A Nudge-based Recommender System Towards Responsible Online Socializing

Rim Ben Salem[a], Esma Aïmeur[a] and Hicham Hage[b]

[a]Department of Computer Science & Operations Research, University of Montreal, Montreal, QC, Canada
[b]Science Department, Notre Dame University-Louaize, Zouk Mosbeh, Lebanon

Abstract

In recent years, the popularity of social media has been on the rise. Driven by a multitude of motivations, users have grown accustomed to sharing online most aspects of their lives. This self-disclosure has not only proven to be dangerous for peoples' privacy and security but also harmful to their personal, professional and intimate relationships. However, in the age of social media, it seems improbable for people to completely discontinue using social media platforms such as Facebook, Twitter and Instagram. Hence, there is an urgent need to find a balanced compromise between online sociability and privacy.

In this work we propose a platform to mitigate the dangers of self-disclosure through the use of a personalized harm-aware recommender system. Specifically, the recommender system balances the requirements for privacy protection with the users' need to share with their social circles. To achieve this, the platform starts by evaluating the risks of disclosing the personal information and then, if necessary, proceeds to recommend to the user how to reduce that risk. While the evaluation of the risk is done in an objective manner, personalization is of the essence since users have different preferences and sharing needs. As such, when performing the recommendation, the systems will provide personalized nudge-based recommendations, raising the users' awareness of the privacy issues stemming from self-disclosure.

Keywords

Harm-aware recommender system, self-disclosure, privacy, personalization, nudge-based recommendations

1. Introduction

Today, Social Networking Sites (SNS) have become very versatile, providing users with a wide range of functionalities: from posting messages, photos and videos, to playing games online, shopping and finding a job. Due to the variety of SNS, the user's information and data can vary greatly from one to another. This variety of SNS platforms and services, along with the broad appeal and extensive use of SNS, creates a wealth of information on users. A report by the Pew research center indicates that about 75% of the American public uses more than one SNS platform, and the typical American uses three of these sites [1]. Moreover, the report also indicates that younger adults tend to use a greater variety of social media platforms.

The initial optimism about the positive potentials of the Internet and social media has given way to concerns about the constant harvesting of personal information. Indeed, SNS actively

OHARS'20: Workshop on Online Misinformation- and Harm-Aware Recommender Systems, September 25, 2020, Virtual Event

EMAIL: kulyabov-ds@rudn.ru (R. B. Salem); i.tiddi@vu.nl (E. Aïmeur); i.tiddi@vu.nl (H. Hage)

encourages the collection and sharing of information, by gradually increasing the variety and types of collected data, as well as relying on more revelatory default visibility settings [2].

While social media platforms are expected to safeguard the user's data and information, this task becomes much harder when the human factors are considered.

Indeed, while users in general report a high concern for their privacy, they tend to have a privacy-compromising online behavior, and an unprecedented level of self- disclosure (the act of voluntarily sharing and disclosing personal information to others) activity is taking place. This contradiction between privacy attitudes and actual behavior is generally referred to as the privacy paradox [3]. Essentially, one would logically expect that users' privacy concerns would restrict the voluntary disclosure of personal information. However, the reverse effect is observed where users tend to share private information in exchange for retail items and personalized services, or through their social network activities [4].

This has existed long before the advent of the internet and social media. In 1973, psychologists Irwin Altman and Dalmas Taylor formulated the theory of social penetration [5]. It theorizes that the more people disclose things about themselves, the closer they get to those with whom they share said information. This applies to friendships [6] and romantic [7] relationships where this reciprocal act is regarded as necessary to build and maintain the interpersonal ties.

Moreover, there are many other reasons why people feel motivated to divulge their personal and sensitive information, ranging from seeking fame, attracting brand sponsorships, release pent-up feelings [8], social validation [9, 10] to non-lucrative altruistic objectives like benefiting others with life experience [11]. These personal motives lead to different assessment of the value of one's personal data [12]. Users looking to connect with people who reside close to them are likely to easily disclose their location over another piece of data such as their medical record. In addition to this, other parameters factor in peoples' choices such as their background, thoughts, interests and prior knowledge of the consequences of publishing said piece of information [13]. Hence self-disclosure is highly dependent on the person's subjective evaluation of their data. As such, approaches to mitigate the dangers of self-disclosure need to be equally user specific and address the individual rather than the crowd.

The following scenario serves to further explain the issue and the proposed solution: Alice is usually careful and aware of dangers on privacy and cybersecurity online. Recently, she has been laid off her current job due to the economic struggle following the COVID-19 outbreak. While browsing through her preferred social media platform, she stumbles upon a post for a "work from home" job offer, providing an email address for applicants and requesting information such as a curriculum vitae, social security number and address. This is one of the biggest scams that have been targeting vulnerable people seeking an alternative income. Alice is usually careful and aware of menaces on privacy and cybersecurity, but her judgement is clouded by her current mental state and she is unable to objectively assess her situation. If she proceeds without being advised otherwise, she might turn into a fraud victim.

In this case, there is a need for a system which already identified Alice's privacy preferences, knows that her current conduct is dangerous and can advise her against this course of action. This system would intervene and nudge Alice telling her "BEWARE! Your are about to share information which can easily lead to your identity being stolen. Consider deleting the information 'social security number' to reduce the risk". This paper proposes a platform for this purpose, to guide users towards aware disclosure. The proposed system considers the user's

assessment of their data as well as an objective evaluation and finds middle ground between the two.

