Privacy Risk Assessment on Email Tracking

Haitao Xu*‡, Shuai Hao†‡, Alparslan Sari†, and Haining Wang†
*Northwestern University, Evanston, IL, USA
†University of Delaware, Newark, DE, USA
Email: hxu@northwestern.edu, {haos,asari,hnw}@udel.edu

***Abstract*—Today's online marketing industry has widely employed email tracking techniques, such as embedding a tiny tracking pixel, to track email opens of potential customers and measure marketing effectiveness. However, email tracking could allow miscreants to collect metadata information associated with email reading without user awareness and then leverage the information for stealthy surveillance, which has raised serious privacy concerns. In this paper, we present an in-depth and comprehensive study on the privacy implications of email tracking. First, we develop an email tracking system and perform real-world tracking on hundreds of solicited crowdsourcing participants. We estimate the amount of privacy-sensitive information available from email reading, assess privacy risks of information leakage, and demonstrate how easy it is to launch a long-term targeted surveillance attack in real scenarios by simply sending an email with tracking capability. Second, we investigate the prevalence of email tracking through a large-scale measurement, which includes more than 44,000 email samples obtained over a period of seven years. Third, we conduct a user study to understand users' perception of privacy infringement caused by email tracking. Finally, we evaluate existing countermeasures against email tracking and propose guidelines for developing more comprehensive and fine-grained prevention solutions.**

I. Introduction

With the increasing importance and maturity of online marketing industry, email has become an ideal channel to achieve an economical and effective marketing solution. To date there are around 4.1 billion email accounts and 2.5 billion email users worldwide [17]. The sheer size of the email user base and the personalized nature of direct email render email tracking the most effective marketing tactic for digital marketers. Email tracking services (ETSes) (e.g., [11], [12]) exist to help email marketers personalize marketing campaigns and collect email access statistics to identify potential customers and drive increased revenues.

ETSes track email opens and allow email marketers to collect personal information of email recipients without their awareness, which raises serious privacy concerns. Due to the open nature of email, any email user could be reached by an email with built-in tracking, and a simple email open could divulge rich metadata information associated with the email reading activity to the sender. The metadata information suffices for miscreants to infer the geolocation, email reading device environment, and even the work and sleep schedule of email recipients. However, the privacy issues with email tracking have not yet been fully studied in previous literature.

In this paper, we conduct an in-depth and comprehensive study on the privacy implications of email tracking. We tackle the issue from different perspectives, including demonstrating its privacy threats, estimating its real-world prevalence, examining users' perception of its privacy risks, and proposing practical countermeasures. First, we develop an email tracking system and perform real-world tracking on hundreds of solicited crowdsourcing participants. We found that reading an email with tracking capability could disclose user metadata information. The seemingly innocent information not only allows miscreants to infer the recipient's privacy information such as real-world identity, email reading environment, real-time whereabouts, and work and sleep schedule, but also suffices for determined miscreants to mount a long-term targeted surveillance attack against email recipients.

Second, we investigate the prevalence of email tracking in the real world with a large-scale, empirical measurement study. We collected 44,449 emails originating from 928 unique email domains. Up to 24.7% of emails were found to be embedded with at least one tracking beacon. The prevalence of email tracking varies with the categories of email domains[1]. About 57.8% of `Travel` emails are equipped with tracking capability. An email domain could use multiple different ETSes for tracking purposes. The domain `staples.com` leverages up to nine different ETSes to track email recipients.

Third, we examine users' email usage behavior and their perception of privacy infringement by email tracking, in a user study conducted through a crowdsourcing platform. We received 291 valid responses from 291 unique participants in 39 countries. Most participants were found to check emails quite often. However, 52.1% of participants did not realize that opening an email could end up with being tracked. 86% of participants consider email tracking as a serious privacy threat and would adopt email tracking prevention tools to protect their privacy.

Finally, we evaluate existing countermeasures against email tracking and propose guidelines for developing more comprehensive and fine-grained prevention solutions. Existing web browsers, add-ons, and email clients have provided functionalities to protect users against email tracking to some extent. We summarize the limitations of all those solutions. Since email clients can be the vantage point to counteract email tracking,

‡ The first two authors contributed equally to this work.

[1] An email domain refers to the domain part next to the `@` symbol in an email address. For instance, `example.com` is considered as the email domain of `John@example.com`. An email domain indicates the origin of an email.

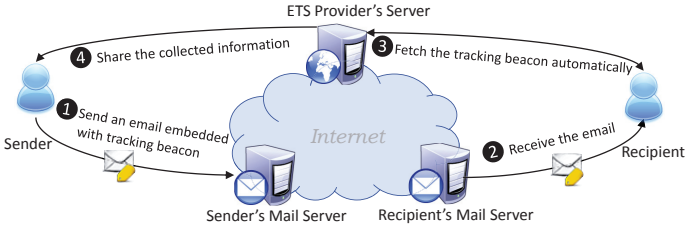


Fig. 1: How email tracking works.

we offer insights into how email clients help in defending users' privacy from email tracking.

II. Background

In this section, we first introduce email tracking techniques, and then address potential ethical concerns in our study.

A. Workflow of Email Tracking

Email tracking provides the technical foundation for email marketing. We illustrate how email tracking typically works in Figure 1. The workflow of email tracking mainly involves three parties: the email sender, the email recipient, and the third-party email tracking service (ETS) provider. Suppose a sender plans to send an email to a recipient and also track the recipient's email reading activity. The sender may turn to a ETS provider, such as third-party plugins [4]–[6] and popular email marketing platforms [7]–[9], and deploy the service on her email composing environment. Thus, every composed email will be automatically embedded with a tracking beacon linking to an object (typically a tiny image) hosted on the ETS server. The sender sends out an email with a tracking beacon embedded (**step 1**). The recipient's mail client retrieves the email from her mail server (**step 2**). The recipient opens the email as usual, which will trigger an automatic HTTP request for the embedded beacon to the ETS server (**step 3**). The ETS server responds with the requested beacon, then notifies its customer, the email sender, of the email reading, and sends the metadata associated with the email reading back to the sender (**step 4**). At this point, the sender completes her initial tracking of the recipient. One important characteristic of email tracking is its persistence. Every time the email is read by the recipient, the activity flows marked in steps 3 and 4 will repeat. Such a characteristic allows the sender to exactly master every email reading of the recipient and makes long-term tracking feasible.

