Understanding users' privacy and security attitudes towards data breaches in Indian public and private online services

Aryan Gupta
Department of Chemical Engineering
IIT Kharagpur
Bokaro Steel City, Jharkhand, India
aryan2601@kgpian.iitkgp.ac.in

Apoorv Modak
Industrial and Systems Engineering
IIT Kharagpur
Pune, Maharashtra, India
apoorvmodak1@kgpian.iitkgp.ac.in

Naveen Sani
Industrial and Systems Engineering
IIT Kharagpur
Kothamangalam, Kerala, India
naveensanip@kgpian.iitkgp.ac.in

MOTIVATION

Data breaches can be far more than a temporary terror – they may change the course of your life. A data breach exposes confidential, sensitive, or protected information to an unauthorized person or a group of people. Anyone can be at risk of a data breach. Recently, the source code and data dumps of many companies have been breached. For example, Dominos India's data of 18 crore orders was breached which included name, mobile number, address, and any other personal information. Another example is of an Indian Brokerage Firm Upstox which suffered a Data Breach leaking 2.5 million users' data.

We were very curious to find out how sensitive the users' will be if this had happened to them. For this, we will be asking them questions about how comfortable they will be to share their personal data with services given that these services had data breaches in recent years. We also tried to find how this is dependent on various factors like age, gender, and other demographic variables, along with related financials of users having possible effects on their decisions. We wanted to notice how this behavior changes for the different users sharing their credentials and other important data across public and private services and does it affect their trust in these services anyhow.

RESEARCH QUESTIONS

We aim to identify a few basic correlations amongst the various demographic variables against the user's consciousness regarding their online presence and data.

• Correlation of consciousness about data shared with public and private services.

• Correlation between the financials of a user and their consciousness regarding their data

Why They Are Interesting

- People generally tend to underestimate the amount of data they have shared with public and private services. So we thought that it would be interesting to test it and see the results.
- We thought it would be interesting to know how the financials of a user relates to the consciousness of their data i.e. whether the amount of assets an individual owns relates to the amount of data they think they share.

Hypothesis

1) Regarding the correlation of the general demographics of users to their choices for the questions asked in the survey, the null hypothesis comes down to more or less similar responses for the questions irrespective of the demographic associated with the user. The alternate hypothesis would state a clear deviation

from these results and yield interesting insights on different demographics displaying a variety of decisions across platforms.

2) The correlation of users' financial association with their important credentials and how that changes across different variables involved. has the simple null hypothesis of the independence of such decisions from the financial aspect. Whereas the alternate hypothesis would conclude a rather skewed decision process for the various financial groups involved.

STUDY DESIGN

Survey

We ask our participants to answer a set of predefined questions based on basic demographics followed by one of the two interfaces as an example and then ask questions regarding their sensitivity towards data breach scenarios or simply their data being publicly available. The survey questions that we use for the purpose can be classified into 5 categories.

1) **General demographic questions (Q14-Q20):-** This part of the survey includes general informative questions which are asked in order to have a background about the participant as well as have a base for the first research question which connects the demographics of the user with their trustability towards the service. The questions that we have asked include the Name of the participant, age group and gender of the participant, educational qualification, and educational institute affiliation of the participant. The age group is divided in such a way to distinguish children, students, adult population, and aged users.
2) **Understanding regarding Privacy (Q1, Q4):**- Along with the general demographic questions, another question asking about the participant's previous knowledge regarding the subject of security and privacy of data has been added to account for the social desirability bias that can possibly be generated in their answers. This question combined with the additional questions on the data breach information vouch for the user's understanding of the crisis, and reduce the possibility of irrational answers.
3) **System Model (Q2, Q3, Q5-Q8):-** This part of the survey contains questions that draw the data regarding the actual usage of different public and private services by the participant (asked directly). This can be used as a rough estimate to determine the amount of monetary assets involved in the system model for a particular participant. (This can be calculated by adding the total sum invested in premium services by checking for the prices of each). This part of the survey also contains questions that help generate a system model of the data that the users consider sharable to the public and private services. The survey asks for the general acceptability of data sharing (financial/ personal) amongst the users. The section also includes a question that identifies a user based on their income group, to have a direct estimate of the financials at stake, in case of a data breach in paid online services such as banking.
4) **Trustability Model (Q9-Q13):-** This model includes questions that ask for the trust of users towards private and public services and the trustability of the users

is measured using Likert scale type input. Along with the trustability of the users, the section also asks for the expectations of the users from public and private online services in the case of a data breach. It also asks users to compare the credibility of the services towards data breach, judging based on whether it is a paid service or an open-source online service.

5) **Consent and Credibility of Answers:-** This part of the survey asks for the self declaration from the users that the responses provided in the survey are true to their belief and that they take responsibility for the credibility of the responses. Furthermore the section also asks consent from the participants, as an acceptance of the data they provided being used for the scope of this research. If the consent is not given by the participant, the response may not be used for the research.

