# Behavioural Analysis of Susceptible Internet Users

Satyam Kumar, Akshay Pathak, D.Prabhu

*B.Tech Undergraduate Student, Computer Science Engineering, SRM University, Tamil Nadu , India*
*B.Tech Undergraduate Student, Computer Science Engineering, SRM University, Tamil Nadu , India*
*Assistant Professor, Computer Science Engineering, SRM University, Tamil Nadu, India*

## ABSTRACT

*The Internet has become an integral part of our everyday life. Unfortunately, not all of us are equally aware of the threats when we use online services. Naive users are generally less aware of security and privacy practices on the Internet and are susceptible to online predators. In this paper, we present a behavioral analysis of Internet users and their susceptibility online malpractices. We have considered the dataset from the Global Internet User Survey for 10789 respondents to perform a security-oriented statistical analysis of correlated user behavior. We constructed logistic regression models to analyze the statistical predictability of susceptible and not-so-susceptible identity theft victims based on their behavior and knowledge of security and privacy practices. We posit that such a study can be used to assess the vulnerability of Internet users and can hence be used to leverage institutional and personal safety on the Internet by promoting online security education, threat awareness, and guided Internet-safe behavior.*

**Keyword: -** *Safety of Internet Users, Susceptible Users, Linear Regression Model , User Model*

## 1. INTRODUCTION

Digital identities vary in different forms, ranging from credit card information of an individual to mere username/password pairs. Online services, such as, banking, bill payments, social media, job searches, and shopping involve the use of digital identities. Hence, in today's world, the security and value of the digital identity of an Internet user has a greater impact than it was a decade ago. Users have different personal behaviour and practices while accessing various Internet-enabled services and the knowledge of the Internet and cognizance of security threats is not equal among these users. As a result, flocks of phishers, spammers, and hackers are preying on these Internet users, based on their different practices and level of awareness. E-Crimes have significantly increased since the last few years, making it increasingly difficult for the authorities dealing with e-crimes. The primary targets are the naive users, who are unaware of online threats. Such online crimes include phishing, viruses, malware bots, social engineering, and privacy breaches, targeting identity thefts on the Internet. According to Moore et al, credit card information are sold at advertised prices of $0.40 to $20.00 per card, and bank account credentials at $10 to $100 per bank account. Social security numbers and other personal details are sold for $1 to $15perperson, while online auction credentials fetches around $1 to $8 per identity. As illustrated by Levchenko et al. The spam value chain has multiple links between the money handling authorities and the spammers.

The susceptibility of naive Internet users being victims of malware and viruses is not new .The proliferation of mobile devices have resulted in the increase of mobile malware. A 2013 study on mobile malware show that 99% of all mobile malware were built to target the Android mobile device platform with an encounter rate of 71% for online

malware . According to an approximate consensus, 5% of online devices on the Internet are susceptible to being infected with malware. At least 10 million personal computers have been assumed to be infected with malware in 2008, the number for which should have had increased significantly over the last few years. According to a study from May 2004, approximately 1.8 million Internet users were tricked by phishing websites into revealing private information study on the responses for the questions pertaining to the usage, behavior, and security practices of Internet users and the consequences (if any) of identity theft incidences.

## 2. DATA SOURCES

The Internet Society conducted the Global Internet User Survey in 2012 to collect reliable information relevant to users on the Internet. The survey was conducted via online panels of a total of 10789 respondents from 20 countries in their corresponding local languages. The survey included over 150 questions regarding their attitudes towards the Internet and their online behaviors. However, we will be focusing our study on the responses for the questions pertaining to the usage, behavior, and security practices of Internet users and the consequences (if any) of identity theft incidences.

## 3. USER STATIC ANALYSIS

### 3.1 Behavioral Features

We propose the following set of characteristics to define a naive Internet user, or naive Alice (AN).

1) Access: AN accesses the Internet in varying frequencies, ranging from many times a day to less than once a week.

2) UsageEM|SM|IC|IM|ST: AN uses email (EM), social media (SM), Internet audio/video conferencing (IC), instant messaging (IM), and/or media streaming (ST) services.

3) LoginEM|SM|IC|IM|ST: AN usually does not log in to use EM, SM, IC, IM, and/or ST services, as the password is saved on the browser and/or web application.

