Who's Got Your Mail? Characterizing Mail Service Provider Usage

Enze Liu
UC San Diego
e7liu@eng.ucsd.edu

Gautam Akiwate
UC San Diego
gakiwate@cs.ucsd.edu

Mattijs Jonker
University of Twente
m.jonker@utwente.nl

Ariana Mirian
UC San Diego
amirian@eng.ucsd.edu

Stefan Savage
UC San Diego
savage@cs.ucsd.edu

Geoffrey M. Voelker
UC San Diego
voelker@cs.ucsd.edu

ABSTRACT

E-mail has long been a critical component of daily communication and the core medium for modern business correspondence. While traditionally e-mail service was provisioned and implemented independently by each Internet-connected organization, increasingly this function has been outsourced to third-party services. As with many pieces of key communications infrastructure, such centralization can bring both economies of scale and shared failure risk. In this paper, we investigate this issue empirically — providing a large-scale measurement and analysis of modern Internet e-mail service provisioning. We develop a reliable methodology to better map domains to mail service providers. We then use this approach to document the dominant and increasing role played by a handful of mail service providers and hosting companies over the past four years. Finally, we briefly explore the extent to which nationality (and hence legal jurisdiction) plays a role in such mail provisioning decisions.

CCS CONCEPTS

• **Information systems → World Wide Web**; • **World Wide Web → Internet communications tools**; • **Internet communications tools → E-mail**.

ACM Reference Format:
Enze Liu, Gautam Akiwate, Mattijs Jonker, Ariana Mirian, Stefan Savage, and Geoffrey M. Voelker. 2021. Who's Got Your Mail? Characterizing Mail Service Provider Usage. In *ACM Internet Measurement Conference (IMC '21), November 2–4, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3487552.3487820

1 INTRODUCTION

Despite the rise of interactive chat and online social messaging applications, e-mail continues to play a central role in communications. By some estimates, close to 300 billion e-mail messages are sent and received each day [34]. In particular, e-mail remains the central modality for modern business correspondence — long since displacing the postal service for such matters over the previous two decades.

However, unlike the postal service (and many other forms of person-to-person communication) e-mail is not centrally administered, but is organized such that each Internet domain owner, by virtue of their DNS MX record, can make unique provisioning decisions about how and where they will accept e-mail delivery. Thus, organizations are free to provision separate e-mail services for each domain they own, to share service among domains they operate, or to outsource e-mail entirely to third-party providers. These choices, in turn, can have significant implications for the resilience, security, legal standing, performance and cost of e-mail service.

In particular, concerns have been raised in recent years about the general risks of increasing Internet service centralization and consolidation [5, 10, 17]. For example, centralization amplifies the impact of (even rare) service failures [4, 15, 25]. Similarly, a single data breach in a widely-used service can put thousands of customers' data at risk.[1] Finally, the legal jurisdiction in which a given service provider operates is implicitly imposed on the data managed by that provider. For instance, as a U.S. company, Google-managed data is subject to the Stored Communications Act, which provides data access to the government under warrant even if the data belongs to a foreign party not residing in the U.S..

Indeed, while historically e-mail was provisioned and implemented independently by each organization (*i.e.*, hosting a local mail server acting as a full-fledged Mail Transfer Agent), the rise of third-party enterprise mail service providers (notably Google and Microsoft) has challenged that assumption; indeed, there are compelling reasons to believe that that global e-mail service is also increasingly subject to a significant degree of centralization. However, in spite of the importance of this issue there has been little empirical analysis of e-mail provisioning choices and how they have been evolving over time.[2]

In this paper, we perform a large-scale measurement and analysis of e-mail service provisioning and configuration. Our study uses three large corpora of domains: one based on all `.gov` domains, another based on a stable subset of the Alexa top 1 million domains observed across nine snapshots between 2017 and 2021, and lastly a similar dataset of one million `.com` domains sampled at random

IMC '21, November 2–4, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-9129-0/21/11.
https://doi.org/10.1145/3487552.3487820

[1] The recent vulnerabilities exploited in Microsoft's Exchange Server were serious [20], and it could have been even worse had attackers been able to penetrate Microsoft's Outlook e-mail service.
[2] One example can be found in Trost's blog post "Mining DNS MX Records for Fun and Profit", although, as our results show, their approach has its limitations [36].

from the same period. We use these datasets to gain insight into the present popularity of e-mail service providers and their longitudinal shifts, and to characterize their makeup. From our data we demonstrate the clear and growing dominance of a handful of third-party e-mail service providers and the shrinking number of domains that provision mail service "in-house" themselves or through their hosting providers.

We make the following contributions:

(1) We detail and justify a methodology to map published MX records to the identity of the mail service provider (providing significant accuracy improvements over approaches that entirely rely on MX record content);
(2) Using our methodology we identify the top e-mail service providers and characterize their market share and customer demographics;
(3) We provide a longitudinal analysis of mail service provider popularity over time and document the source of market share shifts;
(4) We explore the existence of national biases in the choice of mail service provider (i.e., the extent to which mail for domains in country X's top-level domain (TLD) make use of mail service from country Y and hence subject themselves to Y's legal jurisdiction).

Ultimately, our work not only provides a comprehensive analysis of the current state of Internet e-mail provisioning (and the relative role of third-party web mail service providers, mail filtering providers and "in-house" mail services), but also provides a solid foundation on which to base future analyses of e-mail infrastructure.

2 BACKGROUND AND RELATED WORK

2.1 Simple Mail Transfer Protocol

The Simple Mail Transfer Protocol (SMTP) is part of a family of protocols for mail transmission, including SMTP [27], Extended SMTP (ESMTP) [18] and SMTP Service Extension for Authentication (SMTP-AUTH) [33].

In its purest form, as depicted in Figure 1, an e-mail user operates a mail user agent (MUA) that uses ESMTP or SMTP-AUTH to submit e-mail messages to the sender's mail submission agent (MSA) software (e.g., their local mail server). The MSA in turn queues the message for delivery with the sender's mail transfer agent (MTA) for relay to the mail infrastructure of the addressed parties in the *To:*, *CC:* or *Bcc:* lines. Next, the sender's MTA transfers the e-mail to the recipient's MTA, using SMTP or — if supported — ESMTP. It is during this step that the sending MTA uses the recipient's DNS "Mail Exchanger" (MX) record to determine the location of the receiving MTA. Having received the e-mail, the receiving MTA then either delivers the mail locally or places it into a queue for further processing. In practice, the MSA and MTA are often the same piece of software (typically run on a single server in an "in-house" implementation) and in Web mail situations (e.g., Gmail) the MUA is a Web application provided by the same organization as the MSA and MTA.

2.1.1 SMTP Procedures: A Summary. All protocols in the SMTP family follow roughly the same procedure. A session starts when

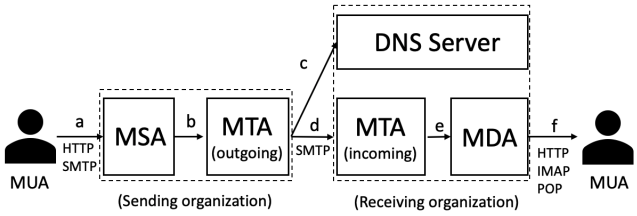

Figure 1: Mail processing model

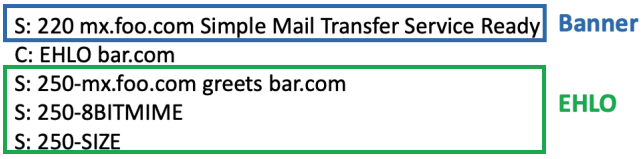

Figure 2: Banner and EHLO message in a typical SMTP session between client (C) and server (S).

an SMTP client (either an MUA seeking to submit mail or an MTA seeking to relay mail) opens a connection to an SMTP server, which responds in kind with a greeting message. This message is informally referred to as the *banner message*, in which the server typically provides either its domain name or IP address [18].

Once the SMTP client has received the greeting message, it normally sends the EHLO (or HELO in earlier versions) command to the SMTP server, signaling its identity, which in turn elicits an EHLO response message containing the SMTP server's domain name and a list of the extensions it supports. Figure 2 illustrates the *banner* and *EHLO message* in a typical SMTP session with the SMTP server (S) having domain `foo.com` and the SMTP client (C) having domain `bar.com`. In this paper, we use *EHLO message* to refer to the second EHLO, i.e., the message elicited from the server.

Depending on the protocol, additional messages may be exchanged between server and client for negotiating configuration options such as authentication. The sending SMTP server can then initiate a mail transaction. These last steps are important for the delivery of message content, but are not relevant to this paper.

2.1.2 Mail submission and mail relaying. When the SMTP protocol is used to submit a new message, e.g., between the sender's MUA and their MSA, the identity of the mail server is typically well-known (i.e., pre-configured) and it is common for the MUA to positively authenticate themselves using the SMTP-AUTH protocol. Thus, the server will not accept SMTP transactions before the sender presents appropriate credentials (also typically protected via a TLS session initiated as part of this protocol step). In this fashion, the customer-facing mail server designated by a broadband Internet Service Provider is able to limit outbound mail submissions to only their customers. In this mail submission mode, servers typically accept connections on TCP port 587, as per RFC 6409 [19]. However, port 465 is also common (although 465 was deprecated in RFC 8314 [24]), and in a number of cases sites may use port 25 for this purpose (typically designating particular hosts to be MSAs and others to be MTAs [19]).

