E-Mail Tracking: Status Quo and Novel Countermeasures

Completed Research Paper

Benedict Bender
Chair of Business Informatics, esp. Processes and Systems
University of Potsdam
August-Bebel-Str. 89,
14482 Potsdam, Germany
benedict.bender@wi.uni-potsdam.de

Benjamin Fabian
Chair of Business Intelligence and Data Science, Hochschule für Telekommunikation Leipzig (HfTL)
Gustav-Freytag-Str. 43-45,
04277 Leipzig, Germany
fabian@hft-leipzig.de

Stefan Lessmann
Chair of Information Systems
Humboldt University of Berlin
Spandauer Str. 1,
10178 Berlin, Germany
stefan.lessmann@hu-berlin.de

Johannes Haupt
Chair of Information Systems
Humboldt University of Berlin
Spandauer Str. 1,
10178 Berlin, Germany
johannes.haupt@hu-berlin.de

Abstract

E-mail advertisement, as one instrument in the marketing mix, allows companies to collect fine-grained behavioural data about individual users' e-mail reading habits realised through sophisticated tracking mechanisms. Such tracking can be harmful for user privacy and security. This problem is especially severe since e-mail tracking techniques gather data without user consent. Striving to increase privacy and security in e-mail communication, the paper makes three contributions. First, a large database of newsletter e-mails is developed. This data facilitates investigating the prevalence of e-mail tracking among 300 global enterprises from Germany, the United Kingdom and the United States. Second, countermeasures are developed for automatically identifying and blocking e-mail tracking mechanisms without impeding the user experience. The approach consists of identifying important tracking descriptors and creating a neural network-based detection model. Last, the effectiveness of the proposed approach is established by means of empirical experimentation. The results suggest a classification accuracy of 99.99%.

Keywords: E-Mail Tracking, Countermeasures, Privacy, Security, Machine Learning

Introduction

Data about e-mail reading behaviour can be used to infer valuable commercial information. In marketing, for example, it allows user preferences to be derived and the reach and effectiveness of e-mail marketing campaigns to be measured (Hasouneh and Alqeed 2010). Contemporary e-mail tracking techniques enable the sender to track how often an e-mail is read, which device the recipient uses, and the time as well as location from which the e-mail is read (Fabian et al. 2015). Importantly, this information is typically gathered without the recipient's consent or acknowledgement. Tracking can also constitute a security threat. Spammers and hackers commonly rely on e-mail tracking to detect and collect active e-mail addresses for their illegal activities. From an end-user perspective, e-mail tracking procedures therefore involve various security and privacy issues.

Mail users should be equipped with effective and reliable protection methods. In order to provide advancements towards user privacy and security protection, this paper follows the design science research paradigm (Peffers et al. 2007). The study starts with a survey of relevant literature and proposes a definition for e-mail tracking. Following that, e-mail tracking technology is explained. Further contributions involve the experimental analysis of information that can be gathered using e-mail tracking, and a critical comparison of currently available protection measures. Then, with regard to problem identification and relevance, our paper presents a large empirical study, confirming e-mail tracking as an important and widespread privacy issue.

This motivates another major contribution of our research: the design of countermeasures, encompassing the development of a novel method for tracking-image identification that is based on machine learning. A demonstration and evaluation are realised through a quantitative and empirical evaluation of the developed detection model based on a large dataset of over 4,500 mails from 300 global companies, including more than 110,000 images. This article will serve to communicate our results.

Definition and Related Work

E-mail tracking and its impact on privacy are often mentioned in the general press, for example in conjunction with scandals that have been uncovered using e-mail tracking technologies (Evers 2006). As far as the academic literature is concerned, surprisingly few papers have looked into the topic (Bonfrer and Drèze 2009; Fabian et al., 2015; Hasouneh and Alqeed 2010) and these do not focus on countermeasures for e-mail tracking. Some initiatives to develop anti-tracking software have been undertaken in corporate practice. However, as we show below, they do not provide sufficient protection. The lack of effective countermeasures motivates this research, which emphasises the technical and process-related aspects of e-mail tracking. In accordance with this focus and with inspiration from Fabian et al. (Fabian et al. 2015), we propose the following definition of the term e-mail tracking: *E-mail tracking allows mail senders to gather information on an individual recipient's reading behaviour of single mails without the need for any further interaction or the recipient's permission*.

Some characteristics of the definition deserve further clarification. *Individual recipient*: Relevant techniques allow the gathering of information on the individual recipient's behaviour. This is important in order to distinguish e-mail tracking procedures from general aggregated traffic measuring techniques. *Reading behaviour*: The minimum requirement is that a technique provides information about whether a single mail has been opened by a specific recipient. *Single mails*: To fulfil the requirement of marketers or other trackers, a tracking mechanism provides information on the level of single mails. In combination with the *individual recipient* requirement, this allows trackers to infer the reading behaviour of every recipient for every mail that was issued. *Without any further interaction*: This emphasises methods that do not require further user actions than simply opening an e-mail. One technical implication of this understanding is that we concentrate on tracking pixels but not on tracking links which users need to click on (Fabian et al., 2015). The important aspect is that simply opening the mail is sufficient to trigger the tracking mechanism. *Without recipient's permission*: This emphasises the fact that the mechanism does not require any acknowledgement of the recipient; the technique therefore distinguishes itself from functions such as mail return receipts. This characteristic involves possibilities of secret surveillance.

From a technological point of view, e-mail tracking can be understood as an adaptation of web tracking mechanisms to HTML-based e-mails. Unlike e-mail tracking, web-tracking mechanisms have received

much attention in the literature. The *use* of web tracking in different situations (Javed 2013; Jensen et al. 2007) as well as their *detection* (Alsaid and Martin 2003; Fonseca et al. 2005) have been analysed. *Prevention* of such mechanisms, including evaluation of software solutions, has also been the topic of studies (Fonseca et al. 2005; Leon et al. 2012). Other research emphasises the technical aspects of web tracking, such as different categories of web bugs (Dobias 2011) or the potential for aggregating multiple server log files (Evans et al. 2003). Yet another stream of research aims at supporting website operators through reviewing and developing criteria for web tracking software selection (Fourie and Bothma 2007; Nakatani and Chuang 2011) or, more generally, evaluating the market for web tracking software (Krishnamurthy and Wills 2009).

Some web tracking papers hint at the possibility of applying tracking techniques to HTML-based e-mails (Bouguettaya and Eltoweissy 2003; Harding et al. 2001; Martin et al. 2003; Moscato and Moscato 2009; Moscato et al. 2013). However, none of these studies provide further details of such an undertaking or discusses the peculiarities of e-mail tracking. Clearly, some similarity between web and e-mail tracking mechanisms exists, especially in terms of tracking technology. From an organisational point of view, a similarity may also be seen in the fact that the tracking infrastructure can be operated by the company or a contracted service provider, though in-house solutions seem to be extremely scarce in web tracking contexts (Burkell and Fortier 2013; Sipior et al. 2011; Waisberg and Kaushik 2009). However, one important difference concerns tracking precision. With e-mail tracking, any information can be easily linked to the user's e-mail address, which is an almost unique identifier of the user. Consequently, tracking users across devices, locations, channels etc. is much easier compared to web tracking, which heavily depends on the browser environment and its configuration. In this sense, e-mail tracking can be considered even more privacy intrusive, which further supports the need for effective countermeasures.

E-Mail Tracking Technology

To give an overview of e-mail tracking methodology and the degree to which it impedes user privacy, the following sections review the e-mail tracking process and detail how and which information is captured about mail recipients. This provides a foundation for developing effective countermeasures.

