On the Privacy Implications of Email Tracking

***Abstract*—Today's online marketing industry has widely employed email tracking techniques, e.g., 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 leverage the information for stealthy surveillance, which has raised serious privacy concerns. In this paper, we present our effort towards an in-depth and comprehensive evaluation of the privacy implications of email tracking. First, we develop a prototype 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 present scenarios demonstrating how easy it is to launch a long-term targeted surveillance attack simply with an email with tracking capability. Second, we investigate the prevalence of email tracking using a large-scale measurement study involving more than 44,000 email samples obtained over a period of seven years. Third, we attempt to understand users' perception of privacy infringement by email tracking with a user study. 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 [18]. The sheer size of the email user base and the personalized nature of direct email advertising 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 revenue.

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 evaluation of the privacy implications of email tracking. We gauge the issue from different angles, including demonstrating its realistic privacy threats, estimating its real-world prevalence, examining users' perception of its posed privacy risks, and proposing practical countermeasures.

First, we develop a prototype 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 ample metadata information. Those innocent-looking 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 beacons. The prevalence of email tracking varies with the categories of email domains[1]. 57.8% `Travel` emails were equipped with tracking capability. An email domain could use multiple different ETSes for tracking purpose. The domain `staples.com` leveraged up to 9 different ETSes to track email recipients.

Third, we examine users' email usage habit and perception of privacy infringement by email tracking, with 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 and conclude that email clients are in the best vantage point to counteract email tracking. We offer insights into how email clients could do better in defending users' privacy from email tracking. To the best of our knowledge, we are the first to conduct such an in-depth and comprehensive evaluation of the privacy implications of email tracking.

The remainder of this paper is organized as follows. We provide background knowledge in §2. We describe measurement methodologies in §3. We evaluate privacy threats posed by email tracking in §4. We estimate its prevalence in the real world in §5. We examine users' privacy perception of email

[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 the email.

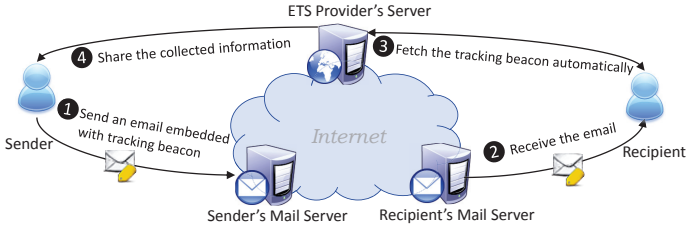

Fig. 1: How email tracking works.

```
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;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.

tracking in §6. We propose guidelines for developing a suitable prevention solution in §7. We discuss related works in §8 and conclude the paper in §9.

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 is the technical foundation of 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 her the metadata associated with the email reading (**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 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, three types of tracking beacons were identified: 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 (pixel image) used by Facebook in Figure 2.

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, with the collected emails, we anonymized the metadata information embedded prior to using them for study.

III. Measurement Methodology

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

A. Experiment 1: Evaluating 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 from one of our previous research studies, individually. Each email was automatically embedded with a tiny image tracking beacon by our system. Technically, each email reading by the recipient could result in sending an HTTP request to us, which contains ample data, such as IP address, user agent, and timestamp. We demonstrate with our empirical results that those innocent-looking information leaked out through email tracking not only allows miscreants to infer the recipient's privacy 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: Estimating the Prevalence of Email Tracking in the 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: Understanding User Perception of Email Tracking

After investigation of the privacy threats arising from email tracking and its prevalence in the real world, next we 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 from 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 in a relatively short period.

A. Performing Real-world Email Tracking

We developed an email tracking system which adopts the similar technology to that of popular ETSes to track email opens. Specifically, the 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, 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 fields including the *email address, sending time, HTTP request arrival time, IP address*, and *user agent*. Although looking trivial and innocent, those fields could allow an email tracking abuser to gather ample 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). Looking up an email address in an OSN site would reveal the social profile of the email owner. 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 those 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, 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 were identified by simply looking up their email addresses in the OSN Google Plus.

