Where is Your Mail? Analysis of US Government Domain Mail Server Geolocation

Kevin Hayes, Jay Park, Hong Zhou

ABSTRACT

Electronic communication is ubiquitous across personal, industrial, and government circles. Indeed, as the prevalence of tools such as instant messaging, online chats and, in particular, emails continues to grow globally, so does the incentive of third parties to tamper and surveil the sensitive information transmitted across the modern internet. With rising political interest in cybersecurity - particularly cyber attacks - and each government's interest in data security and sovereignty, we seek to partially address the issue of mail server security in known government domains. Works in this and adjacent fields have been published in the recent past, with [3] analyzing the roles network architecture and Route Origin Access plays in mail server security. To put these previous findings in a real world context, we geolocated a numerous list of mail server IP addresses found through DNS queries on the country-level. We then mapped the IP addresses with their country of use and the geolocated nation.

1 Introduction

Over the past decade, communication over the internet has become crucial for both personal and professional communication. Email in particular is used everyday to convey private information that could be potentially sensitive for one or more parties. As such, the security of the networks on which these systems relay data is of utmost importance to most major players of the internet. One such prevalent party with interest at stake is the U.S. Government.

As the government places more emphasis on online services, especially due to the COVID-19 pandemic, the trustworthiness of emails from government domains is extremely important for citizens. This increased importance also makes government email servers a crucial part of national security and a clear target for attack. As such, whether or not the email infrastructure of one's government is trustworthy is important, either in order to provide peace of mind when interacting with government organizations over email, or to caution citizens when receiving emails from government domains.

In this paper, we aim to characterize the present usage of mail servers by United States governments. This is done by measuring factors such as the configuration of the mail servers, the redundancy/overlap of servers across different domains, and how extensive the use of third-party servers is by *.gov* domains. This is all done with an emphasis on how the different levels of government manage their infrastructure differently, from city governments to the federal branches. as well as the physical locations which mail servers are located.

We perform DNS lookups on numerous *.gov* domains to extract their associated mail servers and their IP addresses. In order to find where the mail servers are located, we use IPInfo to map IPs to geolocations. If the analysis shows that the mail server is located in a foreign country, it could be cause for concern.

2 Contributions

Our main contributions in this study are as follows:

- Using geolocation to gain preliminary understanding of the physical security of government mail servers. In this area, we found that while the majority of mail server locations used by *.gov* domains reside within the borders of the United States, a small percentage are being hosted outside of the country.
- An analysis of the prevalence of first and third-parties in mail servers and their physical locations. Verifying previous studies, and making new contributions by determining how the different levels of government treat third party providers differently.
- Measuring the redundancy in mail servers used by *.gov* domains as well as the centralization of mail servers to determine the reliability/vulnerability of government services.

3 Background

3.1 Mail Servers

In order to enable e-mail services for a domain, the owner of a domain must add DNS MX record(s) to the set of records for their domain. Each of these records contains an *exchange* field, which contains a domain name, corresponding to a mail server which is responsible for sending and receiving

mail on the behalf of the domain. In this paper we will refer to these servers as *mail servers*. A single domain can be served by multiple mail servers, either by listing multiple MX records, each with a different domain. Or through other load balancing techniques such as mapping multiple IP's to a single mail server domain, or even by mapping multiple servers to a single IP by utilizing IP anycast. In practice though we've found simply listing multiple MX records to be the most common approach used.

In order to differentiate between these different mail server domains, MX records contain an extra *preference* field, which is a single number used to indicate the order in which mail servers should be contacted. For example, if *cia.gov* contains two MX records, one for *mail0.outlook.com* with preference 10, and another for *mail1.outlook.com* with preference 20, then when sending mail to a CIA email account, one should first try to connect to *mail0.outlook.com*. And then only resort to using *mail1.outlook.com* if the first server is unresponsive, because the first server has a lower preference number, and thus a higher priority.

3.2 Prior Work

Studies have been done on mail domain zones as well as the zones containing mail servers for said mail domain (referred to as Direct Zones). Specifically, research on the security of mail servers used by various countries has been done by studying these direct zones, such as the path redundancies of the mail servers, the DNS redundancies for each zone, mail server redundancy, etc. Current studies also examine the existence of "Route Origin Authorization" in the mail domain level, network level, and AS level. However, existing research limits their direct zones only to network and Autonomous Systems.

