Veiled in Clouds? Assessing the Prevalence of Cloud Computing in the Email Landscape

Martin Henze*, Mary Peyton Sanford§, Oliver Hohlfeld*
*Communication and Distributed Systems, RWTH Aachen University, Germany
§College of Arts & Sciences, University of Pennsylvania, USA
{henze, hohlfeld}@comsys.rwth-aachen.de, mars@sas.upenn.edu

***Abstract*—The ongoing adoption of cloud-based email services—mainly run by few operators—transforms the largely decentralized email infrastructure into a more centralized one. Yet, little empirical knowledge on this transition and its implications exists. To address this gap, we assess the prevalence and exposure of Internet users to cloud-based email in a measurement study. In a first step, we study the email infrastructure and detect SMTP servers running in the cloud by analyzing all 154 M .com/.net/.org domains for cloud usage. Informed by this infrastructure assessment, we then study the prevalence of cloud-based SMTP services among actual email exchanges. Here, we analyze 31 M exchanged emails, ranging from public email archives to the personal emails of 20 users. Our results show that as of today, 13% to 25% of received emails utilize cloud services and 30% to 70% of this cloud usage is invisible for users.**

I. Introduction

Email is one of the oldest and most prominent Internet services and remains a significant communication medium. To cope with the steady increase in usage, email is currently experiencing an architectural change from a largely *decentralized* medium towards a more *centralized* one [1]. The reason for this shift is the ongoing trend to outsource email services to external cloud operators, either by hosting email servers inside the cloud or by adopting existing cloud email providers. Compared to the classical decentralized email infrastructure in which each organization operates its own email infrastructure, cloud email offers the potential to run email services in a more flexible, scalable, and cost-efficient manner [2]. Email running in the cloud ranges from email servers running on generic cloud infrastructure over cloud-based email security services such as SPAM and DDoS protection to cloud-hosted email services for end users, e.g., Gmail and Outlook.com.

Despite the popularity of cloud services, little is known about the adoption of cloud email services. That is, it remains unknown how much email is processed by cloud services and if their usage is transparent to users. Answering these questions is relevant to then understand the current email architecture and its impact on email users. Regarding infrastructure robustness, the availability of individual email infrastructure can *increase* when hosted in a large cloud, however, outages can now impact much *larger* user bases [3], [4]. Regarding security, concentrating emails at few large providers renders those to valuable attack targets, as exemplified by the breach of 1 billion Yahoo accounts in 2013 [5]. Also, processing email data by large cloud providers can raise jurisdiction and privacy concerns [6]–[9], especially when their usage is not visible to users, i.e., cannot be inferred from the sender or receiver address. To answer these questions, we posit that a deeper understanding of the prevalence of cloud email is required.

The goal of our study is thus to provide a comprehensive assessment of the prevalence of cloud email. We start by understanding the cloud email *infrastructure*, i.e., the set of email servers hosted in cloud environments. We therefore identify all publicly reachable SMTP servers in the entire IPv4 address space and further analyze email servers configured in the complete set of 154 M .com/.net/.org domains. While this first part provides us with an empirical understanding of email infrastructure hosted in the cloud, it does not provide insights on if and how this infrastructure is actually used. To analyze the *user exposure* to the cloud, we analyze actual email exchanges in the second and main part of our study. We thus analyze both *(i)* a number of public email archives providing longitudinal data and *(ii)* a number of personal mailboxes of volunteers in a user study, totaling to more than 31 M exchanged emails. Our contributions are as follows:

1) We provide a methodology to detect the prevalence of cloud-based email services. This methodology uses information publicly provided by cloud and email providers as well as patterns derived from the Internet infrastructure, such as DNS or BGP routing data, to detect cloud usage.
2) To understand the cloud email *infrastructure* (hit when sending email), we identify email servers running on cloud infrastructures in the entire IPv4 address space and uncover cloud usage for all 154 M .com/.net/.org domains. We find that at least 1% of all email servers on the Internet are operated on public cloud infrastructure and more than 50% of all .com/.net/.org domains use cloud-based email services.
3) To understand cloud email *usage* (for received email), we assemble comprehensive datasets of exchanged emails, including mailing list archives and inboxes of 20 users. We analyze more than 31 M emails and show that 13–25% of received emails are exposed to the cloud in 2016. Notably, 30–70% of this exposure is not visible to users.

Dataset release. To foster future research, we release anonymized and aggregated study data and source code [10].

II. The Cloud-based Email Landscape

Cloud-based email promises to host email in a more flexible and cost efficient manner. Attracted by this promise, large corporations have been shifting their on-premise email infrastructure to the cloud. To understand this trend, we start by dissecting the different types of email services that are realized in the cloud today. Here, we define cloud email infrastructures as *large-scale* hosting infrastructures run by *third-parties* and providing services to a *large* number of users.

Before the emergence of cloud-based email services, outsourced email services could be generally differentiated into email *providers* and email *hosters*. When moving to the cloud, the landscape of email services becomes more diverse:

Email providers. Email providers offer typical email services, i.e., a mailbox with the possibility to send and receive emails. Notably, email addresses served by email providers are bound to the domain of the individual provider (e.g., `@aol.com`). Email providers normally offer services for free and finance their services through advertisements.

Email hosters. Email hosters offer basic email services under the domain of the customer, where each customer will have their own domain (e.g., `@example.com`). Typically, email hosters charge for their services, e.g., based on the size and amount of mailboxes. While private users also use hosters, the majority of customers are corporations and businesses. In contrast to email providers, it is not possible to derive the hoster directly from a hosted email address.

Email on cloud infrastructure. Cloud computing enables the transformation of arbitrary services from own on-premise-hardware to virtualized infrastructure running in a cloud data center. This allows the transfer of previously self-hosted email servers to cloud infrastructure. The main motivations are cost reductions, lower maintenance efforts, and higher scalability and elasticity. As moving an email server to a cloud infrastructure still requires the setup and administration of an email server, this approach is mainly pursued by businesses.

Email security. Mail servers are subject to a number of security threats, ranging from SPAM and malware to DDoS attacks, from which cloud-based email security services promise better protection. This is achieved by relaying email via security proxies for both incoming and outgoing emails.

Email marketing. Cloud-based email marketing services enable the sending of massive amounts of highly personalized emails for marketing purposes, e.g., to advertise products, engage with customers, or solicit donations.