4) LogoutEM|SM|IC|IM|ST: AN does not always log out of EM, SM, IC, IM, and/or ST services after using it.

5) Anonymity: AN usually does not use or is not aware of anonymization services to protect her digital identity.

6) Web Browsing: AN usually does not pay attention to whether the visited websites are legal/authentic/secure, or does not even know how to identify a fraud/fake website

### 3.2 Multi-Conditional Probability Analysis

Next, will analyze the probabilistic correlated behavior for naive Alice (AN). We will refer to the Global Internet User Survey for some target questions. The responses are represented using 1, 2, ...n,where n is the number of options.

### 3.1 Access Frequency

AN may access the Internet in varying frequencies between 1 to 7, with 1 being the most frequent and7beingtheleast.Wesegmentedtheaccessfrequenciesand calculated the following probabilities: many or several times daily (0.8890), at least once daily or several times weekly

(0.1020), and once or less than once weekly (0.0089). The mean frequency was 1.59, which implies a higher probability of AN using the Internet many or several times daily.

### 3.2 Service Usage Frequency

We were interested to find the probable frequency of using Internet enabled services for AN. We calculated probabilities for at least n number of services, where services S ∈{EM,SM,IC,IM,ST}, for varying frequencies between 1 to 5, with 1 being the most frequent and 5 being the least. Figure 1a illustrates the probability distribution. We found that AN is more likely to use less number of services very frequently. However, it also shows that AN has a higher probability of using more services more frequently than more services less frequently. The MANOVA [8] test showed that the dependent service usage on access frequency had a Wilks' Lambda 0.698 and p-value<0.0001, thus establishing a rather strong relation. The access frequency also had p-value<0.0001 for each of the services.

### 3.3 Login Logout Practices

As shown in Figure 1b, the logging in probability of users was lesser with lower access and usage frequency. The behavior pattern does not change drastically for different services for each frequency. We found that frequent users are more cautious in not saving credentials for email and social media. Lesser probability for logging into streaming services also implies that users may prefer using the services anonymously. The Wilks' Lambda value was 0.832 with pvalue<0.0001 for the MANOVA test. Individual tests for the access frequency had p-value<0.0001, which asserts a strong correlation between the features. Figure 1c shows the probability distribution for users rarely or never logging out of services. We found that frequent users are more likely to log out of email and social media, but stay logged in for instant message and other services. We may, however, assume that the respondents replied 'never' even if they did not sign-in in the first place. There is also an increase in the trend for users not logging out with decreasing access and usage frequency.

### 3.4 Anonymity

Browser add-ons and routing protocols for anonymization are available for users on the Internet [12–15]. Anonymization services may not be considered as a prominent indicator of susceptibility. However, we posit that a user's awareness of anonymization services implies a greater knowledge of security concerns. Figure 1d illustrates the distribution for the given conditional probability. We observed that the probability of users not using any form of anonymization services increases with decreasing access and usage frequency. Moreover, the probability of even the most frequent users is still very high, at 0.8047, and the probability of not using anonymization services becomes 1.0 with least frequent users. Therefore, we posit that AN is not security-educated and has a high probability of not protecting her digital identity.

### 3.5 Web Browsing

Users must validate the legitimacy of websites (e.g. HTTPS, server certificate) to avoid phishing, especially when presenting classified information [16]. The probability distribution is illustrated in Figure 1e. The maximum probability for the most frequent users is at 0.5277 and remains fairly constant for all other services. Less frequent users tend to have higher probability of being reluctant to observe the legality of the visited websites and goes up to 1.0 for the least frequent users. Regardless of how regular the users are on the Internet, the numbers show a general lack of security awareness, and therefore, ensues an increasing number of e-crime victims.

## 3. PROPOSED SYSTEM

The proposed system provides the analysis of Susceptible internet users on basis of more classified features providing different activities performed by particular user.

1.Our proposed system uses the regularized algorithms which deals with precision of the analysis.

2.The features used in our analysis are more classified and to overcome the problem of overfitting we have used regularized algorithms.

3.In this problem we deal with more number of features and are implemented by using regularized logistic regression.