When the SMTP protocol is used to relay a message (i.e., from one MTA to another), the sending (i.e., outbound) MTA identifies its partner MTA server by parsing e-mail addresses (i.e., `user@domain`) to extract the associated domain names. For each (unique) domain name in the destination address(es) of an e-mail, the sending MTA will lookup a DNS MX record. This MX record points to the server to which receiving e-mail on behalf of the particular domain name is delegated. By fully resolving this record, the sending MTA server ultimately identifies and establishes a connection with the receiving MTA server. In this mail relay mode, TCP port 25 is typically used (there are other ports that are used occasionally, such as port 2525, but these are not supported by IANA or IETF [39] and so we do not consider them in this paper).

2.2 Mail Exchanger Records

The Mail Exchanger (MX) record specifies which MTAs handle inbound mail for a domain name [18, 24, 26] and is published in the DNS zone of the domain. An MX record should itself contain a valid domain name [23, 26]. Multiple MX records can be configured in a zone, each with an assigned preference number. The lowest preference has highest priority, and multiple MX records can share the same priority for load balancing [18]. An MX record can be made up, in part, of the registered domain name for which it receives e-mail, yet resolve to completely separate infrastructure. For instance, the MX record for our institution `ucsd.edu` contains `inbound.ucsd.edu`, which in turn resolves to an IP address (A record) owned and operated by ProofPoint, a well-established mail filtering company wholly different from `ucsd.edu`.

2.3 STARTTLS and TLS certificates

Modern SMTP implementations opportunistically support the STARTTLS option which, in the mail relay context, allows the sending MTA to initiate a TLS connection with the receiving MTA [11, 16]. If the receiving MTA supports STARTTLS, it will provide a TLS certificate which can be used to bootstrap a TLS session providing session confidentiality. To provide a valid certificate, the receiving MTA must obtain a signed certificate from a trusted certificate authority (CA) for which the MX domain name is either specified in the Common Name (CN) or a Subject Alternative Name (SAN) field. While ideally TLS certificates are validated by the sending MTA, in practice SMTP sessions will continue even if the certificate does not validate [13, 14]. Note that the SAN field is used when a single certificate must support TLS connections across a range of domains. For example, the certificate used by Gmail has Common Name `mx.google.com`, and its SAN specifies other alternate domain names, such as `aspmx2.googlemail.com` and `mx1.smtp.goog`.[3] In these cases, the Common Name (CN) almost always specifies a principal domain operated by the provider of the service.

2.4 Related work

Considering its critical role, remarkably little contemporary analysis exists of e-mail infrastructure and who provides it. Some of the best known modern work in this space is the pair of 2015 papers authored by Durumeric *et al.* and Foster *et al.* which empirically explored the use and configuration of privacy, authentication, and integrity mechanisms at each stage of the e-mail delivery pipeline [13, 14]. Notably, Durumeric *et al.* also provide one estimate of the top mail providers as a part of their study, although their methodology may underestimate the influence of major providers (notably Microsoft). Rijswijk *et al.* [37, 38] investigated the growth of three top mail providers over a relatively short, 50-day period, and demonstrated the phasing out of Windows Live over Office365, among others. Their analysis, unlike ours, considers only the content of MX records, and mail was not the focal point of their work. Finally, in 2005, Afergan *et al.* [2] measured the loss, latency, and errors of e-mail transmission over the course of a month with hundreds of domains.

Somewhat further afield, there is a literature exploring how dangling DNS records impact e-mail security, starting with the work of Liu *et al.* [22], who explored e-mail as a special case of a general analysis of dangling DNS issues. This work was recently expanded by Reed and Reed in their technical report that focuses specifically on dangling DNS MX records and their potential security impact [29]. Another direction of research, notably by Chen *et al.* [9] and Shen *et al.* [32], studies the vulnerabilities of third-party mail providers and how those vulnerabilities could be used to spoof e-mail messages.

In spite of these and related efforts, we have found very little work focused on characterizing which organizations are, in fact, responsible for providing mail service or how this responsibility has changed over time. Indeed, perhaps the closest related work is not from the academic literature, but from the recent Medium post of Jason Trost which describes an analysis of MX records for identifying e-mail security providers [36].

3 IDENTIFYING MAIL PROVIDERS

In this section, we first illustrate the challenges in identifying mail service providers, in particular how MX records alone can be misleading, and the strengths and weaknesses of using alternative features. Given these limitations, we then present our *priority-based* approach for identifying the mail provider for a given domain name. For the purpose of this work, we focus on the primary e-mail provider, which is identified by the MX record with the highest priority. Finally, we evaluate the accuracy of this approach using randomly sampled domains from the three larger datasets of domains on which we base much of our subsequent analysis (described in detail in Section 4.3).

3.1 Challenges in Provider Identification

One approach, exemplified by Trost's analysis [36], relies exclusively on MX records to identify the mail provider. However, this approach can be misleading when the purported MX domain resolves to an IP address operated by a different entity.

Better accuracy can be achieved by incorporating additional features, such as the autonomous system number (ASN) of the IP address to which an MX record resolves, the content of Banner/EHLO messages in the initial SMTP transaction, and TLS certificates learned during an SMTP session. However, using multiple features creates additional complexities. In particular, while SMTP-level information is typically a more reliable indicator of

[3] `mx1.smtp.goog` is a valid and resolvable domain owned by Google.

Domain	MX	MX IP Resolution	ASN of IP
netflix.com	aspmx.l.google.com	172.217.222.26	15169 (Google)
gsipartners.com	mailhost.gsipartners.com	173.194.201.27	15169 (Google)
beats24-7.com	mx10.mailspamprotection.com	35.192.135.139	15169 (Google)
jeniustoto.net	ghs.google.com	172.217.168.243	15169 (Google)

Table 1: Example domains with related mail information.

Domain	Banner/EHLO	Subject CN
netflix.com	mx.google.com	mx.google.com
gsipartners.com	mx.google.com	mx.google.com
beats24-7.com	se26.mailspamprotection.com	*.mailspamprotection.com
jeniustoto.net	N/A	N/A

Table 2: Example domains with additional information retrieved from SMTP sessions.

mail service provider than the hosting party's ASN, the latter is always available while the former is not.

To illustrate these points further, we use the four domains listed in Tables 1 and 2 as examples. Table 1 shows the MX record, the IP address resolution, and the ASN from which the address is announced. Table 2 shows additional information learned by initiating SMTP sessions with the IP addresses listed in Table 1. Specifically, we show the subject Common Name (CN) listed on the certificate presented in STARTTLS (if any) and the Banner/EHLO messages provided during the SMTP session.

3.1.1 MX Record. Using the MX record to infer the mail provider works well when the domain owner explicitly names its provider in the MX record (e.g., netflix.com in Table 1). This is a common practice for domains that outsource their mail services to third-party companies (e.g., Google) to ensure that their providers can property receive e-mail on their behalf [28, 35].

However, this idiom is not always accurate. For example, the MX approach will incorrectly infer that gsipartners.com self-hosts its e-mail delivery because its MX record is mailhost.gsipartners.com. However, this MX name resolves to an IP address announced by Google. When contacted, it emits mx.google.com Banner/EHLO in the SMTP handshake, and the TLS certificate it produces has a subject common name (CN) of mx.google.com. Clearly, gsipartners.com e-mail is handled by Google.

3.1.2 Autonomous System Number (ASN). While the ASN to which the mailhost.gsipartners.com MX leads correctly indicates Google as the mail provider, this inference is not always accurate. Consider the domain beats24-7.com whose MX record also resolves to an IP address owned by Google. In this case e-mail is actually handled by an e-mail security provider that is hosted in Google Cloud's IP space, rather than the internal address space used by Google to host its own services. Another issue with the ASN is that it does not reflect whether an IP address is actually operating an SMTP server and can accept mail. Consider jeniustoto.net in Table 1, which has an MX record that resolves to an IP address in Google's internal address space. However, this IP address is from Google's web hosting service and does not run an SMTP server. In this case, jeniustoto.net does not actually have a mail server (and thus a mail provider), even though it uses a Google IP address.

3.1.3 Banner/EHLO messages. During an SMTP session, the mail server for gsipartners.com identifies itself in its Banner/EHLO handshake as mx.google.com (Table 2). This information is generally reliable for identifying third-party mail providers, as most third-party providers configure their servers to properly identify themselves. However, the Banner/EHLO information need not be mechanically generated and can contain any text configured by the server operator, which makes it unreliable in a small number of scenarios. First, Banner/EHLO messages may not contain valid domain names. For example, instead of having a valid domain name, certain providers put a string (e.g., IP-1-2-3-4) in their servers' Banner/EHLO messages. Second, an individual, who runs their own SMTP server, can falsely claim to be mx.google.com in Banner/EHLO messages. While very rare, we have observed a small number of such cases.

3.1.4 TLS certificate. The gsipartners.com mail server also presents a valid certificate with subject CN mx.google.com, which is a clear indicator of the entity running the mail server (and one attested to by a trusted Certificate Authority) and thus can generally be used to infer the mail provider. In the case of gsipartners.com, we conclude that it uses Google as it presents a valid certificate with subject CN mx.google.com (this certificate is also used by other legitimate Google mail servers).

