Privacy Nudges for Social Media: An Exploratory Facebook Study

Yang Wang*, Pedro Giovanni Leon†, Kevin Scott†
Xiaoxuan Chen‡, Alessandro Acquisti†, Lorrie Faith Cranor†

*Syracuse University
ywang@syr.edu

†Carnegie Mellon University
{pgl,kevinsco, acquisti,lorrie}@andrew.cmu.edu

‡University of Pittsburgh
xic34@pitt.edu

ABSTRACT

Anecdotal evidence and scholarly research have shown that a significant portion of Internet users experience regrets over their online disclosures. To help individuals avoid regrettable online disclosures, we employed lessons from behavioral decision research and research on soft paternalism to design mechanisms that "nudge" users to consider the content and context of their online disclosures before posting them. We developed three such privacy nudges on Facebook. The first nudge provides visual cues about the audience for a post. The second nudge introduces time delays before a post is published. The third nudge gives users feedback about their posts. We tested the nudges in a three-week exploratory field trial with 21 Facebook users, and conducted 13 follow-up interviews. Our system logs, results from exit surveys, and interviews suggest that privacy nudges could be a promising way to prevent unintended disclosure. We discuss limitations of the current nudge designs and future directions for improvement.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation (e.g., HCI)**]: Miscellaneous

Keywords

Facebook, nudge, privacy, online disclosure, soft paternalism.

1. INTRODUCTION

For several decades, social scientists have pointed to the role of heuristics and cognitive or behavioral biases (such as bounded rationality and hyperbolic discounting) in affecting economic decision making [20, 13]. Some of those biases and heuristics are likely to also affect online disclosure habits, explaining why making the "right" privacy decision – a decision an individual will not later regret – is difficult online [1, 4], and why regrettable disclosures may be common. Indeed, privacy blunders in social media offer vivid examples of the hurdles faced by users. Services such as Facebook facilitate the seamless, rapid broadcasting of intimate disclosures to audiences of both friends and strangers, often using interfaces fraught with complex settings. A considerable proportion of users of social media end up sharing online information and feelings that they later regret disclosing. Those disclosures sometimes carry substantial consequences, such as losing a relationship or a job [23].

PSOSM, 2013 Rio de Janeiro, Brazil
ACM 978-1-4503-2038-2/13/05.

In the field of behavioral economics, researchers have proposed soft (or asymmetric or libertarian) paternalistic interventions that nudge (instead of force) individuals toward certain behaviors [21]. Thaler and Sunstein popularized the idea of nudging as a form of soft paternalism to help people overcome cognitive or behavioral biases in decision making [22]. They define a *nudge* as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" [22]. For instance, a radar speed sign that displays the driver's current driving speed (e.g., 85 mph) does not force her to slow down when the speed limit is 60 mph, but rather nudges her to slow down. Inspired by studies of regrettable behavior on social media [23], and by the literature on behavioral decision research, our work explores a novel approach to help people protect their privacy in social media.

Specifically, in this paper, we describe the application of soft paternalistic interventions to mitigate the effects of behavioral and cognitive biases on information disclosure decisions. We designed and evaluated three mechanisms that nudge users to consider more carefully the content and context of their disclosures on Facebook. One nudging mechanism provides visual cues about the audience of a post; a second one includes time delays before a post is published; a third one gives users feedback about their posts. We also developed a platform that enables us to deploy nudges and test them with Facebook users "in the wild."

Using Facebook as an application domain, we explored the possibility of nudging users to make better (that is, less likely to be regretted) decisions about disclosing information in social media. We conducted a three-week exploratory field trial of these nudges with 21 Facebook users. By triangulating system logs of participants' behavioral data with results from an exit survey and follow-up interviews, we found preliminary evidence that the nudges had influenced some users' posting behavior, sometimes mitigating unintended disclosures and potential regret. We also identified limitations of the current nudge designs and future directions for improvement.

2. RELATED WORK

In the offline world, we are typically able to tailor our comments, gestures, and actions to a specific audience [10]. However, on online social media services such as Facebook, communication tends to be flat and lack context. For instance, Facebook users usually have different types of contacts (e.g., family, co-workers) as their Facebook friends. However, Acquisti and Gross found that early

Facebook users had problems configuring Facebook privacy settings according to their expectations [3]—a result more recently confirmed also by Madejski et al [15]. As a result, they may end up sharing content with all of their Facebook friends, a phenomenon called "context collapse" [16].

The consequences of privacy "breaches" in social media may range from simple embarrassment to stalking, identity theft [11], or damaged reputations [5]. Recently, Wang et al. offered empirical evidence of how Facebook disclosures actually led to negative outcomes including damaged personal relationships or problems at work [23].

As noted in the introduction, heuristics and cognitive or behavioral biases can help explain why individuals make decisions (including disclosure decisions) that they may later regret. A number of researchers have investigated or advocated soft paternalistic interventions to help individuals overcome those biases, and nudge them toward behaviors that may increase their welfare [21]. The application of soft paternalistic techniques to online privacy (and security) problems may help users make better online decisions, and avoid regrets.

While there is a large body of research on human behavioral modification (see [17] for an overview), so far little attention has focused on behavioral modifications related to online disclosures, particularly in social media [2, 19].