B. Three Types of Tracking Beacons

Most email tracking services (ETSes) adopt similar email tracking techniques. By deploying several popular ETSes on our computing system and then examining the emails we compose and send out, we can identify three types of tracking beacons: tiny (e.g., 1x1) transparent pixel images, images containing the recipient's email address information, and explicit URL links containing the recipient's email address information. The first two types of tracking beacons would be automatically requested by the recipient's email client upon each email reading. Unlike the first two types or the beacon described in Figure 1, the third type (i.e., the tracking URL) needs the recipient's explicit click action to invoke an HTTP request. We give an example of tracking beacon (a pixel image) used by Facebook in Figure 2.

```
style=""><td height="20" style="line-height:20px;"
colspan="3">&amp;nbsp;</td></tr></table><span style="">
<img src="https://www.facebook.com/email_open_log_pic.php?
cn=3DheLwNDS1h0&amp;amp;mid=3D52ffe21a36b62G5af5cc2b81f1G0G3c2
G6e00d71f"style="border:0;width:1px;height:1px;"/>
</span></td></tr></table></body></html>
```

Fig. 2: An example of tracking beacon used by Facebook.

C. Ethical Consideration

In this study, all the user study, data collection, and experiment plans have been vetted and approved by the Institutional Review Board (IRB) at our institution. In addition, we anonymized the metadata information embedded in the collected emails prior to using them for our study.

III. Measurement Methodology

In this section, we describe our measurement methodology to assess privacy risks caused by email tracking. We conducted three groups of experiments to gauge the issue from different angles. In particular, we first examined the possibility of real-world privacy threats posed by email tracking, by performing real-world email tracking using an email tracking system built by ourselves; second, we investigated the prevalence of email tracking among real-world daily email activities through a large-scale measurement study; last, we attempted to understand individual users' perception of email privacy issues through a two-month long user study on a crowdsourcing platform. With the insights from those experiments, we proposed practical mitigation approaches to addressing email tracking privacy issues.

A. Experiment 1: Privacy Threats from Email Tracking

We developed an email tracking system using the same set of technologies as ETSes. Then we sent a "thank you" email to each of the 715 unique participants individually from one of our previous research studies [38]. Each email was automatically embedded with a tiny image tracking beacon. Technically, each email reading by the recipient could result in sending an HTTP request back to us, which contains the data associated with the recipient, such as IP address, user agent, and timestamp. We demonstrate with our empirical results that the seemingly trivial information leaked out through email tracking not only allows miscreants to infer the recipient's private and sensitive information, such as real-world identity, email reading environment, real-time whereabouts, and work and sleep schedule, but also has the unexpected power for determined miscreants to conduct long-term surveillance attacks on the targeted email users. We detail our experiment results and analysis in Section IV.

B. Experiment 2: Prevalence of Email Tracking in Real World

After evaluating the privacy threats from email tracking, we further study the prevalence of email tracking among real-world email usage. To this end, we collected a large set of more than 44,000 *inward* emails (i.e., the emails received) from two data sources: (1) the emails we received in our own inbox folders from August 2010 through April 2017, and (2) the emails periodically received from our subscription to the top 300 websites in Alexa's list [10].

We examined the HTML source of each email to identify the possibly embedded tracking beacons. We studied the prevalence of each type of tracking techniques. We categorized email domains with the help of a public website categorization database [1] and examined the prevalence of email tracking by domain categories. We found that a significant proportion of email senders leverage third-party ETSes rather than deploying their own tracking systems. We identified the most popular ETSes and studied the number of ETSes used by a sender. We present our detailed analytical results in Section V.

C. Experiment 3: User Perception of Email Tracking

After the investigation of the privacy threats arising from email tracking and its prevalence in the real world, we then attempt to understand the public's email usage habits and user perception of privacy issues with email tracking.

We conducted an online survey through a crowdsourcing platform for about two months, in which participants were asked to answer the questions related to email usage and privacy issues. In total, 291 valid responses were collected from 291 unique participants in 39 countries. We detail the findings in Section VI.

IV. Evaluating Privacy Threats Posed By Email Tracking Abuse

Email tracking could be abused by miscreants to surreptitiously harvest privacy-sensitive information of email users. To evaluate the privacy threats from email tracking abuse while not affecting email users too much, we performed real-world email tracking on a small scale by tracking only a limited number of emails we sent to the participants from one of our previous studies [38] in a relatively short period.

A. Real-world Email Tracking

We developed an email tracking system that adopts the similar technology to that of popular ETSes to track email opens. Specifically, our email tracking system consists of two components: a Chrome browser plugin and a back-end tracking server. The plugin works with any Gmail account and automatically inserts a tiny `1x1` image pixel into each outgoing email. The back-end server records each incoming HTTP request and responds with the requested image.

We deployed the tracking system on our own machine. Then we chose 715 unique email accounts, owned by 715 participants who were previously solicited worldwide through a crowdsourcing platform for one of our previous user studies [38], as our email recipients. On the first day of a time period of one month, we sent a "thank-you" email with a tracking beacon embedded to each of those 715 participants, in which we expressed the appreciation for their participation in our study. On the 15th day, we randomly selected 20 participants and sent each of them a follow-up email also with tracking capability, to mimic a miscreant who attempts to track a target over time. On the 30th day, we finished our data collection and disabled the tracking system.

