaked query traffic in A_y for domains in $S_{A_x} - S_{A_y}$ and also leaked query traffic in A_x for domains in $S_{A_y} - S_{A_x}$. We classify the reason why these domains are not left out in the other AS's attack surface into 4 categories: *No data*, *No recent data*, *Lack periodicity*, *Borderline*. Category *No data* means that none of the domains' leaked queries are observed in the other AS in our 2-year root NXD WPAD dataset, and *No recent data* means none of such queries are observed in one month before the delegation of the corresponding new gTLDs. Category *Lack periodicity* means that the domain's queries appear in less than 50% of the days in 3 months before the delegation of the corresponding new gTLDs, which indicates that they are left out due to low persistence according to our attack surface definition. Category *Borderline* means that we could include them in the other AS's attack surface, but we left them out due to the balancing of persistence level and query ratio as discussed in §V-A.

The breakdown analysis result of the AS-unique attack surface domains is shown in Fig. 12. In the figure, we find that more than 80% of these domains are left out because they have no leaked queries for at least a month before the delegation of the corresponding new gTLDs, which can thus hardly be eligible to be considered as highly vulnerable according to our attack surface definition. For the other 20% domains, almost all of them lack periodicity, and only at most 3% of the domains are left out due to the balancing process in our quantification method. Thus, each victim AS indeed has a large portion of HVDs that are unique to it. This suggests that to deploy effective defense at the AS level, each victim AS, especially those highly-vulnerable ones, should customize its own domain filtering list.

B. Registration Status

Once these HVDs are registered, the actual exploitation can start at any time. Next, we use new gTLD zone files and WHOIS data to characterize the current registration status of these HVDs.

Finding 6. While for some new gTLDs their highly-vulnerable domains have already been fully registered, the overall registration status is still in the early stage. The last 3 columns in Table V include statistics of the registered HVDs as of 2015/09/26 for the 10 highly-vulnerable victim ASes, along with the intersection and union sets. As shown, all 10 victim ASes have some of these HVDs registered, but the registration percentages are in the range of 7% to 13%, which is not high. On the TLD level, approximately 22% to 28% of new gTLDs with attack surface in a victim AS have at least 1 attack surface domain already registered. For most victim ASes, around 10% of them have already had all of their attack surface domains registered, indicating that their attack windows are fully open. Recall that once an HVD is registered, the management of the underlying zones is delegated from the new gTLD registries to the domain registrants, and thus the WPAD name collision attack can be set up at any time outside of the new gTLD registries' control. Fortunately, our results show that even though some new gTLDs' attack surface domains in victim ASes have already been fully-registered, the overall registration has just started, and most HVDs are still under new gTLD registries' control.

Finding 7. For majority of the new gTLDs that have not been fully registered yet, the attack window is open-

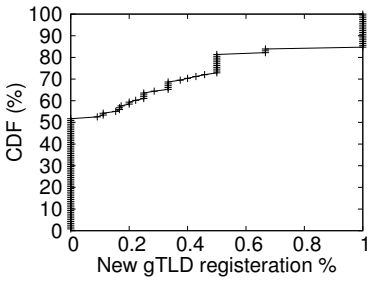

Fig. 13: Attack surface domain registration percentage for new gTLDs in the top vulnerable AS A1.

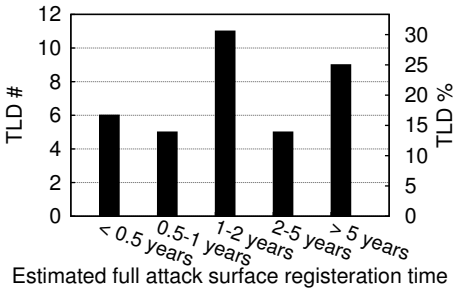

Fig. 14: Linear fitting results for the estimation time for a new gTLD to have all attack surface domains fully-registered.

