XXXXX

***Abstract*—**

I. Introduction

Users disclose their personal data into numerous software systems, such as web sites, mobile applications and even into payment systems when they tap their bank card every time they shop [1]. With the pervasiveness of technology users today disclose a huge amount of personal information, such as their address, credit card and banking details and even their health conditions, into online applications. When users disclose more data into software applications, the ways through which user privacy could be compromised through these software systems increase.

Disclosure of personal details always comes with an associated privacy risk that concerns users. Users are known to be reluctant to disclose data into software systems they cannot trust, because using data for reasons unforeseen by users and loss of data to third parties unknown to the user could result in privacy vulnerabilities. Because of this, privacy conscious users face many problems when they make decisions to disclose their information into software systems such as : How sensitive is this data for me? Who can view this data in this application? For what purposes will this data be used? However the apparent effect these concerns have on the disclosure decisions of the user is not directly observable. Lack of understanding about the aspects that concern users when they disclose data not only makes users uncomfortable when interacting with software systems, but also makes it difficult for privacy researchers and software developers to make data collection decisions when designing software systems.

In order to ensure user privacy preferences are met, developers need to be able to measure and quantify privacy of certain data items are for users. For such a measurement to be done, first one needs to identify the factors that affect privacy of data items for a user and the relationship among these factors. For example, would users be more comfortable sharing their credit card number with their banking application than their social networking account? Would users be equally comfortable sharing their blood group with their banking application? Would users be more comfortable sharing their age than their birthdate with their social networking account? Answering these questions are difficult yet important. A privacy risk metric that could determine the privacy of different data items in a given setting would be beneficial for users, software developers and also privacy researchers. From a user perspective, knowing how private their data are would help them to decide whether or not to disclose certain data items into software systems. From a developer's perspective they could decide which data to collect and which data to avoid when designing software systems in order to protect user privacy within the systems. For privacy researchers this information could be used when making measurements on user privacy among different software systems and provide guidance for law makers and authorities to enforce data protection regulations.

In this research we attempt to extract how users perceive different aspects of data that leads to their decisions in disclosing their data into software systems. Using a survey with 151 respondents we observe how users perceive the sensitivity of the data items to the user, relatedness of the data items to the purpose of an application and the visibility a particular data item gets in an application when they decide to disclose their data into software systems. With this knowledge we attempt to build a relationship relating how sensitivity, visibility and relatedness of data items relate to user privacy. We measure perceived privacy risk when users disclose their data items in different settings. Furthermore, through a qualitative analysis we attempt to identify other factors, if any, that could affect users data disclosure decisions.

This paper has two contributions.

- First, measure privacy of data disclosure decisions. A metric to measure privacy that takes into account the context in which it is disclosed (by considering the relatedness of the data item to the application context) would help software developers to make decisions about collecting, storing and using different data items in software systems.
- Second, understanding factors that affect users' data disclosure decisions. Understanding factors that affect users' decisions to disclose their data into software systems would help organizations to build positive relationships with users when they collect user data for different services and business purposes.

The paper is structured as follows. We first discuss prevoius work that has identified the parameters (sensitivity, visibility and relatedness) in measuring privacy in the background section. Then we describe our user study followed by the results. Next, we discuss the implicatins of our findings followed by the conclusion and future work.

II. Research Methodology

Our goal in this research is to understand how sensitivity, visibility and relatedness affect users' decision making when they disclose their personal data into software systems. For this, we conducted an online survey with 133 users. We designed the survey with 4 simple questions with Likert scales, to understand how users would feel if they are to disclose their personal data into different types of software applications. We used a list of 10 data items and asked users how they would feel disclosing those data items in four different application contexts. We used a five point Likert scale with values, very uncomfortable, somewhat uncomfortable, neutral, somewhat comfortable and very comfortable for users to express their feelings. Following these four questions we also used an open ended question to identify any factors other than what we investigated in the survey, that could affect users data disclosure decisions when they interact with software systems. 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.

The survey design was evaluated with two participants. 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. Before proceeding to the survey participants were given a brief introduction about the survey and the duration of the survey (under 10 minutes, calculated using the participants who evaluated the questionnaire). We also provided the participants with the contact details of the researchers. 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.

We measured the participant adequacy while collecting data and stopped data collection once we reached sample adequacy at KMO = 0.9. We then analyzed the data to obtain results.

A. Data Analysis

We assigned values from 1 to 5 for the answers we received on the Likert scale as follows.

TABLE I
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

The goal of our research was to calculate the relationship between sensitivity, relatedness and visbility of data when users make their disclosure decisions. The data we collected in the study was the feeling of disclosure of data by participants in different scenarios. From this we first calculate the

- Feeling for sensitivity - F_S,
- Feeling of visibility - F_V and the
- Feeling of relatedness - F_R

for each participant when they discose data into software applications. This calculation was done as follows.

$$\text{Feeling of Sensitivity}(F_S) = \frac{\text{Sharing a sensitive data item in an application}}{\text{Sharing a not so sensitive data item in the same application}} \tag{1}$$

We defined sensitive elements following the European Data Protection Regulation's definition for the data analysis. According to their definition we identified, user's credit card details, user's health information (subscribed medicine, blood type) as sensitive data and user's name as a not so sensitive data item.

Similarly,

$$\frac{\text{Sharing a data item in an application where the user cannot control the v}}{\text{sharing the same data item in an application where the user can co}} \tag{2}$$

$$\text{Feeling of Relatedness}(F_R) = \frac{\text{Sharing a related data item into an application}}{\text{sharing an unrelated data item in the same application}} \tag{3}$$

Then we obtain the average values for F_S, F_V and F_R across all participants to get the average feeling. Using the average values for the feelings we calculate the relationship between the parameters as follows,

$$Sensitivity\_vs\_Relatedness = \frac{\text{Feeling of Sensitivity}(F_S)}{\text{Feeling of Relatedness}(F_R)} \tag{4}$$

$$Sensitivity\_vs\_Visibility = \frac{\text{Feeling of Sensitivity}(F_S)}{\text{Feeling of Visibility}(F_V)} \tag{5}$$

$$Visibility\_a\_Relatedness = \frac{\text{Feeling of Visibility}(F_V)}{\text{Feeling of Relatedness}(F_R)} \tag{6}$$

III. RESULTS

We tested the validity of our results with Cronbach's alpha (0.91) and the participant adequacy for correlations with KMO (KMO = 0.82).

IV. DISCUSSION

V. LIMITATIONS

The study had several limitations. Firstly this was a remote survey. While remote surveys are beneficial in getting a large sample of participants in a relatively small period of time, direct engagement with users would disclose more in depth information when it comes to qualitative analysis. However, the main contribution of the paper was the derivation of the privacy measurement formula. Furthermore, for the qualitative component of data analysis we reached data saturation with the codes with 17 participants. The survey had 133 respondents which makes the results valid. Contextual studies involving real time applications that directly engage users in a natural setting could be used to observe how our findings stand in such setups.

VI. Conclusion and Future Work

References

[1] F. T. Commission *et al.*, “Fair information practice principles,” *last modified June*, vol. 25, 2007.

Appendix A
Appendix A Survey Questionnaire