In dealing with this issue, this work makes the following contributions:

- Propose an objective threat assessment model for computing the privacy risk based on an input vector representing the user's disclosed data.
- Detail an adaptive nudge-based recommender that balances on one sidethe objective risk assessment and on the other side the user's sharing needs and privacy preferences.

The paper is organized as follows: section 2 discusses existing work that relates to our proposition. Section 3 details the recommender system and the different submodules of the architecture. Section 4 reports on the evaluation of the platform and section 5 concludes this work with a discussion of the contributions provides pointers on future work.

2. Related Work

Trepte and Masur [14] define privacy as a need to control who has access to personal information and a form of solitude, intimacy, anonymity, or reserve. It is also considered to be the key to fulfilling basic human needs such as the need for autonomy and protected communication [15]. These needs differ from one person to the other and they are highly behavior-dependent [16]. Petronio and Durham [17] relate private information divulgence to their perception of their ownership over the data. This further corroborates that privacy preservation needs to be personalized as the act of disclosure itself is user-based.

People evaluate the risks and the perceived gratification [18] and, depending on their personality as well as the situation, one can outweigh the other resulting in the decision to veil or unveil the data. However, such decisions can be affected by bias and misconceptions [4, 19], as well as the user's background such as skills, experience and education [20]. In fact, some users consider clearing cookies to be the highest form of awareness and privacy preservation [21] while for others, the most frequent protective decision is antivirus scans [22].

It is for this reason that the need for an objective party to help assess the situation has emerged. The user's decision-making abilities should not be discarded but aided instead. Bandura [23] confirms that a certain behavior can be encouraged or deterred provided that the person receives a prediction of positive incentives or detrimental consequences. This is where our proposition plays a major role to either validate or discourage the self-disclosing user post and provide personalized alternatives in the latter case.

Other studies aiming to reduce self-disclosure include research on the correlation between changing privacy settings and revealing personal data [24]. The study concludes that simply limiting the audience does not equal more control over sensitive credentials. This further supports the urgent need to find a solution to manage the user's input rather than focus on platform-specific configurations and antimalware.

In particular, nudge-based mechanisms are garnering interest in awareness raising contexts. Nudges propose positive reinforcements and suggestions. They are used for cybersecurity and privacy preservation in order to encourage users to adopt aware behavior and to reflect on

their behavior in a non-obstructive way. They can be a one-fits-all where they depend on the scenario rather than the user [25] or they can be tailored to the user [26, 27].

Nudged are used in a multitude of contexts, from helping adolescent SNS users avoid privacy and safety threats [28] to Blockchain-based open banking [29]. Another example is the personalized privacy assistant for mobile app permissions [30], which proved its effectiveness with 78.7% of its recommendations being accepted and implemented by the user. Similarly, another work [31] aiming to address the privacy paradox has shown that users tend to reflect on their behavior after receiving nudges. The authors compare the device's general settings with the permissions granted to a specific app and notes the discrepancies. Although this comparison does reveal the user's bias towards one app, if that is the case, the general settings are also subjective. So, this work compares a general level of subjectivity and a specific one and ideally, they are the same. However, this does not truly reveal the user's deviation from advised practices in an effort to correct them.

We believe that having an objective assessment along with the personal judgement has the potential to not only get closer to the user's preferences but also mitigate the risk of self-disclosure, which is why we adopt this approach. The following section details our proposition of the harm-aware recommender system.

3. Personalized Nudge-based Recommender System

This section details the proposed approach for a personalized, nudge-based recommender system. However, the recommender system is one component of a larger platform. Consequently, to help the reader better understand the design and function of the recommender system, the next subsection highlights the general architecture of the platform (Figure 1), briefly introducing the various components.

3.1. General Overview of the Platform

The platform relies on user modeling to study the user's tendencies and preferences over time. This is then utilized along with domain knowledge to empower the personalized recommender system and mitigate the risk of self-disclosure. The user model is used to capture characteristics of the individual including motivation, objectives and cognitive bias. A personalized approach must also consider the user's trust circle (where the user feels comfortable revealing personal data), which includes both the platforms and the human counterparts.

Another major component is the domain knowledge which, contrarily to the user model, does not focus on individuals but rather on the general processes and conclusions. This includes the language model which serves to process the text input and to "understand" it. Other functions include devising the subjective assessment model as well as studying the privacy tendencies on the platform.

The Information Extraction component analyzes the raw text input of the user, determines the exact disclosed data, ultimately generating the threat vector. It has the disclosed information as well as their disclosure rate so each piece of data is represented as: (x_i, y_i).

Finally, the user model, the domain model and the threat vector are sent to the recommender system, the focal point of this paper. The system orchestrator serves as the medium through