E-Mail Tracking Process

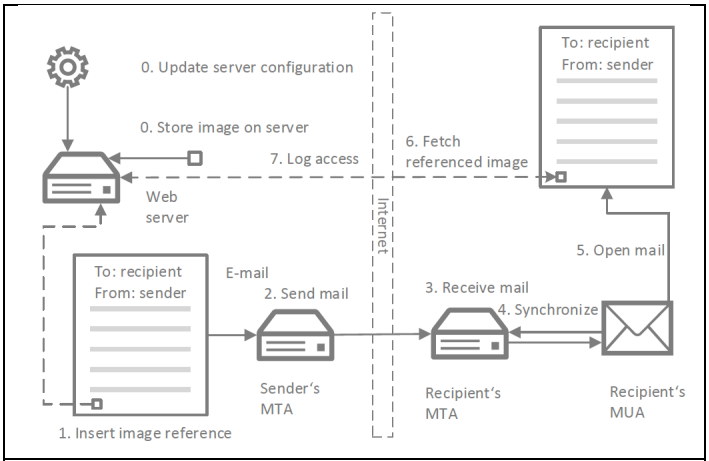

Figure 1: E-Mail Tracking Operation Mode

The tracking process (Figure 1) is based on e-mails that reference external resources. Therefore, it starts with the preparation of an HTML-based e-mail by the sender, since plain-text e-mails do not facilitate such references. This e-mail, which includes a tracking-image reference, passes several mail transfer agents (MTAs) until reaching the receiver's MTA. Next, the recipient opens a mail client, which synchronises the local mail repository with the newest version of the recipient's MTA. When the recipient opens the e-mail with a tracking image, the mail client requests the image from the referenced

destination. The web server logs this request and provides the image to the recipient's client. Afterwards, log analysis allows information to be deducted on the recipient's e-mail reading behaviour. For example, if the e-mail is opened on different devices, every individual access is logged, which allows for cross-device tracking.

Even though the structure of tracking image references varies, the following anonymised reference serves as an example of tracking image references: http://www.example.com/action/view/3827/rtg2ryw3.

Information Gathered by E-Mail Tracking

To elaborate on the collection of behavioural data via e-mail tracking, we distinguish between primary and secondary information. The former is directly extracted from web server access logs, whereas the latter can be derived from combining primary information with auxiliary data sources.

In order to assess the extent of primary information available to a tracker, we constructed a prototypical tracking environment, which includes an Apache webserver to log data relevant to e-mail tracking. The entries in the server log file provide seven major pieces of information: (1) the Internet Protocol (IP) address of the host that requests the image file, (2) the date and time of the file request, (3) the request itself, which includes the URL and GET variables, (4) the status code of the request, (5) the amount of bytes that have been sent in response, (6) the referrer URL from the client, and (7) a string characterising the user agent. Furthermore, when a file is requested multiple times (i.e., it generates multiple entries), it allows information to be derived with respect to a user's reading behaviour. In our test environment, a new log entry was created every time an e-mail was opened.

With respect to secondary information, the first possibility is to induce the fact that the user read or at least opened the e-mail. As we show in more detail below, this follows from the fact that current e-mail clients do not download images before the corresponding e-mail is opened. Accordingly, the existence of multiple log entries allows for the conclusion that the e-mail has been opened multiple times. The combination of multiple entries for one mail, as well as multiple entries from one user for different mails, provides insight into the recipient's e-mail reading behaviour.

Furthermore, possibilities exist for identifying whether an e-mail has been forwarded. The usage of IP geolocation in combination with a log entry aggregation allows the detection of forwarded mails. HP, for example, used this technique to investigate the release of confidential information (Evers 2006). Special log entries allow one to determine whether an e-mail has been printed (Campaign Monitor 2010). It is also possible to gather information about the user environment by analysing the user agent string, which is part of a log entry (Agosti and Di Nunzio 2007). Location-related information can be gathered using geolocation services (Poese et al. 2011). Based on a reverse lookup of an IP address, a log entry may also help determine a user's affiliation to a company or institution. These examples illustrate only some options for trackers and the potential for gaining insights into user behaviour through combining and correlating tracking data with external information.

International Study on E-Mail Tracking Usage

Having established the intrusiveness of e-mail tracking, a relevant follow-up question concerns the prevalence of tracking mechanisms. To answer this, we collected a unique empirical dataset of e-mail newsletters, which allowed us to evaluate the status quo of e-mail tracking. Although potentially not representative of companies' marketing communication in general, newsletter e-mails are a suitable vehicle for this analysis. First, the wide availability of different newsletters simplifies systematic data collection and facilitates the gathering of a large amount of data. Second, it seems likely that companies use e-mail tracking to assess the effectiveness of their newsletters (Hasouneh and Alqeed 2010). To further increase this likelihood, we concentrate on larger companies because these are on average faster to adopt novel technology (Premkumar and Roberts 1999). Contrary to Fabian et al. who performed a comparable analysis among 64 German companies (Fabian et al. 2015), we adopted an international scope and gathered e-mail newsletters from the top-100 companies (ranked by revenue) in Germany (GER), Great Britain (GB), and the United States (USA).

To gather the data, we create two identities and corresponding e-mail addresses using Gmail. With each account, we signed up for the newsletters of the pre-selected 300 companies and collected e-mails in a 13-

week period (calendar week 22-34) in 2015. To identify tracking elements, we compared the e-mails received on each account. That is, we examined the HTML content of each pair of e-mails sent to matched accounts and searched for deviations in image URLs. To obtain ground truth data, we classified images for which the referral URLs do not match as tracking image, and all others as non-tracking. Last, we manually checked all tracking elements to avoid classification errors. However, to avoid bias from senders changing their e-mail policy in response to the reading behaviour of users they were tracking, we ensured that none of the external images were actually requested from the web server.

In total, each artificial identity received 4,505 e-mails, 1,442 (32%) of which came from Germany, 1,489 (33%) from the United Kingdom (UK), and 1,428 (32%) from the United States (US). The remainder, referred to as *various* below, consisted of e-mails sent from multiple countries. This usually applies if externally contracted third parties send mails for several clients in different countries from the same address.

General Statistics

The fraction of HTML compared to plain text e-mails is interesting since only HTML-based e-mails facilitate tracking. Out of 4,505 e-mails, 1,266 (28%) were in plain-text format, while the remaining 3,239 (72%) were HTML-based. Considering the HTML e-mails, 2,920 (90%) contained external image references. These e-mails could facilitate tracking. The HTML e-mails contained references to 110,080 external images, with an average of 38 external images per e-mail. 18% of the e-mails contained a single external image. Figure 2 gives an overview of the key measures relevant for e-mail tracking: 2,292 e-mails contained tracking elements, which equated to a ratio of 51% (71%) among all e-mails (HTML e-mails).

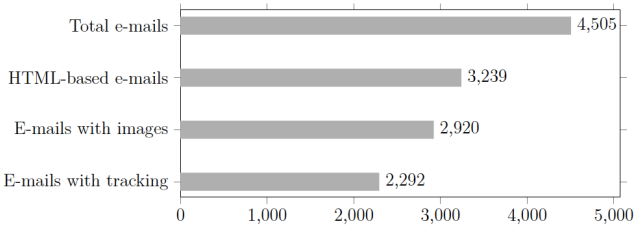

Figure 2: Dataset Overview: Tracking / Non-Tracking

Our results reveal that the tracking quote and the fraction of HTML-based e-mails varied across countries (Figure 3 & Table 1). The proportion of HTML e-mails was 88% (1,375), 34% (513), and 84% (1,205) for Germany, the UK, and the US, respectively. Concentrating on these e-mails, the tracking quote was the highest in Germany (88%), and rather similar for the US (63%) and the UK (62%).

