A model for system developers to measure the perceived privacy risk of users

Abstract

In this paper, we propose a model that could be used by system designers to measure the perceived privacy risk of users in general when they disclose data into a given software system. We first derive a model to measure the perceived privacy risk based on existing knowledge and then we test our model through a survey with 151 participants. Our proposed model showed that how visible data gets in an application by default when the user discloses data into the application had the highest impact on the perceived privacy risk of users. With the increasing concern over embedding privacy into software system designs, our model would help developers to measure the perceived privacy risk of users and how factors like data visibility affects the perceived privacy risk of users. Therefore, it would enable developers to effectively make use of privacy guidelines such as privacy by design and data minimization when they design privacy preserving software systems.

1. Introduction

The new European Data Protection Regulations that came into effect in 2018 has created considerable interest in software system designers towards privacy guidelines such as the Privacy by Design (PbD) principles [1]. However, PbD has been criticized for being incompatible and different to the usual activities developers perform when they design software systems [2]. One such limitation in PbD identified through developer studies is the need to make soft decisions due to the lack of measurable metrics when implementing privacy in software systems [3]. For example, the principle of Data Minimization (DM) requires developers to appropriately implement anonymization and encryption for data in systems. That is developers are expected to decide which data they should anonymize and which data they should not use. However, developers expect quantifiable metrics that would tell them which data they need to protect when they design systems [3]. Because of this, developers find it difficult to follow privacy guidelines as PbD and embed privacy into systems as required by regulations. This results in software systems with inadequate privacy embedded into the system that could invade user privacy [2].

Previous research has shown that the knowledge of the properties of data (such as how sensitive the content is and how visible the content is in a system) could be used [4] to measure the privacy risk of content in software systems. Consequently, previous research has identified that properties of data (such as the user's name, address) being disclosed have an effect on data disclosure decisions made by users. For example, when a user disclose data to a system, how sensitive is this data? and how relevant is this data to the application? are known to have an effect on the data disclosure decisions made by users [5]. Building on this knowledge we propose a metric that would enable software developers to measure the privacy risk of the data they use in the system as perceived by users. Then, using a survey with 151 respondents we observe how good our model fits with the actual privacy risk perceived by users. This metric would help developers to better understand the data they use in the system. It has been said that understand data from a user perspective would enable developers implement data protection in software systems [6].

The paper is structured as follows. We first discuss the background of perceived privacy risk of users, and privacy risk measurement to establish the grounds on which our work stand. Then, building on the existing theoretical knowledge on measuring privacy risk, we first logically build our model to measure user privacy risk associated with disclosing data items in a given software system setting. Then, we describe the experiment we conducted to measure actual privacy risk perceived by users when they disclose their data. Next, we present our results where we show how good out model fits the observations, followed by a discussion of the observed variations of our model. Finally, we present our conclusions and directions for future research

2. Background

Our focus in this research is to develop a metric to measure the users' perceived privacy risk associated with data (such as their name, address and email address) in software systems (such as their baking app, their social networking account etc.). We first identified that users' perceived privacy risk can be interpreted as their discomfort or reluctance for data disclosure in software systems [7–9]. Users are most likely to disclose data into a software system when they feel that the ways through which the system could compromise their privacy (data sharing, data selling) from the data they disclose is minimum [7–9]. Therefore, understanding the data disclosure decisions made by users when interacting with software systems could help understanding their perceived privacy risk. Nevertheless, among many research that attempts to interpret users privacy risk and their data disclosure decisions [5, 8, 10–12], so far no attempt has been made to communicate this perceived privacy risk of users when they disclose data into systems, in a comprehensive way to software developers.

Most research that observe disclosure decisions of users attempt to identify factors that could increase data disclosure. For example, it is said that users are more likely to decide to disclose data when they are shown the decisions made by their friends. [12] or other users [13]. Similarly, Acquisti et al. found that changing the order of intrusiveness of the data being requested also makes users disclose more data when interacting with software systems [14]. Furthermore, testing the effect of the justification provided by the system when requesting data Knijnenburg and Kobsa [15] revealed that when users are told *this data is useful for you* users are more likely to disclose data to the application. Nevertheless, these research focus either on the features of the system that requests data [5, 8, 11] or the personality of the user who discloses data [16] and attempt to find ways to increase user data disclosure to collect and use more data in software systems [12].

Consequently, focusing on the intrinsic properties of the data being shared, Bansal et al. have shown that users' intention to disclose health information is affected by the sensitivity of the data [17]. This intrigued our interest. Malhotra et al. have also shown that consumer willingness to share personal data in commercial platforms is affected by the sensitivity of the data [9]. Similarly, Malheiros et al. [5] have shown that sensitivity of data items such as date of birth and occupation had a significant affect on the decisions of the users to disclose that data into software systems. Interestingly, they have also identified an effect of data relevance for the application context on the disclosure decisions made by users. However, how these parameters correlate when users make their decisions to disclose data and how software developers could make use of this information when they design software systems are not yet known.