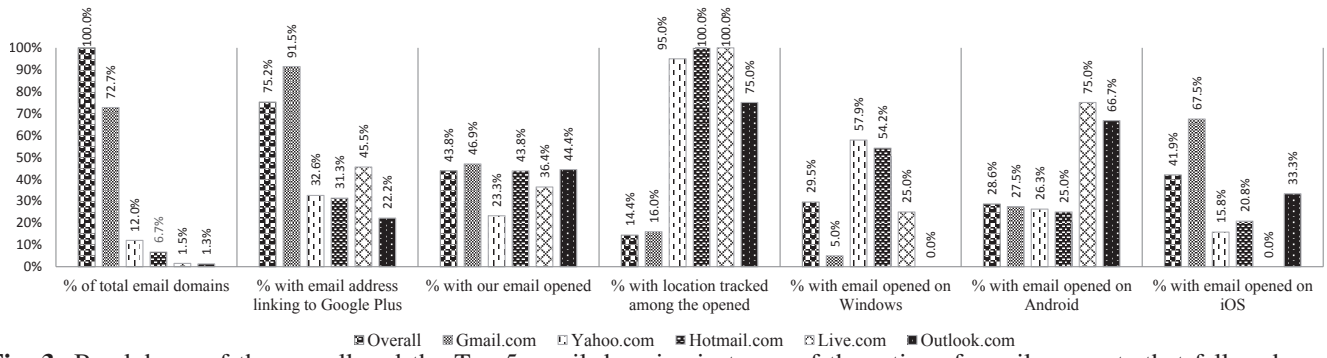

Fig. 3: Breakdown of the overall and the Top 5 email domains in terms of the ratios of email accounts that fall under each email domain, link to Google Plus, have emails opened, have the email recipient's geolocation tracked, and have emails opened on different device types, respectively.

Looking up the email addresses in other OSNs is expected to reveal more cross-referencing profiles. OSN user profile typically covers various information about a user such as her birthdate, living address, occupation, and hobbies. The result suggests the surprising power of an email address in revealing its owner's real online identity. Specifically, 91.5% `Gmail` accounts, 32.6% `Yahoo mail` accounts, 31.3% `Hotmail` accounts, 45.5% `Live` mail accounts, and 22.2% `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 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 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, and the evidence is that those user agents have the similar content like the following one, which clearly shows that the HTTP requests are indeed made by the Google image proxies on behalf of the original email recipients. More specifically, up to 83.7% of the `Gmail` accounts which 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 which directly make the HTTP request for the tracking image without the use of image proxies, their real IP addresses could be harvested and further leveraged to geolocalize the location city of the email recipients with a high accuracy by querying the publicly available GeoIP databases [2]. We have been able to obtain the location city information for the 16.3% `Gmail` accounts, 95% `Yahoo mail` accounts, 100% `Hotmail`, 100% `Live` mail accounts, and 75% `Outlook` accounts, which have viewed our tracking emails. The results demonstrate the effectiveness of image proxies in preventing 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% 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 which cause user agent information leakage on `Gmail` and `Outlook` clients, Windows desktops contribute to only 5% and zero percent, respectively, while mobile devices contribute to the rest 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 not effective on mobile devices, or the two popular email clients do not deploy image proxies for their mobile version. 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. Proposing 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 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 a third (33.2%) recipients read the email at least twice, and one recipient read the email 8 times within 2 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) was easily identified by looking up the email address in the Google Plus. She read our emails 8 times within 12 hours of our email sending. Based on the eight email readings, we have successfully restored her activities during those 12 hours, illustrated in Figure 4. We sent Alice one email at 1:07 pm on one Friday. Alice read the email the first time 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 the second time at 5:45 pm in Endicott, NY, a neighbor city 2.3 miles far from Endwell. She read the email for another 4 times in the city, and the last email reading was at 10:58 pm, labeled as ② in the figure. Less than one hour later, she read the email at 11:47 pm in Endwell, and read it the last time at 12:07 am in Endwell, shown as ③. Alice's eight email readings provided her real-time whereabouts on that Friday afternoon and evening. With those information, one may conjecture that Alice may work and have an apartment at in Endwell, NY; she usually reads email on her iPhone; she goes to a neighbor city Endicott to spend Friday evening, returns back to Endwell very late, and goes to asleep after the midnight. A burglar may calculate the timing that Alice is out of home and perform a home burglary.