Notably, these categories are neither unambiguous nor distinct. For example, larger email *providers* often additionally offer customers to *host* customer domains, e.g., `example.com` (while less known, e.g., Google and Microsoft also offer email hosting). Furthermore, a provider can offer more than one service, e.g., generic cloud infrastructure and email marketing in the case of Amazon. Hence, only an exhaustive picture of the landscape of cloud-based email services ensures a *full* understanding of the impact of cloud computing on email users.

The goal of this paper is to provide an empirical assessment on the prevalence of cloud computing in the current email infrastructure. Shedding light on this question is relevant *(i)* to understand the ongoing change from decentralized to centralized email infrastructures and *(ii)* to better understand potential questions on infrastructure resilience and cloud-related privacy exposures of email. Following the classification derived in this section, we next describe a methodology which we use to assess this prevalence in empirical data.

Cloud Service	**P**	**H**	**I**	**S**	**M**	**Source(s)**
1&1	○	●	○		○	[11]
Adobe					●	[12]
Amazon		○	●		○	[13]
AOL	●					[14]
AppRiver		○		●		[15]
CenturyLink	○	○	●	○		[13]
Cisco		○	○	●		[15]
Comcast	●	○				[14]
Epsilon					●	[12]
Experian					●	[12]
Fujitsu		○	●	○		[13]
GoDaddy		●	○		○	[11]
Google	●	●	●			[11], [13], [14]
IBM (SoftLayer)			●		○	[13]
iCloud	●					[14]
MAX MailProtection				●		[15]
McAfee				●		[15]
Microsoft	●	○	●	○	○	[13], [14]
Mimecast		○		●		[15]
NTT Communications		○	●			[13]
Oracle			○		●	[12]
OVH		●	○			[11]
Proofpoint				●		[15]
Rackspace		○	●			[13]
Salesforce					●	[12]
Strato		●	○			[11]
Symantec				●		[15]
TrendMicro				●		[15]
Virtustream			●			[13]
VMware			●			[13]
Yahoo	●	○				[14]

Table 1. Our selection of 31 major cloud email vendors. We denote the cloud-based email service(s) for which we selected a vendor by ●, while ○ denotes other services offered by this vendor (where it is not a major vendor).

III. Methodology

We start by deriving a methodology that enables us to detect the usage of cloud-based email services based on IP and/or DNS information. It utilizes information publicly provided by cloud and email providers as well as patterns derived from the Internet infrastructure such as DNS or BGP routing data.

A. Representative Set of Cloud Services

To evaluate the prevalence of cloud-based email services, we first derive a representative set of cloud services which we want to classify. To this end, we select the most prominent cloud services for each of the different types of cloud-based email services previously identified in Section II for our analysis. We depict the resulting selected cloud services in Table 1 with filled circles and, in the following, focus on justifying the reasoning behind our selection. Note that one company can offer different types of cloud-based services (e.g., Amazon). In these cases, we merge the different services, which is indicated by multiple circles in the table. Additionally, we depict other services of cloud vendors that we do not classify as one of the most prominent services in their category (e.g., Yahoo's email hosting service) with empty circles in Table 1.

Email providers (P). We base our selection of cloud-based email providers on a survey conducted by Adestra [14]. In our analysis, we include the six most popular email providers which are used by the 1 200 study participants (US residents, all age ranges) as primary email provider. These six providers account for 96% of the participant's primary email providers.

Email hosters (H). For cloud email hosters, we are especially interested in services hosting emails for a large number of domains. We rely on measurements performed by DomainTools on the most popular mail servers based on the number of domains they serve [11]. Based on these results, we include the top five hosters of popular mail servers in our analysis.

Cloud infrastructure (I). Our selection of cloud infrastructure (IaaS) providers builds upon a market analysis performed by Gartner [13]. Based on this analysis, we selected the ten cloud infrastructure services with the highest market share, as those jointly dominate the market [13].

Email security (S). For our selection of cloud-based email security services, we rely on the analysis tools of CloudEmailSecurity.org [15]. We include all eight services that are featured in this survey into our analysis.

Email marketing (M). We base our selection of cloud-based email marketing services on an analysis performed by Forrester [12]. From these results, we derive the five services with the strongest market presence for our analysis.

B. Detection Patterns for Cloud Services

To quantify the prevalence of cloud services among email users, we require patterns enabling this detection. Most notably, this includes IP addresses and DNS names. We next illustrate how these patterns can be derived from public information.

IP addresses. Most, especially larger, *cloud infrastructure* services publish the IP addresses they use, e.g., to allow customers to configure their firewalls [16]. We could retrieve information on used IP addresses for six cloud infrastructures directly from the service. Similarly, all eight cloud-based *email security* services make their IP addresses publicly available, as their customers must restrict their mail servers to only accept incoming emails from these IPs. All cloud-based *email providers* we study publish the IP addresses they use to send emails for two reasons: *(i)* to ease white listing in firewalls or *(ii)* to protect against forging of sender names, e.g., using the Sender Policy Framework [17]. For cloud-based *email hosters*, we were able to directly retrieve IP addresses from two of them. In contrast, we were not able to retrieve information on used IP addresses directly from the service for all five cloud-based *email marketing services*, three email hosters, and four cloud infrastructures. Only in these cases, we looked-up the autonomous system number(s) [18] used by these services and retrieved the associated IP address ranges from the BGP information provided by `ipinfo.io` and `radb.net`. In the end, we were able to retrieve information on the utilized IP addresses for all 31 cloud services.

DNS names. Similar to IP addresses, some cloud-based email services also publish the DNS hostnames they use. However, this fraction of services is significantly smaller. Hence, we require a different approach to obtain information on used hostnames. To this end, we augment the information we were able to retrieve directly from services with information from SenderBase [19]. This enabled us to retrieve the hostnames used by all 31 cloud services under study. In the context of our study, we consider hostnames to be more reliable than IP addresses, as they are more stable over time.

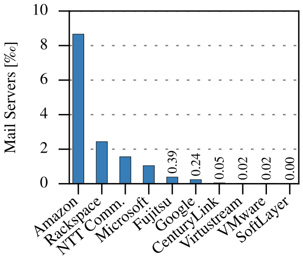


Fig. 1. Cloud usage among publicly reachable SMTP servers (in permil).

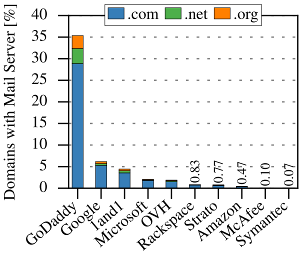


Fig. 2. Cloud usage among .com/.net/.org domains (in percent).

