The Privacy versus Disclosure Appetite Dilemma: Mitigation by Recommendation

Rim Ben Salem[1], Esma Aïmeur[1] and Hicham Hage[2]

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

Abstract

Social Networking Sites (SNS) are growing exponentially and have undoubtedly become an intrinsic part of our lives. This is accompanied by a spike in the amount of time spent by users online, namely teenagers and young adults who allocate an average of three hours a day for various types of SNS. From commenting, posting sharing opinions to selfies and videos, they expose numerous pieces of personal information, jeopardizing their privacy. Specifically, this self-disclosure is driven by various factors including the individual's sharing needs, the information being shared as well as the target audience. As such, it is a multi-layered issue to which the one-size-fits-all interventions, currently being used, have not proven to be most effective.

This paper proposes a novel harm-aware recommender system-based solution to help the user take privacy-preserving decisions while using social media. This novel take on self-disclosure mitigation leverages notions from behavioural economics as well as psychological measurements. One of the contributions of this paper is the conception of the disclosure appetite, which is a user-specific term that encompasses their perception and their drive to reveal their private information. The tailored privacy-aware recommendations rely on the psychometric value that is the disclosure appetite as well as the sensitivity of the data being revealed. Through a trade-off between the two parameters, the system aims to mitigate the privacy compromise while considering the preferences of the user. In the era of oversharing and living virtually, this system that handles private data, with the intention of preserving it, is more crucial than the common uses of recommenders. In most of those cases, the extent of personal information exchanged does not go beyond preferences. Whereas the consequences and preventive potential of this work can have a bigger impact on multiple facets of the users' lives. Finally, an empirical evaluation is conducted using participants from the US, Canada and Europe.

Keywords

Personalized recommender system, disclosure mitigation, harm-aware, behavioural economics, psychological measurements, disclosure appetite

1. Introduction

Recommender systems play a big role in most of the content consumed by users online. From streaming platforms like *Netflix* and shopping services like *Amazon* to personalized assistants like *Siri* and *Cortana*, a multitude of applications have been integrated into our daily lives. Initially, the major focus of these recommenders was to get to know the user and deliver better-tailored suggestions in numerous contexts. The better the results are, the more engagement the system gets from the user and the more lucrative deploying recommender systems becomes for

OHARS'21: Second Workshop on Online Misinformation- and Harm-Aware Recommender Systems, October 2, 2021, Amsterdam, Netherlands

companies. In the 1990s and early 2000s, there was a surge of filtering techniques applied to real-life situations to enhance the weave of information tapestry [1] and improve personalization [2]. It was only recently that the interest widened beyond how well the system knows the user to how well does it protect them and prevent infringements upon their privacy. In fact, pioneering research on *privacy-preserving* recommender systems proposes a system called *ALAMBIC* [3] for e-commerce use and was published in 2008. No longer was privacy an afterthought or a collateral cost for better recommendations but it became a requirement to be fulfilled by these systems. Jeckmans et al. [4] further discuss the privacy concerns in question such as data collection without the users' awareness, long time or even permanent retention and profiting off of personal information. An example of this is the case of the dataset published by Netflix after *anonymization* that was still subject to re-identification and exposure of the original user information [5]. Facing all these threats, the big tech companies delved more into measures of protecting their databases and repositories against breaches. Removing identifiable information while keeping the essence of the preferences used for content personalization is one of the popular techniques known as anonymization. The urgent call for stricter and more user-conscious recommender systems [6] has been answered with numerous approaches including *encryption*, *differential privacy* [7], and *randomization* that consists in adding perturbation to the data. More recently, the focus started shifting to the human side such as the use of *social graphs*, which have proven their potential for privacy preservation [8]. There have also been increasing calls to address *fairness*, *transparency*, and how they impact the user's decisions [9]. This allows users to understand how and why they are being recommended a specific product and make the most informed choice. However, one main issue remains: regardless of how secure the medium and platform are, if the users themselves end up compromising their own privacy, none of the former measures can remedy that. While companies focus on not implicating themselves in legal issues by securing their servers and warehouses, the unaware user is still just as vulnerable and susceptible to becoming a victim. In today's world, the threats to privacy are growing exponentially and people are often caught unprepared to face them. This brings it to the next point, which is what needs to be done once people are made aware of these issues. What to do about the possible scams and fraud in a world where almost every aspect of our life is digitized like shopping, ordering food, watching movies, and due to the COVID19 outbreak, even studying and working? It is not feasible to ask people to stop relying on the services they have grown accustomed to and draw gains from. Hence, a middle ground between not sharing at all and oversharing can help users online. Let us take the example of a user, Alice, who is concerned about her privacy and does not wish to share personal aspects of her life with distant acquaintances or colleagues. However, after having a bad day, Alice, like a lot of social media users, starts writing a post detailing her struggles and complaints and is about to share it on an account visible to her co-workers. Had she been more emotionally stable, she would not share this type of information with them. It would be very convenient and helpful for Alice if she got a prompt before the post is published to guide her and remind her of her own preferences. If that were to happen, she would avoid making a mistake that can compromise her professional position at work. This paper proposes a recommender system whose purpose is to eliminate or at the very least alleviate the problems caused by self-disclosure such as Alice's situation. Her case is but one of many in which users can find themselves as there happen to be numerous pieces of personal data that can face jeopardy in many contexts and scenarios. It is up to the

recommender system to recommend, in a personalized manner, which pieces of information not to disclose, such that to balance the user's disclosure needs with their privacy preferences. This issue is a lot more crucial than the majority of uses of recommender systems where the user only shares preferences in cuisine for example and gets recommended restaurants as a response. Thus, the aim of this paper is to avoid these dire consequences, motivate users and promote better practices on social media. To do that, the proposed harm-aware recommender system mediates between the privacy concerns and the disclosure preferences of internet users specifically on Social Networking Sites (SNS) and contributes the following:

- A model to estimate the user-specific drive for disclosure that we call *disclosure appetite*. It is a term that we introduce as the personalized level of disclosure that is acceptable or necessary to the user. It is inspired by the "risk appetite", which is defined for organizations.
- An objective representation of personal data in order of sensitivity, which is an updated version of a well-established social penetration theory;
- A novel user-centric harm-aware recommender system to mitigate self-disclosure through personal and objective considerations using a *Rasch model* approach. Based on a compromise between the two, the system recommends that the user deletes or keeps each piece of data in the original input.