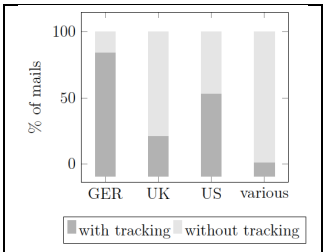

Figure 3: Tracking Elements per Country

Table 1: Tracking Elements per Country

Country	Tracking elements				Total
	0	1	2	3	
GER	231	1206	5		1442
UK	1173	300	16		1489
US	665	636	107	20	1428
Various	144	2			146
Total	2213	2144	128	20	4505

Internal and External Tracking

The main options for performing e-mail tracking are in-house systems and contracted service providers, the latter of which prevail in web tracking. To shed light on the frequency of the two options in e-mail tracking, we differentiated between internal and external tracking. We defined tracking as internal if the web server hosting the tracking image belonged to the company that sent the newsletter, and as external otherwise. To handle ambiguous cases, we defined a third category, 'various', which subsumed e-mails with multiple tracking images from internal and external web servers. Figure 4 and Table 2 show the results of this analysis. It reveals that Germany had the highest internal tracking rate at 86%, followed by the UK with 60%, and the US with 34%. Only mails from the US used internal and external tracking at the same time.

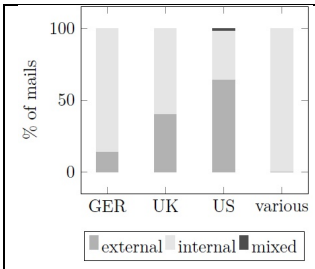

Figure 4: Tracking Type Distribution per Country

Table 2: Tracking Type Distribution per Country

Country	Tracking type			Total
	External	Internal	Mixed	
GER	166	1045		1211
UK	127	189		316
US	485	261	17	763
Various		2		2
Total	778	1497	17	2292

Tracking Location

We used geolocation services to assess the tracking-server location. Table 3 shows the results for all tracking elements and corresponding web servers. Specifically, it depicts the two-letter ISO3166-1:2013 abbreviation of the country and the number of occurrence for main servers, and tracking servers in each country. As noted above, references to external images in a single e-mail may point to multiple servers. We defined the main server as the one that hosts the majority of external image references. According to our data, tracking images are rarely hosted on the main server if multiple servers occur in a single e-mail. For example, this situation may arise when a company stores images on an internal server but has outsourced tracking to an external service provider. To clarify the frequency of such an approach, Table 3 distinguishes between main and tracking servers.

Table 3 suggests that most countries differ in the number of occurrences for tracking and main server locations. One possible explanation is the use of external e-mail tracking providers that operate their business in a different country. Another reason might involve different regulations with regard to tracking technologies. In terms of server locations, Figure 5 classifies countries into three groups according to whether they contain more, the same, or fewer tracking servers than main servers.

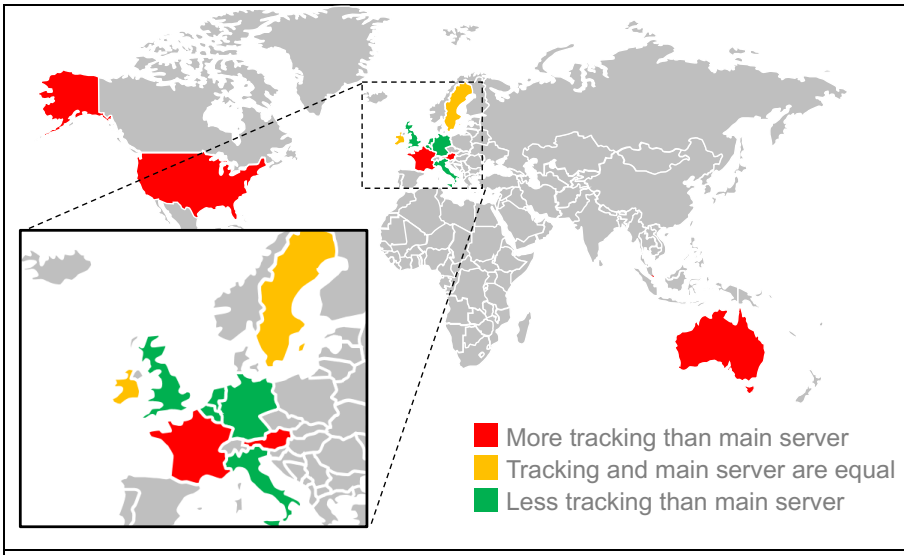

Figure 5: Country Classification

Table 3: Image Server Locations			
Country	ISO 3166	Occurrence tracking	Occurrence main
Austria	AT	77	52
Australia	AU	6	0
Belgium	BE	0	1
Germany	DE	913	1399
France	FR	39	1
Great Britain	GB	68	143
Ireland	IE	14	14
Italy	IT	0	23
Netherlands	NL	9	107
Sweden	SE	1	1
Singapore	SG	2	0
United States	US	1331	719

Countries that host more main servers than tracking servers are Germany, Great Britain, Italy, and the Netherlands. We aim to investigate in future work whether regulations related to tracking techniques are stricter in these countries, or if there are other reasons that tracking services are hosted abroad. In Italy and Belgium, only main images but no tracking images are hosted. Another group of countries host more tracking servers than main servers, such as Austria, Australia, France, and the United States. Singapore and Australia are especially interesting since only tracking images but no main images are hosted in these countries according to our dataset. Foreign tracking servers are least common for e-mails from US companies, where 99.9% of the tracking images are hosted within the nation.

Countermeasure Conceptualisation and Review

The previous analysis shows that e-mail tracking is a common phenomenon, which emphasises the importance of countermeasures to increase privacy and security of individual users. In the following, we conceptualise technological options for countermeasure design. We also review existing implementations and discuss their merits and limitations in order to verify the necessity of developing a novel approach.

Classification of Countermeasures

To summarise the sphere of possible solutions, we distinguished between deceptive and preventive countermeasures. We further divided the latter into holistic and selective approaches. Deceptive countermeasures strive to hide or modify the information sent to a mail sender or a third-party tracking provider so that these obtain selected, modified, or deliberately corrupted information. As Figure 1 shows, this strategy can be implemented through introducing a proxy server in the communication between e-mail sender and receiver, which caches the referenced images. The role of the proxy is to download and cache all images referenced in an in-coming e-mail prior to transferring it to the recipient's mail client. The sender of a tracking e-mail will then recognise the first access of a tracking image through the proxy, but will not be able to observe subsequent requests due to multiple opens. More importantly, the tracked identity is that of the proxy, whereas the reading behaviour of actual recipients remains unobservable. Figure 6 depicts two slightly different versions of this approach. In the first version (left), the proxy modifies the incoming mail so that the formerly externally referenced images are included in the mail. In the second version (right), the proxy server caches the externally referenced images and the references are changed so that they point to the proxy server. Every time the mail is opened, the images are fetched from the proxy and not the original tracking server.