IV. Prevalence of Cloud Email Infrastructures

We begin by assessing the prevalence of cloud services in the global email infrastructure, i.e., the share of email servers hosted in the cloud, hit when *sending* email. To answer this question, we perform two large-scale active measurements.

Email servers running on cloud infrastructure. Our first measurement aims at assessing all publicly reachable mail servers. This study utilizes a trace of a port scan on SMTP port 25/tcp performed on November 19, 2016 covering the entire IPv4 address space and subsequently grabbing SMTP banners [20]. Out of 16.3 M reachable IPs, 6.4 M are classified as valid SMTP servers indicated by a valid `250` status code in the SMTP EHLO banner. We then apply our collection of cloud *infrastructure* IP address ranges (column "I" in Table 1) to identify mail servers hosted by the ten most important cloud infrastructure providers. Our results in Figure 1 show that 1.44% (93 k IPs) of the email servers on the Internet are operated in the networks of these cloud infrastructure providers. Notably, 60.13% (56 k IPs) of these servers are operated on infrastructure provided by Amazon. These results indicate that cloud infrastructure is indeed utilized to provide email services. However, their footprint in terms of IP addresses is rather small and unlike to serve as proxy for usage/popularity.

Cloud usage by .com/.net/.org domains. While the first measurement assesses the cloud usage of all publicly *reachable* mail servers, it does not identify whether the identified IPs are in *use*. That is, while the previously identified IP addresses are publicly reachable mail servers, they do not necessarily have to be configured by any domain as Mail Exchange (MX) to actually receive email. To answer this question, we performed a second measurement querying the MX DNS records of the complete set of 154 M .com/.net/.org domains (DNS zone files provided by Verisign and the Public Interest Registry) on Nov 20, 2016. We obtained MX records for 140 M domains, while 1.2 M were invalid and 12.8 M suffered from authoritative name server errors or timeouts. Out of the obtained 31.9 M distinct MX records, 30.6 M could be resolved to 2.8 M distinct IPs. We remark that the number of detected IPs is lower as compared to the first measurement since *(i)* not the entire DNS space was scanned and *(ii)* not every IP must be configured as MX. The intuition behind this measurement is that any mail server configured as MX in the DNS is intended to receive email.

In contrast to our first measurement, we now have additional DNS information available allowing us to match IP and hostname against the complete set of 31 cloud-based email providers listed in Table 1. In Figure 2, we show the relative share of domains being served by mail servers of one of these 31 cloud-based email services for all 154 M .com/.net/.org. Our results show that, in total, 52.27% of the probed domains use a cloud-based email service. These numbers are largely dominated by GoDaddy, which accounts for 35.36% of the domains served by a small number of servers (34.81 M domains resolving to only 1 732 distinct IP addresses for our vantage point). The dominance of GoDaddy is explained by the fact that it is the world's largest domain registrar, also providing email services to registered domains; whether these are in use is unknown. The other widely used services are the all-purpose services Google and Microsoft, email hosters (1&1, OVH, Strato), cloud infrastructure providers (Rackspace, Amazon), and also email security services (McAfee, Symantec). Surprisingly, the dominance of Amazon in our first IP-based measurement is not reflected in our DNS measurement. Further, email for a large number of domains can be handled by only a small number of public IPs. Subsequent infrastructure (e.g., email redirected to cloud-based security services) is not visible in this analysis since the analyzed MX records denote the *first* server hit when sending mail to a domain. Our DNS analysis shows that an email sent to a random .com/.net/.org address has a more than 50% chance to end up in the cloud.

This first study provides a broad assessment of the *prevalence* of cloud services in the global email infrastructure. It showed that scanning by IP reveals a different cloud provider distribution than probing the DNS. However, it does not provide indications of usage *frequencies* or service popularities, which motivates us to analyze exchanged emails in our second study.

V. Detecting Cloud Usage in Received Emails

To understand the usage *frequencies* of cloud-based email services and hence users' exposure to cloud email, we show how to detect cloud usage in emails solely based on information contained in the email header. Then, we describe how we assemble comprehensive datasets of 31 M emails and discuss limitations as well as privacy considerations of our approach.

A. Dissecting Email Headers to Detect Cloud Usage

To illustrate our approach, we partially depict the header of an email exchanged between a Gmail account and a university account in Listing 1. In the following, we identify the parts of an email header that can be used to detect cloud usage. We differentiate between information that directly allows the detection of cloud usage (green) and information that hints at potential cloud usage based on sender and receiver information (red), which can uncover hidden cloud usage.

```
Received: from mail-qk0-f169.google.com ([209.85.220.169])
   by mx-2.rz.rwth-aachen.de with ESMTP/TLS/AES128-SHA;
   07 Nov 2016 14:37:56 +0100
Received: by mail-qk0-f169.google.com with SMTP id n21so↩
   64861883qka.3 for <[redacted]@comsys.rwth-aachen.de>;
   Mon, 07 Nov 2016 05:37:56 -0800 (PST)
DKIM-Signature: v=1; a=rsa-sha256; c=relaxed/relaxed;
   d=gmail.com; s=20120113; h=mime-version:reply-to:↩
   from:date:message-id:subject:to; bh=0i+V1[...]YJrA=;
   b=bb1p9[...]n0Bw==
X-Google-DKIM-Signature: v=1; a=rsa-sha256; c=relaxed/↩
   relaxed; d=1e100.net; s=20130820; h=x-gm-message-↩
   state:mime-version:reply-to:from:date:message-id:↩
   subject:to; bh=0i+V1[...]YJrA=; b=hTvXs[...]aMA==
X-Gm-Message-State: ABUng[...]DCw==
Received: by [redacted] with HTTP; Mon, 7 Nov 2016 ↩
   05:37:54 -0800 (PST)
From: [redacted] <[redacted]@gmail.com>
Date: Mon, 7 Nov 2016 08:37:54 -0500
Message-ID: <CADLj[...]2b+9g@mail.gmail.com>
Subject: [redacted]
To: [redacted] <[redacted]@comsys.rwth-aachen.de>
```

Listing 1. Information contained in email headers offers different opportunities to detect exposure to cloud-based email services.

Received lines: The main purpose of received lines is to aid debugging of email failures [21]. To this end, each email server that receives an email (either for forwarding or for final delivery) has to prepend a received line to the email's header [21]. While the exact format of received lines can deviate from the specification [21], they typically contain the hostname and IP address of the current and the previous email server in the delivery chain as well as a timestamp (cf. Listing 1). The complete set of received lines in an email enables us to derive the complete path of email servers that this email traversed. Hence, we can use the set of corresponding IP addresses and hostnames to detect usage of cloud services. Notably, the standard forbids removing or modifying any received lines from an email header [21]. While email servers can violate this standard, corresponding countermeasures are widely deployed today [22]. In this work, we hence assume that the information present in email headers has not been tampered with.

