Understanding Perceptions of Email Tracking Services for Senders and Receivers

Shivam Kumar Jha*
17CS30033
shivam.cs.iit.kgp@gmail.com
Indian Institute of Technology
Kharagpur, West Bengal

Veligeti Vineeth†
16CS30039
vvveligeti@gmail.com
Indian Institute of Technology
Kharagpur, West Bengal

Sayantan Bhakat‡
15EC35037
s8bhakat@gmail.com
Indian Institute of Technology
Kharagpur, West Bengal

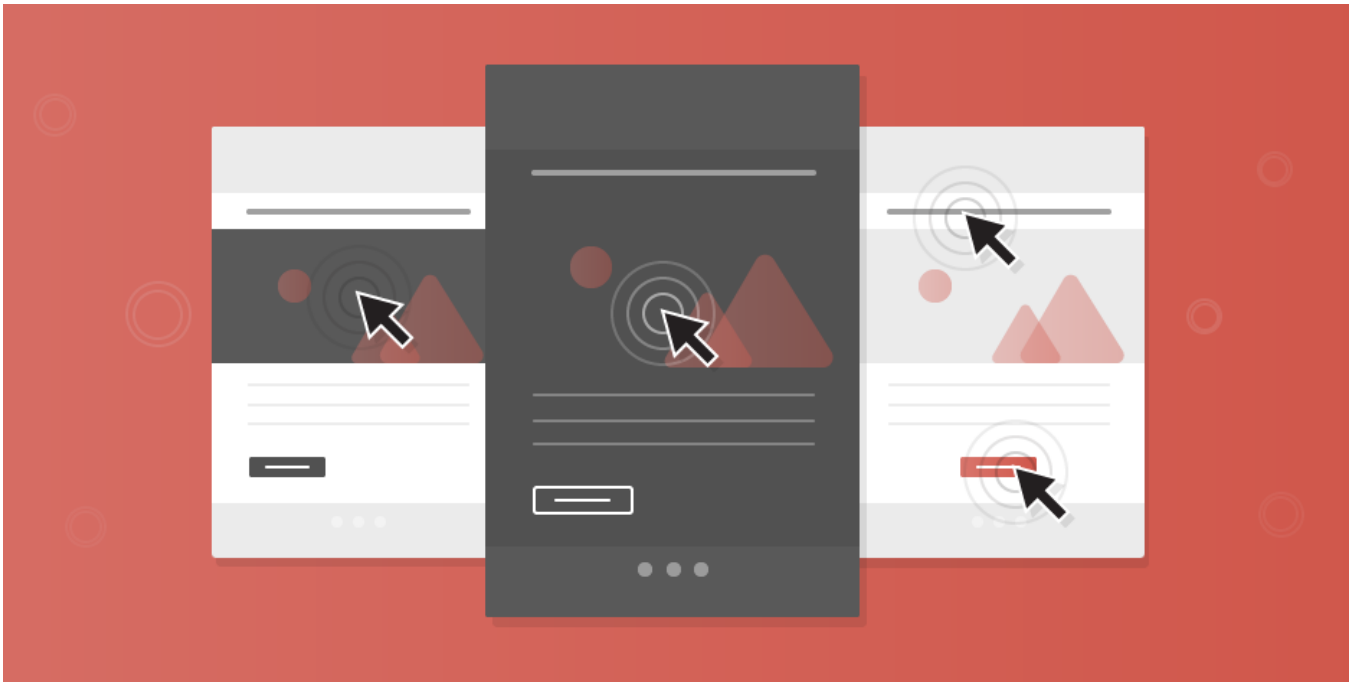

Figure 1: Source: https://moosend.com/blog/email-tracking/

ABSTRACT

In the research paper, we understand the privacy concerns of the users related to email tracking. Email tracking is a growing topic of concern in the field of Usable Security and Privacy, though not much research has been done in the field in the recent past. We have developed a tool Blink [20] that we use to collect data from participants from their Gmail inbox and then using available patterns and link footprint detection methods of Click tracking and Read tracking, classify emails into tracking and non-tracking. We have found that while most of the tracking emails are in the category of "Updates" and "Promotions" sent under marketing campaigns or by automated processes, a tiny fraction of "Personal" emails also have tracking enabled in them. We have also observed an effect of the user's behavior like more email account activity on the percentage of tracked emails that they tend to receive. While categorical analysis of the email tracking has been the focus of the paper, we have delved into how tracked emails are distributed on a time scale of the day, week, and month. It helped us understand how senders target certain periods of the day, week, or month to send emails with tracking embed in them and alter the user behavior. A prominent example of the same we found when we observed that there was a sudden increase in the number of tracked emails received by the participants during the festival season of Durga Puja, i.e. towards the end of October.

*Contributions to Research Question 1
†Contributions to Research Question 2
‡Contributions to Research Question 3

IIT Kharagpur, September 30, 2020, Kharagpur, WB

https://doi.org/xx.xxxx/xxxxxxx.xxxxxxx

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**; *Usability in security and privacy*; • **General and reference** → *Surveys and overviews.*

KEYWORDS

email tracking, tracking, email usage

ACM Reference Format:
Shivam Kumar Jha, Veligeti Vineeth, and Sayantan Bhakat. 2020. Understanding Perceptions of Email Tracking Services for Senders and Receivers. In . ACM, New York, NY, USA, 20 pages. https://doi.org/xx.xxxx/xxxxxxx.xxxxxxx

1 INTRODUCTION

Email tracking is the practice of monitoring opens and clicks of emails to follow up in cases of job applications, marketing, etc. This is achieved by embedding third-party content (images, scripts, etc.) or external links in the content of the email which gets triggered when the user opens or clicks on the email and the sender of the email is notified of the same. Initially, with the increasing importance and maturity of the online marketing industry, email has become an ideal channel to achieve an economical, effective, and targeted marketing solution. In recent times, email tracking has also made to personal emails via easily accessible and usable services to users such as mailtrack [10] that embed tracking content without any intervention of the sender once they sign up for the service.

Due to the accessibility of email, any user could be reached by an email with built-in tracking, and a simple email open could reveal information associated with the email reading activity of the user. The information suffices for miscreants to infer the device, email client, and the work and sleep schedule of the user. Thus, email tracking and the associated privacy concerns from email tracking has inspired a significant and growing body of academic work in the computer security and privacy community, attempting to understand, measure, and defend against email tracking [1, 21].

In this paper, we focus on the following three research questions that, on a broader level, attempt to understand the tracking pattern in user's inbox, their understanding of the term email tracking, and their knowledge about measures to perform tracking as well as protect against email tracking.

