Understanding and Characterizing Intermediate Paths of Email Delivery: The Hidden Dependencies

Ruixuan Li
Tsinghua University
Beijing, China
Beijing National Research Center for Information Science and Technology
Beijing, China
lirx25@mails.tsinghua.edu.cn

Chaoyi Lu
Zhongguancun Laboratory
Beijing, China
lucy@zgclab.edu.cn

Baojun Liu
Tsinghua University
Beijing, China
Beijing National Research Center for Information Science and Technology
Beijing, China
lbj@tsinghua.edu.cn

Yanzhong Lin
Coremail Technology Co. Ltd
Guangdong, China
tim@coremail.cn

Haixin Duan
Tsinghua University
Beijing, China
duanhx@tsinghua.edu.cn

Qingfeng Pan
Coremail Technology Co. Ltd
Guangdong, China
pqf@coremail.cn

Jun Shao
Zhejiang Gongshang University
Hangzhou, China
Zhejiang Key Laboratory of Big Data and Future E-Commerce Technology
Hangzhou, China
chn.junshao@gmail.com

Abstract

In the cloud era, hosting-based email services have become a common business model. Various entities can participate in the email delivery process. However, the intermediate paths of email delivery have received little attention. In particular, the vulnerabilities and centralization of email intermediate paths have already posed real-world security threats.

This paper conducts the first systematic analysis of intermediate paths of email delivery, aiming to understand dependence patterns and characterize the centralization. In collaboration with a large email service provider, we collected *Received* headers from email reception logs spanning nine months and reconstructed the complete intermediate paths of 105M clean emails. Our results reveal that Microsoft is the dominant provider of intermediate paths, participating in 66.4% of emails. We find that 86.9M (82.7%) emails rely on third-party providers in intermediate paths, and 9.1M (8.7%) paths involve multiple providers. Email signature providers frequently appear in cross-vendor intermediate paths. In addition, we reveal significant differences in the regional dependencies and centralization of email intermediate paths across countries and continents. The centralization observed in email intermediate paths also differs from incoming and outgoing servers. We hope our work prompts more attention to email intermediate paths to enhance the security of the email ecosystem.

CCS Concepts

• **Networks** → *Application layer protocols*; *Naming and addressing*; • **Security and privacy** → *Security protocols*; • **General and reference** → *Measurement*.

Keywords

Internet measurement; Email delivery path; Email dependence; Email centralization

ACM Reference Format:
Ruixuan Li, Chaoyi Lu, Baojun Liu, Yanzhong Lin, Haixin Duan, Qingfeng Pan, and Jun Shao. 2025. Understanding and Characterizing Intermediate Paths of Email Delivery: The Hidden Dependencies. In *Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25), October 28–31, 2025,Madison, WI, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3730567.3764488

IMC '25, Madison, WI, USA.

ACM ISBN 979-8-4007-1860-1/25/10
https://doi.org/10.1145/3730567.3764488

1 Introduction

Email plays a vital role in global information exchange and identity authentication [44]. In the cloud era, although email services can still be independently deployed by individual organizations, hosting-based email services have become a common business model. For example, 29% of Alexa Top 1M domains are reportedly relying on Outlook to receive emails [32]. As a result, the traditional "*end-to-end*" delivery model, i.e., emails are sent directly from the sender's server to the recipient's server, is changing. Various types of entities are now involved in the email delivery process [31, 42], such as hosting providers, forwarding servers, and security vendors.

Modern email delivery paths are gradually evolving from an "*end-to-end*" model to a "*segment-to-segment*" model, in which emails traverse multiple middle nodes.

When emails are transmitted sequentially through middle nodes, the security of the entire transmission path will be compromised if any single middle node is vulnerable. Worse, large-scale threats can arise if one vulnerable node becomes a dependency for a large number of important email transmissions (i.e., email services develop *centralized* dependencies on these nodes or providers). For example, the email services of many Fortune 100 companies (e.g., IBM and Disney) have leveraged Proofpoint, a popular provider for spam filtering. By exploiting the lax source verification policies at Proofpoint relays in the email intermediate path, attackers have successfully spoofed emails from many Fortune 100 domains and sent millions of phishing emails [16]. In 2024, Rao et al. [42] also revealed that attackers can abuse the relaxed source restrictions of third-party email filtering services to bypass protections along the email transmission path. Among the 1.6K domains investigated, 80% were found to be vulnerable to such threats.

Research gap. Traditionally, the degree of, and risks behind Internet centralization have been acknowledged and extensively studied in areas including DNS [27, 38] and Web [3, 51]. Regarding email, some works have measured centralization of incoming and outgoing "ends" of the delivery paths (i.e., incoming and outgoing servers) by inspecting data embedded in MX and SPF records [32, 47, 50]. However, traditional datasets have been lacking visibility into intermediate entities in email transmission paths; dependencies and potential risks associated have thus been overlooked by prior studies and remain unevaluated. As a result, we believe investigating intermediate paths can offer new insights into the architecture of email transmission and provide a guide for improving the security of the email ecosystem.

Our study. In this paper, we aim to unveil the picture of intermediate paths of email delivery, find hidden dependencies, and evaluate the degree of centralization. While traditional active methods (e.g., querying MX records) do not reveal the intermediate paths, the task is made possible by inspecting *Received* headers in email content, which record the sequence of middle nodes that handled the email during transmission. To collect real-world email intermediate paths, we collaborated with Coremail [9], a large email service provider, to analyze 2.4B emails received within nine months. By building an email path extractor to retrieve node information from *Received* headers, we successfully recovered intermediate paths for 98.1% of the emails. Eventually, we are able to inspect 105M clean (determined by Coremail) emails with complete intermediate paths. This unique dataset allows us to address the following research questions: *1) What are the identities and distribution of email middle nodes? 2) What is the dependency structure and regionality of email intermediate paths? 3) What are the centralization degrees and cross-country differences of email intermediate paths?*

Major findings. We find that a large proportion of email intermediate paths rely on a small number of entities. Specifically, 42.8% of intermediate paths traverse 5 autonomous systems (ASes), and outlook.com appears as a middle node provider for up to 66.4% of email deliveries. Email signature and security filtering providers also serve as common middle nodes. Furthermore, we analyze the dependency patterns of email intermediate paths, and find that 86.9M (82.7%) email intermediate paths rely on third-party providers. 9.1M (8.7%) emails involve middle nodes operated by multiple providers, with the most frequent email passing relationship occurring between email service providers (e.g., outlook.com) and email signature providers (e.g., exclaimer.net).

At the country level, we observe significant discrepancies in dependency patterns. For example, in Russia and Belarus, around 30% of email intermediate paths are domain self-hosted, significantly higher than in other countries. In addition, the dependence of email intermediate paths on external regions is influenced by sociopolitical, linguistic, and geographical factors. For instance, countries in the Commonwealth of Independent States (CIS), such as Belarus (88%) and Kazakhstan (32%), heavily rely on Russia's email infrastructure. At the continental level, email intermediate paths of Africa show strong dependence on Europe and North America, while South America exhibits a high dependence on North America.

Finally, through the Herfindahl-Hirschman Index (HHI) [48], we find that the market for middle nodes is highly concentrated. While the degree of centralization varies across countries, outlook.com holds a dominant market share in most of them. Moreover, we compare the centralization of middle, incoming (corresponding to MX), and outgoing (corresponding to SPF) nodes. Our results reveal that the incoming node market is the most centralized, and that the dominant providers differ across the three types of nodes.

Contributions. Contributions of this paper include:

- Using a unique and large-scale industrial email dataset, we identify middle nodes and unveil intermediate paths of email delivery, one missing piece from previous studies.
- We systematically analyze hidden dependencies and evaluate the centralization degree of email intermediate paths.
- We publish our email path extractor and intermediate path dataset (at https://github.com/RUI-XUAN-LI/Email_Path) for facilitating future research.

2 Background

In this section, we first describes the email delivery path, with a focus on intermediate paths. Then, we introduce email formats and common email headers, especially the *Received* header. Finally, we outline the current state of email centralization and the email path dependency risks.

2.1 Email Delivery Paths and Middle Nodes

Individuals and organizations can independently deploy their email services. In such cases, an email travels directly from the sender's client to the outgoing server and is then delivered to the incoming server. No intermediate servers are involved in such an email delivery process, and the delivery path length is 1. With the rapid development of cloud services, hosting-based email deployments have been adopted by a large number of domains [32, 47]. As a result, the traditional "*end-to-end*" email delivery model, that is, the outgoing server directly connects to the incoming server, is undergoing changes.

Figure 1 illustrates the email delivery process involving intermediate servers. Emails from the sender's client pass through one or more middle nodes before reaching the outgoing server. Therefore, the email delivery path exhibits a "*segment-to-segment*" mode.

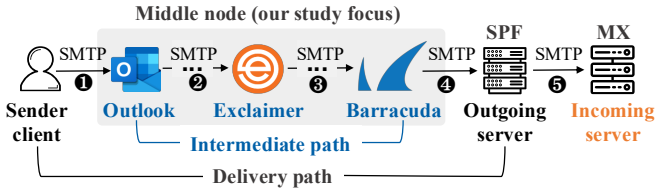

Figure 1: The delivery path of an email, from the sender client to the incoming server.

In this paper, we define the outgoing node as the server that directly establishes a connection with the incoming server (incoming node), and the middle node as the relaying entity located between the sender's client and the outgoing node. In particular, we focus on middle nodes that operate at the application layer (e.g., using SMTP [40]) and are capable of understanding email headers and content. Devices that operate solely at the network layer, such as routers, are excluded from our definition. Below, we introduce four common types of email middle nodes.