Custom header fields: Besides explicitly standardized header fields, email clients and servers can also include arbitrary customized header fields [23]. Traditionally, these custom header fields are prefixed with "X-" (cf. Listing 1). They are utilized especially by larger email services, which enables us to detect these services. Furthermore, previously unstandardized header fields, e.g., DKIM signatures [24], emerged into now standardized and more widely deployed header fields. Such header fields, that are nowadays used by more than one email service, are still valuable as they often contain information on the email service (cf. Listing 1). To identify custom header fields, we manually clustered the header fields present in a subset of our datasets and distilled those header fields unique to a cloud service. As a result, we were able to retrieve custom header fields for seven cloud services, mostly email providers.

Sender and receiver information: Each email contains information on the sender and the receiver(s) of this email (cf. Listing 1). While this information is not reliable (it can easily be spoofed), it provides a visible indicator of cloud usage. For example, if a user receives an email from an @gmail.com address, she would not be surprised that this email has been processed by Google mail servers. Although we cannot use sender and receiver information to detect cloud usage, we can use it to decide whether detected cloud usage is hidden for the user. Sender and receiver information are especially relevant for email providers, as the provider is visible in the email address. We manually identified the hostnames of email addresses used by all six email providers in our study. Additionally, we use hostnames collected for all 31 cloud services to detect associated senders and receivers. This approach is very optimistic and can lead to false positives. As we use senders and receivers merely to preclude hidden cloud usage, false positives will only lower the fraction of hidden cloud usage. This still gives us a lower bound for the prevalence of hidden usage of cloud-based email services.

B. Datasets

A study on the prevalence of cloud computing among email users requires the analysis of a sufficiently large set of exchanged emails. We therefore base our analysis on a set of 31.85 M emails exchanged between 1995 and 2016, obtained from mailing list archives, SPAM traps, WikiLeaks, and 20 volunteer users—representing a diverse user base. Still, the cloud usage derived in this study is dependent on the user base and geographic location. Since these data sets partly begin *before* the emergence of cloud computing, we can observe its widespread adoption. We remark that detection pattern can change over time (see Sect. V-C). To ensure the absence of false positives, we analyzed a random subset. For our analysis, we only consider standard conform emails [21], i.e., containing message ID and date header fields. Further, we only consider emails with at least one received line. By doing so, we eliminate emails only consisting of error messages. We summarize our datasets in Table 2 (number of emails obtained after cleanup). **Mailing lists.** We downloaded the public mailing list archives from the Apache Software Foundation, Dovecot, FreeBSD, the Internet Engineering Task Force, and openSUSE. These emails contain discussions and announcements regarding open source development and standardization efforts. **WikiLeaks.** This dataset contains formerly private emails that have been made public by WikiLeaks [25]. These emails originate from the Turkish Justice and Development Party (AKP), the US Democratic National Committee (DNC), and Hillary Clinton's campaign chair John Podesta. **SPAM.** In this dataset, we combine emails collected by various SPAM traps since 2007 [26], [27]. **Users.** We recruited 20 volunteers (mostly with a technical background) from Germany who agreed to run our analysis tool on their personal and (partly) professional emails. Besides communication with other people, these emails also contain a large number of automatically generated emails such as newsletters, commit messages, and SPAM.

Parts of our datasets are inherently biased to contain significant cloud usage when the *recipient* of the emails uses a cloud-based email service herself. We cope with this bias by ignoring those cloud services that have been used to receive the e-mails under study. Hence, we ignore AppRiver for *WikiLeaks DNC*, Google for *WikiLeaks Podesta*, and 1&1 for *SPAM*. Furthermore, we blacklist Google for *SPAM*, as we observed massive amounts of faked received lines for Google in this dataset. Finally, we asked our volunteers to blacklist those email services that where used to receive their emails.

Dataset	Period	Emails	Public	Comments
Mailing lists	01/95–09/16	22 930 801	●	—
Apache	*02/95–09/16*	*15 516 752*	●	*1507 open source lists*
Dovecot	*07/02–09/16*	*115 007*	●	*3 open source lists*
FreeBSD	*01/95–09/16*	*3 654 624*	●	*160 open source lists*
IETF	*01/95–09/16*	*2 043 606*	●	*949 standardization lists*
openSUSE	*05/06–09/16*	*1 600 812*	●	*85 open source lists*
WikiLeaks	09/07–07/16	254 476	●	—
AKP	*11/09–07/16*	*231 388*	●	*Internal emails*
DNC	*01/15–05/16*	*15 848*	●	*Internal emails*
Podesta	*09/07–03/16*	*7 240*	●	*Internal emails*
SPAM	02/07–09/16	7 788 560		non-public SPAM traps
Users	10/01–09/16	873 587		emails of 20 users

Table 2. We assembled different datasets of emails ranging from mailing lists to private emails of users, in total accumulating to 31.85 M emails.

C. Limitations

Our methodology to quantifying the prevalence of cloud computing by matching patterns in headers of received emails is limited in three ways. First, our approach is inherently restricted to *incoming emails*. As we rely on header information inserted by cloud services, our method cannot be used to detect usage of cloud services in outgoing emails. To partly account for this, our active measurements (Section IV) uncover the cloud usage when sending emails, e.g., the mail servers processing the .com/.net/.org domains. However, emails typically traverse multiple servers and from the outside we can observe only the first hop. Without cooperation of the receiver of an email, this limitation likely cannot be solved. Second, *detection patterns can change* over time. Hence, the patterns we derived to detect cloud usage might not be accurate for the past. However, we observe that information on hostnames and custom header fields remain relatively constant over time. With respect to IP addresses used by cloud services, we observed in past years (for big infrastructure providers), that their IP address ranges constantly grow and previously used IP addresses do not get abandoned. To account for this limitation, we randomly sampled a small subset of very old emails from our mailing lists dataset to verify that no false positives occurred. *Finally*, we limit ourselves to *31 representative cloud services*. Enlarging this set is technically possible but requires manual curation of cloud vendors IPs and hostnames. We remark that service popularity can change between different regions (geographic bias in data and rules). To verify that our selection of services is representative, we manually checked undetected hostnames, custom header fields, and sender names for our mailing lists dataset to ensure that we did not miss any widely used service.