While certificates are ideal for identifying the mail provider of a domain, they are not always available. Some mail servers do not support STARTTLS or they respond with self-signed certificates which are less reliable. Additionally, we note that certain web hosting providers (e.g., GoDaddy with domain name secureserver.net) allow their virtual private servers (VPS) to create certificates using specific subdomains as the subject CN (e.g., vps123.secureserver.net). These servers are operated by individuals renting them instead of the web hosting company providing the infrastructure. Thus, in this case, the subject CN reflects the hosting provider (e.g., GoDaddy) instead of the mail provider (e.g., a self-hosted mail server operated by an individual operating

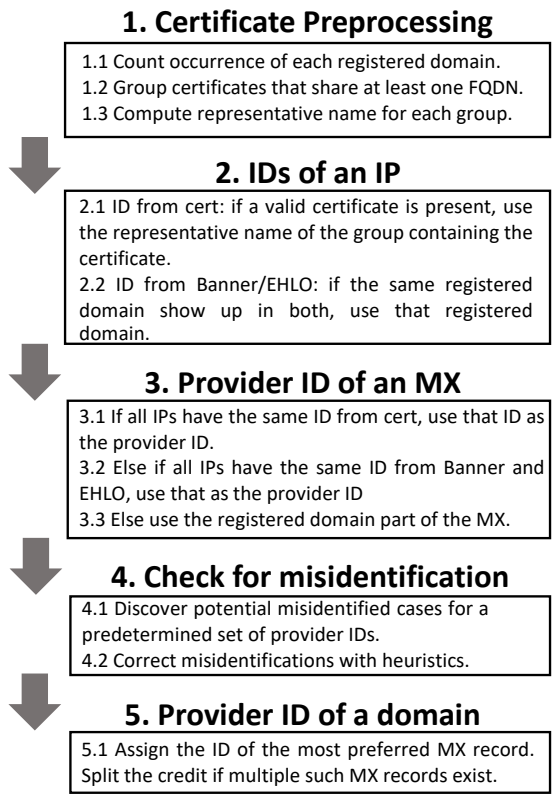

Figure 3: Our five-step approach to infer the provider of an MX record. The approach considers data from MX records, Banner/EHLO messages, and TLS certificates to determine the e-mail provider.

a GoDaddy VPS). Lastly, in a handful of cases, we observe that some third-party mail service providers present the certificates of their customers. For example, the University of Texas (`utexas.edu`) has an MX record (`inbound.utexas.edu`) that resolves to an IP address that, when contacted, presents a valid certificate with CN `inbound.mail.utexas.edu`. However, the ASN of that IP address suggests that mail service is operated by Ironport, an e-mail security company. Additionally, the server indicates in its Banner/EHLO message that it is Ironport. In this case, we can conclude that the University of Texas is using Ironport instead of hosting their own e-mail infrastructure. Thus, the CN presented in the certificate does not indicate the service provider in this instance.

Based on these observations and our experience, we propose an approach that prioritizes SMTP level information when available, and falls back to MX level information in other cases. This approach achieves both good accuracy and avoids the availability issues with SMTP level information. We provide more details below.

3.2 Methodology: A Priority-Based Approach

We propose a methodology, which we term the *priority-based approach*, that takes as input a domain (and relevant information) and outputs a *provider ID* as the inferred primary mail provider responsible for mail service for that domain. Our methodology incorporates data from multiple sources, including MX records, Banner/EHLO messages, and TLS certificates. We achieve high accuracy through prioritizing these sources by reliability: certificates first, then Banner/EHLO messages, and then MX records.

Our methodology consists of five steps shown in Figure 3. First, we preprocess all certificates to find and group certificates that are potentially operated by the same entity. For each group of certificates, we designate a representative name to represent the entity owning these certificates. Second, for each IP address that an MX record resolves to, we try to determine IDs that best represent the mail provider associated with that IP address. Since an MX can resolve to multiple IP addresses, knowing the mail provider operating each IP address is a prerequisite for determining the provider ID of an MX. Next, we assign a *provider ID* to the MX record. We then filter for misidentifications and correct them to the best of our ability. Finally, we assign a provider ID to a domain, which is a registered domain representing the entity operating the mail infrastructure pointed by the MX record.

We detail our five step methodology below, using the examples shown in Table 3, in which domains `third-party1.com` and `third-party2.com` use e-mail services provided by the third-party provider `provider.com`, domain `myvps.com` operates its own e-mail service on a VPS hosted with `provider.com`, and domain `selfhosted.com` operates its own mail service.

3.2.1 Certificate Preprocessing. The goal of the first step — preprocessing — is to find certificates that are potentially operated by the same mail provider. The domains listed in a certificate aid our mail provider inferences. However, certificates also introduce two issues. First, a mail provider can have multiple valid certificates. Additionally, each certificate can contain multiple domain names by using the subject alternative name (SAN) extension. Having multiple certificates, each with multiple domain names, leads to two challenges: which certificates belong to the same mail provider, and which name to use to represent that provider.

We address these two challenges by preprocessing all certificates in our dataset and grouping certificates that likely belong to the same mail provider. We output a *representative name* for each group to represent that group and the mail provider. The process of grouping certificates and producing a representative name has three steps:

(1) **Count Occurrences of Each Registered Domain**: For fully qualified domain names (FQDNs) that appear on a certificate's Subject CN and SANs, we take the registered domain part (e.g., in Table 3 `provider.com` is the registered domain of both `mx1.provider.com` and `mx2.provider.com`) and count occurrences of each registered domain across all certificates. For example, in Table 3, the count for `provider.com` will be 5. We extract the registered domain from the FQDN using the Public Suffix List [21].
(2) **Grouping Certificates**: Providers may use different certificates across their infrastructure, and grouping consolidates

Domain	MX	MX IP	Banner/EHLO	Subject CN	SANs	Provider ID
third-party1.com	mx1.provider.com	1.2.3.4	mx1.provider.com	mx1.provider.com	mx2.provider.com	provider.com
third-party2.com	mx2.provider.com	2.3.4.5	mx2.provider.com	mx2.provider.com	mx1.provider.com	provider.com
myvps.com	mx.myvps.com	3.4.5.6	myvps.provider.com	myvps.provider.com	N/A	provider.com
selfhosted.com	mx.selfhosted.com	4.5.6.7	ip-4-5-6-7	N/A	N/A	selfhosted.com

Table 3: Example domains and relevant information used in our methodology.

them into sets of related FQDNs. We put two certificates into the same group if (and as long as) there is some degree of overlap between their sets of FQDNs. For instance, in Table 3, we would create two groups. We merge the certificates used by third-party1.com and third-party2.com into one group, as they contain the same set of FQDNs: mx1.provider.com and mx2.provider.com. The certificate with subject CN myvps.provider.com is in its own group.

(3) **Selecting a Representative Name**: For each group of certificates, we choose the most common registered domain as the representative name, as it is likely to represent the mail provider best. In our specific example, the representative name for both groups is provider.com.

At the end of this process, certificates are organized into groups and each group will have a representative name.

3.2.2 Identifying IDs for an IP Address. Before assigning a mail provider ID to an MX record, we need to determine the ID(s) that best represent(s) the mail provider for the IP address(es) to which an MX record resolves. We compute one ID with certificates and another ID with Banner/EHLO messages. We also prioritize the ID computed with certificates when using both IDs.

(1) **ID from TLS Certificates**: If a valid certificate is present at the IP address, we use the *representative name* of the group containing the certificate as the ID. We consider a certificate valid if it is trusted by a major browser (e.g., Firefox). In our example, IP addresses 1.2.3.4, 2.3.4.5, 3.4.5.6 would have the ID provider.com from certificates.

(2) **ID from Banner/EHLO Messages**: If the Banner/EHLO message is available and contains a valid FQDN, we use the registered domain part of the FQDN as the ID. In our example, we cannot assign an ID to IP address 4.5.6.7 because it does not present a certificate and its Banner/EHLO message does not contain a valid FQDN. The other three IP addresses have the ID provider.com from Banner/EHLO messages.

3.2.3 Identifying Mail Provider ID for an MX Record. Once we have computed IDs for each IP address, we next analyze the MX records. If all IP addresses of an MX record have the same ID from certificates, we assign that ID as the provider ID to the MX record. In cases where IDs from certificates do not agree or are not available, we check if all IP addresses share the same ID from Banner/EHLO messages. If so, we assign that provider ID to the MX record. Otherwise, we fall back to using the registered domain part of the MX record as the provider ID.

3.2.4 Checking for Misidentifications. While this approach can infer the mail provider of an MX record correctly in most cases, there exist a few that lead to misidentifications. In the above example, for domain myvps.com, we infer that its MX record mx.myvps.com is operated by provider.com using the ID from certificates. However, myvps.com is running its own mail server on a VPS hosted with provider.com. In fact, this example represents a situation that is hard to identify both automatically and correctly: VPS servers hosted with web hosting companies. Certain web hosting companies (e.g., GoDaddy with domain name secureserver.net) allow their VPS servers to create certificates under specific domain names (e.g., vps123.secureserver.net). Similarly, as mentioned above, certificates can be misleading when third-party providers present their customer's certificates. Since there is no good way to automatically detect such cases without prior knowledge, we have to identify such situations manually.

Another source of error comes from Banner/EHLO messages. Recall that Banner/EHLO messages are unrestricted text. Thus, it is possible to falsely claim to be mx.google.com in Banner/EHLO messages. Since our approach prioritizes Banner/EHLO messages over the MX record, we would mislabel it as google.com.

To efficiently find instances of misidentifications, we use the observation that the corner cases mentioned above are for unpopular servers, with few domains pointing at them. For example, IP addresses used by VPS servers (and associated certificates) would only show up a handful of times in our dataset. By contrast, IP addresses (and their associated certificates) used by MX records of popular third-party mail providers would generally be much more common in our dataset, as those MX records would be used by many domains. Thus, it is possible to quickly find potentially misidentified MX records by looking at the number of domains pointing at them.

