2016 IEEE 40th Annual Computer Software and Applications Conference

The Story of Naive Alice: Behavioral Analysis of Susceptible Internet Users

Rasib Khan and Ragib Hasan
SECRETLab, Dept. of CIS, University of Alabama at Birmingham, AL, USA
Email: {rasib, ragib}@cis.uab.edu

***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 to 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.**

***Keywords*-Behavioral Analysis, Naive Alice, Linear Regression Model, Susceptible User, User Model**

I. 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 behavior 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 [1]. 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. [2], 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 \$15 per person, while online auction credentials fetches around \$1 to \$8 per identity. As illustrated by Levchenko et al. [3], 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 [4]. 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 [5]. According to an approximate consensus, 5% of online devices on the Internet are susceptible to being infected with malware [2]. 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 [2]. According to a study from May 2004, approximately 1.8 million Internet users were tricked by phishing websites into revealing private information [6].

The reason for users being victims to e-crimes is primarily due to the lack of knowledge of security threats pertaining to their digital identities [1]. In this paper, we focus on the behavioral pattern of Internet users. We correlate the behavior of users with improper online practices with cases of identity frauds and other such incidence reports. The primary dataset used for our work has been obtained from the Global Internet User Survey 2012 [7]. We have also used data from the Bureau of Justice Statistics on Identity Theft Supplement (ITS) to the National Crime Victimization Survey [1] to depict the severity of the situation. Henceforth, we refer to Naive Alice (A_N) as a naive Internet user, and present a generic probabilistic model to illustrate the susceptibility of naive Alice to identity thefts.

Contributions: We presented a correlated analysis for target questionnaires from the Global Internet User Survey [7]. We created and evaluated five different linear regression models based on the interaction and characteristics of variables to analyze the behavioral class of susceptible and not-so-susceptible Internet users to identity thefts. Finally, we presented a discussion on the proposed models and their applicability to assess, educate, make aware, and monitor users based on their knowledge of security and privacy protection on the Internet.

Organization: The dataset and the target questionnaire is described in Section II. Section III presents the behavioral feature-set and the correlated statistical analysis. Section IV presents the statistical prediction models for naive Alice, A_N, and the counter-class, $\bar{A_N}$. A discussion on the models and user susceptibility is presented in Section V. We present the related work in Section VI, and conclude in Section VII.

II. Data Sources

The Internet Society conducted the Global Internet User Survey in 2012 to collect reliable information relevant to users on the Internet [7]. 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.

The survey questionnaire allowed the respondents to answer with predefined options. The survey results show that upto 80% of users are not aware of privacy policies on the Internet. The difference in sample sizes and stratification is assumed to introduce a margin of error between 3.10% and 4.38% with 95% confidence. The individual survey response statistics can

0730-3157/16 \$31.00 © 2016 IEEE
DOI 10.1109/COMPSAC.2016.206

Variable	Questions	**Response values**: 1 ... n ∈ [number of choices]
x1	On average, how often do you access the Internet?	Many times a day, Several times a day, Once a day, Several times a week, Once a week, Less than once a week, Dont' know
x2 - x6	How often, if at all, you use the following services? – Email, Social Media, Internet-based Conferencing, Instant Messaging, Streaming	At least once a day, Several times a week, Once a week, A few times a month, Less often than once a month, Don't use this
x7 - x11	Which of these services do you log in to use? – Email, Social Media, Internet-based Conferencing, Instant Messaging, Streaming	Yes, No, Never used it
x12 - x16	How often do you log-out of the following services? – Email, Social Media, Internet-based Conferencing, Instant Messaging, Streaming	Always, Often, Sometimes, Rarely, Never
x17	Do you use anonymization services?	Yes, No, Don't know/Not Aware, Would like to but don't know how
x18	I do pay attention to whether websites I visit are legal	Strongly agree, Somewhat agree, Somewhat disagree, Strongly disagree, Don't know
x19	When you are logged in to a service or application do you use privacy protections?	All of the time, Most of the time, Sometimes, Never
x20 - x25	Which of these do you do or have you done to protect your privacy? – Restricted use of location data by Websites or apps, Set sharing permission for friends and family only, Used a separate password for sensitive data, Provided incorrect data (fake name, date of birth, etc) when creating a new account, Downloaded a web browser plug-in, Reused throw-away password for low-value accounts	All of the time, Most of the time, Sometimes, Never
y26	Have you ever disclosed personal information online that was later used in a way you didn't expect? (unsolicited communications, stolen personal data, private data becoming public, impersonation, financial loss)	Yes, No, Don't know