There has been some previous work attempting to apply nudging to computer security. For instance, Brustoloni et al. developed security dialogs in which users were held accountable for their decisions to open email attachments. Those who took unjustified risks could be "subject to a variety of sanctions, such as being unable to use the application for increasing periods of time..." A user study found that these dialogs resulted in significantly fewer unjustified risks [6].

An approach similar to nudging can be found in persuasive computing, a sub-field of computer science concerned with systems intentionally designed to "change a person's attitude or behavior in a predetermined way" [8]. Persuasive technologies have been applied in specific domains including computer security. For instance, Forget et al. built a text password system that encourages users to create stronger text password [9].

A number of mechanisms have been proposed to help users better protect their privacy in social media. Fang et al. designed a privacy wizard that asks users to iterate over privacy settings for some of their friends. Based on this information, a classifier could be built automatically to categorize the remaining friends [7]. Lipford et al. investigated interfaces for social network privacy controls, comparing compact settings in the form of expandable grids to visual policies. They found that both alternatives were usable, but different users appreciated them for different reasons [14]. While most previous work on privacy protections for social media has focused on helping users adjust their settings up front, our work employs tools that encourage real-time adjustments during or immediately following the composition of a Facebook post.

3. EXPERIMENTAL NUDGE PLATFORM

To explore the possibilities of using soft paternalism to help people make better information-disclosure decisions, we designed three types of privacy nudges for Facebook. We used Facebook as a testing application domain because of its popularity and the complexity of privacy issues associated with it. To investigate the effects of these three nudges on Facebook users' posting behavior, we developed an experimental platform that allowed us to integrate the nudges with Facebook and to collect data about users' posting behavior as well as their interactions with the nudges. In this section, we first present the three types of privacy nudges we designed, and then we describe the nudging platform.

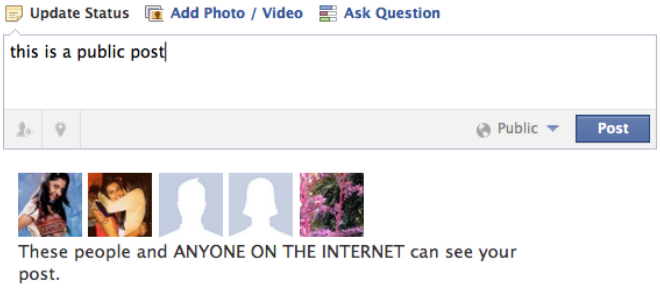

Figure 1: Profile Picture Nudge. A notice about the potential audience for the post and five profile pictures randomly selected from the set of people who will be able to see it are shown under the text box.

3.1 The Privacy Nudges

Inspired by the literature on cognitive and behavioral biases in decision making, past research on online information disclosures, and the concept of soft paternalism, we designed three types of privacy nudges. The general ideas behind the design of our nudges can be applied to various services or domains that involve information disclosure, such as Twitter or FourSquare.

Picture Nudge. Prior research has found that Facebook users often do not think about who is in their audience, and do not have a clear idea of who can see their posts. They also struggle to remember all of their Facebook friends, and often do not understand their privacy settings entirely. As a consequence, Facebook users often post content that can be viewed by unintended audiences; in many cases, this leads to regret [23]. In an attempt to address this, we implemented a nudge designed to lead users to consider the audience for their posts while they are composing them.

Our profile picture nudge attempts to encourage users to pay attention to their audience by displaying five profile pictures, randomly selected from the pool of people who could view the post being created. These profile pictures serve as visual cues to remind the user of the potential audience for their post. As shown in Figure 1, the profile pictures are displayed as a user starts typing in the post text box. The nudge also displays a notice to the user based on the user's current sharing setting. For example, if the post is to be visible only to friends of friends, the notice states, "These people, your friends, AND FRIENDS OF YOUR FRIENDS can see your post."

Timer Nudge. Acquisti has discussed how individuals may trade their personal information for immediate gratification [1]. Prior research on regrettable behavior on social media has also found that people often create regrettable posts "in the heat of the moment" [23]. To encourage users to reflect on their posts, we designed a timer nudge that inserts a short time delay before a post is actually posted.

Figure 2 shows a screenshot of the timer nudge interface before and after the user clicks the "Post" button. When a user starts typing a status update or comment, a message with a yellow background appears stating, "You will have 10 seconds to cancel after you post the update." After the user clicks the "Post" button, the user is given the option to "Cancel" or "Edit" the post during a ten-second countdown before the post gets published on Facebook. There is also an option to circumvent the timer by clicking a "Post now" button.

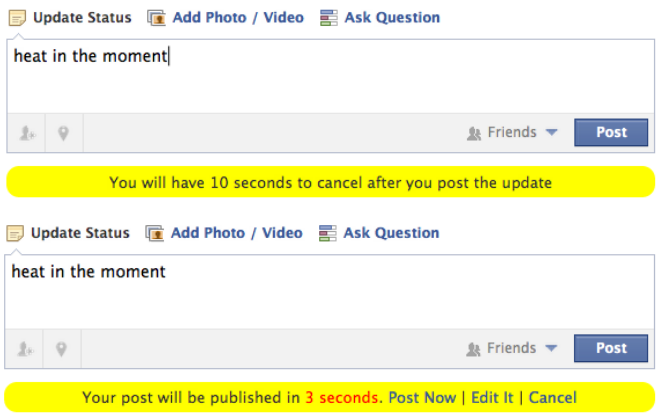

Figure 2: Timer Nudge. Top: timer interface before clicking "Post." Bottom: timer interface after clicking "Post."