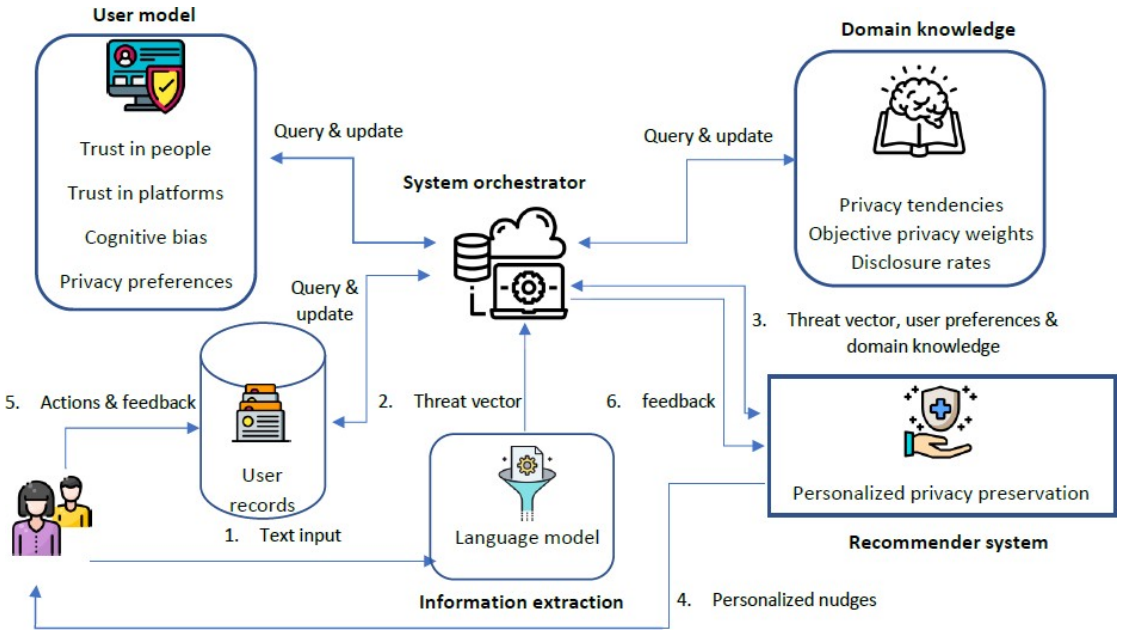

Figure 1: General overview of the platform.

which the different modules communicate.

3.2. The Recommender System Mechanism

This subsection focuses on the main contribution for this paper which stems from the recommender system mechanism. It is worth noting that it can be used as a standalone module (browser plugin for example) or integrated within other platforms. Figure 2 is a view of the recommender system's modules which are detailed next.

3.2.1. Threat assessment module

The objective of this module is to provide an objective assessment of the risk. The value *Risk* is user input-specific but not user preference-specific. Specifically, the value of *Risk* depends solely on the input. As such, if two users, with completely different preferences disclose the same pieces of data, they would have the exact same risk value.

This is done using the objective threat function as follows:

$$Risk = \sum_{i=1}^{n} W_{i,obj} \cdot f(x_i, y_i) \tag{1}$$

x_i corresponds to a component of the data vector X. Each component is a piece of personal data such as location or social security number.

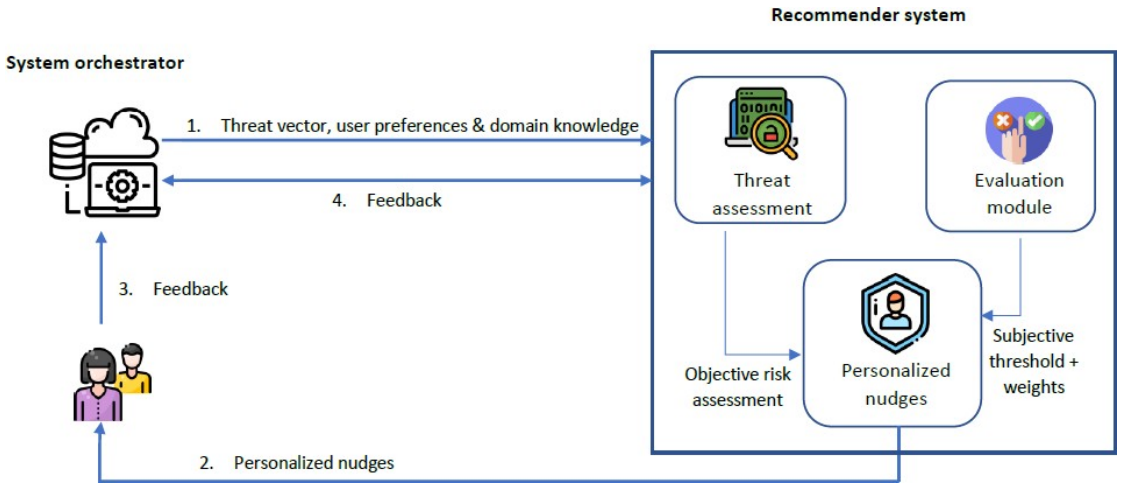

Figure 2: The architecture of the recommender system. The recommender mechanism starts with the threat assessment after receiving data from the orchestrator.

y_i is a component of the disclosure vector Y. For x_i=legal name, y_i= first name or y_i= last name.

$W_{i,obj}$ is the objective weight/risk of disclosing x_i.

$f(x_i, y_i)$ disclosure rate: is the value of disclosing the information y_i out of x_i. This varies between 0 and 1. The lower end means that the user didn't disclose any portion of that information and 1 refers to a complete disclosure of datum x_i.

n the total number of personal data that the system considers to be private.

To the best of our knowledge, there are no formal objective measure to weigh the risk of disclosing certain pieces of personal information. As such, we devised on a novel approach to determine the importance (weight) of personal information $W_{i,obj}$, based on their price on the dark web. A piece of data is considered to be most sensitive and its disclosure most costly when it has the highest price tag. These values were collected from different reports and studies done by various sources including *Experian*, *TransUnion*, *Atlas VPN*, *Safety detectives*, *Keepersecurity*, *Symantec*, *Statista*, and *Pace technical* [32, 33, 34, 35, 36, 37, 38, 39]. Table 1 illustrates some of the data and their values. These values are fixed by the system for all users.

Table 2 shows the results of the raw data preprocessing by binning them into buckets from least to highest sensitivity.