Registrant email	# of total registered domains	# of registered ⟨legacy TLD⟩.⟨new gTLD⟩ domains
⟨email1⟩	19	19 (100%)
⟨email2⟩	7	7 (100%)
⟨email3⟩	2	2 (100%)
⟨email4⟩	16	10 (62.5%)
⟨email5⟩	19	9 (47.4%)
⟨email6⟩	7	3 (42.9%)

TABLE VI: Registration ratio of legacy gTLD string for some registrants, showing potential blind attack attempts. The email addresses are anonymized for privacy reason.

ing quickly. Besides a current snapshot of the vulnerability status, we also analyze the registration trend of these highly-vulnerable domains. In this analysis, we choose the top vulnerable AS, A1, and estimate how fast the attack surface domains for a new gTLD in this AS will be fully-registered. For most new gTLDs, we find that generally the total domain registration numbers increase linearly with time after a big increase at the beginning. Thus, we use a basic linear model to fit the attack surface domain registration trend for a new gTLD, and enumerate different starting dates until the average absolute error of the computed registered attack surface domain number is less than 0.5.

Using this method, we estimate the full registration time for the 38 new gTLDs in A1 which have at least one HVD registered (so that the analysis has input) but still not yet fully registered. Among these 38 new gTLDs, 2 new gTLDs' HVD registration numbers do not change in our zone file data set, and thus our method cannot perform linear fitting for them. For the other 36 new gTLDs, our method is able to find a linear curve with less than 0.5 average absolute error for the registered HVD number. In the fitting, 89.4% (272.1 days) of the available zone file data for a new gTLD are used on average. Fig. 14 shows the estimation results for these 36 new gTLDs. In the figure, 33% of them are likely to be fully registered in 1 year, and this percentage increases to 60% in 2 years. This is just a rough estimation, but does indicate that even though currently most of the new gTLDs' attack surface domains are not yet fully-registered, their attack surface is being registered quickly, suggesting that immediate precautions need to be applied to prevent these vulnerabilities from further expansion.

Finding 8. We did not find strong evidence of adversaries actively registering attack surface domains, but do observe potential blind attack registrations. Given that many of these highly-vulnerable domains have been registered, we next analyze whether some registrants are aware of these highly-vulnerable domains and thus deliberately register them for the WPAD name collision attacks. In this analysis, we also choose the top vulnerable AS, A1. For each new gTLD in this AS, we use 2 time series data each day: the registered number of attack surface domains, and the registered number of other domains shown in root NXD WPAD data for a new gTLD before delegation. For new gTLDs with some of their attack surface domains registered, we compute the Pearson product-moment correlation coefficient, and find an average correlation score of 0.76, showing a very strong correlation. This means that it is just as likely to register attack surface domains as other domains appearing in the root NXD WPAD data, suggesting that there are no strong evidence of adversaries actively registering these HVDs.

However, interestingly, we observed registrations that may be used for malicious purposes, such as name collision attacks. More specifically, we find that there are a number of registrants specifically targeting the registration of legacy TLD strings, e.g., `com`, `net`, etc. as SLDs, under new gTLDs. In this analysis, we refer to the strings of TLDs delegated before the new gTLD program as legacy TLD strings, which include gTLDs such as `.com` and country-code TLDs such as `.uk`. We obtain legacy TLD string list by comparing the TLD list on IANA's root zone database webpage [17], and the new gTLD list on ICANN's website [16]. Using the new gTLD WHOIS dataset, we identify a list of registrants having a very high registration ratio of legacy TLD strings under new gTLDs, which is shown in Table VI. For example, one registrant with email ⟨email1⟩[1] has registered 19 domains as of 2015/09/26, which all contain `com`, `edu`, `gov`, and `org` as SLD strings among over 10 new gTLDs. In our new gTLD WHOIS dataset, only less than 20% of the registrants (identified by email addresses) registered more than 1 domain. Among the 20%, majority of them use corporate email addresses, and the registration targets are usually product related domains, e.g., a registrant with a company email registered 351 domains with a SLD that is the name of their product. The registration behavior in Table VI are very unlikely for brand protection, since (1) they used individual email addresses, and (2) they targeted legacy TLD strings instead of product names, which in combination make such behavior suspicious. One likely reason is that these registrants are trying to exploit one of the earliest reported name collision vulnerability due to an old BIND resolver bug [22]. These results suggest that potential adversaries do exist who are fully aware of the name collision vulnerability. Fortunately at this point, they probably just have