Because previous work has shown the trend among government organizations to utilize mail providers, we'd like to determine where these mail servers are physically located. Effectively answering the question of "where's your mail" to go alongside "who's got your mail" question posed by [6]. If these servers are located in a foreign country, that creates a concern for the security of the servers.

[6] also introduces a methodology for determining the mail provider from a given domain, and applies that methodology to the U.S. government's *.gov* domains, in order to examine the use of third party email providers by different branches of the U.S. government. We make a slight modification to that approach so that it better serves our interests.

4 Methodology

4.1 Data Sets

For each domain in our set of government domains [2], we first run DNS queries to obtain the MX records of that domain, creating a map from government domain to mail server domain. Then we can take those records and run another set of queries to get the IP addresses associated which each mail domain through DNS A records.

This process is relatively straightforward, and we have scripted the entire process, from a list of domains, to running a set of analyses on the final data, using python. The code and final set of data and graphs can be found at [5] in order to facilitate reproducibility and possibly facilitate future analysis of governments other than the U.S.

Once we have our final mapping from government domain to mail server(s) and corresponding IP(s), we then we then can take the set of IP addresses and cross reference them against two different datasets to determine their geographic and network information.

4.2 Geolocation

First we reference each mail server IP against the IPInfo [1] geo location service's database. This gives us a wide range of information about the physical location of the IP, ranging from broad classifications such as which country the IP is thought to be in, to a specific longitude and latitude point on the globe. When determining a *location* in the analysis of the data, we are referring to a specific pair of latitude and longitude, because of the wide range of detail levels in the IPInfo dataset.

There are pros and cons to this approach of determining locations, but in general we believe it should be relatively unbiased, because it can both over and under count in different situations. For example, if two servers are located in different regions of the same state, but IPInfo does not have any more information than the state, then they will both be mapped to the center point of the state, and considered to be the same location, thus an undercount. However if two servers are truly co-located, but IPInfo has a different levels of detail for each, then we would not be able to tell the difference, for example, if one is mapped to the center of the state, and the other is mapped to the center of the specific city.

4.3 ASN's and IP Anycast

The IPInfo dataset also contains entries indicating the Autonomous System Number associated with the IP and whether or not it is an anycast IP. We do utilize IPInfo for analysis related to IP anycast, though we refer to a more reliable CAIDA dataset [4] for mapping IP's to specific ASN's. We are unsure how reliable the IPInfo dataset is for determining anycast IP's, and in the future would like to use a dataset specifically for anycast.

4.4 Mail Server Preferences

Because we query for all MX records of each domain, we can determine the realtive "rank" of each mail server from the perspective of each domain. We do this in order to determine how complex the policies behind domains with multiple MX servers. If we see a wide range of perferences, then that indicates that there is a large number of "backup" mail servers, which are not being used normally. Or if most servers are ranked similarly, and are thus used more or less equally. This also serves as an indication for the importance of each mail server from the perspective of the government

domain, and restrict some analysis to only servers which are ranked higher.

4.5 Third Party Mail Servers

Finally, we measure the usage of third party mail servers, as opposed to the government using their own servers to facilitate mail transfer. We used an extremely coarse method for determining third party servers, analyzing the mail domain name itself, and classifying each mail domain as either, first party, third party, or unknown.

We use a set of regular expressions trying to match common third party mail providers domain names to each server. And concluding it's a third party if we find a match. If we cannot find a match, we then check to see if the domain is itself within the top level *.gov* domain, and if we conclude it is a first party server, hosted by the US government. If both methods fail, then we mark the mail domain as "unknown".

This method has major drawbacks: the third party analysis would not be comprehensive, as our hard-coded list of third parties would not be exhaustive. And it makes the task of associating servers with a specific company more difficult, as each rule needs to be manually created, and checked to make sure it is owned by the correct company. It also does not allow us to conclude which entity within the *.gov* domain is hosting the server. So it is possible that one part of the U.S. government is hosting a mail server for another, and hence not truly a first party server.

However, despite these drawbacks, we were able to classify upwards of 80% of domains, so we chose to use this relatively simple approach.