D. Ethical and Privacy Considerations

As we operate on potentially sensitive data of individual users, all our experiments where designed following the basic

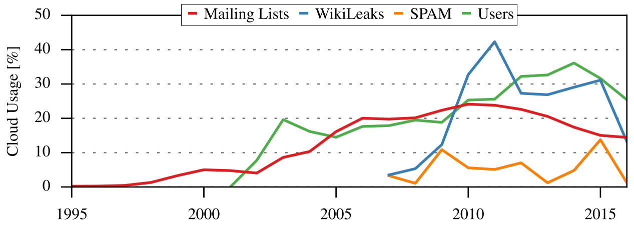


Fig. 3. In the past, the cloud usage of emails steadily increased to 20–40%, but now shows a remittent tendency with a cloud usage of 15–25% in 2016.

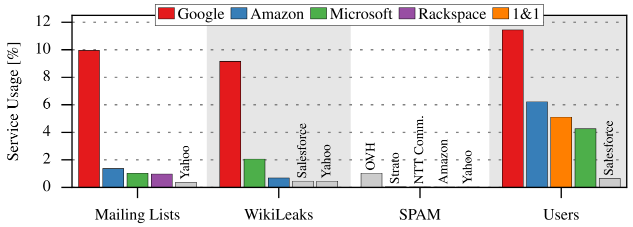


Fig. 4. The cloud services with the highest usage in 2016 vary between our datasets, but Google clearly plays an important role.

principles of ethical research. The goal of this work is to understand the prevalence of cloud computing among email users. Having this goal in mind, we designed all experiments such that the risk of (inadvertently) harming the privacy of users is minimized. Hence, we excluded exact sender identifiers and the actual content of emails from our analysis. This way, we unlink potentially sensitive information from identities. Furthermore, we aggregate all our results in a way that prevents drawing conclusions about individuals who contributed data.

VI. Impact of Cloud Computing on Email Users

We begin by studying the cloud usage for *individual* emails.

The rise of cloud-based email services. The first question is how large the usage of cloud services is and how it has developed over time. To study this development, we show the sum of emails processed by detected cloud services per year for each data set in Figure 3. We consider an email to be processed by a cloud service if it was processed by an SMTP server [21] of a detected cloud service as listed in Table 1.

When looking at *mailing lists* (by far the largest dataset in our analysis with nearly 23 M emails), we observe that the rise of cloud-based email services first gains traction in the late 1990s with the email offers of AOL, Microsoft, and Apple. This rise increases in 2004 when Google's Gmail was launched, peaking at 24.12% in 2010. Since then, we observe a decrease of cloud usage, leading to a usage of cloud email services of 14.45% in 2016. For the emails of our volunteer *users*, we observe a quite similar trend until 2010, with early-adopters of cloud email leading to a first peak of 19.63% cloud usage already in 2003. In contrast to the mailing lists dataset, cloud usage of our volunteers continues to grow beyond 2010 to 36.11% in 2014 before surprisingly dropping to 25.41% in 2016. While our data does not allow us to derive a reason for this observation, one possible explanation might be that persons involved in open source development and standardization efforts could become more privacy-sensitive and avoid large email services. The *WikiLeaks* dataset shows a similar, yet more extreme trend with a peak of 42.31% cloud usage in 2011. Here, the sudden decrease in cloud usage (to 13.21% in 2016) can mostly be attributed to a decreasing cloud use of AKP emails in 2016.

For *SPAM* mails, we assumed a lower fraction of cloud usage, as cloud-based email service providers have a strong interest in spam prevention. Indeed, we see little impact of cloud computing on SPAM emails, far less than in our other datasets. The spike for 2015 corresponds to a significant increase in SPAM emails (apparently) received from the hoster OVH. Overall, there is no significant impact of cloud computing on SPAM emails with a cloud usage of only 1.22% in 2016.

Cloud services with highest usage. Considering the general trend of a rising cloud email usage, an immediate question is which services contribute most to this cloud usage. To study this question, we consider the usage of *individual* cloud services in each of our datasets in the year 2016. Figure 4 depicts the fraction of emails exposed to a specific service for each dataset.

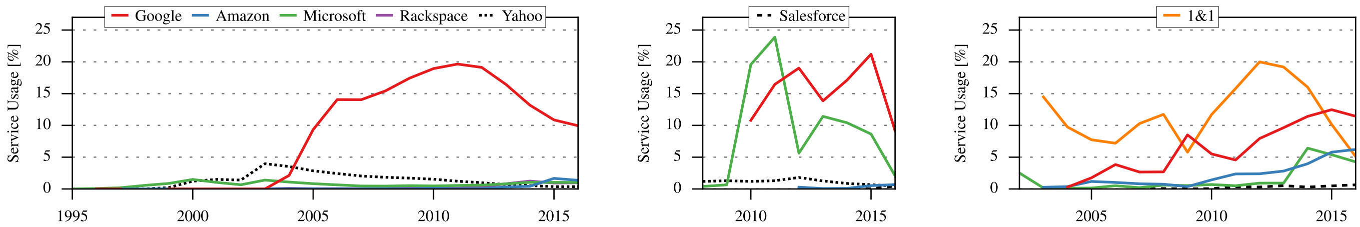
For the *mailing lists* dataset, we can clearly identify Google as the service with the highest cloud usage: 9.95% of mailing list emails were processed by Google in 2016. Amazon, Microsoft, Rackspace, and Yahoo already show a notable distance with a usage between 0.37% and 1.37%. Similar observations can be made for *WikiLeaks*, with Google (9.16%) clearly leading in front of Microsoft (2.06%) followed by Amazon, Salesforce, and Yahoo, each well below 1%. Given the overall low cloud usage for the *SPAM* dataset in 2016, the results for the individual services provide limited insight. The top-infrastructure used for spam is OVH (1.03%).

For the *users* emails, we again observe the highest cloud usage for Google (11.44%), this time followed closer by Amazon (6.22%), 1&1 (5.11%), and Microsoft (4.26%). The comparable high usage of 1&1 likely corresponds to our users being from Germany, where 1&1 is one of the leading email hosters and email providers. The higher usage of Amazon services can partly be attributed to emails sent by Amazon's Simple Email Service, e.g., newsletters and other marketing emails, which naturally are more relevant for the users dataset.

These results highlight that the usage of individual cloud services is dataset dependent. To gain a clear picture of the prevalence of cloud computing in email, we thus consider all four datasets to derive the most used services for 2016. This gives us Google, Amazon, Microsoft, Rackspace, and 1&1 as the five services with the highest fraction of emails exposed to them across all datasets.