Interestingly, from a perspective of privacy risk measurement, Maximilien et al. [4] have shown that a metric for privacy risk in a given context can be obtained by multiplying the measurement for sensitivity of a data item with the measurement for visibility the data item gets in an application. They define their metric for privacy risk as "a measurement that determines their [the user's]willingness to disclose information associated with this item" [4]. Using this metric, Minkus and Memon [18] have attempted to measure the privateness of Facebook users from their privacy settings. However, privacy risk is a contextual measurements. The context in which data is being disclosed [16, 19] is known to have an effect on user disclosure decisions [10]. For example, it is said that users have a negative attitude towards rewards for data disclosure when the requested data appears irrelevant for a system [8], whereas they accepted the rewards if the data is relevant for the system. However, in the current metric for privacy risk measurement of content, this measurement of data relatedness is missing.

In this research, we focus on the effect of data sensitivity, the relevance of the data for an application and the visibility the data gets in the application on the perceived privacy risk of users. With this we focus on obtaining a privacy risk metric that could communicate the effect of data sensitivity, visibility and the relatedness of data for a particular application on the perceived privacy risk of users to software developers and privacy researchers. By software developers, we refer to all those who are involved in making the decisions on collecting data, designing and implementing software systems. This metric would help them to understand and incorporate perceived privacy risk of users into the software system designs and assist the development of privacy preserving software systems. For example, they could identify which data users are most concerned about and which data users would feel most uncomfortable sharing. This knowledge could help them implement better security for data in system designs and communicate it to the user in order to actively reduce the perceived privacy risk of users when they interact with software systems.

3. Research Methodology

In this section we first introduce the parameters of data we are interested in. Then using these parameters we derive and propose a model to measure privacy risk

of data items based on existing theoretical knowledge.

3.1. Model to calculate privacy risk of data elements

The goal of our research was to develop a measurement to calculate the perceived privacy risk of users when they disclose data into software systems. Referring to previous research we identified data sensitivity (S), relatedness (R) and visibility (V) of data on the perceived privacy risk of users when they make the disclosure decisions. For the context of this research we define data sensitivity, visibility and relatedness of data to be parameters that depend only on a particular data item D_i and the application context in which it is being used C_j.

3.1.1. Data Sensitivity We define the sensitivity of a particular data item to be a parameter that is dependent on the data item D_i itself. That is inherently for a user their credit card number is more sensitive than their age. We define sensitivity of a data item to be the perceived impact of loss of that particular data item. We define sensitivity in three categorical values. These categories are defined based on the definition of sensitive data in the European Data Protection Regulations (GDPR) [1] and logical reasoning. In defining categories thee categories is considered to be cognitively more manageable than complex scales with more levels of categorization [20]. Therefore, we based our categorization into three categories as given in table 1.

Table 1. Data Sensitivity

Category	Description	Sensitivity Value
Category I - Highest sensitivity	Data that could be used to identify a unique characteristic of a person. For example, a person's race, religion or HIV status.	3
Category II - Moderate sensitivity	Personally Identifiable information about the person. For example, a person's name, address, mobile number	2
Category III - Low sensitivity	Any other detail about a person that may have an impact of loss, however, would not affect the person. For example, a person's high school	1

Therefore according to our definition the sensitivity of a data element D_i takes categorical values $S_i \in \{1,2,3\}$.

3.1.2. Data visibility We define the visibility of a data element to be an inherent property gained by a particular data element D_i in a particular application context C_j due to the design of the application. That is how visible the data item would be by default once the user disclose the data item to the application. If the application by default allows the data to be seen only by the user, we define that data item has the lowest visibility. These categories are defined on the basis of the survey conducted by Minkus et al. [18] in their attempts to scale Facebook privacy settings according to their visibility, they have asked participants questions that investigate the users perception of visibility of their content in Facebook. Building on their reasoning we logically form the three visibility categories presented in Table 2.

Table 2. Data Visibility

Category	Description	Visibility Value
Category I - Highest visibility	Data would be seen by any one by default. Data is visible in the application by default. For example the name of a user in Facebook	3
Category II - Moderate visibility	Data would be seen by a controlled set of users by default. For example, content that can be only see by the friends of the user in Facebook	2
Category III - Low visibility	Data would be seen by any one by default. Data is visible in the application by default. For example, your pin number in the banking app will not be visible to anyone	1

Therefore according to our definition the visibility of a data element D_i in an application context C_j takes categorical values $V_{i,j} \in \{1,2,3\}$.

3.1.3. Data Relatedness We define the relatedness of a data element D_i to be a property that is defined by the application context C_j. That is based on the requirements of the application, the data could be highly related to the application (For example, your bank account number for your banking application) or not related at all. This is determined by the primary functionality of the application defined by the application requirements. We build this categorization based on logical reasoning. While it has been widely accepted that the relatedness of data affects the privacy risk perceived by users when they disclose data into software systems [16, 19], so far there is no evidence as to how related a data item should be in order to make users feel comfortable sharing those data into the system. Therefore, based on logical rea-

soning, we propose the categorization present in table 3 for scaling data relatedness to a software system.