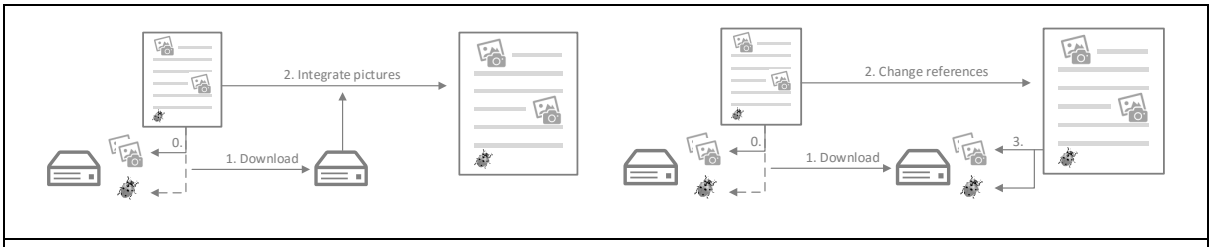

Figure 6: Deceptive Prevention Approach

The strength of the proxy-based approach is the possibility to conceal some tracking information. Weaknesses are related to the fact that the referenced images, including tracking images, are still accessed. The sender is therefore able to gather some, though potentially fuzzy, information. Another weakness is the necessity of server-side support. Operating a proxy server appears prohibitively expensive for an individual user. Even among e-mail clients, we were unable to identify an actual implementation of the proxy-based approach and, more generally, any deceptive countermeasure specific to e-mail tracking. Even though Google Mail introduced proxy technology for their web client in 2013 and hides some information such as the IP address from trackers, we conducted experiments that showed that every mail open is still registered by the tracking software. Therefore, proxy-based deception approaches do not yet provide full tracking protection.

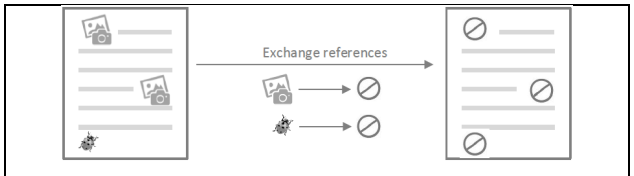

Figure 7: Holistic Prevention Approach

A second option for countering tracking is to block all external content referenced in an e-mail. We call this approach holistic prevention (Figure 7). Given that all tracking procedures known today depend on references to external content, the strength of the block-all approach is a fully-reliable protection against tracking. From a technical point of view, ease of implementation on either the server's or the client's side may be considered as another advantage. In fact, most contemporary mail clients allow users to block external content. However, the massive loss of information and decrease in user experience resulting from the exclusion of all referenced images and the corresponding content constitute severe disadvantages. For example, consequences of the holistic prevention approach include incorrect formatting, loss of styling elements, and potentially misinterpretation if external images convey important information and/or are crucial to correctly interpret an e-mail message.

Finally, selective countermeasures are based on identifying and blocking tracking elements within an e-mail (Figure 8). The idea is to categorise the referenced images into content-providing and tracking images. Assuming that images of the former category do not provide tracking functionality, they remain untouched, whereas tracking images are removed from the e-mail. The strength of the identify-and-block concept is the combination of systematic prevention of tracking images while preserving the full user-experience. The possible pitfall is the risk of misidentification. The solution will work only as well as the algorithm for tracking-image identification.

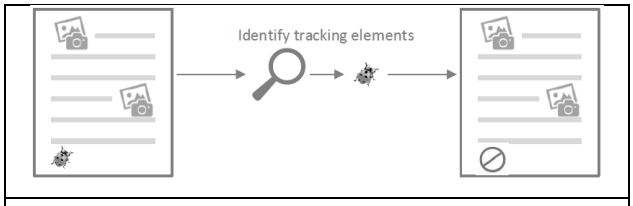

Figure 8: Selective Prevention Approach

Based on the review of alternative options for countermeasure design, we argue that a selective identify-and-block strategy provides the best balance between preventing user tracking and sustaining user experience. Accordingly, we concentrate on this approach in the remainder of the paper. In the next section, we report the results of an empirical evaluation of available implementations of the identify-and-block concept. Given the criticality of distinguishing content-providing and tracking images with high accuracy in this approach, we will then present a prototype of an identification engine and an empirically test of its effectiveness.

Selective Prevention – Empirical Experiments

Some mail clients and add-ons to mail clients support selective prevention. To examine the appropriateness of existing implementations, we created a testing infrastructure including e-mail accounts, e-mail clients, operating systems, hardware, and a third-party analysis tool (https://emailprivacytester.com/). The latter sent e-mails with tracking elements to our accounts and reported the information gathered through tracking. Using this infrastructure, we assessed the reactions of nine popular mail clients. More specifically, the selection of clients drew inspiration from a study by Litmus Labs and encompassed desktop, mobile, and web clients. We selected the three most popular mail clients in each category for analysis. Together, these clients were responsible for 92% of all mail openings (Litmus Labs 2015). It is worth noting that the mobile category had the greatest share with 47% of mail openings, which highlights the importance of mobile devices in today's e-mail communication.

Table 4 reports the results of our mail client evaluation, which distinguishes three process steps: before mail open, after mail open, and after displaying images. The first step characterises the point at which an e-mail client has synchronised with the mail server and the mail is visible in the list of incoming mails but has not been opened. After mail open is the moment when the mail is selected and opened but no further action has been performed. After display images refers to the time when the e-mail has been opened and the display of external images has been allowed by the user. This is only applicable if the mail client does not download external images by default. A check mark means that the mail client does not fetch the tracking image and therefore provides sufficient protection against tracking. In contrast, a cross indicates that the tracking images have been downloaded and tracking data can be gathered.

Table 4 reveals that no mail client fetched referenced images prior to opening an e-mail. Afterwards, however, four out of nine clients fetched referenced images directly. The other five mail clients fetched the tracking image in the third step. Given that none of the mail clients filtered tracking elements, we concluded that they failed to provide sufficient tracking protection.

Table 4: E-Mail Clients Usage Share and Protection against Tracking Images

Category	Mail Client	Usage Share	Before Mail Open	After Mail Open	After Display Images	Sufficient Protection
Desktop	Outlook	8%	✓	✓	×	×
	Apple Mail	8%	✓	×	-	×
	Windows Live Mail	2%	✓	✓	×	×
Mobile	iPhone Mail	28%	✓	×	-	×
	iPad Mail	11%	✓	×	-	×
	Android Mail	8%	✓	✓	×	×
Webmail	Gmail	18%	✓	×	-	×
	Outlook.com	5%	✓	✓	×	×
	Yahoo!	4%	✓	✓	×	×

With respect to add-ons, the second approach in the selective prevention category, we identified two representatives called *Uglyemail* and *Pixelblock*. The Uglyemail browser add-on allows the detection of tracking pixels from nine third-party tracking providers, but does not offer functionality to selectively block these images. Pixelblock blocks images with a one-pixel size, but fails to identify tracking images of different size. Both add-ons share the additional disadvantage that usage is restricted to the Google Chrome browser and Google's mail service Gmail. Given their restricted applicability, we conclude that browser add-ons also fail to provide appropriate tracking protection.

In summary, none of the tracking prevention solutions provided sufficient, practical and universal protection against e-mail tracking, which confirms the demand for novel solutions. Therefore, we proceeded with designing a prototype for a selective countermeasure and the corresponding identification engine in particular.

Tracking Image Detection

Tracking image detection can be interpreted as the classification of unknown images. All images that are referenced within an e-mail are used as input, and the classification process needs to decide whether or not each individual image is a tracking image. The detection process can be conceptualised in two steps. Frist, the identification of essential characteristics that distinguish normal from tracking images will result in a detection model that is based on various image attributes. Second, a classification decision needs to be made based on the detection model. This two-step approach ensures a certain independence of the detection model from the decision model. As a result, multiple different detection and decision techniques can be used and easily compared.

The detection model proposed in this paper is based on six different categories of data. The first three categories (image attributes, reference structure, and e-mail structure) subsume aspects that are directly associated with the source code of an e-mail which is part of the e-mail body. The fourth category (image server) is associated with the servers that host the images. The fifth category uses information from the whole dataset to assess whether a server is a tracking server. The sixth category covers the e-mail header.