TABLE I: Summary of Target Behavioral Questionnaire for the Global Internet User Survey 2012

be found online at [7]. We have selected a total of 26 questions related to security practices, summarized in Table I.

III. User Statistic Analysis

In this section, we present the statistical analysis for naive Alice, based on a set of behavioral features.

A. Behavioral Features

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

1) **Access**: A_N accesses the Internet in varying frequencies, ranging from many times a day to less than once a week.
2) **Usage**$_{EM|SM|IC|IM|ST}$: A_N uses email (EM), social media (SM), Internet audio/video conferencing (IC), instant messaging (IM), and/or media streaming (ST) services.
3) **Login**$_{EM|SM|IC|IM|ST}$: A_N 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) **Logout**$_{EM|SM|IC|IM|ST}$: A_N does not always log out of EM, SM, IC, IM, and/or ST services after using it.
5) **Anonymity**: A_N usually does not use or is not aware of anonymization services to protect her digital identity.
6) **Web Browsing**: A_N 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.
7) **Password**$_{SEP|RE}$: A_N does not always use a separate password (SEP) for sensitive data, and reuses the same password (RE) for different random low-value services.
8) **Identity**: A_N does not always provide incorrect data (fake name, date of birth, address) while registering with online services for the sake of protecting her identity.
9) **Privacy**$_{LOC|SHR}$: A_N does not always use privacy protection settings for restricting location data (LOC) and/or shared information (SHR) when using online services.

B. Multi-Conditional Probability Analysis

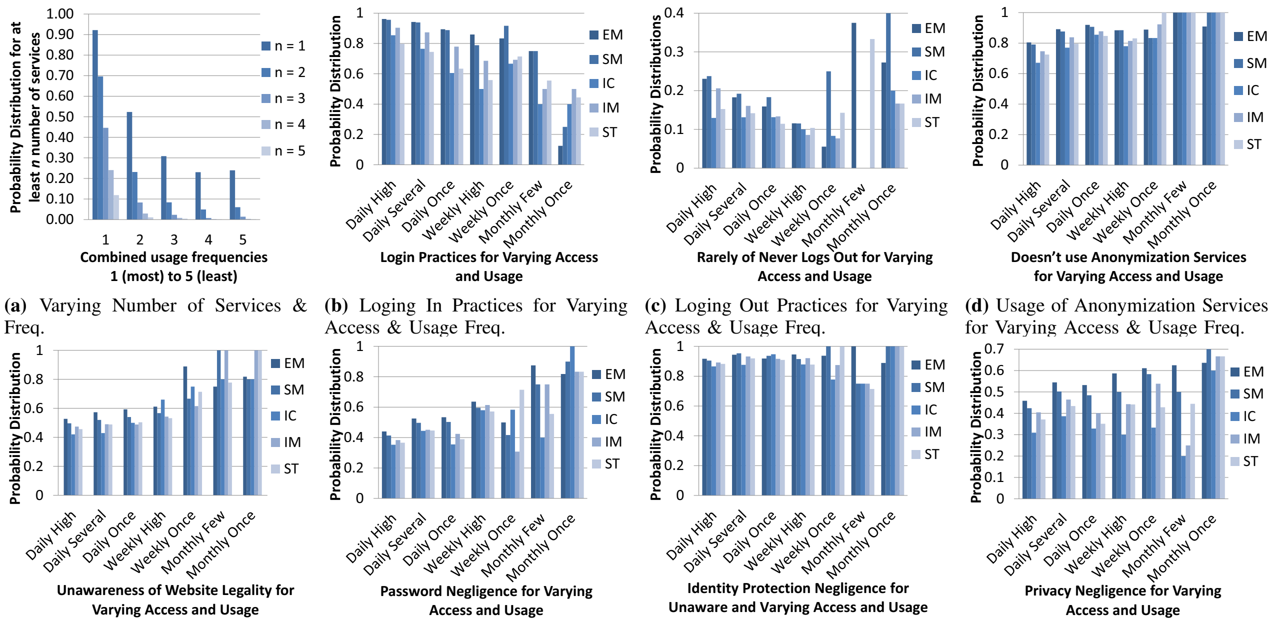
Next, will analyze the probabilistic correlated behavior for naive Alice (A_N). We will refer to the Global Internet User Survey [7] for some target questions. The responses are represented using 1, 2, ... n, where n is the number of options.