We identify potential instances of misidentifications using the observation above. We keep two counters globally. We keep track of the number of domains that point to each IP address (num_{IP}) and each certificate (num_{Cert}). For each IP address, the confidence score of its mail provider ID inference is $max(num_{IP}, num_{Cert})$. If an IP address does not have certificate information, num_{Cert} is ignored. For any dataset of a reasonable size, this score largely reduces the number of cases we need to examine. That said, it is still unrealistic to perform such manual work for all the providers on large datasets. Thus, we only check for misidentifications for large providers.

Once we have identified potential candidates to examine, we employ various heuristics to ease the process of manually going through all of them. For example, we can quickly determine a server is falsely claiming to be google.com if it does not reside in Google's AS. Similarly, we observe that GoDaddy uses specific hostnames for their dedicated servers (e.g., mailstore1.secureserver.net) and

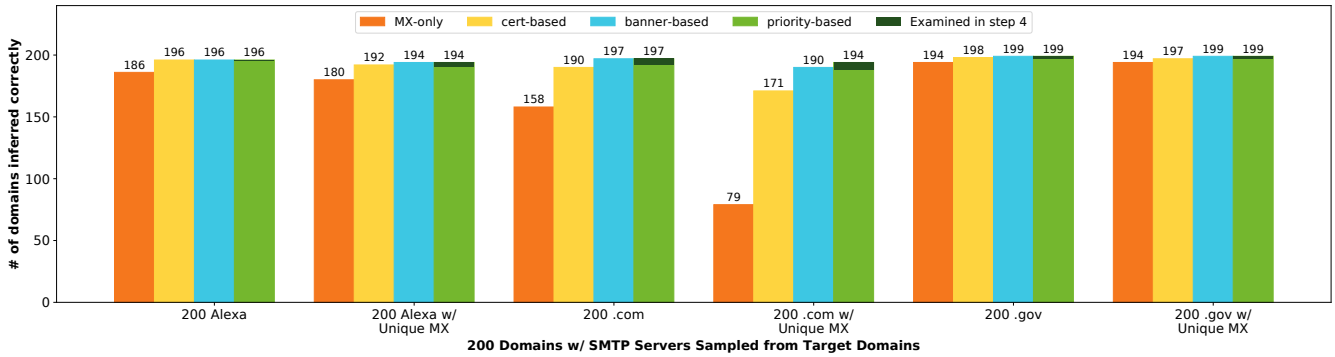

Figure 4: Accuracy of different approaches on 200 domains sampled from the three lists of target domains.

different patterns for VPS servers (e.g., `s1-2-3.secureserver.net`). Such observations can help us quickly sift through all candidates.

3.2.5 Identifying Mail Provider ID for Domain. At the end, every MX record will have an assigned mail provider ID. This assignment could be either based on TLS certificate information, Banner/EHLO messages, or the MX record itself. Based on the MX record that a domain uses we can assign a mail provider to that domain. In the case that a domain has more than one primary MX record (multiple MX records with the same priority but different provider IDs, which happens occasionally), we split the domain across the multiple providers.

3.3 Relative Accuracy of Approaches

The priority-based approach combines the use of TLS certificates, Banner/EHLO messages, and MX records. Each of these sources could be independently used to determine the mail provider for a domain. As such, we have four potential approaches: (1) the MX-only approach [36], (2) a *cert-based* approach that combines TLS certificates and MX records, (3) a *banner-based* approach that combines Banner/EHLO messages and MX records, (4) the *priority-based* approach that combines TLS certificates, Banner/EHLO messages and MX records.

We evaluate the four approaches and their relative accuracy using 200 random domains sampled from three sets of domains in two ways, resulting in an evaluation set of 1,200 domains. The three sets of domains we randomly sample are: all `.gov` domains, a stable set of domains from the Alexa list, and a stable set of 1 million `.com` domains (see Section 4.1 for how we define stable domains). We sample (a) 200 domains and (b) 200 domains with unique MX records from the three datasets.

Since there is no ground truth for mail providers, we use domains with SMTP servers, scan the relevant information ourselves, and manually label their providers.[4] We then use this labeled data to compare the results of the different methods.

Figure 4 shows the results. The dark green part of the priority-based approach highlights the total number of candidates manually examined in step 4 (check for misidentifications) of our approach. In general, the priority-based approach works the best among all four approaches for the two sets of domains, with an accuracy of at least 97%. In total, it missed 21 domains (1.8%) out of 1200 domains sampled and required us to manually examine 20 (1.7%) domains.

Among 21 domains it missed, we cannot decide the providers of 4 domains. Three of these four domains are hosted on servers with unpopular web hosting companies. We do not have enough information and confidence to decide if the servers are VPS instances rented from the web hosting companies or directly managed by them. One presents a valid certificate of company A, but indicates that it is company B in Banner/EHLO messages (a situation much like `utexas.edu` described above). However, unlike `utexas.edu` which is hosted with a well-known provider, both company A and B are relatively unpopular and we are not confident enough to decide whether company A or B is running the mail server. Out of 17 domains for which we decide the provider, 11 are VPS servers that use subdomains of the web hosting companies in their certificates or Banner/EHLO messages (like the GoDaddy example mentioned above),[5] 4 are poorly configured servers with Banner/EHLO messages containing strings like `localhost` operated by web hosting companies, and 2 are poorly configured local servers that supply FQDNs that are misleading in their Banner/EHLO messages. For the 20 domains that require manual examination, our heuristic, which we publish together with our code, can automatically determine if they need to be corrected. The amount of labor required in the step is small.

The MX-only approach, on the other hand, relies upon just one data source, and consequently performs the worst among all four approaches (notably with an accuracy of only 40% for 200 random `.com` domains with unique MX records). We also observe that its performance is significantly better on Alexa and `.gov` domains than `.com` domains. We suspect two factors contribute to this phenomenon. On the one hand, if a domain (e.g., `foo.com`) is hosted with a web hosting company, often its MX record will be configured as `mx.foo.com` (a default configuration employed by many web hosting companies), leading the MX approach to believe that the domain runs its own mail infrastructure. On the other hand, stable Alexa and `.gov` domains are generally well-configured and more likely to

[4]Note that we select 200 domains with SMTP servers to ensure a fair comparison across different methods. Some methods (e.g., the MX-only approach) are oblivious to SMTP server presence, and their accuracy drops considerably if domains with MX records but without SMTP servers are in the sample.

[5]Recall that we only check for misidentifications for large providers.

name their mail providers in the MX records, in which cases the MX approach works well.

Considering information from certificates and Banner/EHLO messages increases accuracy by at least a few percent. Note that the banner-based approach performs better than the cert-based approach. This is because, as mentioned in Section 3.1, while more reliable, certificates information is less often available than Banner/EHLO messages. Finally, we note that the banner-based approach achieves an accuracy that is close to the priority-based approach in most cases. These results suggest that the banner-based approach is a good fallback in cases where certificates are not available.

Overall, the priority-based approach performs the best among these four approaches, identifying at least 5 and at most 115 more domains than the MX approach on the 200 sampled domains.

3.4 Limitations

The priority-based approach does have several limitations. First, the flow of exchanging e-mail could involve multiple hops, and we only observe the first step of delivery using DNS MX records. As a result, our inference result may not always reflect the eventual e-mail provider used by users of a domain. Certain heuristics, such as SPF records, might help discover the eventual e-mail provider. However, this is not the focus our work and we leave this as future work. Second, the MX records of a domain could point to any arbitrary server, and there is no guarantee that the server is actually the one responsible for handling the domain's incoming mail. However, this is a limitation that all approaches share. Furthermore, we develop a generic inference method based on IPv4 addresses. We imagine future work extending this method to incorporate IPv6 addresses and better handle corner cases in an automatic way (e.g., with machine learning techniques). Finally, the priority-based approach relies on both DNS data and active measurement data. To carry out the longitudinal analysis in Section 5, we rely on scanning information made available by third-party services like OpenINTEL and Censys. As such, our results can have blind spots (e.g., Censys may not scan IP addresses if certain providers choose to opt out of scans or if it has a bug).

4 LARGE-SCALE IDENTIFICATION OF MAIL PROVIDERS

We now apply the priority-based approach to three lists of target domains collected from OpenINTEL [38] and Censys [12]. For each list we consider nine separate days of data (except for the `.gov` domains, for which we only had seven snapshots), equally spaced over a four-year period between June 2017 and June 2021.

4.1 Target Domains

The first set of domains consists of the Alexa Top 1M domains [3] that have an MX record in their DNS zone. To capture long-term dynamics in mail provider use, we only consider stable domains that consistently appear on the Alexa Top lists across the four years of our study. Considering only the domains that are stable across the years also eliminates noise from the churn [31] in the Alexa Top 1M rankings.

Since the Alexa domains are by definition popular domains, for comparison we also use a set of stable, random `.com` domains as a second list. As with the Alexa domains, we consider `.com` domains with MX records that are registered across the four years. We start by randomly choosing 1M `.com` domains on June 8, 2017 (the first day we consider) and then filter out domains that expire before June 8, 2021 (the last day we consider) or do not have MX records. We remove Alexa domains that also appear in this dataset to create a disjoint view.

The last dataset consists of all `.gov` domains that have an MX record in their DNS zone. Since OpenINTEL does not have coverage of all `.gov` domains in 2017, our measurement data of `.gov` domains starts in June 2018 and consists of seven snapshots instead of nine. Similar to the `.com` domains, we remove Alexa domains that also appear in this dataset to create a disjoint view.

