# A Quantitative Analysis of the Traffic Patterns on the Brigham Young University Servers

Chester Davison
*Brigham Young University*

Jeff Hutchings
*Brigham Young University*

Warda Usman
*Brigham Young University*

## 1 Introduction

Recursive servers are valuable parts of the Domain Name System (DNS). There are many domain names paired with their own unique IP addresses. The DNS allows users to type in the domain name (such as google.com) of the website rather than the IP address. A typical user of a web browser enters the name of a desired website into the browser. Then, the browser makes a connection with a recursive server in the DNS. The recursive server then serves the IP address for the domain name that the user provided. For example, the IP address 172.217.2.14 would be returned if the domain name google.com were entered.

Sometimes, a domain's IP address is cached on the recursive server, in which case the recursive server can send the IP address back to the browser without any additional work. Other times, however, the domain's IP address is not cached on the recursive server, and the recursive server must make a request to an authoritative server that stores domain names and their IP addresses.

In this paper, logs that contain data regarding queries that originated from recursive servers were sent to authoritative servers are analyzed in order to gain a better understanding of internet traffic patterns and how to model such patterns. The Covid-19 pandemic began in early 2020 and is ongoing. It is interesting to consider how the pandemic may have affected internet traffic patterns.

Several empirical analyses are presented in this paper that quantify different types of data associated with a DNS query, including spacial-temporal differences between data before COVID-19 (April 2019) and data after COVID-19 (April 2020), QName distribution, server IPs, client IPs, IPv4 vs IPv6, port distribution, and flag distribution.

Several challenges needed to be overcome if these goals were to be accomplished. First, processing the large amounts of data present in several months of DNS query logs on our computers was laborious and time-consuming. Third party APIs facilitated analysis of the data, but limitations to free accounts and restricted query amounts per day, preventing the data from being analyzed as thoroughly as it could have been.

## 2 Background

There are several different types of data associated with an internet query. In this section, we explain the meaning and purpose of each type of data on which we run our analyses.

### 2.1 Timestamp

The UNIX timestamp is a way to track time as a running total of seconds [3]. The start time is defined to be January 1st 1970 at UTC. Since this is a running total, the time stamp is not affected by time zones. Currently the time stamp is a 32 bit number but will need to be converted to a 64 bit number on January 19, 2038, or it will overflow.

### 2.2 QName

The QName is a qualified name contained in the query. Recently there has been a strong push to QName minimisation, where the DNS resolver no longer sends the full original QName to the upstream name server. For example, a resolver receives a request to resolve foo.bar.baz.example and it already knows that ns1.nic.example is authoritative for .example, it will send the query QNAME=baz.example to ns1.nic.example. In this way it will minimise the amount of data sent from the DNS resolver to the authoritative name server, and thus improve DNS privacy. [2]

### 2.3 QType

The Domain Name System (DNS) specifies a query type (QType) and can be set to a default of "ANY" (or asterisk). There are dozens of resource record types and include pseudo resource records as well as obsolete record types. The full list can be found here. The most common types are A, AAAA, CNAME, NS, MX, TXT, and PTR.

- A is an address record and will return a 32-bit IPv4 address.
- AAAA is an IPv6 address record and will return a 128-bit address.
- CNAME is a canonical name record and is an alias of one name to another.
- NS is a name server record and will delegate a DNS zone to use the given authoritative name servers.
- MX is a mail exchange record and maps a domain name to a list of message transfer agents for that domain.
- TXT is a text record and was originally used for human-readable text, but is now used more often to carry machine-readable data.
- PTR is a pointer to a canonical name. Unlike CNAME, DNS processing stops and just the name is returned.

## 2.4 Server IP and Client IP

An IP address is a unique numerical label that identifies each computer using the internet protocol to communicate over a network. This address serves to identify the host or network interface and location addressing. An IPv4 uses a 32-bit number and IPv6 uses 128 bits.