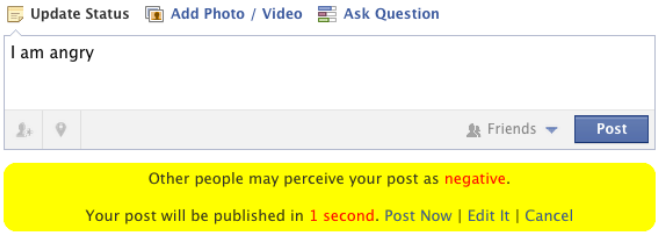

Figure 3: Sentiment Nudge. Different sentiment notices are shown depending on the overall sentiment of the post content.

Sentiment Nudge. Past research has found that regrettable posts on Facebook often contain negativity, profanity, or sensitive topics like alcohol and sex [23]. Our third nudge sought to provide users with immediate feedback on the content of their posts. We designed a sentiment nudge that combines a countdown timer with a notice regarding the content of the post, as shown in Figure 3. After the user clicks "Post," the timer and a notice highlighted with a yellow background will appear below the text box.

For this nudge, we used an open-source sentiment-analysis module to analyze the content of each post.[1]This module uses AFINN-111—a list of 2,477 English words and phrases manually rated as negative or positive, on a scale between -5 (negative or very negative) and 5 (positive or very positive) [12, 18]. For each post, any words in the wordlist are scored, creating a weighted sum for the entire post. A text message corresponding to this sum is shown to the user. For example, a slightly negative weighted sum would lead to the message, "Other people may perceive your post as *Negative*."

3.2 Nudging Platform and Data Collection

We implemented an experimental platform to display nudges on users' Facebook pages and monitor how users interact with these nudges. We used a Facebook application to access the users' data from Facebook and a Chrome browser plug-in to integrate the nudge interfaces into the Facebook pages.

The experimental platform stores the following data: each participant's information, including her Facebook ID, current Facebook privacy settings, history status updates, comments, and likes on Facebook (i.e., before the study); anything the participant typed in the status update or comment box (even if she didn't post it) during the study; any changes of privacy settings; her interactions with the nudge (e.g., clicking "Cancel" in the timer and sentiment nudges); and the Facebook IDs of the participant's Facebook friends. We use public key encryption to encrypt sensitive data, such as Facebook IDs and the content of posts, and store the data in a MySQL database on our server. The decryption key is kept offline and is only accessible to the research team.

In addition to the above data items, the browser plug-in and the Facebook application access and use the following data to enable the nudge features: the participant's Facebook profile picture, her Facebook friends' profile pictures, the participant's name, and her Facebookfriends' names.[2]

4. STUDY METHODOLOGY

To investigate how nudges would be perceived by active Facebook users and could impact their disclosures on Facebook, we conducted an exploratory field study with 21 participants, complemented with survey questionnaires and follow-up interviews. Participants remotely downloaded and installed a Chrome browser plug-in and a Facebook application, which they used over a period of three weeks. The study took place in Pittsburgh (PA) and Syracuse (NY) during the summer of 2012. It was approved by the Carnegie Mellon University (CMU) IRB.

4.1 Recruitment

We sought active Facebook users who were also native English speakers. Since our plug-in was designed for the Chrome browser, we recruited participants who primarily used that web browser to access Facebook. Participants were recruited using Craigslist, flyers, email distribution lists, and a CMU research recruitment system. Participants were given $10 Amazon gift cards for each week they remained in the study,[3] plus a $10 bonus for participating through the end of the study period and completing the final survey. Each participant who conducted an optional interview received an additional $10 Amazon gift card.

Recruitment material directed prospective participants to a screening survey. We invited via email 51 prospective participants, 31 of whom agreed to the online consent form and installed the Chrome plug-in and the Facebook application. Once participants had installed the plug-in, we verified that their self-reported Facebook usage was similar to their actual usage. We dropped one participant who in the screening survey self-reported posting several times a day but had only three posts recorded in the last 30 days. Two participants quit the study due to technical difficulties, and three more were dropped half-way through the study for not having answered the midterm survey. Four more participants never saw the profile picture nudge during the treatment period. We present results from the 21 participants who completed the field study and 13 of them participated in a follow-up interview.

Using a round-robin scheme, participants were randomly assigned to one of the three nudging interfaces: Profile Picture, Timer, and Sentiment.

4.2 Study Protocol

Study participants were required to install our plug-in and Facebook application, which allowed us to monitor participants' behavior on Facebook, and to enable or disable the corresponding nudge treatment for each participant. The field study comprised two main

[1]https://github.com/thinkroth/Sentimental

[2]We did not collect and store these data on our server.

[3]Either three or four weeks.

stages. During the first stage, the control stage, data collection took place without changes to the Facebook user interface. At the end of this stage, a midterm survey was administered to better understand the context in which each participant was making his or her posts and to identify external factors that could have affected participants' posting behaviors during the control period. During the second stage, the treatment stage, in addition to data collection, each participant was shown one of the three nudges. On average, participants remained in the control and treatment conditions for 11 and 12 days, respectively. The specific time each participant remained in the study depended on their response time to our midterm survey and the nudge they were assigned. In particular, participants in the profile picture nudge condition remained in the study for a longer time since we experienced technical difficulties showing profile pictures for posts with custom privacy settings and comments on posts originally made using custom settings. Leaving the participants more time in the study allowed us to resolve some of these issues and increased the chance that users would use a different setting (e.g., friends only) for some of their posts, allowing them to see the profile pictures.

