Unintentional Darknet: An Analysis of Scanning Behavior through the Lens of Abandoned Domains

Christopher Kitras and Bryson Schiel
Department of Electrical and Computer Engineering
Brigham Young University
Provo, UT, United States
{chkitras,schielb}@byu.edu

1 INTRODUCTION

As the internet continues to mature, many privately hosted sites remain active, even as their overall usage drops. In the case of organizations like universities, some of these sites might represent projects that have been abandoned, but that remain available to public inquiry. These sites may become harder to find, especially if the hosting services make changes to the DNS route by which a user may access the site. This is the case with a project site from the BYU Electrical and Computer Engineering Department, a project known as Java Hardware Design Language (JHDL).

In this report, we illustrate how misconfigured or abandoned domains can behave as a sort of darknet. Instead of relying on potentially thousands of unconnected IP addresses within a large subnet, we utilize DNS logs from an authoritative server that records all incoming queries. For our project, we specifically observe the aforementioned JHDL project. We discuss its peculiar place in the top 15 QNAMEs searched for on the BYU DNS that did not contain the substring "byu". We then try to characterize from which locations most of the queries originate and whether or not they are normally trustworthy. Finally, we discuss the relevance of our findings and what future steps could be taken to further utilize abandoned domain names as a means of detecting suspicious activity.

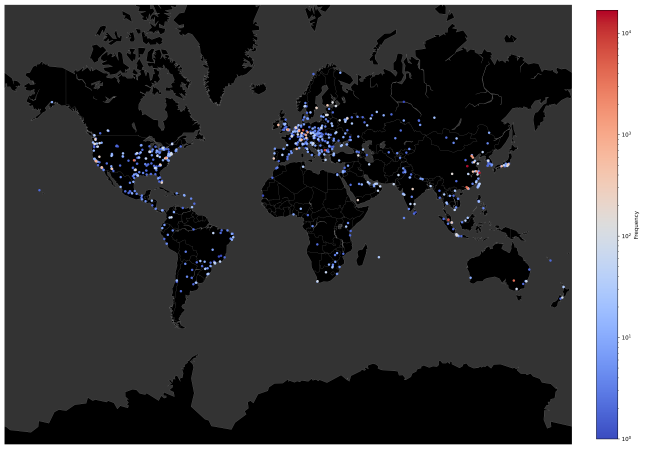


Figure 1: Geographical locations of IP addresses that made queries with "jhdl" in the QNAME.

2 BACKGROUND

JHDL combines object-oriented programming (OOP) from Java and a traditional hardware design language like VHDL [2]. Hardware design languages like VHDL and Verilog use a programming-like syntax to describe the flow of inter-connected circuits meant to be synthesized into a bitstream and uploaded to and executed on an FPGA. While these design languages are contemporaries to other more widely recognized low-level languages like C and Java, they focus more on the synthesis of circuits and therefore, do not adhere naturally to many of the paradigms in software-oriented programming languages. Today, many computer engineers who want to design circuits feel like the hardware design languages of the late ’90s and early 2000s are missing many familiar paradigms for coding up logic.

JHDL is a high-level tool that combines the familiar Java language and its OOP concepts and compiles them to more traditional HDL logic, creating the bitstreams needed to emulate circuits on an FPGA. JHDL is a project designed at BYU in the Configurable Computing Lab (CCL). Its first release was in 1998, with a website as a landing page for the project created in 1999 and last modified in 2006. At the time of this report, the website itself has not been modified in over 18 years. Despite its 97 citations in papers and its 21 citations in patents, the metric data on IEEE Xplore shows the project as receiving only a handful of views per month. It appears that the once burgeoning project of BYU’s CCL has all but ceased.

However, as mentioned previously, just because a project goes dormant at an institution, the website advertising the project does not. Abandoned sites are normally viewed as a security flaw by which attackers may gain access to an organization, mainly due to misconfigured or outdated software. Attackers use scanners to enumerate domains and potentially vulnerable hosts. While no such direct claims are made

in this report, we acknowledge that an elevated interest in a defunct website may be suspicious. As elaborated in 3.1, this proves interesting since domains associated with the JHDL website are in the top 10 QNAMEs without the substring "byu".