Table 3. Data Relatedness

Category	Description	Relatedness Value
Category I - Highest relatedness	Data the application cannot do without. These data are absolutely necessary for the primary functionality of the application	3
Category II - Moderate relatedness	Data could add additional functionality to the application. For example, data that could deliver benefits through data analysis techniques	2
Category III - Low relatedness	Data the application can do without. For example, data that is not needed for the functionality of the application	1

Therefore according to our definition the relatedness of a data element D_i in an application context C_j also takes categorical values $R_{i,j} \in \{1,2,3\}$.

3.1.4. Model to calculate privacy risk of a data element D_i in an application context C_j We define the calculated privacy risk $P_{i,j}$ of a data element D_i in an application context C_j as follows.

Building up on the relationship proposed by Maximilien et al. [4] we define that the privacy risk $P_{i,j}$ of a data element D_i in an application context C_j monotonically increases with the sensitivity of a data item S_i and the visibility of a data item in a given context $V_{(i,j)}$. This has been previously used by Minkus and Memon [18] in determining the privacy level of Facebook privacy settings for a particular user. Then, we propose that the privacy risk P_c of a data element D_i in an application context C_j is in a monotonically decremental relationship with the relatedness of the data element D_i to the application context C_j. This is based on the knowledge that users perceive low privacy risk when disclosing data items that are relevant to the application as opposed to data elements that do not appear relevant [15]. Therefore, we propose that an approximation for the privacy risk $P_{i,j}$ of a data element D_i in an application context C_j can be obtained by,

$$\text{Privacy Risk } P_{(i,j)} = \frac{S_i^a \times V_{(i,j)}^b}{R_{(i,j)}^c}$$

where a,b and c values could take any real number. However, as we are aiming for an approximation we limit a,b,c to whole numbers.

According to this calculation Privacy Risk $P_{(i,j)}$ of a data element D_i in an application context $C_j \in \{\text{x}|\ \text{x}\ in\ \mathbb{R}$ where, $0 < \text{x}\}$. This relationship could be used to measure the privacy risk of data in a given application context so that developers and system designers could get an idea as to how appropriate privacy measures should be implemented for data items in an application design. We argue that this numeric measurement of privacy measurement would be meaningful for software developers than the soft measurements developers are expected to make in most scenarios that involve user privacy. For example, it has been previously coined that when implementing privacy in software systems, developers find it difficult to interpret the requirements to anonymize appropriate data, encrypt sensitive data, when decisions are not measurable [3]. Next, in order to see how closely the proposed model fit the actual perceived privacy risk of users when they disclose data we conducted a survey study.

3.2. Research Study

Our goal in conducting the research study is to observe how close the relationship we proposed using data sensitivity, visibility and relatedness approximate the actual perceived privacy risk by users. Building on the work of Maximilien et al. [4] we define perceived privacy risk $P_{i,j}$ to be "a measurement that determines the user's feeling of discomfort in disclosing an data item D_i in an application context C_j". We conducted two separate user studies for this research.

3.2.1. Study I : The first study was an online survey with 151 internet users to obtain the dependent variable of our model, the perceived privacy risk of users $P_{i,j}$. For this we defined three data disclosure scenarios.

- Health Care application that allows remote consultancy with doctors - with data being visible to the user and the doctor.
- Social Networking application - with no control over data visibility (Cannot control who can view the data once disclosed)
- Banking application - with the data being visible only to the user (and the bank)

We communicated three different visibility levels in the three application contexts. We used ten data items including demographic data and sensitive data following the European Data Protection Regulations [1]. The data items we provided are name, age, address, mobile number, email address, occupation, blood type, credit card

number, medicine taken, and birthday. We asked the participants how they would feel if they are to disclose these 10 data items in the four application contexts. We define a five point Likert scale to express their *feeling of disclosure* F_d, with values, very uncomfortable, somewhat uncomfortable, neutral, somewhat comfortable and very comfortable. We alternatively used reverse ordered Likert scales to ensure the validity of the answers. We consider F_d to be a function of the sensitivity of the data item i (S_i), visibility of the data item in the application j ($V_{i,j}$) and the relatedness of the data item to the context of the software system j ($R_{i,j}$). Our goal is to determine how close the relationship we proposed approximate F_d.

Following these four questions we also included an open ended question in the questionnaire to further observe the reasons for the difference in the feeling of discomfort (F_d) users expressed. With this we aimed to obtain further insights as to why users demonstrate different discomfort levels when they disclose different data items into different application contexts.

At the end of the survey, we included questions to extract the demographics of the participants. However, we included an option *prefer not to say* in all these questions, so that users could avoid disclosing their age, gender and educational background. Tables 4 provides the basic profile of the participants;

Table 4. Participants

Gender	**No. of Participants**
Male	87
Female	64
Education	
Completed School Education	5
Professional Diploma	9
Bachelor's Degree	87
Masters/PhD	50
Age	
18-24	31
25-32	101
33-40	13
41 or above	6

The survey design was evaluated with two participants (graduate students in the university not connected to the research). We fine tuned the wording of the questionnaire with the feedback of these two participants. Then the survey was distributed using social media platforms (Facebook, LinkedIn and Twitter) and personal connections of the authors. The research methodology (survey design, participant recruitment and results collection) was approved by the university ethic committee responsible for ethical conduction of studies that involve human subjects.