## 2.5 Ports

In computer networking, a port is a communication endpoint. A port is identified for each transport protocol and addressed by a 16-bit unsigned number, known as a port number. [1]

## 2.6 Flags

The query log contains a flag field. From the Bind documentation we can get what each one of these flags mean. The Recursion Desired flag will be set with a + if set and - if not set. If the query was signed (S), if an Extension mechanism for DNS (EDNS) was in use (E), if TCP was used (T), if DO (DNSSEC Ok) was set (D), or if CD (Checking Disabled) was set (C).

From the documentation[1], it appears that stub resolvers almost always never set the recursion bit. Generally only full service resolvers such as named resolvers flip this bit, since they are capable of taking the referral and using them to follow the delegation chain all the way to an authoritative nameserver. Stub resolvers, in general, do not have this capacity and therefore leave this bit unflipped. It also appears from the documentation that using this feature is considered bad form.[2] [2]

# 3 Methodology

One of the first challenges we faced as a team was the large size of the datasets. In order to facilitate all future parsing and analysis of the data, we added all the details to a SQLite database. Even then the datasets were so large that it could take more than 20 minutes to parse a file. To speed up the analysis we read all the data into a database 1000 lines at a time, then ran queries against the database. This dropped the file analysis time from 20 minutes to less than 20 seconds. With each file, we noticed that there were about a dozen lines that could not be read. Since the files contained over three million lines and the unread lines constituted about 0.0004% we decided this loss of data would have no significant change on our analysis and was therefore ignored.

## 3.1 Dataset

We used an anonymized dataset of queries (available here) to BYU's authoritative DNS servers. The dataset consisted of query data collected in March-April period in 2019 and 2020, and January 2020 and 2021. We used several different subsets of the data depending on the analysis we were running. Each dataset instance consisted of a timestamp, qname, qtype, server IP, client IP, client port, and flag type. The data for a single day often exceeded four million queries.

## 3.2 Research Questions

After looking at the dataset instances and understanding the kind of information we had at hand, we decided on our research questions. The first things we wanted to do was look at the distribution of qname, ports, flags, and qtype.

We also postulated that over time we would be able to see more traffic using IPv6 over IPv4. To test this, we wanted to look at all three years and choose a day that was the most similar on the calendar.

The biggest and most complex part of our analysis was to compare and contrast the data over a time period. Specifically, we wanted to look at traffic patterns over the course of a day, a week, and a month. We were also interested in exploring these traffic patterns from one year to the next, pre-COVID and post-COVID. Further, we hypothesized that there would be a significant shift in the geographical locations the queries originated from. It was our belief that post-COVID data would show a very clear shift to more queries made outside of Utah.

# 4 Analysis

There was the expectation that there would be a significant shift in traffic patterns pre-COVID and post-COVID. However, after analyzing the data, we were not able to confirm anything conclusive.