Trending email services. We next study the question of how cloud services with the highest usage in 2016 emerged over time, shown in Figure 5 for the mailing lists, Wikileaks, and users datasets (we omit SPAM given its low cloud usage).

Cloud usage of the mailing lists dataset (Figure 5a) is nearly exclusively dominated by Google, surpassing Yahoo quickly after Gmail's launch in 2004. For the Wikileaks dataset (Figure 5b), Google and Microsoft are on par, each accounting



(a) Mailing lists

(b) Wikileaks

(c) Users

Fig. 5. The usage of individual cloud services differs for the mailing lists, Wikileaks, and users dataset, but a small number of services clearly dominate.

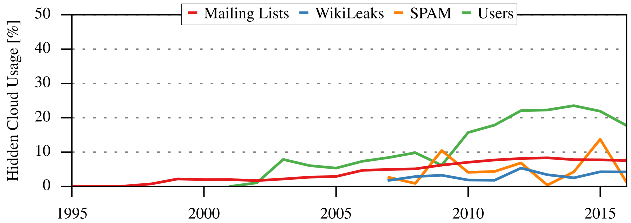


Fig. 6. Emails with hidden cloud usage among the *total* set of emails. Hidden usage of cloud services follows a similar trend as cloud usage in general.

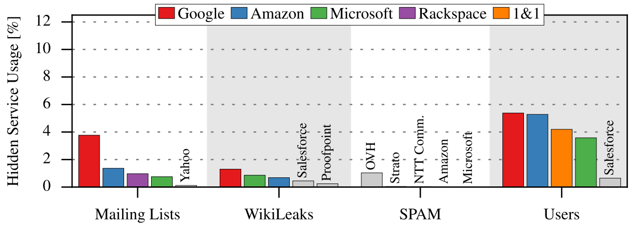


Fig. 7. Hidden cloud usage across our four datasets mainly results from email hosters and cloud infrastructure offers as well as hybrid offers.

for more than 20% of the email traffic. While the users dataset (Figure 5c) initially is dominated by 1&1 (see above), we observe a steady increase for emails from Google and Amazon.

To conclude, we observe a surprisingly high usage of cloud computing for email exchanges. Between 13.21% (WikiLeaks) and 25.41% (users) of received emails are processed by at least one cloud service in 2016. It is important to remark that we only account for cloud services that are *not* utilized by the user herself (e.g., to host her emails), but for cloud services hit on the way to the recipient. Depending on the dataset, between 9.16% and 11.44% of received emails are processed by a *single* cloud service in 2016 (most notably Google, Amazon, and Microsoft). Hence, the (partial) outage of these services can impact large user bases. These services also learn about a large fraction of the users' (potentially sensitive) email communication, stimulating privacy concerns.

VII. Hidden Usage of Cloud-based Services

The usage of cloud email services on the way to the recipient can be *hidden* to the (non-expert) user. We define the usage of a cloud service as hidden if this cloud service is not obviously used as the email provider of sender or any recipient, i.e., the cloud service cannot be inferred from email addresses in the sender or recipient fields. For example, if any sender or recipient address ends with `@gmail.com`, the usage of *all* servers attributed to Google is *not* hidden. Hidden usage of cloud resources can raise privacy concerns, e.g., when communication (meta) data should not be exposed to a third-party operator [28]. We therefore aim at understanding the extent to which hidden exposure of emails to cloud services happens and to which services we can attribute the most hidden cloud usage.

General trend of hidden cloud exposure. Again, we first study the general evolution of hidden cloud exposure over time for our different datasets in Figure 6. For each dataset, we plot the overall fraction of hidden cloud usage among the entire set of emails per year. We define cloud usage to be hidden if at least one of the utilized cloud services is neither detectable from the sender field nor from any of the recipient fields.

The hidden cloud exposure for the *mailing lists* dataset shows a steady increase, similar to the overall increase in cloud exposure. In 2016, we observe that 7.53% of all emails in our datasets use cloud services hidden to the user (see Fig. 6), which amounts to 52% of all emails using the cloud (i.e., 14.45% of all emails in our datasets, see Fig. 3). Similarly, we observe that 70% of cloud usage remains oblivious to users for the *users* dataset (i.e., 17.72% emails with hidden usage vs. 25.41% cloud emails in 2016). For *WikiLeaks* emails, we observe a lower hidden cloud usage than for the mailing lists and users datasets. Given the overall high fraction of cloud usage for the WikiLeaks emails, these results indicate that the cloud usage during this period can mostly be attributed to emails that originate from cloud-based email *providers*. Here, we observe that 32% of the cloud usage cannot be observed by users (i.e., 4.21% vs. 13.21% in 2016). In contrast, for *SPAM* emails cloud usage happens nearly exclusively hidden, as seen by the nearly identical curves in Figures 3 and 6. This suggests that the cloud portion of SPAM does not originate from (potentially hacked) cloud email accounts but instead from email hosters or cloud infrastructure.

Cloud services with highest hidden usage. Knowing that a large part of the cloud exposure is hidden to the user, the immediate question is to which services the most hidden cloud usage can be attributed. We thus study the hidden usage of individual cloud services in 2016 for each dataset in Figure 7. Again, we observe that it is important to cover different sources of emails, as the results for hidden cloud usage of specific services vary across the datasets. Nevertheless, we can derive what types of cloud email services (cf. Section II) account for hidden cloud usage: *(i)* email hosters (e.g., 1&1 with 4.20% in the users dataset) and *(ii)* cloud infrastructure (e.g., Amazon with 5.28% in the users dataset). Furthermore, hybrid services such as Google and Microsoft that offer email hosting and cloud infrastructure have a significant impact on hidden cloud usage. As expected, we do not observe hidden usage of cloud-based email *providers* (e.g., AOL or Comcast).

In summary, we observe that non-expert users remain oblivious to the hidden cloud usage of 30% to 70% of emails exposed to the cloud. This hidden usage is predominantly caused by email hosters and cloud infrastructure. Some hidden usage could be uncovered by email software by analyzing DNS MX records of the recipient domains. Other cloud exposure (e.g., use of security services) cannot be detected by the sender.