In the invitation email we sent to participants, we included a brief introduction about the survey and the duration of the survey (under 10 minutes, calculated using the participants who evaluated the questionnaire). We provided the participants with the contact details of the researchers in case they wanted to contact us for more information. Before proceeding with the survey participants were given an introduction to the survey with details about the survey and the type of data we collect. We also informed the participants that they could exit the survey at any time without submitting their answers. Participants were asked to proceed with the survey if they give us (the researchers) consent to collect and store the details they submit with the survey.

We measured the participant adequacy while collecting data and stopped data collection when we reached sample adequacy at KMO = 0.8 (A KMO value 0.8 is considered good in calculating correlations [21]). We had 157 responses at that point. We then analyzed the data and eliminated 6 responses that were either incomplete or invalid as the participant had selected the same choice in the Likert scale for all options.

To transform the likert scale input into a measurement of the feeling of discomfort of the participants, we assigned values from 1 to 5 for the answers we received on the Likert scale as given in Table 7.

Table 5. Assigning values to Likert Scale preferences

Likert Scale Preference	Value Assigned
Very Comfortable	1
Somewhat Comfortable	2
Neutral	3
Somewhat Uncomfortable	4
Very Uncomfortable	5

Through this we obtained $F_d \in \{1,2,3,4,5\}$ of users for the 30 scenarios (ten data items in three application contexts) that represent the user's feeling of discomfort in disclosing data.

3.2.2. Study II : The second study was a focus group with 4 software developers, to obtain the dependent variables of the model (sensitivity, visibility and relatedness) for the three data disclosure scenarios we used in the survey. As our goal is to introduce a metric for software developers to evaluate the perceived privacy risk of users, we calculated $P_{(i,j)}$ through a focus group with 4 participants with a software development experience. We believe this approach would closely represent the context in which software developers would discuss

and evaluate the sensitivity, visibility and the relatedness of the data elements they use in software systems, at design stage. The focus group took 40 minutes and the participants were volunteers.

In the focus group we first discussed the data items as individual elements and categorize them according to the sensitivity of the data item. For this we provided the participants with the three categorical definitions we defined in table I. Next, for all three application scenarios, we asked the developers to categorize the ten data items according to their relatedness to the application context and provided them with table III. We encouraged the participants to raise arguments and discuss and clarify different opinions in categorizing data. As visibility was pre-determined when we defined the three application scenarios in the survey and communicated to users in the user study we did not evaluate it here. During the focus group, we also evaluated our model for data categorization presented in Table I, II and III. We encouraged the participants to argue and raise any concerns they had on the three categories we defined and their appropriateness in categorizing the data. We discuss the concerns raised by the participants in the focus group when we discuss our findings.

3.2.3. Data Analysis After obtaining the S,V,R combinations for the 30 scenarios, we first calculated the perceived privacy risk for the 30 scenarios using the model we propose. Then, we tested in order to test our model we tested the calculated perceived privacy risk against the actual perceived privacy risk values we obtained through the user study. We first attempted to fit our model on the raw data available (151 users and 30 instances, altogether 4530 instances). However, due to the relatively high variation of data, it was not possible to fit a model to the data set. That is, the same combination of S,V, R values had multiple perceived privacy risks varying from 1 to 5. This is expected because users have very different perceived privacy risks. Therefore, we then averaged the perceived privacy risk of all 151 users to obtain 30 distinct mean perceived privacy risk values for the 30 scenarios tested. Then we used these values to observe the goodness of fit of our proposed model in Matlab.

Then, we used qualitative methods to analyse the answers to the open ended question using two independent coders. We followed the grounded theory approach where the coders coded data by eliciting codes from the data available without any prejudice [22]. This was done in NVivo [23]. Coders reached code saturation at 49 and 103 respectively. The two coders came up with 6 common codes and 7 and 20 codes present in either of the coders at the end. This was because one coder had very granular level codes while the other code had coded data at a higher level. For example, one coder had a code saying *concerns on controlling data visibility*, while the other coder had three separate codes for the same content as *controlling who can see my data*, *application providing tools to hide data from public* and *controlling data in the app*. Then both coders iteratively evaluated their codes and merged similar codes together to come up with 11 final codes that explain the differences in perceived privacy risks in the participants.

4. Results

We tested the validity of our results with Cronbach's alpha (0.91) (a Cronbach's alpha > 0.7 is considered acceptable [24]) and the participant adequacy for correlations with KMO (KMO = 0.8269).

Following charts (image 1-4) shows the averages of the disclosure feeling of the 151 participants on the 10 data items across the three scenarios. It can be seen that in all scenarios except for the banking app users had the highest discomfort in sharing their credit card information, and this was followed by medical information except for the medical application, which suggests users feel higher risk when disclosing sensitive data yet, it was reduced when they felt that the data was related to the application.