Each class is then represented by a normalized interval's mean value to reduce the noise of the raw data. This newly calculated value corresponds to the objective weight of each information belonging to said class. For example, the weight of an information belonging to the "low sensitivity" bucket is 2.25.

In Table 3, scenarios are defined, and their objective risk values are calculated using the weights in Table 2 and the estimated values of the disclosure rate $f(x_i)$. This is the method

Table 1
Prices of personal data on the dark web.

Personal data	Min Value ($)	Max Value ($)	Sensitivity
Date of birth	1	1	Least
Address	1	1	Least
Social security number	1	4	Least
E-mail and password	5	30	Low
Passport (1) / ID (2) / License (3)	200	500	High
Diploma	300	656	Highest

Table 2
Data binning.

Classes	$1 \leq cost \leq 26$	$26 < cost \leq 101$	$101 < cost \leq 201$	$201 < cost \leq 352$	$352 \leq cost \leq 453$
Normalized classes	$1 \leq cost \leq 1.5$	$1.51 \leq cost \leq 3$	$3.01 \leq cost \leq 5$	$5.01 \leq cost \leq 8$	$8.01 \leq cost \leq 10$
Sensitivity	Least	Low	Medium	High	Highest
Representative element	1.25	2.25	4	6.5	9

used to estimate values of $f(x_i)$: If the data is fully disclosed: $f(x_i) = 1$ and if the data is fully undisclosed: $f(x_i)$=0.

If a portion of the data is disclosed: Suppose that the complete data x_i has n pieces y_i ranked from least to most user specific, example for x_i =credit card number, the first 4 digits (y_1) that are bank-specific are worth less than the rest of the number (y_2) which are more user-specific. To put a value of how much of x_i does y_1 account for, we define the function f in general as follows:

$$
\begin{aligned}
&f(x_i, y_{1+k}) = 2 * f(x_i, y_k), \forall\ 1 \leq k < n \\
&f(x_i, y_1) + f(x_i, y_2) + \ldots\ldots\ldots + f(x_i, y_n) = f(x_i, x_i) = 1 \\
&f(x_i, y_1) + 2f(x_i, y_1) + \ldots\ldots\ldots + f(x_i, y_1) = 1 \\
&f(x_i, y_1)\left(1 + 2 + 4 + \ldots + 2^{n-1}\right) = 1 \\
&f(x_i, y_k) = \frac{1}{2^n - 1} \cdot 2^{k-1}
\end{aligned} \tag{2}
$$

Example: in Table 3, scenario 2, the user disclosed their year of birth which is the most specific in comparison with the day and month of birth. $y_3 = 2^2 \cdot \frac{1}{2^3-1} = 0.57$

The exception to this would be the address because disclosing the zip code for example, even without explicitly stating the city, province and country, is an implicit disclosure of all of these pieces of data. As such, we define the f values for country, province, city, zip code, building consecutively as 0.125, 0.25, 0.5, 0.75,1. This is applied to scenario 1 in Table 3.

Going back to the main scenario in the introduction where Alice is about to disclose her address and social security number, x_1 is the address and x_2 is the SSN, $Risk = W_{1,obj} \cdot f(x_1, y_1) + W_{2,obj} \cdot f(x_2, y_2) = 1.25 * 1 + 1.25 * 1 = 2.5$.

Table 3
Experimental scenarios

	Scenario	Data sensitivity	$f(x_i)$	Risk value
1	Posting a FB status: The government just announced that our **province of Quebec** was hit the worst during the covid19 pandemic.	Least	0.25	0.31
2	Posting a tweet: I could see the accident that was reported on the news at **Newstreet** from my window at home and it was the most horrific thing I've seen in my **30 years of existence**.	Least	0.75 Location 0.33 date of birth	1.65
3	Sending a message to a friend: I'm expecting an email response to my job interview but I will be visiting my parents in the countryside and the internet service is bad there. Can you please check my email daily for me ? Here are my **email and password**: xxxxxxxx xxxx	Low	1	2.25
4	Phishing message on LinkedIn asking for professional email: The malicious person pretends to request professional email for future cooperation and the user discloses it.	Medium	0.33	1.32
5	Responding to phishing scam email claiming to be from the Immigration, Refugees and Citizenship department asking for a full scan of the receiver's **passport**. The user sends his full passport.	High	1	6.5

The aforementioned and recorded values are non-user specific as they define the objective values. The following module shows how the user preferences are used along with these objective weights to personalize the nudges.

3.2.2. Personalized nudge module

This module takes as input the threat vector and the calculated *Risk*. The latter is then compared with the user's *subjective threshold*. Specifically, this threshold represents the user's privacy tolerance, that is how much information they are willing to share. This is an important aspect of personalization since different users have different sharing/privacy needs. If ($Risk < Threshold$), it means that the disclosure is within the user's tolerance, and the process terminates. However, if ($Risk \geq Threshold$), it means that the risk exceeds the user's tolerance, and the system must notify the user and recommend (nudge) an appropriate course of action to reduce the risk below the threshold. Another important aspect of personalization is the user preferences for each piece of data that are called subjective weights $W_{i,user}$. Each data x_i has an objective weight $W_{i,obj}$ and a subjective $W_{i,user}$.

