Surveying DNS Wildcard Usage Among the Good, the Bad, and the Ugly

Andrew Kalafut, Minaxi Gupta, Pairoj Rattadilok, and Pragneshkumar Patel

School of Informatics and Computing, Indiana University
{akalafut,minaxi,prattadi,patel27}@cs.indiana.edu

Abstract. A DNS wildcard can be used to point arbitrary requests for host names within a domain to a specific host name or IP address. Wildcards offer administrators the convenience of not having to change DNS entries when host names change. However, we are not aware of any work that documents how wildcards are used in practice. Such a study is particularly important now, because Internet miscreants are starting to exploit DNS wildcards for convenience and possibly for evading blacklists based on exact host names. In this paper, we study the prevalence and uses of wildcards among the good, bad, and ugly domains in the Internet. We find that wildcards are in extensive use among businesses that monetize unregistered domains, domains hosted by large web-hosting providers, blogging sites, and websites connected to scam, phishing, and malware.

Key words: DNS, Wildcard, Security

1 Introduction

The Domain Name System (DNS) [16] serves the basic purpose of translating human-readable host names into the IP addresses. While conceptually a simple mapping, the DNS is complex in reality. Several record types exist for different types of mappings, and several features exist to improve convenience and functionality beyond this basic description. One such feature, which we examine in this paper, is *wildcards*.

Wildcards are one of the original features of DNS, defined in the original standard. The role of wildcards in DNS is a many to one mapping, allowing all names within a single domain or subdomain to map to a single value. This can be used for example to map all host names in a domain to a single IP address, or to assign a single DNS or mail server to all possible subdomains of a domain. In both of these and other similar cases, the single catch-all wildcard record saves the DNS administrator from having to maintain many different records that all return the same value.

Despite their usefulness, very little is known about who uses wildcards, and for what purposes. There are signs that Internet miscreants have discovered the convenience of wildcards. Recently, Netcraft released two advisories that point to the use of wildcards in setting up phishing campaigns [15,19]. Wildcards may be attractive to miscreants because they allow mapping multiple host names in their campaigns to the same IP address, for example. This can be useful in evading host name based blacklists with minimal effort. Given this, understanding the use of wildcards becomes even more important.

In this paper, we undertake the first systematic study to investigate the use of wildcards in the Internet. We specifically work towards two goals. Our primary goal is to survey wildcard usage among good, bad, and ugly domains in the Internet. Toward this goal, we query approximately 8 million domains for wildcard entries in the four most popular DNS record types. Our second goal is to investigate if malicious uses of wildcards can be differentiated from their benign uses. The ability to do so may be helpful in identifying and effectively blacklisting malicious domains.

Working towards these goals, we arrive at the following key results:

- **Prevalence:** We find that a surprisingly large percentage of Internet domains use wildcards. Specifically, 25-75% of domains in various data sets use wildcards, making this a much more popular DNS feature than one would expect.
- **Type:** An overwhelming majority of domains using wildcards use them in their address records, which map arbitrary host names to a IP address.
- **Uses:** Prominent users of wildcards include domain-parking businesses that wish to monetize unregistered domains and subdomains, web-hosting companies, and blogging and social-networking sites.
- **Malicious sites:** Malicious sites also make extensive use of wildcards, with spammers leading the pack with 75% of the scam-related domains in our data wildcarded. We also find that Google knows more host names matching wildcards in malicious domains than are in our data sets, implying that the coverage of blacklists could be improved by including wildcard entries.
- **Distinguishing malicious uses:** Our preliminary investigation shows that IP addresses contained in wildcard records typically spread across many ASes. Additionally, they tend to have lower TTLs than wildcards for benign purposes. These features can be used to differentiate wildcard usage among malicious domains, particularly those associated with spam, from the benign ones

The rest of this paper proceeds as follows: Section 2 presents background on the syntax and behavior of DNS wildcards. The data we use throughout the paper is described in Section 3. We discuss the prevalence of wildcards in Section 4. We examine what they are being used for in Section 5. Section 6 explores differences between wildcards used for malicious purposes and others. We discuss related work in Section 7 and conclude in Section 8.

2 Background

The primary goal of DNS is to translate host names into IP addresses. The most popular types of host names resolved are mail servers, DNS servers (also known as name servers) and all other types of servers, including web servers. Mail and

DNS servers have dedicated DNS record types, MX and NS respectively, that map the queries for those servers to host names. These host names are then mapped to IP addresses through A records[1]. Web and other kinds of servers do not have dedicated types of DNS records and a mapping between their exact name and the corresponding IP addresses is accomplished directly through the A records. A fourth popular DNS record type is the CNAME record, which aliases a host name to another host name. In total, 59 DNS record types are defined as of now but only 42 of these are in wide use [9]. Figure 1(a) shows an example of DNS provisioning for a domain with one MX, one NS, and three A records.

www.foo.com	A	129.79.245.53
foo.com	MX	mail.foo.com
foo.com	NS	ns1.foo.com
mail.foo.com	A	129.79.247.191
ns1.foo.com	A	129.79.247.191

(a) without wildcards

www.foo.com	A	129.79.245.53
foo.com	MX	mail.foo.com
foo.com	NS	ns1.foo.com
*.foo.com	A	129.79.247.191