Recruitment Method

We ask students and professors of many colleges to take part in our survey. Diversity is hard to achieve but will give quite a good statistic on the correlation of variables like age or gender regarding their online life. The survey that is being used to obtain the data is shared amongst various social media platforms like Facebook, Instagram, Linkedin, etc. and a direct approach with the Kharagpur community is also involved in the process.

STUDY INSTRUMENTS

We have created a survey questionnaire for the participating users. We have included demographic questions like age, gender, assets owned, and employment status. Then their question in the survey is to gauge the trust model of users in the existing systems.

We also started making the interface but realizing upon further analysis that the average Indian user is not that privacy-aware to check for the breach status of his credentials, it simply decreases the ecological validity of this research and the study.

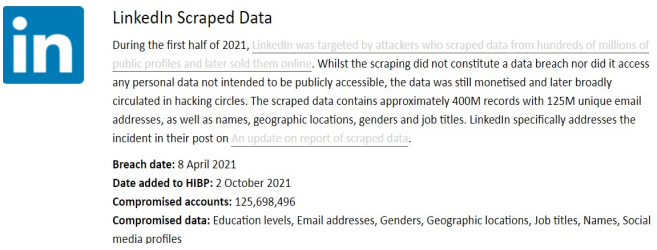

Figure 1: Interface after they press submit

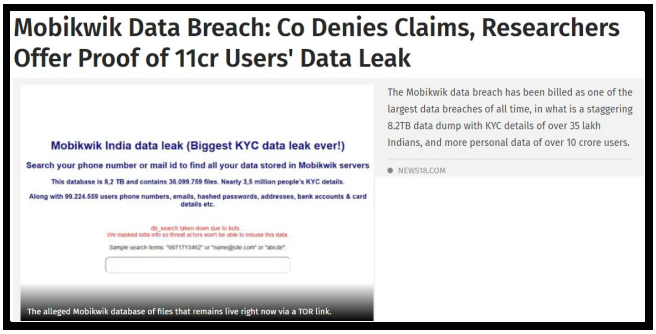

Figure 2: Interface showing news article regarding a private service

Figure 3: Interface showing a news article about a public service

ANALYSIS PLAN

We needed a technique that can be used for numerical response data and numerical or categorical input data. The one-way analysis of variance (One way ANOVA) is a useful test for the analysis of the hypothesis involved to check the possible correlation amongst the various variables involved.

To analyze the survey responses each demographic variable is compared with the total responses for each categorical question asked against the total number of responses. The mean, squared sum, and degrees of freedom are calculated. For example the degree of freedom for the Age variable = 4-1 =3. These values are calculated for within group and between group to calculate the F ratio. The significance of the alternate hypothesis comes from the Pr values of the statistic (i.e. that less than 0.005 for example) and that is used to obtain the significance of the factor questions on the demographics.

→ Example of applying one way ANOVA using R and interpreting the results.

```
one.way <- aov(yield ~ fertilizer, data = crop.data)
```

```
summary(one.way)
```

→ **Statistical values for the tests (example) and their interpretation**

```
Df Sum Sq Mean Sq F value Pr(>F)
  2   6.07  3.0340   7.863  7e-04 ***
 93  35.89  0.3859
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

JUSTIFICATION

The benefit of taking the survey is that we can gather data from a large number of people quickly and we can determine how prevalent an issue is. Our demographic survey questions such as assets owned help us to know the independent variables. The main research question helps us to know the level of comfort of people in sharing their data with a recently breached service. Combining these we will be able to answer our research questions. The effect of age or possibly gender on the thought process of users regarding their data will be concluded to some extent. The attached financial assets of users is another dimension other than general demographic correlations that we although feel to give a result most likely favoring high asset value leading to more privacy-conscious individuals but might lead to a different result depending on the recruited subjects. The survey analysis will yield many correlation results of the trust model factors with the users, which will help in understanding user mental model of services and their functioning with respect to the user, irrespective of the technicalities.

FUTURE SCOPE

This project has a lot of future scopes. We are living in a digital world right now where data has become the most powerful tool. As the years pass, the amount of data that is produced will keep on increasing. Hence there is an urgent need to know and make others aware of the data breaches and their harmful effects.

We can also ask more research questions like,

” Correlation between the educational qualification of the user and their consciousness regarding their data.”

” Correlation between the number of years of internet usage of the user and their consciousness regarding their data.”

DIVISION OF WORK

Aryan Gupta - Made the research questions, corrected the survey questions and the analysis of the possible tests for hypothesis testing and conclusions

Apoorv Modak – Made the research questions and survey questions and designed the system model and gauged the trustability model for users.

Naveen Sani – Wrote the report