At the end of the field study, we administered a final survey that collected participants' opinions on the nudge they were shown. We further asked whether they were interested in participating in a follow-up interview. We extended this invitation to all participants who expressed interest, except the four participants in the profile picture treatment who, due to technical difficulties, never saw the profile pictures during the study.

The purpose of the follow-up interviews was to understand participants' attitudes and perceptions about as well as experiences with the nudges. We asked participants about their main motivations for using Facebook, knowledge of Facebook privacy settings, first impressions with the nudge interface, and perceived benefits and drawbacks of that nudge. We then showed them three posts or comments they had made and asked them about the contexts of those posts and whether the nudge had affected their posting decision in any way. Towards the end of the interview, we asked them to log into their Facebook accounts using their own laptops or a lab computer with the Chrome plug-in installed. We reactivated the nudge they had seen during the field study and collected their ideas for design improvements while seeing the nudge on their FB page. Towards the end of the interview, we showed them a different nudge from the one they had used during the field study and collected their opinions about that other nudge.[4] We interviewed 13 participants, and each interview took about 30 to 45 minutes.

4.3 Analysis

We analyzed participants' responses to Likert questions, behavioral data collected using the Chrome plug-in, and interview data to explore the impact of our three nudges.

The final survey included Likert questions that queried participants' opinions about the usefulness of the nudges, their willingness to use these nudges, and their level of comfort with the nudges. The purpose of these analyses was not to compare statistically the results across the three nudge treatments, but to show a quantitative summary of opinions about these treatments.

We used the data collected with the Chrome plug-in to investigate whether there was any evidence of changes in Facebook usage before and after the participants started seeing the nudges. The metrics that we used to investigate behavioral changes included: number of changes in inline privacy settings, number of canceled or edited posts, post frequency, and topic sensitivity. We focused on sensitive topics that previous research identified as problematic on Facebook [23]. Given the number of factors other than our nudges that could have affected participants' behaviors during the study period, we do not claim any causality, but only show instances that could have signaled an impact of a nudge on users' behavior. Similarly, given the exploratory nature of our study, the small sample size, and the uncontrolled environment of the study, we did not attempt to perform any statistical tests. If we had a larger sample size, we could have analyzed the results using a number of statistical techniques based on the distribution of the collected metrics. For example, we could use t-tests or Wilcoxon Rank Sum tests to perform both between- (across treatments) and within-subjects (control versus treatment) comparisons using the collected metrics as dependent variables.

Finally, we performed a qualitative analysis of the interview data. We developed a code book of the comments that participants made during the follow-up interviews. We then grouped these comments into thematic strands, including perceived benefits and drawbacks, context in which the nudges could have a positive effect on users, and opportunities for design improvement. We report comments that were common among participants, as well as those that were unique. We illustrate these comments with a number of interview quotes.

5. RESULTS

In this section we first describe our participants' demographics and overall posting behavior. Then we discuss participants' first impressions of the nudges, which were collected at the beginning of each interview. After that, we use system logs and interview data to describe the impact of these nudges on participants' posting decisions. We further discuss the participants' perceptions of benefits and drawbacks of these nudges. Finally, we discuss the results of the survey administered at the end of the field study.

5.1 Participants' Demographics

Seven of our 21 participants were undergraduate students, five were graduate students, two were unemployed, and seven were employed in a variety of occupations. They included 13 females and eight males between the ages of 18 and 48 (mean age 24). Our participants' demographics are summarized in Table 1. We use a combination of a letter and a number to refer to each participant. The letter represents the initial for the nudge treatment, and the number refers to the sequence within each treatment group. For instance, T-1 denotes the first participant in the time nudge group.

During the 3-week study period, our Chrome browser plug-ins stored a total of 1,209 posts (353 status updates and 856 comments) made by the 21 participants. On average, each participant made about 2 posts per day. Table 1 also includes summaries of participants' posting behavior. For participants in the sentiment nudge, the number of nudge appearances include both positive ("Other people may perceive your post as [*positive / very positive*]) and negative ("Other people may perceive your post as [*negative / very negative*]) messages. The sentiment warning did not appear if the post was considered *neutral* by the sentiment analysis algorithm.

5.2 Participants' First Impressions of Nudges

During the interviews, we asked participants "What was your impression when you first noticed the new interface on your Facebook page?"

Three of four interviewees in the timer nudge treatment commented that they thought the delay was a new feature introduced by Facebook, although they wondered why Facebook would want

[4]Participants in the profile picture treatment were shown the sentiment nudge and participants in the sentiment or timer treatments were shown the profile picture nudge.