In order to gather information on tracking images and to allow for a later validation of the developed model, the total dataset was divided into a training and a test set. The training set was used to gather

information on tracking pixels and to train the detection model. The training set accounted for 76% of the total mails, while the other 24% were assigned to the test set, which was used for validation purposes.

Image Attributes

The first category covers image attributes that are directly associated with an image element as well as attributes referring to centrally defined style information from Cascading Style Sheets (CSS). Typically, images are embedded using the <img> tag of the Hypertext Markup Language (HTML) (Musciano and Kennedy 2006). The analysis and model conception were restricted to attributes occurring in at least 1% of the images, since it would be difficult to derive representative results from a smaller amount of cases. Furthermore, the more attributes there were to be analysed, the slower the performance of the whole solution.

The *border* attribute defines the thickness of the border that is drawn around an image (Musciano and Kennedy 2006). While the border attribute only occurred in one third of all tracking images, it occurs in more than three-quarters of all non-tracking images. In our dataset, all values different from zero (or an empty value) occurred in non-tracking images only. The border attribute could therefore be used to identify images clearly as non-tracking images if their border was provided and deviated from zero.

The *width* attribute allows the horizontal size of a displayed image to be specified once the website or e-mail has been rendered (Musciano and Kennedy 2006). 65% of the tracking pixels with a specified width had a width of one. The *height*, similar to the width, allows the height to be defined at which an image is displayed. Another result of our empirical analysis is that the vast majority of tracking pixels are quadratic. One assumption from related work was that tracking pixels have an area less than 10 and are usually very small (Fabian et al. 2015). However, our dataset also contains 27 tracking pixels with a specified area of more than 10, which therefore does not match the former model criteria.

The *style properties* that were considered in the analysis are composed of the style attribute tag and centrally provided CSS commands. In order to avoid an individual discussion of each attribute and to be able to dynamically expand the attribute classification, categories of CSS commands were set up, subsuming commands which fulfil several criteria and allow us to either identify tracking images or to confidently release an image from the suspicion of being a tracking element. This is similar to a black- and whitelisting approach. Through the processing of the centrally provided CSS information for each image, an additional 24% of non-tracking images could be classified.

The *title* attribute can be used for various HTML elements. It was only used with one tracking image in our dataset and was in this case empty. Therefore, it can be assumed that if the title attribute with content is provided, the image is not a tracking image.

The attributes *vspace* and *hspace* define white spaces around images. The analysis showed that the tracking images have only zero as a value for both attributes. Therefore, all images that have a hspace or vspace value larger than zero could be classified as non-tracking images. Similarly, the attributes *align*, *id* and *usemap* only occur in non-tracking images. An image that uses any of the three attributes can be classified as a non-tracking image.

The *alt* and *class* attributes showed very mixed analysis results. We decided to leave them out since they could result in an overfitting optimisation of the detection model to the specific dataset, not leading to generally applicable results.

Reference Structure

The second category includes aspects that relate to the referencing link that points to the image (i.e., URL). This category focuses on the textual and structural analysis of the reference. The usage of individualised links allows both the user and the e-mail to be identified. This is a necessary condition for tracking images.

Therefore, an important aspect is the detection of *individualised references*. In a first approach, we identified several aspects that distinguish tracking from non-tracking references. One example criterion involves combinations of letters, numbers, and again letters. Based on the insights from this initial analysis, we developed a scoring model indicating the likelihood that a link is individualised. Tracking

references often fulfil the corresponding textual characteristics. Nonetheless, they are occasionally also encountered in non-tracking images. Therefore, several aspects need to be used in combination for detection.

The *wordlist* approach tries to identify tracking images by using a dictionary lookup to identify individualised parts of references. The idea is that non-tracking references (e.g., http://www.SLD.TLD/common/images/general/spacer.gif) usually contain a lot of common words that should be found in a dictionary, in contrast to tracking references that contain fewer or no common words. For the analysis, an English wordlist with 250,000 entries and a German wordlist with 190,000 entries were used. Each reference was split up into single parts that were checked in both wordlists. Afterwards, a measure was calculated that expresses the ratio of parts that were found in any of the dictionaries in relation to the total number of parts. However, it turned out that this approach did not provide results that clearly supported the decision process.

The *letter distribution* approach, similar to the wordlist, tries to identify how "normal" the image reference is, compared to typical distributions of letter occurrences within texts. Since links could be in both languages, distributions for the German (Beutelspacher 2015) and English languages (Lewand 2000) were considered. An analysis of the calculated measures shows that the deviations of the values for tracking images in relation to the rest of the images varied in both directions, higher as well as lower, while the deviation in the higher direction occurred more often. This deviation information can be used for tracking image identification.

Another aspect that could assist the identification of tracking images is the *similarity to other references*. Tracking references often distinguish themselves from other images in the mail with regard to their structure. A literature review on URL-similarity analysis revealed that various approaches have been developed. Often, the similarity of nodes (URLs) within a graph is used, which usually represents a subgraph of the World Wide Web (Benczúr et al. 2006; Cho et al. 1998; Lin et al. 2006; Maurer and Höfer 2012; Menczer 2004; Qi et al. 2007; Wu et al. 2012). Most approaches have been designed for hypertext web pages with two major requirements: the first is some machine interpretable text that can be used for textual word analysis, and the second involves hyperlinks that point to other web pages. However, both requirements are not fulfilled for image elements referenced in e-mails, which means that the graph-based approaches are currently not applicable for the given problem setting.

Therefore, another appropriate link-similarity measure was conceptualised. It is important to keep in mind that the measure should express structural similarity and not direct equality of each occurring character. In order to achieve this, the similarity of two links was defined as the amount of identical characters except digits, where the longer link is used as a comparison basis. If digits occur at the same position in each link, they are interpreted as equal characters, regardless of whether or not they are actually identical.

The *keyword filter* evaluates whether the use of keywords is useful for the identification of tracking images. The idea is that specific words are only used in the context of tracking pixels, while some other keywords might only be used in non-tracking images. This approach is similar to the black-/whitelist approach. Since the goal was to identify keywords that are independent of specific senders and recipients, further filtering was necessary. Finally, a whitelist of 14 entries and a blacklist with 32 entries were created.

The *user-id as part of a reference* is an aspect that combines the advantages from the huge dataset and the reference analysis. The term 'user-id' is defined as a unique identifier which each tracking link of a specific sender contains and which is not part of any other non-tracking image link. It is assumed that the user identifier is included in every tracking link for the same recipient, while the e-mail or content identifier is different for each e-mail and would therefore only occur in a single mail. It turned out that user-ids could be determined for 41% of the tracking-mail senders.

The *text only* aspect analyses whether the image references are only composed of alphabetic letters, except special characters that are used for separation. Tracking links often contain randomly generated components including numbers for identification purposes. And, in fact, all tracking elements within our training dataset contained at least one digit.

The *file extension* aspect focuses on the file extensions of the images that are referenced within the dataset. While the majority of file extensions are used for both tracking and non-tracking images, some file extensions occur in only one category. For example, the extensions "cfm", "php" and "ssp" only occur in tracking images, while the extensions "io", "jpe", "ver" and "xiti" only occur in non-tracking images.

The *regular expression patterns* describe the structure of tracking-image references by means of regular expressions. Our analysis shows that tracking pixels from different senders are quite similar in terms of their structure. For the training dataset, 79 different types of patterns could be identified. After optimisation of the regular expressions, all tracking image references could be detected, while no non-tracking images matched. The regular expression approach would be sufficient to detect all tracking image references within the given dataset, but could result in heavy overfitting.

Further aspects that we analysed, but which did not improve results, included the amount of special characters, the actual number of numerical digits in links, and capital letters.