Access Frequency: A_N may access the Internet in varying frequencies between 1 to 7, with 1 being the most frequent and 7 being the least. We segmented the access frequencies and 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 A_N using the Internet many or several times daily.

Service Usage Frequency: We were interested to find the probable frequency of using Internet enabled services for A_N. We calculated probabilities for at least n number of services, where services $S \in \{EM, SM, IC, IM, ST\}$, 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 A_N is more likely to use less number of services very frequently. However, it also shows that A_N 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.

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 p-value<**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.

Flexible authentication mechanisms using identity federations and single-sign-on features improves the seamless service access for the users [9, 10]. Unfortunately, online attackers can exploit the weakest link to gain access to one of the services and gain access to a landslide of accounts for a user [11]. Users not logging in partially implies that the credentials are saved on their web-applications. Users not logging out mostly implies that they remain logged in from their devices.



(a) Varying Number of Services & Freq.

(b) Loging In Practices for Varying Access & Usage Freq.

(c) Loging Out Practices for Varying Access & Usage Freq.

(d) Usage of Anonymization Services for Varying Access & Usage Freq.

(e) Unawareness of Website Legality for Varying Access & Usage Freq.

(f) Password Negligence for Varying Access & Usage Freq.

(g) Identity Negligence Given True for Legality Unaware & Varying Access & Usage Freq.

(h) Privacy Negligence for Varying Access & Usage Freq.

Fig. 1: Probability Distributions for Multi-Conditional User Behavior

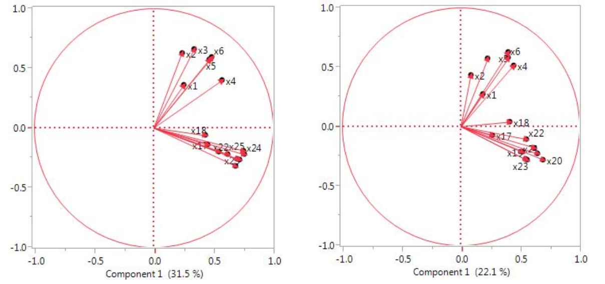
The security of systems are as secure as their weakest link. Therefore, the lowest conditional probability of logging in for users with access and usage frequency of at least once a day or more is **0.6053**. This connotes that the users are at **0.3947** probable risk of losing their credentials. Similarly, the highest probability is at **0.2307** for the set of users with the same access and usage frequency, resulting in the probability of losing credentials at **0.7693**.

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 A_N is not security-educated and has a high probability of not protecting her digital identity.

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.

Password Usage: Improper use of passwords allows online criminals to exploit users [17, 18]. Conversely, not reusing throw-away passwords for low-value online services indicates that users are more likely using the same passwords as their high-value accounts [19]. An average user uses 8 different passwords in a day, which is much lower than the number of websites visited, implying usage of similar passwords across multiple accounts. The average strength of passwords for is well-below the notion of strong passwords in online security [19]. The probabilities for negligent password usage is illustrated in Figure 1f. The highest probability of password negligence is **0.4403** for high frequency access and usage. The probability of the risks increase to **0.6363** for users with weekly high access/usage and goes up to **1.0** for the least frequent users. The uniformity of negligent behavior of the users is observed for most classes of users. Users who are not always using different passwords and not reusing low-value passwords are the most vulnerable in terms of password theft.