To do so, we start by defining the set $Y = \{x_i, Risk - threshold \leq W_{i,obj} \cdot f(x_i)\}$. Specifically, Y is a set of candidates (pieces of data x_i) whose elimination alone reduces the value of objective threat function *Risk* below the personalized threshold. Y can either be:

- $Y \neq \emptyset$, meaning that there exists at least one piece of data x_i that, when removed, will reduce the risk below the threshold. If multiple candidates exist, we choose the one to delete x_{nudge} using this user preference-based formula:

$$x_{nudge} = \arg \min_{1 \leq i \leq |Y|} \left(W_{i,obj} \cdot f(x_i) \cdot W_{i,user} \right) \tag{3}$$

- $Y = \emptyset$, meaning that there is not one single piece of data that would reduce the risk below the threshold. In this case, the system recommends to the user to delete multiple pieces

of disclosed data to reach the risk tolerance level. In this case we execute the following pseudo code:

$$
\begin{aligned}
&Y \leftarrow X \\
&repeat\ :\ x_{nudge} = \arg\min_{1 \le i \le |Y|} \left(W_{i,obj} \cdot f(x_i) \cdot W_{i,user}\right) \\
&\qquad X_{nudge} \leftarrow x_{nudge} \\
&\qquad Risk \leftarrow Risk - W_{nudge,obj} \cdot f\left(x_{nudge}\right) \\
&\qquad Y \leftarrow Y \setminus \left\{x_{nudge}\right\} \\
&until\ \ (Risk \le threshold)
\end{aligned}
$$

X_{nudge} is the final output and it's the set of data to be recommended to the user. In particular, Alice who is highly aware of privacy menaces, has a threshold equal to 2, and the risk of her current action is 2.5 as calculated in the previous section. If her preference for SNN is 1 and for location is 2, in this case the nudge would be to delete SSN.

At this point, Alice has received the recommendation and the process that follows her choice is detailed in the evaluation module section.

3.2.3. Evaluation module

This process is based on the user's action after being recommended the personalized nudges. The user can refuse the nudge or accept it. In the first case, the user shares the content as is, without modifications. In the second case, the user either accepts the recommendation and reduces the risk as proposed by the system, or might reduce the risk based on their own discretion. All these cases are considered as a form of implicit feedback which is used to update the values of the threshold and $W_{i,user}$. Both these values are user-specific parameters on which the personalization is based.

i Begin with reevaluating the *Risk* after the user's action:

- User doesn't delete any piece of data: *Risk* remains the same
- User deletes one or more pieces of data. Supposing that the user deletes the following Deleted_data $= \{x_k, \dots, x_j\}$:

$$Risk = Risk - \sum_{i=k}^{j} W_{i,obj} \cdot f(x_i) \tag{4}$$

ii Then we update the threshold as follows:

$Threshold = Threshold + \text{softsign}\,(risk - Threshold) \cdot \alpha, 0 < \alpha < 1$ is the learning rate

- if *Risk–threshold* < 0, either the risk of the user's original input was already below the threshold or the user fully accepted the recommendation to delete the pieces of data whose weight exceeds the threshold: $-1 <$ softsign(*Risk–threshold*) < 0. As a result, the threshold is reduced.

Table 4
The result of applying this process to the user Alice

User	Initial values	Risk	X_{nudge}	User's decision
Alice	Threshold=2 Subjective weights: SSN =1 Address =2	2.5	Risk > threshold Recommend deleting social security number	Accept nudge: $New\ Risk = 2.5 - 1.25 = 1.25$ $W_{i,nudge} = 2.5 + \text{softsign}(1.25 - 2) = 1.57$ Refuse nudge: $Risk = 2.5$ $Threshold = 2.5 - \text{softsign}(1.25 - 2) = 2.42$

- if $Risk–threshold \geq 0$, the user either deleted partial pieces of data or completely ignored the recommendation and proceeded to post the input as it is. $0 < \text{softsign}(Risk–threshold) < 1$. As a result, the threshold is increased.

iii We update the weights using a similar formula, for each $x_i \in X_{nudge}$:

- If the user deletes x_i:

$$W_{i,user} = W_{i,user} + \text{softsign}\left(W_{i,obj} - W_{i,user}\right) \cdot \alpha, 0 < \alpha < 1 \text{ is the learning rate}$$

- If the user does not delete x_i:

$$W_{i,user} = W_{i,user} - \text{softsign}\left(W_{i,obj} - W_{i,user}\right) \cdot \alpha, 0 < \alpha < 1 \text{ is the learning rate}$$

In Table 4, this process is applied to Alice's scenario and calculations are made for both cases when she does and when she does not accept the nudge.

To simplify the weight update , we use $\alpha = 1$ but in the evaluation section we discuss the impact of the value of this parameter on the system's recommendations. Table 4 shows how the threshold is increased when the user ignores the nudge and vice versa. Finally, the algorithms for recommendations and updating the user preferences are detailed below.

Algorithm 1 Recommendation algorithm

Input: vector X, W_{obj}, function f, float *threshold*

$Risk \leftarrow Risk - \sum_{i=1}^{n} W_{i,obj} \cdot f(x_i)$
if $Risk \leq threshold$ **then**
end
end if
$Y = \{\}$
for all i *in* $[1, size(X)]$ **do**
if $\left(Risk - threshold \leq W_{i,obj} \cdot f(x_i)\right)$ **then**
$Y \leftarrow Y \cup \{x_i\}$
end if
end for
if $Y \neq \emptyset$ **then**
$x_{nudge} \leftarrow \arg\min_{1 \leq i \leq |Y|} \left(W_{i,obj} \cdot f(x_i) \cdot W_{i,user}\right)$
else
$Y \leftarrow X$
$X_{nudge} \leftarrow x_1$
while $Risk \geq threshold$ **do**
$x_{nudge} \leftarrow \arg\min_{1 \leq i \leq |Y|} \left(W_{i,obj} \cdot f(x_i) \cdot W_{i,user}\right)$
$X_{nudge} \leftarrow x_{nudge}$
$Risk \leftarrow Risk - W_{nudge,obj} \cdot f(x_{nudge})$
$Y \leftarrow Y \backslash \left\{x_{nudge}\right\}$
end while
end if