(1) **What percentage of the emails, divided into five categories as per G-mail - personal, updates, promotions, forums, and social, received by participants having a different level of exposure and understanding of the term "email tracking", have email tracking (precisely open tracking and click tracking) embedded with them?**

(2) **How knowledgeable are users about email tracking in general and different types of email tracking, detecting and blocking email tracking? How comfortable are users for sending and receiving tracked emails in different scenarios?**

(3) **What are the various impacts of the tracking system on both senders and receivers? What are the utilities of the tracking system? What are the privacy issues over multiple users of the email tracking system?**

The first research question. attempts at gaining an understanding of the pattern of the exposure of users to email tracking from a receiver's point of view. We attempt to understand the distribution of different categories of emails that the user receives, the quantification of the tracking emails in the inbox, and the relation between the user's email tracking usage and the tracking emails received. The aim is to draw a comparison between the user's knowledge and understand if more email activity, higher online presence, usage of email tracking, etc. impact the amount of tracking emails an individual is exposed to. Through our analysis, we attempt to establish the following hypotheses.

(1) Emails received with category other than "personal" have a higher percentage of emails (among total emails in the concerned category only) with tracking compared to personal emails received.
(2) Promotion emails have the highest percentage of emails (among total emails in the concerned category only) with email tracker enabled compared to any of the four other categories.
(3) Participants whose background requires a lower response time (defined as avg. time lapse between reply from or to the sender) have a higher percentage (among all received emails) of tracking emails received compared to participants with higher response time compared to them.
(4) Participants who send or receive more emails (defined as the number of emails present in their inbox for the concerned time) have a higher percentage of tracking emails (among all emails of the concerned time frame) received compared to users with lesser received or sent emails.
(5) Participants who send emails with tracking have a higher percentage of tracking emails (among all emails of the concerned time frame) received compared to users who don't send emails with tracking.
(6) Click tracking is the prominent (defined as the presence in more number of emails) form of tracking among click tracking and read tracking.

2 RELATED WORK

Trackers on the web. While we have focused on email tracking in specifics, today's world is laid with complex trackers [1, 2, 4, 11] and Google has been the leader for a while now with tracking users across 80% of sites [8]. The complexity has been marked by the shift from a focus on stateful tracking to stateless tracking. Device fingerprinting has been the pioneer in the same [3, 13].

On the contrary, email tracking has remained fairly simple and, since email clients prohibit JavaScript, does not use device fingerprinting. The email tracking in most cases is done via two means - click tracking [9] and pixel tracking [6]. We create a list of tracking URL patterns from two existing email tracking blockers - EmailTrackerBlocker [7] and ugly-email-website [12] - to detect the presence of tracking using Regex matching.

Privacy implications of email tracking. The literature on email tracking and related privacy issues have been concerned about the content of the email being leaked out and available to unauthorized or unintended third parties. For example, Holz et al. found that email is poorly secured in transit, often due to configuration errors [15]. In the paper titled, "I never signed up for this! Privacy implications of email tracking", Stephen et al. found that securing email in transit will not defend against email tracking and vice versa [17]. They emphasized that "the simple act of viewing emails contains privacy pitfalls for the unwary".

Haitao et al. [21] have estimated the amount of privacy-sensitive information available from email reading, assessed privacy risks of information leakage, and demonstrated how easy it is to launch a long-term targeted surveillance attack in real scenarios by simply sending an email with tracking capability. They have also investigated the prevalence of email tracking through a large-scale measurement, which includes more than 44,000 email samples obtained over seven years. At last, they have presented the users' perception of privacy infringement caused by email tracking.

Our work is complementary to Stephen et al. and Haitao et al. [21] as we have tried to explore the analytical aspect of how much tracking a user is exposed to and their usage pattern of tracking. We have also, with research question 2 and research question 3, looked into the emotional aspects and privacy concerns associated with the participant's perspective - without concern of the technical outlook or data breach concerns.

3 METHODOLOGY

We use a within-subject design approach; each participant is provided with the same executable script that reads the emails sent and received in a specified period from their Gmail Inbox. The entire process (figure 2) of reading emails, performing labelization, extracting timestamp, etc. happens on the local system of the participant itself. The emails are parsed, categorized, and analyzed for the presence of tracking elements (namely read tracking and click-tracking website REGEX pattern), following which a report (with one hot encoded data or timestamp, not any of the email content) is generated. This information is collected from all participants with their consent (refer appendix II of the attached IRB form) and post recruitment (refer append I of the attached IRB form).

3.1 Blink: One Tool To Do It All

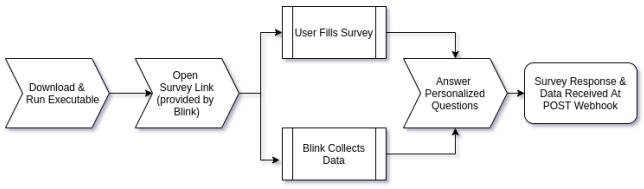

Figure 2: Flow of the survey process assisted by Blink.

Using open-source tools [19] to interact with the participant's G-mail account, we have developed a tool, Blink [20]. Blink uses the list of tracking services used by notable open source anti-tracking tools and finds tracking elements in an email's body. Blink is written in Python and is packaged as an executable using PyInstaller [18] for Microsoft Windows and Linux platforms. This makes it easy for participants to download and run the script, without performing any intermediate steps or installations.

Using Blink, the participant's emails are read for a predefined period on their system. Using the same tool, we also perform observation and data analysis of participants' mail - inbox and outbox. The participant's emails are parsed, analyzed, and labeled with Gmail smart labels [16] - category (namely personal, updates, promotions, forums, and social), have read tracking, has click tracking, thread average time, thread length, and time-stamp related data (see figure 3). We perform labeling for a category by using the smart labels that G-mail associates with the email - this is far more accurate compared to any data-set we can train a model on for the same purpose [16].

We use existing solutions [7, 12], and lists that provide regular expressions to match in the HTML & plain-text body of the email. The lists contain all the major tracking websites and the patterns that they use in their click tracking or pixel tracking code. We run a Regex match over the entire email body and accordingly decide if tracking is present or not based on the matches.

Thread length is determined by counting the number of emails that belong to the same thread. Accordingly, with all the individual emails belonging to the thread, we calculate the sum of all the time differences between each consecutive email (separate counters for sent and received) and divide by the total number of emails present in the thread to get the average thread interaction time in minutes. In this manner, we generate figures for the average delay between emails both for the reply (sent from participant) and receive (sent to participant).