[1]We anonymize the email addresses of the registrants for privacy considerations.

not found an effective way of identifying highly-vulnerable domains.

C. Exploit Status

For the registered HVDs, we are also wondering whether the domain registrants have already started exploiting the vulnerability by serving a valid MitM proxy. Since the domain registrants have full control of the zone after the registration, it is not possible for a 3rd party like us to get an accurate list of subdomains under these HVDs. In our experiment, we use the list of query names in previous WPAD queries to these domains before the delegation of their TLDs as a guess of potential attack subdomains. For each WPAD domain query name $qname$ in the list, we issue request using `wget http://`$qname$`/wpad.dat` and check whether we can get a valid proxy configuration file. Note that even with this list, this experiment can still have false negatives since our probing queries can be intentionally filtered by attackers for only targeted attacks (i.e., only resolve the queries from certain AS, IP, etc.) in order to prevent external detection.

We perform such probing several times for all the domains in the union set of the 10 victim ASes' attack surface domains, but are not able to find valid proxy files. This indicates that the registrants of the highly-vulnerable domains may not realize this attack vector yet, implying that now would be a good time to start deploying remediation strategies, which is discussed in the next section.

VII. Remediation Strategy Discussion

Considering that the overall vulnerability registration and exploitation are still in the early stage, it presents an opportunity to proactively mitigate this attack. In this section, we discuss the potential remediation strategies by 3 different parties involved in the DNS ecosystem: new gTLD registries, victim ASes, and end users.

Table VII summarizes the results for the estimated effectiveness and deployment difficulties for these remediation strategies. In contrast to the previous sections, which focused on the 10 highly-vulnerable ASes in the U.S., here we consider estimations based on the attack surface quantification using all ASes with leaked WPAD queries in our 2-year root WPAD NXD dataset. This allows us to present more accurate global vulnerability reduction percentages and deployment numbers.

New gTLD registry level remediation. To reduce the chance of an attack, the new gTLD registries, especially the ones we find to have large attack surface (shown in Fig. 11), need to ensure that these HVDs are not registered, or treat them more carefully and propose policies to scrutinize their registrations. A naïve approach is to reserve the registrations of all domains seen in NXD traffic. However, according to the experience of deploying the block list in ICANN's Alternate Path to Delegation (APD) [4], merely using 2 days of root NXD data for 3 years, each new gTLD registry needs to block 7449.3 domains on average, and 7 new gTLDs need to block over 100,000 domains. Preventing such a large number of them from being registered, especially those popular ones, is in conflict with the original goal of providing more registration choices, and also hurts new gTLD registries' revenue model. ICANN now changes the policy to allowing their registrations after a 90-day "controlled interruption" period instead of blocking them forever [3].

Level	Remediation strategy	Effectiveness	Deploy #
New gTLD registry	Scrutinize the registration of the union set of highly-vulnerable domains	97.4%	494
Victim AS	Filter the intersection set of highly-vulnerable domains	36.4%	11305
	Filter AS-specific highly-vulnerable domains	97.4%	
	Filter responses w/ public IP	Not evaluated	
End user	Disable WPAD service (if not used in internal networks)	Not evaluated	$>$ 6.6 million
	Update OS, no hardcoding	$\sim$100.0% (in theory)	
	Filter device-level leaks		

TABLE VII: Effectiveness and deploy number estimation for remediation strategy at new gTLD registry, victim AS, and end user levels. "Not evaluated" means that we cannot evaluate its effectiveness using current dataset.