Algorithm 2 User preference adaptation algorithm

Input: Set Deleted_data, X_{nudge}, float $Risk$, $Threshold$, α, Vector W_{user}, W_{obj}

for all x_i *in* Delete_data **do**
$Risk \leftarrow Risk - W_{i,obj} \cdot f(x_i)$
$W_{i,user} \leftarrow W_{i,user} + \text{softsign}(W_{i,obj} - W_{i,user})$
end for
for all x_i *in* $X_{nudge} \backslash$Deleted_data **do**
$W_{i,user} \leftarrow W_{i,user} - \text{softsign}(W_{i,obj} - W_{i,user})$
end for
$Threshold \leftarrow Threshold + \text{softsign}(Risk - Threshold)$

4. Evaluation

In order to evaluate the proposed approach, a proof-of-concept experiment was designed. The main goal of the evaluation is to study how the user's choices impact the process of personalization. The criterion for testing the quality of personalization is how fast does the initial threshold converge to the user's actual threshold as in his actual preferences. The

Table 5
Updated threshold depending on the user, scenario and learning rate.

	scenario	Risk value	α	John	Bob	Anna	Sam	Catherine
Initial threshold				1	4.25	6	7.5	8.75
Actual threshold				2	3	8	5	6
	1	0.31	0.5	1.29	3.84	6.17	7.06	8.30
	3	2,25	0.5	1.33	3.44	6.34	6.62	7.87
	7	6,5	0.5	1.21	4.40	6.11	6.91	7.87

simulated data represents all types of users from low awareness, medium to high. One of the main concerns when using recommender systems is the cold start problem in which the user has just been introduced to the platform and his preferences are unknown. To overcome this and set initial threshold values to kickstart the recommender system-based awareness raising process, the user's preferences are estimated though a short survey. Essentially, it consists of a straightforward process where the individual estimates how much their own assets are worth to them on a scale of 1 to 10. These values as given by the user are averaged and considered to be the initial subjective weights. This is also used to compute the initial threshold which equals $10 - average - (subjective\ weights)$. After doing this, the system has the initial threshold and as mentioned previously, the system is judged on how fast this converges to the actual threshold

The next step is to see how these initial thresholds evolve over time. This depends on the scenario that defines the objective risk value, the system's set learning rate α and the user's responses to the recommendations. We select 3 scenarios out of a total of 20 conducted tests that have a significant objective risk values to demonstrate the different outcomes depending on the user.

In Table 5, users' decisions are simulated as follows:

- Users who have a low threshold value are the most concerned about their privacy and the most likely to accept to delete data as recommended by the system. The higher the risk the more likely they are to accept the nudge. For example, user 3 refused the recommendations with a 25% probability, accepted them with a 75% probability.
- Users with medium thresholds (Bob and Anna) accept recommendations with a probability closer to 50%.
- Users with high threshold are the least conscious about their privacy and as a result they refuse the recommendations with a 75% probability and accept them with a 25% probability.

In each iteration (scenario), the risk value is evaluated (values taken from Table 3) and compared to the personal threshold which is initialized based on the survey filled by each user. This comparison allows the system to determine which type of nudge it needs to make: recommend modification or encourage the post as it is if the risk does not surpass the personal threshold. Example: in scenario 7, Sam whose threshold is 6.62 after the previous scenario is encouraged by the system to reduce the disclosure but upon his refusal, the threshold is increased as it corresponds to his tolerance to risk. The threshold goes from 6.62 to 6.91 which gets the user closer to their actual threshold 5.

For the same scenario, John who displays the highest aversion to risk, accepts the recommendation which results in his threshold going down from 1.33 to 1.21 which is further from the actual threshold which is equal 2. However, that can be explained by the fact that the risk is 6.5 which is high, and a highly aware person is likely to accept to reduce the risk. The threshold converges more quickly with low risk scenarios.

Another important thing worth pointing out is that α is an experimental value that is characteristic of the system and not the users. Meaning that once we set its value, the thresholds of all users are computed using the same α value. In our evaluation, we tested α values between 0.1 and 0.9 and compared the impact of each on the convergence of the threshold. A good choice of α is imperative. A large value for α would introduce a considerable change, and this would cause the threshold to exceed and miss the targeted value. A small value of α would introduce very small changes to the threshold, and that is not desirable as well, since it will take longer to converge to the targeted value. So, if the threshold difference from one iteration to the next is negligible then the awareness-raising purpose.

We tested the outcome with values of α from 0.1 to 0.9 in increments of 0.1. For example, Catherine whose initial threshold is 8.75 , has given the address the highest score during the quiz so he might be inclined to eliminate his province after second thought. with a value α =0.9, Catherine's new threshold is 7.94 which is a 0.81 reduction from the earlier value. This seems like an improvement in terms of the user's awareness. However, it is too optimistic as in fact, this user has a minimal awareness of all of her assets. So, the next time we nudge her, unless the risk is address-related, we predict that she would ignore the system. In conclusion, maximizing the threshold reduction can have a negative impact because large jumps result in frequent unwelcomed nudges that do not reflect the user's preferences. Similarly to this, fixing $\alpha = 0.1$ results in $threshold = 8.67$ for the same user Catherine, after she performs the same action. This is not a good choice either because a user with a high threshold such as Catherine would almost never receive nudges which counters the privacy preserving purpose of the platform. After several tests, α was set to 0.5, as it provides an optimal value that has shown to converge to the user's preferences more accurately and quickly, with Low (0.31) , Medium (2.25) and High (6.5) risk levels alike.