(b) with wildcards

Fig. 1. Example of DNS provisioning of a domain with and without wildcards

Wildcards in DNS were first defined in RFCs 1034 [16]. Later, RFC 4592 [10] updated and clarified the specification, providing more details and examples of intended behavior, and the interactions of wildcards with specific record types. A wildcard record is a DNS record of any type with a minor change to the left hand side of the record. In a wildcarded DNS record, instead of the name being an exact host name, its least significant (leftmost) label in the name consists of a single asterisk character, as shown in Figure 1(b). Conceptually, the asterisk matches one or more labels at the left end of the DNS name. In this example, *.foo.com is being used in place of mail.foo.com and ns1.foo.com. When a DNS query is made for mail.foo.com, seeing no match, the server will return results for *.foo.com, substituting mail for the *. Specific records override the wildcard records. Since the record for www.foo.com is still present, the wildcard would not be considered when responding to a query for this host name.

The client receiving a DNS response can not directly tell if the response was generated from a wildcard record or not; their use is transparent to the client systems. If a query for host name name.foo.com were matched from the wildcard record *.foo.com, the name on the record returned in the response will still be name.foo.com instead of *.foo.com as it is stored on the DNS server. We can however still tell if a wildcard is in use by directly querying for the wildcard name, in this case, *.foo.com. Since the wildcard record is the only one that would match such a query, if a response is given to such a query, it would let us know a wildcard record is present. Note that wildcard matches only work in one direction. Although the query for *.foo.com looks like the client has a wildcard

[1] AAAA records are used to map host names to IPv6 addresses.

in the query, it will only match an explicit wildcard record, not an arbitrary name in `foo.com` on the server.

3 Data Sets

Our goal is to study DNS wildcard usage in three contexts: domains judged as worthwhile or useful by Internet users (*the good*), domains from several blacklists (*the bad*), and a large general collection of domains, including both good and bad domains (*the ugly*). Table 1 shows an overview of the data sets.

Table 1. Overview of the data sets

	`DMOZ`	`ZONE_FILES`	`PHISH`	`MALWARE`	`SPAM`
Start Date	Sept. 17	Sept. 27	Sept. 22	Sept. 22	Sept. 22
End Date	N/A	N/A	Oct. 21	Oct. 21	Oct. 21
Frequency	Once	Once	Daily	Daily	Daily
Hosts	3,038,928	N/A	16,496	18,570	N/A
Domains	2,737,326	5,536,475	10,575	12,854	548,041
TLDs	3,235	7	306	259	327

The good: One context in which we study wildcards is the domains determined to be useful by Internet users. We use data from the DMOZ Open Directory Project [3] for this purpose. The Open Directory Project is a large directory of user submitted and editor approved Web URLs. We assume that those links submitted and approved are those someone has judged to be worthwhile, and are therefore in some sense good. We consider 2.7 million domains contained in this data set on September 17th, 2009. We refer to this data set as `DMOZ` throughout this paper.

The Bad: Another context in which we study wildcards is domains known to be associated with malicious activity. For this context, we use host names extracted from two real-time feeds of known phishing URLs [2, 22], three feeds of known malware-serving URLs [5, 11, 20], and one feed of domains for scam sites seen in spam mails [26]. We examine each of these feeds every day for a period of 30 days, extracting a total of 571,470 domains that were alive at the time of our receiving the feed. We refer to these data sets respectively as `PHISH`, `MALWARE`, and `SPAM` throughout this paper.

The Ugly: The last context we consider is a large general list of domains on the Internet. We refer to these as "ugly" since they could be used for any purpose, good or bad or something in between. The data source for this context is the zone files[2] from seven generic top-level domains (gTLDs), `.asia`, `.biz`, `.com`, `.info`, `.mobi`, `.net`, and `.org` [1,4,18,21,24,28], on September 27th, 2009. There were 110,728,143 domains contained in these TLDs, 58% of the total 192 million domains in the Internet at the time of our study [27]. From these, we

[2] Zone files are text files listing all DNS records directly contained in a domain.

randomly sample at a rate of 5%, or 5,536,475, which we examine in this paper. We refer to this data set as `ZONE_FILES` throughout this paper.

4 Wildcard Prevalence

Wildcards can occur at all levels of the DNS hierarchy. We concentrate on the domain level, instead of TLDs or subdomains, since this is generally the start or administrative control. We look for wildcards in four DNS record types: `A`, `NS`, `MX`, and `CNAME`. From the entries in each data set, we determine the domain name part of each host name using the Public Suffix List [17]. For all domain names in these data sets, for example, `foo.com`, we query for `*.foo.com` for the four record types. All queries were run once for each domain in the `DMOZ` and `ZONE_FILES` data sets, but daily for the others that changed often in real-time. We also query for the `NS` record for each domain to ensure that the domain exists at the time of the query.

A large fraction of domains we surveyed used wildcards. Table 2 presents an overview of the number and types of wildcards present in each data set at the domain level. *Between 1/4 and 3/4 of domains use wildcards, with the `DMOZ` data set showing the least prevalence of wildcards and the `SPAM` data set showing the most.* Not only is the `A` wildcard overwhelmingly popular, its usage mimics general wildcard usage trends. Some domains have more than one type of wildcard, causing the percentages in the last four rows of Table 2 to exceed the total percentage of domains using wildcards.