E-Mail Structure

The third category, e-mail structure, focuses on aspects that describe the occurrence of images based on their position within the structure of an e-mail. Furthermore, the number of occurrences is considered.

The *position* of images within e-mails is an important criterion. During the data analysis phase, it was noted that tracking pixels seem to occur often at the beginning or end of an e-mail. This seems reasonable, since they do not provide actual content. Another explanation might be the use of external services or software that just appends the tracking image to the top or end of an e-mail. Extended analysis reveals that, indeed, the vast majority of tracking images occurred at either the beginning or the end of an e-mail. If the first and last three images were always taken together, they accounted for 98.9% of all tracking images within the training set.

The second aspect is related to the number of occurrences of an image within an e-mail. The analysis of our data shows that no image that was referenced at least three times within the same mail was a tracking pixel. This information alone could be used to classify 39,994 (48%) image occurrences (3,134 different images) as non-tracking images.

Image Server

The fourth category, image server, describes aspects related to the server hosting the referenced images. The first aspect is the *occurrence of servers* within an e-mail. The relative occurrence of servers can provide valuable information with regard to tracking image detection. For example: If a company uses an external tracking provider, the regular images for the e-mail may be hosted by the company itself, while the tracking pixel is hosted by the tracking server provider on a different server. Therefore, the previously introduced term "main server" is useful. A short analysis showed that in the training dataset, 94% of the tracking images were not hosted on the main server. This supports the hypothesis that tracking images are usually not hosted on the main server if one exists.

The second aspect concerns the *location of the server*. First, the location of all image servers is determined via domain-name resolution and IP address geolocation (Wang et al. 2011). Then, the main server is used as a reference for all server assessments within a single mail. Location points are assigned as an approximation of the distance between the server in question and the main server. It turns out that more than two-thirds (68%) of the servers with the highest point score per email hosted tracking pixels.

Server Black-/Whitelisting

The fifth category is based on the entire dataset and distinguishes whether the image server in question is hosting only tracking images, only non-tracking images, or both. This idea of black- and whitelisting is borrowed from the detection and prevention of unsolicited emails (SPAM) (Cormack 2007). For the application in our context, the elements of the lists are servers providing images that are referenced in the e-mails. This is an important difference from SPAM classification, where usually the sender or MTA is the object of investigation.

Our *blacklist* contains all servers that host tracking images but no non-tracking images. The second list is the *whitelist* with servers that only provide non-tracking images. The third case, *mixed hosts*, contains all servers that are part of neither the first nor the second list. This procedure was executed for the entire dataset. The majority of servers (57%) were part of the whitelist. A little more than one-third (34%) of the hosts were part of the blacklist. The remaining 9% of hosts were part of the mixed category, since they hosted both types of images. It has to be noted that the usage of servers could change over time. The approach is therefore only as up-to-date as the black-/whitelist itself, and regular updates should be ensured in order to minimise misclassifications.

Header Components

Any e-mail is composed of an e-mail body and an e-mail header. Usually, the e-mail header contains technical information and is not visible to the end user. The sixth category analyses fields of the e-mail header. Here, the header fields "list-unsubscribe", "reply-To", "message-ID", "content-type", "return-path", and "received-SPF" have been selected based on initial tests. A method was developed which allowed already classified image links to be correlated with the six header fields. The majority of classifications could be realised through a customised text search in the list-unsubscribe header field. The header fields reply-to, content-type, return-path, and received-spf did not cause any misidentifications. Overall, 99.8% of the occurring matches were indeed tracking images. Therefore, header analysis proved very useful.

Detection Model Summary and Dataset Dependency

We now turn to discussing the dataset dependency of the detection model in order to estimate how suitable the individual attributes are for detecting tracking elements in new, unknown datasets. A distinction will be made through the association to one of the following groups: low, medium, or high dataset dependency. Table 5 gives an overview of the model aspects and their dataset dependency.

Table 5: Model Summary and Dataset Dependency

Model Category	Model Aspects (Dataset Dependency)
Image Attributes	width, height, area, border, alt, style_blacklist, style_whitelist, vspace, hspace, title, class, align, id, usemap (all low dataset dependency)
Reference Structure	text-only (low), numbers_ratio (low), upperletters_ratio (low), exceptional_reference (low), match_blacklist (medium), match_whitelist (medium), file_extension_blacklist (medium), file_extension_whitelist (medium), match_user-id (high), match_regular_expression (high)
E-Mail Structure	image_occurrence, internal_image, image_position (all low dataset dependency)
Image Server	server_location_points, main_server (both medium dataset dependency)
Server Black-/ Whitelisting	server_blacklist, server_whitelist (both medium dataset dependency)
Header Components	unsubscribe_link, reply_to, message_id, content_type, return_path, received_spf (all low dataset dependency)

We assume that the aspects of *image attributes*, *e-mail structure*, *image server* and *header components* have low dataset dependency, since they represent properties that are characteristic for tracking images in general. The *black- / whitelisting* of image servers is assumed to be medium dataset dependent. Lists of tracking servers of external provider are applicable for all other e-mails that use the same provider.

The majority of *reference structure* attributes are assumed to have a low dataset dependency since they describe characteristics of tracking references that need to be fulfilled for the unique identification of e-mail and recipient. The reference keyword matching and the file extension analysis are assumed to have a medium dataset dependency, since they are relatively dependent on the data from which they have been derived, but are applicable to new datasets as well. The sender identifiers (user-id) are highly dependent

on the dataset, since they were generated based on the dataset and are specific to the e-mail recipients. Even though the regular expressions are applicable to other e-mails as well, it is assumed that they are not characterising all possibly occurring tracking image references and might therefore mislead future detection processes for entirely different e-mails.

Taken together, the developed model seems to be very applicable to different e-mails. Since the aforementioned regular expression attribute seems to be highly dataset dependent and shows a high explanatory power at the same time, we do not consider the attribute for future classifications, since it might distort the classification results and accuracy evaluation for future e-mails. This step was taken in order to make the detection process more independent from the specific dataset that was used.

Validation

This section evaluates the classification accuracy and execution time of the proposed mechanism for detecting tracking images in e-mails. The approach consists of the detection model (described above) and a decision model, which aggregates the model aspects and forms an overall decision. In order to perform the validation, the total dataset was divided into a training set and a test set. With this separation, a distribution similar to the total dataset was for the goal, in particular with respect to the tracking elements. The separation of the 4,505 mails resulted in a distribution of 1,086 (24.1%) mails (26781 images) in the test set and 3,419 (75.9%) mails (83299 images) in the training set.

An artificial neural network (ANN) was used as a decision technique. Classification tasks are a popular application area of ANNs (Fausett 1994). The attributes of the decision model are given as inputs to the ANN, which classifies whether or not the image in question is a tracking image.

It is expected that the ANN approach benefits from its capacity to detect various complex patterns, as well as the ability to handle incomplete information (Fausett 1994). For the given setting, multilayer perceptrons (MLPs) are appropriate. They can process categorical and numerical input and deliver categorical output values. They can also capture complex – nonlinear – relationships between inputs and outputs (Haykin 1999).

The network is generated using IBM Statistics in version 21. Since it was not our main goal to study the different possibilities for structural variations of the ANN, an automatic procedure was used for the network setup. As a training approach, the batch method was applied in order to assure that the network was directly optimised in terms of a global optimum (IBM Corporation 2013).

The *neural network* shows very good results regarding the classification of image elements in both training and test sets. The ANN was able to classify all images in the training set correctly. With regard to the classification of the unknown images in the test set, the overall accuracy was 99.99%.

Table 6 shows the so-called confusion matrix of the classification. A particularly important result is that no false negative classifications were generated that could threaten user privacy.

