Towards Transparency in Email Tracking

Max Maass, Stephan Schwär, and Matthias Hollick

Secure Mobile Networking Lab, TU Darmstadt, Darmstadt, Germany
{mmaass,sschwaer,mhollick}@seemoo.tu-darmstadt.de

Abstract. Tracking technologies have become ubiquitous, not only on websites but also in email messages. However, while protection and transparency tools exist for the web, no such tools exist for email messages, thus obscuring privacy violations. We introduce the PrivacyMail platform to assist with the automated analysis of email messages. The platform automatically analyzes commercial mailing lists, making it easier to detect different forms of tracking. Our platform introduces transparency about the practices of companies, and serves as a tool for regulators, data protection professionals and consumers alike. Our preliminary results show widespread email tracking, where opening an email can result in information being sent to up to 13 third parties, in some cases disclosing the users' email address in the process.

Keywords: Scanner · Tracking · Compliance · Email · Privacy.

1 Introduction

While discussions about tracking on websites have entered the mainstream, one issue that has received far less attention is the prevalence of tracking in email communication. Here, a large ecosystem of commercial tracking companies offers services that allow marketing professionals and private individuals alike to monitor if their emails are being viewed and which of their links are being clicked. The used techniques include *tracking pixels* and personalized links, which will in some cases leak the email addresses of the affected users to third parties [4, 7]. At the same time, fewer protections for end-users exist—while web tracking can be countered to a certain degree by using ad-blockers and tracking protection systems such as PrivacyBadger [3], no such tools exist to protect against email tracking. This problem is exacerbated by the fact that emails are often being opened repeatedly on multiple devices, using different clients (Webmail, Thunderbird, Outlook, iOS Mail, ...), which allows trackers to link these devices to the same owner, and makes defense more difficult, as every client needs to be protected separately. There are few technologies providing transparency in this space, leading to a lack of awareness about the tracking practices of commercial mailing services.

Previous studies have sought to quantify the prevalence of tracking in commercial emails through a variety of methods [4, 6, 7], and investigated the potential privacy implications and user acceptance of these methods [12]. However, so far, they can only provide an aggregate analysis at a specific point in time.

Author version. The final authenticated version is available online at https://doi.org/10.1007/978-3-030-21752-5_2.

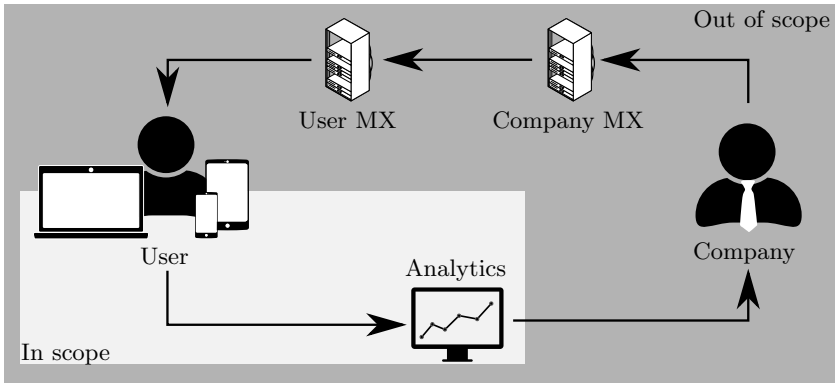

Fig. 1. Potential areas of privacy concern.

Their ability to provide public transparency about the practices of individual companies over time, a common approach in the area of online tracking [1, 9], is thus limited.

We aim to fill this gap by designing and developing a **public email privacy benchmarking system called PrivacyMail**. The system allows anyone to register special email addresses for commercial mailing lists, and will analyze incoming emails for common tracking techniques. It will also attempt to detect the disclosure of personally identifiable information (PII), like the email address, to third parties. This information can be used by *data protection officers (DPOs)* to check the compliance of companies with relevant regulation, by *individual users* to inform themselves about the risks of subscribing to mailings from specific companies, and by *researchers* to gain more insight into the practices of a large, crowd-sourced set of companies that send out these mailings. A beta version is available at https://PrivacyMail.info.