Case study II. In the second case, 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 far or 4 hours and 27 minutes driving from Sarajevo. Bob read the second email the second time 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 collected information, he 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 of 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) 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) 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 use.

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

Analysis of our email dataset reveals that email tracking techniques are commonly used in everyday email communication. As shown in Figure 6, up to 24.7% emails are found to include at least one kind of tracking beacons. Specifically, 17.3% emails were embedded with "invisible" pixel images[2]; 9.9% emails contain regular images with either the recipient's email address or its MD5 or SHA1 hash value embedded; 13.3% 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 7 shows the distribution of emails in each domain category. Specifically, `News/Media` contributes the most emails (about 23%) among all domain categories. About 75% emails originate from the domains in the five categories, `News/Media`, `Shopping`, `Education`, `Social Networking`, and `Financial Services`.

[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.

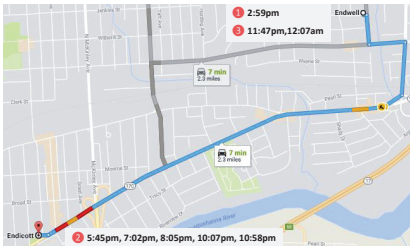

Fig. 4: Restoring the activities of Alice on Friday, July 14, 2017 from 2:59pm local time to 12:07am (midnight) with email tracking.

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

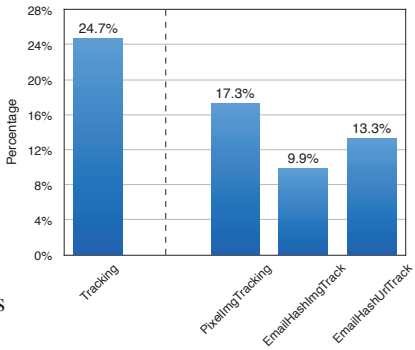

Fig. 6: Percentage of emails with tracking beacons embedded.

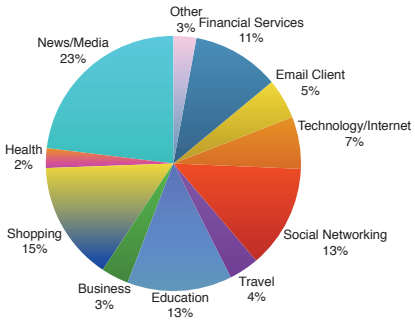

Fig. 7: Distribution of emails by domain category.

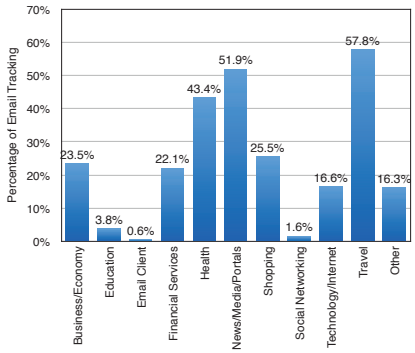

Fig. 8: Percentage of emails with tracking beacons by domain category.

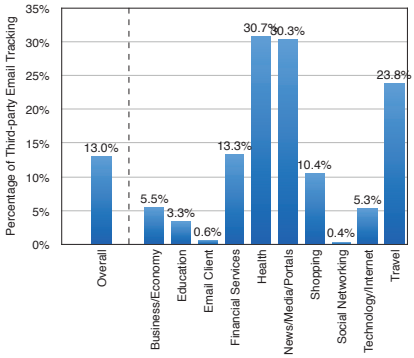

Fig. 9: Percentage of emails with third-party tracking.