An arguably better, though more complicated, method is instead to utilize an IP to ASN mapping, and then using WHOIS queries to determine which entity owns the specific ASN. We can check if a third-party provider is being used if the ASN does not belong to the US Government and concretely confirm the identity of the third party itself. This approach assumes that the owner of the mail server also owns the AS which the server is located within, which may not be true, especially for lower levels of goverment such as smaller cities and counties. In the future we would like to utilize a combination of domain name analysis and ASN ownership to make our determination.

5 Results

5.1 IP Anycast

We would like to begin our analysis by noting the unknown degree of influence which anycast IPs affected our data. During data collection, we noticed a total of [Insert number here] servers using IP anycast. Through anycast, a single domain could potentially use MX exchanges in multiple locations with the same IP. This means that the count of physical locations that any single domain relies on is actually an undercount due to the influence of IP Anycast. This is especially true for several specific locations, such as San Francisco, Seattle, Kansas City, San Antonio, and Jacksonville.

5.2 Mail Server Locations

The vast majority of mail servers' locations found are within the US. However, there are a total of ??? locations beyond the borders of the United States. Most of these locations are in US-friendly countries, notably France, England, Japan, and South Korea. However, there are notable outliers such as one in the coastlines of China.

5.3 Mail Server Centralization

We created a CDF of the number of physical locations a single *.gov* domain has listed as a mail exchange. We see that 11% of *.gov* domains were mapped to a single physical server, while the remaining 88% map to at least 2 physical locations. We also see that the most prevalent number of locations used is 4 locations, followed by 7 locations. This shows that most *.gov* domains have redundant systems, which is good for reliability.

6 Limitations

Our results revolve heavily around IPInfo's geolocation service, which means that any inaccuracy in the part of IPInfo influences our findings as well. As mentioned before, the influence of IP Anycast is also something that we fail to account for, meaning that our count of mail servers in a single location is actually a lower bound to the true count. Finally, using the preference fields for MX records for preference analysis leaves a degree of ambiguity, as we cannot tell whether a domain puts a primary server and multiple backup servers, or whether a domain load balances equally among the servers.

7 Future Works

While this study is a strong step towards better understanding of the geolocation of government mail servers, there are nonetheless further avenues for a more robust understanding of the geography of domain, IP, and mailserver distribution. Most notably is a longitudinal expansion into analyzing the geo locations of mail servers beyond the officially-listed *.gov* domains. This would include the analysis of the domains of other governments, as well as non-government entities. Afterall, concern for security is not limited to public services, as other major firms such as Microsoft, Google, Vanguard, etc. all hold sensitive information that could be transmitted across the email architecture.

Furthermore, a thorough analysis of the mail servers themselves beyond geolocation could provide insight beyond physical security into the relevant cyber security of mail servers. This could be done by applying the methodology outlined in [3] in the case of ROA protection. Only through considered evaluation of both physical and cyber security could the modern email system be adequately secure.

Beyond mail servers, accurate country-level geolocation of critical virtual services should be assessed. This includes major databases and other infrastructure backing the modern internet. As technology and the Internet continues to evolve

and change, it is vital to continuously curate and understand the relationship between our physical world and the virtual one to ensure that our private and sensitive information remains private.

8 Conclusion

9 References

[1] Ip info geolocation api free account. https://ipinfo.io/products/ip-geolocation-api, 2024. Accessed: 2024-2-29.

[2] List of registered .gov domains. https://raw.githubusercontent.com/cisagov/dotgov-data/main/current-full.csv, 2024. Accessed: 2024-2-9.

[3] A. Bartoli. Network architecture and roa protection of government mail domains: A case study. *Computer Communications*, 201:143–161, 2023.

[4] E. Carisimo. Ip to asn mapper using caida. https://github.com/estcarisimo/map-ip-to-asn, 2024. Accessed: 2024-3-4.

[5] K. Hayes, J. Park, and H. Zhou. Project source code. https://github.com/hongZhou443/MSS, 2024.

[6] E. Liu, G. Akiwate, M. Jonker, A. Mirian, S. Savage, and G. M. Voelker. Who's got your mail? characterizing mail service provider usage. In *ACM Internet Measurement Conference (IMC '21)*, page 15, Virtual Event, USA. ACM, New York, NY, USA, Nov. 2021.