The paper is organized as follows: Section 2 discusses existing work that relates to our research. Section 3 details the recommender system, Section 4 reports on the evaluation of the platform and in Section 5, a discussion ensues from the findings. Finally, we conclude this paper and shed light on future directions.

2. Related work

Solove [10] define privacy as a need to control who has access to personal information and a form of solitude, intimacy, anonymity, or reserve. Disclosing private information can be seen as the outcome of a risk-benefit analysis known as *privacy calculus* [11]. However, people are often susceptible to the *privacy paradox* [12]. The latter is defined as the discrepancy between the users' reported concerns about privacy and the actual acts they commit that compromise it. This inconsistency, which already existed in offline contexts, became even more prevalent with the growth of the virtual universe in general and social media platforms in specific. As a result, discussions arose on whether virtual citizens even care about privacy as much as they claim to [13] or are their actions more reflective of their perception of the matter. In addition to trying to determine the roots of self-disclosure, existing research has been tackling the issue of oversharing personal information online and how to mitigate it. Notably, several studies highlighted the need for reasoning assistants to guide users and limit their personal data disclosure.

Although there are user-centric approaches on the system side such as developing privacy-preserving protocols [14], focusing on the user side is just as crucial. For instance, the privacy assistant *PACMAN* [15] exists as a social media assistant and a reminder to users to carefully choose the audience of their posts. The proposed solution takes as input the user's information

to utilize the existing yet often ignored or forgotten options to limit self-disclosure such as limiting the audience of certain posts. In fact, Facebook users for example have always had the option to share either with the public or with friends but rarely do people assess to whom every post of theirs should be visible. That is the case especially for users who are very active and remain engaged on a plethora of pages and groups.

Another way to remedy the issue is the use of *nudges* to discourage users from harmful actions [16], an approach that has become a popular practice. Nudges are either positive behaviour reinforcements or ill-advised activity deterrence called preventive nudges and in both cases, the aim is the wellbeing of the person. Prompts, cues, notifications are all forms of nudges, which can rely on general knowledge or user-specific parameters. Their mechanisms can be clustered into four major categories: decision information, decision structure, decision assistance, and social decision appeal [17]. Existing applications of these concepts to disclosure mitigation include the platform [18] that relies on a user's disclosure preferences, and risk aversion to deliver nudges and to auto-adapt following the user's reaction. User-centric solutions [19] have been garnering attention as the belief grows that privacy protection is not a one size fits all concept [20]. In more critical contexts such as mobile payment, privacy-preserving recommendations play a very important role in reinforcing the opinion of experts on secure online purchases [21]. There is a gap between the ubiquity of card fraud cases and the buyer negligence to adopt the recommended measures, and the research concluded that implementing nudges can help users make better and more informed decisions. Regarding privacy-concerned recommender systems that aim to alleviate incidents on social media [22], the example of *YourPrivacyProtector* [23] fits this category. Based on simple machine learning techniques, the proposed solution shows great promise in assisting Facebook users by suggesting a different privacy setting than the default one. Another example of privacy settings recommendations is *SPAC* [24], which learns the user's behavioural patterns and uses their history and profiles in order to predict the best configuration.

Most of these assistants and recommenders consider the privacy of a post being shared as a whole rather than put a value on each personal piece of information. Other approaches [18] put metaphorical price tags on different bits of data in order of sensitivity. This paper falls within the latter scope and instead of opting for an all or nothing method, aims to mitigate harms in order of sensitivity while also considering personal preferences. For this reason, data is regarded as a group of items appraised differently by the system and recommendations are based on the Rasch model [25] that incorporates singular objective values and a person measure. This duality of recommender systems and the Rasch model has been explored in other contexts before. Existing work includes tailoring goals for nutrition assistance systems [26], in which the proposed app offers dietary tracking, visual feedback, and personalized recipe recommendations and showed higher success in comparison with the non-user-specific alternative. Similarly to this research and in the domain of energy conservation and health, the Rasch model, once again is used and proves its fit for modelling user behaviour [27]. Knijnenburg [28] proposes a user-tailored approach that considers the benefits and privacy loss for commercial use. The paper demonstrates and justifies the failure of the existing generic nudges, persuasive cues, and all one-size-fits-all approaches as effective privacy-preserving solutions. Our work falls within the same scope of personalizing harm-aware recommendations with the aim of increasing their acceptance rates amongst users. However, to the best of our knowledge, it presents

the first Rasch model-based recommender system for disclosure mitigation on social media. Moreover, it has consideration for different contexts, which are the social circles involved in the disclosure. Thanks to our proposed approach, privacy preservation can progress beyond the restrictive all or nothing solutions, but more importantly, the emphasis is put on the need for a user-centric approach. The need and desire for disclosure is a highly subjective matter, as a result, the mitigation is equally tailored by the individual's preferences. The next section sets the foundations for user modelling and provides the disclosure appetite measure that is used along with the objective data parameters to limit the disclosure and its infringement upon privacy; an approach that shows potential in addressing self-disclosure.

3. Recommender system-based approach for disclosure mitigation

This section details the harm-aware recommender system. Specifically, when potential data disclosures are detected, the private data shared are reported along with the context. The recommender system receives this information then, retrieves the user-specific disclosure appetite and the sensitivity of the data in question. Upon receiving the disclosure mitigating recommendation, the user acts by either accepting or rejecting the prompt. The proposed system is *agnostic* to the mechanism and technology behind disclosure detection. Various techniques could be used, ranging from a simple search using keywords to more sophisticated and cutting-edge solutions such as Google Cloud Natural Language. However, disclosure detection is not in the scope of the current paper and is not detailed further. Specifically, what is of interest is the personalized privacy-preserving and disclosure-mitigating recommendations that come *after a disclosure is detected.*

3.1. Personalized harm-aware recommender system

Disclosure mitigating recommendations are based on the trade-off between the user's preferences, motivations and data sensitivity. This can also be interpreted using the privacy calculus theory that compares risks and benefits. To achieve this 2-parameter based system, we choose to work with *Item Response Theory* (IRT) as this paradigm, by design, is user-centric and fits our design. The model takes as input "*user ability*" and "*item difficulty*" and the output is the probability of them answering the question correctly. The proposed system is built on the foundations of the *Rasch model*, a special case of IRT based on personal and objective parameters.