Overall, the Alexa set contains 93,538 domains, the `.com` set contains 580,537, and the `.gov` set contains 3,496 domains. The three sets of domains provide insight into the changing mail provider landscape for popular domains, random domains sampled from the full distribution of registrants in `.com`, and domains in a restricted TLD.[6]

4.2 External Data Sources

To enable our longitudinal and large-scale identification of mail providers, we use two external data sources: OpenINTEL [38] and Censys [7, 12].

4.2.1 OpenINTEL: Active DNS Measurement Data. OpenINTEL is a DNS measurement platform that collects snapshots of a large part of the DNS on a daily basis. It does so by structurally querying substantial lists of domain names for sets of Resource Records (RRs). These lists include, for example, all registered domain names under specific zones such as `.com`. Other sources of names, such as the Alexa Top 1M, are also targeted for measurement. The resulting data accounts for MX records as well as for IP addresses (i.e., A records) associated with the names found inside MX records. By using OpenINTEL data, which allows us to look years into the past, we can investigate MX configuration at scale and perform a longitudinal analysis.

4.2.2 Censys: Internet Scanning Data. Censys is a service that performs regular Internet-wide scans on a wide range of ports in the IPv4 address space, and publishes the data collected. For example, Censys regularly scans IP addresses on port 25 and, if hosts respond, collects application-layer information. For our study, we use the port 25 scans that capture the banner and EHLO messages, as well as any certificates discovered from the SMTP or STARTTLS handshake. It is worth noting that, though Censys performs Internet-wide scans, it may not have data for all IP addresses: the IP address may not publicly accessible, the IP address may be blocked due to requests from the address owner, the host may not listen (or have open) the specific port on the day the scan was performed, or the Censys scan may have failed to cover certain IP addresses intermittently. These issues may skew results for methods that rely upon certificates and Banner/EHLO messages. We also note that

[6] Note that we randomly sampled 400 domains each from these three lists to evaluate our methodology in Section 3.3.

Category	Alexa Domains	COM Domains	GOV Domains
No MX IP	1,692	23,040	49
No Censys	3,215	17,842	160
No Port 25 Data	8,419	63,042	200
No Valid SSL Cert.	19,920	279,002	665
No Valid Banner/EHLO	2,074	9,992	342
No Missing Data	58,218	187,619	2,080
Total	93,538	580,537	3,496

Table 4: Breakdown of data from the June 2021 snapshot of the Alexa domains and random .com domains. These domains have MX records and exist across nine snapshots spanning four years.

Company	Provider ID	ASN
Microsoft	outlook.com office365.us hotmail.com outlook.cn outlook.de	8075 (Microsoft) 200517 (MS Deutschland) 58593 (Blue Cloud)
ProofPoint	gpphosted.com ppops.net pphosted.com ppe-hosted.com	52129 (ProofPoint) 26211 (ProofPoint) 22843 (ProofPoint) 13916 (ProofPoint) 15830 (Telecity Group)

Table 5: Provider IDs operated by Microsoft and ProofPoint identified in our datasets.

Censys recently rolled out an upgraded scanning system, which reportedly fixed some bugs and should have better coverage [8]. However, for consistency reasons, all of our data is taken from the previous system.

4.3 Data Gathering

We start with the *target* list of domain names (e.g., stable domains in Alexa top 1M list) as well as one or more *dates* for which to gather data. We then extract from OpenINTEL the relevant DNS records for domains in the target list on the selected dates. The extracted data includes the MX records associated with the target domains, as well as the IP addresses to which the names in those MX records resolved. We use CAIDA's IPv4 prefix-to-AS data [6] to augment the IP addresses with routing information such as AS number. For each IP address obtained from OpenINTEL, we query Censys for the associated scanning information related to port 25. This data includes the state of the port and data from SMTP and STARTTLS handshakes, including Banner/EHLO messages and certificates. Table 4 shows how we filter data collected for a day's snapshot.

4.4 Providers and Companies

On the data thus gathered, we then use the priority-based approach (Section 3) to determine the mail providers for the domains. Our methodology outputs provider IDs (in the form of registered domains) as mail providers. For example, our methodology tags `google.com` as the provider ID for `netflix.com` (as seen in Table 1). The provider ID `google.com` can then be associated with the mail service provider company, which is Google in this case. However, a single company may have multiple provider IDs, which can either be the result of different services operated by the company or different sources of data (certificates, Banner/EHLO messages, or MX) used to derive the provider ID. Table 5 shows various provider IDs used by Microsoft and ProofPoint identified in our datasets as well as the ASN information of the mail infrastructure.

For our analyses, we ultimately want to aggregate the registered domains that make up provider IDs into the companies that operate these names. This step requires a certain amount of manual work, which makes a blanket analysis of providers infeasible. Instead, we focus on the most prominent mail providers. We investigate frequently-occurring names to identify prominent provider IDs. We then map these provider IDs to companies by examining relevant information (e.g., ASN and the provider ID itself) and searching on the Internet.[7] We use the resulting company information as input for our analyses in Section 5.

5 ANALYSIS

In this section we characterize various aspects of mail providers identified for our target set of popular and random domains (Section 4). We characterize the market share, infrastructure and services provided by the dominant companies in e-mail delivery, their trends over time with particular focus on e-mail security services and web hosting companies, the dynamics of domains switching companies over the span of our data set, and mail provider preferences across different countries.

5.1 Market Share of Top Companies

We start by examining the most popular companies that MX records refer to. We use the priority-based approach from Section 3 to identify the provider IDs most prevalent. We then associate these provider IDs with companies (Section 4.4).

Figure 5 shows the top five companies for the three sets of domains in the most recent snapshot in our dataset (June 2021).[8] Since prior work [30, 31] has demonstrated that the nature of domains in Alexa vary with ranks, we also present the five top companies for domains in the Alexa Top 1k, 10k and 100k. Finally, for `.gov` domains we also identify the top five companies separately for federal and non-federal domains.

For Alexa domains of different ranks, the top two are consistently mail hosting providers (Google and Microsoft). For the top 1k, 10k and 100k domains, the third most popular company is ProofPoint, an e-mail security company. However, when considering all Alexa domains, the third company is Yandex, a Russian mail hosting

[7] Note that our list of provider IDs associated with a company is never meant to be exhaustive. Identifying domain names owned by the same company is a research question by itself [40].

[8] The appendix contains a longer table that lists the number and percentage of the top 15 companies in each set of domains. Provider IDs associated with each company can be found in our GitHub repository: https://github.com/ucsdsysnet/mx_inference

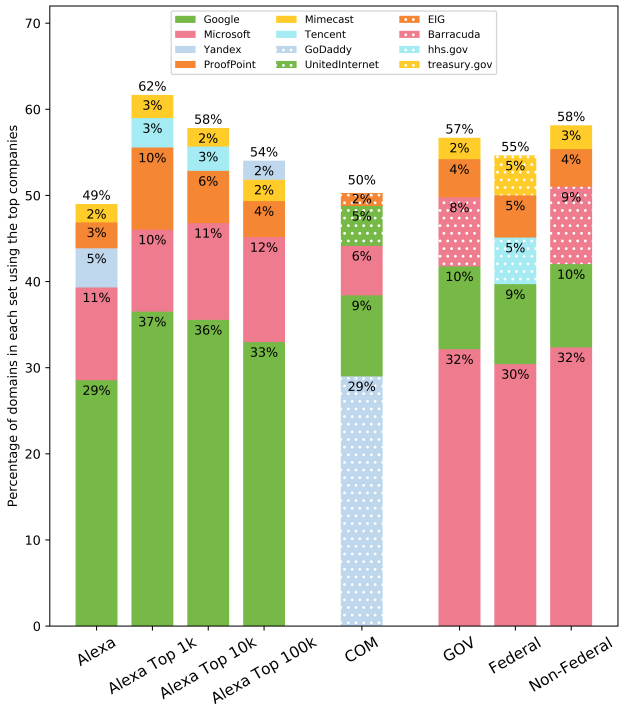

Figure 5: Top providers and the number and percentage of domains using these companies in different sets of domain names (Jun. 2021).

provider. We suspect this likely reflects the presence of many `.ru` domains in the long tail of Alexa domains.

We observe similar phenomena in `.gov` domains: Microsoft and Google are the most prominent companies (although their market shares are reversed), followed by several e-mail security companies (Barracuda, ProofPoint and Mimecast). That said, we observe a non-negligible amount of domains pointing at mail servers operated by the US Department of Health (`hhs.gov`) and the US Department of Treasury (`treasury.gov`) among federal domains. Manually checking a random sample suggests that most of these domains are either directly operated by or closely related to the two departments.

Finally, for `.com` domains, we note a slightly different company distribution. While Google and Microsoft still have a significant presence, the other companies are web hosting providers (GoDaddy, UnitedInternet, and EIG). Indeed, GoDaddy by far has the dominant market share among the random `.com` domains.[9] In contrast to the Alexa and `.gov` sets, the random domains reflect the full distribution of sites using MX records. This distribution has a long tail with many small sites, and it is not surprising that many of them operate using the infrastructure of their hosting provider. Finally, while e-mail security services such as ProofPoint and Mimecast do not rank highly among the random domain set, Section 5.2.2 shows that such services are increasing in popularity over time.

5.2 Longitudinal Trends

5.2.1 Top Companies. While Figure 5 shows the most recent breakdown for the top companies, we now use the full data set to examine the breakdown for top companies longitudinally over time.