TABLE I: Information typically disclosed via email tracking

Raw Field	Inferred Information
Email address	(*Who*) Online identity
HTTP request arrival time	(*When*) Email opening time
IP address	(*Where*) Location on a city level
User agent	(*How*) Device type, browser type, OS type
Number of HTTP request	Number of views

B. Privacy Risks of the Collected Data

For each incoming HTTP request, implying one time of email open at the recipient side, our back-end tracking server creates one record, which contains the five fields to store the information of email address, sending time, HTTP request arrival time, IP address, and user agent. Although seemingly trivial and innocent, those fields could allow an email tracking abuser to gather enough privacy-sensitive information about the email recipient and even launch a long-term surveillance attack.

Table I summarizes the information that could be disclosed as a result of email tracking. We highlight that the three fields, *email address, IP address*, and *user agent*, are quite privacy-sensitive. An email address is usually linked to online social networks (OSNs) as the unique account identification (ID). Checking an email address in an OSN site would reveal the social profile of the email owner. An IP address can be used to locate the email recipient with about 90% accuracy on a city level within a radius of tens of kilometers [2]. A user agent could reveal the email reading environment. All the inferred information together could allow an attacker to piece together the profile of email owners for further targeted attacks.

C. Experiment Results

Based on our experiment of tracking the email opening activities on 715 unique email accounts, we report our analytical results below, which are also illustrated in Figure 3.

Email accounts breakdown by domain. Grouped by the email domain, the top 5 email domains with most unique email accounts are `Gmail.com` with 520 (72.7%) email accounts, `Yahoo.com` with 86 (12%) accounts, `Hotmail.com` with 48 (6.7%) accounts, `Live.com` with 11 (1.5%), and `Outlook.com` with 9 (1.3%) email accounts. All these email domains except `Hotmail` are actually among the top 10 email clients by market share [3].

Email accounts linking to OSN profiles. We also checked each email address on Google Plus using its feature "find people by email." Overall, for up to 538 (75.2%) of the 715 email account owners, their social profiles can be identified by simply checking their email addresses in the OSN Google Plus. Checking the email addresses in other OSNs is expected

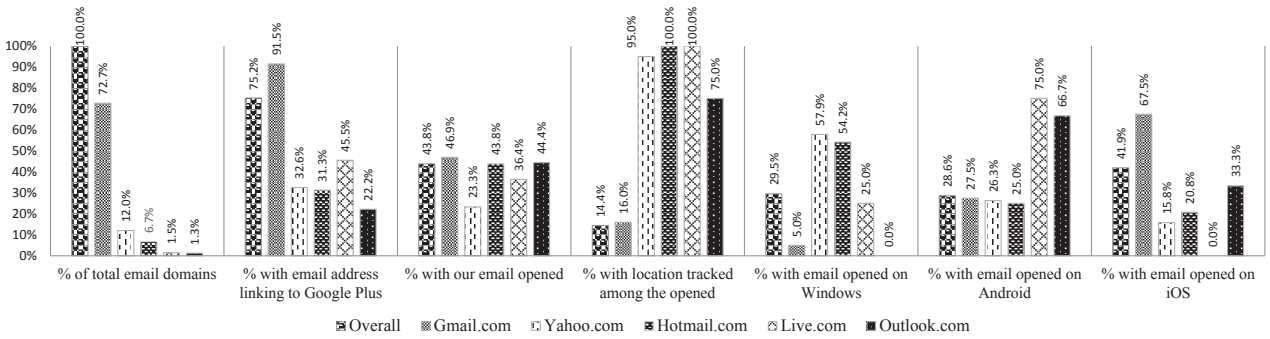


Fig. 3: Breakdown of the overall and the top 5 email domains in terms of the ratios of email accounts that fall under each email domain, link to Google Plus, have emails opened, have the email recipient's geolocation tracked, and have emails opened on different device types, respectively.

to reveal more cross-referencing profiles. The profile of an OSN user typically covers various information about the user such as her/his birthdate, living address, occupation, and hobbies. The result shows the surprising power of an email address in revealing its owner's real online identity. Specifically, 91.5% of `Gmail` accounts, 32.6% of `Yahoo mail` accounts, 31.3% of `Hotmail` accounts, 45.5% of `Live` mail accounts, and 22.2% of `Outlook` accounts were being associated with their Google Plus profiles.

Email open rates. Among the 715 unique email recipients, 313 (43.8%) of them have opened our email at least once, according to our tracking server logs. The rest of email recipients either may never open our "thank-you" emails or adopt some email tracking prevention mechanisms (discussed in detail in Section VII). The email open rates for the top 5 email clients vary from 23.3% to 46.9%. In the following analysis, we mainly focus on the 313 email recipients who definitely have triggered our email tracking beacons.

Image proxies introduced by email clients. Some email clients, such as `Gmail` and `Outlook`, have introduced image proxies to prevent email senders from geolocating a recipient and detecting her/his email reading environment, by masking the IP address and user agent of the email recipient. By examining the user agents in the HTTP request headers, we found that 210 (67.1%) out of the 313 email clients have deployed image proxies for protecting user privacy. The evidence is that those user agents include the content, which clearly shows that the HTTP requests are indeed made by the Google image proxies on behalf of the original email recipients, as shown below. More specifically, up to 83.7% of the `Gmail` accounts that have triggered our email tracking beacons were identified to adopt image proxies, and the percentages for `Outlook` accounts and `Yahoo mail` accounts are 25% and 5%, respectively. None of `Hotmail` accounts and `Live` mail accounts were deployed with image proxies.

```
Mozilla/5.0 (Windows NT 5.1; rv:11.0) Gecko
Firefox/11.0 (via ggpht.com GoogleImageProxy)
```

Geolocalization with IP address. For any email accounts that directly make HTTP requests for the tracking image without using image proxies, their real IP addresses could be harvested and further leveraged to geolocalize the location city of the email recipients with high accuracy by querying the publicly available GeoIP databases [2]. We can obtain the location city information for the 16.3% of `Gmail` accounts, 95% of `Yahoo mail` accounts, 100% of `Hotmail`, 100% of `Live` mail accounts, and 75% of `Outlook` accounts, which viewed our tracking emails. The results demonstrate the effectiveness of image proxies in thwarting IP-based geolocalization.