The original Rasch model:

$X_{ni} = x \in \{0, 1\}$ is the dichotomous random variable where, $x = 1$ denotes the correct answer, $x = 0$ an incorrect response to a given assessment item. In the Rasch model for dichotomous data, the probability of outcome $X_{ni} = 1$, which is a correct answer, is given by:

$$P\{X_{ni} = 1\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \tag{1}$$

Where: β_n is the ability of person n and δ_i is the difficulty of item i.

The adapted Rasch model:

Before presenting the mathematical aspect of the model, it is worth noting that one of the main strengths of the proposed system is the fact that recommendations pertain to *specific pieces* of data rather than the content as a whole. It is based on mitigation instead of total elimination. In the context of the proposed recommender system, $X_{ni} = x \in \{0, 1\}$ is the dichotomous random variable where, $x = 1$ denotes an accepted recommendation by user n concerning item i (user accepts to delete the item/piece of data). P is the probability of outcome $X_{ni} = 0$ as if the user would disclose the information (reject the recommendation to delete piece of data i):

$$P\{X_{ni} = 0\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \tag{2'}$$

In terms of accepting the recommendation:

$$P\{X_{ni} = 1\} = \frac{e^{\delta_i - \beta_n}}{1 + e^{\delta_i - \beta_n}} \tag{2"}$$

Where: β_n: user-specific disclosure appetite, δ_i: sensitivity/value of a piece of data.

β_n is the n^{th} user's drive, motivation and preference to disclose. It is a measure inspired by its economics counterpart: the *risk appetite*, the level of risk that an individual or organization is willing to accept while pursuing its objectives. In the proposed system, three contexts are considered and they represent the *social circle*, with whom the users are sharing their private data. These social circles are grouped into the following three categories: close friends & family, colleagues and acquaintances and the general public. As such, for each user, a separate β_n value, one for each of social circle is maintained, reflecting the 3 variants of the disclosure appetite depending on the context. The rationale behind this approach resides in the fact that different contexts affect the sharing motivation and preferences [29], a person might have a very high drive for disclosure with family members but would be reluctant to do so with the public. A different person might be careless with their own data and have a similar disclosure attitude across all contexts. δ_i differs from β_n in the sense that the former is the same for the entire population. This concludes the general presentation of the user-specific system.

3.2. Configuring the disclosure appetite and data sensitivity

The proposed approach relies on the Rasch model's representation of the trade-off between the data sensitivity and the disclosure appetite. Figure 1 represents the steps taken by the recommender system.

The following subsections detail each step of the process, starting with the estimation of the disclosure appetite and data sensitivity to tuning/fitting the model and finally performing the recommendations.

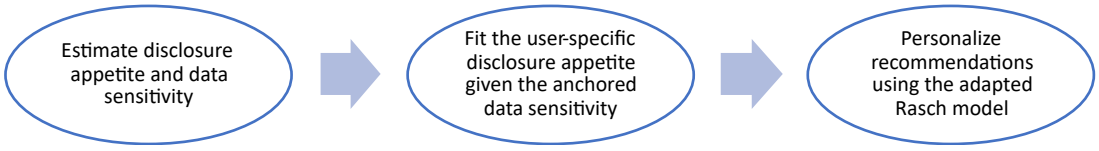


Figure 1: The proposed process to get the personalized harm-aware recommendation

3.2.1. Estimate disclosure appetite and data sensitivity

At this stage, the system does not yet have preestablished values of data sensitivity nor disclosure appetite. So, the first step is to estimate these values. There are numerous Rasch estimation methods. As such, a comparison of the following four methods was performed: *Joint Maximum Likelihood Estimation (JMLE)*, *Marginal Maximum Likelihood (MMLE) Estimation*, *Conditional Maximum Likelihood Estimation (CMLE)*, *Pairwise Maximum Likelihood Estimation (PMLE)* [30]. Throughout this estimation process as well as the calibration and anchoring of the data sensitivity the data used is provided by participants whom we detail in the evaluation (Section 4). A "user", in this paper, refers to an individual amongst the participants in the survey.

Optimal estimation method: The estimation methods are judged based on 2 criteria: first, how well does the model fit the current data and second, how well would it fit future data that is yet to be seen. The first criterion can be measured through the goodness of fit as well as the *Root Mean Square Error (RMSE)*. The second can be observed though the reliability value that indicates reproducibility. The reliability coefficient reflects the correlation between the true measure and the observed measure. Throughout this paper, *Winsteps* version 4.8.1 [31] is used for the estimations and evaluations and it relies on the *Cronbach's Alpha* to calculate this metric for the disclosure appetite reliability and all the results are reported in Table 1. The *global fit* measure relies on *Pearson chi-square* and the output: the *p-value* is a number between 0 and 1 with higher values indicating better fit. JMLE returns the highest value of 0.82 while CMLE has the lowest 0.61.

Table 1
Comparison of estimation methods

Estimation method	Global fit (p-value)	Data sensitivity = columns RMSE	Reliability	Disclosure appetite= rows RMSE	Reliability
JMLE	**0.82**	**0.023**	**0.79**	**0.017**	**0.74**
MMLE	0.78	0.024	0.74	0.038	0.71
CMLE	0.61	0.034	0.68	0.019	0.63
PMLE	0.67	0.030	0.63	0.031	0.59

The reliability of data sensitivity, in theory, should be at least > 0.70 to be considered acceptable and for the disclosure appetite, the reliability is considered good if it's > 0.70 and very good if it's > 0.80. As for RMSE, a measure of 0.05 or smaller is good fit, between 0.06 and 0.08 is reasonable and > 0.1 = poor fit. Taking all these metrics into account, JMLE outperformed MMLE, CMLE and PMLE and reported good results. Thus, the data sensitivity values that are

fixed henceforth in upcoming sections are the ones estimated using JMLE.