For each of the companies from the Alexa data set in Figure 5, Figure 6 shows the percentage and number of domains whose MX records point to those companies over the four years of our data set. Each curve corresponds to one of the companies. While not dramatic, the trends are all steady increases over time. The top five companies combined are used by 40.1% of MX records in 2017, and the total increases to 49.0% by June 2021. Google dominates the market with Gmail, with Microsoft and Outlook a notable second, and both continue to steadily increase market share. Google increases from 26.2% to 28.5% from 2017–2021, and Microsoft likewise increases from 7.9% to 10.8%.

Notably, ProofPoint and Mimecast are both in the top five and increase their market share over the past four years. These companies are not mail providers, but instead provide an e-mail security service. We explore the rise of such e-mail security services in more detail in Section 5.2.2.

The Self-Hosting curve shows the percentage of domains that host their own SMTP server, rather than using a separate provider. We estimate the number of domains that are self-hosted by looking for domains whose provider ID is the same as its registered domain name. The trend for self-hosting is the opposite that of the top companies. The percentage of domains that self-host steadily decreased over the four years of our data set, falling from 11.7% in 2017 to 7.9% in 2021. Section 5.3 below explores where they switch to in more detail.

Figures 6d and 6g similarly shows the trends over time for the top companies serving the random `.com` and `.gov` data sets in Figure 5. We note that Censys is only intermittently successful in scanning EIG for unknown reasons. Thus, for the longitudinal results we show OVH instead, which is the six largest company in `.com` domains and scanned reliably over time. As with the Alexa data set, the market share of the dominant mail providers (Google and Microsoft) increases over time for .com and .gov domains.[10] The consolidation of Google and Microsoft applies not only to popular domains, but domains across the full distribution. In contrast, though, the market share of hosting providers is steadily decreasing (GoDaddy and UnitedInternet) or flat (OVH) over time. Either there are fewer customers of hosting providers overall, or more of their customers switch away from using the default mail service of the hosting provider.

While the random `.com` data set has over five times the number of Alexa domains (580,537 vs. 93,538), the number of self-hosted domains in `.com` is significantly smaller than that of the Alexa domains (1,836 vs. 7,407 in June 2021) and this number slightly decreased over the last four years. This result matches our expectation that most `.com` domains are small sites hosted with other companies.

[9] Given GoDaddy's dominance, we performed sanity checks to ensure that the domains using GoDaddy are not simply parked domains. Indeed, when we registered a domain or published a website using the registered domain with GoDaddy, GoDaddy did not automatically set up the MX record for the domain. Instead, GoDaddy only configured an MX record for the domain when an e-mail address was created and associated with the domain.

[10] Not for Google in .gov dataset from 2019-12 to 2021-06. A quick sanity check suggests that the majority of the domains moving away from Google were moving to Microsoft.

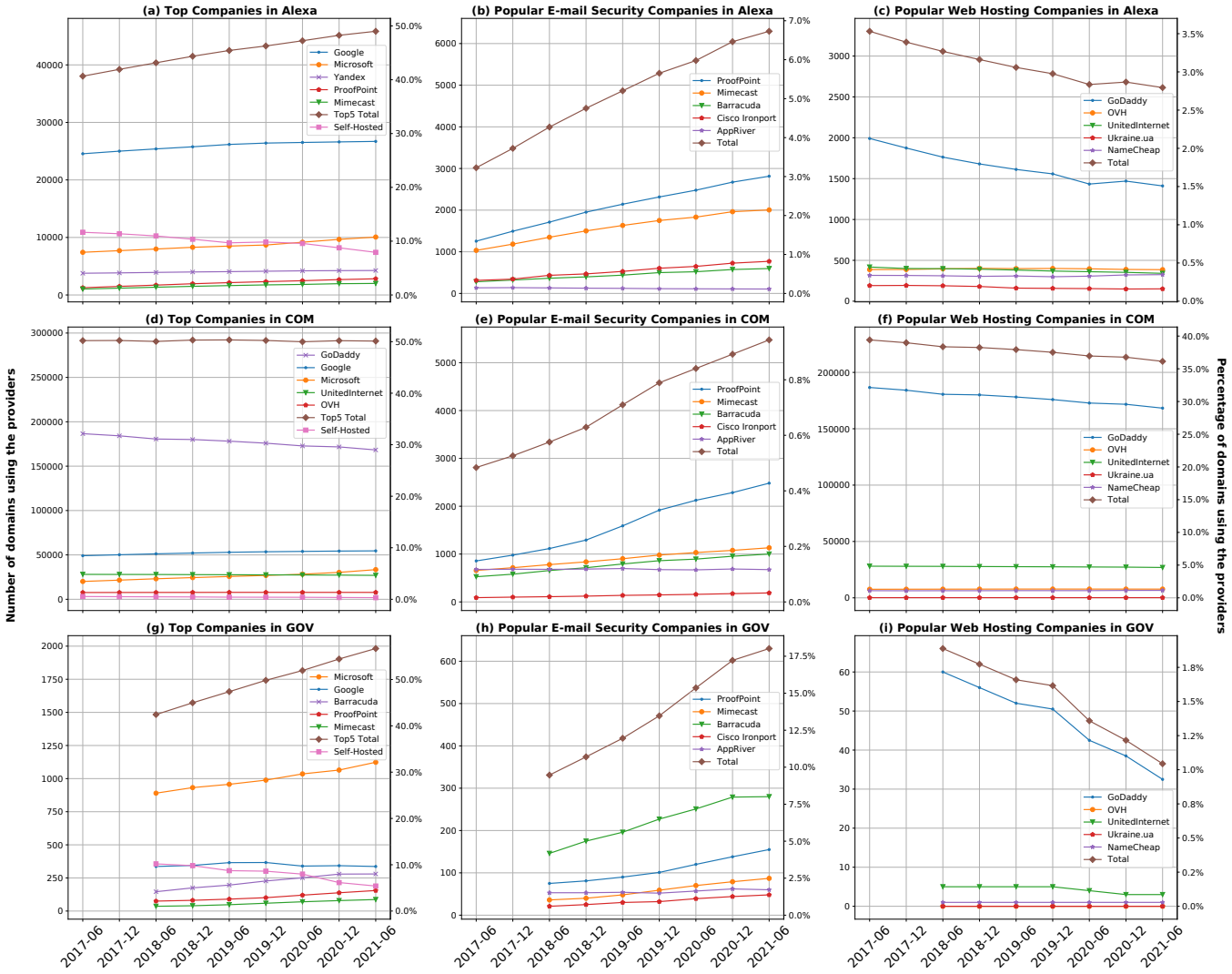

Figure 6: Market share of different types of services from 2017 to 2021. Note that the y-axes of all graphs show the same quantities, but the value ranges are distinct to each graph.

5.2.2 E-mail Security Services. Figure 6a highlighted ProofPoint and Mimecast in the top five companies used by popular domains. These companies provide e-mail security services that can operate as a third-party filter for inbound e-mail delivery, removing the need to purchase and manage a local appliance. Customers use MX records to direct mail agents to deliver mail intended for the customer to the security provider instead, either by explicitly using a provider domain in the MX record (e.g., `ge.com`, which has MX `mx0a-00176a02.pphosted.com`) or by using a customer domain whose A record uses a provider IP address (e.g., `albabotanica.com`, which has MX record `mx1.haingrp.com` that resolves to a ProofPoint IP). The provider then performs spam filtering, phishing detection, URL rewriting, etc., on behalf of the customer, and subsequently forwards the customer's mail to the customer's servers.

The rise of ProofPoint and Mimecast suggests that such companies are becoming a more attractive service option. To explore this point further, in addition to ProofPoint and Mimecast, we manually identified three other popular companies in the third-party e-mail security market across our data sets. Figures 6b, 6e, and 6h show the percentage of MX records that refer to each of five prominent third-party e-mail security companies over time for the Alexa, `.com`, and `.gov` domains, respectively. The results confirm that these services are becoming increasingly attractive for both popular and random domains, as security incidents via e-mail continue to be a major concern.

5.2.3 Web Hosting Companies. Web hosting companies like GoDaddy make it convenient for hosted domains to use company infrastructure for a variety of services including e-mail delivery. As

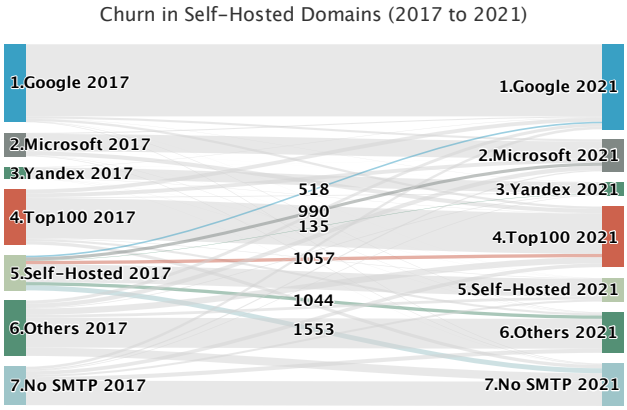

Figure 7: Sankey graph that demonstrates churn in Mail Providers for Alexa domains from 2017 to 2021

we saw in Figure 6d, though, fewer domains over time are taking advantage of hosting company e-mail delivery. We expand upon these results by manually identifying the top five Web hosting companies in both data sets.