3.2 Storing Data: Anonymously And Securely

interaction_type	has_open_tracking	has_click_tracking	content_type	time_quadrant	hour	minute	weekday	date	month	thread_length	thread_avg_reply_delay	thread_avg_receive_delay	identifier
RECEIVED	False	False	PROMOTIONS	3	12	41	Saturday	24	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
SENT	True	False	PROMOTIONS	1	5	12	Saturday	24	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	UPDATES	1	4	41	Saturday	24	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	UPDATES	1	4	40	Saturday	24	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	UPDATES	1	4	39	Saturday	24	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	True	True	PROMOTIONS	1	3	54	Saturday	24	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	UPDATES	4	23	2	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
SENT	False	False	UPDATES	4	18	9	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
SENT	False	False	PERSONAL	4	22	38	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	PERSONAL	3	15	44	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
SENT	False	False	UPDATES	2	11	20	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
SENT	True	False	UPDATES	2	8	2	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	UPDATES	4	22	13	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	True	True	UPDATES	3	14	29	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	True	False	UPDATES	4	19	21	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	UPDATES	3	12	50	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	UPDATES	3	13	10	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	UPDATES	2	11	8	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	UPDATES	2	10	53	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	UPDATES	3	16	19	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	UPDATES	3	15	51	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb
RECEIVED	False	False	UPDATES	2	7	26	Friday	23	10	1	0		0x6ca72cddd77e2dfc9ef1d47ed743edb

Figure 3: Data collected by the script from a participant of the test pilot run conducted on 17 October 2020.

Data collected from each participant is identified by a hashed identifier. The hashed identifier is performed using the MD5 hash algorithm and cannot be reversed easily to get the email id of the participant [5] and thus ensures that the responses and collected information cannot be associated with and used to identify the individual. The same hashed identifier is entered by the user in the survey conducted and, in this manner, each participant's labeled email data and survey response are linked.

This labeled data, an example of which is present in figure 3, is sent to a private POST webhook [14]. Once the data from all participants is collected, it is collated into a single spreadsheet using a personal Google Sheet shared only among the researchers. The sheet contains rows of data from a particular participant identified by their hashed identifier.

3.3 Pilot Study And Analysis Approach

To perform our pilot study, we chose 8 participants with the only compulsory requirement for the participant is that they should have an active email account with at least 100 emails (individual messages and not separate threads) sent or received (combined) in the specified period, that is from 00:01 1 October 2020 to 23:59 31 October 2020. The participants are sent the recruitment form (Appendix I of the attached IRB form) and consent form (Appendix II of the attached IRB form) through email communication.

Once the data is collected, we perform inter-subject and intra-subject comparisons to test the aforementioned hypotheses. The labeled data is also analyzed to conclude email tracking concerning the participants' email usage - collectively from all participants to have an ecological perspective. For numerical magnitude comparison, we use a simple calculation of percentages to conclude. We have used the Chi-Square test and correlation to draw the relationship of the categorical dependent variable - categorical independent variable and the quantitative dependent variable - quantitative independent variable. For testing our hypothesis concerning the relation between participants sending tracking emails and receiving a larger percentage of emails with tracking, we have used t-Test since we have categorical independent variable and quantitative dependent variable.

4 RESULTS

4.1 Category Effects and Distribution of Tracked Emails

Type	Updates	Promo	Personal	Forums	Social
Tracked	193	162	10	27	0
Non-Tra.	261	792	278	344	1033
Track (%)	42.51	16.98	3.47	7.27	0

Table 1: Comparison of tracked v/s non-tracked email for all categories and percent of tracked emails in the category.

As we can see in table 1 and figure 4, the category in which tracking was most prominent is "Updates" where 42.51% of the total emails in the category had tracking embed in them. A second far is the category "Promotions" with 16.98% emails having the presence of tracking; we had hypothesized the "Promotions" category to have the highest tracking. On the far last, we have the category "Social" with the least amount of tracking at 0% tracked emails.

This gives (figure 4) us an important insight that emails with promotional and update content are most tracked since they contain emails that companies use to track information such as bounce rate, click rate, conversion from email to opening website, etc. Strangely, the category "Social" does not contain tracking in any mail as we believe the information - that how many users open after seeing a friend request update (from Facebook) - is very important to a company.

From table 2, we can infer that the usage of email tracking is very less in "Personal" emails compared to all other categories combined as we had hypothesized. Even among all categories taken separately

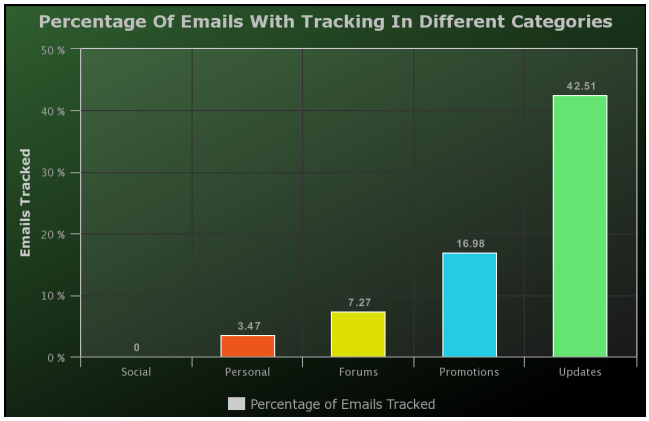

Figure 4: Percentage of email tracking present in each category in comparison to total emails in the same category.

Type	Personal	Others
Tracked	10	382
Non-Tracked	278	2430
Tracked (%)	3.47	13.58

Table 2: Comparison of tracked v/s non-tracked email for "Personal" category with others.

(figure 4), "Personal" emails are only second to "Social" in terms of emails having the least amount of tracking. This can be interpreted as most people generally are not concerned with tracking their emails when sending to friends, family, and other personal contacts as the urgency to get a response are less, and the statistics related to email opening and clicking are not important. Overall (figure 5) also we see that out of all tracking emails we parsed, "Updates" has a gigantic share indicating that within the category as well as among all tracked emails, updates are the most prominent category where tracking is used.