[1] RFC1035, RFC 1034 sec4.3.2, 5.3.3

[2] RFC 8499

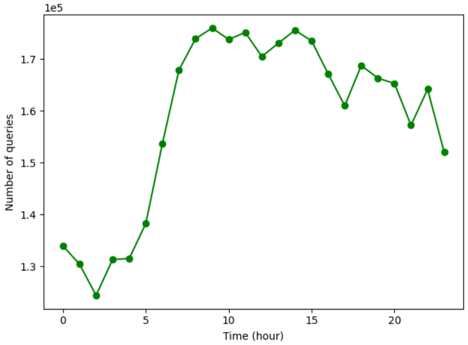

Figure 1: Query rate per hour on April 03, 2019

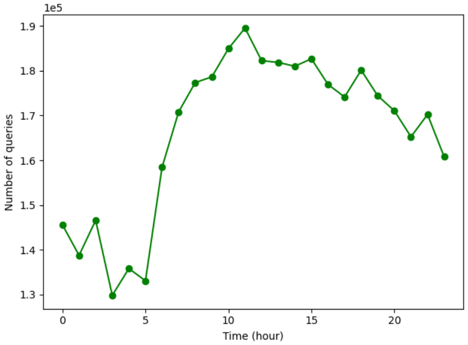

Figure 2: Query rate per hour on April 01, 2020

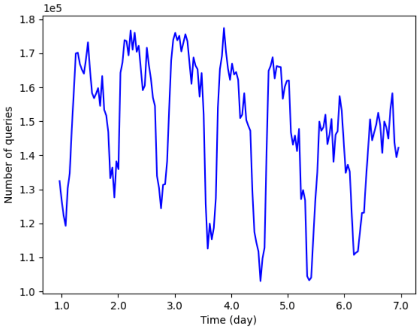

Figure 3: Monday April 1 - Sunday April 7, 2019

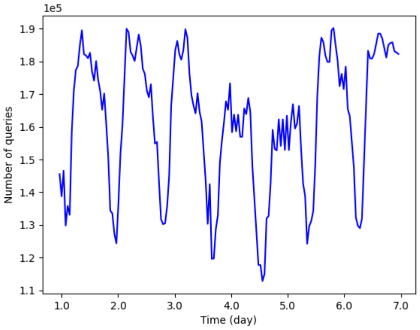

Figure 4: Wednesday April 1 - Tuesday April 7, 2020

## 4.1 Time Analysis

To analyze the query rates with respect to time, we visualized the temporal distribution of the available dataset. To achieve this, we first converted the timestamp associated with each instance to human-readable date and time. We then separated the time from the date and grouped all queries that were generated in the same hour. We now present and discuss the results from our analyses.

**Hourly Analysis** To explore if the traffic amount fluctuated over the course of 24 hours, we graphed the number of queries vs time for each day in the first week of April 2019. We noticed that the traffic had a certain pattern to it. The traffic patterns seemed to correspond with human sleep patterns. Figure 1 shows the number of queries against each hour for 03 April, 2019, a Wednesday. We saw higher query rates during the day and lower query rates later at night. Another thing common for each day in our week's data was that there were two traffic peaks: one around 11AM-12PM and the other around 10PM. These findings were consistent for all days of the week, even weekends.

To see if COVID-19 had any effect on these diurnal patterns, we plotted hourly traffic for each day in April 2020. Our results show that this double peak pattern as well as the diurnal pattern continued over from one year to next. Figure 2 shows the number of queries against each hour for April 1, 2020, which to stay consistent is also a Wednesday. Wang et al. model the cellular traffic patterns of large scale towers deployed in a metropolitan city [4]. We can see the same pattern in their results as we see in ours. This shows that traffic patterns – whether cellular or internet – correspond to human sleep patterns around the world.

**Daily Analysis** We also plotted the number of queries per day over the course of one full week (April 1-7, 2019, as in Figure 3). We observed patterns similar to the hourly analysis, showing that the highest traffic times are during mid-day and at night around 10-11PM. The graphs for one full week make daily patterns more prominent.

We graphed the first week of April 2020. As Figure 4 represents, we found the exact same trend in 2020 as in 2019. Since the first week of April starts on a Wednesday, we see two shorter peaks in the center for this graph as they represent Saturday and Sunday.