Table 2. % of active domains with wildcards of each record type in each data set

	`DMOZ`	`ZONE_FILES`	`PHISH`	`MALWARE`	`SPAM`
Domains					
Checked	2,737,326	5,536,475	10,575	12,854	548,041
Active	2,717,186	4,861,053	9,044	11,312	226,060
Inactive (%)	0.73%	12.2%	14.5%	12%	58.7%
Wildcards					
total %	24.52%	45.15%	32.09%	31.39%	75.10%
% `A`	18.76%	42.72%	27.79%	26.59%	72.30%
% `NS`	0.32%	5.53%	0.20%	0.19%	1.60%
% `MX`	5.72%	6.44%	4.10%	6.14%	6.83%
% `CNAME`	3.40%	3.75%	3.37%	4.49%	2.34%

4.1 Overridden Wildcards

Some wildcards may be overridden by specific entries. For example, a domain `foo.com`, may have a wildcard entry for `*.foo.com`, and a more specific entry for host name `a.b.foo.com`. This allows the domain to point `a.b.foo.com` to a different value than any other host name fitting `*.foo.com`. Now that we have seen how often wildcards are occurring, an important consideration is if they are overridden by a more specific DNS entry. If latter, then our conclusion about wildcard usage in the Internet would be different.

Toward the goal of identifying overrides, we proceed as following. For the DMOZ, PHISH, and MALWARE data sets where we have host names in the feeds, we query the DNS for A and CNAME records corresponding to the host names and check if the results of this lookup match the results of the wildcard lookup. If they are not the same answers, we consider the exact match to be overriding the wildcard. If we have multiple host names for one wildcard, we count it as an override if any of them do not match the wildcard entry. For ZONE_FILES and SPAM, we do not have exact host names, so we simply prepend www to the domain name. Though we do not know for sure that the host name so generated is being used, it is commonly used for web servers and may catch some overrides.

Notice that since our data sets are for web servers only, they do not contain name servers or mail servers. As a result, we cannot establish the presence of overrides for MX and NS wildcards by querying each domain for MX and NS wildcards and comparing the result to host names in the feed. This limitation is not severe since MX and NS are the two least popular type of wildcards per Table 2.

Table 3 shows the percentage of A and CNAME wildcards being overridden in each data set. Wildcards are overridden in 2.8-31.6% of cases. The SPAM data set sees the least overrides.

Table 3. Percentage of A and CNAME wildcards being overridden by specific entries

	DMOZ	ZONE_FILES	PHISH	MALWARE	SPAM
A	10.7%	31.6%	19.0%	19.9%	6.7%
CNAME	17.4%	8.8%	17.5%	30.0%	2.8%

Some data sets witness overrides for CNAME wildcards more often than those for A wildcards and vice versa. The difference is most striking for the ZONE_FILES data set. Examining the overrides in this data set closely, we find that 25.4% (557,949) of wildcards in the ZONE_FILES data set are hosted on name servers in domaincontrol.com. Of these, 99.7% are A wildcards being overridden by a specific CNAME record. These account for 88.9% of the overrides of A wildcards in this data set. If we ignore wildcard entries on this name server, only 6.6% of remaining A wildcards in this data set are overridden, much closer the percentage of overridden CNAME wildcards in this data set. *We conclude that wildcards are not frequently overridden in most data sets.*

5 Wildcard Usage

We now investigate the specific uses of wildcards by the good, bad, and ugly domains. To group related wildcarded domains, we considered several options and found it best to aggregate them by the DNS servers serving them. This grouping is intuitive because provider of DNS services, for example hosting companies, often provide a default configuration which most domains may choose. Similarly, large organizations running many of their own domains are likely to use similarly-provisioned servers. In fact, we aggregate even more by grouping wildcarded domains in terms of the domain of the DNS server.

5.1 Wildcard Usage Among Good Domains

The first data set we analyze is DMOZ, our set of good domains from a user edited directory. From this data set, we saw a total of 666,334 domains (24.5%) using wildcards. These were served by DNS servers belonging to 28,883 domains. Figure 2 shows a CDF of the wildcarded domains and the corresponding DNS server domains for this and ZONE_FILES data sets. *A key observation from this Figure is that just a few DNS servers are responsible for a disproportionate number of wildcarded domains.* Specifically, 29.1% of domains in the DMOZ data set are served by just top ten DNS server domains.

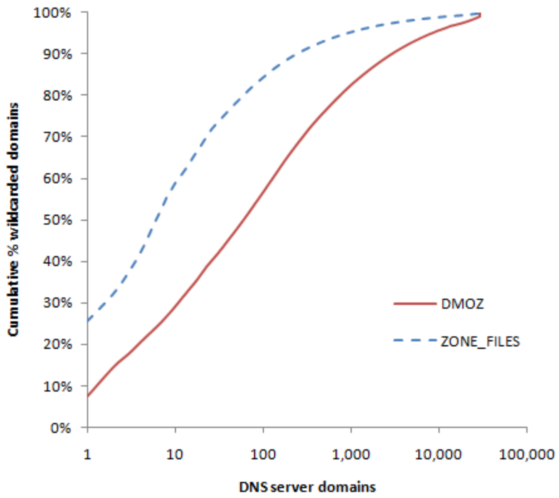

Fig. 2. CDF of wildcarded domains served by each DNS server domain

We now consider the top ten DNS server domains serving the most wildcarded domains. Table 4 shows the total domains and wildcarded domains served by each. *In looking over the domains accounting for most wildcard usage, we find that all are operated by registrars or web-hosting providers. Both these entities tend to provide a default configuration to users which includes a wildcard record.* Even users who override these with specific records for individual hosts may choose to keep the wildcard record.