We will proceed by reviewing related work in Section 2 and providing an overview about our system in Section 3. In Section 4, we discuss preliminary results from the operation of an early version on a limited dataset to demonstrate the capabilities of the current prototype. We will close the paper by discussing future work in Section 5 before concluding in Section 6.

2 Related Work

Research on email privacy can be split into two areas (cf. Figure 1): privacy against intermediaries (like email providers or third parties eavesdropping on communication between the mail exchanges (MXs)), and privacy against tracking by the sender, which is the topic of this paper. Privacy against intermediaries can be ensured through transport- and end-to-end encryption, and has been studied in some detail (cf. [2] for an overview). In contrast, privacy against the sender has not received a lot of attention so far.

Englehardt *et al.* developed the system that serves as the conceptual basis for our own. They used OpenWPM [5] to scan a dataset consisting of over 12 500 emails from 902 different senders (a mix of popular shopping and news websites) [4]. They found that 85% of emails contained at least one embedded image from a third party, and 19% of senders contained embedded external content that leaked the email address of the recipient to a third party (by encoding it in the URL, cf. Section 3). They also found that repeatedly opening the same email changed which third parties were embedded in 21% of the cases. Finally, they showed that existing tracker blocking lists, designed for use against website-based trackers, missed a significant portion of third parties commonly embedded in emails.

Xu *et al.* [12] analyzed a corpus of over 44 000 emails, collected over a period of 7 years, and found widespread use of tracking in a large variety of different sectors. They also investigated the potential privacy implications of email tracking, and found that sending a small number of emails is sufficient to track some users for several weeks, including their geographical location. Finally, they performed a user study and found that users are generally unaware of the privacy risks of email tracking, and a vast majority of users were interested in protecting themselves from it after they learned of its existence.

Hu *et al.* [7] analyzed a large corpus of emails collected from disposable email services. They crawled public mailboxes of several popular providers of anonymous, temporary email accounts and collected a corpus of 2.3 million messages from over 200 000 distinct domains. They again confirmed that email tracking is a common practice, and is disproportionately used by large companies. They also found that the market for email tracking is not yet dominated by a single company.

Haupt *et al.* [6] collected a dataset of over 60 000 emails from the newsletters of a variety of different companies. They investigated the properties of tracking images, and proposed an automated approach to detect and block them using a machine learning classifier, achieving a detection rate of 92%.

All of these studies have in common that they provide only an aggregate analysis of a snapshot of the current state of email newsletters, thus making it impossible to draw conclusions about an individual users' exposure to tracking. Such an analysis for the area of web privacy is being offered by two projects: *Webbkoll* [1] and our own prior work, *PrivacyScore.org* [9]. Both perform automated scans of websites to determine their privacy properties, and PrivacyScore also seeks to create public transparency about the practices of website operators to incentivize them to change their behavior [8]. To the best of our knowledge, aside from the platform developed in this paper, no similar system exists in the domain of email messages.

3 System Overview

In this section, we give an overview of our system, the analyses we perform, and the challenges we encountered. The platform is built using Python and the Django framework. The overall process of using PrivacyMail is shown in Figure 2.

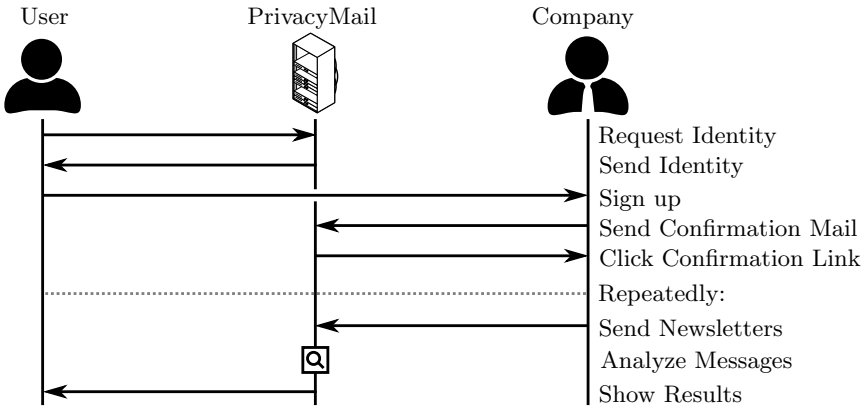

Fig. 2. Usage of the PrivacyMail platform.