Modeling the data sensitivity: It is important to keep in mind that the recommender system relies on the Rasch model to establish a trade-off between "disclosure appetite" (user's gain from disclosing information) and "data sensitivity" (user's loss of privacy). In this section, the focus is on the latter parameter. Based on the choice of the estimation method JMLE, Table 2 reports on examples of these results per personal data type.

Table 2
Data sensitivity measure per examples of pieces of data

Piece of data	Sensitivity
Banking information	0.98
Religious opinion	0.56
Biographical data	0.17

Overall, the model was tested on 13 pieces of data; they are, in order of highest to lowest sensitivity: *Banking information*, *ID-Passport*, *Medical records*, *Home address*, *License plate*, *Photos of diploma*, *Religious belief*, *Political belief*, *Travel plans*, *Selfies*, *Charity contributions*, *Preferences* (products and services), and *Biographical data*. These values are calculated based on answers given by users to disclosure scenarios (see examples in the Evaluation section).

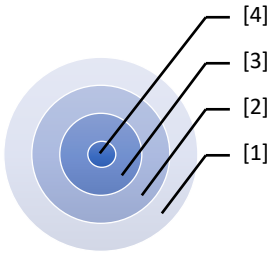
Now that these values are obtained, it is important to proceed to the categorization or dichotomization process, which entails grouping several pieces of data together rather than viewing each one uniquely. This is crucial to prepare for generalization. For example, in the present paper, 2 types of official documents are considered: *ID and passport* so knowing how sensitive they are can give an insight into the value of a driver license. That can be achieved thanks to the generalization of the model performed in this step. While this can be achieved using simple approaches (such as equal bins), it will not really tend to the specificities of the values of data reported in Table 2. How the private information is attributed numeric levels of sensitivity and categorized is crucial as this defines the "item value" in the Rasch model and is in direct correlation with the probability of the user accepting/refusing the recommendation. Table 3 highlights the layer bounds using *Andrich thresholds*, which are the boundaries separating the categories. The representative element is the sensitivity value of each layer. Based on these results, this paper proposes a four-layer model for private data in order of sensitivity.

Table 3
Categorization or binning based on Andrich thresholds

Layer	Upper bound	Representative element
1	0.258	0.176
2	0.543	0.457
3	0.747	0.643
4	1.000	0.882

The proposed approach, in this work, to update the representation of personal data in a layered manner based on sensitivity is inspired by the *Social Penetration Theory* (SPT) [32]. Specifically, SPT explains the trade-off people make while they consider revealing their private

information to gain a sense of intimacy with others. SPT proposes a six-layer model to order data based on lowest to highest sensitivity: *Biographical data*, *preference in clothes*, *food and music*, *goals and aspirations*, *religious convictions*, *deeply held fears and fantasies*, and *concept of self*. Figure 2 illustrates the proposed 4 layers in correspondence with the thresholds from Table 3 and the different measures of data sensitivity such as the ones reported in Table 2.



1. Biographical data
2. Preferences, Selfie, charity contribution
3. Travel plans, religious and political opinions, license plate, photos of diploma
4. Banking information, official documents (ID, passport), Medical records, home address

Figure 2: Social penetration-inspired classification of personal data.

At this point, the Rasch model-based probability can be computed using equation 2" as the system has data sensitivity values as well as disclosure appetite values using JMLE. The purpose of fitting the model (or the calibration process) is to obtain a better, more personalized estimation of the disclosure appetite.

3.2.2. Fit the user-specific disclosure appetite

This is another estimation process similar to the earlier conducted JMLE. However, in this section, the data sensitivity is anchored as detailed in subsection 3.2.1. Its values are considered to be anchored and the fitting module takes them as input along with the estimated disclosure appetite by JMLE. The *Anchored Maximum Likelihood Estimation* (AMLE), a method based on the concept of fixing "item difficulties" and estimating "person values", is used. It is worth reiterating that "item difficulty" corresponds to "data sensitivity" in the proposed disclosure mitigating recommender system. Table 4 shows scenarios that users respond to on a 6-point Likert scale: "*Definitely no*" - "*No*" - "*Somewhat no*" - "*Somewhat yes*" - "*Yes*" - "*Definitely yes*".

Table 4
Scenarios used as samples for AMLE

Piece of data	Context	Scenario
Political belief	Close friends & family	During the presidential election, would you add a banner on your profile picture that displays to your friends the candidate you are supporting?
Travel plans	Colleagues & acquaintances	You are about to go on your most awaited trip to your favourite destination. You have booked a 5-star hotel with great accommodations and you have already planned every detail of the vacation. Would you share this with your colleagues?

Going back to the example of Alice, the user used as an example in the introduction, prior to the model fitting process, her disclosure appetite was calculated using a basic preference elicitation, and her measured disclosure appetite was 0.3. When given realistic scenarios for the system to get to know her better, she responded that with "*definitely yes*" to revealing her travel

plans with her colleagues and acquaintances. That data, as specified in the previous section belongs to *layer 3* of the proposed personal data classification in order of sensitivity. Given this information and the context "*with colleagues and acquaintances*", her disclosure appetite increases from 0.3 to 0.34. Another user, Bob, who when presented the same hypothetical scenario but answers "*Strongly no*", sees a decrease in his disclosure appetite from 0.5 to 0.46. For both Alice and Bob, the values adapted here are the disclosure appetites for the specific context "colleagues and acquaintances" and none of the other 2 variables ("public" and "close friends and family") is modified. As seen in Figure 1, at this point, the remaining step is the personalization of the recommendations.