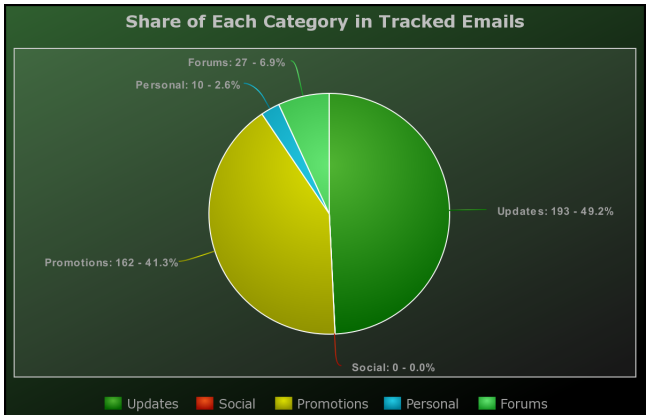

Figure 5: Distribution of total tracking emails across different categories.

Test	χ^2-value	p-value
Between Each Category (numbers)	563.986	0.00963
Personal (%) vs Others (%)	5.3212	0.02154

Table 3: Chi-Square Test values for the two tests.

In our χ^2 test results (table 3) that we ran to test our hypothesis of a relationship between the categories and the presence of tracking information in the emails in the category, we found p-values to be 0.009631 and 0.02154. Both the obtained p-values indicate that the relationship observed in the distribution of emails with tracking across the two tests - between each category and between personal v/s non-personal - have a high probability of holding and thus reject the null hypothesis that there is no effect of the category on the presence of tracking in the email. The lower p-value in the former test indicates a higher probability of the distribution repeating between categories compared to the distribution between personal and other categories.

4.2 User Behaviour and Percentage of Tracked Emails

Identifier	Total Mails	Tracked (%)	Resp Delay	Sent Tracked
e6ca72	470	17.23	56.12	Yes
341e36	461	5.63	3.58	No
2c2c8c	491	17.31	0.70	Yes
04372d	178	11.79	0.18	No
698cbd	426	23.94	0.35	Yes
f38c93	462	9.74	1.04	Yes
26094f	483	3.31	3.28	No
9f4181	129	12.40	2.33	Yes

Table 4: User data used for correlation and T-test from the pilot study with identifier truncated to first 6 characters.

Using the data in table 4, we conducted tests to understand the correlation between the quantitative variables. As per the results (table 5), we obtained interesting values of 0.22 and -0.43 for correlation between the percentage of tracked emails and the average delay between two consecutive messages in a tracked thread and percentage of tracked emails and total emails that are present in the user's inbox.

Independent Variable	Dependent Variable	Correlation Value
Tracked Emails (%)	Resp Delay (hours)	0.22
Tracked Emails (%)	Total Emails	-0.43

Table 5: Correlation values as obtained after running tests on quantitative variables.

As evident from table 5, there is a positive correlation of 0.22 between the percentage of emails received and delay in response indicating that user's who received a larger number of emails with tracking replied less often to the emails. With the negative correlation between the percentage of emails and the total number of emails received, we see that as users receive more and more emails the fraction of tracked emails keeps decreasing.

Independent Var	Dependent Var	Statistics Value	p-value
Sends Tracked	Tracked Emails (%)	2.46	0.048

Table 6: T-test value obtained for testing quantitative dependent variable and categorical independent variable.

To test if the users who send emails with tracking (self-reported) embed tend to receive more emails with tracking, we ran a T-test and the results are available in table 6. As can be ascertained from the low p-value of 0.048, the relationship between the data does not occur by chance and hence there is an effect on the number of emails with tracking received by a user if they send tracking emails or not. Since this was self-reported, the data also validates that users were probably reporting true since their sent emails with tracking information are also collected from their inbox and added to the number of tracked emails in our analysis, hence the effect.

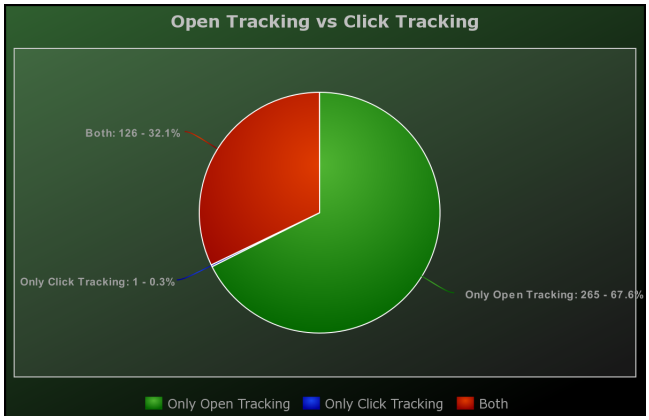

Figure 6: Distribution of total tracking emails across open tracking and read tracking.

An interesting observation in figure 6, contrary to our hypothesis, is that open tracking is the most common form of tracking element embed in tracked emails. Almost exclusively, click tracking is always used in pair with open tracking. This is interesting as it makes us aware of the fact that when senders are interested to know which links are clicked, they are also (with some exceptions) interested to look into whether the email is opened in the first place or not. For instance, we researched and found that the same behavior is observed in the service Sendgrid wherein marketing email senders have the option to embed open tracking or open tracking and link (click) tracking both - without an option to enable only link tracking.

4.3 Time Distribution of Tracked Emails

We conducted a simple analysis of the data collected to observe the distribution of the number of tracked emails over the month, a week, and the day.

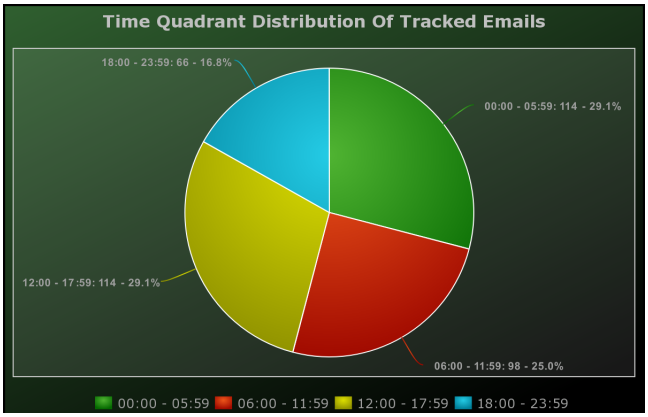

Figure 7: Distribution of total tracked emails across the day divided into four time quadrants of 6 hours each.

As presented in figure 7, we observed that most of the tracked emails are present in either the active hours (from 06:00 to 17:59) or wee hours of the morning (00:00 to 05:59). While it is obvious that active hours are targeted for better reach, we hypothesize that wee hours are targeted since that allows the emails to stay at the top of the user's inbox when they wake up in the morning and check their emails. In this manner, the chances of activity regarding the email from the user increase.