3.1 Adding a service

Any service that sends out newsletters can be registered with the system by entering its URL into the system. PrivacyMail will generate a unique identity with an email address (hosted by PrivacyMail), name, and gender (as some newsletter providers ask for this upon registration), and display it to the user performing the registration. The user will then enter that email and other required information into the newsletter sign-up form. The resulting email confirmation will be received by PrivacyMail.

The user will also be invited to add additional metadata about the service. This includes a canonical name (e.g. "Spiegel Online" for spiegel.de, or "Annual Privacy Forum" for privacyforum.eu) to facilitate a search using human-readable terms, and information about the country and industry sector of the website. This metadata can later be used for further analyses.

Each new identity must be manually confirmed by an administrator, and no automated processing takes place until then. This ensures that the email address was signed up at the correct website. If everything is in order, the administrator will confirm the registration by clicking the email confirmation link. Any future emails from the sending domain will be automatically processed without human interaction.

3.2 Analyzing emails

When a new email from a permitted sender for a confirmed identity arrives, it is automatically processed. First, the email is saved to the database, including all relevant headers. Next, all external links (but not the embedded external resources, like images) are extracted from the email. The system attempts to deduce which of the detected links are management links (e.g., links to change subscription settings or view the email in the browser), and which are regular content links (e.g., links to news articles, products, etc.), using a mix of heuristics,

including word lists and link clustering. Once all likely management links have been excluded, the system chooses one of the remaining links and marks it for later investigation.

External resource analysis: Email tracking is usually performed with one of two goals: The sender wants to determine if the recipient opened the email, and/or if she clicked any links embedded in the email. Depending on the goal, different techniques are used. Commonly, a small image hosted by a tracking company is embedded in the email, using a personalized URL that can be linked to the recipient. Upon opening the message, most email clients[1] will automatically load this external resource from the servers of the company, thereby notifying the company that the email was opened by this specific recipient. As this *tracking pixel* is too small to be seen, the user will not notice its presence in the email. Alternatively, the same technique can be used with visible images (e.g., product photos in an ecommerce newsletter). Such requests to tracking providers not only inform them about the fact that the email was opened, but also leak the IP address and user agent (i.e., the used browser or email client) to them. This information can be used to obtain a (coarse) geographic location of the user [12].

More than one tracker can be included using a single tracking pixel through the use of HTTP forwarding. In this case, the first tracking service will forward the user to a second one, which forwards the user even further, until all desired tracking services have been informed and the request is answered by the final destination (i.e., the 1x1 pixel image). This allows an unlimited number of trackers to be included using a single image.

To detect this tracking, we save the message to an HTML file and host it on a machine-local web server. This allows us to view it with OpenWPM, an automated Firefox browser intended for research [5]. Viewing the email like this approximates opening it in a webmail system with remote content enabled. OpenWPM will log all requests and responses generated by viewing the email, thus giving us an accurate representation of what will happen when a user views this email without clicking any links. Using this (instead of a static analysis of embedded external content) allows us to see not only the embedded external trackers, but also any additional trackers contacted through HTTP redirects. All requests and responses and the relations between them are saved in the database.

Link analysis: If the sender wants to know if links from the email were clicked (e.g., to judge the click-through rate of advertising campaigns), they can also personalize the links. In this case, the links will point to a special URL, hosted by the tracking service, which will log the visit and forward the user to the actual target of the link (e.g., a product or news article). This tracking only becomes active when the user clicks the link, but cannot be prevented by not loading external resources. Again, more than one tracker can be informed through HTTP forwarding.