3.2.3. Personalized recommendations

The Rasch model introduced and explained at the beginning of section 3.1 can now be used to perform recommendations. Suppose that the system is acting as an assistant to Alice and Bob whose disclosure appetites are $\beta_A = 0.34$ and $\beta_B = 0.46$ respectively. These values were updated (as highlighted in section 3.2.2) following the fitting process. Both of them face the same exact scenario such as "*sharing a selfie in which they are acting funny or maybe they are drinking*" and they are about to share it with colleagues and acquaintances. This piece of information belongs to layer 2 of the data sensitivity model and is given the value $\delta_i = 0.457$ in Table 3. Alice whose disclosure appetite in this context is 0.34, has a probability of $\frac{e^{0.457-0.340}}{1+e^{0.457-0.340}} = 0.529$ of accepting the recommendation advising her not to share the photo. Bob, on the other hand, has a job with less formality and his colleagues are comfortable sharing selfies while partying. This explains his relatively higher disclosure appetite that is 0.46 and as the system got to know his preferences in the steps prior to the recommendation, the probability that he would accept to not share the data in question is 0.49 (the probability of rejecting the recommendation is $1 - 0.49 = 0.51$). As a result, the system does not intervene by nudging him to reduce the disclosure. Following the recommendation, for each user who was nudged, the disclosure appetite is adapted according to their decision so if Alice accepts the recommendation she got, the value goes down but if she refuses if increases. The next step is to assess how well does the disclosure mitigating system perform beyond the theory and using real participants.

4. Evaluation

The proposed recommender system estimates the disclosure appetites of users and the personal data sensitivity, then adapts the person's value, and provides more personalized, privacy-preserving recommendations. This section provides insights on the different points to evaluate and corroborate:

- The calibration and validation of the disclosure appetite values (to determine how accurate the user model is)
- The context-specific disclosure appetite consideration (to corroborate the need to consider disclosure context)
- The evaluation of the overall personalized recommender system as an effective harm-aware system.

4.1. Survey administration

There is no preexisting available dataset that can be used for the evaluation. Hence, *800 people* were recruited over a period of 4 days using *Amazon's Mechanical Turk* [33]. The demographics are reported in Table 5. The participants are equally divided from two geographical locations: *Europe* and *North America* (*US* and *Canada*). The survey has a total of *40 questions* and was hosted on Limesurvey [34].

Table 5
Demographic characteristics of the study sample

Demographic variable	Category	Frequency %
Gender	Female	58%
	Male	41%
	I prefer not to answer	< 1%
Age	18-24	25%
	25-34	25%
	35-44	25%
	45+	25%
Occupation	Student	49%
	Employee	19%
	Business owner /Self-employed/professional	12%
	Manager/Official	9%
	Researcher	< 1%
	Retired/Unemployed	9%
	I prefer not to answer	<1%

Before completing the survey, respondents are presented with a consent form providing the purpose of the research, affiliation of the researchers and information on the anonymity of the responses and the right of withdrawal. The participants are not required to submit any uniquely identifiable information (real name, email, phone number, etc.) and they remain anonymous throughout the entire process.

The survey is composed of three linked sections shown in Figure 3. The evaluation is based on a *10-fold cross-validation* meaning that 9 folds are used for estimation and calibration and 1 fold for the tests.

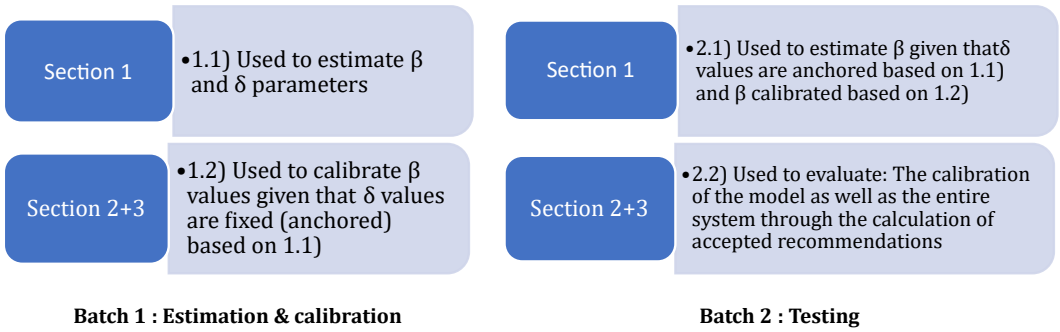


Figure 3: A batch breakdown of the evaluation dataset

Table 6 shows an example of a question flow that the participant answers in the survey. Section 1, 2 and 3 follow the reasoning "what does the user prefer to share?", then: "What would they actually share given specific situation?" and finally: "Do they accept the personalized disclosure mitigating recommendations?".

Table 6
Example of a scenario flow throughout the survey sections

Section 1 question	Section 2 question	Section 3 question
I have shared my **medical records** publicly before.	You are healing from a severe illness. You developed specific habits like your diet or your bedtime that could help other patients who suffer from the same disease. Would you be prepared to share this with **everyone** as well as your **medical record** for the benefit of the **public**?	You have answered yes, but **previously you indicated that your medical records are sensitive and you would not share them with the public**. Based on your own pre-set preferences, the system advises you not to share this. What would you do?

Now that the premise of the evaluation is explained, next is a report on the results.

4.2. Results

This section is dedicated to observations made and results recorded through the evaluation process of the system. Figure 4 shows that the probability to accept the recommendation depends on:

- The user-specific disclosure appetite: the higher the disclosure appetite the less likely the user is to accept.
- The item sensitivity: the higher the sensitivity, the more likely the participant is to accept the recommendation.

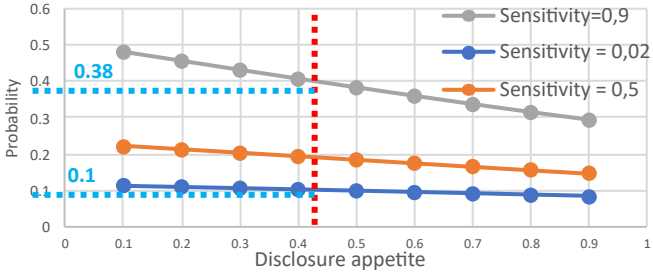


Figure 4: Probability of accepting a disclosure mitigating recommendation as a function of disclosure appetite

For a person whose disclosure appetite is 0.5, the probability of them accepting the recommendation and proceeding to conceal a piece of information with the sensitivity 0.02 is 0.1. For the same person, the probability is 0.38 if the piece of data in question is highly sensitive (0.9 sensitivity). So the probability increases with the increase of sensitivity. Comparing a person with a disclosure appetite of 0.8 and someone with a 0.2 measure, Figure 4 shows that the higher

the appetite, the less likely it is that the recommendation would be accepted. Previously, it was detailed that each section serves a different purpose in the evaluation process. Section 2 of batch 2 (testing batch), which presents the hypothetical scenarios as in Table 6, is used to better fit the model and tailor it. To assess the impact of this process, an evaluation of how well the recommendations are received by the user (accepting is the desired outcome) with or without calibration is used and reported in Table 7.