• **Email hosting provider** offers hosted mailbox services for enterprises or individuals, managing the reception, storage, sending, and account administration of emails. These services typically integrate Webmail access and IMAP/POP3 protocols. Common public cloud-based hosting providers include Google Workspace [15], Microsoft 365 [35], etc.

• **Email forwarding provider** is responsible for automatically redirecting received emails to another designated email address. Many companies support custom forwarding domains and rules for their subscribers, such as GoDaddy [13]. Additionally, most email service providers (e.g., Gmail [14]) allow users to configure their mailboxes to automatically forward incoming emails.

• **Email signature provider** typically offers branding and signature management for outbound corporate emails. An email signature refers to the consistent content appended to the end of each email body, often including personalized text or graphics such as company logos, job titles, and contact information. Companies like Exclaimer [12] and CodeTwo [8] offer professional email signature services and are used by many Fortune 500 enterprises.

• **Email filtering provider** performs security checks on inbound and outbound emails, such as spam and virus, to protect customers' email security. Popular email filtering providers include Proofpoint [41], Barracuda [4], etc.

2.2 Email Formats and Common Headers

The email typically consists of three parts: the SMTP envelope, the email header, and the email body [7]. Figure 2 shows a simplified example of an email. The SMTP envelope contains email routing information, including the sender address (*Mail From*) and the recipient address (*Rcpt To*). Email servers use the envelope to determine the source and destination of the email. The email headers and email body form the content of the email. The headers contain metadata, such as sender, recipient, subject, timestamp, and are composed of multiple "Field: Value" lines, each separated by "\r\n". The email body is the actual content written by the user.

The *Received* header records each node that an email passes through from the sender's client to the incoming server, and plays a

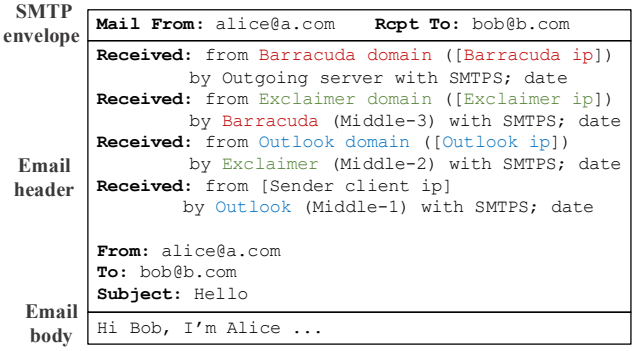

Figure 2: Example of a simplified email.

critical role in email path tracing, troubleshooting, and performance analysis [34]. Figure 2 shows an example of *Received* headers, corresponding to the email delivery path in Figure 1. Every server that processes the email, including both middle nodes and the outgoing node, adds its own *Received* line to the top of the email content. Therefore, *Received* headers are arranged in reverse path order, with the last node appearing at the top. Each *Received* header generally has two key components: the *from part* and the *by part*. The *from part* records the information of the source server (previous node), and the *by part* records the information of the receiving server (current node). The information in the *Received* header usually includes: domain name, IP address, timestamp, encryption suite, email protocol, etc.

2.3 Email Concentration and Dependency Risks

With the rise of third-party hosting services, email systems, much like other Internet infrastructure, such as DNS, have exhibited a clear trend toward centralization [3, 38]. A key driver is that hosted providers significantly lower the technical barrier for users to operate email services. These providers also offer high service quality, advanced anti-spam capabilities, DDoS protection, and more. Typically, users only need to configure a few DNS records for their domain to enable email hosting services. The centralization of email services introduces significant risks, primarily in terms of user privacy and single point failure [6, 24, 28].

Previous research on email dependence and centralization mainly relied on MX and SPF records. In 2021, Liu et al. [32] analyzed the providers behind MX records and revealed that 29% of the email reception services for Alexa Top 1M domains relied on Google, and 11% on Microsoft. Some scholars also used SPF records to analyze the centralization of email delivery services [26]. For example, Wang et al. [47] reported that 20.1% of Tranco Top 1M domains relied on Outlook for delivering emails, and 15.7% depended on Google.

Currently, only a few studies have preliminarily investigated the email intermediate path, but none of them have systematically revealed the dependency patterns and centralization characteristics. For example, Luo et al. [34] used *Received* headers to uncover the infrastructure of phishing emails. Rao et al. [42] revealed the security risk of incomplete protection in email transmission paths, i.e., attackers can bypass email filtering providers and send spam

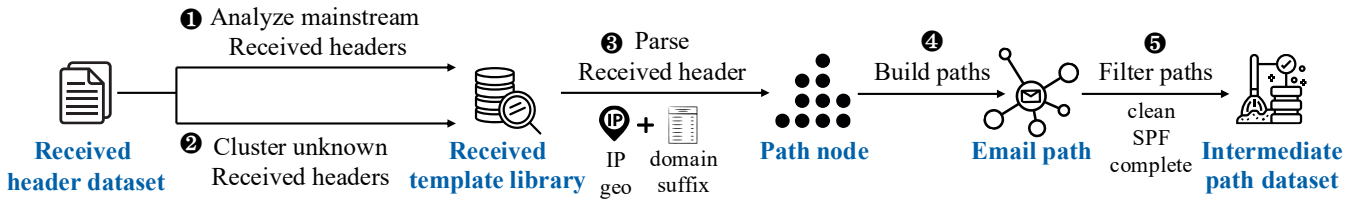

Figure 3: Workflow of constructing email intermediate path dataset.

directly to victims. By analyzing 15 popular cloud-based email filtering services, they found that 80% of 1.6K examined domains were vulnerable to incomplete path protection.

Unfortunately, attackers have already exploited the dependencies and centralization of email intermediate paths to carry out malicious activities. As an example, Guardio Labs disclosed the *EchoSpoofing* vulnerability in 2024 [16], which allows attackers to exploit the relaxed source restrictions of Proofpoint relays in the email intermediate path to spoof victim domains. More seriously, Proofpoint offers security protection for 87 Fortune 100 companies, such as Disney, IBM, and Nike, allowing attackers to impersonate a wide range of well-known brands. Ultimately, the EchoSpoofing vulnerability led to the distribution of millions of highly deceptive phishing emails.

3 Methodology

In this section, we first introduce our email *Received* header dataset from a large email service provider. Then, we describe the process of constructing email delivery paths. After that, we present an overview of our intermediate path dataset.

3.1 Email Received Header Dataset

This paper aims to analyze the email intermediate paths between the sender's client and the outgoing node. MX and SPF records can only roughly represent the incoming and outgoing nodes of domain-associated emails. Moreover, many SPF records contain overly large ranges of IP addresses [10], not all of which actually work for delivery emails. To this end, we use the *Received* header to understand the email delivery path, which has been adopted by some studies [30, 33, 34].

The *Received* header is included in the email content received by the incoming server. It is impractical to actively instruct numerous domains to send emails to our controlled server. The main reason is that it is difficult to obtain email accounts for numerous domains, especially considering that many email services are internal. Previous studies obtained outgoing emails from domains by recruiting volunteers [19], which cannot support large-scale measurements.

In this paper, we cooperate with Coremail [9], a large email service provider in China, to obtain a realistic and unique view of email delivery paths. Coremail offers email services for more than 20K organizations. The original dataset is the email reception log of Coremail within nine months, from May 1, 2024 to November 30, 2024. We removed emails without *Received* headers, and the outgoing IP address belongs to a reserved or private address range (vendor's internal emails). For ethical reasons, we only extracted the minimum data required for our study and did not obtain the user's email address or email body. Specifically, the information in our email *Received* header dataset includes: the domain in the *Mail From* and the *Rcpt To* field, the IP address of the outgoing server, all *Received* headers, the time of receiving the email, the SPF verification results, and the email compliance check results by Coremail (e.g., clean, spam).

3.2 Construct Email Intermediate Paths

The format and content of the *Received* header are not strictly standardized and vary by software and provider. Therefore, we build a template library to parse *Received* headers and extract node information from them. Figure 3 illustrates the workflow for constructing the intermediate path dataset.

Parse *Received* header. To obtain the email path information as accurately as possible, we choose to use regular expressions to exactly match the *Received* header, instead of directly extracting key text. First, we select *Received* headers of emails from the top 100 sender domains (ranked by email number) in our email *Received* header dataset. Then, we generate regular expressions for these *Received* headers through manual analysis (❶). Through verification, we can match 93.2% of the *Received* headers in our dataset. For the remaining unmatched *Received* headers, we apply the Drain algorithm [18] to perform text clustering on them, and then construct regular expressions for the 100 clusters containing the largest number of *Received* headers (❷). At last, we built a *Received* header template library with 54 regular expressions, which can match 96.8% of *Received* headers in our dataset.

Next, we use the template library to parse *Received* headers to obtain email path nodes (❸). Specifically, the path nodes are the IP address and domain name of the *from* and *by* parts in each *Received* header. For *Received* headers that cannot be covered by our template library, we directly extract the domain name and IP address of the *from part* and the *by part*. We also use geographical databases [22] and domain suffix lists [2, 21] to obtain the AS and second-level domain (SLD) corresponding to the email path nodes.

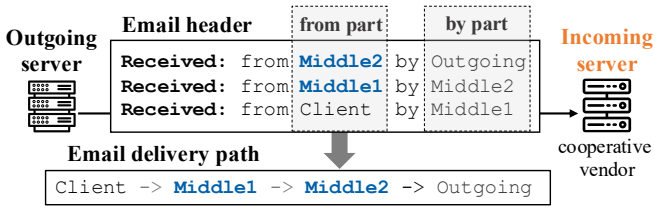

Figure 4: Process of building an email delivery path.