Figures 6c, 6f and 6i show the number and percentage of MX records referring to each of these companies in the Alexa, `.com`, and `.gov` data sets, respectively. In both cases the trends are the same. The most popular hosting companies (GoDaddy and UnitedInternet) have fewer domains using their e-mail delivery services over time, and the trend is particularly pronounced among the large sites using popular domains in the Alexa data set. The remaining hosting companies are comparatively flat.

5.3 Churn

Recall that the set of domains we study have valid MX records for the entire duration of our data set. During this time there is churn in the values of the MX records that reflect administrative decisions about mail delivery. Some domains that initially used Google, for instance, may switch to Microsoft during the four years. Similarly, other domains that were self-hosting might switch to Google.

Figure 7 is a Sankey diagram illustrating changes in MX records between the first snapshot in the Alexa data set (June 2017) and the last (June 2021). The diagram groups the domains into various categories: the top three third-party mail hosting providers (Google, Microsoft, Yandex); the remaining top 100 providers; self-hosted domains; all other providers; and the residual set that either had no responding SMTP server or timed out during a Censys scan. For each category, the diagram shows the number of domains using that company that did not change, the number of domains that used the company in 2017 but switched to another by 2021 (outgoing flows), and the number of domains that switched to use the company by 2021 (incoming flows).

While the use of the top companies increased over time, the diagram shows that domains from all of the various categories contributed to this increase (e.g., the incoming flows to Google). From the perspective of domains that switched providers, we in particular highlight the changes that occurred to self-hosted domains between 2017 and 2021. While self-hosted domains switched to providers across all categories, more than a quarter of them changed their mail provider to Google or Microsoft — a quantity larger than the sum of domains that switched to providers ranked in the remaining top 100.

5.4 Mail Provider Preferences by Country

Finally, we explore the existence of national biases in e-mail service provider choice. Since we have no easy mechanical way to classify the national origin of individual gTLD domains (such as those in `.com`) we focused on country code top-level domains (ccTLDs) found in our stable subset of the Alexa top 1M list as a proxy. We consider fifteen ccTLDs, namely: `.br` (Brazil), `.ar` (Argentina), `.uk` (the United Kingdom), `.fr` (France), `.de` (Germany), `.it` (Italy), `.es` (Spain), `.ro` (Romania), `.ca` (Canada), `.au` (Australia), `.ru` (Russia), `.cn` (China), `.jp` (Japan), `.in` (India) and `.sg` (Singapore); thus we assume, for example, that domains under .ru are likely Russian in origin.[11] Among the domains in these ccTLDs, we focus on the use of four popular e-mail service providers: Google, Microsoft, Tencent and Yandex, representing the two dominant e-mail service providers in the US and each of the dominant e-mail service providers in China and Russia, respectively. For each of these four providers, Figure 8 shows the percentage (and absolute number) of domains in each of our ccTLD sets that that make use of the service (June 2021).

There are two clear takeaways. First, Google and Microsoft, the two dominant US-based e-mail service providers, appear to be in wide use by organizations outside the US — particularly across Europe, North America, South America, large parts of Asia and, to a lesser extent, Russia (but not China). For example, 65% of the `.br` domains in our set host mail with Google or Microsoft (significantly exceeding even the baseline market share for our stable Alexa domains of 39.3%). This is of note because under US law (particularly as clarified by the recent Cloud Act's modification to the Stored Communications Act [1]) providers operating in the US can be legally compelled to provide information under their control (including e-mail content) to US law enforcement regardless of the location of the data, or the nationality or residency of the customer using the data. The second clear result is that Yandex and Tencent are comparatively isolated — primarily serving domains only from the ccTLD matching their own country of origin. Indeed, the handful of deviations from these patterns primarily reflect domains for companies whose national origin is not reflected by their choice of ccTLD.[12]

It is an open question the extent to which this discrepancy is driven entirely by market power and infrastructure deployment (i.e., that domain holders do not consider the jurisdictional risk of hosting mail service with a foreign-owned company and are simply picking those who are best able to support their feature, performance, availability and price requirements) or if it also reflects an explicit trust decision (i.e., that European and Brazilian companies are sufficiently comfortable being subject to US jurisdiction that

[11] While there are individual instances that may deviate from this assumption (e.g., `google.ru`), we believe it is predominately true in aggregate (i.e., across the 10,000 plus `.ru` domains we consider, the majority are Russian-operated).

[12] For example, Shein is a Chinese-owned apparel company that operates in the UK under `shein.co.uk`. Similarly, bitrix24 is a Russian-owned Cloud collaboration service that operates under a number of ccTLD aliases including `bitrix24.fr`.

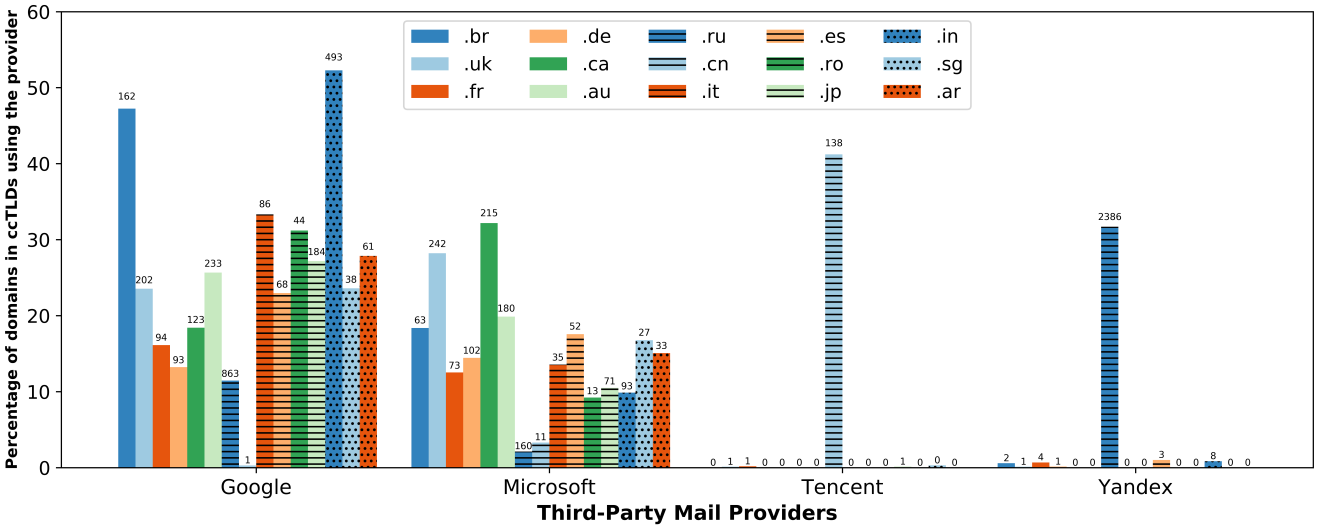

Figure 8: Mail Provider Preferences by Country (ccTLD)

they do not seek local alternatives). Similarly, the dominance of Tencent and Yandex in their local markets may, in part, reflect marketing and infrastructure deployment advantages in their home countries. However, a key role may also be played by state-imposed security review requirements in those countries that US service providers are unwilling or unable to meet. Regardless of reason, the key result is the same: the centralization of e-mail service has been heterogeneous across the globe, with certain providers dominating certain markets. However, it is primarily US-based e-mail service providers who have been effective in attracting foreign customers, despite the additional legal risks posed by this arrangement.

6 CONCLUSION

In this paper, we have presented a methodology for mapping Internet domains to mail service providers. Our methodology combines DNS data with active measurement data to significantly improve accuracy. We have applied this technique to large sets of domains to identify and characterize the current distribution of dominant mail providers. Additionally, our longitudinal study over four years has empirically documented the steady consolidation of Internet e-mail service towards a small number of providers. Finally, we explore the extent to which nationality (and hence legal jurisdiction) plays a role in such mail provisioning decisions.

The analysis code and results for this paper are available at https://github.com/ucsdsysnet/mx_inference.

7 ACKNOWLEDGMENTS

We thank our anonymous shepherd and reviewers for their insightful and constructive suggestions and feedback. We also thank Cindy Moore for her support of the software and hardware infrastructure necessary for this project, and Stewart Grant for his suggestions and feedback. Funding for this work was provided in part by National Science Foundation grants CNS-1629973 and CNS-1705050, the UCSD CSE Postdoctoral Fellows program, the Irwin Mark and Joan Klein Jacobs Chair in Information and Computer Science, the EU H2020 CONCORDIA project (830927), generous support from Google, and operational support from the UCSD Center for Networked Systems. This research used data from OpenINTEL, a project of the University of Twente, SURF, SIDN, and NLnet Labs.