To detect this tracking, we delete the local state (cookies, sessions, ...) of the OpenWPM browser and instruct it to visit the link we have previously selected

[1] See Table 12 of [4] for an overview of eMail client behavior.

in the email. Again, we log all requests and responses and identify the chain of HTTP redirects that takes place when visiting the link, until the final destination is reached.

Email disclosure analysis: Trackers use different techniques to identify email recipients in these links, however, identifiers derived from the email address are common. Previous work has shown that in many cases, hashes or encoded versions of the email address are used by tracking services [4, 7], in some cases nesting different encodings or hash algorithms (e.g., `md5(sha1(email))`). This shows that the email addresses of recipients are widely shared with third parties, either intentionally by the sender of the newsletter, or implicitly by the tracking services. Previous work has shown that simple hashing of such personally-identifiable information is insufficient to guarantee privacy [10].

To detect this eMail leakage, we compute a series of hashes and encodings of the address, nested to a depth of 2, and check if any of them are found in any of the recorded request URLs for the eMail. If so, we assume that this request discloses the email address, and save this fact in the database. After this, processing of the email is finished.

Further personalization detection: Not all personalization uses identifier derived from the email address. Users may be identified by a different identifier that is linked to their identity on the server. To detect this type of personalization, we offer the option to register more than one identity per service. The system then uses a combination of email timestamps and subject lines to match newsletter messages between different identities. Once a pair has been found, the links are extracted from both and compared. If no personalization is used, the links in both messages should be identical when excluding subscription management links. Thus, if (partially) different links are detected, this is a strong indicator that they are personalized.

Another possibility for differing links may be the use of A/B testing, in which different versions of emails are sent out to recipients to determine which headlines are more effective at generating clicks. These practices have been observed by Englehardt *et al.* [4]. To distinguish A/B testing from other forms of personalization, we also compare the text of the messages to see how similar they are. A high similarity indicates that the same message was sent to both identities, while a low similarity indicates A/B testing.

Further analyses: Having a large archive of emails, both for a single service over time and for a large, crowdsourced collection of different services, will also allow us to perform additional analyses. For example, does the number of trackers increase or decrease over time? What is the influence of regulatory changes like the upcoming ePrivacy directive? For services annotated with additional metadata, we can compare tracking practices between countries and industry sectors, where Haupt *et al.* found significant differences [6].

3.3 Providing Transparency

The results for all newsletters are made available using a searchable frontend on the project website (currently in development). This allows users to check if the newsletter they are interested in has already been analyzed, and if so, which trackers it uses and to which the email addresses are disclosed. We do not republish the content of newsletters, only the results of our analysis, to avoid allegations of copyright infringement.

3.4 Challenges

One concern is the handling of identities that receive emails from sources that are not affiliated with the original newsletter provider. This could be spam (e.g. due to a data breach at the newsletter provider), or due to a user registering the generated identity with more than one website. We solve part of this problem by only processing emails that come from an approved sender for the identity. emails not sent by the expected sender are held back for manual verification, at which point a decision can be made on how to handle them (e.g., set the sender as a new approved sender, mark the message as spam, or discard it).

The processing time for a single email is on the order of several seconds to half a minute. This makes the analysis a bottleneck for the performance of the system. We are already working on distributing this work to enable PrivacyMail to scale horizontally with demand.

Finally, service providers may not want their newsletters to be analyzed. As we would like to avoid unilateral action from the service providers (i.e., identifying and unsubscribing identities linked to PrivacyMail based on the used email domains), we provide them with the option to opt out of being analyzed by contacting us. To make this transparent to the users, their services will then be listed as *excluded from analysis*.