Table 4. Top 10 DNS server domains serving the most wildcarded domains in the DMOZ data set

	Domains served	Wildcarded domains
worldnic.com	55,947	48,484
rzone.de	47,913	47,771
yahoo.com	23,194	21,835
namespace4you.de	17,611	17,409
kasserver.com	13,313	13,227
name-services.com	17,529	10,595
b-one.nu	9,471	9,406
ipower.com	9,058	9,057
register.com	13,853	8,077
mediatemple.net	7,869	7,705

Table 5. Top 10 DNS server domains serving the most wildcarded domains in the ZONE_FILES data set

	Domains served	Wildcarded domains
domaincontrol.com	1,138,877	557,949
name-services.com	179,130	147,697
worldnic.com	137,696	116,759
sedoparking.com	96,790	96,789
dsredirection.com	91,796	91,796
yahoo.com	82,747	80,669
register.com	72,827	62,137
secureserver.net	62,672	60,063
fabulous.com	39,166	39,137
parked.com	37,529	37,522

5.2 Usage Among the Ugly

We now change our focus to the ZONE_FILES data set, a large collection of domains taken from several TLD zone files. In this data set, we saw 2,194,565 domains using wildcards (45.2%). These domains are also served by a small number of DNS server domains, only 32,644. *Overall, we find that wildcarded domains are even more concentrated at a few name server domains than in the DMOZ data set.* Table 5 shows the top ten name server domains serving the most wildcarded domains in the ZONE_FILES data set. Four of the domains listed are in common with those in Table 4.

100% of the domains served by `sedoparking.com` and `dsredirection.com` are wildcarded. *These two, along with the two others that have the highest percentage of wildcarded domains, `fabulous.com` and `parked.com`, belong to companies involved in domain parking*[3]. Wildcards are very useful for parked domains. By directing visitors to a parking page, they allow monetization of all possible subdomains of a domain. However, not all parked domains are wildcarded. At least one provider of parking services we know of serves over 700,000

[3] A parked domain is a domain with no actual useful content, just a template page filled with ads redirecting the user to other pages, mostly for the purpose of monetizing chance-visitors to the domain.

domains but uses wildcards on less than 1%. The other major user of wildcards in this data set are web-hosting providers, as we saw in DMOZ. In fact, four of these are the same ones we saw in the top 10 from the DMOZ data set.

5.3 Usage Among the Bad

Next, we look for wildcard usage in bad data sets, PHISH, MALWARE, and SPAM. As we saw in Table 2, 32.1% of active phishing domains, 31.4% of active malware hosting domains, and 75.1% of active spam domains were using wildcards. The top ten DNS server domains serving wildcarded domains in PHISH, MALWARE, and SPAM, are shown in Tables 6, 7, and 8. They account for 21.67%, 22.95%, and 21.82% of the wildcard domains in these data sets respectively. This indicates a slightly lower concentration on the top name servers than we saw in the DMOZ data set, and much lower than we saw in the ZONE_FILES data set. *The other key observation from these tables is that many of the top-10 domains serving wildcarded domains are shared across all data sets.* This happens because many of the registrars and web-hosting providers are common across the three types of data sets.

Table 6. Top 10 DNS server domains serving the most wildcarded domains in the PHISH data set

	Domains served	Wildcarded domains
ixwebhosting.com	151	151
nshost.com.ve	139	139
rzone.de	63	60
yahoo.com	98	55
name-services.com	54	47
hosteurope.com	100	44
worldnic.com	48	42
hrnoc.net	33	32
register.com	32	30
namebay.com	128	29

Table 7. Top 10 DNS server domains serving the most wildcarded domains in the MALWARE data set

	Domains served	Wildcarded domains
freeservers.com	203	203
ixwebhosting.com	93	92
ipower.com	83	83
name-services.com	101	81
northsky.com	73	73
everydns.net	173	67
yahoo.com	63	59
servage.net	58	57
sorpresor.com	51	51
sitelutions.com	54	49

Table 8. Top 10 DNS server domains serving the most wildcarded domains in the SPAM data set

	Domains served	Wildcarded domains
name-services.com	16,699	14,764
tutby.com	6,167	5,966
domainservice.com	4,640	4,555
domainsite.com	3,202	3,200
domaincontrol.com	6,278	2,045
dsredirection.com	1,778	1,777
sedoparking.com	1,323	1,323
netstandardconsulting.com	1,296	1,296
peak-communications.net	1,180	1,180
dzcamera.net	941	940

Examining the domains listed in these three tables, some of the top ten from these data sets are in common with the top ten from the other two data sets. Some of these from the SPAM data set are associated with domain parking, and are probably there due to spam domains that have been taken down but still appear in our data set. These are less than 5% of the wildcards in SPAM so are certainly not the primary reason it has a higher proportion of wildcards than the others. Others are present because they are hosting providers. The most prominent example of this is name-services.com, which appears in the top ten from every data set. This and the few others from the three malicious data sets that are also top users in the other data sets may be large providers of malicious wildcards just because they are large providers who use wildcards by default and miscreants happen to use them. However, a majority that are the top users in these three data sets are not among the top users in the other two, making it likely that the miscreants are configuring wildcards intentionally.