Table 7
Valuation of the disclosure appetite calibration

Disclosure layer	errors prior to calibration	errors post-calibration
Layer 1	0.04	0.03
Layer 2	0.11	0.05
Layer 3	0.07	0.02
Layer 4	0.05	0.01

The probability of the user accepting to delete a specific piece of data is calculated prior to and post-calibration and is then compared to the actually reported action of the user. An error occurs when the two do not match, namely if the probability is $0.6 > 0.5$ but the user does not accept the recommendation, this is an estimation error that the system reports and Table 7 records these values across the different layers of personal data specified in Figure 2.

One of the foundations of the proposed system is the consideration of the context, which is the audience with whom the data is shared. Table 4, shows examples of different scenarios with the corresponding context. Each user has three different disclosure appetites depending on whether the private information is being shared with "close friends and family", "colleagues and acquaintances" or the "public". The results reported in Figure 5 corroborate this approach. Figure 5 (a) highlights the results of using one single β value for each user across all contexts, which means that if the user decides to share data with their family members, their overall disclosure appetite increases as a result, the system assumes that the user's drive to share with the public has increased as well. This has led to a lack of understanding of the user's attitude, which is backed up by the results namely the percentage of accepted recommendations regarding disclosure with close friends and family in particular. Associating a value with each context in Figure 5 (b) has led to an increase from 3% to 59% for the age group 18-24, from 27% to 47% for people aged 25-34 and a 20% difference for the category 45+.

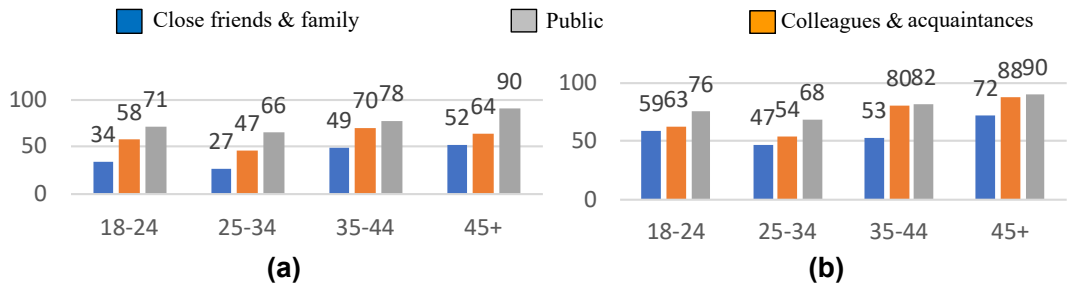


Figure 5: Percentage of accepted recommendations per age group depending on context considerations, graph (a) shows results using a single value for disclosure appetite across all contexts and (b) corresponds to context-specific values.

This concludes the validation of the first two evaluations and the remaining one is to assess the overall performance. In other words: how well are these user-centric recommendations received by the users. To do so, it is also interesting to look at how the different demographic pieces of information play part in the user's decision. The results highlighted in Figure 6 show that the percentage of accepted recommendations does not differ a lot based on whether the participant is from North America or Europe. Before going into more details, it is important to note that overall, *891 recommendations* were pushed across all 800 participants. Specifically, some users displayed an attitude more inclined to disclose and were given 8 recommendations across the entire survey, for example, while other participants were less inclined, and got fewer recommendations as per the proposed Rasch model. As pointed out throughout the paper, the recommendations are tailored and not given to all users in the same situations. Moving on to the results: for the most sensitive data (belonging to layer 4), *461 recommendations* were given, amongst which, 177 for European participants and 284 for North Americans. For the former group, 93% showed a willingness to accept the disclosure mitigating recommendations while the percentage is 86% for North Americans (Figure 6 (c)). There was not a large gap based on gender either. Comparing students to non-students and based on the age category yielded more noticeable differences such as a 52% difference between participants aged 45+ and 18-24 when it comes to accepting recommendations concerning layer 2. Similarly, non-students showed a 61% willingness to accept recommendations centred around data with the same sensitivity (same layer) while students were only 14% agreeable to the prompt.

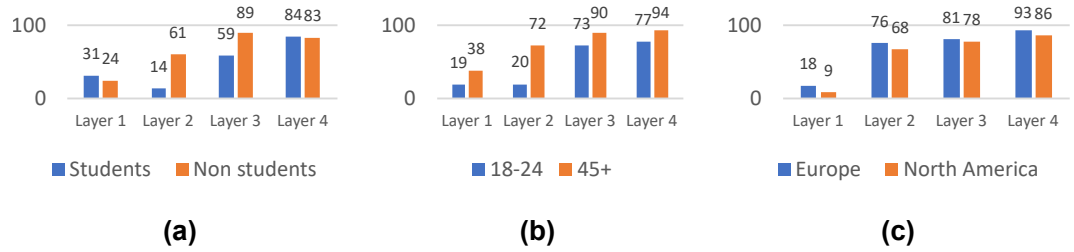


Figure 6: Percentage of accepted recommendations per demographic criterion

Specifically, recommendations are only pushed to users if they declared in section 1 (of the survey) that they would not disclose a piece of data but end up saying they would when given a realistic scenario in section 2.

4.3. Discussion

The evaluation has featured encouraging preliminary results. In particular, it demonstrated the utility and need for such a user-centric disclosure mitigating recommender system and its potential is corroborated by the response of the individuals to recommendations. When it comes to protecting data from the most sensitive layer (layer 4), the percentage of accepted recommendations ranged from 77% to 94% ((a) and (b) from Figure 6). These results translate to privacy preservation meaning that 94% of users were going to disclosure very crucial data but were deterred by the system, which fulfils the aim of this work. It is worth mentioning that we acknowledge the fact that in the testing setup, users are given hypothetical scenarios to respond

to and are not naturally using their social media accounts and getting real-time recommendations, on their actual personal data. While this can introduce a bias in the results, the scenarios that were given in the survey were crafted such that users could relate to them, effectively mitigating and reducing the effect of this setup on the users' responses. Nonetheless, validating this kind of system with real users and real data has its own intricacies and considerations and not just from a technical perspective. The findings of this paper further validate the privacy paradox. In fact, in the first section of the survey, individuals show a high consideration for their privacy expressing that they have never shared a piece of data only to respond in section 2 (Figure 3) saying that they would reveal that information when a specific situation arises.