VIII. Related Work

Understanding email traffic. Ramachandran et al. [29] study the properties of SPAM emails based on network-level observations, e.g., IP address ranges used to send SPAM emails. They find that network-level characteristics can indeed be used to tell SPAM and legitimate email apart. Motivated by these findings, Hao et al. [30] propose a reputation engine for emails based on network-level characteristics. They report that their fully automated approach achieves comparable SPAM classification rates to hand-labeled blacklists. From a different line of research, Schatzmann et al. [31] strive to classify *web*mail traffic to gain a comprehensive view of the Internet email infrastructure. To this end, they develop flow-level techniques operating solely on passive network measurements to reliably tell webmail traffic and other HTTPS traffic apart.

Understanding cloud traffic. Bermudez et al. [32] utilize DNS responses to detect cloud services based on network traffic. Their results reveal that the vast majority of traffic generated by Amazon Web Services originates from a single data center. Similarly, Drago et al. [33] study the properties of personal cloud storage services. They perform passive measurements and distinguish between different cloud storage services based on information contained in DNS and TLS network packets. He et al. [34] present a measurement study to understand the deployment of web service on cloud infrastructure. They rely on DNS probing to identify which popular web services use Amazon's and Microsoft's cloud offer and conclude that 4% of the most popular web services run on this infrastructure. Likewise, Fiadino et al. [35] discuss an analysis of WhatsApp based on passive measurements from the core of a cellular network and geo-distributed active measurements. They find that WhatsApp is hosted by a single cloud service in the US. Finally, Henze et al. [16] propose the analysis of smartphone network traffic to uncover apps' cloud service usage.

Understanding cloud-based email. Willett et al. [36] performed a survey to quantify the adoption of cloud-based email services at higher education institutions in South Africa. They observe that the majority of institutions are using cloud-based email services or plan to. A study performed by Hsu et al. [37] targets the cloud adoption of the largest Taiwanese companies. Their results indicate that 44% of the companies have migrated their email system to the cloud or plan to do so within one year. Gartner [38] analyzed the DNS records of nearly 40 000 companies to check for Google or Microsoft usage as an email hoster. They find that about 13% of the studied companies use one of the two email providers. Hotmail traces were analyzed to identify dynamic IP addresses for spam filtering [39]. Finally, van Rijswijk-Deij et al. [40] analyze the growth of cloud-based email services based on DNS records for the .com zone. They observe that the largest (by number of domains) cloud-based email *providers* are Google, Microsoft, and Yahoo.

While these works highlight the importance of understanding the impact of cloud computing on email users, an empirical evaluation of both the cloud usage among the email infrastructure and the users' exposure to cloud services, is missing.

IX. Key Observations and Conclusion

The goal of this paper is to provide a first comprehensive understanding on the prevalence of cloud email. This is a relevant view since the ongoing transformation of the email architecture from a largely decentralized one towards a more centralized one can have consequences for privacy, security, and robustness aspects. To tackle this problem we design a methodology based on public information to detect cloud-based email services in two studies. Our first study analyzes email *infrastructures* hosted in the cloud, i.e., servers hit when sending email. We analyzed all publicly reachable mail servers obtained by scans of the entire IPv4 address space and by querying the complete set of 154 M .com/.net/.org domains. Our second study then focuses on understanding the *user* exposure to these infrastructures when exchanging email by analyzing more than 31 M emails. We make the following key observations:

Exchanged emails tell a different story than infrastructure measurements. With regard to measurement studies, we show the difference of three views on email: *(i)* size of the public-facing infrastructure (i.e., number of SMTP IPs hosted in cloud infrastructures), *(ii)* email servers configured for domains (i.e., DNS MX records), and *(iii)* exchanged emails. All three perspectives tell a different story. For example, a large number of domains can point to few mail server IPs hosted in a cloud. Since configuring a cloud mail infrastructure in a domain name cannot tell how often email is sent to this domain, the cloud usage derived from exchanged email again differs. All three perspectives can, however, provide interesting insights: infrastructure studies yield insights into the (ongoing) adoption of cloud mail services, while the analysis of exchanged emails yield insights into the cloud exposure experienced by users. Thus, all these perspectives are relevant for future studies.

Email users are exposed to the cloud. We observe that users' emails are processed by the cloud: between 13.21% (WikiLeaks) to 25.41% (users) of all emails received in 2016 were processed by cloud services. Regarding the email infrastructure, our DNS analysis shows that email sent to a random .com/.net/.org address has a more than 50% chance to end up in the cloud. While the concrete services and their exposure level varies between the datasets (and users), we observe a concentration of few large infrastructures. None of these infrastructures is large enough to dominate, e.g., to easily roll out new features and force adoption. This concentration can, however, motivate the use of encryption. It thus opens questions on privacy and security implications of email becoming more centralized, i.e., single providers having access to large fractions of the overall exchanged emails. It also opens the question of whether or not email becomes more robust when processed in the cloud (e.g., increasing availability or eased spam filtering). **Usage of cloud-based email services happens unobservable.** Surprisingly, for 30%–70% of the emails that are processed by the cloud, this cloud usage is unobservable for (non-expert) users. That is, it cannot be inferred from email addresses, e.g., `@gmail.com`. One reason for hidden cloud exposure is the ability to have a domains mail exchange record configured to a cloud mail server (e.g., email for a state-owned university can be managed by a third-party cloud operator). This hidden exposure can in principle be detected by implementing our methodology in email programs, thereby raising users' awareness for hidden cloud usage. Another reason for hidden cloud exposure is that email can be transparently forwarded to cloud services, e.g., to security cloud solutions for virus checking by the operator or to private cloud mail accounts by the receiver. Such forwarding cannot be inferred by the sender and can only be detected by the receiver with email header analysis. One implication is the question of whether email routing and processing should or can be made controllable. For example, by granting security services only access to selected parts of an email (e.g., to perform virus checking on executable attachments) security and privacy concerns could by moderated.

Our work to understand the prevalence of cloud email provides the foundation for such highly necessary countermeasures.

Acknowledgments

We thank the anonymous reviewers and our shepherd Cristian Lumezanu for their valuable comments and suggestions to improve this manuscript. This work has received funding from the German Federal Ministry of Education and Research (BMBF) under project funding reference no. 16KIS0351 (TRINICS) and the European Union's Horizon 2020 research and innovation program 2014-2018 under grant agreement no. 644866 (SSICLOPS). This manuscript reflects only the authors' views and the funding agencies are not responsible for any use that may be made of the information it contains.

References

[1] S. Marston, Z. Li, S. Bandyopadhyay, J. Zhang, and A. Ghalsasi, "Cloud computing – The business perspective," *Decision Support Systems*, 2011.