4 Preliminary Results

To demonstrate the capabilities of the current prototype, we performed a small-scale analysis on a non-representative dataset, obtained by signing up for newsletters from 20 ecommerce and news websites in Germany, the United States, France, Italy and Poland. They were chosen partially based on popularity, partially on personal familiarity, and not informed about the analysis. In total, the dataset contains approximately 2000 emails. More detailed analyses on a larger set of services will be presented at the Annual Privacy Forum.

16 of 20 companies (80%) sent emails containing at least one resource hosted by a 3rd party (i.e., a domain not directly associated with the sending company), with an average of 118.4 resources per email (median 111, min 0, max 363). These may represent tracking, but also more benign purposes, such as the use of content distribution networks (CDNs) to host article pictures.[2] In total, 43

[2] Differentiating between these cases automatically is challenging, as standard tracker blocking lists have been shown to be unreliable when applied to email tracking [4].

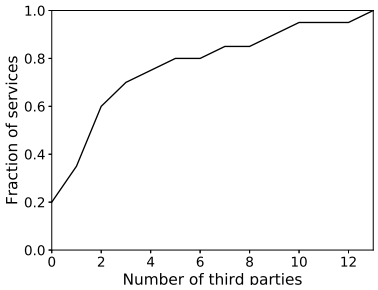

Fig. 3. CDF of third party count per service ($n = 20$).

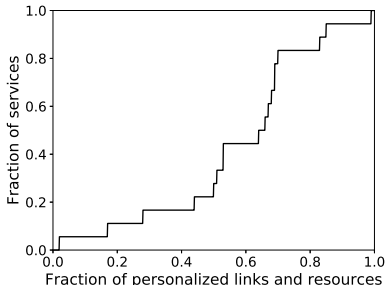

Fig. 4. CDF of fraction of personalized links per service ($n = 18$).

distinct 3rd party domains were contacted, with an average of 1.56 third parties per email (median 1, min 0, max 13, cf. Figure 3 for the cumulative distribution function (CDF)).

When opening the emails, at least 4 out of 20 services (20%) leak the email address of recipient to at least one website (including their own), and 13 different websites receive them from at least one service, often hashed using the md5 algorithm. Some of the receiving websites belong to the company sending the newsletter, while at least 9 of them belong to tracking companies, many of them located in the United States.

4.1 Case Study 1: Individual service analysis

For our first case study, we examine the daily newsletter sent by a major French newspaper. When opening one email, mail clients that load external content will access 70 external resources, 24 of which are loaded from 3rd party domains, including a French tracking company and an advertising subsidiary of Google. Some requests are also forwarded to additional external 3rd parties, leading to the inclusion of another company. Interestingly, some requests are forwarded to the local machine of the user (`http://localhost/`), which may be either due to a misconfiguration, or used to ensure that no content is loaded.

22 of the embedded external resources contain the md5-hashed email address of the recipient, which is sent to the website of the newspaper and forwarded to a 3rd party, `ivitrack.com`.[3] The URLs also contain what is likely a message identifier, indicating that this is used to track which users have actually read the newsletter.

When clicking the link to a news article, the user is forwarded to a subdomain of the newspaper, which forwards the user via the same French tracking company that was previous included through embedded images. This is likely used to

[3] We were unable to find details about this company, but hashed email address leaks to this company have also been observed by Englehardt *et al.* [4].

track which links are being opened by users, although it does not carry any user identities derived from the email address.

4.2 Case Study 2: A/B testing

Our second and third case study show the added possibilities enabled by having more than one recipient for each newsletter. Comparing emails sent by the same service to different recipients allows us to detect if the service is performing A/B testing (cf. Section 3.2). In our dataset, we observed two services performing A/B testing, both of them German shopping companies. Each used a base email with a set of products or banners common to both emails, with one being extended with additional banners or product offers. Due to the probabilistic nature of A/B testing, these numbers should be considered a lower bound. The confidence can be increased by adding additional identities to the service under test.

4.3 Case Study 3: Link Personalization

We also compare the links sent to different identities registered for the same newsletter to detect tracking identifiers that are not derived from the email address. In our dataset, we observed different degrees of personalization. In some newsletters, almost all links and external resources were personalized, some only personalized links to their homepage, but not to individual articles from the newsletter. Only one service in the dataset used personalization only for the subscription management links. The CDF of the degree of personalization is shown in Figure 4.