Identity Protection: Negligence towards personal information is one of the critical factors triggering the number of online identity thefts [18]. Figure 1g illustrates the probability distribution for negligent personal information usage for unaware web browsing. We assume the responses were inclusive of instances for presenting personal information during account registration. However, the calculations show an alarmingly high probability distribution. The most frequent Internet users, given that they do not verify website legality, have a probability of **0.9170**. Additionally, almost all other access and usage frequencies have probabilities between **0.8** and **1.0**.



(a) Principal Components for Class 1 **(b)** Principal Components for Class 2

Fig. 2: Principal Components for Classes *"Yes"* (1) and *"No"* (2) for y26

Content Privacy: Social networks are gaining popularity without users gaining proper knowledge of content sharing and privacy [20–23]. Users are unaware of the circulation/accessibility of these data that they are sharing. The probability distribution is illustrated in Figure 1h. Frequent users have a probability of **0.4586** for being negligent towards maintaining the privacy of shared content, with a gradual increase in the probability for lesser frequent users. Users who are at least regular on a weekly basis have a probability of negligence between **0.6111** and **0.7**. As a result, online criminals can easily exploit the privacy of naive users on the Internet.

IV. Behavioral Modeling of Internet Users

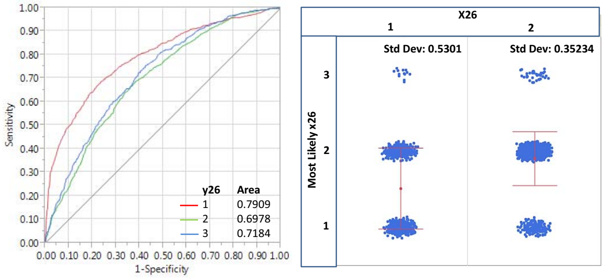
In this section, we investigate statistical models for predicting the susceptible A_N versus the not-so-susceptible $\bar{A_N}$. As shown in Table I, *y26* is our dependent variable to foretell the vulnerability of A_N. The response can be used to identity the victims of identity thefts, including unwanted communications, losing personal data, loss of privacy, impersonation, and financial losses. The statistical analysis was performed using SPSS [24], R [25], and JMP [26].

A. Nominal Logistic Models

The nominal logistic model was used to address nominal data used for the responses (1 to $n \in$ [*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.

B. Principal Component Analysis

We performed a Principal Component Analysis (PCA) on the set of variables S_T, (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 A_N being a victim of identity theft (*"Yes"* responses) were determined by the set of variables $S_Y \subseteq$ [*x1, x2, x3, x4, x5, x6, x17, x18, x19, x20, x21, x22, x23, x24, x25*] $\subseteq S_T$. The *"No"* responses were determined by the set of variables S_N = [S_Y - x20]. However, the order of the effects varied for the two classes. For example, x18 affects the responses in different directions and magnitude. The PCA therefore helped us identify the strong factors for analyzing A_N's behavior to construct the regression models.



(a) ROC **(b)** Model Contingency

Fig. 3: ROC and Model Contingency Plot for Singular Variables Model

Model	**Significant Predictors** (p-value)	**ROC** (1,2,3)	**TP Rate** (1,2)
Singular Variables	**x3** (0.0196), **x17** (<0.0001), **x18** (0.0135), **x19** (0.0169), **x20** (0.0012), **x23** (<0.0001), **x25** (<0.0001)	0.7909, 0.6978, 0.7184	49.88%, 86.76%
Minimal Interaction	**x5** (0.0288), **x17** (<0.001), **x19** (0.0327), **x20** (0.0027), **x23** (0.0007), **x25** (0.0001)	0.8199, 0.7418, 0.7732	53.45%, 85.27%