[2] K. Barlow and J. Lane, "Like Technology from an Advanced Alien Culture: Google Apps for Education at ASU," in *ACM SIGUCCS*, 2007.

[3] C. Donnelly, "Microsoft and Google cloud users suffer service outages," Computerweekly, http://www.computerweekly.com/news/450304333/Microsoft-and-Google-cloud-users-suffer-service-outages, 2016.

[4] C. Albanesius, "Google: Software Bug Caused Gmail Deletions," PCMag, http://www.pcmag.com/article2/0,2817,2381168,00.asp, 2011.

[5] S. Thielman, "Yahoo hack: 1bn accounts compromised by biggest data breach in history," The Guardian, https://gu.com/technology/2016/dec/14/yahoo-hack-security-of-one-billion-accounts-breached, 2016.

[6] M. Henze, J. Hiller, O. Hohlfeld, and K. Wehrle, "Moving Privacy-Sensitive Services from Public Clouds to Decentralized Private Clouds," in *CLaw Workshop*, 2016.

[7] I. Ion, N. Sachdeva, P. Kumaraguru, and S. Čapkun, "Home is safer than the cloud!: Privacy concerns for consumer cloud storage," in *SOUPS*, 2011.

[8] X. Fan, E. Katz-Bassett, and J. Heidemann, "Assessing Affinity Between Users and CDN Sites," in *TMA*, 2015.

[9] M. Henze, R. Matzutt, J. Hiller, E. Mühmer, J. H. Ziegeldorf, J. van der Giet, and K. Wehrle, "Practical data compliance for cloud storage," in *IEEE IC2E*, 2017.

[10] RWTH Aachen University, "MailAnalyzer – Source code and study dataset," https://github.com/COMSYS/MailAnalyzer.

[11] DomainTools, "Statistics About Mail Servers," http://research.domaintools.com/statistics/mailservers/.

[12] R. McAdams, "The Forrester Wave^{TM}: Email Marketing Service Providers, Q3 2016," Forrester Research, Inc., 2016.

[13] L. Leong, G. Petri, B. Gill, and M. Dorosh, "Magic Quadrant for Cloud Infrastructure as a Service, Worldwide," Gartner Rep. G00278620, 2016.

[14] Adestra, "2016 Adestra Consumer Adoption & Usage Study," 2016.

[15] "Cloud Email Security Comparison," http://cloudemailsecurity.org/.

[16] M. Henze, D. Kerpen, J. Hiller, M. Eggert, D. Hellmanns, E. Mühmer, O. Renuli, H. Maier, C. Stüble, R. Häußling, and K. Wehrle, "Towards Transparent Information on Individual Cloud Service Usage," in *IEEE CloudCom*, 2016.

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

[18] J. Hawkinson and T. Bates, "Guidelines for creation, selection, and registration of an Autonomous System (AS)," IETF RFC 1930, 1996.

[19] Cisco Systems, Inc., "SenderBase," http://www.senderbase.org/.

[20] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman, "A search engine backed by internet-wide scanning," in *ACM CCS*, 2015.

[21] J. Klensin, "Simple Mail Transfer Protocol," IETF RFC 5321, 2008.

[22] E. Bursztein and V. Eranti, "Internet-wide efforts to fight email phishing are working," Google Security Blog, https://security.googleblog.com/2013/12/internet-wide-efforts-to-fight-email.html, 2013.

[23] P. Resnick, "Internet Message Format," IETF RFC 2822, 2001.

[24] D. Crocker, T. Hansen, and M. Kucherawy, "DomainKeys Identified Mail (DKIM) Signatures," IETF RFC 6376, 2011.

[25] "WikiLeaks," http://wikileaks.org/.

[26] O. Hohlfeld, T. Graf, and F. Ciucu, "Longtime Behavior of Harvesting Spam Bots," in *ACM IMC*, 2012.

[27] G. Stringhini, O. Hohlfeld, C. Kruegel, and G. Vigna, "The harvester, the botmaster, and the spammer: On the relations between the different actors in the spam landscape," in *ACM AsiaCCS*, 2014.

[28] Y.-A. de Montjoye, E. Shmueli, S. S. Wang, and A. S. Pentland, "openPDS: Protecting the Privacy of Metadata through SafeAnswers," *PLOS ONE*, vol. 9, no. 7, 2014.

[29] A. Ramachandran and N. Feamster, "Understanding the Network-level Behavior of Spammers," in *ACM SIGCOMM*, 2006.

[30] S. Hao, N. A. Syed, N. Feamster, A. G. Gray, and S. Krasser, "Detecting Spammers with SNARE: Spatio-temporal Network-level Automatic Reputation Engine," in *USENIX Sec*, 2009.

[31] D. Schatzmann, W. Mühlbauer, T. Spyropoulos, and X. Dimitropoulos, "Digging into HTTPS: Flow-based Classification of Webmail Traffic," in *ACM IMC*, 2010.

[32] I. Bermudez, S. Traverso, M. Mellia, and M. Munafò, "Exploring the Cloud from Passive Measurements: the Amazon AWS Case," in *IEEE INFOCOM*, 2013.

[33] I. Drago, M. Mellia, M. M. Munafo, A. Sperotto, R. Sadre, and A. Pras, "Inside Dropbox: Understanding Personal Cloud Storage Services," in *ACM IMC*, 2012.

[34] K. He, A. Fisher, L. Wang, A. Gember, A. Akella, and T. Ristenpart, "Next Stop, the Cloud: Understanding Modern Web Service Deployment in EC2 and Azure," in *ACM IMC*, 2013.

[35] P. Fiadino, M. Schiavone, and P. Casas, "Vivisecting WhatsApp in cellular networks: Servers, flows, and quality of experience," in *TMA*, 2015.

[36] M. Willett and R. Von Solms, "Cloud-based Email Adoption at Higher Education Institutions in South Africa," *JITIM*, 2014.

[37] P.-F. Hsu, S. Ray, and Y.-Y. Li-Hsieh, "Examining cloud computing adoption intention, pricing mechanism, and deployment model," *IJIM*, 2014.

[38] N. Drakos and J. Mann, "Survey Analysis: Microsoft Dominates Cloud Email in Large Public Companies but Shares the Rest With Google," Gartner Rep. G00292300, 2016.

[39] Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt, and T. Wobber, "How dynamic are ip addresses," in *ACM SIGCOMM*, 2007.

[40] R. van Rijswijk-Deij, M. Jonker, A. Sperotto, and A. Pras, "A high-performance, scalable infrastructure for large-scale active DNS measurements," *IEEE J-SAC*, 2016.