Finally, in Table 6, a comparison is drawn between this paper's proposition and existing work on nudges for privacy preservation.

In this section, we have evaluated the platform based on users with different initial thresholds and it is then compared in terms of mechanism with existing approaches.

5. Conclusion and Future Work

Today, the proliferation of social networking sites, their reachability, ease of use as well as their integration within many aspects of our daily activities, has given rise to an unprecedented level of self-disclosure activity. Various solutions exist to help the user navigate the maze of online self-disclose, however, such systems focus mainly either on the data being disclosed (ignoring the user's needs and preferences) or on the user (ignoring the data being disclosed). As such, this paper presents a new platform that balances the risks of self-disclosure and the user's sharing needs and privacy preferences. It is a user-centric proposition based on a personalized

Table 6
Comparing our approach with existing work.

	AppOps [25]	Tailored privacy nudges [26]	Nudge Me Right [27]	Our approach
Personal data	location	Scenario-based where users are given scenarios and choose whether they disclose or not	password	12 types of data with varying degrees of sensitivity
Nudge mechanism	The nudge offers predefined responses for the user to choose with the highlighted best choice	2 nudges corresponding to disclosure options	The user is nudged to strengthen their password is it is deemed weak and the user is free to enter a new password and have it tested again. (no predefined choices)	The user is completely free to delete the data as recommended by the system or another piece of information, or do nothing, the risk is recalculated. (no predefined choices)
Personalized	no	yes	yes	yes

risk-aware recommender system. Users are guided towards privacy preservation through nudges tailored according to their perception to maximize the probability of accepting the recommendation. The contributions of this work can be summarized into two main aspects. First, this work provides an objective approach to evaluate the risk of various types of user data. Second, this work provides a novel method of modeling and evaluating the user's privacy preferences at two different levels, one for a general disclosure (the threshold), and the second more specific ($W_{i,user}$), based on the different types of data being disclosed. Finally, an evaluation of the personalization component of the recommender system is performed, to validate how well the system adapts and converges to the user's preferences.

As for future works, various aspects still require further development. First, currently, the recommender system considers each disclosure on its own and keeps track only of the user preferences. This can represent a threat for users who small nuggets of information, one at a time. These might not trigger the recommender system and go undetected. As such, the risk evaluation function proposed in this work should also incorporate other aspects from previous disclosures by the user. Second, both the risk and the user preference, in reality, are very dependent on the sharing environment as well as the people involved. As such we intend to investigate further how to incorporate these aspects within the objective evaluation of risk, and the personalized recommendation and nudges. Also, the language model needs implementation and more development to ensure both the understanding of the user input and the generation of the nudge. Finally, it would be interesting to explore how one can extend this approach to other domains, such as fake news. Specifically, an approach that would combine an objective evaluation of the post with the user's preferences and cognitive biases to nudge them away from fake news.

References

[1] A. Smith, M. Anderson, Social media use in 2018: Pew research center; 2018, Available from: htpp://www. pewinternet. org/2018/03/01/social-media-use-in-2018/.[Last accessed on 2018 May 20] (2019) 1–17.

[2] A. Acquisti, L. Brandimarte, G. Loewenstein, Privacy and human behavior in the age of information, Science 347 (2015) 509–514.

[3] A. Acquisti, Privacy in electronic commerce and the economics of immediate gratification, in: Proceedings of the 5th ACM Conference on Electronic Commerce, EC '04, Association for Computing Machinery, New York, NY, USA, 2004, p. 21–29.

[4] S. Barth, M. D. de Jong, The privacy paradox – investigating discrepancies between expressed privacy concerns and actual online behavior – a systematic literature review, Telematics and Informatics 34 (2017) 1038–1058.

[5] I. Altman, D. A. Taylor, Social penetration: The development of interpersonal relationships., Holt, Rinehart & Winston, 1973.

[6] S. M. Jourard, The transparent self: Self-disclosure and well-being, volume 17, Van Nostrand Princeton, NJ, 1964.

[7] J.-P. Laurenceau, L. F. Barrett, P. R. Pietromonaco, Intimacy as an interpersonal process: The importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges, Journal of personality and social psychology 74 (1998) 1238–51.

[8] R. Zhang, The stress-buffering effect of self-disclosure on facebook: An examination of stressful life events, social support, and mental health among college students, Computers in Human Behavior 75 (2017) 527–537.

[9] N. N. Bazarova, Y. H. Choi, Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites, Journal of Communication 64 (2014) 635–657.

[10] K. Greene, V. J. Derlega, A. Mathews, Self-Disclosure in Personal Relationships, Cambridge Handbooks in Psychology, Cambridge University Press, 2006, p. 409–428.

[11] E. Aïmeur, N. Díaz Ferreyra, H. Hage, Manipulation and malicious personalization: Exploring the self-disclosure biases exploited by deceptive attackers on social media, Frontiers in Artificial Intelligence 2 (2019) 26.