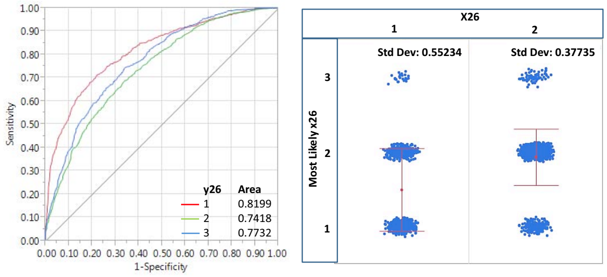
TABLE II: Summary of Nominal Logistic Models for Naive Alice

C. Singular Variables Model

We created a logistic regression model for the susceptibility of A_N to identity thefts (y26) using all of the independent variables (x1 – x25), summarized in Table II. We only considered the cases where all the questions (x1 – x25) had their response values. The gradient converged in 15 iterations with seven primary predictors. The receiver operating characteristics (ROC) graph of a model illustrates the relation between sensitivity and specificity for the predictions and is measured using the area bounded by the curve. The ROC for the model is illustrated in Figure 3a. The users who were and were not victims of identity thefts can be modeled using the significant classes with an ROC area of **0.7909** and **0.6978** respectively. The true positive (TP) rate for the corresponding classes are also mentioned in Table II. The susceptibility of A_N can be predicted with a probability of **49.88%**. The contingency plot for the prediction model can be visualized in Figure 3b, with the mean value falling in between classes 1 and 2, and **0.5301** standard deviation (StDev). However, $\bar{A_N}$, belonging to class 2, had a higher TP at **86.76%** as well as **0.6978** ROC area cover. The predicted model contingency is also shown in Figure 3b, where the mean lies within the cluster of corresponding instances with StDev **0.3523**.

D. Minimal Interaction Model

Next, we created a nominal logistic model based on singular variables (x1 – x25) and minimal interaction variables with non-null values. The interactions considered were (x3*x4*x6), (x17*x18), and (x22*x23*x25), and were chosen based on the PCA and the variance of their individual effects. The interaction groups were based on: (a) similar effects in the PCA, and (b) iterative adding and subtracting of interactions. The interactions implied that A_N is likely to have correlated behavioral pattern among some particular service types, and is not aware of privacy preservation techniques for anonymity, web browsing, and identity protection. The logistic model converged in 18 iterations and is summarized in Table II. The ROC plot for the model is illustrated in Figure 4a. The primary predictors changed their accuracy for the model with the new interaction variables. The ROC area have also expanded for the three classes, *"Yes"* (1), *"No"* (2), *"Don't know"* (3), and were **0.8199**, **0.7418**, and **0.7732** respectively, and the TP rate for identifying A_N increased to **53.45%**. This implies



(a) ROC **(b)** Model Contingency

Fig. 4: ROC and Model Contingency Plot for Minimal Interaction Model

Model	**Condition**	**TP** (Class-1)	**TN** (Class 2)
Logistic Lasso Regression	Minimum cross-validation error	20.8191%	98.2346% *(2 + 3)*
Logistic Stepwise Selection Regression	Minimum BIC criteria	22.9644%	97.2762% *(2 + 3)*
Multinomial Lasso Regression	Minimum cross validation error	24.1346%	96.3914%

TABLE III: Summary of Modified Logistic Model Fitting for Naive Alice showing Condition, True Positive (TP), and True Negative (TN)

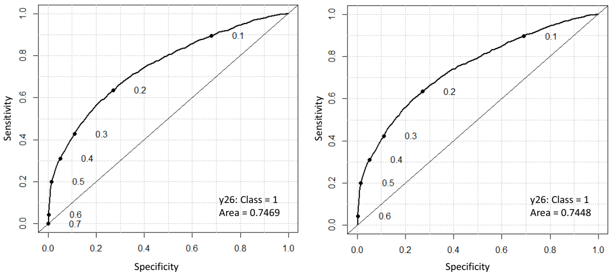
that the behavior of A_N can now be identified with a greater probability than 50%. The TP for $\bar{A}_N$ (*"No"*) remained fairly constant at **85.27%**. Figure 4b shows the contingency plot for the given model. As seen in the figure, the *"Yes"* instances fall closer to the predicted *"Yes"* instances, with **0.5523** StDev. The *"No"* instances remained fairly similar, with the mean within the corresponding cluster and StDev **0.3773**.