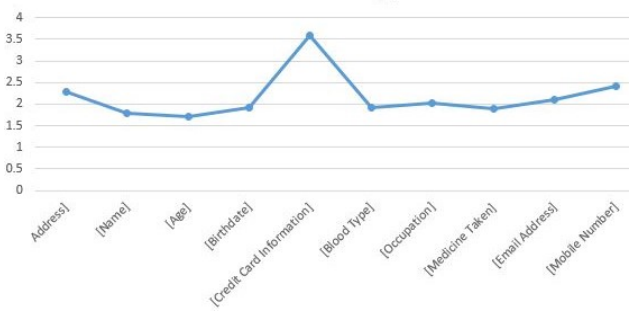

Figure 1. Feeling of Discomfort in Disclosure - Health application

Table 5 shows the results when we tested our model for calculated privacy risk $P_{(i,j)}$ against the perceived privacy risk F_d. In this table SSE : Sum of Squares due to Error, R^2 : Square of the correlation between the calculated $P_{i,j}$ and the observed $P_{i,j}$, and RMSE : Root Mean Squared Error.

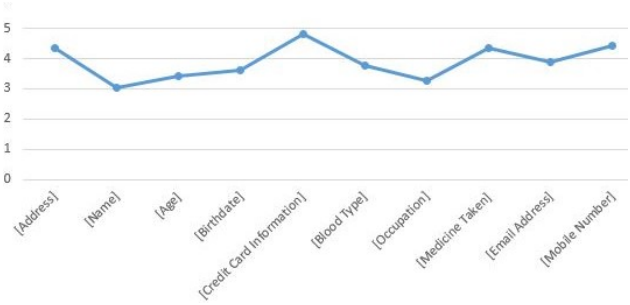

Figure 2. Feeling of Discomfort in Disclosure - Social Networking application

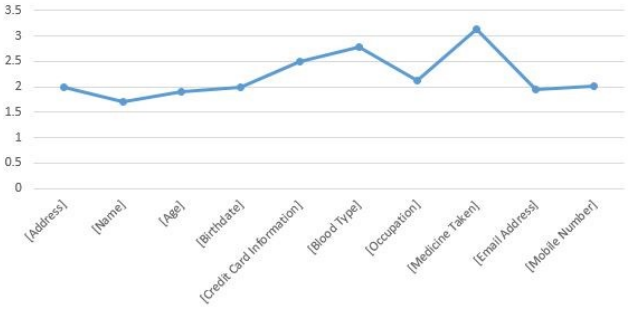

Figure 3. Feeling of Discomfort in Disclosure - Banking application

Table 6. Model Fitting - basic model

Model	a (95% CI)	Goodness of fit		
		SSE	R^2	RMSE
$\frac{S_i^1 \times V_{(i,j)}^1}{R_{(i,j)}^1}$	0.24	67.8	0.6	1.5

As seen in the table, that when we give the same power to all three parameters in the relationship the error is relatively high with a low R-square value. Therefore, we tried all 27 combinations of the powers 1,2 and 3 for S,V,R combinations (Table 6) without the combinations where all parameters got the same power. That is we ignored the combinations (1,1,1), (2,2,2) and (3,3,3).

Table 7. Model Fitting

Model	a (95% CI)	Goodness of fit		
		SSE	R^2	RMSE
$a(\frac{S_i^1 \times V_{(i,j)}^1}{R_{(i,j)}^2})$	0.24	15.22	0.4353	0.7373
$a(\frac{S_i^1 \times V_{(i,j)}^1}{R_{(i,j)}^3})$	0.21	16.7	0.3803	0.7723
$a(\frac{S_i^1 \times V_{(i,j)}^2}{R_{(i,j)}^1})$	0.10	9.335	0.6536	0.5774
$a(\frac{S_i^1 \times V_{(i,j)}^2}{R_{(i,j)}^2})$	0.08	12.57	0.5336	0.67
$a(\frac{S_i^1 \times V_{(i,j)}^2}{R_{(i,j)}^3})$	0.08	14.4	0.4657	0.7171
$a(\frac{S_i^1 \times V_{(i,j)}^3}{R_{(i,j)}^1})$	0.03	8.285	0.6926	0.544
$a(\frac{S_i^1 \times V_{(i,j)}^3}{R_{(i,j)}^2})$	0.03	11.73	0.5646	0.5646
$a(\frac{S_i^1 \times V_{(i,j)}^3}{R_{(i,j)}^3})$	0.02	13.71	0.4912	0.6998
$a(\frac{S_i^2 \times V_{(i,j)}^1}{R_{(i,j)}^1})$	0.08	13.94	0.4828	0.7055
$a(\frac{S_i^2 \times V_{(i,j)}^1}{R_{(i,j)}^2})$	0.07	15.38	0.4294	0.7411
$a(\frac{S_i^2 \times V_{(i,j)}^1}{R_{(i,j)}^3})$	0.07	16.45	0.3895	0.7666
$a(\frac{S_i^2 \times V_{(i,j)}^2}{R_{(i,j)}^1})$	0.03	11.06	0.5897	0.6284
$a(\frac{S_i^2 \times V_{(i,j)}^2}{R_{(i,j)}^3})$	0.02	14.78	0.4515	0.7266
$a(\frac{S_i^2 \times V_{(i,j)}^3}{R_{(i,j)}^1})$	0.01	10.07	0.6264	0.5996
$a(\frac{S_i^2 \times V_{(i,j)}^3}{R_{(i,j)}^2})$	0.009	12.74	0.5271	0.6746
$a(\frac{S_i^2 \times V_{(i,j)}^3}{R_{(i,j)}^3})$	0.009	14.38	0.4665	0.7166
$a(\frac{S_i^3 \times V_{(i,j)}^1}{R_{(i,j)}^1})$	0.02	14.37	0.4669	0.7163
$a(\frac{S_i^3 \times V_{(i,j)}^1}{R_{(i,j)}^2})$	0.02	15.31	0.432	0.7394
$a(\frac{S_i^3 \times V_{(i,j)}^1}{R_{(i,j)}^3})$	0.02	16.22	0.3982	0.7611
$a(\frac{S_i^3 \times V_{(i,j)}^2}{R_{(i,j)}^1})$	0.009	11.68	0.5664	0.646
$a(\frac{S_i^3 \times V_{(i,j)}^2}{R_{(i,j)}^2})$	0.009	13.56	0.497	0.6958
$a(\frac{S_i^3 \times V_{(i,j)}^2}{R_{(i,j)}^3})$	0.008	14.86	0.4485	0.7286
$a(\frac{S_i^3 \times V_{(i,j)}^3}{R_{(i,j)}^1})$	0.003	10.78	0.5998	0.6206
$a(\frac{S_i^3 \times V_{(i,j)}^3}{R_{(i,j)}^2})$	0.003	13.12	0.513	0.6846