Inferring the device type with user agent. The email reading environment (mainly computers and smartphones) is usually revealed in the user agent field. As shown in Figure 3, email recipients who divulge their user agent information without awareness are found to read emails on Windows desktops and mobile devices. Specifically, 29.5% of email recipients read emails on Windows desktop, 28.6% read emails on Android devices, and 41.9% on iOS devices. The results may imply that reading emails on mobile devices is more likely to cause privacy leakage than reading emails on desktops. We performed similar analyses on the top 5 popular email clients. One very interesting observation is that reading emails on desktops with `Gmail` or `Outlook` email clients could largely prevent information leakage. More precisely, among the email readings that cause user agent information leakage on `Gmail` and `Outlook` clients, Windows desktops only contribute to 5% and zero percent, respectively, while mobile devices contribute to 95% and 100%, respectively. It indicates that the image proxy practice adopted by `Gmail` and `Outlook` performs well on desktop computers in protecting users' privacy but ineffective on mobile devices; or the two popular email clients do not deploy image proxies for their mobile versions. The other three popular email clients do not exhibit such characteristics, and for them, both Windows desktops and mobile devices contribute to significant proportions of information leakage due to email reading.

D. Long-term Surveillance Attack with Email Tracking

An email address itself could reveal the email user's identity in the real world. A simple email reading could disclose the current living city, the email reading device, and even the work and sleep schedule of the targeted email user. The user identity and the real-time whereabouts information inferred could raise great privacy concerns and may cause security threats to the

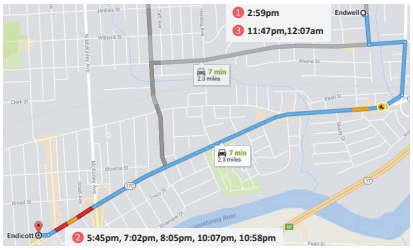


Fig. 4: Restoring activities of Alice on one Friday from 2:59pm local time to 12:07am (midnight) with email tracking.

Fig. 5: Tracking Bob over three weeks simply with sending him two emails with tracking capability.

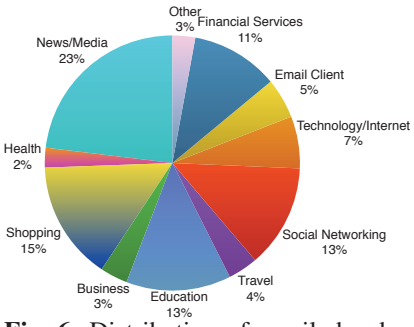


Fig. 6: Distribution of emails by domain category.

recipient. For instance, terrorists and criminals would use such surveillance information to plan and execute a targeted attack.

In our experiment, each of the 715 participants received one email from us, and 20 of them received a second one. We found that about one third (33.2%) of recipients read the email at least twice, and one recipient read the email eight times within two days. We use two case studies to demonstrate the unexpected powerfulness of tracking people through email tracking.

Case study I. In the first case, the email recipient is referred to as Alice, and her real name and social profile (including occupation, marriage status, and home address information) can be easily identified by checking the email address in the Google Plus. She read our email 8 times within 12 hours of the email sending. Based on the eight email readings, we have successfully restored her activities during those 12 hours, as shown in Figure 4. We sent Alice one email at 1:07 pm on one Friday. Alice read the email at 2:59 pm on an iPhone in Endwell, NY for the first time, labeled as ① in the figure. Two hours and 46 minutes later, Alice read the email again at 5:45 pm in Endicott, NY, a nearby city 2.3 miles away from Endwell. She read the email for another 4 times in Endicott, and the 6th email reading was at 10:58 pm, labeled as ②. Less than one hour later, she read the email again at 11:47 pm in Endwell, and read it the last time at 12:07 am in Endwell, as shown as ③. The Alice's eight email readings reveal her real-time whereabouts on that Friday afternoon and evening. With such information, one may conjecture that Alice may work and live in Endwell, NY; she usually reads email on her iPhone; she goes to a nearby city Endicott to spend Friday evening, returns back to Endwell very late, and goes to asleep after the midnight. A burglar may calculate the time when Alice is out of home and commit a home burglary.

Case study II. In the second case, as illustrated in Figure 5, we refer to the email recipient as Bob. Similarly, we successfully located his Google Plus profile. We sent Bob the first email on Day 1, and he read it on an Android Tablet on the same day in the city Sarajevo of Bosnia and Herzegovina. We sent Bob the second email on Day 15, and the email was read on an Android phone on the second day in Podgorica, Montenegro, a city in the neighbor country with 240 Kilometers away or 4 hours and 27 minutes driving distance from Sarajevo. Bob read the second email again on a Windows desktop on Day 21 in Sarajevo, Bosnia and Herzegovina. In summary, two emails have allowed us to track Bob in a time period of several weeks.

A potentially more sophisticated attack methodology with email tracking. In a real attack, an adversary could conduct far more sophisticated surveillance attacks with email tracking. An adversary could first perform reconnaissance with the victim's email address to collect the information from the associated OSN profile, such as the name, gender, hobbies, occupation, affiliated company, and home address. Next, with the collected information, the adversary fabricates some "must-read" targeted emails with tracking beacons embedded. Then he periodically sends one of those bogus emails to the victim and tempts the latter to read it. In addition, in order not to be blocked, the adversary could use a sufficient number of free, disposable, and temporary email addresses from disposable email address services online for this purpose. The above proposed attack strategy would allow the adversary to track the victim for a long time and conduct further attacks at will.

V. Estimating the Prevalence of Email Tracking in the Real World

We performed a large-scale empirical measurement study on the prevalence of email tracking among daily email activities. We present our dataset and analytical results below.