The prevalence of email tracking is found to vary with the categories of email domains. As depicted in Figure 8, `Travel` and `News/Media/Portals` are the top two domain categories with high percentages of 57.8% and 51.9% emails equipped with tracking capability, respectively. More than 40% `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. In contrast, only 0.6% 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 which 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 case, the *email domain* is believed to be using third-party tracking. Figure 9 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 8, 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.

Next, we pay special attention to the third-party *tracking domains* which serve at least 2 unique *email domains* and actually provide email tracking services to email marketers. Based on both the number of email domains they are serving and the domain categories they are covering, the top 10 ETSes are determined and shown in Figure 10. 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 purpose. Figure 11 shows the CDF of the number of ETSes used per email domain. 29% 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 from 39 countries. Table II lists the demographic

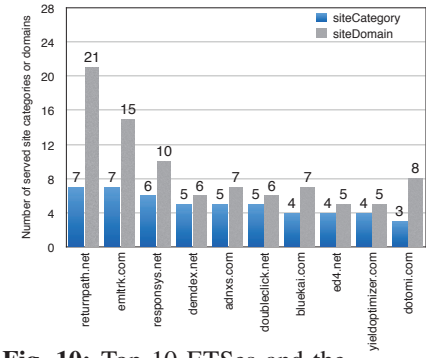

Fig. 10: Top 10 ETSes and the number of domain categories and unique domains they serve.

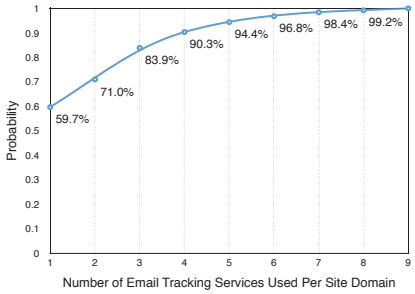

Fig. 11: CDF of the number of ETSes used per email domain.

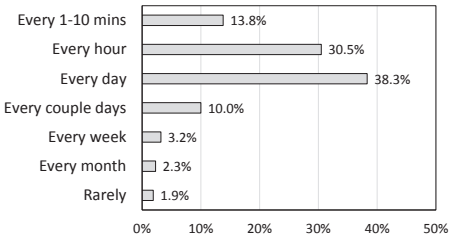

Fig. 12: Distribution of email checking frequency.

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% 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 12 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% users would read an email multiple times.

Willingness to read email or click email URL link. Email reading or clicking on the embedded URL link is typically required for email tracking to work. Figure 13 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% users often open the email; 13.5% and 28.9% 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. 24.3% and 37.3% users choose to never read such email or click on the link, respectively.

B. User Perception of Email Privacy Violations

When told that email reading could cause the recipients to be tracked, 52.1% participants had no awareness. We then measured user perception of possible email privacy violations and presented the results in Figure 14. 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 do not care about online privacy (ratings 1 and 2).

We then asked participants to rate for each kind of possibly disclosed information on a scale from 1 (no privacy concern) to 5 (extremely large privacy concern). As shown in Figure 15, the location information is what people are most concerned about and 57.1% of participants gave high rating scores of 4 and 5. Disclosing the device type being used also provokes much privacy concern. 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 propose a more suitable intervention approach.

A. Evaluating Existing Countermeasures Against Email Tracking

Existing web browsers, add-ons, and email clients have provided functionalities to protect users against email 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 want to block the tracking images rather than all images.

Some add-ons, such as the PixelBlock [17] for Google Chrome, are developed exclusively to tackle the email tracking problem. However, such tools usually prevent pixel-based email tracking with the help of a blacklist of well-known email

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%

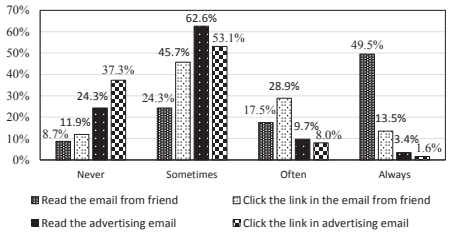

Fig. 13: Distribution of users' willingness of email opening and clicking the URLs.