ID	Sex	Age	Days in Control/ Treatment	Nudges Shown	Posts in Control	Posts in Treatment
Picture						
P-1*	F	29	11 / 12	4	18	21
P-2*	F	18	11 / 18	10	68	63
P-3*	F	23	12 / 18	2	17	23
P-4*	M	27	14 / 16	33	47	116
P-5*	F	35	14 / 16	34	40	32
P-6	M	48	12 / 10	10	25	26
Timer						
T-1*	F	21	10 / 11	20	16	20
T-2*	F	30	10 / 11	4	22	4
T-3*	M	24	11 / 12	32	38	32
T-4*	M	18	13 / 16	17	114	17
T-5	M	20	10 / 11	6	6	6
T-6	M	22	11 / 11	27	45	27
T-7	F	21	11 / 12	2	8	2
Sentiment				**All (Pos. / Neg.)**		
S-1*	F	25	10 / 11	3 (3 / 0)	2	8
S-2*	F	27	10 / 11	3 (2 / 1)	4	6
S-3*	F	26	13 / 13	20 (13 / 7)	69	24
S-4*	M	31	11 / 12	3 (3 / 0)	11	3
S-5	F	22	10 / 10	4 (4 / 0)	20	5
S-6	F	18	10 / 11	3 (3 / 0)	26	10
S-7	M	20	13 / 13	13 (6 / 7)	24	40
S-8	F	22	14 / 16	33 (23 / 10)	59	45
	Min	18	10 / 10	2	2	10
	Max	48	14 / 18	34	114	116
	Mean	24	11 / 12	13	32	25

Table 1: Participants' demographics and posting behavior. The nudges were shown only in the treatment period. The profile picture nudge appeared only on some of the posts and the timer nudge showed up on every post. In the sentiment nudge, the countdown timer appeared on every post, in addition to positive or negative messages, depending on the content of the post. The * denotes participants who were interviewed.

to introduce the new feature. T-2 explained that the first time she saw it she was annoyed by the time delay: "Why would it make me wait?" Later, she noticed that "post now," "edit," and "cancel" were clickable options and started to like the features because they allowed her to review her posts before making them public. When we switched participants to the timer and sentiment nudges, we experienced a few technical difficulties that caused some of the participants' posts not being posted. T-4, who experienced this problem, also expressed a negative feeling. "The application was eating my posts," he said. Nevertheless, this participant later explained that once the problem was fixed, the timer nudge prevented him from posting trivial statements such as "hahaha," which he perceived as a benefit from the timer.

P-1 and P-4 wondered whether the profile pictures were a new Facebook feature or part of the user study. Another participant, P-2 thought it was a new Facebook feature that would allow her to tag people easily, but she soon realized that was not the case. She was surprised when she read that her post could be seen by such a large number of people. "It reminded me that I should probably clean up my friends list," she said.

S-3 immediately associated the sentiment nudge with the study. Both S-1 and S-3 wondered how the sentiment of their posts had been determined when they saw the "Other people may perceive" warning message. However, while S-1 expressed that "it made me think," S-3 mentioned she completely disregarded it. S-3 further explained, "I was like why would it think it's negative? Oh whatever, post now." She further elaborated that she did not like the warnings because "I'm giving a legitimate statement or opinion on something or I'm being sarcastic and my friends know that." This participant's comment highlights an important challenge of content or sentiment analysis: it should consider or understand the context around a post, not only the content of the post itself.

5.3 Impact on Posting Behavior

We logged participants' posting behavior on Facebook and their interactions with the nudges during the study. We analyzed participants' posting behavior during both the control and treatment periods. We found evidence of changes in posting behavior for some of our participants, and we combined those results with the interview data to better understand whether those behavioral changes could be associated with the nudges. We use concrete instances to illustrate the kinds of impacts that each nudge has on some participants' posting behavior.

5.3.1 Profile Picture Nudge

Both P-2 and P-3 reported the profile picture nudge made them think about their privacy settings and the content of their posts. P-3 reported having changed the privacy settings of one post because she saw a picture of a person she did not recognize. When looking at her behavioral data in our system logs, we noticed that during the treatment period, she changed her privacy settings from "Friends" to "Friends except acquaintances" when she posted "Survived one of the craziest, most exhausting days ever!" Based on the stored typing history of this post, we also found that the post was edited from the original, "Definitely just had one of the craziest/most exhausting days ever."

P-2 reported that she ended up canceling "a couple of posts" because of the profile picture nudge. She explained that she once canceled a negative post: "There wasn't any swear words or anything but it was a snide remark and then one of the pictures that popped up was one of the people I work with. It is probably not the best idea." She volunteered that she is often careless when posting on Facebook and the nudge "made me change, it did make me think." She added that she could probably benefit from the sentiment nudge as well, especially if she could configure a dictionary of curse words she normally uses. In contrast, although P-5 recognized that the profile picture nudge creates awareness about the audience of one's posts and encourages people to be more cautious, she did not believe that the nudge had a significant impact on her posting decisions. P-1 and P-4 both volunteered that they were ignoring the profile pictures for most of the study. P-4 explained, "I only make my posts available to friends," and he claimed he knew which people he had placed on his friends list. He added, "If I were using different lists, [the profile pictures] would be very useful."

5.3.2 Timer Nudge

T-3 mentioned that the timer was "at times annoying and at times handy." He explained that it was annoying when he "knew exactly what I wanted to say" but had to wait for the timer to expire or hit "post now," which required extra time and effort. He also said it was handy because sometimes he edited his post to "make it a bit more publicly acceptable when it was a venting post" or to fix typos. He also said he canceled posts rather than wait for the timer "if I didn't need to say it." He further volunteered that he posted less often due to the time delay. However, we did not observe a change in the frequency of his posts during the study period.