Build email intermediate path. Figure 4 shows the process of constructing an email delivery path. Considering that email servers may hide or falsify their identities, it is unreliable to use the information in the *by part* of the *Received* header to identify nodes [34]. Therefore, we use the *from part* of each *Received* header to indicate the information of the previous node (❹). For the outgoing node, we use the IP address and domain name of the outgoing server recorded by our cooperative vendor. Through the above process, we can construct the delivery path for the email.

To ensure the reasonableness of our results, we further filter the email delivery path (❺). First, this paper focuses on the intermediate paths associated with clean email, and the spam infrastructure is not within our consideration. As such, we removed the emails that were judged as spam by Coremail. Second, this paper uses the domain name in *Mail From* field as the sender of the intermediate path for analysis. Therefore, we removed the emails that did not pass SPF verification to ensure the authenticity of the sender's domain. Third, we removed email delivery paths without middle nodes. Fourth, we removed the incomplete email intermediate path, which includes middle nodes that cannot find valid identity information. Valid identity information defined in this paper includes the IP address and the domain name in legal format. However, we find that many middle nodes with invalid identity information are "local" nodes. We identify nodes as "local" if they contain only "local"/"localhost" or exclusively private/unallocated IP addresses [43]. Directly removing all intermediate paths containing "local" nodes would be inappropriate, since such nodes may represent internal relays or gateway devices within the email system. Therefore, we allow "local" nodes to appear in intermediate paths. Nevertheless, if an intermediate path consists entirely of "local" nodes, it provides no analytical value for our study on path dependency and centralization, and thus such paths are removed.

3.3 Overview of Intermediate Path Dataset

In our nine-month email *Received* header dataset, we observed a total of 2.4B emails. Among them, we successfully parsed the *Received* header in 98.1% of emails. Following this, we extracted emails that are marked as clean by our cooperative vendor and passed the SPF verification. After this step, the number of emails decreased significantly, leaving only 380M (15.6%). Specifically, 51.27% of emails are spam, and 42.87% of emails fail the SPF verification. Considering that this paper focuses on analyzing the email intermediate paths, we select emails with middle nodes and complete paths. In addition, we find that 15.34% of the email intermediate paths pass through the "local" node; only 0.9% of email intermediate paths consist entirely of "local" nodes, and these paths are removed from our dataset. At last, we obtain the intermediate path of 105M (4.3%) emails, which we refer to as the intermediate path dataset in this paper.

Among the intermediate path dataset, we observe 412,197 sender SLDs, 42,478 middle node SLDs, 679,507 outgoing node IP addresses, and 881,669 middle node IP addresses. Note that the sender SLD is obtained from the *Mail From* field, and the middle node SLD is obtained from the *Received* header. Among the middle nodes that can obtain valid IP addresses, 846.4K (96.0%) nodes are IPv4 and 35.3K (4.0%) are IPv6. Considering that our dataset originated from a Chinese provider, we analyzed IP addresses recorded in *Received* headers to determine the share of emails originating within China versus outside China. The results show that 32.8% of the emails were transmitted exclusively within China ("domestic email"), while the rest were from outside China ("international email").

Table 1: Statistics on the processing of the email Received header dataset.

Dataset	Number of emails
Email *Received* header dataset	2,446,933,441 (100%)
# Email *Received* header parsable	2,399,932,266 (98.1%)
# Clean and SPF pass	380,840,897 (15.6%)
# With middle node and complete intermediate path	**105,175,093 (4.3%) (Intermediate path dataset)**

Discussion of data filtering. To construct an accurate and clean dataset, this paper applies strict criteria that resulted in the removal of approximately 95% of emails. We acknowledge that this may cause some benign emails to be excluded, and we discuss the impact on our results. More than half of the emails were discarded because our cooperative provider classified them as spam. Through discussions with Coremail, we learned that Coremail's spam detection engine evaluates email compliance by comprehensively considering multiple factors such as sender reputation, delivery behavior, and message content, achieving an accuracy rate of 99.8% in detecting spam, including phishing, pornography, gambling, etc. However, the specific detection algorithms and processes were not explained. The use of commercial spam detection systems for dataset construction and filtering has also been adopted by other researchers [17, 34]. Furthermore, given Coremail's large-scale email service scope, we believe that its detection systems are reliable and unlikely to significantly affect the validity of our research results. Furthermore, we removed emails that failed SPF verification. We acknowledge that this may exclude some clean emails. For example, emails from domains with misconfigured/missing SPF records or emails that failed SPF verification due to forwarding. To ensure that intermediate paths can be reliably associated with the actual sender domains (*Mail From* field), we adopted SPF verification as a dataset filtering criterion.

4 What are the Identities and Distribution of Middle Nodes?

In this section, we first investigate the number of middle nodes contained in the email intermediate path. Then, we analyze the distribution of middle nodes in terms of AS and provider.

Intermediate path length. In our dataset, we find that most of the email intermediate paths contain only one middle node, accounting for 70.37%. Furthermore, 21.4M (20.39%) emails pass through two middle nodes, and only 746.3K (0.71%) emails pass through more than five middle nodes. We also analyze 481 emails with intermediate path lengths of more than 10. The results show that most of the

middle nodes in a path have the same SLD, so we infer that most of these emails are internal email relays for domains.

AS distribution. Table 2 shows the top 5 ASes in both middle and outgoing nodes, ranked by the number of sender SLDs that depend on them. We find that for both middle nodes and outgoing nodes, *8075 MICROSOFT-CORP-MSN-AS-BLOCK* has the highest domain share of 20.9%. For middle nodes, most of the mainstream ASes belong to email service providers (e.g., *MICROSOFT*) and local ISPs (e.g., *Chinanet*). In contrast, the mainstream ASes of outgoing nodes are more located in cloud service providers (e.g., *Alibaba*). One possible reason is that large email service providers offer (free) email forwarding services [14, 36], making many domains dependent on them for relaying email. For example, we observe that the email intermediate paths of 6,545 sender SLDs pass through *YANDEX LLC*, but only the outgoing nodes of 206 sender SLDs belong to *YANDEX LLC*.

Table 2: Statistics on Top 5 ASes of middle and outgoing nodes with high sender SLD dependencies.

Top 5 ASes	# SLD	# Email
Middle node		
8075 MICROSOFT-CORP-MSN-AS-BLOCK	20.9%	36.8%
15169 GOOGLE	3.7%	1.7%
13238 YANDEX LLC	2.7%	1.4%
16509 AMAZON-02	2.1%	1.5%
4134 Chinanet	2.1%	1.3%
Outgoing node		
8075 MICROSOFT-CORP-MSN-AS-BLOCK	23.4%	14.8%
37963 Hangzhou Alibaba Advertising ...	6.3%	9.2%
15169 GOOGLE	5.9%	7.5%
45090 Shenzhen Tencent Computer ...	5.5%	3.4%
16509 AMAZON-02	3.9%	6.3%

Provider distribution. We identify the provider behind the middle node based on its SLD. Table 3 presents the top 10 middle node providers ranked by the number of sender domains that depend on them. In this paper, ESP refers to the email service provider that offers integrated email-related services, including user mailboxes, domain hosting, email forwarding, etc. We find that most middle nodes belong to ESPs, with outlook.com accounting for more than half of the emails. Among the top 10 providers, we also identify vendors offering email signature (exclaimer.net and codetwo.com) and security filtering services (secureserver.net).

When a provider offers email services through multiple SLDs, our method of identifying providers by SLD leads to misclassification. To assess the impact of this issue, we manually examined the top 200 SLDs ranked by email dependency, which together account for 94.2% of the total email volume. We find that only 1.2% of emails would be misclassified, with the most common cases involving large providers offering mail services under different top-level domains (TLDs), such as outlook.cn, outlook.fr, and yandex.ru. Therefore, identifying providers based on SLD does not significantly affect our overall conclusions.

Table 3: Statistics on Top 10 providers of middle nodes with high sender SLD dependencies.

Top 10 providers	Type	# SLD	# Email
outlook.com	ESP	295.9K (51.5%)	84.6M (66.4%)
exchangelabs.com	ESP	25.0K (4.4%)	5.9M (4.6%)
icoremail.net	ESP	13.0K (2.3%)	0.5M (0.4%)
yandex.net	ESP	9.8K (1.7%)	0.6M (0.5%)
exclaimer.net	Signature	9.0K (1.6%)	1.7M (1.3%)
google.com	ESP	8.1K (1.4%)	0.7M (0.6%)
codetwo.com	Signature	7.1K (1.2%)	1.0M (0.8%)
qq.com	ESP	2.9K (0.5%)	0.2M (0.2%)
aliyun.com	ESP	2.4K (0.4%)	0.2M (0.2%)
secureserver.net	Security	2.3K (0.4%)	0.1M (0.1%)

5 What is the Dependency Structure and Regionality of Email Intermediate Paths?

This section explores the dependency structure and regionality of email intermediate paths. We conduct the analysis from three aspects: path dependency patterns, dependency passing relationships, regional dependency characteristics.

5.1 Dependency Pattern of Intermediate Paths

To better characterize the email intermediate path structure, we define the dependency patterns of email intermediate paths from the following two perspectives.