According to our attack surface characterization, for most of the new gTLDs, relatively few SLD are highly vulnerable to the WPAD name collision attack and need scrutinized registration. For example, for `.network`, 96% of its domains in NXD traffic have very low volume and/or low persistence of WPAD queries. This is why a general-purpose block list is counterproductive, as opposed to per-SLD and per-TLD analysis performed in this paper. Thus, the attack surface defined and quantified in this paper offers a cost-effective way of deploying new gTLD registry level domain registration scrutinization. With the attack surface quantification results for all victim ASes, we take the union of the attack surface domains, and find that in total 494 new gTLDs among the 738 ones delegated before 2015/08/25 have HVDs. If all of them have registration scrutinization, 97.4% of the global leaked WPAD queries in our dataset can be protected. Consistent with our findings in §VI, most of the new gTLDs have only very few HVDs which need protection – among the 494 new gTLDs, 302 (61.3%) of them have less than 10 HVDs. Thus, for majority of new gTLD registries, this defense can be deployed with very little sacrifice of the business revenue while still being highly effective.

Considering that having all 494 new gTLD registries agreeing on the deployment may be difficult in practice, we also evaluate the effectiveness of a partial deployment. In this analysis, we rank the 494 new gTLDs by the protected leaked WPAD query percentages if they deploy scrutinized registration of HVDs, and the CDF is shown in Fig. 15. As shown, deployment at only the top 18 (3.6%) new gTLDs can already protect 80% of the leaked WPAD query globally. Thus, in the deployment, a more feasible and also very effective strategy is to start with the most important 20–40 new gTLDs.

Victim AS level remediation. As shown in §IV, majority of the leaked WPAD queries come from a few home access network ASes. In addition to new gTLD registry level defense,

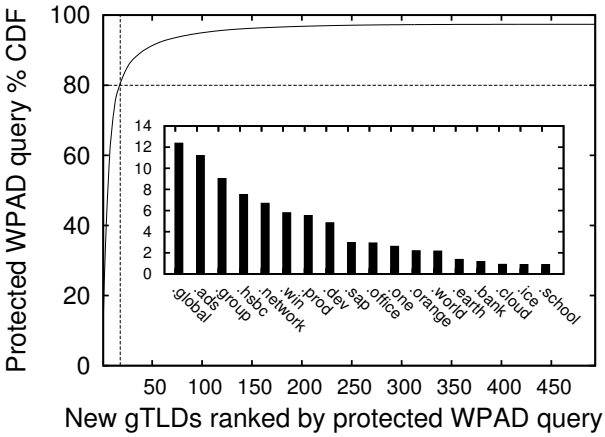

Fig. 15: Protected leaked WPAD query percentage CDF for partial deployment of new gTLD registry level defense. The figure inset lists the top 18 new gTLDs which can protect 80% of total leaked queries if the defense is deployed.

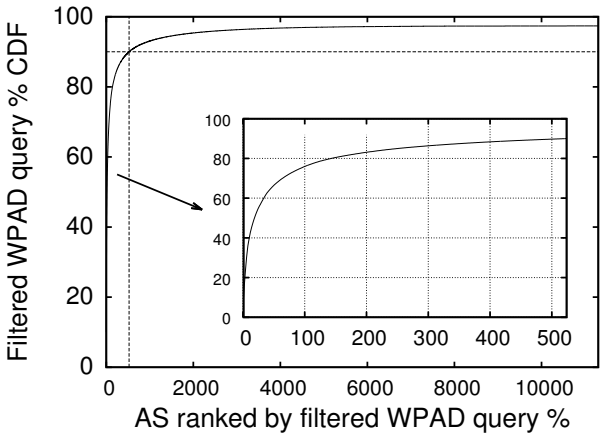

Fig. 16: Filtered leaked WPAD query percentage CDF for partial deployment of AS level defense. The figure inset shows the CDF for the top 524 ASes.