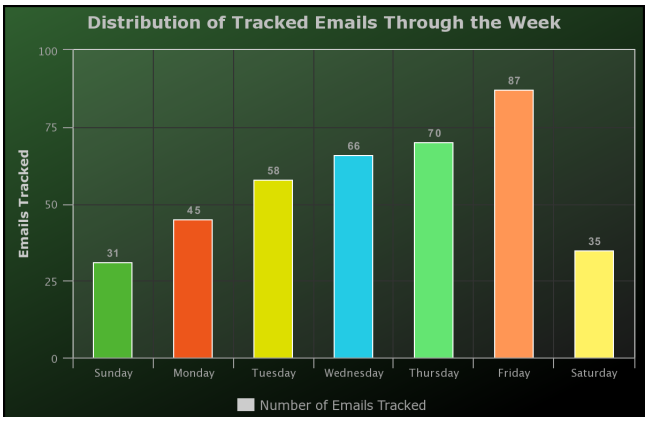

Figure 8: Distribution of total tracked emails across the week.

From the week based distribution (figure 8), we observe that most of the tracked emails are sent during the active days of the week,i.e. Monday to Friday. The frequency is increasing from Monday to Friday indicating a higher activity and more conditioning to the user as weekends approach. A possible explanation for the same could that users are more active on their email inboxes during the active days of the week, and while they may not open their inboxes so frequently on weekends, the emails that have been delivered to them on the days closer to weekend tend to influence their choices of online shopping, weekend getaways, etc. Hence, the increasing frequency of tracked emails from Monday to Friday and then a sudden drop.

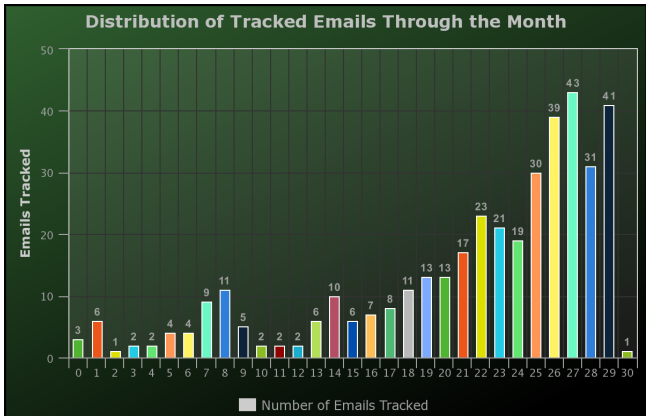

Figure 9: Distribution of total tracked emails across the month.
Note: 0 implies 1st date of the month and so on.

For the month analysis (figure 9, we observe that most of the emails that have been tracked are concentrated towards the last week of the month. This is not a mere coincidence as it overlaps with the festivities of Durga Puja, and hence a drastic increase in the marketing emails, order updates, etc. from different companies. Hence, we see a concentrated distribution of tracked emails towards the end of the month.

5 DISCUSSION

We observed in the responses that 7 out of 8 users were able to correctly guess which of the emails in their inbox had tracking, we were intrigued, and looking at the responses we realized that the users were associated mainly with the sender and content type of the email while making the decision. For instance, one of the users had the choice of emails - a marketing email selling a new mutual fund scheme from UTI Mutual Funds and a reminder for an upcoming exam from their professor. The user, using the email sender and content as the marker, chose the former as the one with email tracking. While this is an effective and intuitive method of performing segregation and guessing the presence of email tracking, it is far from the actual markers that a user should look for such as external images, links with a random string representing some form of identification, etc.

As we observed a stacking of tracked emails towards the end of October, we can find similar patterns for other festivities to understand the pattern better. A similar approach can be applied to understand with more precision the distribution of emails across the day (as we have done in four quadrants) to establish the times when users are more susceptible to received emails with tracking embed.

6 LIMITATIONS

As we have conducted a pilot study and not a full-fledged study with participants from diverse backgrounds with varying patterns

in email usage, email tracking, and different exposure to email tracking, the external validity of our study is limited. As we had participants, all of whom were students, our data, in general, represents a bias towards more email activity and tracking as students tend to use tools such as mailtrack to send emails for internships, jobs, etc. They are also more active on social media, exposed to e-commerce, and other online platforms that perform heavy marketing through email. Students in general due to their higher online activity are more aware of tools that allow or prohibit email tracking and hence, conclusions cannot be extended to the general email user base.

In our study, we have not considered emails that might have landed in spam, have been archived by the user, deleted by the user, or are not in the user's inbox due to any reason. Also, most of the marketing emails are likely marked as spam or deleted upon reception by the user, and they contain tracking mostly. This makes our categorical analysis skewed due to the absence of those emails which have tracking but are not present in the user's inbox.

We have also not considered the usage of email tracker blocking tools [7, 12] by the participant. In our study, if a participant is aware of the tools and uses the tools to not open, not reply, or any other action influenced by his knowledge of the presence of tracking in the email, we will not be able to know the uninfluenced behavior. In such scenarios, a user with the knowledge may have a relatively higher response time to tracked emails while in general users respond quickly to tracked email. These outlier participants in the study affect our conclusion and are thus undesirable and diminish the generalizability of the results.

Due to the nature of the data collected, some senders such as Amazon, Myntra, Facebook, etc. who tend to make only one-sided communication through email for order updates, marketing, etc. affect the thread response time. Their communication, if too quick or too spaced out, dictates the average response time of the thread without interaction by the user. In this manner, it affects the correlations we draw from the average response time and the presence of email tracking in the email. Hence, the external validity of the correlation cannot be guaranteed.

7 FUTURE WORK

Through our work and reading the academic research that has been done in the field of understanding privacy concerns with email tracking, we observe that there is a lack of study in the field linking how users perceive email tracking and what kind of content they consider more likely to have email tracking. As we observed, the general and intuitive markers that users associate while applying to explicit scenarios, are far from actual markers that reveal information about email tracking presence such as external images and external links with IDs among others. An area of future research would be understanding that once the users have been informed about the markers and taught how to spot them, are users able to identify emails with trackers better or they still rely on intuitive markers.

While we have considered intuitive markers such as content and sender to not represent the presence of tracking accurately, it is an interesting area to understand that if a unique fingerprint of multiple intuitive and understandable markers correlate with the presence of tracking. These markers may include but are not limited to content, sender, language, and tone of the content, subject, urgency of the information, etc. This would be an interesting amalgamation of Natural Language Processing and semantic analysis to correlate if emails with a particular pattern of markers are more likely to contain email tracking.

We can also utilize the feature of manual labels that users assign to their emails to extend the study towards observing if users understand that particular manually created labels may have more concentration of emails with tracking. The users can be shown emails with different labels they assigned and asked to choose if these emails contain tracking. This would help us understand their perception and bias towards expecting tracking emails, for example, do they expect mail from a friend to have a tracking component or mail from a student applying for an internship under them to have email tracking.