• **Hosting pattern** describes the relationship between middle nodes and the sender domain, reflecting the extent to which a domain relies on third-party providers in the email intermediate path. We categorize hosting patterns into three types. 1) *Self-hosting*: The domain uses its own infrastructure to handle the email intermediate path, meaning all middle node SLDs are the same as the sender SLD. 2) *Third-party hosting*: The email intermediate path of the domain is completely dependent on third-party providers, meaning the middle node SLDs all differ from the sender SLD. 3) *Hybrid hosting*: The email intermediate path involves both self-hosted and third-party infrastructure, meaning the middle nodes include both the sender SLD and other SLDs.

• **Reliance pattern** refers to the number of distinct providers involved in an email intermediate path, reflecting the complexity of the intermediate path. We define two types of reliance patterns. 1) *Single reliance*: The intermediate path involves only one provider, meaning all middle nodes share the same SLD. 2) *Multiple reliance*: The intermediate path involves multiple providers, meaning the SLDs of middle nodes include more than one unique value.

Overview. Table 4 summarizes the hosting and reliance patterns observed for emails and sender domains in our intermediate path dataset. Note that a single sender domain may correspond to multiple email intermediate paths with different patterns. Regarding the hosting pattern, we find that 82.7% of email intermediate paths fall into the *Third-party hosting* category, indicating a complete reliance on external providers for email relaying. Only 4.3% of sender domains rely solely on their own infrastructure for email intermediate transmission, reflecting the dominance of third-party hosting

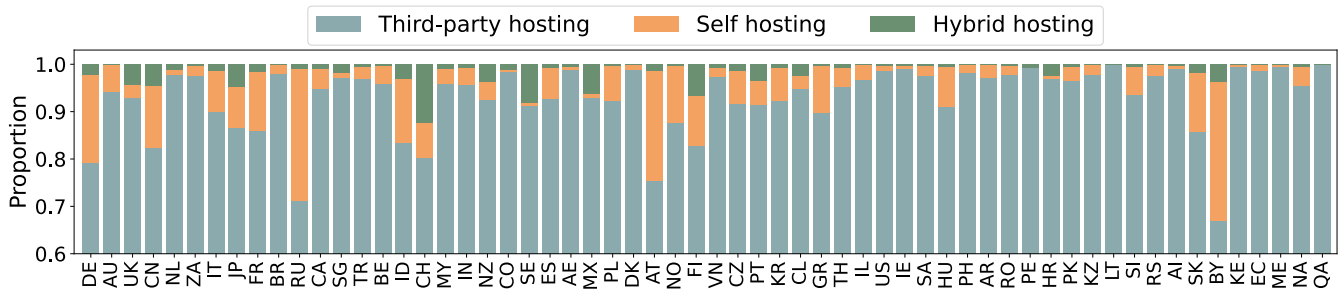

Figure 5: The hosting patterns of email intermediate paths for domains of different countries.

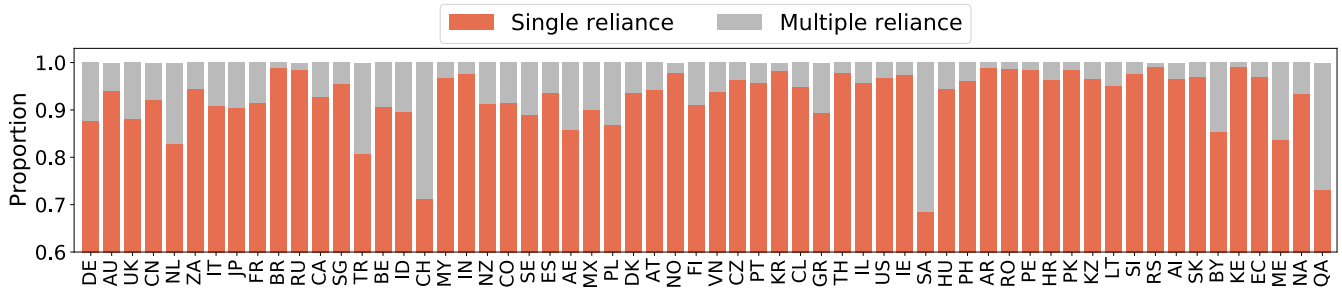

Figure 6: The reliance patterns of email intermediate paths for domains of different countries.

Table 4: Statistics on dependency patterns of email intermediate paths.

	# SLD	# Email
Hosting pattern		
Self hosting	17.7K (4.3%)	15.1M (14.3%)
Third-party hosting	399.1K (96.8%)	86.9M (82.7%)
Hybrid hosting	7.5K (1.8%)	3.2M (3.0%)
Reliance pattern		
Single reliance	384.5K (93.3%)	96.0M (91.3%)
Multiple reliance	52.8K (12.8%)	9.1M (8.7%)

services in email delivery. Additionally, 3.2M emails (3.0%) exhibit *Hybrid hosting*, where both self-hosted and third-party nodes appear in the same path. This is expected given that many different roles of participants in the email ecosystem are capable of relaying emails. We provide a more detailed analysis in Section 5.2.

Regarding the reliance pattern, we observe that 91.3% of intermediate paths involve only a single provider. There are two potential reasons for this phenomenon. On the one hand, large hosting providers typically integrate various email functionalities, such as email delivery, management, detection, and forwarding services, allowing domains to rely on a provider for complete email service. On the other hand, additional email functions, such as email signature and security filtering, are not necessary for most domains.

Country domain. Next, we analyze the dependency patterns of email intermediate paths of country-specific domains. Using the country code top-level domain (ccTLD) list [2, 21], we select emails from different countries' sender SLDs within our email intermediate path dataset. Figure 5 and Figure 6 display the hosting and reliance patterns of email intermediate paths for domains from the top 60 countries, ranked by the number of SLDs we observed. The top five countries each have more than 12K SLDs, and all the top 60 countries have more than 300 SLDs.

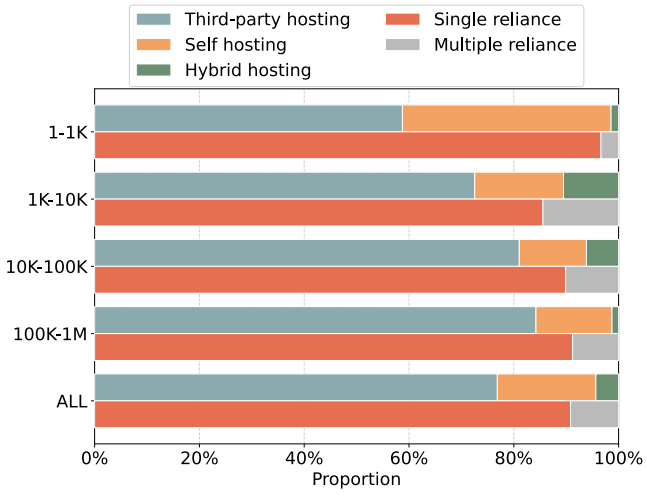

Figure 7: The dependency patterns of email intermediate paths for domains of different popularity ranges.

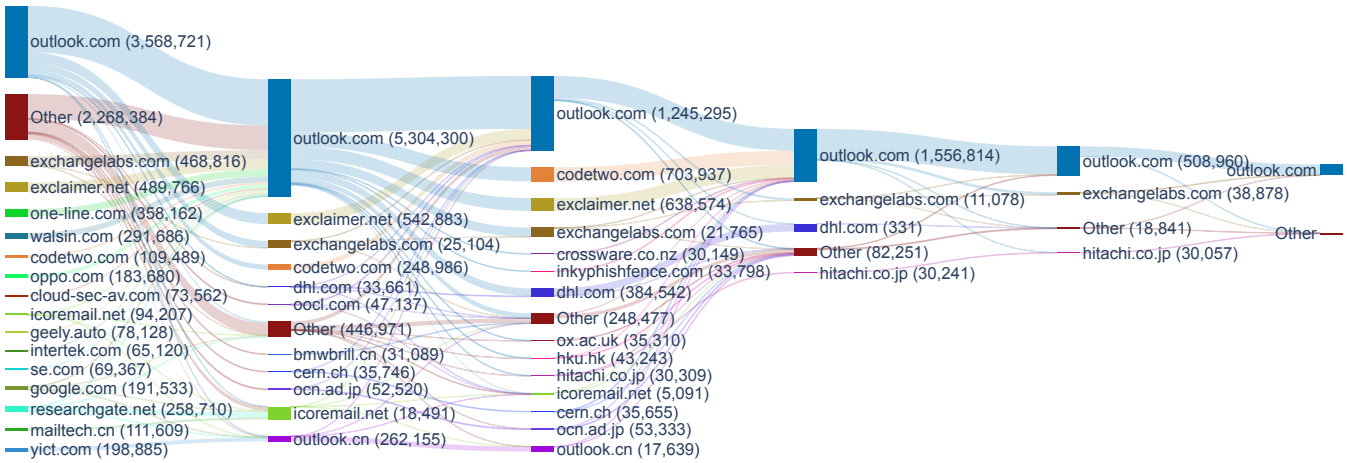

Figure 8: Dependency passing in *Multiple reliance* email intermediate paths (show up to six hops).

Our results show that the proportion of *Third-party hosting* in email intermediate paths for various countries exceeds 60%, highlighting the email dependency on hosting providers. Notably, email intermediate paths from Russia and Belarus exhibit the *Self hosting* proportion of about 30%, which is significantly higher than most other countries. Using a URL-type classifier [39], we observe that 42.9% of the *Self hosting* domains in Russia belong to commercial companies, and 18.2% are educational institutions. Measurements by Jonker et al. [23] in 2022 also reported that, following the Russia-Ukraine conflict, Russia indeed reduced its Internet dependency (e.g., DNS and PKI) on foreign hosting services.