another direction is to prevent their leaks at the victim AS level. Each victim AS can distribute a black list of vulnerable domains to their DNS resolvers, and filter the queries to these domains before sending them to the public namespace. To create such a list for all ASes, one quick approach is to find the common vulnerable domains using the intersection set of the attack surface domains for the victim ASes. We estimate the effectiveness of this approach using the HVD intersection set for 10 highly-vulnerable ASes, which contains 90 domains as shown in Table V. We find that if all ASes adopt this black list, approximately 36.4% of the leaked WPAD queries globally in our 2-year dataset can be filtered. Thus, even though the creation of the black list is convenient without AS-specific customizations, this approach has limited effectiveness, mainly because many HVDs are AS-specific as characterized in §VI.

To increase the effectiveness, each victim AS should customize its black lists based on their own query traffic patterns. This can be enabled by DNS traffic monitoring and filtering in the recently-proposed name collision risk management framework [26]. One candidate approach to create such list is to use the attack surface quantification method proposed in §V based on NXD query data, which can be obtained either by collecting DNS queries on their own, or collaborating with DNS root server operators. The deployment locations are the ASes with HVDs, including 11,305 ASes globally according to our quantification results. If every AS deploys this, it is capable of filtering 97.4% of the leaked WPAD queries globally in our dataset. Compared to the new gTLD registry level defense, this approach can achieve the same level of high effectiveness, but may have higher deployment challenges due to significantly more deployment locations. Thus, we also evaluate partial deployment strategy, shown in Fig. 16. In this figure, the X-axis is the 11,305 victim ASes ranked by their leaked WPAD query percentages. As shown, deploying at the top 143 (1.2%) ASes can effectively filter more than 80% of the leaked queries. Thus, similar to new gTLD registry level defense, it is not entirely necessary to cover all 11,305 victim ASes, and targeting the top 1–5% ASes can already achieve a relatively high level of effectiveness.

Victim AS level filtering can also be IP based. In the WPAD discovery process, the leaked WPAD queries are intended to get internal proxy server IP addresses, while in the WPAD name collision attack the attacker needs to return public proxy IP addresses. Thus, victim AS resolvers can prevent the attack by filtering the DNS responses with public IP addresses. The effectiveness of this approach cannot be evaluated using our root NXD dataset, which is left as future work.

End user level remediation. As shown in §IV, the major cause of the WPAD query leaks is using devices with internal domains outside of the internal network. Thus, to fundamentally solve this problem, this unintended client-side behavior needs to be fixed. If WPAD proxy discovery service is not actually used in the internal network, we suggest that the local network administrator, e.g., the IT department in a company, disable this feature in the supported browsers and OSes (Table I) during corporate device setup process. To more efficiently enforce this policy without the need of enumerating the configurations of all installed browsers or other related software, the administrator can change OS-level domain name mapping files such as `/etc/hosts` to map all permutations of WPAD URLs within the internal namespace to `127.0.0.1`. In this paper, the effectiveness of this approach is not evaluated since it is difficult to measure the amount of leaked WPAD queries belonging to local networks that do not use WPAD service internally.

For the corporate devices depending on WPAD for internal network proxy discovery, the WPAD feature in OSes and browsers still needs to be enabled. To prevent WPAD query leaks for these devices, leveraging our insights of the device-side causes found in §IV, companies or other entities with internal domains need to stop hardcoding the internal domain search list on their devices. If Windows OS is used, they need

to stop setting the Windows PC domain, and also upgrade their OSes. As we mentioned before, these causes may just be the tip of the iceberg, and there might exist plenty of other causes under different conditions. Moreover, considering the large variety of software on the device, new causes, for example domain hardcoding behavior in certain applications, can be created at any point in the future. Thus, these are only short-term solutions and not future proof.