## 4. BEHAVIORAL MODELING OF INTERNET USERS

In this section, we investigate statistical models for predicting the susceptible AN versus the not-so-susceptible AN. As shown in Table I, y26 is our dependent variable to foretell the vulnerability of AN. The response can be used to identity the victims of identity thefts, including unwanted communications, losing personal data, loss of privacy, impersonation, and financial losses.

### 4.1 Nominal Logistic Models

The nominal logistic model was used to address nominal data used for the responses (1 to n∈[number of choices]). The dependent variable, y26, had three response classes, "Yes", "No", and "Don't know", which are labeled as 1, 2, and 3 respectively. In the models, we primarily focused on the classes 1 and 2 (Yes and No). The p-value is considered significant if lesser than 0.05.

### 4.2 Principal Component Analysis

We performed a Principal Component Analysis (PCA) on the set of variables ST, (x1 – x25) to determine the observations which are responsible for maximal data variance. Figure 2a and Figure 2b show the PCA plot for classes 1 and 2 respectively. It was seen that the cases of AN being a victim of identity theft ("Yes" responses) were determined by the set of variables SY ⊆ [x1, x2, x3, x4, x5, x6, x17, x18, x19, x20, x21, x22, x23, x24, x25]⊆ ST. The "No" responses were determined by the set of variables SN =[SY - x20]. However, the order of the effects varied for the two classes.

### 4.1 Singular Variables Model

We created a logistic regression model for the susceptibility of AN to identity thefts (y26) using all of the independent variables (x1 – x25).
The susceptibility of AN can be predicted with a probability of 49.88%. The contingency plot for the prediction model can be visualized , with the mean value falling in between classes 1 and 2, and 0.5301 standard deviation (StDev). However, ¯ AN, belonging to class 2, had a higher TP at 86.76% as well as 0.6978 ROC area cover.

### 4.1 Modified Logistic Model Fitting

Next, we created three more logistic models to investigate AN: a logistic Lasso regression model, a logistic stepwise selection regression model, and a multinomial Lasso regression model. However, unlike before, we populated the missing values using probabilistic imputed values. We observed that classes "No" (2) and "Don't know" (3) had similar patterns for y26. Hence, we merged classes 2 and 3 for y26 for the logistic Lasso regression and stepwise selection model. The summary of the modified models is presented in Table III. Logistic Lasso Regression: The Lasso model is helpful for predicting variables with missing values. We incorporated certain interactions based on our earlier PCA and iterative addition and subtraction of variables. The introduced interactions were (x3*x5*x6), (x17*x18), (x22*x23), (x22*x25), and (x23* x25). We specified a minimum cross-validation error threshold of 0.0001.

## 4. DISCUSSION

We utilized the Global Internet User Survey to analyze the susceptibility of naive Alice, AN, to identity thefts using security-oriented behavioral patterns.

Unfortunately, our models could identify AN with only a TP of 53.45%. After discarding the missing data, the minimal interaction model gave us the highest TP for identifying AN. The missing data is a limitation of our dataset. We may assume that given AN answers all questions regarding her behavior, the models will perform better in predicting AN. The prediction for ¯ AN improved using the modified logistic models. Both the logistic Lasso and stepwise selection regression performed better in identifying ¯AN. However, the evaluated cases had merged the "No" and "Don't know" response cases. This is not a strong assumption, as we believe that the users responding "Don't know" will, in reality, belong to either "Yes' or "No" classes. Therefore, the multinomial Lasso regression model may be considered the best for predicting ¯ AN.

## 5. CONCLUSION

The Internet has become a major target for online criminals to exploit naive users. In this paper, we have considered the Global Internet User Survey dataset to perform statistical tests and constructed 5 different models to analyze the behavioral features of susceptible naive users and their counter-class. The investigation revealed a moderately performing model for classifying naive users, but had a very high performance for modelling the not-so-susceptible users. We therefore suggest using logical double negation to ensure secure Internet practices using iterative reporting, monitoring, and security education. Our future work includes enhancement of the statistical analysis using learning-based algorithms to develop suggestive security frameworks.

## 6. REFERENCES

[1]. Coursera (https://www.coursera.org)
[2]. Analytics Vidhya (https://www.analyticsvidhya.com)
[3]. Kaggle (https://www.analyticsvidhya.com)