Churn of Hosts Among Bad Wildcarded Domains: Miscreants can exploit the flexibility of wildcards to their advantage by simply swapping a blacklisted host name with a new one without having to change DNS entries. This can be useful in evading blacklists, which are based on exact host names today. We now attempt to determine if such is the case. In this analysis, we focus on the PHISH and MALWARE data sets, since the SPAM data set only includes domains, not host names.

We examine if new host names matching an existing wildcard entry are being added to our feed of bad data sets over time. Toward this goal, we calculate the daily churn of host names for each wildcarded domain in PHISH and MALWARE data sets. For this, we compare the host names for a domain with those listed the previous day. The sum of the additions and deletions is the churn rate for the day. We average this over all days the domain is alive. We do not count the initial set up or take down of the domain since some domains may have existed before or continued to exist after our data collection. Domains only seen for one day are also not counted since there is no second day to compare to derive a churn. Figure 3 depicts the CDF of churn over a period of 30 days.

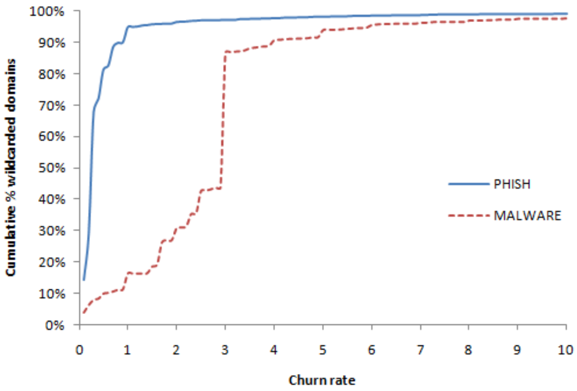

Fig. 3. CDF of churn rate of malicious domains over 30 days

For the PHISH data set, the average churn rate is 0.64, a little more than one change every 2 days, and the maximum is 52. For malware, the average is 2.87 with a maximum of 32.5. *Clearly, these numbers indicate that miscreants whose domains are active for more than a day, especially those serving malware, are taking advantage of the wildcard records to use new host names over time.*

6 Identifying Malicious Wildcard Usage

Thus far, we have seen that wildcards are in wide-spread use among all types of domains in the Internet. Even though some types of bad domains use wildcards more commonly than good or ugly domains, it is unclear if there are any trends that would distinguish wildcard usage among such domains from others. The primary reason for this is that the largest wildcard users are domain registrars and web-hosting providers and many of them are common across all data sets. This is somewhat unsurprising, given that a recent report examining phishing attacks from the first half of 2009 [25], found that only 14.5% of domains used in phishing were actually registered by the phishers, the remaining were compromised domains that could belong to a known service provider.

In this section, we take the first step and examine features of wildcard usage in an attempt to find ones that can distinguish their malicious usage from benign ones. Specifically, we examine time to live (TTL) values on wildcard DNS records and autonomous systems (ASes) corresponding to the wildcarded domains. We also use the Google search engine to discover new hosts matching known wildcards.

6.1 TTLs of Wildcarded Records

We examined the TTLs for each type of wildcarded records to see if malicious domains set different TTLs on their DNS records than benign ones. We compared the TTLs across the three data sets focusing on A wildcards since they were the most common type. A histogram of TTLs for A records is shown in Figure 4.

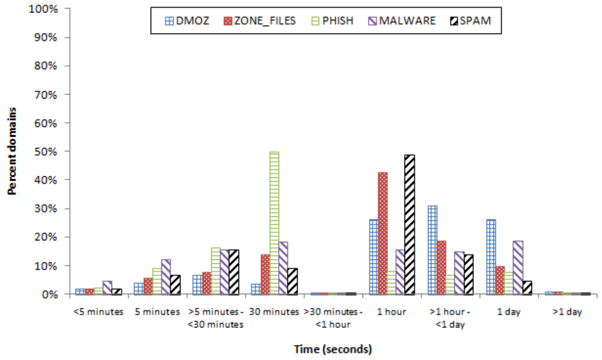

Fig. 4. TTLs for A records in each data set

A few TTLs are most popular in our data sets: 5 minutes, 30 minutes, 1 hour, and 1 day. The most significant difference we see among data sets is at 30 minutes, where the PHISH domains have a large spike but none other do. *In general, we find that wildcards in the PHISH, MALWARE, and SPAM data sets have shorter TTLs than those in good and ugly data sets, with 30 minutes and 1 hour being most popular values for malicious wildcarded domains.* This is intuitive because shorter TTLs allow miscreants to quickly update the IP addresses corresponding to malicious host names. Given that it is well known that miscreants are increasingly *fluxing* through IP addresses using short TTLs in an attempt to escape detection [12], examining TTLs corresponding to wildcarded records appears to be a promising avenue for investigation.