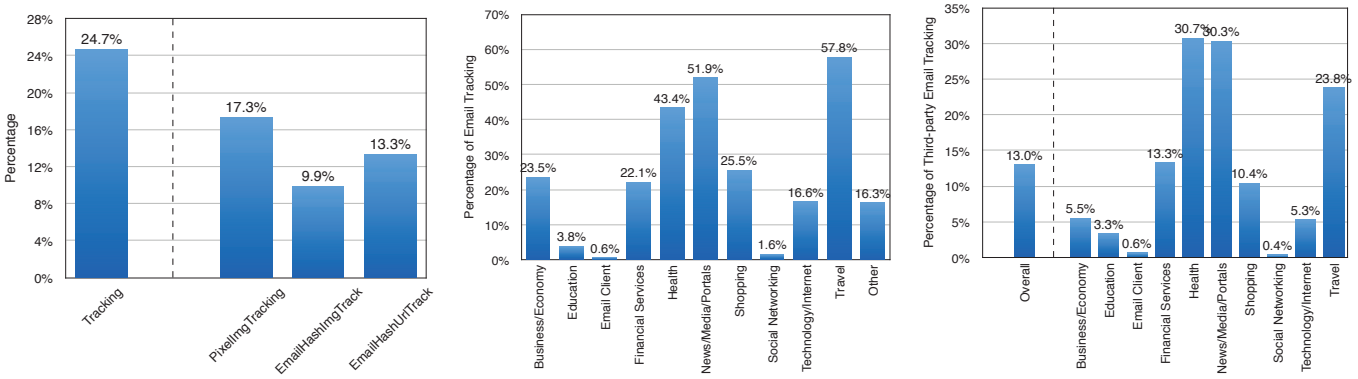
A. Data Collection

We collected a total number of 44,449 emails originating from 928 unique email domains. Specifically, 59.9% (26,643) of emails originating from 713 unique email domains were addressed to seven personal email accounts belonging to five individuals in the past seven years from 2010 through 2017. The rest 40.1% (17,806) of emails from 266 unique email domains were periodic updates addressed to one honeypot email address from the subscription to the top 300 Alexa site domains. Note that these two data sources have 51 overlapping email domains. Our dataset is representative given the diverse email domains and many years of email account usage.

The HTML source code of an email is the key to determine whether the email is tracked or not. As mentioned in Section II, there are mainly three types of tracking beacons, each of which has one email equipped with the tracking capability.

B. Prevalence of Email Tracking

The analysis of our email dataset reveals that email tracking techniques are commonly used in everyday email commu-



(a) **(b)** **(c)**

Fig. 7: (a) Percentage of emails with tracking beacons embedded. (b) Percentage of emails with tracking beacons by domain category. (c) Percentage of emails with third-party tracking.

nication. As shown in Figure 7a, up to 24.7% of emails are found to include at least one kind of tracking beacons. Specifically, 17.3% of emails were embedded with "invisible" pixel images[2]; 9.9% of emails contain regular images with either the recipient's email address or its MD5 or SHA1 hash value embedded; 13.3% of emails contain explicit URL links with the recipient's email address or its hash value embedded. Invisible pixel images are found to be the most popular beacons used for email tracking. Note that it is quite common that two or three types of email tracking beacons are used in one email.

Next, we check whether the prevalence of email tracking varies with the categories of email domains. To this end, we first categorized all 928 domains in our dataset into ten categories and one `Other` category by looking up each domain in Symantec Corporation's website categorization database [1]. Figure 6 shows the distribution of emails in each domain category. Specifically, `News/Media` contributes the most emails (about 23%) among all domain categories. About 75% of emails originate from the domains in the following five categories, `News/Media`, `Shopping`, `Education`, `Social Networking`, and `Financial Services`.

The prevalence of email tracking is found to vary with the categories of email domains. As depicted in Figure 7b, `Travel` and `News/Media/Portals` are the top two domain categories with high percentages of emails equipped with tracking capability, 57.8% and 51.9%, respectively. More than 40% of `Health` emails are also found to track the recipients. About a fifth to a quarter of `Shopping`, `Business/Economy` and `Financial Services` emails were embedded with tracking beacons. By contrast, only 0.6% of emails from the `Email Client` category track their recipients, which is reasonable since such emails (e.g., Gmail and Hotmail) could be safely regarded as personal emails, and people seldom track recipients in personal email communications.

C. Popularity of Email Tracking Services

Every email tracking beacon contains a URL linking to some external resources (e.g., images or webpages), and the domain name of such a URL indicates the real domain that performs email tracking. One interesting observation is that the domain performing actual tracking (termed as *tracking domain*) is not always the same as the domain sending the email (termed as *email domain*). In such a scenario, the *email domain* is believed to be using third-party tracking. Figure 7c presents the prevalence of third-party tracking. Specifically, 13.0% of all emails are observed to use third-party email tracking. `Health`, `News/Media/Portals`, and `Travel` are the top 3 domain categories using third-party tracking, with up to 30.7%, 30.3% and 23.8% of their emails, respectively. In combination with Figure 7b, these three domain categories represent the top 3 categories most likely to use email tracking and also third-party email tracking in their outgoing emails.

Then we pay special attention to the third-party *tracking domains*, which serve at least two unique *email domains* and actually provide email tracking services to email marketers. Based on both the number of email domains being served and the domain categories being covered, the top 10 ETSes are determined and shown in Figure 8. Each of the top 10 ETSes covers 5 domain categories and serves 9 unique email domains on average. The top 3 most popular ETSes are `returnpath.net`, `emltrk.com`, and `responsys.net` owned by the Oracle Corporation. Oracle's another tracking service `bluekai.com`, Adobe's `demdex.net`, and Google's `doubleclick.net` are quite popular too. In addition, an email domain could use multiple different ETSes for tracking purposes. Figure 9 shows the CDF of the number of ETSes used per email domain. About 29% of email domains leverage at least two ETSes for email tracking and 5.6% use more than 5 ETSes. The two email domains `staples.com` and `united.com` are found to leverage 9 and 8 different ETSes to track email recipients, respectively. The results demonstrate that email tracking is highly valued by digital marketers.