From the above result we can see that the goodness of fit increases with the increase of the power of visibility and decreases when the power of sensitivity and relatedness increases. Therefore, we then gradually increased the power of visibility and tested the goodness of fit while keeping the power of sensitivity and relatedness at 1. Table 7 shows the values we received.

We can see that the error increases again the power of visibility increases beyond 7. Therefore, the optimal relationship with the best goodness of fit is in the model where visibility is raised to the power of 7 with a co-

Table 8. Model Fitting - increasing visibility

Model	a (95% CI)	Goodness of fit		
		SSE	R^2	RMSE
$a(\frac{S_i^1 \times V_{(i,j)}^4}{R_{(i,j)}^1})$	0.01	7.872	0.7079	0.5302
$a(\frac{S_i^1 \times V_{(i,j)}^5}{R_{(i,j)}^1})$	0.003	7.723	0.7134	0.5252
$a(\frac{S_i^1 \times V_{(i,j)}^6}{R_{(i,j)}^1})$	0.001	7.682	0.7149	0.5238
$a(\frac{S_i^1 \times V_{(i,j)}^7}{R_{(i,j)}^1})$	0.01	7.682	0.715	0.5238
$a(\frac{S_i^1 \times V_{(i,j)}^8}{R_{(i,j)}^1})$	0.01	7.693	0.7145	0.5242
$a(\frac{S_i^1 \times V_{(i,j)}^9}{R_{(i,j)}^1})$	4.378e-05	7.706	0.7141	0.5246

efficient of 0.01. This had a SSE of 7.6 and an R^2 of 71.5%. However the increase of R^2 from the model with visibility to the power three to visibility to the power 7 is only almost 1%. Therefore, one could safely assume the model,

$$\frac{0.03 \times S_i \times V_{(i,j)}^3}{R_{(i,j)}}$$

gives a good enough approximation of the perceived privacy risk of users for a data item *i* in a software application *j*. From the results, it is apparent that the visibility has the largest effect on the perceived privacy risk of users.

In order to further observe why users felt differently when they disclosed data in the three scenarios we described, in the next section we present the qualitative analysis of the reasons users gave.

4.0.1. Qualitative analysis on factors that affect the feeling of discomfort in data disclosure

Table 8 gives the summary of the codes we generated through the qualitative analysis. We developed a total of 11 codes.

When it comes to the properties of data, participants mentioned only sensitivity, relevance and visibility of the data items that affect their disclosure decisions. We could not identify any other attribute related to the data itself that affected the perceived privacy risk of users when they disclosed data. Participants were most concerned about the relevance of data (26%) followed by sensitivity of data (15%) and visibility (12%). Nevertheless, our model showed that the visibility of data had the highest impact on the perceived privacy risk of users. For example, in the descriptive answers P146 said *If the application provides some tools to hide private information from public, it is fine* and P87 said *the controls on the data we disclosed are important*. Therefore, when designing software systems, if the system could control the visibility of data in the system and communicate how visible the data would be once the user disclose data into the system, it would help reducing the perceived privacy risk of users.

Consequently, we identified that users are concerned about the trust towards the organization that develop and publish applications (19%). Participants said that they are comfortable sharing data as long as the application is developed and owned by a trusted organization. This was observed in the mean perceived privacy risk of users we calculated for the three application contexts. We observed a relatively low mean perceived privacy risk for the scenario with the banking app, probably because users trusted their bank more. Some participants spoke about the trust with the application itself rather than the organization (11%). Some participants also raised concerns about personal safety (12%). Their concerns on personal safety was two fold. One was on financial and reputation loss on data being accessed by unknown parties. The other was their concern on being subjected to unwanted marketing via phone and email. They said that they consider this as a personal threat and hence they think twice before disclosing data to any application. A small number of participants were concerned about the previous personal experience and also about the benefit of sharing the data.