Additionally, we observe that the majority of countries' email intermediate paths rely on a single provider, with the proportion of *Single reliance* typically exceeding 80%. However, in countries such as Switzerland, Saudi Arabia, and Qatar, the proportion of *Multiple reliance* exceeds 30%. Upon further analysis, we find that this is primarily due to the inclusion of email signatures and security filtering providers in intermediate paths. For example, in Switzerland, 21.6% of intermediate paths with *Multiple reliance* include email signature providers, and 22.9% include email filtering providers.

Popular domain. We also analyze the dependency patterns of email intermediate paths of popular domains. The domain rankings are based on the Tranco Top 1M list [46] obtained on December 1, 2024. Figure 7 illustrates the reliance patterns of domains with different levels of popularity. We observe that more popular domains tend to rely less on third-party hosting providers, suggesting that large companies have the capacity to deploy their own email services. For example, approximately 60% of the email intermediate paths for domains ranked between 1-1K are categorized as *Third-party hosting*, while this percentage increases to over 80% for domains ranked between 100K-1M. In addition, regardless of popularity tier, over 80% of email paths for domains include only a single provider (*Single reliance*).

5.2 Dependency Passing in Intermediate Paths

The email intermediate path involves different SLDs, meaning that dependencies are passed between various suppliers. However, the interactions between email middle nodes may harbor potential security risks. For example, the EchoSpoofing vulnerability abuses the fragile dependency between the victim domain and the middle node (Proofpoint) to distribute spoofed emails [16]. In this section, we conduct an in-depth analysis of the dependency passing in 9.1M *Multiple reliance* intermediate paths, revealing the interactive relationships present during the email delivery process.

If two email intermediate paths contain the same set of middle node SLDs (regardless of order), we consider them to belong to the same dependency passing relationship. In total, we identify 28,359 distinct dependency passing relationships, among which 15,386 (55.8%) involve two SLDs, 7,304 (25.8%) involve three SLDs, and 5,219 (18.4%) involve more than three SLDs. Figure 8 illustrates the dependency passing flows for each hop in *Multiple reliance* intermediate paths and annotates each node with its corresponding email out-degree. Note that providers with fewer than 50K email out-degrees in each hop are merged into the "Other" node, and the width of the flow lines is scaled using the log_2 value of the email out-degree. We can see that in intermediate paths of each hop, a significant proportion of the emails rely on outlook.com for transmission. Excluding internal relays within the same provider (e.g., outlook.com to outlook.com), the three most common dependency passing relationships are: "outlook.com to exclaimer.net" (1.5M emails, 17.3%), "outlook.com to codetwo.com" (939.9K emails, 10.9%), and "outlook.com to exchangelabs.com" (732.4K emails, 8.5%). The first two relationships stem from outbound email signature attachment services, while the last one represents internal relaying between two Microsoft-operated email services. In addition, we observe that third-party email additional services, such as email signature and filtering, frequently appear in the first three hops of outbound email intermediate paths.

By manually analyzing the top 50 dependency passing paths with the highest email volumes, we identify six common types of dependency passing relationships, as summarized in Table 5. We find that the most prevalent dependency passing occurs between ESPs and email signature providers ("ESP-Signature"), accounting for 29.7% of the emails. In addition, 13.3% of emails fall into the "ESP-ESP"

Table 5: Statistics of the main types of email dependency passing relationships.

Type	# SLD	# Email
ESP-Signature	16,468 (31.2%)	2,716,390 (29.7%)
ESP-ESP	8,339 (15.8%)	1,216,430 (13.3%)
ESP-Security	2,850 (5.4%)	237,798 (2.6%)
ESP-Signature-ESP	1,530 (2.9%)	192,068 (2.1%)
ESP-Security-ESP	950 (1.8%)	146,337 (1.6%)
ESP-Signature-Security	580 (1.1%)	82,314 (0.9%)

type, typically arising from email forwarding, replies, or bounce between ESPs. Dependency transfers between ESPs and security filtering providers ("ESP-Security") are relatively rare, constituting less than 5% of the emails.

5.3 Regional Dependency of Intermediate Paths

Prior studies [3, 27] have shown that domains from many countries exhibit distinct regional dependency patterns in terms of website, DNS, and certificate authority (CA). In the following, we analyze the regional dependencies in email intermediate paths, focusing on the reliance of emails from different countries or continents on external regions.

Cross-regional path volume. We first analyze how many email intermediate paths traverse different regions. Our results show that over 95% of email intermediate paths involve only a single region, whether in terms of country, AS, or continent. This indicates that email transmission in the intermediate path stage rarely crosses multiple regions.

Regional dependence across countries. Next, we analyze the dependence on external countries in the email intermediate paths of different countries. Referring to the ccTLD list [2, 21], we extract the intermediate paths of sender domains from various countries in our dataset. To ensure the representativeness of the results, we exclude countries with fewer than 10K emails and 300 SLDs. Figure 9 shows the regional dependency of email intermediate paths in 60 countries. Specifically, if email middle nodes belong to the same country as the sender domain, it is marked as "Same". In the Figure 9, we only display countries with a proportion of emails exceeding 15%, and the remaining countries are grouped under "Other". We rank countries in descending order of their dependence on external countries.

The result shows that regional dependency patterns vary significantly across countries. In some countries, over 90% of email intermediate paths rely solely on domestic infrastructure, such as Russia and Malaysia. However, in some countries, email intermediate paths almost entirely depend on foreign infrastructure, such as Montenegro and Morocco. Additionally, 23 countries show no single external country accounting for more than 15% of their intermediate paths, meaning only "Same" and "Other" categories are shown in Figure 9. This indicates that these countries do not excessively rely on any specific foreign nation for email transmission, and thus have a relatively high degree of independence. In contrast, the remaining 37 countries show a significant proportion

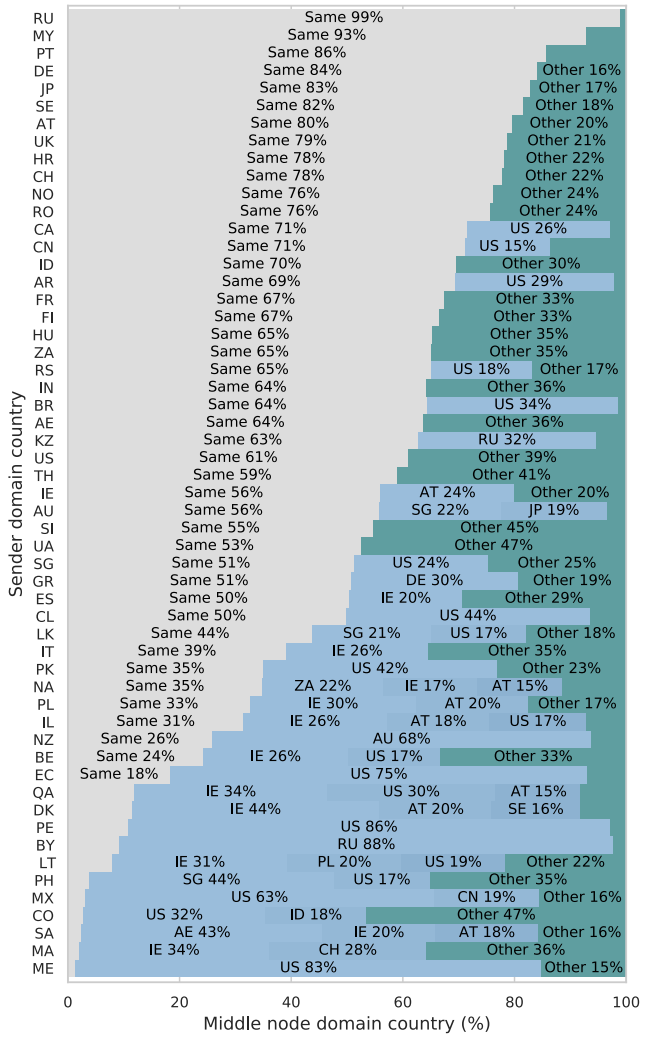

Figure 9: Regional dependence of email intermediate path in different countries. Only countries with a proportion of more than 15% are displayed.

of dependency on external countries. For example, 88% of intermediate paths from Belarus include nodes located in Russia, and 83% of paths from Montenegro include nodes in the United States. This implies that email relays in these countries are susceptible to geopolitical influences.

Furthermore, we analyze the reasons behind the differences in regional dependency patterns across countries. We observe that countries belonging to the Commonwealth of Independent States (CIS), formed after the collapse of the Soviet Union, significantly rely on Russia's email infrastructure. For example, we find that a substantial proportion of email intermediate paths from Kazakhstan (32%) and Belarus (88%) included nodes located in Russia. In contrast, no other countries show a similarly significant dependency on Russia's email infrastructure. Moreover, we observe that email

intermediate paths often reflect dependencies between geographically proximate or linguistically similar countries. For example, 68% of email paths from New Zealand include middle nodes located in Australia, and 43% of email paths from Saudi Arabia include middle nodes located in the United Arab Emirates. Surprisingly, we find that the email intermediate paths of several European countries heavily rely on Ireland (IE). For example, 26% of paths from Italy, 30% from Poland, 26% from Belgium, and 44% from Denmark include middle nodes located in Ireland. Upon further investigation, we discovered that the ASes and SLDs of these Ireland middle nodes are almost all associated with Microsoft and outlook.com. We speculate that Microsoft's hosting service may use data centers located in Ireland to relay emails from these national domains to our cooperative provider in China.