3 METHODOLOGY

3.1 Data Sources

The data used for these findings are BYU's DNS logs from August 26, 2024 to September 15, 2024. These contain many fields associated with a query sent to the server such as timestamp, QNAMEs, source IP address, and much more. While we initially considered most of these fields to be interesting, especially before we focused on a specific aspect of our project, we ultimately filtered out most of the rows that did not include the substring "jhdl". Specific analyses led us to filter this subset further and will be noted for each respective test.

3.2 Finding Unexpected Domains

While initially working on the project, we reviewed the prompt research questions to try and find any interesting trends. When we got to the question, *What names are being queried of BYU authoritative servers that are not expected (e.g., outside of* `byu.edu` *or* `187.128.in-addr.arpa`*)?*, we filtered out all QNAMEs that did not contain any reference to BYU. We initially saw that most QNAMEs were from a sister college to BYU: Ensign College (previously known as LDS Business College). Eight out of the top 15 either pointed to the main address for the `ensign.edu` or various subdomains of technologies used by the university such as Okta [5] and Brightspot [3].

However, as referenced in Table 1, the rest of the top 10 most unexpected QNAMEs were for JHDL. Unaware of what JHDL was at this point, we used `whois` to try and determine more information about the owner of the JHDL domains. For all three major TLDs, we found that the Registrant Organization was not just BYU, but specifically the Configurable Computing Lab in the Electrical and Computer Engineering Department.

After discovering that the administrators of the JHDL website are in our department, we spoke with the associated faculty members who are still supporting the website to gain further insight into the project. Most of what they told us can be found in 2. Furthermore, the faculty member who is over the actual administration of the website admitted that it is somewhat of a department heirloom and that he had received responsibility over it once the last faculty originally associated with it had retired. Now certain that this project was functionally abandoned, we decided to further discover why there was so much traffic in the given dataset for JHDL.

Table 1: Top 15 QNAMEs not including "byu" or its IPv4 Prefix

Query Name	# of Queries
www.ensign.edu.	2109088
ensign.edu.	804684
jhdl.com.	**779440**
ip6 arpa	683525
ntp-nist.ldsbc.net.	606853
msoid.ensign.edu.	616685
rp.ensign.edu.	559277
hcsys.ensign.byu.edu.	319234
ip6 arpa one	163838
okta.ensign.edu.	150022
brightspot.ensign.edu	133090
ip6 arpa	114386
jhdl.org.	**97143**
jhdl.net.	**94732**
ip6 arpa	90627

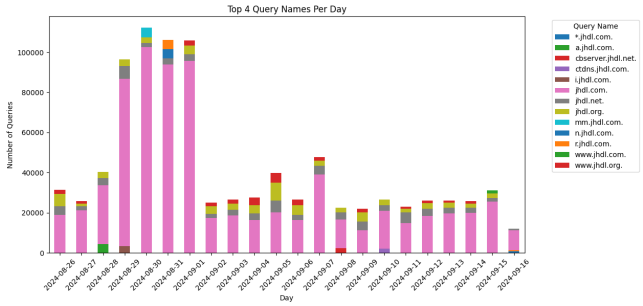


Figure 2: Top 4 QNAMEs per day that did not include "byu" but did include "jhdl".

3.3 Volume of Queries

To characterize the interesting volume of queries we observed for JHDL-related domains, we take all the data that has filtered out mentions of BYU but still mentions JHDL and find the top four QNAMEs queried for that day. As is noted in Figure 2, a few interesting subdomains were searched for as well such as `cbserver` [4] and `ctdns`, which may refer to certificate transparency [6]. However, other queries look simply for seemingly random single-character subdomains. When probing all of these domains (including those without subdomains specified) with `nslookup`, we received SERVFAIL errors, suggesting that there are no servers who claim responsibility for those domains. While we have no empirical evidence to suggest as such, we believe that probing single-character domains could be part of an effort to scan all possible subdomains for `jhdl.com`.

We also noticed that the dates between August 29 - September 1 received unusually high traffic compared to the rest of the days. We originally hoped for some sort of trend since this date range takes place during BYU's New Student Orientation for the Fall semester and the following weekend. However, upon initial inspection, we were disappointed to see that there were no obvious correlations. Due to tight time constraints for this project, we did not pursue this vector any further and focused on other aspects of our analysis.