Table 6: Confusion Matrix for Image Classification

		Predicted Class	
		Tracking	Non-Tracking
Actual Class	Tracking	577 (TP)	0 (FN)
	Non-Tracking	2 (FP)	26,202 (TN)

The execution time is also a major aspect with regard to the practicability of the prototypical solution. The process involves a setup and a usage phase. In the setup (training) phase, the model is built and optimised for the data basis. In the usage phase (classification), the model is applied to classify unknown images. The training procedure needs to be conducted only rarely and is independent from the classification.

In our test environment, performance was determined separately for training and classification. The measurements were conducted under non-optimised conditions (virtual machine and limited main memory), leaving opportunities for performance improvements in subsequent applications. The measured times included the feature extraction, but without some fast pre-processing steps.

The training time (Figure 9, left) for the ANN shows a relatively linear relationship to the number of images that were used for training. The *classification time* (Figure 9, right) displays a nearly perfect linearity. Most importantly, classification was very fast with less than 0.2 ms per image. Full classification for a typical mail with 38 external images requires less than 7 ms. This result indicates a high practicality for real world application, even when many e-mails have to be processed.

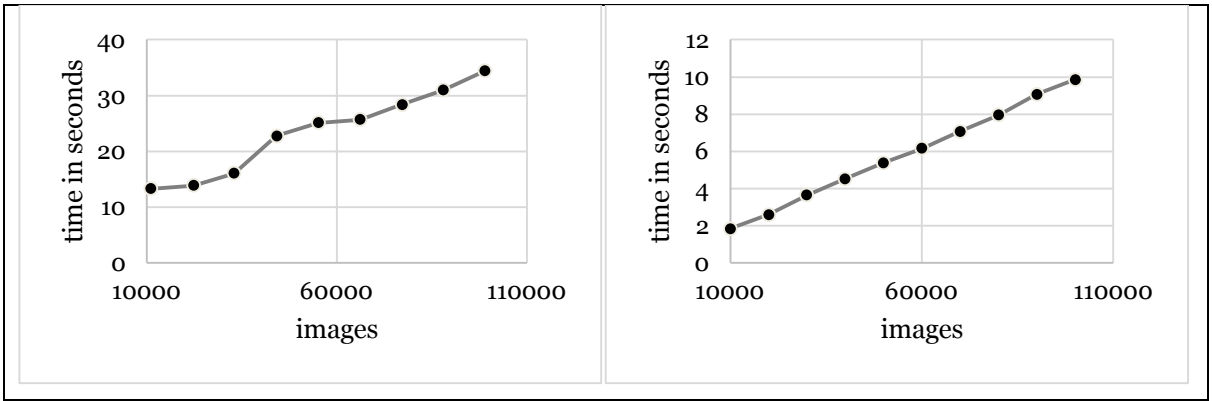

Figure 9: Training Time (left) / Classification Time (right)

Limitations

Finally, some important limitations of our research should be considered. The study focuses on *tracking images,* which represent a commonly applied and effective e-mail tracking mechanism. Other techniques such as tracking links have so far not been considered in the analysis and countermeasure design, even though similar approaches could be adopted.

The data used in this study only involve *professional e-mail newsletters,* which carry some important advantages for this research setting. Nonetheless, typical mail usage involves additional mail categories, especially individual mails. While we expect e-mail tracking to be less common, further studies are required on tracking mechanisms in these types of e-mails. Another aspect is the internationality of the dataset. Even though three representative countries were used for gathering the newsletters, tracking techniques adopted in other countries might slightly deviate and are currently not represented in the data set. The *location* analysis might not be highly accurate, since only free geolocation services were used for this study. Nonetheless, the country classification should be correct (Poese et al. 2011).

The *dataset dependency* of the tracking-image identification has already been discussed. Furthermore, this study uses data gathered in the year 2015. The results therefore reflect the current technological development at this time. It is very likely that the use of countermeasures will lead to an enhancement of tracking technologies, and will therefore make further advanced detection techniques necessary.

This study uses an *artificial neural network* for image classification. Alternative machine learning techniques could potentially show better results and should also be considered in further developments.

Conclusion

E-mail tracking can be used to gather sensitive information without user control, which raises several security and privacy concerns for end-users. Our empirical analysis of over 4,500 e-mails from the top 100 companies in Germany, Great Britain, and the United States showed that e-mail tracking is widely applied in all of these countries. Out of all the e-mails that could potentially contain tracking elements, 71% actually used tracking images. The tracking quota of German mails was the highest at 88%, followed by the mails from the US with 63% and UK with 62%. While the tracking in German mails heavily relied on internal tracking, the tracking mails from the US mostly relied on external providers.

Our evaluation of current countermeasures showed that currently no general, reliable, and sufficient protection against e-mail tracking exists. This constitutes a demand for a universal and reliable protection method.

As a first step in the direction of countermeasure realisation, several concepts have been developed and discussed in this study. With regard to end-user demands, the *identify & block* solution seems to be the most suitable, since it aims at selectively identifying and blocking tracking images while permitting other referenced images. Since e-mail tracking images so far do not provide the receiver with any content and are typically invisible, this solution does not reduce functionality or usability for the recipient.

Based on the analysis of our large empirical dataset with of over 110,000 images, a detection model was developed which encompasses six categories of important aspects that are useful for classifying unknown images. An artificial neural network was created based on the detection model and used as the decision technique for image classification.

Finally, the usefulness of our approach was evaluated using experiments on a test dataset. The experimental results showed that the neural network classified 99.99% of the images correctly and that no problematic false negatives occurred. Moreover, the execution speed of our classification algorithm was fast, indicating its practical usefulness for future work on a fully implemented countermeasure solution.

References

Agosti, M., and Di Nunzio, G. M. 2007. "Gathering and Mining Information from Web Log Files," in *Proceedings of the 1st International Conference on Digital Libraries: Research and Development*, C. Thanos, F. Borri and L. Candela (eds.), Berlin, Heidelberg: Springer, pp. 104–113.

Alsaid, A., and Martin, D. 2003. "Detecting Web Bugs with Bugnosis: Privacy Advocacy through Education," in *Privacy Enhancing Technologies*, R. Dingledine and P. Syverson (eds.), Berlin, Heidelberg: Springer, pp. 13–26.

Benczúr, A. A., Csalogány, K., and Sarlós, T. 2006. "Link-Based Similarity Search to Fight Web Spam Information Retrieval on the Web," in *Proceedings of the Second International Workshop on Adversarial Information Retrieval on the Web, AIRWeb 2006, Seattle*, B. D. Davison, M. Najork and T. Converse (eds.), pp. 9–16.

Beutelspacher, A. 2015. Kryptologie: Eine Einführung in die Wissenschaft vom Verschlüsseln, Verbergen und Verheimlichen, Wiesbaden: Springer Spektrum.

Bonfrer, A., and Drèze, X. 2009. "Real-time Evaluation of E-Mail Campaign Performance," *Marketing Science* (28:2), pp. 251–263.

Bouguettaya, A. R. A., and Eltoweissy, M. Y. 2003. "Privacy on the Web: Facts, Challenges, and Solutions," *Security & Privacy, IEEE* (1:6), pp. 40–49.

Burkell, J., and Fortier, A. 2013. "Privacy Policy Disclosures of Behavioural Tracking on Consumer Health Websites," in *Proceedings of the American Society for Information Science and Technology* (50:1), A. Grove (ed.), pp. 1–9.

Campaign Monitor 2010. *How Do I Create a Printer-Friendly Email Newsletter?* https://www.campaignmonitor.com/blog/post/3232/how-do-i-create-a-printer-friendly-email-newsletter/. Accessed 5 May 2016.