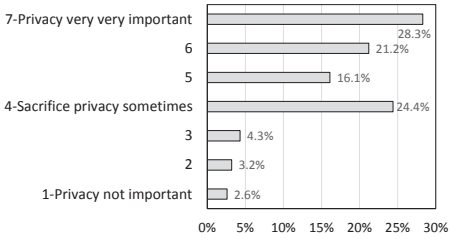

Fig. 14: Distribution of users' perception of privacy violations.

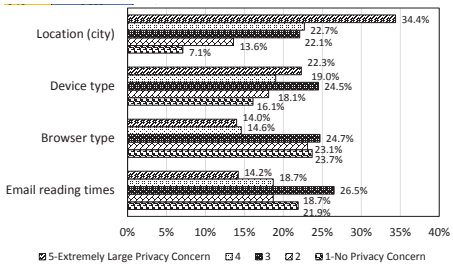

Fig. 15: Distribution of users' perception of the privacy level with information.

tracking services such as Mailtrack. One inherent 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, are already utilizing proxy servers to request the embedded images in an email on behalf of email recipients. Such practice can indeed hide users' IP addresses and user agents and thus prevent geolocalization and location tracking, but 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 Solutions

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 with email tracking. Nevertheless, we propose that several guidelines towards designing a practical tool to mitigate the privacy concerns.

First, email clients are in the best vantage point to counteract email tracking. As an entity responsible for accessing and managing a user's email, the email client is more suitable to take a non-intrusive way against email tracking. Otherwise, users have to take explicit efforts to survey various online tools which 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. Users have the right to know about everything that happens with their emails. When an email client detects a tracking beacon embedded within a new email, the user should be notified with either placing a small icon next to the email subject or highlighting the email with a predefined label.

Third, email clients need to provide more comprehensive preventions besides their state-of-the-art features (i.e., deploying proxy servers). As discussed above, proxy servers cannot prevent miscreants from inferring the email reading time and the number of times. One possible solution that email clients get rid of the embedded tracking beacons before the emails are read, with the acknowledgment from users.

Last, the defense mechanisms adopted by email clients for email tracking should provide more fine-grained opt-in features for users able 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 any one she designates as trusted.

VIII. Related Work

Privacy Concerns on Online Tracking. The online tracking [24], [30], e.g., the *Web Bugs* [19], [31] which leverages the invisible third-party images to track the page viewing, has been used in a variety of purposes, such as targeted advertisements, customized recommendations and search results, the analysis of user preference, and the surveillance of user activities. Meanwhile, the privacy concerns [36] raised by such tracking activities have received the attentions from research community.

Gross *et al.* [27] discussed the information revelation in social networks and related privacy issues. They pointed out possible attacks on privacy. They showed that most of the users are not aware of necessary privacy settings to prevent possible attacks. Goldfarb *et al.* [25] studied the impact of privacy regulation on web-based advertisement. They showed that the advertisers track consumer's behavior and browser history to deliver target Ads to consumers, and the consumers do not aware of the information collection process. Greengard

[26] explained how companies utilizes predictive analysis on patterns and behaviors of user buying and discussed the cookie-based tracking prevention. Mayer *et al.* [32] discussed the technologies and public policy related to the tracking activities of third-party web services. Subsequently, Datta *et al.* [23] developed an automated tool called "AdFisher" to explore the user behavior and advertisement interactions utilizing statistical methods. Melicher *et al.* [33] collected the browser histories and interviewed the 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 are more problematic in web surveillance since their URLs can be easily associated with user's email address. Also, HTML web bugs are normally based on the 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], [28], [29]. To the best of our knowledge, however, there is no study to examine the privacy risk of pixel based email tracking. Sipior *et al.* [35] discussed the legal issues regarding employee and employer related privacy issues in 1995. They also studied the U.S. legal system regarding protecting email privacy. However, the tracking for email activities was not mentioned in this study. According to a recent New York Times report [15], the legality of the email tracking is still unclear.