3.4 Access both `jhdl.com` and `jhdl.ee.byu.edu`

While trying to sort through the different addresses querying JHDL services, we identified a still-active URL for the web service: `jhdl.ee.byu.edu`. Any of the addresses reaching out to `jhdl.com` could also eventually identify this site as the location of the web server, and indeed many addresses do find this. In our study of the two weeks' data, we identified 47 addresses that query both `jhdl.ee.byu.edu` and some other query with "jhdl" in the query name. This, however, is a small minority of all sites that query a JHDL site. For a fair comparison, however, we decided to pull the 40 most active addresses to query some JHDL sites but that *never* request information on `jhdl.ee.byu.edu`. We also stuck with the 40 most active addresses that eventually *did* request information on `jhdl.ee.byu.edu`, and we compared these two groups in several ways.

One reason for this distinction in the dataset is that once a host has successfully queried a site like `jhdl.ee.byu.edu`, that single DNS response will guide all future traffic from that address for the time being. For those sites that continuously query sites that lead nowhere, such repeated and futile behavior could indicate malicious behaviors on the network. As part of our analysis, we dig deeper into the differences between these two groups and try to understand the reasons for their specific behaviors.

4 ANALYSIS

In this section, we start by reviewing general trends we see from the entire two weeks' measurement of JHDL-related DNS queries, and then we dig deeper into various sub-trends that match the other behaviors mentioned.

4.1 Autonomous Systems

We start our analysis of the addresses reaching out to `jhdl.*` by seeing how many there are and how often each address queries one of the different JHDL domains. While this is one of the most queried domains (outside of any `byu.edu` site), it quickly becomes apparent that the traffic for this site is very uneven. Observing Fig 3, we see a cumulative distribution function (CDF) of the count of Autonomous Systems (AS)

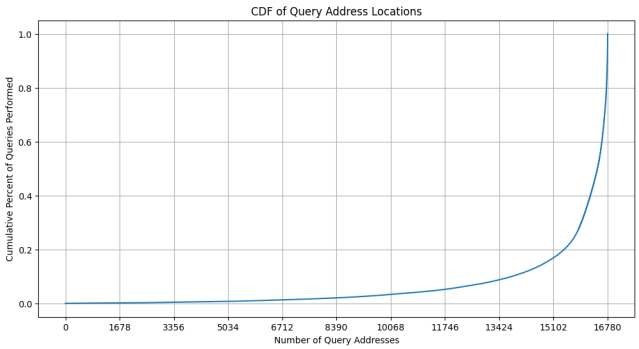


Figure 3: CDF of all ASes that queried any `*.jhdl.*` domain.

that query JHDL. In this list, it is clear that a relatively small portion of the ASes that queried JHDL domains had the highest rate of performing the query in the first place (you can see this with the steep curve at the right end of the CDF).

This matches what we see in our divided dataset, based on which addresses eventually queried `jhdl.ee.byu.edu`, and which did not. In all, looking at the most active addresses to query JHDL data from the DNS servers, there seem to be very few entities with DNS query counts about a couple thousand, while some have tens of thousands who ignore this official domain.

4.2 Geolocation

After looking at the distribution of queries to domain names, we look at the locations of some of the most active query addresses in this dataset. A plot of this data is found in Fig 1. We identified a couple of hotspots looking for our deprecated site, primarily from East Asian locations like China and Malaysia. We also saw quite a bit of traffic from Belgium and Germany as well as visible in Table 2. At first glance, it seems that these countries are in line with what we would expect from SoC and FPGA manufacturers. We also note that it is curious that a technology that was developed and is actively maintained in the United States, specifically in Utah, did not even make the list of top 10 hosts accessing these domains.

With this elevated rate of unexpected traffic from international regions, we are aware that this data could indicate a possibility of cyberattacks or at the least, suspicious activity. It is for this reason we decided to further characterize the security experiences of our dataset in Section 4.3. At this point, we will return to the mention of addresses that eventually queried `jhdl.ee.byu.edu` vs. those addresses that never did.