T-4 reported that the timer made him think about the utility of his posts, explaining that he canceled several posts because the timer made him realize it was not really necessary to post them. Indeed, our collected data about him show that, on average, he reduced his posting activities in the treatment period by more than seven posts per day. In addition, while he did not post sensitive content during the treatment period, there were ten instances of sensitive content during the control period. He also edited a few of his posts in the treatment period. For example, one of his comments was "Wow." Upon reviewing the typing history we stored for this comment, we found that he typed, "God damn. That's so cool man," and then deleted this sentence from the comment.

Both T-1 and T-2 agreed that the edit option was very convenient. They were using the time delay to review their posts, and they started liking the nudge after having used it for several days. T-1 reported caring about what she writes on Facebook and paying attention to grammar and spelling. She volunteered that she clicked "edit" several times to improve the wording of her posts. Similarly, T-2 mentioned she used the "edit" option a few times. For example, once when she posted a link to a movie cover, she edited out "this is the movie" because she felt it was redundant.

5.3.3 Sentiment Nudge

S-2 said the nudge reminded her that she was in the study, but that most of the time the sentiment meter was very sensitive or missing the context. Regardless, she remembered that the first time she saw the negative sentiment warning was when posting "damn the Steelers rock," and she decided to use the word "dang" instead. She further explained that she usually does not swear and she does not want to be perceived as a negative person.

Both S-1 and S-2 said that the nudges made them "stop and think" and review and edit their posts. Although, S-3 volunteered that she only paid attention the first few times she saw the warning, ignoring it afterwards; she said she edited a few of her posts because of typos during the timer countdown. She also remembered canceling a post: "It was a link to a funny story. I just realized other friends had already posted it so I canceled the post." Her collected data further shows that her post frequency was reduced on average by almost four posts per day. We also found fewer (7) instances of sensitive content during the treatment period than during the control period (13). In contrast, S-4 commented that each time he saw the sentiment warning he was given a positive score, which he thought was nice since "I do not want to be perceived as a jerk," but it did not have any effects on his posting habits. He further explained that as he is usually careful with what he posts, the sentiment nudge was not particularly useful to him. Behavioral data collected through the plug-in aligns with his claims, since no sensitive content was found nor were changes in posting habits detected.

S-7 became annoyed when he saw the negative sentiment warning. He posted, "Also, apparently if I cuss on facebook I now get a warning that some people may find my post negative. As if I give a fuck." In another post that he ended up canceling, he claimed, "Now I just want to post a shit-ton of bad words and see how facebook reacts to each one." These remarks show the potential negative effects of a sentiment warning and the importance of considering the form, style, and tone of the feedback given to users.

5.4 Perceived Benefits and Drawbacks

We asked participants, "Do you see any benefits from a Facebook interface like the one you tested?" Four out of seven interviewees in the timer or sentiment nudge mentioned the opportunity to stop and think as a benefit. Two of those participants also mentioned that it could deter people from posting trivial things. T-4 explained that the timer nudge helped him to post "better quality versus quantity." The same participant added that the timer nudge could prevent people from posting "politically incorrect statements." T-1 and T-2 also mentioned the timer nudge could be useful to correct typos. Three out of four interviewees who tested the profile picture nudge mentioned that it could be useful to remind those users who use customized groups to select the right group for each post. P-1 further mentioned that it could help to remember who is in each group. Moreover, P-3 mentioned that it was useful at creating awareness about who can see her posts, and P-2 thought it was a good reminder to clean up her friends list and to be cautious about what to post.

Apart from encouraging users to stop and think because of the time delay, the sentiment nudge was not perceived as being as useful as the other two nudges. Overall, users believed that the sentiment algorithm was taking isolated words and missing the context. However, S-3 recognized that it could be useful for people posting while in an emotional state. Towards the end of the interview, when the sentiment nudge was shown and explained to T-1, she disliked it because "sometimes people post things that might sound negative, but they need others' empathy and support." P-3 also thought the sentiment meter was not very useful for her; she added that the algorithm could "misinterpret sarcastic comments." However, she said it could be useful for people who had problems controlling their emotions. She mentioned children with autism as an example of those who could benefit from the sentiment nudge. P-4 also commented that the timer could help to cool people down when they engaged in a heated exchange of posts.

The downsides mentioned by our interviewees were mainly associated with performance issues such as Facebook page lag, posts not getting through or delayed posting. Nevertheless, participants appreciated the benefit of the nudges. In the words of P-2 "[there were] some technical things but the concept of having something there to remind you was fine."

5.5 Exit Survey Opinions

In the final survey, we used both open-ended and Likert questions to collect participants' opinions about the nudges they were shown. From the responses to the open-ended questions, we noticed that participants' opinions were significantly affected by some of the performance issues they experienced with the nudges. This distracted their attention from the actual functionalities of the nudges. In particular, due to technical difficulties that arose from changes rolled out by Facebook, the timer and sentiment nudges temporarily prevented posts from showing up.

Nevertheless, some of the participants valued the options offered by the timer nudge. In particular, when answering the survey question about whether our Facebook application was helpful in any way, T-3 typed, "[I] Had time to think about what I posted and whether or not I really wanted to be represented in that way." T-7 further reported that the option to cancel "was interesting." Similarly, S-1 also believed the time delay was particularly useful; she said, "I liked the time available to cancel or edit a post."