E. Modified Logistic Model Fitting

Next, we created three more logistic models to investigate A_N: 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**. As mentioned earlier, we merged the *"No"* and *"Don't know"* responses. The given model was able to predict A_N with a poor TP rate of **20.8191%** and **0.7469** ROC shown in Figure 5a. However, we observed a very high TP rate of **98.2346%** for the *"No"* responses. This implies that we were very successfully able to identify the $\bar{A}_N$ users.

Logistic Stepwise Selection Regression: The model was generated using 1000 iterations and a minimum Bayesian information criterion (BIC) for selecting the best features. The best predictors were x8, x9, x12, x17, x23, x25, with a minimized Akaike information criterion (AIC) of **9358.35**. The stepwise selection was able to improve the TP rate to **22.9644%** for predicting A_N, with an ROC area of **0.7448**, as shown in Figure 5b. The TP for $\bar{A}_N$ reduced from the earlier model to **97.2762%**.



(a) Logistic Lasso Regression **(b)** Logistic Stepwise Regression

Fig. 5: ROC for *"Yes"* (1) Responses in Modified Logistic Models

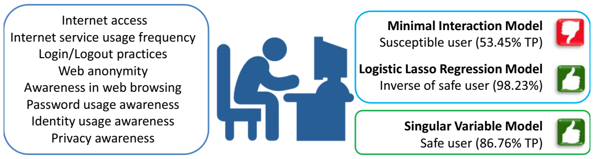

Internet access
Internet service usage frequency
Login/Logout practices
Web anonymity
Awareness in web browsing
Password usage awareness
Identity usage awareness
Privacy awareness
Minimal Interaction Model
Susceptible user (53.45% TP)
Logistic Lasso Regression Model
Inverse of safe user (98.23%)
Singular Variable Model
Safe user (86.76% TP)

Fig. 6: Infographic for Naive Alice User Model

Multinomial Lasso Regression: We included a minimum cross-validation error threshold at **0.0001** and (x3*x5*x6), (x17*x18), (x22*x23), (x22*x25), and (x23* x25) interactions for three separate classes. The separation of the classes allowed us to increase the TP for A_N based on the *"Yes"* responses upto **24.1346%**. Conversely, the TP prediction for $\bar{A}_N$ based on the *"No"* responses decreased to **96.3914%**.

V. Discussion

We utilized the Global Internet User Survey [7] to analyze the susceptibility of naive Alice, A_N, to identity thefts using security-oriented behavioral patterns. A descriptive infographic is illustrated in Figure 6. Unfortunately, our models could identify A_N with only a TP of **53.45%**. After discarding the missing data, the *minimal interaction model* gave us the highest TP for identifying A_N. The missing data is a limitation of our dataset. We may assume that given A_N answers all questions regarding her behavior, the models will perform better in predicting A_N. The prediction for $\bar{A}_N$ improved using the modified logistic models. Both the logistic Lasso and stepwise selection regression performed better in identifying $\bar{A}_N$. 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 $\bar{A}_N$.

Identity protection is a major concern in fighting against e-crimes. A model for A_N and $\bar{A}_N$ can be valuable in various contexts. Employees can be evaluated for work place Internet usage safety and can be trained accordingly. Users may undertake targeted surveys to evaluate their vulnerability and obtain an online safety score. An overview of such a model is illustrated in Figure 7. A user can be asked to take security surveys, which will be analyzed by a security monitoring server. Subsequently, the system can introduce appropriate access control for the user while connecting to the Internet. The security monitor can also place behavioral monitors for reporting users' actions, prompt education materials to the

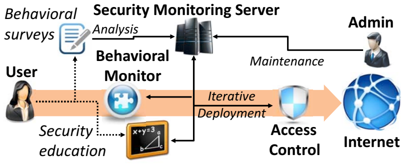


Fig. 7: A Learning Framework for Secure Internet Practices

users, and iteratively revise the rigidity of access control. This guided behavior will also allow A_N to become aware of online security practices. Given that we were able to predict $\bar{A_N}$ with a higher probability, the logical double negation, that is, $\bar{\bar{A_N}}$, should be considered to be the same as A_N, a susceptible user.