6.2 Autonomous Systems Pointed to by Wildcards

Many of the malicious domains are hosted on bots in geographically diverse Internet locations. ASes are one way to measure such diversity. Here, we examine how often the IP addresses corresponding to the wildcard records for each domain are spread over multiple ASes. This test only applies when there are multiple wildcard records for the same name pointing to different IP addresses. This is straightforward to do for A wildcards, since the right hand side of these records directly provides an IP address. For CNAME, MX, and NS records, which point to a host name instead of an IP address, we simply resolve the hosts on the right hand side to IP addresses. For all wildcard types, we see some difference in the results among the various data sets, however, we focus on A and CNAME wildcards in this discussion since these show the greatest difference, enough that they could be used to distinguish benign and malicious use of wildcards.

A histogram of the ratio of ASNs to IP addresses for wildcarded A records is shown in Figure 5. *The most notable observation here is that a majority of*

SPAM wildcard domains with multiple IP addresses have a ratio of ASNs to IP addresses between 0.6 and 0.7. Very few of the good data sets are in this range. In fact, PHISH and MALWARE A wildcards are much more likely than ZONE_FILES and somewhat more likely than DMOZ to be in the 0.9 to 1.0 range.

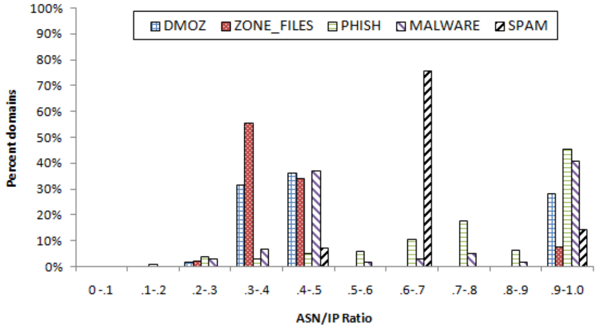

Fig. 5. Ratio of number of ASNs associated with each wildcard A record to number of IP addresses pointed to by record

Figure 6 shows the ASN/IP ratio for wildcarded CNAME records. Here, the SPAM data set almost all ends up in the 0.9-1.0 range, while less than 10% of the good data sets do so. Phishing and malware sites are significantly more likely than good ones to fall into the ranges from 0.1 to 0.4.

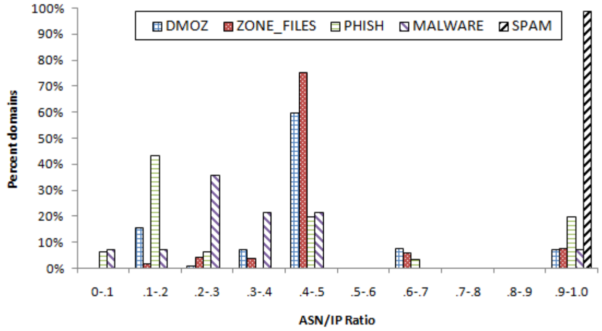

Fig. 6. Ratio of number of ASNs associated with each wildcard CNAME record to number of IP addresses pointed to by record

Overall, this method looks like a good one for identifying wildcards associated with spam sites, and can also be used with wildcards associated with phishing and malware sites. The only issue with it is that it relies on the wildcard entry pointing to multiple IP addresses, since otherwise, the notion of geographical diversity makes no sense. Wildcard entries point to multiple IP addresses in 1.6 - 4.2% of domains with CNAME wildcards and 0.5 - 27.2% of A wildcards depending on the data set. In the SPAM data set, this happens for 18.2% of A wildcards and 41.2% of CNAME wildcards. This data set is also the one where the ratio is most different from the good data sets, indicating that it would be effective a significant amount of the time for identifying wildcards associated with spam.

6.3 Host Names Represented by Wildcards

Technically, a wildcard entry in the DNS can match any host name. However, in practice, a site may only use some of these host names. Blogging and social networking sites often provide a subdomain for each user. Out of 170 such sites we investigated, 52 support subdomains for each user and all do so using wildcard entries. Of these 52, 37 use A wildcards, and the rest use CNAME. As a specific example of this, Windows Live Spaces provides a subdomain for each user, all handled by a single wildcard entry, and claims 175 million users [14]. Even the smallest blog site we have found using wildcards supports over 10,000 subdomains.

We now investigate if Google searches can reveal new host names covered by our wildcards. Using this method, we would like to see if malicious and benign domains use the wildcard for differing numbers of host names in practice. Toward this goal, we queried the Google search API [6] for a sampling of the domains with A or CNAME wildcards from each data set, using site restriction to make sure all responses were from the domain we were interested in, not external pages with the domain in their text. This gives us an idea of how the wildcard is being used, subject to a maximum of 64 results imposed by the Google API. Table 9 shows how many domains were queried from each data set, and what percentage were found in the Google index.