VI. Study on Email Usage and Privacy Perception

We conducted an online survey through a crowdsourcing platform for about two months. Participants were asked to answer the questions related to email usage and privacy issues. In total, 291 valid responses were collected from 291 unique participants in 39 countries. Table II lists the demographic

[2]Defined as the images with one of the three possible dimensions (width x height), `0x0`, `1x1`, and `1x3`, based on our statistical results of the dataset.

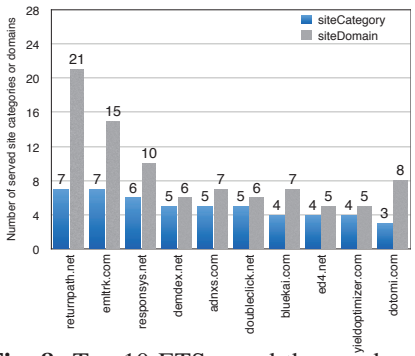


Fig. 8: Top 10 ETSes and the number of domain categories and unique domains they serve.

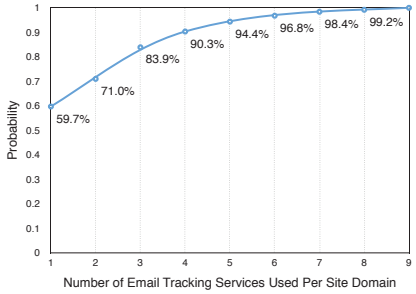


Fig. 9: CDF of the number of ETSes used per email domain.

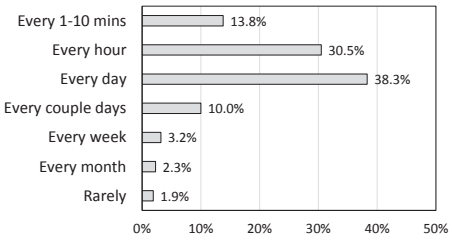


Fig. 10: Distribution of email checking frequency.

TABLE II: Demographic statistics of survey respondents

Gender	Percent	Country	Percent	Age	Percent	Education	Percent	Field	Percent
Male	70.1%	USA	47.77%	<=17	17.7%	Graduate	32.8%	Information Tech.	27.7%
Female	29.9%	Serbia	10.65%	18−24	21.5%	Bachelor	26.0%	Math. and Science	17.7%
NA	NA	Turkey	4.12%	25−34	38.3%	High Sch.	37.0%	Business	14.5%
NA	NA	UK	3.78%	35−44	17.0%	Middle Sch.	2.9%	Arts	7.7%
NA	NA	India	3.09%	>=45	5.5%	Elementary	1.3%	Other	32.5%

statistics about participants: (1) 70.1% were male, (2) 69.4% of participants were from the top five countries, (3) 76.8% of participants were between the ages of 18-44, (4) 95.8% of participants received a high school degree or higher, and (5) 27.7% of participants received the degree in the field of information technology. These participants could well represent the primary email users.

A. Email Usage Habits

Frequency of email checking. We examined how often a user checks for new emails and whether she reads one email multiple times. Figure 10 shows that most participants check emails quite often: 13.8% check for new emails every 5 or 10 minutes, 30.5% check emails every hour, and 38.3% every day. The responses also reveal that up to 42.8% of users would read an email multiple times.

Willingness to read emails or click email URL links. Email reading or clicking on the embedded URL link is typically required for email tracking to work. Figure 11a depicts people's willingness to open an email or click through the embedded URLs when receiving two kinds of emails: regular emails from a friend and promotion emails from advertising companies. It shows that when receiving an email from a friend, about a half of (49.5%) users always open the email and 17.5% of users often open the email; 13.5% and 28.9% of users would always or often click on the embedded link, respectively. People have a relatively low willingness to visit the embedded links even for an email from a friend. Comparatively, people have a much lower willingness to read a promotion email or click the embedded links. About 24.3% and 37.3% of users choose to never read such emails or click on the links, respectively.

B. User Perception of Email Privacy Violations

When told that email reading could cause a recipient to be tracked, 52.1% of participants were not aware of this privacy risk. We then measured user perception of possible email privacy violations and presented the results in Figure 11b. Participants were asked to respond on a scale from 1 (privacy not important at all) to 7 (privacy very important) on whether they would sacrifice the privacy for benefits like reading a new email. About one half of participants gave ratings of 6 and 7 to express that they value online privacy very much. Only 5.8% of participants did not care about online privacy (ratings 1/2).

We then asked participants to rate for each kind of possibly disclosed information on a scale from 1 (no privacy concern) to 5 (very serious privacy concern). As shown in Figure 11c, the location information is what people are most concerned about and 57.1% of participants gave high rating scores of 4 and 5. The disclose of the device type being used also raises privacy concerns. Participants are less concerned about the browser type information or email reading times.

C. User Demand for Email Tracking Prevention Tools

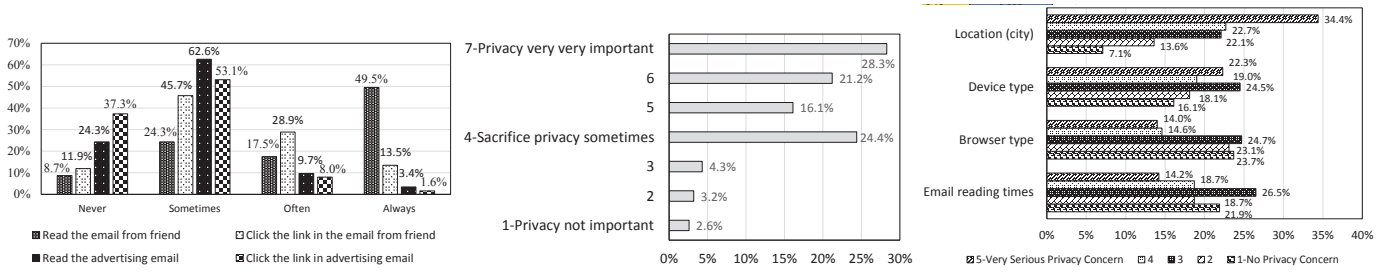
The above results demonstrate that people are indeed concerned about possible privacy violations caused by email tracking. We then studied if people have deployed email tracking prevention tools to protect their privacy. Up to 93% participants did not use any such tool. It makes sense considering the fact that more than 50% participants had no awareness of email tracking. However, people indeed have the demand for reliable email tracking prevention tools, and 86% expressed interests in using such tools in the future.