5 Future Work

Over the coming weeks we plan to include more analyses in the platform, and expose them in the frontend. We will enhance security and performance by using a distributed and containerized system for the analysis of the emails, and potentially switching from OpenWPM [5] to Privacyscanner [11]. Finally, we would like to discuss the feature wishes and requirements of practitioners in the field at the conference, and incorporate them to make the platform more useful for their purposes.

6 Conclusion

In this paper, we presented *PrivacyMail*. Similar to our PrivacyScore platform [9], we aim to shine a light on a type of privacy invasion that has traditionally been invisible. To facilitate this we designed a system that automatically analyses emails for tracking and personalization, and presented an example evaluation of a small set of services, finding evidence of email address leakage, tracking through personalized links, and A/B testing.

The platform is intended to be a public resource. Anyone can add new services to be analyzed, and the results will be made publicly available on the project homepage. The platform is available at `https://PrivacyMail.info`, and the source code will be released under an open license. By providing transparency, we hope to inform end users about the privacy impact of the newsletters they consume, support data protection professionals in their task of testing companies for compliance with relevant regulation, and provide interesting datasets for future research.

Acknowledgements

This work has been co-funded by the DFG as part of project C.1 within the RTG 2050 "Privacy and Trust for Mobile Users".

References

1. Andersdotter, A., Jensen-Urstad, A.: Evaluating Websites and Their Adherence to Data Protection Principles: Tools and Experiences. In: IFIP Advances in Information and Communication Technology, vol. 498, pp. 39–51 (2016)
2. Clark, J., van Oorschot, P.C., Ruoti, S., Seamons, K., Zappala, D.: Securing Email. ArXiv preprint (2018), http://arxiv.org/abs/1804.07706
3. Electronic Frontier Foundation: PrivacyBadger, https://eff.org/privacybadger, [accessed 2019-01-28]
4. Englehardt, S., Han, J., Narayanan, A.: I never signed up for this! Privacy implications of email tracking. Proceedings on Privacy Enhancing Technologies **2018**(1), 109–126 (2018)
5. Englehardt, S., Narayanan, A.: Online Tracking: A 1-million-site Measurement and Analysis. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16. pp. 1388–1401. No. 1, ACM Press, New York, New York, USA (2016)
6. Haupt, J., Bender, B., Fabian, B., Lessmann, S.: Robust identification of email tracking: A machine learning approach. European Journal of Operational Research **271**(1), 341–356 (2018)
7. Hu, H., Peng, P., Wang, G.: Characterizing Pixel Tracking through the Lens of Disposable Email Services. In: IEEE Security & Privacy 2019. pp. 545–559 (2019)
8. Maass, M., Walter, N., Herrmann, D., Hollick, M.: On the Difficulties of Incentivizing Online Privacy through Transparency: A Qualitative Survey of the German Health Insurance Market. In: 14. Internationale Tagung Wirtschaftsinformatik (2019)
9. Maass, M., Wichmann, P., Pridöhl, H., Herrmann, D.: PrivacyScore: Improving privacy and security via crowd-sourced benchmarks of websites. Lecture Notes in Computer Science **10518 LNCS**, 178–191 (2017)
10. Marx, M., Zimer, E., Mueller, T., Blochberger, M., Federrath, H.: Hashing of personally identifiable information is not sufficient. In: Sicherheit 2018. pp. 55–68 (2018)
11. Pridöhl, H.: Privacyscanner. https://doi.org/10.5281/zenodo.2555037, https://github.com/PrivacyScore/Privacyscanner, [accessed 2019-01-28]
12. Xu, H., Hao, S., Sari, A., Wang, H.: Privacy Risk Assessment on Email Tracking. In: IEEE INFOCOM. vol. 2018-April, pp. 2519–2527 (2018)