Table 9. Wildcarded domains from each data set queried at Google

	DMOZ	ZONE_FILES	PHISH	MALWARE	SPAM
Domains checked	6,717	9,867	1825	2321	4,057
Domains responding	6,587	4,596	1089	1263	475
% indexed	98.1%	46.6%	59.7%	54.4%	11.7%

We find that a large percentage of domains we queried were indexed by Google. Over half from ZONE_FILES were not indexed, probably due to the large amount of sites devoid of useful content, such as parking pages. Over half of the MALWARE pages were indexed. From SPAM, a large majority were not indexed by Google. This is perhaps because the URLs associated with them are only advertised though email, so the Google crawler would have never seen them. It is possible that Google to intentionally excludes some pages with known malicious content.

Out of the domains that did return Google results we examine how many results were returned. Results are shown in Figure 7. The most notable result here is that wildcards from PHISH correspond to a higher number of hosts known by

Google than wildcards from other data sets. For the other data sets, meaningful distinctions are hard to make, since SPAM and ZONE_FILES results are similar to each other, as are MALWARE and DMOZ. While it can not be said with certainty that wildcards representing large numbers of host names are associated with phishing, it is certainly an indication that further scrutiny is required to see if they are phishing sites. While a client could not directly determine the number of hosts a wildcard represents, any organization who crawls the Web should be able to provide data on how many host names they have seen in a domain name, making this check practical.

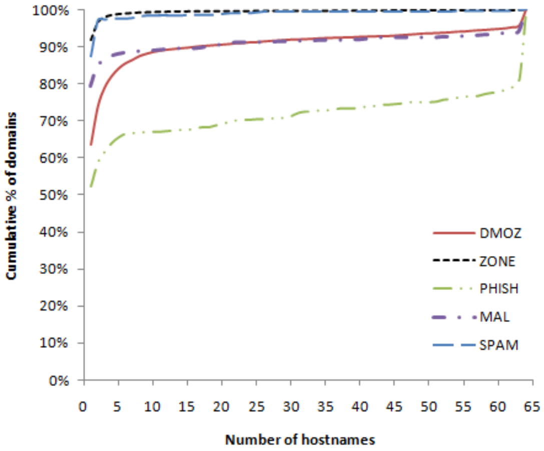

Fig. 7. Cumulative percent of wildcarded domains in each data set with the given number of host names found in the Google index

Figure 8 shows how many host names Google returned that were not found in out data sets. Here, only those data feeds that contained host names are considered since no conclusions can be drawn from data sets containing only domain names. *For most domains Google indexed in the PHISH and MALWARE data sets, it knows of several host names not in our data set. This indicates that blacklisting could be improved by directly including wildcard entries instead of exact host names.*

Since overridden wildcards may indicate the wildcard entry itself is not actually used, just present as a default, we wanted to see if the wildcarded domains where we found specific DNS records overriding the wildcarded entry appeared less than others in Google. This does in fact appear to be true. 8.8% of overridden domains we looked up from DMOZ, 20.4% from PHISH, 6.8% from MALWARE, and 1.0% from SPAM appear in Google. Compared to the percent responding for wildcarded domain in general from Table 9, the percent responding for domains with overridden wildcards is an order of magnitude less, for all but PHISH, which is still significantly lower. The fact that Google does not know about these

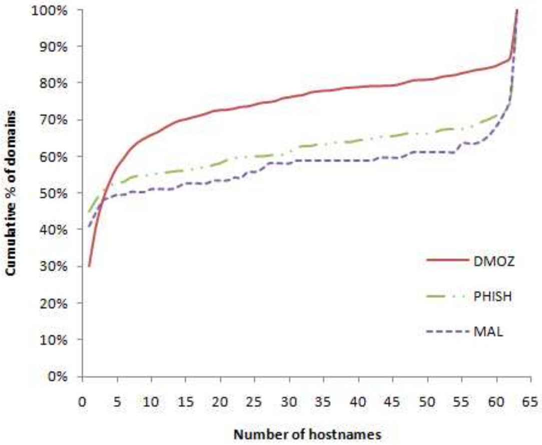

Fig. 8. Cumulative percent of wildcarded domains in each data set with new host names found in the Google index

domains indicates their wildcard may not be used to represent as many hosts, giving Google less of a chance to find them. This is further reinforced by the observation that of the ones with overriding found in Google, only 4.8% from `DMOZ`, 7.2% from `PHISH` and 2.4% from `MALWARE` return any new results not from our data feed. This is far lower than seen in Figure 8. *This observation implies that the presence of DNS records overriding wildcarded records may be an indication that the wildcard is not used for purposes such as evading blacklists.*

7 Related work

Wildcard records have been a part of DNS from the original specification [16]. This specification is ambiguous and unintuitive, so RFC 4592 [10] was created to clarify the intended behaviors of wildcard records. In addition to issues arising from the specification being non-intuitive, it has been argued that they violate common assumptions on how DNS should operate. An Internet Architecture Board (IAB) commentary [8] describes the way wildcards violate this assumptions and the issues that can arise from it. It recommends only using `MX` type wildcards since they are the only ones that only affect a single protocol. It also recommends not ever using wildcards for domains that have subdomains. Nonetheless, wildcards are in widespread use, as our study finds.