VII. Protecting Against Privacy Infringement

In this section, we evaluate existing potential countermeasures against email tracking and then provide guidelines for more effective intervention at email clients.

A. Evaluating Existing Defense Against Email Tracking

Existing web browsers, add-ons, and email clients have already provided functionalities to protect users against email



(a) **(b)** **(c)**

Fig. 11: (a) Distribution of users' willingness of email opening and clicking the URLs. (b) Distribution of users' perception of privacy violations. (c) Distribution of users' perception of the privacy level with information.

tracking to some extent. Popular browsers, such as Chrome, Mozilla Firefox, and Microsoft Edge, have built-in privacy settings that users can enable to block all images when rendering web pages. However, most users are not aware of such browser settings. Furthermore, users just need to block the tracking images rather than all images.

Some add-ons, such as the PixelBlock [16] for Google Chrome, are developed exclusively to address the email tracking problem. However, such tools usually prevent pixel-based email tracking under the help of a blacklist of well-known email tracking services like Mailtrack. One limitation is that such blacklist-based tools can only be effective to detect known tracking systems and could be easily circumvented by newly deployed systems, such as the simple tracking system we built for this study.

Nearly all popular email clients have an option for turning off image display, but still the major drawback is that users have to choose between displaying all images or blocking them all. Text-only email clients such as Alpine do not support any image display and thus render email tracking impossible. Although they provide one of the simplest and strongest prevention of email tracking, the inability of displaying any image hinders their wide use. Some email clients, such as Gmail and Outlook, already utilize proxy servers to request the embedded images in an email on behalf of email recipients. Such a practice can indeed hide users' IP addresses and user agents, and thus prevent geolocalization and location tracking. However, email tracking abusers can still infer when and how many times the email recipients read the email based on the metadata information associated with the HTTP requests from those proxy servers.

B. Guidelines for Developing Comprehensive and Fine-grained Prevention at Email Clients

The above evaluation shows that there is still no silver bullet for tackling email tracking. It is challenging to design a comprehensive prevention method to address all privacy issues caused by email tracking. Thus, we provide several guidelines towards designing a practical tool to mitigate the privacy concerns.

First, as an entity responsible for accessing and managing a user's email, email clients are more suitable to provide a non-intrusive defense against email tracking. Otherwise, users have to take explicit efforts to survey various online tools that claim to defend against email tracking and install one on their end systems.

Second, keeping users aware of an email tracking attempt should be an option-in feature of email clients. When an email client detects a tracking beacon embedded within a new email, its user should be notified by placing a small icon next to the email subject or highlighting the email with a predefined label.

Third, email clients need to provide more active and comprehensive prevention besides their existing features (e.g., deploying proxy servers). As discussed above, proxy servers cannot prevent miscreants from inferring the email reading time and frequency. To address this problem, email clients can either directly remove the embedded tracking beacons before the emails are read or proactively block the request that could be issued by a tracking beacon.

Finally, the defense mechanisms adopted by email clients should enable users to customize their own email tracking prevention policies based on their own system preferences and personal expertise. For example, a user should be allowed to only display the images in an email from a sender in her contact list or from anyone she designates as trusted.

VIII. Related Work

Privacy Concerns on Online Tracking. The online tracking [19], [24], [36], [37], e.g., the *Web Bugs* [20], [32], which leverages invisible third-party images to track page viewing, has been used for a variety of purposes, such as targeted advertisements, customized recommendations and search results, analysis of user preference, and surveillance of user activities. Meanwhile, the prevalence of various online tracking activities has raised significant privacy concerns [25], [31], [33].

Gross *et al.* [29] discussed the information revelation in social networks and related privacy issues. They pointed out possible attacks on privacy. They demonstrated that most of users are not aware of necessary privacy settings to prevent potential attacks. Goldfarb *et al.* [27] studied the impact of privacy regulation on web-based advertisements. They showed that advertisers track consumers' behaviors and browser history to deliver target Ads to consumers, and consumers are not aware of the information collection process. Greengard [28] explained how companies utilize the predictive analysis on purchasing patterns and behaviors of users, and discussed the cookie-based tracking prevention. Mayer *et al.* [33] discussed the technologies and public policies related to the tracking

activities of third-party web services. Datta *et al.* [23] developed an automated tool called "AdFisher" to explore the user behavior and advertisement interactions through statistical methods. Melicher *et al.* [34] collected the browser histories and interviewed a group of users to understand the users' perception on the online tracking, and examined the efficacy and user preference on the "controlling tracking."

The tracking pixels in emails can cause more serious privacy issues than in web surveillance since their URLs can be easily associated with a user's email address. Also, HTML web bugs are normally based on browser cookies, while the tracking pixels embedded in emails typically do not require the collaboration from web browsers.

Email Privacy Concerns. The privacy concerns for email communications have been investigated from various aspects [21], [22], [30]. Sipior *et al.* [18] discussed the legal issues regarding employee and employer related email privacy issues, and they also studied the U.S. legal system regarding p email privacy protection. However, the tracking of email activities was not mentioned in this study.

It is challenging to launch an effective email campaign due to privacy concerns. Cases *et al.* [22] studied the role of privacy and customers' attitude towards an email campaign. Narayanan *et al.* [35] considered that email addresses may not be used as the "Personally Identifiable Information". Nevertheless, the email tracking could still pose privacy risk due to the exposure of email reading behavior. Zhao *et al.* [39] demonstrated the possibility of combining the email tracking with a phishing attack, where an attacker can mount much more sophisticated attacks based on email open rate.