As discussed earlier, we were unable to show profile pictures for every post that participants made. As a result, participants in that treatment were not exposed to the nudge as often as participants in the other two treatments. This issue probably prevented them from giving a completely informed opinion. For example, even when the system logs allowed us to determine that the pictures had showed up several times on some participants' Facebook pages, these participants forgot having seen them.

Towards the end of the final survey, we asked participants to rank their opinions about the likelihood of using the nudge application in their daily Facebook usage, and recommending it to a friend. We also asked about their perceived level of usefulness and comfort with it during the study period.

Overall, participants had a positive perception of the timer nudge. They were both willing to use it and believed it could be useful. In contrast, opinions of the sentiment and profile pictures nudges were mixed. Participants perceived benefits from the sentiment nudge, but they mainly stemmed from the time delay and the opportunity it provided to stop and think. Participants mostly did not like the sentiment warnings, which we will discuss in details in the next section.

Opinions captured from Likert questions about the profile picture nudge did not show a particular positive or negative trend. We attribute this result to the fact that participants in this treatment only saw the profile pictures a few times, making it difficult for them to make an informed judgment about the nudge. However, as we discussed in the previous subsections, participants expressed a more positive opinion of the profile picture nudge during the interviews.

6. DISCUSSION

The objective of our nudges was to help prevent users from making online disclosure that they will later regret. Consistent with the tenets of soft-paternalism, our nudges did not limit participants' ability to post on Facebook. Instead, they encouraged participants to reflect on their posts and their audience. In general, we found that our nudges did not greatly inconvenience our participants, and induced positive behavioral changes in some of them.

6.1 Stop and think

Our timer nudge was designed to encourage users to stop and think, so as to avoid regrettable, "heat of the moment" posts. We observed that this nudge was often successful in helping users reconsider their posts. It had an additional benefit of helping users catch typos and minor errors in their posts. Some participants rephrased or even canceled their posts during the timer delay. However, this benefit comes at the cost of delaying every post participants made. Although we did provide a "post now" button, some participants wished it were more salient. Increasing the saliency of this button might lead users to get into the habit of clicking it without thinking, which would undermine the effectiveness of the nudge. Further research on time delay nudges might explore adjusting the duration of the delay, allowing users to customize this duration, or varying the delay automatically based on factors such as number of words in a post. Research might also consider other mechanisms that might nudge users to stop and think without imposing a delay.

6.2 Content feedback

Our sentiment nudge was designed to help make users more aware of how others might perceive their posts, since past research has found that posts that are perceived as very negative or contain sensitive topics are among those most regretted [23]. However, participants who received sentiment warnings did not find them useful. Participants seeing only positive scores believed the feedback was needless since they were already being careful with their posts. Participants who saw negative scores often disliked the negative feedback because it did not account for the post's context; in addition, they tended to dislike the feeling of being judged. Other difficulties with our sentiment nudge implementation were its inability to identify sarcasm and its inability to distinguish potentially damaging negativity in posts from more benign expressions of negativity. However, a number of participants agreed that a similar nudge could be useful for younger, less mature Facebook users. Further work might focus on improving the feedback algorithm; or by allowing users to customize it based on their past posts and typical vocabulary, or by providing a list of words they would like to avoid posting.

6.3 Pay attention to the audience

Our picture nudge was designed to remind Facebook users of who can see their posts, as prior research has found that users often forget who their Facebook friends are or have trouble understanding their privacy settings [23]. This feature was positively received by participants, and seemed to have improved some participants' behavior. Showing profile pictures of people who might see a given post encouraged users to be more aware of and more cautious about their posts. For example, one participant adjusted her privacy settings in response to the nudge, and another reconsidered the size of her friend list. These anecdotes suggest that this nudge can assist users with making better privacy decisions at least in some situations. This nudge might be further improved by refining the number of pictures, the algorithm for selecting pictures, and the proximity of the pictures to the posts; or by providing additional cues about the audience.

6.4 Limitations

Conducting our investigation as a field study provided the advantage of users interacting with our nudges in a natural environment. However, it also introduced difficulties, such as external factors influencing participants' posting behavior. Further, while we were able to observe posts made using our Chrome plug-in and Facebook application, we were unable to analyze posts the participants may have made using other browsers. We also experienced technical difficulties when Facebook implemented changes to its interface.

Our recruitment was affected by biases. Our plug-in was designed for users of the Chrome web browser, and participants were informed that their Facebook activities would be monitored. Therefore, our sample might be biased towards users with fewer privacy concerns and and with browser preferences different from that of the general population of Facebook users.

Measuring the effectiveness of our nudges in preventing regret is challenging because only a small fraction of the posts made by users lead to regret, and arguably even fewer lead to the short-term regret we could detect in this study. Instead, we could measure when a participant modified his or her post in response to a nudge. In addition, it is often difficult to measure the effect of a nudge; users may not react to them in a noticeable way, or the reaction might be gradual.

Some of our participants reported that they began to ignore our nudges after several days. Future work might investigate this habituation effect and how to mitigate it—for example, by varying the presence or content of the warning messages. Nudges could also be designed to appear only when a warning is needed (e.g., a post contains controversial topics), rather than appear for every post. However, determining when to display a warning is in itself a challenging research question. Alternatively, a more interactive system, similar to ELIZA, could be used to make nudges more engaging so that users will not ignore it after some time [24].