As a long-term solution, we propose to design an OS-level daemon which can filter queries based on the network environment. This daemon is a background process which intercepts DNS queries issued by all applications on the device, and can correctly tell and filter queries with domains not belonging to current network environment. In order to distinguish unintended queries, it tracks the network environment at each network status change, and stores a list of intended domains suffixes for each network environment, either by learning from DHCP configuration messages, or directly being configured by the user. To realize this approach, there are still some design challenges, for example how to accurately tell network environments apart when they use same IP address prefixes, which we leave as future work.

For the short-term and long-term device-side solution above, in theory they can fundamentally solve the problem; thus, we consider its maximum effectiveness as 100% in Table VII. However, the downside is that it is extremely difficult to reach and apply these solutions to all end user devices, which is estimated to have at least 6.6 million deployment points using the number of distinct $\langle IP, sld.tld \rangle$ pairs in our 2-year root NXD WPAD dataset, where IP is the resolver IP sending WPAD queries, and $sld.tld$ is the WPAD query domain. This is only a lower bound estimation as there might be more than one user device with domain $sld.tld$ behind a resolver, but it is already at least 2 orders of magnitude larger than the new gTLD registry and victim AS level defenses described above.

To help facilitate the deployment process, OSes and browsers can displaying warning messages when detecting potential name collision risks. For example, if the issued WPAD query is leaked to the public namespace, the response will include a special IP address, `127.0.53.53`, during the 90-day "controlled interruption" period [3]. Browsers and OSes can thus leverage this to display risk warnings and recommend the users to consult their IT department immediately to resolve the problem. Note that the 90-day "controlled interruption" period [3] was ineffective to mitigate such issue since the victim machines automatically perform the vulnerable operations even without user awareness [33]. With more support from OS and browser sides, end users can be better notified of the imminent threat to help with the mitigation progress.

To summarize, no single defense approach discussed here can easily solve the problem. To maximize the chance of preventing the attack in practice, the best choice would be using these approaches jointly. Considering the serious and disseminated nature of this vulnerability as shown in this paper, actions need to be taken as soon as possible.

VIII. Related Work

DNS spoofing attacks. Like the WPAD name collision attack studied in this paper, some previous DNS spoofing attacks also try to deceive victims using malicious DNS response. One attack category assumes that the attacker is MitM and thus replies forged response when observing a query. This can be achieved through attacking the network configurations of the victim devices. For example, prior work [2], [35] show that scripts on web pages can change home routers' DNS configurations and point the client resolver IP to attacker's servers. Another category of attacks assumes that the attacker is off path. One such example is DNS cache poisoning attack [29], [34], which corrupts the resolver's cache with spoofed DNS responses, causing all downstream devices to be redirected to the attacker's IP addresses. These previous attacks exist since the victim cannot determine whether the received DNS responses are legitimate or manipulated, which can be solved by DNSSEC protocol [18], [19]. Compared to them, the attacker in the WPAD name collision attack is actually authoritative for the request domains. This means that she can legitimately give malicious response and launch MitM attack without the need of spoofing, making it exploitable even if DNSSEC is used.

New TLD delegation study. The addition of new gTLDs into the DNS root zone usually requires considerable debate about the extent to which new TLDs will actually serve a real need. Before the New gTLD Program, the growth of gTLD set maintained a very slow and steady rate. Some previous work studied the impact of certain early gTLD delegation, e.g., for `.biz` [25] and `.xxx` [24], and recently Halvorson et al. perform the first study targeting the New gTLD Program [23]. These studies mostly focus on characterizing the registration intent, and in comparison, our work targets the security problem exposed by the new gTLD delegation.