In parallel with our study, Englehardt *et al.* [26] also tackled the privacy issues caused by email tracking. They assembled the emails from commercial mailing-lists and also revealed the prevalence of email tracking activities. They then proposed a defense mechanism by stripping tracking tags based on tracking protection lists. Different from their work, we demonstrated the potential attacks to infringe user privacy or even pose real-world risks. Also, we performed detailed user studies to understand a user's perception of email tracking.

IX. Conclusion

Email tracking collects privacy-sensitive user information and raises great privacy concerns. In this work, we conducted an in-depth and comprehensive evaluation of the privacy implications of email tracking. We developed an email tracking system, performed real-world tracking, and demonstrated with real scenarios in which the information divulged due to email tracking suffices for miscreants to mount a long-term stealthy surveillance attack. We estimated its real-world prevalence through a large-scale measurement study involving more than 44,000 email samples, and we found that up to 24.7% of emails track their recipients and some email domains adopt nine different email tracking services to track email recipients. We examined a user's privacy perception of email tracking with a crowdsourcing study, and we found that more than a half of users have no awareness of email tracking but 86% of them deem it as a serious privacy threat. We also surveyed the existing countermeasures and proposed guidelines for building a more comprehensive and fine-grained prevention solution.

References

[1] Bluecoat Site Review. http://sitereview.bluecoat.com/sitereview.jsp.
[2] GeoIP2 DB. https://www.maxmind.com/en/geoip2-enterprise-database.
[3] Email Client Market Share. https://emailclientmarketshare.com.
[4] Streak: Email View Tracking. www.streak.com/email-tracking-in-gmail.
[5] Mailtrack. https://mailtrack.io/en/.
[6] HubSpot: Email Tracking. hubspot.com/products/sales/email-tracking.
[7] Return Path: We Know Email. https://returnpath.com/.
[8] Emltrk. http://emltrk.com/.
[9] Responsys. http://www.oracle.com/us/corporate/acquisitions/responsys.
[10] Alexa Top Sites. http://www.alexa.com/topsites.
[11] Bananatag: Email Tracking. https://www.bananatag.com/.
[12] Boomerang for Gmail. http://www.boomeranggmail.com/.
[13] The Hosting Platform of Choice. *cPanel, Inc*. https://cpanel.com/.
[14] Image Block: Add-ons for Firefox. hemantvats.com/image-block/.
[15] Picture Blocker: Add-ons for Firefox. https://goo.gl/6WZCiu.
[16] PixelBlock - Chrome Web Store. https://goo.gl/VAhiX6.
[17] Email Statistics Report 2014-2018. https://goo.gl/TmeYoy.
[18] J. C. Sipior and B. T. Ward. The Ethical and Legal Quandary of Email Privacy. In *Communications of the ACM*, 1995.
[19] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *ACM CCS*, 2014.
[20] A. Alsaid and D. Martin. Detecting Web Bugs with Bugnosis: Privacy Advocacy Through Education. In *PETS*, 2002.
[21] H. Berghel. Email: the Good, the Bad, and the Ugly. In *Communications of the ACM*, 1997.
[22] A. Cases, C. Fournier, P. Dubois and J F. Tanner Jr. Web Site Spill Over to Email Campaigns: The Role of Privacy, Trust and Shoppers' Attitudes. In *Journal of Business Research*, 2010.
[23] A. Datta, M. Tschantz, and A. Datta. Automated Experiments on Ad Privacy Settings. In *PETS*, 2015.
[24] P. Eckersley. How Unique is Your Web Browser? In *PETS*, 2010.
[25] S. Englehardt and A. Narayanan. Online Tracking: A 1-million-site Measurement and Analysis. In *ACM CCS*, 2016.
[26] S. Englehardt, J. Han, and A. Narayanan. I Never Signed Up for This! Privacy Implications of Email Tracking. In *PETS*, 2018.
[27] A. Goldfarb and C. E. Tucker. Online Advertising, Behavioral Targeting, and Privacy. In *Communications of the ACM*, 2011.
[28] S. Greengard, Advertising Gets Personal. In *Communications of the ACM*, 2012.
[29] R. Gross and A. Acquisti. Information Revelation and Privacy in Online Social Networks. In *ACM WPES*, 2005.
[30] D. Kafura, D. Gracanin, M. Perez-Quinones, T. DeHart, and S. Codio. An Approach to Community-Oriented Email Privacy. In *PASSAT*, 2011.
[31] A. Lerner, A. K. Simpson, T. Kohno, and F. Roesner. Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016. In *USENIX Security*, 2016.
[32] D. Martin, H. Wu, and A. Alsaid. Hidden Surveillance by Web Sites: Web Bugs in Contemporary Use. In *Communication of the ACM*, 2003.
[33] J. R. Mayer and J. C. Mitchell. Third-Party Web Tracking: Policy and Technology. In *IEEE S&P*, 2012.
[34] W. Melicher, M. Sharif, J. Tan, L, Bauer, M. Christodorescu, and P. G. Leon. (Do Not) Track Me Sometimes: Users' Contextual Preferences for Web Tracking. In *PETS*, 2016.
[35] A. Narayanan, V. Shmatikov. Myths and Fallacies of "Personally Identifiable Information". In *Communications of the ACM*, 2010.
[36] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting. In *IEEE S&P*, 2013.
[37] F. Roesner, T. Kohno, and D. Wetherall. Detecting and Defending Against Third-party Tracking on the Web. In *NSDI*, 2012.
[38] H. Xu, H. Wang, and A. Stavrou. Privacy Risk Assessment on Online Photos. In *RAID*, 2015.
[39] M. Zhao, B. An, and C. Kiekintveld. Optimizing Personalized Email Filtering Thresholds to Mitigate Sequential Spear Phishing Attacks. In *AAAI*, 2016.