5. Discussion

The model we tested in this research was derived based on the theoretical knowledge presented by Maximilien et al. [4]. They propose that privacy could be measured by sensitivity and visibility where their combination is any arbitrary expression that results in a monotonically incremental result for privacy risk. However, their model has been applied on the assumption that both sensitivity and visibility of content has the same effect on the privacy risk of the content [18]. At the same time their model did not account for the relatedness of the content. In our model we introduced a term for relatedness of the content and through the user study we were able to identify that the content visibility had more impact on the privacy risk of the content than data sensitivity and relatedness. Therefore, when measuring privacy risk of content in similar environments using either the original relationship by Maximilien et al. [4] we suggest that content visibility should be taken at a higher power to closely approximate the perceived privacy risk of users. We also argue that for a more accurate measurement, the relatedness of the content to the application should also be taken into account.

For the categorization of data according to S, V and R we used three categories. In the workshop to deter-

Table 9. Issues participants faced when embedding privacy into the designs

Code	Representative Quotes	Coverage
Benefit to me	how it benefits myself/ how useful it is for me.	2.64% (4)
How much I need the app	based on my requirements from the application	7.2%(11)
News I see	by considering cyber crimes and all that	0.66%(1)
Personal experience	I was in couple of these situations which gave me an idea	2%(3)
Personal Safety	Some data are highly confidential and could end up in a reputation and/or financial loss/ don't like to see unwanted advertisements and messages	12% (19)
Relevance of data to the purpose	if I don't think such applications needs the data. For instance my blood group for a banking app	26% (40)
Visibility of Data - who can see it	audience with access to the data/ as in whether I could control what others see	12% (19)
Sensitivity of Data	As long as the requested information is not sensitive/ some sensitive information can't be disclosed irrespective of the application	15% (23)
Transparency - knowing how the data is used	Depends on what they are going to do with the information/ when privacy is not guaranteed	6.6% (10)
Trust with the application	every online application cannot be trusted/ random Facebook applications are not safe	11% (17)
Trust with the organization	If it is a reputed or a government institution there is less doubt and more trust on data security	19% (29)

mine S,V and R values with software developers, we encouraged the developers to further define categories if they felt three categories were not sufficient. We also asked them to challenge and argue on the definitions we have provided. While the participants agreed with the categories for V and R, they said that S may require more categories to identify sensitive data and extremely sensitive data. However, when we agreed and created one more category, they ended up moving all data in the sensitive category to the extremely sensitive category and hence ending up with three categories at the end. Therefore, the participants agreed that the three categories we defined sufficiently captures the S,V R variation in data.

Consequently, the model we derived here does not account for the human attributes of users that affect their perceived privacy risk when interacting with software systems. Previous research has shown that the personality of users affects the perceived privacy risk of users when they interact with software systems. For example, Westin's privacy personality scale [25] shows that users could be divided into privacy fundamentalists, who are extremely concerned about their privacy, privacy pragmatists, who understand that privacy needs to be compromised according to situations and privacy unconcerned, who are either little not concerned about their privacy [25]. Indicating the effect of such personalities on their perceived privacy risk, in our survey P41 said *Basically I feel comfortable giving information on a need to know basis only* and P114 said *nothing* implying he did not feel different disclosing data into different application settings. This could be explained by the theory of psychometry, which explains why people's perception of external factors such as privacy is dependent on their psychological differences [9, 26]. There is a lot of work done in this area where privacy psychometry is scaled and defined. For example IUIPC is one such scale that defines how people differ in their privacy attitudes [9]. Consequently, there could exist other attributes such as previous experience of users, their age and the nature of work they do that may affect their perceived privacy risk. For example, P5 said *With the experiences when surfing in the internet made me to answer above questions so* and P89 said *I was in couple of these situations which gave me an idea to answer these questions easily*. However, in this research our focus was to model the perceived privacy risk eliminating the personality traits of a person. Therefore, by design we did not capture the privacy profile of our participants. The model we tested had an SSE value of 7.682 and an R_2 value of 71%. This could be taken as an acceptable goodness of fit in a human study. While the variations in the model could probably be explained by human factors, for the purpose of deriving a model for software developers to approximate the perceived privacy risk of the data used in software systems, we believe our model is appropriate.

As future work, we aim to improve our study with privacy profiling of participants incorporating the models that capture psychometric measurements [9, 25, 26], in order to observe how our model could cater for users with different privacy personalities.

6. Conclusion and Future Work

In this research we used the sensitivity of data, the visibility data gets in a system design and the relatedness of data to the system as the independent variables in the model and proposed the model based on existing theoretical knowledge. We then tested our model against actual perceived privacy risk of users in three different application settings. Our results indicate that both sensitivity and visibility of content must be in a monotonically increasing combination to represent privacy where visibility of content is given a higher power. At the same time relatedness of the content should be in a combination with sensitivity and visibility such that privacy risk monotonically decrease with the relatedness. We believe that this knowledge could be used by software developers (those who are involved in developing, designing and defining software systems) to measure the perceived privacy risk of the data they use in the systems they design. With this knowledge, they could implement better security for data with higher perceived privacy risk and communicate the system functionalities to users in order to reduce the perceived privacy risk of users.