Name collision from new gTLD delegation. Before our work, concerns from the domain name industry have already been raised about potential name collision problem from new gTLD delegation [30]. Several studies have measured the leaked DNS queries to the DNS root servers and shown the potential risks of information leakage, denial of service, and MitM attack [31], [33]. The discussions resulted in a name collision management framework from ICANN in 2013 [4], which allows the majority of new gTLD strings to be delegated by following an Alternate Path to Delegation (APD). In APD, the new gTLD registries are required to block large numbers of high-risk SLDs according to measurement of DITL (Day in the Life of the Internet) dataset. Later on in 2014 a new framework allows releasing these blocked names after a 90-day period called "controlled interruption" for testing and resolving name collision problem [3]. However, previous studies have shown that the block list is ineffective due to the statistical limitation of DITL dataset [36]. In addition, the controlled interruption period is unlikely to change anything for problems similar to the WPAD name collision attack, since the victim machines automatically perform the vulnerable operations even without

user awareness [33]. This indicates the lack of a systematic approach to understand and find effective solutions for the newly-exposed name collision problem. Our work uses in-depth cause analysis and attack surface quantification to fill this critical gap.

IX. Conclusion

In this work, we perform a systematic study of the underlying problem cause and the vulnerability status for WPAD name collision attack in the new gTLD era. We first characterize the severity of the problem, and uncover that the major cause of the fundamental leakage problem is very likely devices used in their non-intended network, such as work laptops at home. Then, using a candidate attack surface definition and a quantification method, we systematically assess the vulnerability of the attack in the wild. We find that even though some attack surface domains have already been registered, the overall registration and exploitation status are still in the early stage, indicating that proactive protection strategies are still feasible. Based on these insights, we discuss remediation strategies at the new gTLD registry, AS, and end user levels, and estimate their effectiveness and deployment difficulties. Our work demonstrates the importance of addressing known security vulnerabilities, which might become more exploitable as assumptions change. This work also serves as the first in-depth study of one type of name collision problem in the new gTLD era, hopefully inspiring other follow-up studies.

Acknowledgments

We would like to thank Danny McPherson, Nick Feamster, Andy Simpson, Yannis Labrou, Aziz Mohaisen, Shumon Huque, Duane Wessels, Burt Kaliski, Yunhan Jack Jia, our shepherd, Sam King, and the anonymous reviewers for providing valuable feedback on our work. The University of Michigan authors were supported in part by the National Science Foundation under grants CNS-1345226, CNS-1318306, and CNS-1526455, as well as by the Office of Naval Research under grant N00014-14-1-0440.