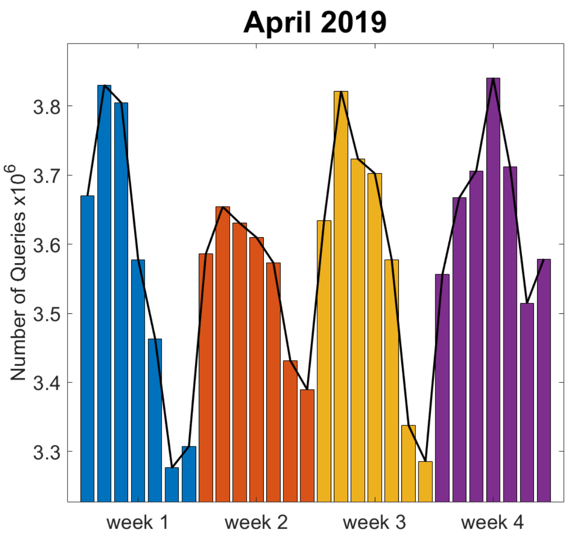

Figure 5: April 2019

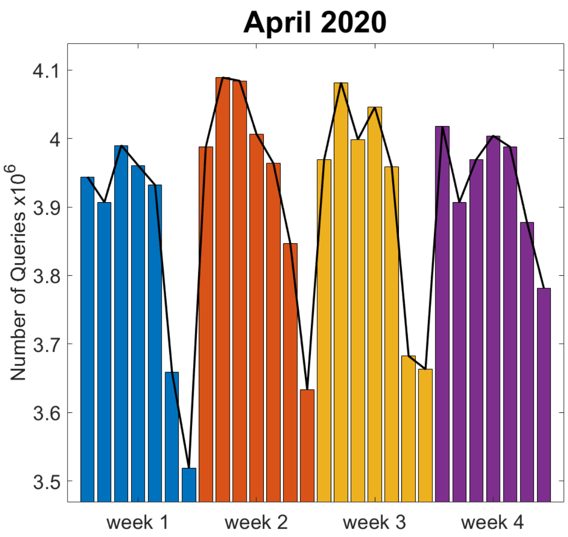

Figure 6: April 2020

**Weekly Analysis** To analyze the weekly patterns in traffic, we graphed the traffic per day for the entire month of April, 2019 (Figure 5). For each week, we saw that the traffic was maximum midweek and minimum during the weekend. As BYU's working days are from Monday through Friday, the traffic patterns align well with that. This does not necessarily have to do with BYU's working days only. As Wang et al. show [4], this generalizes to China too. It is possible that people spend their weekends doing activities that don't rely very heavily on the internet, such as outdoor recreation and family activities.

We then graphed the traffic per day for the entire month of April 2020 (Figure 6). To start with a Monday for consistency, we included March 30-31, 2020.

We see the traffic in 2020 following the same trend: higher during the weekdays and lower during the weekends. This is intuitive too. We also wanted to compare any differences in the total number of traffic per day in April 2019 vs 2020. We found that for each day in April 2020, the traffic on average exceeded that of April 2019 by 300,000. This tells us that internet traffic volume increases every year. To confirm this further, we compare traffic rates per day for 01-21 January 2020 and 2021. We find an even larger increment for this year – about 610,000 per day! (Figure 7).

## 4.2 QType

An analysis of the QType (Figure 8) shows us that the majority of the requests are address requests with 86% of the requests being either of type "A" or the IPv6 version "AAAA". The next largest percentage is of type PTR. PTR is a pointer to a canonical name, but unlike CNAME, with a PTR requests, DNS processing stops and just the name is returned. MX or mail exchange accounted for almost 3% of the traffic, which on four million queries is 120,000 emails, which is inline with our expectations. All other QTypes account for very little traffic.

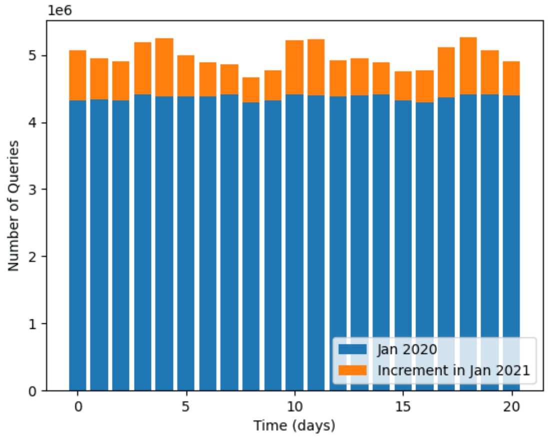

Figure 7: Increase in Traffic January 2020 vs 2021

## 4.3 Server IP Analysis

BYU only has a set number of resolver servers some of them are IPv4 and some are IPv6. In order to see which servers were being queried, a data source was chosen at random and the queries were parsed and each destination IP address query was counted. From figure 9, it is easy to see that IPv4 resolvers