Our approach to addressing the privacy paradox aims to reduce the discrepancy between the users' judgement (how they evaluate the data sensitivity) and their behaviour when put in a particular situation (proceed with the disclosure or accept the recommendation). Bridging this gap, as proposed in this paper, is done through nudging the users, such that their behaviour conforms to their judgement. This is achieved through providing personalized recommendations that are in fact aligned with the user's judgment. The preliminary results using a proof of concept of the personalized harm-aware recommender system show its potential, especially when dealing with the most sensitive data: 93% of Europeans and 86% of North Americans agreed with recommendations to reduce disclosure (Figure 6 (c)). When calculating the correlation between each demographic parameter and the decision to accept or refuse the recommendation, *Pearson correlation* is used. A high degree is reported for the criteria "age group" and "being a student vs non-student" (respectively 0.62 and 0.53 correlation), moderate for the gender (0.41) and low degree for the country (0.17). This offers insight on the criteria to focus on in the future, to further improve the recommendations. However, in its current state, the personalized harm-mitigating recommender system already is an improvement compared to a previous preliminary study that we conducted. In fact, the purpose of that one was to investigate the need for mitigation over elimination through evaluating the acceptance rates of nudges such as "delete only this part" against "delete the entire post". North American participants were presented with various scenarios to respond to, but the same recommendations were given to all of them (without any personalization). In the preliminary study, 68% of participants accepted the disclosure-reducing recommendations involving data that is classified as belonging to layers 2, 3 and 4. If we consider the same North American region and recommendations from the same layers, the present paper reports a 77% acceptance rate (averaging the results from Figure 6 (c)) thanks to the addition of personalization using the Rasch model.

Moving on to the context, the results in Figure 5 show that dividing the audience into three categories allows for a better tailored user-specific disclosure appetite value and as a result a better understanding of the disclosure behaviour and the probability of accepting recommendations. But another consideration is the fact that between the three social circles, users were less likely to accept to reduce oversharing with "close friends and family". An example of this is the recorded 53% in comparison with an 82% percentage of accepted recommendations for people aged between 35-44. This can be explained by the trust that users place in closer social circles. These results are still encouraging as 53% is an improvement showing that participants are willing to change their decisions when prompted in a way that considers their preferences. Had they not received the recommendation, they responded that they would have proceeded with the disclosure, which proves the efficacy of the system. This validates the

proposed user-centric approach but also highlights a future direction, which is to consider a more specific way to approach disclosure with more emotional value and trust attached to it. Figure 5 shows that the highest percentage of accepted recommendations is associated with the context "public" and the lowest with the context "close friends and family". This further corroborates existing research [35] on friendship and how it can be a catalyst for disclosure and compromising one's privacy. Finally, a major highlight of this paper is, not only its encouraging results during the evaluation but also the fact that the proposed user-centric recommender system can be generalized to a multitude of uses as it is agnostic to the type of data being shared. This is particularly important seeing as the field of application concerns privacy and sensitive data unlike most common uses of recommendations that are mainly concerned with preferences.

5. Conclusion and future work

Today's social media users are facing bigger threats than ever. The most concerning fact is that although the menaces are becoming more prevalent, users themselves, who are potential victims, are not becoming equally more aware. They need help to protect themselves as well as others around them such as the use of a security training platform [36]. This paper proposes a novel user-centric recommender system to help people achieve two conflicting goals: achieve the personal preferences for disclosure of information while protecting their privacy. The scope of this work extends beyond the generic preference-based recommendations to a more critical application. What is at stake is not simply whether the person's preference in food or clothes is leaked but the fact that their uniquely identifiable information can be compromised. The proposed system estimates the user-specific drive for sharing their data and defines it as the disclosure appetite inspired by its economic counterpart called the risk appetite. Furthermore, this work updates the well-established social penetration and extends it to social media disclosure. The privacy-preserving recommender system uses a trade-off between the disclosure motivation and the data sensitivity to guide users through the privacy versus disclosure appetite dilemma. Being agnostic to the specific data that is shared (text, video etc.) paves the way for reusability in other applications and forms of privacy preservation. From this point onwards, the research is headed towards testing with real users in real-life scenarios and evaluating their responses to the personalized recommendations while attempting to share something on SNS.

The reported results are encouraging as a measure of the system's effectiveness (accepted recommendations). However, we do intend to investigate the other metric: user satisfaction. Although our system considers implicit feedback in the form of accepting or rejecting the recommendation and adapts to this action, explicit feedback can further improve recommendations. An updated version of the 6-item scale by Knijnenburg et al. [37] be adapted to the social context and enhance the user experience and subsequently, the harm awareness goal. Moreover, this paper acknowledges that the issue is not simply a matter of revealing information about oneself but can extend to others. Sharing a photo that has another person in it is a form of disclosure of others that brings forth another aspect called *multiparty privacy* [38]. The existing techniques in this scope such as the auction, negotiation and aggregation-based mechanisms are far from

ideal. Thus, it could prove to be quite challenging to extend the user-centric recommender system proposed in this paper to an optimal multiparty-aware disclosure-mitigating model.

References

[1] D. Goldberg, D. Nichols, B. M. Oki, D. Terry, Using collaborative filtering to weave an information tapestry, Communications of the ACM 35 (1992) 61–70.

[2] S. S. Anand, B. Mobasher, Intelligent techniques for web personalization, in: IJCAI Workshop on Intelligent Techniques for Web Personalization, 2003, pp. 1–36.

[3] E. Aïmeur, G. Brassard, J. M. Fernandez, F. S. Mani Onana, ALAMBIC: A privacy-preserving recommender system for electronic commerce, International Journal of Information Security 7 (2008) 307–334.

[4] A. J. P. Jeckmans, M. Beye, Z. Erkin, P. Hartel, R. L. Lagendijk, Q. Tang, Privacy in recommender systems, in: Social Media Retrieval, 2013, pp. 263–281.

