Risk prediction of malware victimization based on user behavior

Fanny Lalonde Lévesque
Ecole Polytechnique de Montréal
Montréal, Canada
fanny.lalonde-levesque@polymtl.ca

José M. Fernandez
Ecole Polytechnique de Montréal
Montréal, Canada
jose.fernandez@polymtl.ca

Anil Somayaji
Carleton University
Ottawa, Canada
soma@scs.carleton.ca

Abstract

Understanding what types of users and usage are more conducive to malware infections is crucial if we want to establish adequate strategies for dealing and mitigating the effects of computer crime in its various forms. Real-usage data is therefore essential to make better evidence-based decisions that will improve users' security. To this end, we performed a 4-month field study with 50 subjects and collected real-usage data by monitoring possible infections and gathering data on user behavior. In this paper, we present a first attempt at predicting risk of malware victimization based on user behavior. Using neural networks we developed a predictive model that has an accuracy of up to 80% at predicting user's likelihood of being infected.

1. Introduction

Risk assessment is a universal problem. Whether choosing where to walk, what to eat, or who to love, we make decisions based in part on an ongoing assessment of risk. As more of our lives are affected by our actions online, we all find ourselves making risk assessments as part of choosing what we do online.

Yet as individuals, it is becoming increasingly hard to know when we are at risk online. Threats such as drive-by downloads and spearphishing emails can fool even the most sophisticated of users. Nevertheless, we do know that there is a relationship between user choice and online risk. For example, studies of web browsing behavior have shown that malware exposure can be predicted based upon the types of websites that a user visits [2].

While such risk assessment may be useful for vendors of anti-malware products, in most other contexts risk must be assessed using much coarser data that has fewer privacy implications. For example, automobile insurance rates—which are directly proportional to the predicted risk for an individual—are largely based upon a very small number of factors such as age, gender, and automobile type. While finer-grained data (such as continuous location tracking) can be used to make more precise risk models, use of such data brings up a number of privacy concerns and other ethical issues [9]. Similarly, there is potential utility in assessing the relative risks of individuals for education, marketing, and insurance purposes based upon coarse non-invasive information such as user demographics and high-level behavior information.

In this work we report on a first attempt to predict user risk using user demographics and characteristics combined with online user behavior. The data for our work comes from our previous work on a first clinical trial of antivirus software [4, 5, 7], a study of fifty users over a four month period that gathered data using specially instrumented computers and multiple in-person interviews. We have previously reported a correlation between user demographics and risk that we observed in this study [7]; here we use the study data to build and evaluate a neural network-based model of user risk.

While our sample size of fifty users is small relative to larger scale studies of web [2] and network [3] traffic, such studies cannot be used to study the relationship between demographics and online risk because they have no reliable way of connecting user behavior with computer activity. Indeed, they cannot even reliably determine how many users they are studying, as IP addresses, accounts, and computers are only partially correlated with distinct users.

As we show here, computer security clinical trials [12], with their clear connections between user reported information, observed user behavior, and system behavior, provide

978-1-4799-7329-3/14/$31.00 ©2014 IEEE

just the right kind of data needed for building and evaluating models of user risk based upon user demographics and online behavior. Our key contribution here, then, is a first published user risk model based upon user demographics and online behavior that is validated using records of actual user exposures to malware. We believe this work is a key step towards validated models of demographics and online user risk that could be used for insurance, education, and other purposes.

The remainder of the paper is organized as follows. Section 2 presents related work in the identification of risk factors related to malware victimization; Section 3 discusses work related to risk prediction of malware infection. We review in Section 4 the field study we performed. In Section 5, we describe and evaluate our neural-network based model of risk. Section 6 discusses the limitations and implications of our work. Section 7 concludes.