An important field of research, as we have presented in the section before, is the analysis for a better understanding of the pattern of tracking of emails throughout the year with a focus on festivities around the world. This will help us make users more aware of the targeted concentration of emails around the corner and, a further behavioral study may help users understand and protect from how the emails condition them to spend more during the festival time frame.

ACKNOWLEDGMENTS

We sincerely express our gratitude to Prof. Mainack Mondal (CSE, IIT Kharagpur) for investing his time in providing us with valuable feedback and guidance during the entire process of formation of our research questions and related details.

We also thank our friends and course-mates who were kind enough to volunteer for our pilot studies and help us gather valuable data and survey responses. It would not have been possible for us to complete the study and make observations without their time investment in helping us improve the study design.

REFERENCES

[1] Tadayoshi Kohno Adam Lerner, Anna Kornfeld Simpson and Franziska Roesner. 2016. Internet Jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016.
[2] Justin Rao Ceren Budak, Sharad Goel and Georgios Zervas. 2016. Understanding emerging threats to online advertising.
[3] Peter Eckersley. 2010. How unique is your web browser? , 18 pages.
[4] Tadayoshi Kohno Franziska Roesner and David Wetherall. 2012. Detecting and defending against third-party tracking on the web.
[5] Z. Ge and Z. Yong-Xia. 2010. MD5 Research. In *MultiMedia and Information Technology, International Conference on*, Vol. 2. IEEE Computer Society, Los Alamitos, CA, USA, 271–273. https://doi.org/10.1109/MMIT.2010.186
[6] H. Hu, P. Peng, and G. Wang. 2019. Characterizing Pixel Tracking through the Lens of Disposable Email Services. In *2019 IEEE Symposium on Security and Privacy (SP)*. 365–379. https://doi.org/10.1109/SP.2019.00033
[7] JannikArndt. 2018. E-Mail Tracker Blocker. Retrieved October 26, 2020 from https://github.com/JannikArndt/EMailTrackerBlocker

[8] Timothy Libert. 2015. Exposing the invisible web: An analysis of third-party http requests on 1 million websites.
[9] Dustin Long. 2009. Click tracking using link styles. https://patents.google.com/patent/US8543668B1/en
[10] Mailtrack. [n.d.]. Know When Your Emails Are Opened. https://mailtrack.io/en/
[11] Jonathan R Mayer and John C Mitchell. 2012. Third-party web tracking: Policy and technology.
[12] OneClickLab. 2020. ugly-email-website. Retrieved October 26, 2020 from https://github.com/OneClickLab/ugly-email-website
[13] Walter Rudametkin Pierre Laperdrix and Benoit Baudry. 2016. Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints.
[14] pipedream. [n.d.]. Request Bin. https://requestbin.com/
[15] Olivier Mehani Mohamed Ali Kâafar Ralph Holz, Johanna Amann and Matthias Wachs. 2016. TLS in the wild: An internetwide analysis of tls-based protocols for electronic communication.
[16] Kerry Rodden and Michael Leggett. 2010. Best of both worlds: improving gmail labels with the affordances of folders. *CHI EA '10: CHI '10 Extended Abstracts on Human Factors in Computing Systems* 50, 1 (April 2010), 4587–4596. https://doi.org/10.1145/1753846.1754199
[17] Jeffrey Han Steven Englehardt and Arvind Narayanan. 2018. I never signed up for this! Privacy implications of email tracking.. In *Privacy Enhancing Technologies*. 109–126.
[18] PyInstaller Development Team. [n.d.]. PyInstaller. https://www.pyinstaller.org/
[19] thealphadollar. [n.d.]. Blink Open Source Requirements. https://github.com/thealphadollar/Blink/blob/master/requirements.txt
[20] thealphadollar. 2020. Blink. https://github.com/thealphadollar/Blink
[21] H. Xu, S. Hao, A. Sari, and H. Wang. 2018. Privacy Risk Assessment on Email Tracking. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*. 2519–2527. https://doi.org/10.1109/INFOCOM.2018.8486432

A APPENDIX: IRB SUBMISSION

Please find the attached IRB document(s) from the next page.

STUDY TITLE:

Understanding Perceptions of Email Tracking Services for Senders and Receivers.

PRINCIPAL INVESTIGATOR:

Name: Dr. Mainack Mondal
Department: Computer Science and Engineering

STUDENT INVESTIGATOR (complete this section only if the project is student-initiated):

Name: Shivam Kumar Jha
Department: Computer Science and Engineering

Are you an:
☒ Undergraduate Student
☐ Graduate Student or Medical Student

Name: Vineeth Veligeti
Department: Computer Science and Engineering

Are you an:
☒ Undergraduate Student
☐ Graduate Student or Medical Student

Name: Sayantan Bhakat
Department: Electronics and Electrical Communications Engineering

Are you an:
☒ Undergraduate Student
☐ Graduate Student or Medical Student

VERSION DATE:

13/10/2020

RELATED STUDIES:

- Robust Identification of Email Tracking: A Machine Learning Approach [*Johannes Haupta, Benedict Benderb, Benjamin Fabianc Stefan Lessmanna*]
- Privacy Risk Assessment on Email Tracking [*Haitao Xu, Shuai Hao, Alparslan Sari, and Haining Wang*]
- I never signed up for this! Privacy implications of email tracking [*Steven Englehardt, Jeffrey Han, and Arvind Narayanan*]

Check any **applicable** boxes in the table below – you will be asked for further detail on these topics later in the protocol form:

Indicate Vulnerable Population(s) to be Enrolled	☐ Children (you must **complete Appendix A** in addition to this protocol document if you plan to enroll children) ☐ Cognitively Impaired Adults ☐ Pregnant Women (IF the research activities will affect the pregnancy or the fetus) ☐ Prisoners (or other detained/paroled individuals)
International Research (check this box if you will collect data from individuals located outside the India)	☐
Research involving external collaborators (some research activities will be carried out by individuals not employed by IIT Kharagpur or it's affiliates)	☐
Research has Indian government funding via direct award or a sub-award (e.g., NIH, NSF, other federal agencies or departments)	☐

1.0 Purpose and rationale of the study:

Purpose of the study:

The study is undertaken to understand the awareness of email tracking, the pattern of email tracking, and privacy concerns that users have with email tracking. The study will help us, and subsequent studies on how systems can be developed that take into account the concerns of the users while maintaining the functionality - for instance, development of a tracking system which explicitly gives the control in the receiver's hand to whether to send read receipt or not. The analysis of pattern and purpose in the user's inbox will help us understand the motivation for tracking and what services and what type of users (in terms of email activity) are more exposed to tracking services.