REFERENCES

[1] Stored Communications Act. 2018. 18 USC 2713. Required preservation and disclosure of communications and records.
[2] Mike Afergan and Robert Beverly. 2005. The state of the email address. *ACM SIGCOMM Computer Communication Review* 35, 1 (2005), 29–36.
[3] Alexa. 2021. Top 1M sites. https://toplists.net.in.tum.de/archive/alexa/
[4] Mark Allman. 2018. Comments on DNS Robustness. In *2018 Internet Measurement Conference*. ACM, Boston, MA.
[5] J. Arkko, B. Trammell, M. Nottingham, C. Huitema, M. Thomson, J. Tantsura, and N. ten Oever. 2019. Considerations on Internet Consolidation and the Internet Architecture. https://tools.ietf.org/html/draft-arkko-iab-internet-consolidation-02
[6] CAIDA. 2021. Routeviews Prefix to AS mappings Dataset for IPv4 and IPv6. http://www.caida.org/data/routing/routeviews-prefix2as.xml
[7] Censys. 2020. *Bulk Data.* Censys. https://censys.io/data
[8] Censys. 2021. *Censys Search 2.0 Official Announcement.* https://support.censys.io/hc/en-us/articles/360060941211-Censys-Search-2-0-Official-Announcement
[9] Jianjun Chen, Vern Paxson, and Jian Jiang. 2020. Composition kills: A case study of email sender authentication. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 2183–2199.
[10] Constance Bommelaer de Leusse and Carl Gahnberg. 2019. *The Global Internet Report: Consolidation in the Internet Economy.* Internet Society. https://www.internetsociety.org/blog/2019/02/is-the-internet-shrinking-the-global-internet-report-consolidation-in-the-internet-economy-explores-this-question/
[11] Viktor Dukhovni and Wes Hardaker. 2015. SMTP Security via Opportunistic DNS-Based Authentication of Named Entities (DANE) Transport Layer Security (TLS). RFC 7672. , 34 pages. https://doi.org/10.17487/RFC7672
[12] Zakir Durumeric, David Adrian, Ariana Mirian, Michael Bailey, and J. Alex Halderman. 2015. A Search Engine Backed by Internet-Wide Scanning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA) *(CCS '15)*. ACM, New York, NY, USA, 542–553. https://doi.org/10.1145/2810103.2813703

[13] Zakir Durumeric, David Adrian, Ariana Mirian, James Kasten, Elie Bursztein, Nicolas Lidzborski, Kurt Thomas, Vijay Eranti, Michael Bailey, and J Alex Halderman. 2015. Neither snow nor rain nor MITM... an empirical analysis of email delivery security. In *Proceedings of the 2015 Internet Measurement Conference*. ACM, New York, NY, USA, 27–39.

[14] Ian D Foster, Jon Larson, Max Masich, Alex C Snoeren, Stefan Savage, and Kirill Levchenko. 2015. Security by any other name: On the effectiveness of provider based email security. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, New York, NY, USA, 450–464.

[15] Alex Hern. 2020. Google suffers global outage with Gmail, YouTube and majority of services affected – The Guardian. https://www.theguardian.com/technology/2020/dec/14/google-suffers-worldwide-outage-with-gmail-youtube-and-other-services-down

[16] Paul E. Hoffman. 2002. SMTP Service Extension for Secure SMTP over Transport Layer Security. RFC 3207. , 9 pages. https://doi.org/10.17487/RFC3207

[17] Cecilia Kang and David McCabe. 2020. *Lawmakers, United in Their Ire, Lash Out at Big Tech's Leaders - The New York Times.* The New York Times. https://www.nytimes.com/2020/07/29/technology/big-tech-hearing-apple-amazon-facebook-google.html

[18] Dr. John C. Klensin. 2008. Simple Mail Transfer Protocol. RFC 5321. https://doi.org/10.17487/RFC5321

[19] Dr. John C. Klensin and Randall Gellens. 2011. Message Submission for Mail. RFC 6409. https://doi.org/10.17487/RFC6409

[20] Brian Krebs. 2017. *At Least 30,000 U.S. Organizations Newly Hacked Via Holes in Microsoft's Email Software Krebs on Security.* https://krebsonsecurity.com/2021/03/at-least-30000-u-s-organizations-newly-hacked-via-holes-in-microsofts-email-software/

[21] Public Suffix List. 2021. *Public Suffix List.* https://publicsuffix.org/

[22] D. Liu, S. Hao, and H. Wang. 2016. All Your DNS Records Point to Us: Understanding the Security Threats of Dangling DNS Records. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Vienna, Austria) *(CCS)*. ACM, New York, NY, USA, 1414–1425. https://doi.org/10.1145/2976749.2978387

[23] P. Mockapetris. 1987. Domain Names - Implementation and Specification. RFC 1035. https://rfc-editor.org/rfc/rfc1035.txt

[24] Keith Moore and Chris Newman. 2018. Cleartext Considered Obsolete: Use of Transport Layer Security (TLS) for Email Submission and Access. RFC 8314. https://doi.org/10.17487/RFC8314

[25] Giovane CM Moura, Sebastian Castro, Wes Hardaker, Maarten Wullink, and Cristian Hesselman. 2020. Clouding up the Internet: how centralized is DNS traffic becoming?. In *Proceedings of the ACM Internet Measurement Conference*. ACM, New York, NY, USA, 42–49.

[26] Craig Partridge. 1986. Mail routing and the domain system. RFC 974. https://doi.org/10.17487/RFC0974

[27] Jonathan B. Postel. 1982. Simple Mail Transfer Protocol. RFC 821. https://doi.org/10.17487/RFC0821

[28] Protonmail. 2021. *Verify your custom domain and set MX record.* https://protonmail.com/support/knowledge-base/dns-records/

[29] Joshua Avery Reed and JC Reed. 2020. Potential Email Compromise via Dangling DNS MX.

[30] Walter Rweyemamu, Tobias Lauinger, Christo Wilson, William K. Robertson, and E. Kirda. 2019. Clustering and the Weekend Effect: Recommendations for the Use of Top Domain Lists in Security Research. In *PAM*.

[31] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. 2018. A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. In *Proceedings of the Internet Measurement Conference 2018* (Boston, MA, USA) *(IMC '18)*. Association for Computing Machinery, New York, NY, USA, 478–493. https://doi.org/10.1145/3278532.3278574

[32] Kaiwen Shen, Chuhan Wang, Minglei Guo, Xiaofeng Zheng, Chaoyi Lu, Baojun Liu, Yuxuan Zhao, Shuang Hao, Haixin Duan, Qingfeng Pan, et al. 2020. Weak Links in Authentication Chains: A Large-scale Analysis of Email Sender Spoofing Attacks. *arXiv preprint arXiv:2011.08420* (2020).

[33] Rob Siemborski and Alexey Melnikov. 2007. SMTP Service Extension for Authentication. RFC 4954. https://doi.org/10.17487/RFC4954

[34] Statistica. 2021. Number of sent and received e-mails per day worldwide from 2017 to 2024. https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/

[35] Google Support. 2021. *Set up MX records for Google Workspace email - Google Workspace Admin Help.* https://support.google.com/a/answer/140034?hl=en

[36] Jason Trost. 2020. *Mining DNS MX Records for Fun and Profit – Medium.* https://medium.com/@jason_trost/mining-dns-mx-records-for-fun-and-profit-7a069da9ee2d

[37] Roland van Rijswijk-Deij, Mattijs Jonker, Anna Sperotto, and Aiko Pras. 2015. The Internet of Names: A DNS Big Dataset. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. 91–92. https://doi.org/10.1145/2785956.2789996

[38] Roland van Rijswijk-Deij, Mattijs Jonker, Anna Sperotto, and Aiko Pras. 2016. A high-performance, scalable infrastructure for large-scale active DNS measurements. *IEEE Journal on Selected Areas in Communications* 34, 6 (2016), 1877–1888.

[39] Wikipedia. 2020. *Simple Mail Transfer Protocol.* Wikipedia. https://en.wikipedia.org/wiki/Simple_Mail_Transfer_Protocol

[40] Maya Ziv, Liz Izhikevich, Kimberly Ruth, Katherine Izhikevich, and Zakir Durumeric. 2021. ASdb: A System for Classifying Owners of Autonomous Systems. In *ACM Internet Measurement Conference (IMC'21)*.

Rank	Alexa		COM		GOV	
1	Google	26,697 (28.5%)	GoDaddy	168,287 (29.0%)	Microsoft	1,124 (32.1%)
2	Microsoft	10,072 (10.8%)	Google	54,564 (9.4%)	Google	336 (9.6%)
3	Yandex	4,253 (4.5%)	Microsoft	33,406 (5.8%)	Barracuda	280 (8.0%)
4	ProofPoint	2,815 (3.0%)	UnitedInternet	26,939 (4.6%)	ProofPoint	155 (4.4%)
5	Mimecast	2,005 (2.1%)	EIG	8,714 (1.5%)	Mimecast	87 (2.5%)
6	GoDaddy	1,411 (1.5%)	OVH	7,752 (1.3%)	AppRiver	60 (1.7%)
7	Zoho	1,229 (1.3%)	NameCheap	6,620 (1.1%)	Rackspace	48 (1.4%)
8	Tencent	826 (0.9%)	Tucows	5,517 (1.0%)	Cisco	48 (1.4%)
9	Cisco	771 (0.8%)	Strato	5,025 (0.9%)	GoDaddy	32 (0.9%)
10	Rackspace	752 (0.8%)	Rackspace	4,930 (0.8%)	Sophos	29 (0.8%)
11	Barracuda	598 (0.6%)	Web.com Group	4,200 (0.7%)	Solarwinds	28 (0.8%)
12	Mail.Ru	555 (0.6%)	Aruba	3,842 (0.7%)	IntermediaCloud	24 (0.7%)
13	Beget	420 (0.4%)	Yahoo	3,652 (0.6%)	TrendMicro	22 (0.6%)
14	MessageLabs	412 (0.4%)	SiteGround	3,461 (0.6%)	hhs.gov	21 (0.6%)
15	OVH	386 (0.4%)	Tencent	3,451 (0.6%)	treasury.gov	18 (0.5%)
Total		53,201 (56.9%)		340,362 (58.6%)		2,312 (66.1%)

Table 6: Top 15 companies identified in the three datasets (Jun. 2021)

A TOP 15 COMPANIES IN EACH DATASET

Table 6 lists the top 15 companies identified in the three datasets and their market share: the number and percentage of domains in each dataset using services from these companies.