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 [2] 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 assess the trustworthiness of a mail server by seeing if third-party mail servers are used, as well as looking at the country in 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 IP-Info 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

3 Background/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 [3]. If these servers are located in a foreign country, that creates a concern for the security of the servers.

[3] 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 Collection

For each domain, we first run nslookup to obtain the MX records of that domain, which contains information about the mail servers that the domain is using. However, these MX records could be misleading; [3] describes how, for example, we may infer that gsipartners.com self-hosts their email servers because their MX record lookup returns mailhost.gsipartners.com. Upon further inspection however, the mail server domain mailhost.gsipartners.com resolves to an IP address announced by Google, which means that the mail service is actually hosted by Google. Following the MX record, we ran A record lookups on the mail server domains to get the IPs associated with the domain names. We compile these mail server domains and their corresponding IP addresses into a csv file for further analysis.

Getting the IP address is especially helpful, as we can use a geolocation service to map IPs to their approximate physical locations. Our geolocation service of choice is IPInfo, chosen out of convenience.

The entire process is automated using Python scripts that run terminal commands such as nslookup by spawning a process. We then obtain the list of *.gov* domains from a repository maintained by [1]. Our script loops through the list of *.gov* domains and runs the analysis suite, creating a csv file with our data collection results.

4.2 Analysis

4.2.1 *Mail Server Preferences*

The purpose of analyzing the mail server preference of each government domain is to see if certain third party providers are being preferred over another third party provider (or even a first party provider). In order to do this, we utilize the MX records' "preference" field included in our data collection of all domains. The lower the preference number, the higher the priority. We first sort the preferences of the mail servers, grouped by the government URL. Using this sorted data, we collect the highest priority mail server domain for each government domain. We also collect the second highest priority mail server domain (if it exists). We then analyze the frequency of each third party provider within both of these lists.

4.2.2 *Mail Server Centralization*

We then measure the centralization of mail server locations by creating two mappings. The first is the number of domains to a physical location, which allows us to determine whether there exists an over reliance by many different government services, to any single mail server location. The second is the mapping of physical location to a single *.gov* domain, which allows us to analyze the physical redundancy of mail servers utilized by any single government service. In essence, we ask the question "if a crucial mail server went down, how many government servers would be rendered unavailable?". We also see if these distributions change significantly within and without the borders of the United States. Our analysis anonymizes the data garnered from our automated script.

4.2.3 *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. One option for determining the third party of choice given the domain name is to analyze the name itself, looking through a hard-coded list of third party providers and searching for a match. However, 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. Furthermore, if a third party provider has different variations of domain names, our list would have to capture all possible variations of said domain names. Even a slight variation (such as "cloudflare.us" to "cloudflare.usa") would cause the program to mark it as "not a third party provider" if it is not in our list.

Our method of choice is instead to utilize an IP to ASN mapping, and then look up what third party provider is associated with that ASN. Looking up which provider is associated with an ASN is more effective compared to a static analysis of the domain name. 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. Through this alternative, we avoid the pitfalls of having a hard-coded list of third party providers while guaranteeing a more accurate assessment.

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

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

[3] 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.