Regional dependence across continents. Below, we analyze the regional dependence of email intermediate paths at the continental level. We only consider sender domains under ccTLDs within each continent. Figure 10 illustrates the intercontinental dependencies of email intermediate paths. We find that the majority of emails originating from Asia, Europe, and North America have middle nodes located within the same continent, with Europe accounting for as much as 93.1%. In contrast, email intermediate paths from Africa heavily depend on Europe and North America, while those from South America are highly dependent on North America. These observations align with dependency patterns in web hosting. For instance, Akiwate et al. [3] reported that approximately 70% of websites from African countries are hosted in North America. In addition, middle nodes in Africa, South America, and Oceania are almost only responsible for the emails of their continents, and emails from other continents do not pass through them.

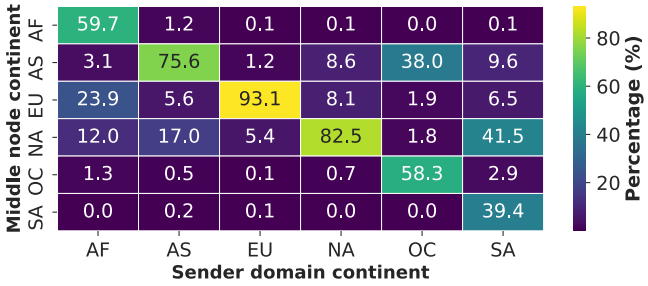

Figure 10: Regional dependence of email intermediate path in different continents.

6 What are Centralization and Cross-country Differences of Email Intermediate Paths?

In this section, we first evaluate the degree of centralization of email intermediate paths and examine the market share of large email providers. Then, we investigate the differences in the centralization of email intermediate paths across countries. Finally, we compare the centralization among middle, outgoing, and incoming nodes.

6.1 Market Share of Large Providers

We use the Herfindahl-Hirschman Index (HHI) [48] to evaluate the market concentration of email middle nodes. HHI is calculated by summing the squares of the market shares of all entities, and it is widely used for assessing centralization in internet infrastructure [5, 20]. A higher HHI indicates a more monopolistic market structure: an HHI of 10% indicates moderate concentration, while a value above 25% indicates high concentration. Considering all email intermediate paths, we obtain an HHI of 40% for the middle node market, which indicates a highly concentrated market. As shown in Table 3, Microsoft dominates the overall middle node market, participating in about 70% of the intermediate paths.

In addition, we investigate the popularity of domains served by large middle node providers. Specifically, we select the domains relying on the Top 10 middle node providers that appeared in the Tranco Top 1M list. The violin plot of Figure 11 depicts the popularity distribution of domains relying on five large providers. The remaining providers are not shown due to insufficient sample sizes. The area of each violin is proportional to the number of dependent domains; for visual clarity, we apply linear scaling to the areas. The shape of each violin reflects the distribution of popularity rankings, with the median ranking indicated by a white circle. We observe that outlook.com is relied upon by the largest number of popular domains, with 25,844 in total and a median popularity ranking of 278K. Furthermore, the dependent domain popularity distributions for outlook.com, exchangelabs.com, and exclaimer.net are relatively broad, while those for icoremail.net and google.com are more concentrated. Overall, the above findings further underscore Microsoft's dominant position in the email middle node market.

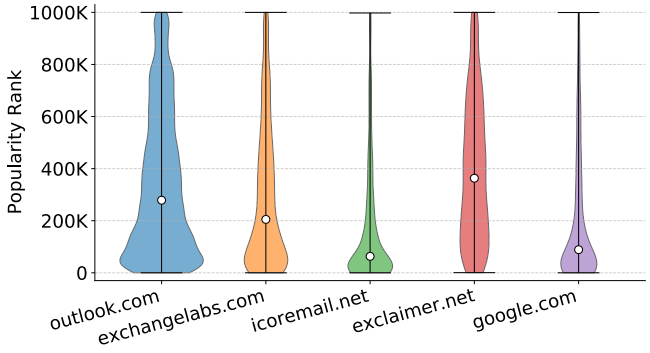

Figure 11: The popularity distribution of domains that rely on 5 large middle node providers.

6.2 Centralization of Email Intermediate Paths Across Countries

In the following, we analyze the centralization of email intermediate paths across countries. In our data analysis, we excluded countries with fewer than 10K intermediate paths and fewer than 300 SLDs. Figure 12 presents the HHI of the middle node providers for each country, with the provider holding the largest market share in each country marked by a circle. In addition, the country codes on the x-axis are color-coded according to their respective continents.

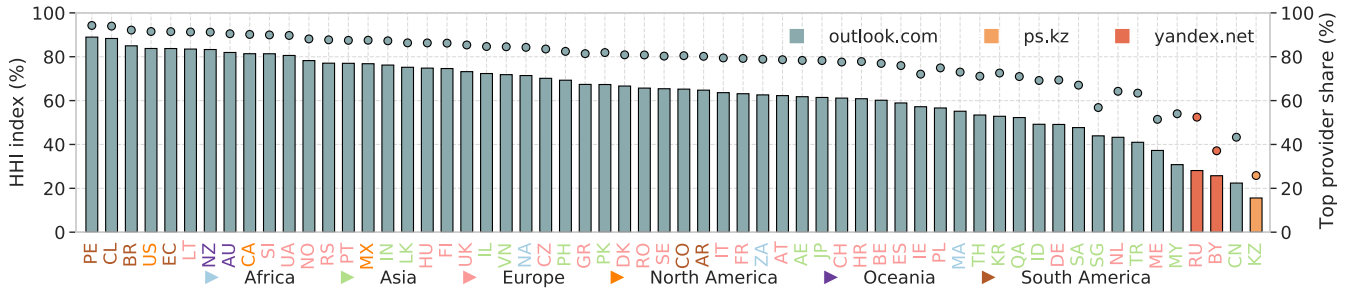

Figure 12: The HHI of middle node providers by country (bars, left Y-axis), and the most popular provider in each country by market share (circles, right Y-axis).

We observe considerable variation in the HHI across countries. Peru exhibits the highest HHI at 88%, while Kazakhstan shows the lowest at 16%. Moreover, outlook.com dominates the email intermediate path market share in most countries, typically exceeding 60%. Notable exceptions include Russia and Belarus, where yandex.net is the primary provider; yandex.net also accounts for 21% of the market in Kazakhstan. Given that Yandex is a Russian email provider, the dependence of CIS countries on it is understandable. In the case of Kazakhstan, ps.kz, a local cloud service provider, holds 26% of the intermediate email path market. We also note that countries in South America and Oceania generally have high HHI, with all countries exceeding 60%.

6.3 Compare the Centralization of Incoming and Outgoing Nodes

Previous studies on email system centralization have primarily used MX and SPF records [32, 47, 51], representing the perspectives of incoming and outgoing nodes, respectively. To better understand the differences in centralization across various segments of the email delivery paths, we conducted a comparative analysis of middle, incoming, and outgoing nodes. On May 1, 2025, we actively scan the MX and SPF records of 412,197 sender SLDs in our email intermediate path dataset. Following prior work [47, 51], we identified the incoming providers by extracting the SLDs of the MX records, and the outgoing providers by extracting the SLDs from the include fields in the SPF records.

In total, we identify 42,478 middle node providers, 48,597 incoming node providers, and 23,853 outgoing node providers. By calculating the number of dependent domains for each provider, we find that the HHI for the middle node provider market is 29%, for the incoming node is 37%, and for the outgoing node is 18%. Therefore, we conclude that the markets for incoming node and middle node providers are highly concentrated, and the incoming node market is more concentrated. In contrast, the outgoing node provider market is moderately concentrated.

Furthermore, we investigate the popularity and market share of large providers across middle, incoming, and outgoing nodes. Our results reveal that there are differences in the market concentration of different types of node providers. As shown in Figure 13, we mark the top 10 middle node providers with colored dots, and then analyze their rankings and shares among the top 100 incoming and

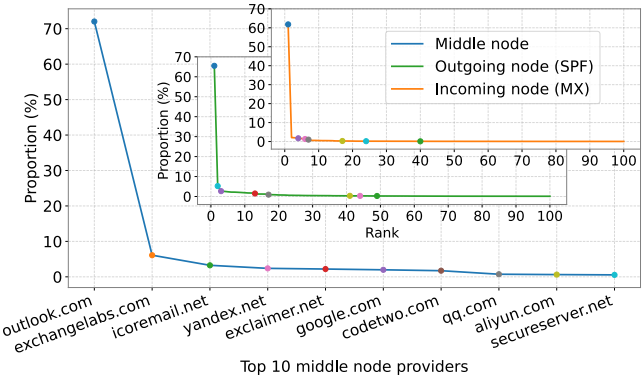

Figure 13: Rank and market share of the Top 10 middle node providers in incoming and outgoing nodes.

outgoing node providers. We find that outlook.com is the dominant provider across all node types, accounting for more than 60% of the market share in each category. In contrast, the distribution of other large providers is highly fragmented. Specifically, email service providers (e.g., google.com) and security filtering providers (e.g., secureserver.net) typically appear among the top 100 providers across all three node types. However, email signature providers tend to serve only as middle and outgoing nodes. For example, no domain in our dataset sets its MX record to codetwo.com or exclaimer.net. Moreover, exchangelabs.com, a domain belonging to Microsoft, only appears in the middle node providers, indicating its primary role in email relaying. Among the top 100 middle node providers, we find that 41 do not appear in the incoming and outgoing node provider lists.

7 Discussion

In this section, we first present the insights of our work for the email community and highlight several potential directions for future research. We then report the ethical considerations addressed during our study.