VI. Related Work

Stanton et al. [27] presents a study on 110 interviewees and evaluated their intentions of information misuse and password-oriented practices based on a two-factor taxonomy for classification. Florencio et al. [19] conducted a large-scale survey on the different password-oriented habits of Internet users. Hull et al. [28] has analyzed the contextual privacy issues on Facebook and the way social media effects privacy issues. User behavior regarding disclosing the identity on micro-blogs and the relative factors have been studied by Lee et al. [29]. Riek et al. [30] have studied a pan-European sample to investigate the effects of cybercrimes and the perceived awareness based on victimization and media reports. E. Litt [31] features a study on the current complexities in measuring users' Internet skills and why the evaluation should evolve over time for the survey methodology. Wagner et al. [32] presents an interesting work on malware infected Twitter users and their actions. According to most studies, the primary factors influencing the behavior of users on the Internet are age, education, gender, technology experience, content creation and sharing, online activities, income group, amount of leisure time, and the type of job. Preferential anonymity on the Internet was studied by Kang et al. [33]. Unlike most works so far, we faced a high variation in the data and the behavioral aspect of Internet users, and showed how they can be used to leverage the analysis for susceptible users. User model ontology, such as GUMO [34], can be helpful in describing the feature space of security oriented behavior to protect the users based on model-based intelligent agents while using the Internet [35].

VII. 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 [7] 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 modeling 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.

Acknowledgment

This research was supported by the National Science Foundation CAREER Award CNS-1351038.

References

[1] Bureau of Justice Statistics, "Identity Theft Supplement (ITS) to the National Crime Victimization Survey," Online at http://www.bjs.gov/content/pub/pdf/vit12.pdf, 2012.

[2] T. Moore, R. Clayton, and R. Anderson, "The economics of online crime," *The Journal of Economic Perspectives*, vol. 23, no. 3, pp. 3–20, 2009.

[3] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, N. Weaver, V. Paxson, G. M. Voelker, and S. Savage, "Click trajectories: End-to-end analysis of the spam value chain," in *Proceedings of the S & P*. IEEE, 2011, pp. 431–446.

[4] J. K. Reynolds, "RFC1135: The Helminthiasis of the Internet," Online at http://tools.ietf.org/html/rfc1135, Dec 1989.

[5] CISCO Security Intelligence Operations, "Annual security report," CISCO, Tech. Rep., 2014.

[6] S. Granneman, "Phishing for savvy users," Security Focus, Online at http://www.securityfocus.com/columnists/274, Nov 2004.

[7] Internet Society, "Global Internet User Survey 2012," Online at https://www.internetsociety.org/internet/global-internet-user-survey-2012, 2012.

[8] C. J. Huberty and S. Olejnik, *Applied MANOVA and discriminant analysis*. John Wiley & Sons, 2006, vol. 498.

[9] T. S. Dare, E. B. Ek, and G. L. Luckenbaugh, "Method and system for authenticating users to multiple computer servers via a single sign-on," Nov 1997, US Patent 5,684,950.

[10] S. Shim, G. Bhalla, and V. Pendyala, "Federated identity management," *Computer*, vol. 38, no. 12, pp. 120–122, Dec 2005.

[11] Y.-Y. Chan, "Weakest link attack on single sign-on and its case in saml v2. 0 web sso," in *Computational Science and Its Applications*. Springer, 2006, pp. 507–516.

[12] L. F. Cranor, "Internet privacy," *Commun. of the ACM*, vol. 42, no. 2, pp. 28–38, Feb 1999.