[5] A. Narayanan, V. Shmatikov, How to break anonymity of the Netflix prize dataset, 2006. arXiv:cs/0610105.

[6] R. Bourassa, Building recommender systems with strict privacy boundaries, in: Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18), ACM, Vancouver, Canada, 2018, p. 486.

[7] A. Berlioz, A. Friedman, M. A. Kaafar, R. Boreli, S. Berkovsky, Applying differential privacy to matrix factorization, in: Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15), ACM, Vienna, Austria, 2015, pp. 107–114.

[8] A. Wainakh, T. Grube, J. Daubert, M. Mühlhäuser, Efficient privacy-preserving recommendations based on social graphs, in: Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19), ACM, Copenhagen, Denmark, 2019, pp. 78–86.

[9] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, The unfairness of popularity bias in recommendation, in: Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys '19), Copenhagen, Denmark, 2019.

[10] D. J. Solove, Understanding privacy, Harvard University Press, 2008.

[11] R. S. Laufer, M. Wolfe, Privacy as a concept and a social issue: A multidimensional developmental theory, Journal of Social Issues 33 (1977) 22–42.

[12] S. Kokolakis, Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon, Computers & Security 64 (2017) 122–134.

[13] A. Acquisti, L. Brandimarte, G. Loewenstein, Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age, Journal of Consumer Psychology 30 (2020) 736–758.

[14] S. Badsha, X. Yi, I. Khalil, E. Bertino, Privacy preserving user-based recommender system, in: IEEE 37th International Conference on Distributed Computing Systems (ICDCS), Atlanta, USA, 2017, pp. 1074–1083.

[15] G. Misra, J. M. Such, PACMAN: Personal agent for access control in social media, IEEE Internet Computing 21 (2017) 18–26.

[16] A. Acquisti, L. Brandimarte, J. Hancock, A Sense of Privacy (2021). doi:10.1184/R1/14200199.v1.
[17] M. Jesse, D. Jannach, Digital nudging with recommender systems: Survey and future directions, Computers in Human Behavior Reports 3 (2021) 100052.
[18] R. Ben Salem, E. Aïmeur, H. Hage, A nudge-based recommender system towards responsible online socializing, in: Proceedings of the Workshop on Online Misinformation- and Harm-Aware Recommender Systems (OHARS 2020), Virtual Event, Brazil, 2020.
[19] E. Duriakova, E. Z. Tragos, B. Smyth, N. Hurley, F. J. Peña, P. Symeonidis, J. Geraci, A. Lawlor, PDMFRec: A decentralised matrix factorisation with tunable user-centric privacy, in: Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19), ACM, Copenhagen, Denmark, 2019, pp. 457–461.
[20] D. Wilkinson, M. Namara, K. Badillo-Urquiola, P. J. Wisniewski, B. P. Knijnenburg, X. Page, E. Toch, J. Romano-Bergstrom, Moving beyond a "one-size fits all": Exploring individual differences in privacy, in: CHI Conference on Human Factors in Computing Systems, ACM, Montreal, Canada, 2018, pp. 1–8.
[21] P. Story, D. Smullen, A. Acquisti, L. F. Cranor, N. Sadeh, F. Schaub, From intent to action: Nudging users towards secure mobile payments, in: Symposium on Usable Privacy and Security (SOUPS), 2020, pp. 379–415.
[22] T. Khazaei, L. Xiao, R. E. Mercer, A. Khan, Understanding privacy dichotomy in Twitter, in: Proceedings of the 29th on Hypertext and Social Media (HT '18), ACM, Baltimore, USA, 2018, pp. 156–164.
[23] K. Ghazinour, S. Matwin, M. Sokolova, YourPrivacyProtector: A recommender system for privacy settings in social networks, International Journal of Security, Privacy and Trust Management 2 (2013) 11–25.
[24] L. Li, T. Sun, T. Li, Personal social screen–a dynamic privacy assignment system for social sharing in complex social object networks, in: IEEE Third International Conference on Privacy, Security, Risk and Trust and IEEE Third International Conference on Social Computing, 2011, pp. 1403–1408.
[25] G. Rasch, Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests, Nielsen & Lydiche, 1960.
[26] H. Schäfer, M. C. Willemsen, Rasch-based tailored goals for nutrition assistance systems, in: Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19), ACM, Marina del Ray, USA, 2019, pp. 18–29.
[27] A. Starke, The effectiveness of advice solicitation and social peers in an energy recommender system, in: Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2019), Copenhagen, Denmark, 2019, pp. 65–71.
[28] B. P. Knijnenburg, A user-tailored approach to privacy decision support, Ph.D. thesis, University of California, Irvine, 2015.
[29] E. Aïmeur, Z. Sahnoune, Privacy, trust, and manipulation in online relationships, Journal of Technology in Human Services 38 (2020) 159–183.
[30] D. Andrich, Rasch Models for Measurement (Quantitative Applications in the Social Sciences), Sage Publications, 1988.
[31] [Online], https://www.winsteps.com/winsteps.htm, [Accessed 2021/07/29].
[32] R. Altman, D. A. Taylor, Social penetration: the development of interpersonal relationships,

1973.
[33] [Online], https://www.mturk.com/, [Accessed 2021/07/30].
[34] [Online], https://www.limesurvey.org/, [Accessed 2021/08/01].
[35] L. Yu, S. M. Motipalli, D. Lee, P. Liu, H. Xu, Q. Liu, J. Tan, B. Luo, My friend leaks my privacy: Modeling and analyzing privacy in social networks, in: Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies (SACMAT '18), ACM, Indianapolis, USA, 2018, pp. 93–104.
[36] A. Hamoud, E. Aïmeur, Handling user-oriented cyber-attacks: STRIM, a user-based security training model, Frontiers in Computer Science 2 (2020) 25.
[37] B. Knijnenburg, M. Willemsen, Z. Gantner, H. Soncu, C. Newell, Explaining the user experience of recommender systems, User Modeling and User-Adapted Interaction 22 (2012) 441–504.
[38] J. M. Such, N. Criado, Multiparty privacy in social media, Communications of the ACM 61 (2018) 74–81.