[12] V. J. Derlega, J. Grzelak, Appropriateness of self-disclosure, in: G. J. Chelun (Ed.), Self-disclosure: Origins, Patterns, and Implications of Openness in Interpersonal Relationships, Jossey-Bass, 1979, pp. 151–176.

[13] S. M. Jourard, P. Lasakow, Some factors in self-disclosure, The Journal of Abnormal and Social Psychology 56 (1958) 91–8.

[14] S. Trepte, P. Masur, Need for privacy, in: V. Zeigler-Hill, T. Shakelford (Eds.), Encyclopedia of personality and individual differences, Springer, 2020.

[15] A. F. Westin, Special report: Legal safeguards to insure privacy in a computer society, Communications of ACM 10 (1967) 533–537.

[16] M. H. Millham, D. Atkin, Managing the virtual boundaries: Online social networks, disclosure, and privacy behaviors, New Media & Society 20 (2018) 50–67.

[17] S. Petronio, W. T. Durham, Communication privacy management theory: significance for interpersonal communication, SAGE Publications, Inc., 2008, pp. 309–322.

[18] T. Dienlin, M. J. Metzger, An extended privacy calculus model for SNSs: Analyzing self-disclosure and self-withdrawal in a representative u.s. sample, Journal of Computer-Mediated Communication 21 (2016) 368–383.
[19] E. L. Spottswood, J. T. Hancock, Should I share that? prompting social norms that influence privacy behaviors on a social networking site, Journal of Computer-Mediated Communication 22 (2017) 55–70.
[20] L. Baruh, E. Secinti, Z. Cemalcilar, Online privacy concerns and privacy management: A meta-analytical review, Journal of Communication 67 (2017) 26–53.
[21] M. Büchi, N. Just, M. Latzer, Caring is not enough: the importance of internet skills for online privacy protection, Information, Communication & Society 20 (2017) 1261–1278.
[22] E. G. Smit, G. Van Noort, H. A. Voorveld, Understanding online behavioural advertising: User knowledge, privacy concerns and online coping behaviour in europe, Computers in Human Behavior 32 (2014) 15–22.
[23] A. Bandura, Social foundations of thought and action: a social cognitive theory, Prentice-Hall series in social learning theory, Prentice-Hall, 1986.
[24] H. Chen, W. Chen, Couldn't or wouldn't? the influence of privacy concerns and self-efficacy in privacy management on privacy protection, Cyberpsychology, behavior and social networking 18 1 (2015) 13–9.
[25] H. Almuhimedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. F. Cranor, Y. Agarwal, Your location has been shared 5,398 times! a field study on mobile app privacy nudging, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, ACM, New York, NY, USA, 2015, p. 787–796.
[26] L. Warberg, A. Acquisti, D. Sicker, Can privacy nudges be tailored to individuals' decision making and personality traits?, in: Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society, WPES'19, ACM, New York, NY, USA, 2019, p. 175–197.
[27] E. Peer, S. Egelman, M. Harbach, N. Malkin, A. Mathur, A. Frik, Nudge me right: Personalizing online security nudges to people's decision-making styles, Computers in Human Behavior 109 (2020) 106347.
[28] H. Masaki, K. Shibata, S. Hoshino, T. Ishihama, N. Saito, K. Yatani, Exploring nudge designs to help adolescent sns users avoid privacy and safety threats, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, ACM, New York, NY, USA, 2020, p. 1–11.
[29] H. Wang, S. Ma, H.-N. Dai, M. Imran, T. Wang, Blockchain-based data privacy management with nudge theory in open banking, Future Generation Computer Systems 110 (2020) 812–823.
[30] B. Liu, M. S. Andersen, F. Schaub, H. Almuhimedi, S. Zhang, N. Sadeh, Y. Agarwal, A. Acquisti, Follow my recommendations: A personalized privacy assistant for mobile app permissions, in: SOUPS, 2016.
[31] C. B. Jackson, Y. Wang, Addressing the privacy paradox through personalized privacy notifications, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2 (2018).
[32] [Online] The Dark Web & Your Data: Facts to Know, 2020, Accessed: 2020-07-28. URL: https://www.transunion.com/blog/identity-protection/the-dark-web-your-data-facts-to-know.
[33] [Online] The Dark Web explained - what does it mean for online security?, 2020,

Accessed: 2020-07-28. URL: https://www.equifax.co.uk/resources/identity_protection/dark-web-explained.html.

[34] [Online] Your SSN costs less than a Starbucks coffee on the dark web, 2020, Accessed: 2020-07-28. URL: https://atlasvpn.com/blog/your-ssn-costs-less-than-a-starbucks-coffee-on-the-dark-web.

[35] [Online] Dark Web: The Average Cost of Buying a New Identity in 2020, 2020, Accessed: 2020-07-28. URL: https://www.safetydetectives.com/blog/dark-web-the-average-cost-of-buying-a-new-identity/.

[36] [Online] How Cybercriminals Make Money, 2020, Accessed: 2020-07-28. URL: https://www.keepersecurity.com/en_GB/how-much-is-my-information-worth-to-hacker-dark-web.html.

[37] [Online] Internet Security Threat Report, 2020, Accessed: 2020-07-28. URL: https://docs.broadcom.com/doc/istr-24-2019-en.

[38] [Online] Dark web market price for stolen credentials 2019, 2020, Accessed: 2020-07-28. URL: https://www.statista.com/statistics/1007470/stolen-credentials-dark-web-market-price/.

[39] [Online] How Much Is Your Identity Worth on the Black Market?, 2020, Accessed: 2020-07-28. URL: https://www.pacetechnical.com/much-identity-worth-black-market/.