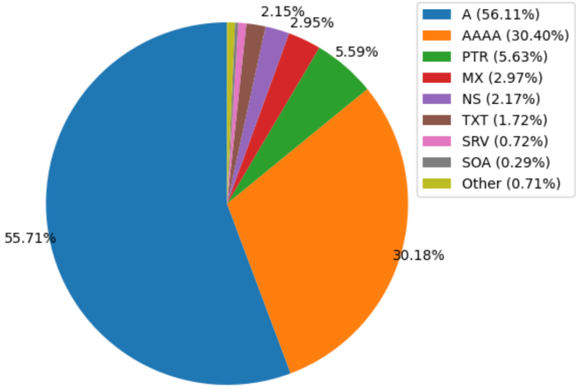

Figure 8: QType Distribution

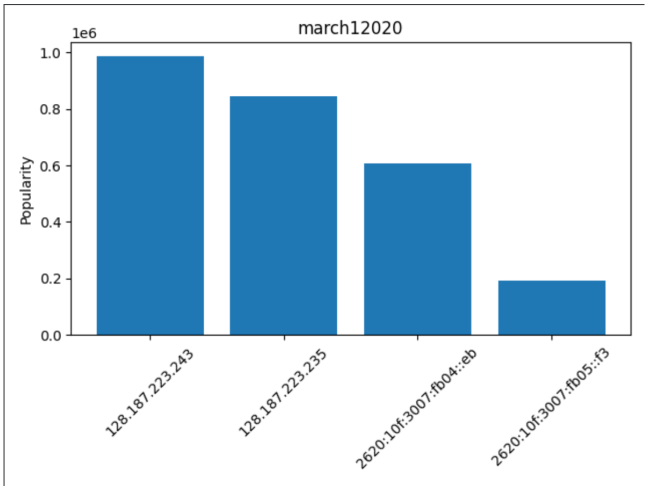

Figure 9: Server IPs

receive the majority of the traffic with over 70% of the queries.

## 4.4 IP Analysis Over Time and Location

We wanted to analyze where client IP's were located. To do this, we first registered an account with an IP-to-address provider in order to get an API key. We then queried the API key, passing the individual IP's and storing the response in a JSON file. One of the problems we had with this method was that we were using a free account which limited our API queries to 1000 per day, so we had to be judicious with our queries. We decided that the most advantageous way to map this would be to rank all the IP clients based on how many times the same client sent a query, then only query the API with the top 1000 clients.

Once we had a JSON file with all 1000 locations, we plotted the results on a heat map (Figures 10).

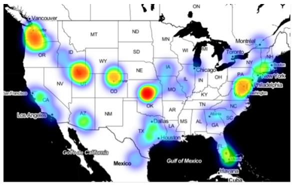

Figure 10: with Kansas

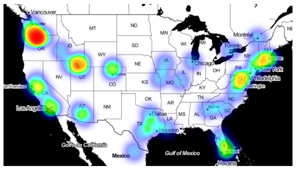

Figure 11: without Kansas

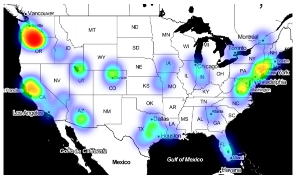

Figure 12: March 4, 2019

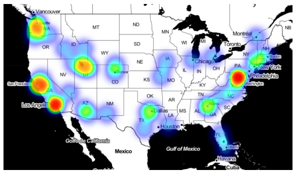

Figure 13: March 1, 2020

From this heat map, it appears that the majority of request on the BYU network are coming from Kansas. This is misleading because when the API cannot determine an exact location but can determine a country, it places the latitude and longitude coordinates in the middle of that country, which happens to be in Kansas. To get a better representation, we decided to remove all coordinates that defaulted to the center of the country. The resulting heat map (figure (Figures 11)aligned much better with our expectations with the exception of the density centered around Portland, Oregon.

We were really interested in seeing if there was a visible correlation in the data from pre-COVID to post-COVID. We postulated that due to COVID and the amount of students that now access classes remotely, the percentage of request coming from outside of Utah would increase, and we would see a broader distribution on the map. However, from our analysis, there is nothing definitive. As shown in figures 12 and 13 there is no discernible difference between pre-COVID and post-COVID results. We are skeptical that there is not a broader geographical distribution post-COVID. Instead, we believe that there is a fault in processing the data. Since we can only get IP locations for the top 1000 queries, we suspect that our analysis is not capturing the whole picture, and is misleading. It is very possible that if we could map all IPs in the dataset as well as all IPs that are defaulting to Kansas, we would have a much better view of what is happening.