The following Research Questions will guide our study:

1) What percentage of the emails, divided into two categories - personal and commercial, received by participants, belonging to diverse backgrounds, have email tracking (vis-à-vis read tracking, click tracking, and attachment tracking) embedded with them?
2) How knowledgeable are users from different backgrounds about email tracking and different types of email tracking?
3) How comfortable are users for sending and receiving tracked emails and why?

2.0 Enrollment Criteria (who can be in your study and who would not be eligible to participate in your study):

Inclusion criteria:

1) should have an email account
2) should be using email for primary electronic communications in cases such as job applications, e-commerce transactions, business, etc.
3) age range: 18 - 50
4) gender: any

Exclusion criteria:

1) no email account
2) have not used email for the past one month

3.0 Sample Size:

Sample Size: 8 participants.

This is the minimum required participants to answer our research questions as we have to consider people from different backgrounds and draw conclusions based on their responses and data.

As this is a small number, there will be some limits to the generalization of our research, which will be mentioned in the project report.

4.0 Recruitment and Screening Methods:

Since the eligibility criteria of our study are trivial, we will be contacting participants who are known to us on a personal or professional level. Thereafter we will be sending them the recruitment form (Annexure I) and consent form (Annexure-II). If they have any doubts, we will be answering the doubts through the mail and keeping track of the conversation for future reference. If considered appropriate and required, we may choose to create a FAQ of common doubts from all participants.

Once the consent form and recruitment form is filled, we will be having a short 5 minutes telephonic conversation (detailed below) to reassure that they meet the basic requirements and explain to them the process once again to make it clearer.

Eligibility screening activities:

We will be mentioning favorable candidates in the recruitment form. To be more sure, a 5-minutes telephonic conversation with the participant will be done and the following questions will be asked:

- Do they use email services for business and e-commerce purposes?
- What is the average number of emails they receive in a day?

The answer to the above questions will not be recorded in any manner but used to ensure that the participant has a primary email account which is active (at least 5 sent and received emails combined in a day) and used for at least one purpose involving interaction with individuals out of their personal circle (individuals they know personally).

5.0 Research Locations:

No geographical research location required since the surveys are conducted through Typeform surveys and the user can run the provided script on their personal computer system.

6.0 Multi-site Research (research that involves external collaborating institutions and individuals):

Not Applicable

7.0 International Research (where data collection will occur outside India, including online activities)

The data collection will be done in a DigitalOcean server located in DigitalOcean's data servers in Bangalore, India. The collected data will be archived, without relation to the individual, for verification purposes.

8.0 Procedures Involved:

Please check the boxes for all applicable data collection procedures you plan to use:

☐One-on-one interviews

☐Focus Groups

☒Questionnaires/surveys

☒Analysis of secondary data (medical record data, educational records, government or private sector datasets, etc.)

☐Ethnographic observation

☐Physiological measurements (e.g., EEG, EKG, MRI)

☐Biospecimen collection (saliva samples, blood draws, hair samples, etc.)

☐Mobile applications/data collection devices (e.g., Fitbits, actigraphs, etc.)

☐Behavioral decision-making tasks (e.g., puzzles, interactive games, etc.)

☐Physical activities such as walking and other forms of exercise

☐Other procedures (briefly list types of procedures here if not covered by the check-boxes above): _____

A participant's email data will be collected using a script that will be provided as an executable. This data collection will only read the emails for an informed period of time in the past and all processing will happen on the user's system itself. Google OAuth will be used to gain read access to the user's email inbox. The executable will involve minimal interaction with the participant and will be finished within 5 minutes of the start. The following variables associated with an email will be recorded as labels without any personal information:

- type of email (business or personal)
- contains tracking: if yes, type of tracking (read, click, or attachment)
- interaction type (sent or received)
- sent/received timestamp
- thread length (number of following replies or receives)
- time difference b/w each reply

After the completion of data collection, users will be provided with hashed identifiers, comparison email links (to be referred to during the survey), and the survey link which will take around 8-10 minutes to complete.

9.0 Research with Vulnerable Populations (if children are the ONLY vulnerable population you plan to enroll, do NOT complete this section -- instead fill out Appendix A)

Not Applicable

10.0 Incomplete Disclosure or Deception:

Not Applicable

11.0 Consent Process:

The participants will be presented with a consent form (Annexure-II) that will be sent to them along with the recruitment email (Annexure I). They'll be required to read the consent form, digitally sign the document using an online platform (due to Covid restrictions, the physical signature process is being skipped), and revert back with the signed document. Any query of the participant will be answered via email and documented for future reference.

12.0 Waiver of Participant Signature on Consent Form:

Due to covid restrictions and the remote nature of the research, we will not be able to obtain physical signatures of the participant.

13.0 Waivers and Alterations of Consent Information:

Not Applicable

14.0 Financial Compensation:

Not Applicable

15.0 Audio/Video Recording/Photography

Not Applicable

16.0 Potential Benefits of this Research:

Potential benefits to participants:

Participants could gain a basic understanding of tracking mails and tracking in particular. They will be able to rethink their perspectives about tracking and tracking emails. We will also be sharing with them methods and tools using which they can be more aware and disable email tracking.

Potential benefits to society:

This research can be used in the future to design better email clients and tracking in mailing systems in particular, which will be beneficial and comfortable to receivers and senders. The analysis and finding can be used to understand and, thereafter, create tools that will facilitate tracking in necessary scenarios without risking privacy and with consent.

17.0 Potential Risks to Participants:

The only potential risk to participants in the identification of their data by reversing the hashed email ID. Currently, we are not aware of any practical tool that can perform the same, and hence the risk is negligible in the contemporary scenario.

18.0 Provisions to Protect Participant Privacy and Data Confidentiality:

Participant Privacy:
A participant's data will be associated with them using the hash of their email ID, which will be done on the system of the participant itself. Thereafter, all relations and associations within and b/w different datasets will be made using this hashed value.

Surveys will be conducted online, so the responses of participants (relating to the hashed email identifier) will be available only to investigators, which they consented to.

Confidentiality of data/biospecimens:

• Participant identifier (email) will be stripped and replaced on the participant's system itself before any or all data reaches us.

• We will be using an MD5 hash with a random secret created on the participant's system (ad-hoc generation, for instance using /dev/random on Linux systems).