Table 2: Top 7 Autonomous Systems and Countries that query JHDL domains

AS Name	Country	ASN	≥ # of Queries
UNICOM-SHFT-IDC, Shanghai	China	140979	44838
TENCENT-NET-AP	China	45090	11908
CHINANET BACKBONE	China	4134	9457
UNICOM-SHFT-IDC, Guangdong	China	135061	9297
TELEPOINT, BG	Belgium	31083	7388
TTSSB-MY TM TECHNOLOGY SERVICES	Malaysia	4788	7160
DE-FIRSTCOLO	Germany	44066	5470

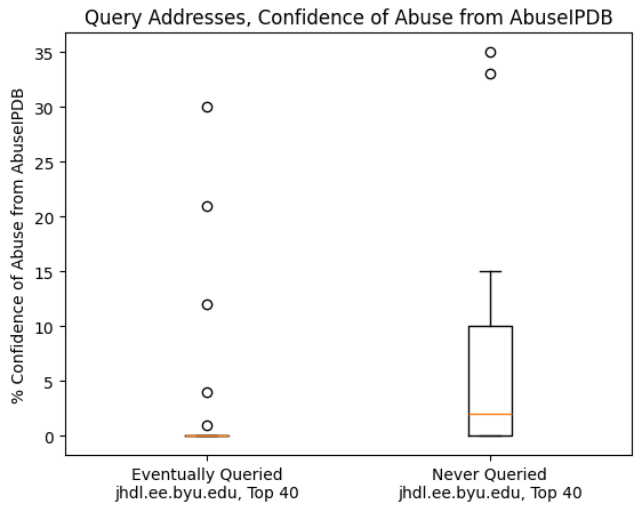


Figure 4: Confidence of abuse according to AbuseIPDB comparing hosts that queried `jhdl.ee.byu.edu` and those that did not.

Table 3: Top 40 Hosts Visiting JHDL Sites Excluding `jhdl.ee.byu.edu` with non-zero probability of abuse

Query Address	% of Abuse
84.200.x.x	35
84.200.x.x	33
2001:1608:x:x::x:x	15
2001:1608:x:x::x:x	13
113.96.x.x	11
210.22.x.x	10
198.142.x.x	10
198.142.x.x	10
223.166.x.x	2
58.251.x.x	2
106.54.x.x	2
122.51.x.x	2
101.43.x.x	2
81.71.x.x	2
101.91.x.x	2
198.142.x.x	2

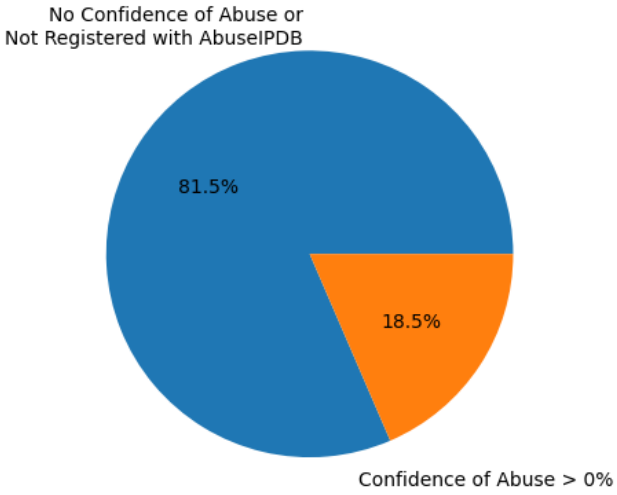
4.3 AbuseIPDB

One tool we use in our analysis is the Abuse IP Database (AbuseIPDB) [1]. This tool keeps a record of addresses that are reported to have performed suspicious online activities such as scanning, phishing attacks, denial of service, and so on. Based on the number and types of behaviors reported, AbuseIPDB will assign a confidence value of that host being suspicious or malicious.

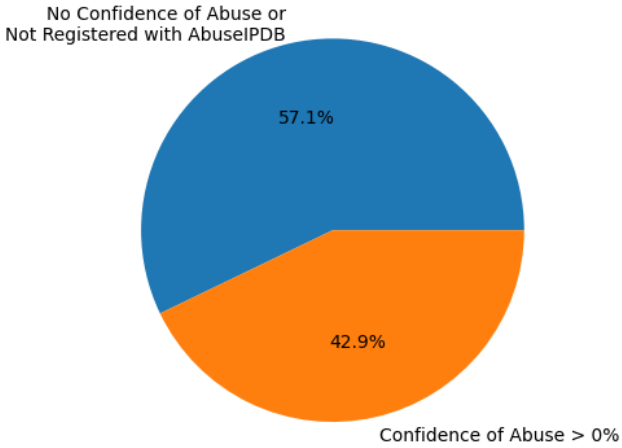
For our analysis, we requested information from AbuseIPDB on the top 40 addresses that eventually did reach out to `jhdl.ee.byu.edu` as well as the top 40 that did not ever query `jhdl.ee.byu.edu`. These rankings can be found in Table 4 and Table 3 respectively.