## 4.5 IPv6 vs IPv4 Analysis

A very simple analysis that we did was to see if there was a trend towards more clients using IPv6 addresses. We looked at data from 2019, 2020 and 2021, choosing days of the week and times of the month that aligned the best. From Figure 14, we see that there is a slight increase in IPv6 queries each year. Even though 2020 had a lower percentage of IPv6 queries

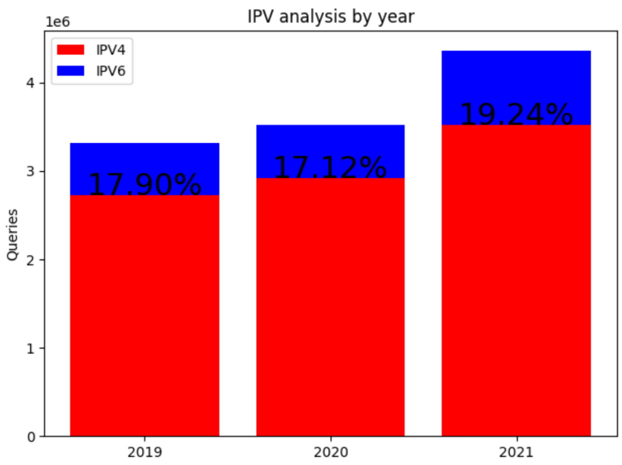

Figure 14: IP Version Analysis

compared with 2019, there is possibly a trend towards more IPv6 queries with 2021 reaching almost 20%.

### 4.6 Port Distribution

Ports are a 16-bit number giving us 65,535 unique port numbers. Some ports are used more frequently than others and different operating systems tend to favor some ports over another. When the Port distribution was graphed (Figure 15) it was clear that even though there are some specific ports that are selected four to five times more frequently, the vast majority fall within a very clear probability distribution. The other trend that is easy to see from the figure, is that higher ports are chosen at a slightly higher frequency than lower ports.

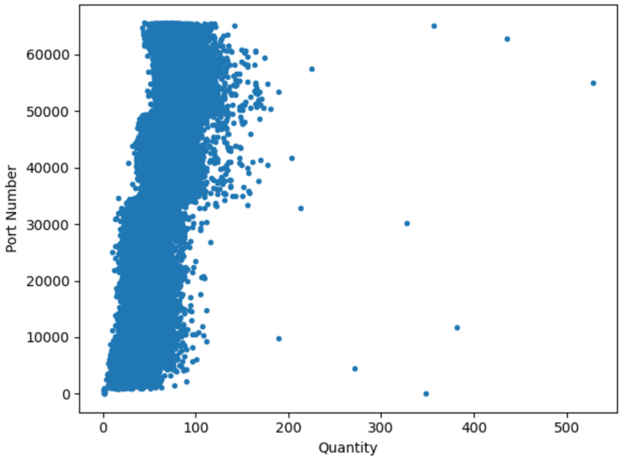

Figure 15: Port Distribution

### 4.7 Flag Distribution

The query log contains a flag field. From the Bind documentation [2] we get what each one of these flags means which has already been discussed in Section 2.6.

From Figure 16 we notice that Recursion Desired is almost never set. From the documentation [3] it appears that stub resolvers almost never set this bit. Generally only full service resolvers such as named resolvers flip this bit, since they are capable of taking the referral and using them to follow the delegation chain all the way to an authoritative nameserver. Stub resolvers, in general, do not have this capacity and therefore leave this bit unflipped. [2]

If we sum up all the E flags from -E, -ED, -EDC, we notice that almost all the queries are using EDNS, a hop-by-hop extension to DNS [2]