• We will be using HTTPS REST communication protocol to transfer data b/w participants' system and DigitalOcean web server.

• The data will be stored for a period of one month (from the date of collection) in which it'll be used for verification of the research. Thereafter the same will be destroyed without any backups.

19.0 Data Monitoring Plan to Ensure the Safety of Participants:

Not Applicable

20.0 Long-term Data and Specimen Storage and Sharing:

If data/specimens will be stored and shared long-term for future research studies, explain the plan for storing and sharing the data. If you plan to place your data in a data repository, explain which repository/database and why.

Explain whether identifiers will be included with the data/specimens when they are shared.

21.0 Qualifications of Research Team to Conduct the Research:

The Principal Investigator has good experience in usable security and privacy research, his research works are present in this link.

The student Investigators have relevant experience in designing required software systems and scripts, have good and sufficient knowledge about email tracking to conduct this research. The student investigators are pursuing the course Usable Security And Privacy (CS60081) offered by Dr. Mainack Mondal which is providing them with the foundation to build the research on.

Annexure-I

Indian Institute of Technology
Kharagpur

Recruitment Email

Date: _ _/_ _/ 2020

Re: **Understanding Perceptions of Email Tracking Services for Senders and Receivers**

Dear (*Name*)

I am writing to let you know about an opportunity to participate in a voluntary research study about "**Understanding Perceptions of Email Tracking Services for Senders and Receivers"***.* This study is being jointly conducted by *Veligeti Vineeth*, *Shivam Kumar Jha*, and *Sayantan Bhakat* respectively at the Indian Institute Of Technology, Kharagpur.

Participants will be provided with a script (as an executable) which they can run in their computer system; the script will read through the last one month (or other pre-informed duration) emails, analyze them for category segregation, and observe the presence of tracking. No data will be stored on the server except for numerical data such category wise total number of emails w/ and w/o tracker(s). For the privacy of the participants, each data set will be anonymous and only the MD5 hash (near impossible to retrieve email id from the hash) of the participants' email will be associated with the data. The data will be exported, verified by the participant for any confidential issue before being sent to us by the script.

After completing the run of the script, participants will be given a survey link to be filled; the survey will take roughly 8-10 minutes. Questions will inquire the participants on the lines of their exposure to the understanding of email tracking, frequency of using email services, usage of email tracking, distinguishing b/w some emails from their inbox, etc.

As per our criteria, each participant should have an email account that they use as the primary mode of electronic communication for business, online shopping, job applications, etc. As the project is about tracking services, each participant should have heard of the term *tracking* and, it is better if they have heard of the specific term "*email*

tracking". We are committed to equal participation opportunities regardless of sex. And at Last, preferably participants should belong to an age group of 18-50 Years.

As a token of appreciation, we would be helping the participant understand their email tracking information as collected by our script and sharing with them resources and guidelines, following which they can have a more aware and secure presence on email platforms.

If you would like to know any additional information about this study or have some questions or concerns about the study, you can contact the researchers at the below-mentioned mail-ids respectively:

1.Veligeti Vineeth [vvveligeti@gmail.com]
2.Shivam Kumar Jha [shivam.cs.iit.kgp@gmail.com]
3.Sayantan Bhakat [s8bhakat@gmail.com]

Thank you for your consideration, and once again, please do not hesitate to contact us if you are interested in learning more about this Institutional Review Board approved project.

Dr. Mainack Mondal
Principal Investigator
Assistant Professor
Department of Computer Science and Engineering
Indian Institute Of Technology, Kharagpur

Annexure-II

Consent Form for Research Participation

Study Number: *ABC-1234*
Study Title: *Understanding Perceptions of Email Tracking Services for Senders and Receivers*
Researchers:

1.Veligeti Vineeth
2.Shivam Kumar Jha
3.Sayantan Bhakat

Principal Investigator: *Dr. Mainack Mondal*

Description: We are researchers at the Indian Institute of Technology, Kharagpur doing a research study about understanding perceptions of email tracking services for senders and receivers. The study is undertaken to understand the awareness of email tracking, the pattern of email tracking, and privacy concerns that users have with email tracking. The study will help us, and subsequent studies on how systems can be developed that take into account the concerns of the users while maintaining the functionality - for instance, development of a tracking system which explicitly gives the control in the receiver's hand to whether to send read receipt or not. The analysis of pattern and purpose in the user's inbox will help us understand the motivation for tracking and what services and what type of users (in terms of email activity) are more exposed to tracking services. At first, the participants will be presented with a consent form that will be sent to them along with the recruitment email. They'll be required to read the consent form, digitally sign the document using an online platform (due to Covid restrictions, the physical signature process is being skipped), and revert back with the signed document. Any query of the participant will be answered via email and documented for future reference. Participants will not be asked any personal question or need not share confidential information throughout the survey. Participation should take about roughly 8-10 minutes. Your participation is voluntary.

Incentives: You will not be paid for participating in this study.

Risks: The only potential risk to participants in the identification of their data by reversing the hashed email ID. Currently, we are not aware of any tool that can perform the same in practical time limits, and hence the risk is negligible in the contemporary scenario.

Benefits: Participants will gain a basic understanding of tracking mails and tracking in particular. They will be able to rethink their perspectives about tracking and tracking

emails. We will also be sharing with them methods and tools using which they can be more aware and disable email tracking.

Confidentiality: A participant's data will be associated with them using the hash of their email ID, which will be the sole identifier and will be done on the system of the participant itself. Thereafter, all relations and associations within and b/w different datasets will be made using this hashed value.
Surveys will be conducted online, so the responses of participants (relating to the hashed email identifier) will be available only to investigators, which they consented to.

Contacts & Questions:
If you have questions or concerns about the study, you can contact the researchers at the below-mentioned mail-ids respectively,

1.Veligeti Vineeth [vvveligeti@gmail.com]
2.Shivam Kumar Jha [shivam.cs.iit.kgp@gmail.com]
3.Sayantan Bhakat [s8bhakat@gmail.com]

If you have any questions about your rights as a participant in this research, feel you have been harmed, or wish to discuss other study-related concerns with someone who is not part of the research team, you can contact the Indian Institute of Technology Institutional Review Board (IRB).

Consent:
Participation is voluntary. Refusal to participate or withdrawing from the research will involve no penalty or loss of benefits to which you might otherwise be entitled.

By checking the "I agree to participate in the research" box below, you confirm that you have read the consent form, are at least 18 years old, and agree to participate in the research. Please print or save a copy of this page for your records.

□ I DO NOT agree to participate in the research
□ I agree to participate in the research