Despite these limitations, this study provides interesting preliminary results and directions for future work. With further refinements, our experimental platform will be useful for conducting large-scale, longitudinal field trials, testing a variety of nudges.

7. CONCLUSIONS AND FUTURE WORK

We designed three mechanisms that nudge users to consider their online disclosures in social media more carefully. These mechanisms, Based on the concepts of soft paternalism and choice architecture, provide visual cues about audience, time delays, and feedback. We developed these three privacy nudges on Facebook by implementing a nudging platform comprised of a Chrome plug-in and a Facebook application. We conducted an exploratory field trial and follow-up interviews to investigate the impact of these nudges on Facebook users' posting behavior. We found that two of our nudges, introducing a delay when a user attempts to create a post on Facebook and showing the profile pictures of other people who might see the post, were perceived as useful and had a positive effect on some users' posting behavior.

While our study was exploratory, our results suggest that privacy nudges could potentially be a powerful mechanism to discourage unintended disclosures in social media that may lead to regret. Although we provide a Facebook case study, this idea of privacy nudges can be extended to other domains such as e-commerce, location sharing, and mobile applications.

We collected participants' ideas to improve the design of the nudges and are currently working on both improving the nudge designs and making our system more resilient to Facebook changes. We plan to conduct more field experiments on both Facebook and other application platforms. We believe that larger-scale and longitudinal field studies are desirable to quantitatively assess the impact of these interventions in both the short and long term. Finally, we advocate the privacy nudging approach and encourage other researchers to explore the rich design space of nudging to help protect people's privacy.

8. ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation under Grant CNS-1012763 (Nudging Users Towards Privacy), by the IWT SBO Project on Security and Privacy for Online Social Networks (SPION), and by Google under a Focused Research Award on Privacy Nudges.

9. REFERENCES

[1] A. Acquisti. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM conference on Electronic commerce*, pages 21–29. ACM, 2004.

[2] A. Acquisti. Nudging privacy: The behavioral economics of personal information. *IEEE Security and Privacy*, 7(6):82–85, 2009.

[3] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Privacy enhancing technologies*, pages 36–58. Springer, 2006.

[4] A. Acquisti and J. Grossklags. Privacy and rationality in individual decision making. *Security & Privacy, IEEE*, 3(1):26–33, 2005.

[5] d. boyd and N. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 2007.

[6] J. C. Brustoloni and R. Villamarín-Salomón. Improving security decisions with polymorphic and audited dialogs. In *Proceedings of the 3rd symposium on Usable privacy and security*, SOUPS '07, pages 76–85, New York, NY, USA, 2007. ACM.

[7] L. Fang and K. LeFevre. Privacy wizards for social networking sites. In *Proceedings of the 19th international conference on World wide web*, pages 351–360. ACM, 2010.

[8] B. J. Fogg. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann, 1 edition, Dec. 2002.

[9] A. Forget, S. Chiasson, P. C. van Oorschot, and R. Biddle. Improving text passwords through persuasion. In *Proceedings of the 4th symposium on Usable privacy and security*, SOUPS '08, New York, NY, USA, 2008. ACM.

[10] E. Goffman. *The Presentation of Self in Everyday Life*. Anchor, 1 edition, June 1959.

[11] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *WPES '05: Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, 2005.

[12] L. K. Hansen, A. Arvidsson, F. F. Å. Nielsen, E. Colleoni, and M. Etter. Good friends, bad news - affect and virality in twitter. In *The 2011 International Workshop on Social Computing, Network, and Services*, Jan. 2011.

[13] D. Laibson. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478, May 1997.

[14] H. R. Lipford, J. Watson, M. Whitney, N. Carolina, H. Lipford, K. Froiland, and R. W. Reeder. Visual vs . Compact : A Comparison of Privacy Policy Interfaces. *Interfaces*, pages 1111–1114, 2010.

[15] M. Madejski, M. Johnson, and S. M. Bellovin. The failure of online social network privacy settings. Technical Report CUCS-010-11, Department of Computer Science, Columbia University, February 2011. In submission.

[16] A. Marwick and d. boyd. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114, 2011.

[17] R. G. Miltenberger. *Behavior modification: Principles and procedures*. Wadsworth Publishing Company, 2011.

[18] F. Å. Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *ESWC2011 Workshop on Making Sense of Microposts*, Mar. 2011.

[19] Rebecca Balebako, Pedro Leon, Hazim Almuhimedi, Patrick Kelley, Jonathan Mugan, Alessandro Acquisti, Lorrie Cranor, and Norman Sadeh. Nudging users towards privacy on mobile devices. In *Proceedings of the 2nd International Workshop on Persuasion, Influence, Nudge & Coercion through mobile devices*, 2011.

[20] H. A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118, Feb. 1955.

[21] R. H. Thaler and C. R. Sunstein. Libertarian paternalism. *American Economic Review*, 93(2):175–179, May 2003.

[22] R. H. Thaler and C. R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, 1 edition, Apr. 2008.

[23] Y. Wang, S. Komanduri, P. G. Leon, G. Norcie, A. Acquisti, and L. F. Cranor. "I regretted the minute I pressed share": A qualitative study of regrets on facebook. In *Proceedings of the 7th Symposium on Usable Privacy and Security*, 2011.

[24] J. Weizenbaum. ELIZA - a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, Jan. 1966.