What we found took both of us by surprise. By a factor of 2.5, a site is more likely to have a suspiciousness rating on AbuseIPDB if it only ever queries these superficial, unsupported domains than if it does eventually query `jhdl.ee.byu.edu`. We further represent this difference in Fig 5a and Fig 5b. We note that the site list that never queried the true domain has a significantly higher percentage of potentially malicious actors in its group. Our results show that only 18.5% of hosts that did eventually visit the canonical website for the JHDL project were considered somewhat suspicious whereas an increased 42.9% of the hosts that never visit the canonical website were marked as suspicious.

Not only are there more potential sites that are registered on AbuseIPDB if they never query the legitimate site, but their overall score on AbuseIPDB indicates even further reason to distrust these sites. Referring to Fig 4, it can be seen that the trend of suspicion scores on AbuseIPDB is significantly higher for those sites that never query the legitimate `jhdl.ee.byu.edu` site.



(a) Top 40 addresses that eventually queried `jhdl.ee.byu.edu`



(b) Top 40 addresses that never queried `jhdl.ee.byu.edu`

Table 4: Top 40 Hosts Visiting JHDL Sites Including `jhdl.ee.byu.edu` with non-zero probability of abuse

Query Address	% of Abuse
66.249.x.x	30
172.253.x.x	21
95.217.x.x	12
112.65.x.x	4
217.195.x.x	1

5 DISCUSSION

As noted in Section 2, the aim of this paper is not to ascertain whether the still-active JHDL project website is an active target of cyberattacks. That would require a deeper analysis of its server's logs and traffic. However, we have found some unique and potentially concerning behaviors in just the DNS queries towards this project site.

First and foremost, we have a convincing pointer to the idea that you can use DNS queries to help determine a querying site's likelihood of being a malicious user, particularly a network scanner. The JHDL project has an active domain (`jhdl.ee.byu.edu`), and any DNS query source that eventually queries that address can quite probably be looking for a genuine connection to that project (whether that be malicious or not still has yet to be determined). If instead a DNS query source address never attempts to access the legitimate site, but rather only queries alternative locations, that behavior can help indicate malicious scanning activity on the network. While there is much more data that can and should be collected to support this claim, it is readily inferrable by reviewing the data in this work.

6 CONCLUSION

The work done in this paper is a preliminary attempt to characterize DNS querying addresses based on whether or not they ever attempt to connect with the actual server they appear to be reaching out to. More work needs to be done, and there stands a need to correlate data from the DNS logs with data from the JHDL server logs, but this offers a promising start to the research.

Abandoned project sites in general require careful review as they remain active well after a project is shelved. There is much that can be gleaned from observing public interest in these projects after their conclusion date, and it is worth exploring the potential threat that these sites can pose as unintentional darknet access points.

REFERENCES

[1] AbuseIPDB LLC. 2024. Abuse IP Data Base. Retrieved from https://www.abuseipdb.com/.

[2] P. Bellows and B. Hutchings. 1998. JHDL-an HDL for reconfigurable systems. In *Proceedings. IEEE Symposium on FPGAs for Custom Computing Machines (Cat. No.98TB100251)*. 175–184. https://doi.org/10.1109/FPGA.1998.707895

[3] DBA Brightspot. 2024. Enterprise CMS and headless CMS solutions. Retrieved from https://www.brightspot.com/.

[4] M. Jeusfeld. 2024. ConceptBase.cc - A System for Metamodeling and Method Engineering. Retrieved from https://conceptbase.sourceforge.net/.

[5] Okta, Inc. 2024. Employee and Customer Identity Solutions. Retrieved from https://www.okta.com/.

[6] Francisco Ros. 2018. Certificate transparency for domain owners. https://moss.sh/certificate-transparency-for-domain-owners/