7.1 Hidden Intermediate Dependencies

Our measurement study uncovers several notable findings regarding the dependency structure and centralization of email intermediate paths in the real world. These findings not only expand current understanding of the email ecosystem, but also highlight critical areas requiring further exploration and community attention.

First, our results reveal the presence of previously underexplored middle nodes in email delivery paths, such as email signature suppliers and security filtering providers. These intermediaries, often acting as relays or email processors, appear in a significant portion of email traffic. We suggest that future work conduct in-depth analyses of these nodes, focusing on their operational roles and potential implications for security and resilience in global email infrastructure.

Second, we find that the transmission of many emails involves multiple intermediate entities. Prior studies have demonstrated that such interactive relationships between nodes along the email path expose unexpected risks [42]. In light of this, we recommend that the community further develop systematic methods for measuring the structural risk of email transmission interactions. Moreover, domain operators should rigorously review their email middle node configurations, especially when relying on third-party intermediaries, to guard against vulnerabilities like EchoSpoofing [16].

Third, we observe significant differences in the dependency pattern and centralization of email intermediate paths across different countries. While some regions exhibit highly concentrated reliance on a small number of foreign providers, others demonstrate more diversified structures. This geographic disparity calls for deeper investigation to uncover the underlying economic, technical, or policy-driven factors. In parallel, we suggest that stakeholders pay closer attention to critical points of dependency along intermediate paths, as they may pose significant risks of service disruption under geopolitical tensions or cross-border regulatory shifts.

Finally, our findings point to a neglected aspect of email delivery security: the end-to-end security consistency across all segments of the email transmission path. For example, in our intermediate path dataset, 27K emails include segments that use both outdated (1.0 and 1.1) and secure (1.2 and 1.3) versions of TLS [37], as indicated in the *Received* headers. This protection inconsistency may undermine the security of the entire email transmission process. Overall, we advocate that future security research should consider fine-grained segment-level email transport security, and promote the uniform adoption of secure standards throughout the entire email transmission path. To help other scholars reprise and expand our research, we publish our email path extractor and intermediate path dataset at https://github.com/RUI-XUAN-LI/Email_Path.

7.2 Ethical Considerations

Our study relies on real-world email reception logs from a large email service provider, and we have carefully considered the associated ethical risks. Although our institution does not have an Institutional Review Board (IRB), our research was authorized and overseen by the administrative department of our cooperative vendor. Our ethical considerations were guided by prior research using similar datasets [29, 34] and by established principles of research ethics [1, 25].

Construction and processing of datasets. In this paper, we collect and analyze the minimal data necessary to study email intermediate paths, including domains in the *Mail From* and *Rcpt To* fields, IP addresses of outgoing servers, *Received* headers, the time of email reception, SPF verification results, and the email compliance label. Coremail stores *Received* headers for the purpose of analyzing transmission delays and diagnosing network issues. For this study, Coremail transferred the minimal set of required data fields to a separate secure server within its organization and provided us with controlled access. All data processing and storage during our research were carried out on secure servers to prevent data leakage. Consequently, none of the researchers involved in this paper had access to the raw email logs containing, for example, email content or complete email headers. We strictly adhered to the data protection and server usage guidelines agreed upon with Coremail and did not share any data with individuals outside our research team. Upon completion of the study, we deleted all data and intermediate results related to this paper from the secure server.

PII involved in the study. In this paper, we did not collect the email subject and body, as well as the usernames in the *Mail From* and *Rcpt To* fields. The PII that may be involved in our research is mainly related to the email address in the *Received* headers. In the process of parsing *Received* headers through regular expressions, we find that 28.35% of *Received* headers contain email addresses. Therefore, during data extraction, we only used regular expressions to match the email address format, without retrieving the corresponding values. Regarding the share of research artifacts, the released dataset includes only the domains and IP addresses of middle nodes, and does not contain email addresses.

Risk-benefit analysis. This paper relies on large-scale real-world datasets and therefore requires careful consideration of the potential risks and benefits of the study. The primary ethical risks stem from dataset construction and the presence of PII (e.g., email addresses). As discussed earlier, we have taken measures to minimize the involvement of sensitive user information and to prevent data leakage. Overall, the ethical risks of our study are limited and manageable. Compared to the potential risks, we believe that our work offers greater benefits by advancing the community's understanding of the email intermediate ecosystem and by facilitating future security research on email transmission paths.

8 Limitation

Although our study offers a large-scale measurement of email intermediate paths, several limitations should be acknowledged. We first discuss the limitations related to the data perspective. Our dataset is sourced from a single email service provider and does not comprehensively capture global email delivery paths. It is extracted from inbound email logs, primarily reflecting the sender-side perspective, while offering limited visibility into recipient-side middle nodes. Moreover, our cooperative provider is a Chinese enterprise, our dataset cannot capture intra-country delivery paths of various countries, and email delivery paths may vary depending on the geographic location of recipient servers.

This study relies on *Received* headers to reconstruct email intermediate paths. We recognize that our criteria for filtering *Received* headers can result in the removal of some benign emails,

and discuss the impact of data filtering on our study in Section 3.3. Moreover, we acknowledge that forged *Received* headers would affect our results. However, according to the research of Luo et al. [34], the forged *Received* header in the email content is almost non-existent in the wild, even for phishing emails. In particular, the target of our study is clean emails, and they do not have enough motivation to forge email headers. Moreover, our work aims to reveal the dominant dependency patterns and centralization in real-world intermediate paths, which are not significantly influenced by a few forged paths. Overall, the impact of the forged *Received* header for our study is minimal.

Furthermore, if providers offer relay services via multiple SLDs, relying on SLDs to identify them would result in misclassification. We also recognize that the underlying causes of email dependency between countries and continents are complex and challenging to accurately identify. This paper provides only reasonable hypotheses regarding possible causes, without conducting a systematic causal analysis. We hope that our work inspires the community to further discuss the underlying reasons behind regional dependencies.

9 Related Work

9.1 Internet Centralization

The Internet centralization has become a widely discussed topic in recent years, and researchers have focused on critical components such as DNS, CAs, and cloud services [3, 27, 38]. Their results all indicate the concentration of many core Internet functions in the hands of a few dominant providers.

Research on DNS centralization is mainly divided into two aspects: recursive resolution and authoritative service. On the DNS recursive resolution side, Doan et al. [11] analyzed the recursive resolvers used by RIPE Atlas probes and found that 78.4% of the probes rely on Google's public DNS service. Xu et al. [49] observed that 90% of DNS forwarders depend on only 5% of public resolvers. On the authoritative service side, many studies have examined the centralization of domain hosting services by actively collecting authoritative name servers. For example, Kumar et al. [27] discovered significant variation in countries' reliance on third-party DNS providers, ranging from 36% in the Czech Republic to 72% in Singapore. In addition, some studies analyzed real-world DNS resolution traffic to assess centralization. Moura et al. [38] examined DNS traffic from two TLDs and the B root server and reported that 30% of DNS requests originated from five cloud providers.

In terms of domain dependency on CAs, Kumar et al. [27] analyzed the CAs of popular websites from 50 countries. They showed that DigiCert is the most popular CA among these websites, accounting for 36% of all websites using third-party CAs. Akiwate et al. [3] examined the level of CA centralization across 150 countries. Their findings indicate that a small number of CAs, such as Let's Encrypt, dominate the CA ecosystem, and the leading CA providers vary across countries. Researchers have also measured domain reliance on third-party CDN providers and web hosting services [3, 27, 51]. These studies highlight the trend of Internet centralization and show that historical, political, and linguistic factors shape the centralization patterns in different countries.

9.2 Email Dependence and Centralization

Only a few studies have analyzed the centralization of the email ecosystem. Scholars typically use MX and SPF records to evaluate the dependence of domains on third-party email providers. Specifically, Liu et al. [32] analyzed the centralization of incoming nodes by examining the MX records of Alexa Top 1M domains. Their results showed that among the Alexa Top 1K domains, 37% relied on Google, 10% on Microsoft, and 10% on Proofpoint. They also noted that from 2017 to 2021, the market shares of Google and Microsoft steadily increased. In another study, Zembruzki et al. [51] actively scanned the MX records of the Tranco Top 1M domains and found that 10 ASes provided email services for 50% of the domains. The market was dominated by Google, Microsoft, and Amazon, with Google accounting for the highest share at 21%. Regarding outgoing nodes, Wang et al. [47] identified outgoing providers by analyzing the *include* field in SPF records. Among the Tranco Top 1M domains, they found that outlook.com provided email delivery services for 20% of the domains, and google.com served 16%.

Only a few studies analyzed the transmission paths of emails, usually using the *Received* headers in the email content. For instance, Luo et al. [34] examined *Received* headers in phishing emails to characterize the infrastructure used by spammers, and Sanchez et al. [45] investigated the length and the number of IP addresses of spam delivery paths. However, to the best of our knowledge, the dependency structure and centralization of email middle paths have never been systematically explored. Moreover, vulnerabilities in the email path have been shown to cause real-world security risks, such as sender spoofing [16] and the circumvention of security protections [42].

10 Conclusion

This paper systematically analyzes the dependency and centralization of email intermediate paths. We find that the ecosystem of email middle nodes is highly centralized, with Microsoft occupying a dominant market share. In total, 86.9M (82.7%) emails rely on third-party middle node providers, and 9.1M (8.7%) involve multiple providers. Email signature and security filtering vendors are involved in many cross-vendor interactions in email intermediate paths. The dependency patterns in email intermediate paths vary across regions. Overall, our study can offer new insights into the email transmission process and contribute to improving the overall security of the email ecosystem.