References

[1] BestWhois service. https://www.whoisxmlapi.com/terms-of-service.php.
[2] Home routers come under attack from new DNS redirection tool. http://www.enyo.de/fw/notes/the-great-corp-renaming.html.
[3] ICANN: Mitigating the Risk of DNS Namespace Collisions Phase One. https://www.icann.org/news/announcement-2-2014-06-10-en.
[4] ICANN: Proposal to Mitigate Name Collision Risks. https://www.icann.org/public-comments/name-collision-2013-08-05-en.
[5] Naming an internal (private) Active Directory LAN. http://arstechnica.com/civis/viewtopic.php?f=17&t=394734.
[6] Root server distribution. http://root-servers.org.
[7] Second Level Domain (SLD). http://icannwiki.com/SLD.
[8] The DNS Operations, Analysis, and Research Center (DNS-OARC). https://www.dns-oarc.net/.
[9] Top Level Domain (TLD). http://icannwiki.com/TLD.
[10] WHOIS database. http://whois.icann.org/en.
[11] Setting up Web Proxy Autodiscovery Protocol (WPAD) using DNS. http://tektab.com/2012/09/26/setting-up-web-proxy-autodiscovery-protocol-wpad-using-dns, 2012.
[12] The New gTLD Program. https://newgtlds.icann.org/en/about/program, 2013.
[13] Use a Custom Tld for Local Development. http://blog.bfontaine.net/2013/08/26/use-a-custom-tld-for-local-development, 2013.
[14] What is a Windows Domain and How Does It Affect My PC? http://www.howtogeek.com/194069/what-is-a-windows-domain-and-how-does-it-affect-my-pc, 2014.
[15] Centralized Zone Data Service. https://czds.icann.org/en, 2015.
[16] New delegated TLD strings. http://newgtlds.icann.org/en/program-status/delegated-strings, 2015.
[17] Root Zone Database. http://www.iana.org/domains/root/db, 2015.
[18] R. Arends, R. Austein, M. Larson, D. Massey, and S. Rose. Protocol Modifications for the DNS Security Extensions. rfc4035, 2005.
[19] R. Arends, R. Austein, M. Larson, D. Massey, and S. Rose. Resource Records for the DNS Security Extensions. rfc4034, 2005.
[20] C. Deccio. Whats in a Name (Collision)? Modeling and Quantifying Collision Potential. In *Workshop and Prize on Root Causes and Mitigation of Name Collisions (WPNC)*, 2014.
[21] D. Eastlake 3rd and A. Panitz. Reserved Top Level DNS Names. rfc2606, 1999.
[22] Gavron, Ehud. A Security Problem and Proposed Correction With Widely Deployed DNS Software. rfc1535, 1993.
[23] T. Halvorson, M. F. Der, I. Foster, S. Savage, L. K. Saul, and G. M. Voelker. From .academy to .zone: An Analysis of the New TLD Land Rush. In *ACM IMC*, 2015.
[24] T. Halvorson, K. Levchenko, S. Savage, and G. M. Voelker. XXXtortion? Inferring Registration Intent in the .XXX TLD. In *ACM WWW*, 2014.
[25] T. Halvorson, J. Szurdi, G. Maier, M. Felegyhazi, C. Kreibich, N. Weaver, K. Levchenko, and V. Paxson. The BIZ Top-Level Domain: Ten Years Later. In *Passive and Active Measurement*, 2012.
[26] B. S. Kaliski Jr. and A. Mankin. United States Patent Application 20150256424: Name Collision Risk Manager. http://www.freepatentsonline.com/y2015/0256424.html.
[27] P. Mockapetris and K. J. Dunlap. Development of the Domain Name System. In *SIGCOMM '88*, 1988.
[28] Mockapetris, Paul. Domain Names Implementation and Specification. rfc1035, 2004.
[29] B. Muller. Whitepaper: Improved DNS Spoofing Using Node Redelegation. `https://www.sec-consult.com/fxdata/seccons/prod/downloads/whitepaper-dns-node-redelegation.pdf`.
[30] E. Osterweil and D. McPherson. New gTLD Security and Stability Considerations. Technical Report 1130007 version 1, 2013. http://techreports.verisignlabs.com/docs/tr-1160018-1.pdf.
[31] E. Osterweil, M. Thomas, A. Simpson, and D. McPherson. New gTLD Security, Stability, Resiliency Update: Exploratory Consumer Impact Analysis. Technical Report 1130008 version 1, 2013. http://techreports.verisignlabs.com/docs/tr-1130008-1.pdf.
[32] E. Osterweil and L. Zhang. Interadministrative Challenges in Managing DNSKEYs. *IEEE Security and Privacy*, 7(5):44–51, 2009.
[33] A. Simpson. Detecting Search Lists in Authoritative DNS. In *Workshop and Prize on Root Causes and Mitigation of Name Collisions (WPNC)*, 2014.
[34] S. Son and V. Shmatikov. The Hitchhiker's Guide to DNS Cache Poisoning. In *Security and Privacy in Communication Networks*, pages 466–483. Springer, 2010.
[35] S. Stamm, Z. Ramzan, and M. Jakobsson. Drive-by pharming. In *Information and Communications Security*, pages 495–506. Springer, 2007.
[36] M. Thomas, Y. Labrou, and A. Simpson. The Effectiveness of Block Lists to Prevent Collisions. In *Workshop and Prize on Root Causes and Mitigation of Name Collisions (WPNC)*, 2014.