References

[1] J. Wagner and A. Benecke, "National legislation within the framework of the gdpr," *European Data Protection Law Review*, vol. 2, no. 3, pp. 353–361, 2016.

[2] A. Senarath, N. A. Arachchilage, and J. Slay, "Designing privacy for you: A user centric approach for privacy," *19th International Conference on Human-Computer Interaction, Vancouver, Canada*, 2017.

[3] A. Senarath and N. A. G. Arachchilage, "Why developers cannot embed privacy into software systems? an empirical investigation," in *Proceedings of the 22nd conference of Evaluation and Assessment in Software Engineering (EASE) 2018*, p. to appear, ACM, 2018.

[4] E. M. Maximilien, T. Grandison, T. Sun, D. Richardson, S. Guo, and K. Liu, "Privacy-as-a-service: Models, algorithms, and results on the facebook platform," in *Proceedings of Web*, vol. 2, 2009.

[5] M. Malheiros, S. Preibusch, and M. A. Sasse, "fairly truthful: The impact of perceived effort, fairness, relevance, and sensitivity on personal data disclosure," in *International Conference on Trust and Trustworthy Computing*, pp. 250–266, Springer, 2013.

[6] B. Marr, *Big Data: Using SMART big data, analytics and metrics to make better decisions and improve performance*. John Wiley & Sons, 2015.

[7] A. Kobsa, "Privacy-enhanced web personalization," in *The adaptive web*, pp. 628–670, Springer-Verlag, 2007.

[8] H. Li, R. Sarathy, and H. Xu, "Understanding situational online information disclosure as a privacy calculus," *Journal of Computer Information Systems*, vol. 51, no. 1, pp. 62–71, 2010.

[9] N. K. Malhotra, S. S. Kim, and J. Agarwal, "Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model," *Information systems research*, vol. 15, no. 4, pp. 336–355, 2004.

[10] B. P. Knijnenburg and A. Kobsa, "Making decisions about privacy: information disclosure in context-aware recommender systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 3, no. 3, p. 20, 2013.

[11] J. Wang, N. Wang, and H. Jin, "Context matters?: How adding the obfuscation option affects end users' data disclosure decisions," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pp. 299–304, ACM, 2016.

[12] D. Dennett, "With a little help from my friends," *Dennetts philosophy: A comprehensive assessment*, pp. 327–388, 2000.

[13] A. Besmer, J. Watson, and H. R. Lipford, "The impact of social navigation on privacy policy configuration," in *Proceedings of the Sixth Symposium on Usable Privacy and Security*, p. 7, ACM, 2010.

[14] A. Acquisti, L. K. John, and G. Loewenstein, "The impact of relative standards on the propensity to disclose," *Journal of Marketing Research*, vol. 49, no. 2, pp. 160–174, 2012.

[15] B. P. Knijnenburg and A. Kobsa, "Helping users with information disclosure decisions: potential for adaptation," in *Proceedings of the 2013 international conference on Intelligent user interfaces*, pp. 407–416, ACM, 2013.

[16] H. Nissenbaum, *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

[17] G. Bansal, D. Gefen, *et al.*, "The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online," *Decision support systems*, vol. 49, no. 2, pp. 138–150, 2010.

[18] T. Minkus and N. Memon, "On a scale from 1 to 10, how private are you? scoring facebook privacy settings," in *Proceedings of the Workshop on Usable Security (USEC 2014). Internet Society*, 2014.

[19] L. K. John, A. Acquisti, and G. Loewenstein, "Strangers on a plane: Context-dependent willingness to divulge sensitive information," *Journal of consumer research*, vol. 37, no. 5, pp. 858–873, 2010.

[20] M. C. Oetzel and S. Spiekermann, "A systematic methodology for privacy impact assessments: a design science approach," *European Journal of Information Systems*, vol. 23, no. 2, pp. 126–150, 2014.

[21] J.-O. Kim and C. W. Mueller, *Factor analysis: Statistical methods and practical issues*, vol. 14. Sage, 1978.

[22] R. Y. Wong, D. K. Mulligan, E. Van Wyk, J. Pierce, and J. Chuang, "Eliciting values reflections by engaging privacy futures using design workbooks," *Proceedings of the ACM on Human Computer Interaction*, vol. 1, no. 2, 2017.

[23] J. Saldaña, *The coding manual for qualitative researchers*. Sage, 2015.

[24] J. C. Nunnally, I. H. Bernstein, and J. M. t. Berge, *Psychometric theory*, vol. 226. McGraw-hill New York, 1967.

[25] A. Westin, H. LOUIS, *et al.*, "Equifax-harris consumer privacy survey," *Conducted for Equifax Inc*, 1991.

[26] S. Egelman and E. Peer, "Predicting privacy and security attitudes," *ACM SIGCAS Computers and Society*, vol. 45, no. 1, pp. 22–28, 2015.