Acknowledgments

We thank all anonymous reviewers for their valuable and constructive feedback. This work is supported by the National Key Research and Development Program of China (No. 2023YFB3105600), the National Natural Science Foundation of China (Grant No. 62102218, 62272413), and the "Pioneer" and "Leading Goose" R&D Program of Zhejiang, China (Grant No. 2024C03288). Baojun Liu and Jun Shao are both corresponding authors.

References

[1] 1979. The Belmont report: ethical principles and guidelines for the protection of human subjects of research. United States. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. Department of Health, Education and Welfare.

[2] 2025. List of Country Code TLDs. https://gist.github.com/derlin/421d2bb55018a1538271227ff6b1299d.

[3] G. Akiwate, K. Ruth, R. Habib, and Z. Durumeric. 2024. On the Centralization and Regionalization of the Web. *CoRR* abs/2406.19569 (2024).

[4] Barracuda. 2025. https://www.barracuda.com/.

[5] S. Bates, J. Bowers, S. Greenstein, J. Weinstock, Y. Xu, and J. Zittrain. 2018. Evidence of decreasing internet entropy: the lack of redundancy in dns resolution by major websites and services. In *National Bureau of Economic Research*.

[6] CBS. 2020. Google services including Gmail and YouTube suffer outage. https://www.cbsnews.com/news/google-services-restored-global-outage-today-2020-12-14/.

[7] J. Chen, V. Paxson, and J. Jiang. 2020. Composition Kills: A Case Study of Email Sender Authentication. In *USENIX Security Symposium*. 2183–2199.

[8] CodeTwo. 2025. https://www.codetwo.com/.

[9] Coremail. 2025. https://www.coremail.cn/.

[10] S. Czybik, M. Horlboge, and K. Rieck. 2023. Lazy Gatekeepers: A Large-Scale Study on SPF Configuration in the Wild. In *IMC*. ACM, 344–355.

[11] T. Doan, J. Fries, and V. Bajpai. 2021. Evaluating Public DNS Services in the Wake of Increasing Centralization of DNS. In *IFIP*. IEEE, 1–9.

[12] Exclaimer. 2025. https://exclaimer.com/.

[13] GoDaddy. 2025. https://www.godaddy.com/en/offers/email/sem-email-domain-bundle.

[14] Google. 2025. Automatically forward Gmail messages to another account. https://support.google.com/mail/answer/10957?hl=en.

[15] Google. 2025. Google Workspace. https://workspace.google.com/.

[16] Guardio. 2024. EchoSpoofing — A Massive Phishing Campaign Exploiting Proofpoint's Email Protection to Dispatch Millions of Perfectly Spoofed Emails. https://labs.guard.io/echospoofing-a-massive-phishing-campaign-exploiting-proofpoints-email-protection-to-dispatch-3dd6b5417db6.

[17] Wei H., Van T., Vincent R., Z. Wang, A. Dasbach-Prisk, M. H. Afifi, J. Yang, E. Katz-Bassett, and A. Cidon G. Ho. 2025. Do Spammers Dream of Electric Sheep? Characterizing the Prevalence of LLM-Generated Malicious Emails. In *IMC*. ACM.

[18] P. He, J. Zhu, Z. Zheng, and M. Lyu. 2017. Drain: An Online Log Parsing Approach with Fixed Depth Tree. In *ICWS*. IEEE, 33–40.

[19] F. Holzbauer, J. Ullrich, M. Lindorfer, and T. Fiebig. 2022. Not that Simple: Email Delivery in the 21st Century. In *USENIX ATC*. USENIX Association, 295–308.

[20] G. Huston. 2022. Looking at centrality in the DNS. https://blog.apnic.net/2022/11/22/looking-at-centrality-in-the-dns/.

[21] IANA. 2025. Root Zone Database. https://www.iana.org/domains/root/db.

[22] ip api. 2023. IP Geolocation API. https://ip-api.com/.

[23] M. Jonker, G. Akiwate, A. Affinito, k. claffy, A. Botta, G. Voelker, R. Rijswijk-Deij, and S. Savage. 2022. Where .ru?: assessing the impact of conflict on russian domain infrastructure. In *IMC*. ACM, 159–165.

[24] A. Kashaf, V. Sekar, and Y. Agarwal. 2020. Analyzing Third Party Service Dependencies in Modern Web Services: Have We Learned from the Mirai-Dyn Incident?. In *IMC*. ACM, 634–647.

[25] E. Kenneally and D. Dittrich. 2012. The menlo report: Ethical principles guiding information and communication technology research.

[26] S. Kitterman. 2014. Sender Policy Framework (SPF) for Authorizing Use of Domains in Email, Version 1. RFC 7208.

[27] R. Kumar, S. Asif, E. Lee, and F. Bustamante. 2023. Each at its Own Pace: Third-Party Dependency and Centralization Around the World. In *SIGMETRICS*. ACM, 43–44.

[28] R. Li, X. Jia, Z. Zhang, J. Shao, R. Lu, J. Lin, X. Jia, and G. Wei. 2023. A Longitudinal and Comprehensive Measurement of DNS Strict Privacy. *IEEE/ACM Trans. Netw.* 31, 6 (2023), 2793–2808.

[29] R. Li, S. Xiao, B. Liu, Y. Lin, H. Duan, Q. Pan, J. Chen, J. Zhang, X. Liu, X. Lu, and J. Shao. 2024. Bounce in the Wild: A Deep Dive into Email Delivery Failures from a Large Email Service Provider. In *IMC*. ACM, 659–673.

[30] X. Li, Y. Zhu, X. Zhang, Z. Jie, and Q. Liu. 2023. SpoofingGuard: A Content-agnostic Framework for Email Spoofing Detection via Delivery Graph. In *CSCWD*. IEEE, 1098–1105.

[31] E. Liu, G. Akiwate, M. Jonker, A. Mirian, G. Ho, G. Voelker, and S. Savage. 2023. Forward Pass: On the Security Implications of Email Forwarding Mechanism and Policy. In *EuroS&P*. IEEE, 373–391.

[32] E. Liu, G. Akiwate, M. Jonker, A. Mirian, S. Savage, and G. Voelker. 2021. Who's got your mail?: characterizing mail service provider usage. In *IMC*. ACM, 122–136.

[33] E. Lochin. 2010. STAMP: SMTP Server Topological Analysis by Message Headers Parsing. In *CCNC*. IEEE, 1–2.

[34] E. Luo, L. Young, G. Ho, M. Afifi, M. Schweighauser, E. Katz-Bassett, and A. Cidon. 2025. Characterizing the Networks Sending Enterprise Phishing Emails. In *PAM*, Vol. 15567. Springer, 437–466.

[35] Microsoft. 2025. Microsoft 365. https://www.microsoft.com/en-us/microsoft-365.

[36] Microsoft. 2025. Use rules to automatically forward messages. https://support.microsoft.com/en-us/office/use-rules-to-automatically-forward-messages-45aa9664-4911-4f96-9663-ece42816d746.

[37] K. Moriarty and S. Farrell. 2021. Deprecating tls 1.0 and tls 1.1. RFC 8996.

[38] G. Moura, S. Castro, W. Hardaker, M. Wullink, and C. Hesselman. 2020. Clouding up the Internet: how centralized is DNS traffic becoming?. In *IMC*. ACM, 42–49.

[39] Netstar. 2025. URL Categorization and Threat Intelligence Solution. https://incompass.netstar-inc.com/urlsearch.

[40] J. Postel. 1982. Simple Mail Transfer Protocol. RFC 821.

[41] Proofpoint. 2025. https://www.proofpoint.com/.

[42] S. Rao, E. Liu, G. Ho, G. M. Voelker, and S. Savage. 2024. Unfiltered: Measuring Cloud-based Email Filtering Bypasses. In *The Web Conference (WWW)*. ACM, 1702–1711.

[43] Y. Rekhter, B. Moskowitz, D. Karrenberg, G. J. de Groot, and E. Lear. 1996. Address Allocation for Private Internets. RFC 1918.

[44] J. Rijn. 2023. Email is not dead. But email IS changing. https://www.emailisnotdead.com/.

[45] F. Sanchez, Z. Duan, and Y. Dong. 2010. Understanding Forgery Properties of Spam Delivery Paths. https://www.cs.fsu.edu/~duan/publications/2010_ceas_sdp.pdf. In *CEAS*.

[46] Tranco. 2024. Top 1M Domains. https://tranco-list.eu/.

[47] C. Wang, Y. Kuranaga, Y. Wang, M. Zhang, L. Zheng, X. li, J. Chen, H. Duan, Y. Lin, and Q. Pan. 2024. BreakSPF: How Shared Infrastructures Magnify SPF Vulnerabilities Across the Internet. In *NDSS*.

[48] Wikipedia. 2025. Herfindahl–Hirschman index. https://en.wikipedia.org/wiki/Herfindahl-Hirschman_index.

[49] C. Xu, Y. Zhang, F. Shi, H. Shan, B. Guo, Y. Li, and P. Xue. 2023. Measuring the Centrality of DNS Infrastructure in the Wild. *Applied Sciences* 13, 9 (2023). doi:10.3390/app13095739

[50] L. Zembruzki, A. Jacobs, L. Granville, and R. Pfitscher. 2023. Examining the Centralization of Email Industry: A Landscape Analysis for IPv4 and IPv6. In *ISCC*. IEEE, 360–366.

[51] L. Zembruzki, R. Sommese, L., A. Jacobs, M. Jonker, and G. Moura. 2022. Hosting Industry Centralization and Consolidation. In *NOMS*. IEEE, 1–9.