Previous work by Kalafut *et al.* [9] took a more general look at the contents of all DNS records in 5 million DNS domains. In this paper, we focus in more detail on a specific subset of the records, just the wildcards, and search for such records in a larger set of domains. Pappas *et al.* [23] also examined DNS configurations, looking specifically for three types of errors that could impact availability. The

Measurement Factory also has done surveys of DNS configurations [13], focusing on software version and deployment of features such a source port randomization.

8 Discussion

Our study found that wildcards are popular among all types of Internet domains, including those involved in malicious behaviors. Among malicious users, spammers use wildcards the most. They are also the least likely to override them. There is a significant churn among host names matching wildcards belonging to malware and phishing domains, implying that they too are likely taking advantage of the wildcards to escape exact host-name-based blacklists.

We found some distinguishing features of the malicious wildcards, such as short TTLs, distinct ratios of IP addresses to ASes when the wildcard pointed to multiple IP addresses, and a low likelihood of appearing in Google results, especially for wildcard domains associated with spam. None of these observations on their own may be enough to distinguish a benign wildcard use from a malicious one. However, these characteristics may be useful in conjunction with others and with each other to identify some malicious sites, a direction we plan to pursue in future work.

Finally, the observations in this paper point to a specific immediate improvement that can be made in blacklists. Many blacklists currently list individual host names. One prominent example, the Google Safe Browsing API [7], uses a system somewhat similar to regular expressions, but this is not common for other blacklists. In the blacklists we use for data feeds, only individual host names were listed, which often ended up matching wildcards. Such blacklists could be easily improved by checking for a wildcard DNS entry and adding it instead of the host name where appropriate. Miscreants could evade detection based on wildcards and still use a large number of host names by creating separate DNS entries for each. However, these cases could still be dealt with by a wildcard entry in the blacklist, added once some threshold number of individual host names in a domain has been seen involved in malicious activity.

References

1. Afilias Limited: How can I get access to Afilias' TLD zone file for .INFO domains?, http://www.info.info/faq/how-can-i-get-access-afilias-tld-zone-file-info-domains
2. APWG: Anti-phishing working group. http://www.antiphishing.org/
3. DMOZ: Open directory project, http://www.dmoz.org/
4. DotAsia Organization Limited: .ASIA Zone File Access Agreement, http://www.dotasia.org/info/DAO.ZONE-2007-10-24.pdf
5. eSoft Inc.: http://www.esoft.com/
6. Google: Google AJAX Search API, http://code.google.com/apis/ajaxsearch/
7. Google: Google Safe Browsing API, http://code.google.com/apis/safebrowsing

8. Internet Architecture Board: Architectural concerns on the use of DNS wildcards. IAB Commentary (Sep 2003), www.iab.org/documents/docs/2003-09-20-dns-wildcards.html
9. Kalafut, A., Shue, C., Gupta, M.: Understanding implications of DNS zone provisioning. In: ACM SIGCOMM Internet Measurement Conference (IMC) (2008)
10. Lewis, E.: The role of wildcards in the domain name system (Jul 2006)
11. MalwarePatrol: Malwarepatrol - malware block list. http://www.malwarepatrol.net/lists.shtml
12. McGrath, D.K., Kalafut, A., Gupta, M.: Phishing infrastructure fluxes all the way. IEEE Security and Privacy Magazine Special Issue on DNS Security (2009)
13. Measurement Factory: DNS survey: October 2008. http://dns.measurement-factory.com/surveys/200810.html
14. Microsoft: Windows Live Fact Sheet, http://www.microsoft.com/presspass/newsroom/msn/factsheet/WindowsLive.mspx
15. Miller, R.: Phishers use wildcard DNS to build convincing bait URLs (Mar 2005)
16. Mockapetris, P.: Domain names - concepts and facilities. IETF RFC 1034 (Nov 1987)
17. Mozilla Foundation: Public suffix list. http://publicsuffix.org
18. mTLD, Ltd.: dotMobi Zone File Access Agreement, http://mtld.mobi/domain/zonefile
19. Mutton, P.: New phishing attacks combine wildcard DNS and XSS. http://news.netcraft.com/archives/2009/02/17/new_phishing_attacks_combine_wildcard_dns_and_xss.html (Feb 2009)
20. NETpilot GmbH: Viruswatch mailing list. http://lists.clean-mx.com/cgi-bin/mailman/listinfo/viruswatch
21. NeuStar Registry Services: .BIZ Zone File Distribution, https://www.neulevel.biz/zonefile/
22. OpenDNS: PhishTank. http://www.phishtank.com/
23. Pappas, V., Xu, Z., Lu, S., Massey, D., Terzis, A., Zhang, L.: Impact of configuration errors on DNS robustness (2004)
24. Public Interest Registry: .ORG Registry - Zone File Access, http://pir.org/index.php?db=content/Website&tbl=Registrars&id=7
25. Rasmussen, R., Aaron, G.: Apwg global phsihing survey: Trends and domain name use in 1h2009 (Oct 2009)
26. SURBL: http://www.surbl.org/
27. VeriSign: Domain name industry brief (Feb 2010), http://www.verisign.com/domain-name-services/domain-information-center/domain-name-resources/domain-name-report-feb10.pdf
28. VeriSign, Inc: TLD Zone Access Program, http://www.versign.com/information-services/naming-services/page_001052.html