Preparing an effective email campaign is challenging due to privacy concerns. Cases *et al.* [22] studied the role of privacy, trust and customers attitude towards an email campaign. Narayanan *et al.* [34] considered that the email address may not be the "Personally Identifiable Information" (PII). Nevertheless, the email tracking may still cause privacy risk due to the exposure of email reading behavior. In addition, Zhao *et al.* [37] demonstrated the possibility of combining the email tracking with a phishing attack. In such case the attackers can craft much more sophisticated attacks based on email open rate. Moreover, based on open rate email subject can be modified into a successful bait.

IX. Conclusion

Email tracking collects privacy-sensitive user information and raises great privacy concern. In this work, we are the first to conduct such an in-depth and comprehensive evaluation of the privacy implications of email tracking. We developed a prototype email tracking system, performed real-world tracking, and demonstrated with realistic scenarios that the information divulged due to email tracking suffices for miscreants to mount a long-term stealthy surveillance attack. We estimated its real-world prevalence with a large-scale measurement study involving more than 44,000 email samples, and found that up to 24.7% emails track their recipients and some email domains adopt 9 different email tracking services to track email recipients. We examined users' privacy perception of email tracking with a crowdsourcing study, and found that more than a half of users have no awareness of email tracking and 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. We hope that this work could provide insights into shaping the further development of email tracking technologies for better user privacy protection.

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] Kate Murphy. Ways to Avoid Email Tracking. https://goo.gl/gDr1f3.
[16] Picture Blocker: Add-ons for Firefox. https://goo.gl/6WZCiu.
[17] PixelBlock - Chrome Web Store. https://goo.gl/VAhiX6.
[18] Email Statistics Report 2014-2018. https://goo.gl/TmeYoy.
[19] A. Alsaid and D. Martin. Detecting Web Bugs with Bugnosis: Privacy Advocacy Through Education, in *PETS*, 2002.
[20] K. Bajaj and J. Pieprzyk. A Case Study of User-Level Spam Filtering, in *AISC*, 2014.
[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] S. Englehardt and A. Narayanan. Online Tracking: A 1-million-site Measurement and Analysis, in *CCS*, 2016.
[25] A. Goldfarb and C. E. Tucker. Online Advertising, Behavioral Targeting, and Privacy, in *Communications of the ACM*, 2011.
[26] S. Greengard, Advertising Gets Personal, in *Communications of the ACM*, 2012.
[27] R. Gross and A. Acquisti. Information Revelation and Privacy in Online Social Networks, in *ACM Workshop on Privacy in the Electronic Society*, 2005.
[28] J. Jang, S. Nepal, and J. Zic. Trusted Email Protocol: Dealing with Privacy Concerns from Malicious Email Intermediaries, in *IEEE CIT*, 2008.
[29] D. Kafura, D. Gracanin, M. Perez-Quinones, T. DeHart, and S. Codio. An Approach to Community-Oriented Email Privacy. , in *PASSAT*, 2011.
[30] 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.
[31] D. Martin, H. Wu, and A. Alsaid. Hidden Surveillance by Web Sites: Web Bugs in Contemporary Use, in *Communication of the ACM*, 2003.
[32] J. R. Mayer and J. C. Mitchell. Third-Party Web Tracking: Policy and Technology, in *IEEE S&P*, 2012.
[33] 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.
[34] A. Narayanan, V. Shmatikov. Myths and Fallacies of "Personally Identifiable Information", in *Communications of the ACM*, 2010.
[35] J. C. Sipior and B. T. Ward. The Ethical and Legal Quandary of Email Privacy, in *Communications of the ACM*, 1995.
[36] S. Spiekermann, The Challenges of Privacy by Design, in *Communications of the ACM*, 2012.
[37] M. Zhao, B. An, and C. Kiekintveld. Optimizing Personalized Email Filtering Thresholds to Mitigate Sequential Spear Phishing Attacks, in *AAAI*, 2016.