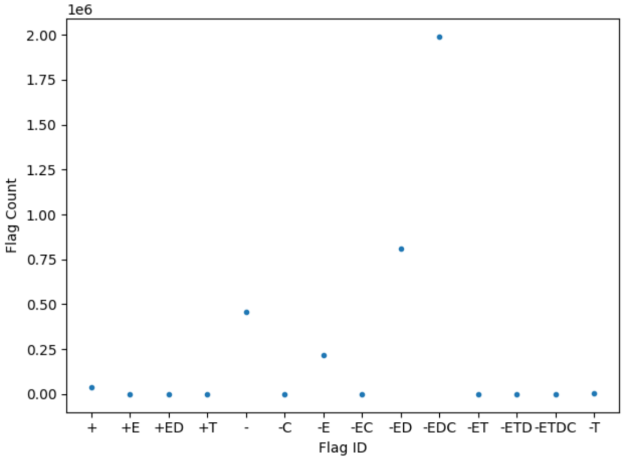

Figure 16: Flag Distribution

## 5 Future Work

The query data that we were provided is detailed enough to allow many different analyses. However, time did not permit us to perform all of the analyses that we wanted to perform. It is possible that a recursive server might make the same request using IPv4 and IPv6 at about the same time, but we were unable to perform this analysis due to time constraints.

Making the same request using both IPv4 and IPv6 raises some interesting questions. First of all, why would a recursive server make two such requests rather than simply making one request using either IPv4 or IPv6? Is it possible that one query will fail where the other would succeed? If so, which is more likely to fail? Is it possible that one type of query tends to return results faster or more securely than the other? Also, although IPv6 accommodates a greater number of devices to be connected to a network, IPv4 is still commonly used. At

[3] RFC1035, RFC 1034 sec4.3.2, 5.3.3

some point, dropping IPv4 entirely in favor of IPv6 might be worth it. Exploring recursive servers' IPv4 and IPv6 query trends might provide valuable information in determining when the switch from IPv4 to IPv6 could be made.

## 6 Conclusion

Interesting spatial-temporal patterns are identifiable by carefully analyzing DNS query logs. Peaks and valleys in the number of queries are cyclic and appear to reflect both biological and cultural norms of human behavior. For example, over the course of a 24 hour period, the number of queries from around midnight to 6:00 AM lowers significantly compared to the daylight hours, with a gradual transition between the two time periods. This diurnal pattern matches diurnal human nature. Also, the number of queries is lower during the weekend than during weekdays, reflecting the culture's use of weekdays for tasks such as school and work and emphasis on social and personal matters on the weekends.

The spatial patterns presented in this paper cannot directly reveal spatial trends of end users, since the query logs did not contain information regarding the location of the user who initiated the request. Rather, spatial patterns represent the recursive servers that make requests to authoritative servers.

Stub resolvers appear to be the primary senders of queries represented by the data, since the Recursion Desired flag is rarely set. Also, the overwhelming usage of -E indicates that the vast majority of queries utilize an EDNS.

Lastly, IPv4 is used at a much higher rate than IPv6. The percentage of IPv6 queries decreased slightly from 2019 to 2020, but increased significantly in 2021. It is also notable that the number of queries overall increased each year and that the increase between 2020 and 2021 was significantly larger than that of the previous year. This might be an effect of the Covid-19 pandemic forcing people to spend more time at home, where the internet is a readily available facilitator of entertainment, shopping, and working. Since query logs were only collected from the first month of 2021, it would be interesting to see if March and April show a similar increase in internet activity compared to 2019 and 2020, keeping in mind the state of the Covid-19 pandemic during those upcoming months.

## References

[1] IANA assignments, *https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml*, Feb 2021.

[2] IETF documents, *https://tools.ietf.org/html/*, Feb 2021.

[3] UNIX timestamp, *https://www.unixtimestamp.com//*, Feb 2021.

[4] Huandong Wang, Fengli Xu, Yong Li, Pengyu Zhang, and Depeng Jin. Understanding mobile traffic patterns of large scale cellular towers in urban environment. In *Proceedings of the 2015 Internet Measurement Conference*, IMC '15, page 225–238, New York, NY, USA, 2015. Association for Computing Machinery.