Cho, J., Garcia-Molina, H., and Page, L. 1998. "Efficient Crawling Through URL Ordering," in *Proceedings of the Seventh International Conference on World Wide Web 7*, P. H. Enslow and A. Ellis (eds.), Amsterdam: Elsevier Science Publishers, pp. 161–172.

Cormack, G. V. 2007. "Email Spam Filtering: A Systematic Review," *Foundations and Trends in Information Retrieval* (1:4), pp. 335–455.

Dobias, J. 2011. "Privacy Effects of Web Bugs Amplified by Web 2.0," in *Privacy and Identity Management for Life*, J. Camenisch, S. Fischer-Hübner and K. Rannenberg (eds.), Berlin, Heidelberg: Springer, pp. 244–257.

Evans, M. P., and Furnell, S. M. 2003. "A Model for Monitoring and Migrating Web Resources," *Campus-Wide Information Systems* (20:2), pp. 67–74.

Evers, J. 2006. How HP Bugged E-Mail: Commercial online service was used to track e-mail sent to a reporter in Hewlett-Packard's leak probe, investigator testifies. Images: Commercial e-mail tracking. http://www.cnet.com/news/how-hp-bugged-e-mail/. Accessed 15 July 2015.

Fabian, B., Bender, B., and Weimann, L. 2015. "E-Mail Tracking in Online Marketing - Methods, Detection, and Usage," in *Proceedings of the 12. International Conference on Wirtschaftsinformatik (WI 2015)*, O. Thomas and F. Teuteberg (eds.), pp. 1100–1114.

Fausett, L. (ed.) 1994. Fundamentals of Neural Networks: Architectures, Algorithms, and Applications, Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Fonseca, F., Pinto, R., and Meira Jr., W. 2005. "Increasing User's Privacy Control Through Flexible Web Bug Detection," in *Proceedings of the Third Latin American Web Congress*, R. Bilof (ed.), Washington: IEEE Computer Society, pp. 205–213.

Fourie, I., and Bothma, T. 2007. "Information Seeking: An Overview of Web Tracking and the Criteria for Tracking Software," *Aslib Proceedings* (59:3), pp. 264–284.

Harding, W. T., Reed, A. J., and Gray, R. L. 2001. "Cookies and Web Bugs: What They Are and How They Work Together," *Information Systems Management* (18:3), pp. 17–24.

Hasouneh, A. B. I., and Alqeed, M. A. 2010. "Measuring the Effectiveness of E-Mail Direct Marketing in Building Customer Relationship," *International Journal of Marketing Studies* (2:1), pp. 48–64.

Haykin, S. S. 1999. Neural Networks: A Comprehensive Foundation, Upper Saddle River, N.J.: Prentice Hall.

IBM Corporation 2013. IBM SPSS Neural Networks.

Javed, A. 2013. "POSTER: A Footprint of Third-party Tracking on Mobile Web," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, A.-R. Sadeghi, V. Gligor and M. Yung (eds.), New York: ACM, pp. 1441–1444.

Jensen, C., Sarkar, C., Jensen, C., and Potts, C. 2007. "Tracking Website Data-Collection and Privacy Practices with the iWatch Web Crawler," in *Proceedings of the 3rd Symposium on Usable Privacy and Security*, L. F. Cranor (ed.), New York: ACM, pp. 29–40.

Krishnamurthy, B., and Wills, C. 2009. "Privacy Diffusion on the Web: A Longitudinal Perspective," in *Proceedings of the 18th International Conference on World Wide Web*, J. Quemada, G. León, Y. Maarek and W. Nejdl (eds.), New York: ACM, pp. 541–550.

Leon, P., Ur, B., Shay, R., Wang, Y., Balebako, R., and Cranor, L. 2012. "Why Johnny Can't Opt Out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, J. A. Konstan, E. H. Chi and K. Höök (eds.), New York: ACM, pp. 589–598.

Lewand, R. E. 2000. *Cryptological Mathematics:* Mathematical Association of America Textbooks.

Lin, Z., Lyu, M. R., and King, I. 2006. "PageSim: A Novel Link-based Measure of Web Page Similarity," in *Proceedings of the 15th International Conference on World Wide Web* , L. Carr, D. De Roure and A. Iyengar (eds.), New York: ACM, pp. 1019–1020.

Litmus labs 2015. Email Client Market Share: Email Client Usage Worldwide, Collected from 1.02 Billion Email Opens. http://emailclientmarketshare.com/. Accessed 16 June 2015.

Martin, D., Wu, H., and Alsaid, A. 2003. "Hidden Surveillance by Web Sites: Web Bugs in Contemporary Use," *Communications of the ACM* (46:12), pp. 258–264.

Maurer, M.-E., and Höfer, L. 2012. "Sophisticated Phishers Make More Spelling Mistakes: Using URL Similarity Against Phishing," in *Proceedings of the 4th International Conference on Cyberspace Safety and Security, Y. Xiang, J. Lopez, C.-C. J. Kuo and W. Zhou (eds.), Berlin, Heidelberg: Springer, pp. 414–426*.

Menczer, F. 2004. "Combining Link and Content Analysis to Estimate Semantic Similarity," in Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters, S. Feldman and M. Uretsky (eds.), New York: ACM, pp. 452–453.

Moscato, D. R., Altschuller, S., and Moscato, E. D. 2013. "Privacy Policies on Global Banks' Websites: Does Culture Matter?" *Communications of the IIMA* (13:4), pp. 91–109.

Moscato, D. R., and Moscato, E. D. 2009. "Information Security Awareness in E-commerce Activities of B-to-C Travel Industry Companies," *International Journal of the Academic Business World*, pp. 39–46.

Musciano, C., and Kennedy, B. 2006. *HTML & XHTML: The Definitive Guide (6th Edition):* O'Reilly Media, Inc.

Nakatani, K., and Chuang, T. 2011. "A Web Analytics Tool Selection Method: An Analytical Hierarchy Process Approach," *Internet Research* (21:2), pp. 171–186.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems* (24:3), pp. 45–77.

Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., and Gueye, B. 2011. "IP Geolocation Databases: Unreliable?" *ACM SIGCOMM Computer Communication Review* (41:2), pp. 53–56.

Premkumar, G., and Roberts, M. 1999. "Adoption of New Information Technologies in Rural Small Businesses," *Omega* (27:4), pp. 467–484.

Qi, X., Nie, L., and Davison, B. D. 2007. "Measuring Similarity to Detect Qualified Links," in *Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*, C. Castillo, K. Chellapilla and B. D. Davison (eds.), New York: ACM, pp. 49–56..

Schmidt, J. 2013. "E-Mail im Visier: Tracking im Alltag aufspüren und abbestellen," *Magazin für Computertechnik* (22), pp. 130–135.

Sipior, J. C., Ward, B. T., and Mendoza, R. A. 2011. "Online Privacy Concerns Associated with Cookies, Flash Cookies, and Web Beacons," *Journal of Internet Commerce* (10:1), pp. 1–16.

Waisberg, D., and Kaushik, A. 2009. "Web Analytics 2.0: Empowering Customer Centricity," *The Original Search Engine Marketing Journal* (2:1), pp. 5–11.

Wang, Y., Burgener, D., Flores, M., Kuzmanovic, A., and Huang, C. 2011. „Towards Street-Level Client-Independent IP Geolocation," in *Proceedings of the 8th USENIX Symposium on Networked Systems Design and Implementation,* USENIX Association.

Wu, Y.-J., Huang, H., Hao, Z.-F., and Chen, F. 2012. "Local Community Detection Using Link Similarity," *Journal of Computer Science and Technology* (27:6), pp. 1261–1268.