[13] O. Berthold, H. Federrath, and S. Köpsell, "Web mixes: A system for anonymous and unobservable internet access," in *Designing Privacy Enhancing Technologies*. Springer, 2001, pp. 115–129.

[14] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," DTIC Document, Tech. Rep., 2004.

[15] T. Sochor, "Anonymization of web client traffic efficiency study," in *Computer Networks*, ser. Communications in Computer and Information Science. Springer, 2012, vol. 291, pp. 237–246.

[16] F. Toolan and J. Carthy, "Feature selection for spam and phishing detection," in *Proc. of the eCrime*. IEEE, Oct 2010.

[17] E. Hayashi and J. Hong, "A diary study of password usage in daily life," in *Proc. of the SIGCHI*. ACM, 2011.

[18] R. Khan, M. Mizan, R. Hasan, and A. Sprague, "Hot zone identification: Analyzing effects of data sampling on spam clustering," *Journal of Digital Forensics, Security and Law*, vol. 9, no. 1, pp. 67–82, 2014.

[19] D. Florencio and C. Herley, "A large-scale study of web password habits," in *Proc. of the WWW*. ACM, 2007.

[20] A. Lenhart, "Adults and social network websites," *Pew Research Internet Project*, Jan 2009, online at http://www.pewinternet.org/2009/01/14/adults-and-social-network-websites/.

[21] Privacy Rights Clearinghouse, "Social networking privacy: How to be safe, secure and social," Fact Sheet 35: Online at https://www.privacyrights.org/social-networking-privacy-how-be-safe-secure-and-social, Aug 2014.

[22] S. B. Barnes, "A privacy paradox: Social networking in the united states," *First Monday*, vol. 11, no. 9, 2006.

[23] J. McDermott, "Foursquare selling its location data through ad targeting firm turn," Online at http://adage.com/article/digital/foursquare-selling-data-ad-targeting-firm-turn/243398/, July 2013.

[24] IBM Corp, *IBM SPSS Statistics for Windows v 22.0, Armonk, NY: IBM Corp*, Released 2013.

[25] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org

[26] SAS Institute Inc., *JMP, Version 11.2.0*, Cary, NC, 1989-2007.

[27] J. M. Stanton, K. R. Stam, P. Mastrangelo, and J. Jolton, "Analysis of end user security behaviors," *Computer Security*, vol. 24, no. 2, pp. 124–133, Mar. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.cose.2004.07.001

[28] G. Hull, H. R. Lipford, and C. Latulipe, "Contextual gaps: Privacy issues on facebook," *Ethics and information technology*, vol. 13, no. 4, pp. 289–302, 2011.

[29] S. Lee, Y. Kim, and B. G. Lee, "Determinants of voluntary self-disclosure in the usage of micro-blog," in *Proc. of the ICONI*. KSII, Dec 2010.

[30] M. Riek, R. Böhme, and T. Moore, "Understanding the influence of cybercrime risk on the e-service adoption of european internet users," in *Proc. of the WEIS*, Jun 2014.

[31] E. Litt, "Measuring users' internet skills: A review of past assessments and a look toward the future," *New Media & Society*, vol. 15, no. 4, pp. 612–630, 2013.

[32] C. Wagner, S. Mitter, C. Körner, and M. Strohmaier, "When social bots attack: Modeling susceptibility of users in online social networks," in *Proc. of the MSM*. Citeseer, 2012, p. 2.

[33] R. Kang, S. Brown, and S. Kiesler, "Why do people seek anonymity on the internet?: Informing policy and design," in *Proc. of the SIGCHI*. ACM, 2013.

[34] D. Heckmann, T. Schwartz, B. Brandherm, M. Schmitz, and M. von Wilamowitz-Moellendorff, "Gumo–the general user model ontology," in *User Modeling 2005*. Springer, 2005, pp. 428–432.

[35] F. A. Asnicar and C. Tasso, "ifWeb: a prototype of user model-based intelligent agent for document filtering and navigation in the world wide web," in *Proc. of the UM*, Jun 1997.