2. Risk factors of malware victimization

While numerous studies aimed to evaluate risk factors related to various cyber threats, few of them focused on the risk of malware victimization. Among those, Milne *et al.*. [10] conducted a national online survey of 449 non-student respondents. They found that gender, age and number of hours spent online, excluding email, have a significant impact on users' likelihood to adopt risky behaviors. They concluded that male, younger users and users who spend many hours online were more at risk. Ngo *et al.* [11] applied the general theory of crime and lifestyle/routine to identify risk factors on seven types of cybercrime victimization. They used an online self-report survey on 295 students in the U.S. and found that age is a significant predictor of computer virus victimization, where younger users are most at risk. Although these studies have identified potential risk factors of malware victimization, their results are subjects to limitations as they are based on self-reported infection rates rather than upon actual detections of malware.

Microsoft used telemetry data from Microsoft Software Removal Tool (MSRT) to identify technical risk factors related to malware infections [1]. They found that users who are not using any AV product, or who are using a non up-to-date AV, expired AV or snoozed AV, are 5.6 times more at risk of getting infected than users running an up-to-date protection. Symantec analyzed emails collected by the Symantec.cloud mail scanning service to identify putative risk factors that are associated with individuals subjected to targeted attacks [8]. They focused their analysis using emails from academic researchers and applied epidemiological techniques to evaluate the risk associated with certain work domains. They identified work area that are at a statistically significant increased risk of being subjected to targeted attacks, supporting the hypothesis that it is an individual's area of expertise that leads them to become of interest to attackers. In another study, Canali *et al.* used telemetry data from Symantec on URLs visited by more than 100 000 users during a period of three months to predict their risk of visiting a malicious Web site [2]. Their results confirmed that the more a user surfs the Internet, the more he is at risk of encountering a malicious page. They also found that the category of Web site does not seem to matter much, with exception of adult and pornography categories. Using traffic of a large set of real ADSL customers, Carlinet *et al.* [3] build profiles of network usage to identify users who are more at risk of generating malicious traffic. They shown evidence that the type of operating system (OS) and using the Web a lot for chatting and streaming was also a risk factor.

3. User-based risk prediction

User-based risk prediction is widely used, specifically by bank institutions and insurance companies. While one will try to predict user's credit score the other will try to predict his risk of having an accident. For example, an 18 year-old man driving a new sport car might pay higher insurance fees than a 40 year-old mother driving a minivan. These companies have been able to identify risk factors based on historical data and they are now able to predict customers' risk.

One way to predict risk is by using data mining techniques, like neural networks, machine learning, random forest tree, decisions tree, etc. Even though these techniques have proven their efficiency in risk prediction, to the best of our knowledge, only the Canali *et al.* study [2] applied data mining in computer security to predict users' likelihood of visiting a malicious Web site. Using logistic regression, they developed a predictive model based on 74 features related to web browser usage and achieved a performance of up to 87%. While our approach is similar, one of our main contribution is to combine both social-demographic and behavioral factors to predict the risk of malware victimization based on real-usage data.

4. Study description

We describe in this section the design of the study that provided the data modeled in Section 5. For a more detailed description of our study design and methodology, see [4, 6]; for our original data analysis, see [5, 7].

The 4-month study we conducted involved 50 subjects whose laptops were instrumented to monitor possible infections and gather data on user behaviour. We monitored real-world computer usage through diagnostics and logging tools, monthly interviews and questionnaires, and in-depth investigation of any potential infections. By conducting this first study, we wanted to: 1) develop and test the validity and viability of a new methodology for the evaluation of security products; 2) determine how malware infects computer systems and characterize sources of malware infections; and 3) determine how factors such as the configuration of the system, the environment in which the system is used, and user behavior affect the probability of infection of a system.

The 50 participants were recruited through posters and newspaper advertisements on the Université de Montréal campus where the École Polytechnique is located. A short on-line questionnaire was used to collect initial demographic information. Using these profiles, we categorised interested volunteers based on their gender, age group, status and field of work/study. We randomly chose a sample from each category in order to have a diverse and representative sample of Internet users that included both students and employees from various fields.

4.1. Equipment

The laptops we provided to the subjects had all an identical configurations, with the following software installed: Windows 7 Home Premium; Trend Micro's OfficeScan; monitoring and diagnostic tools including HijackThis, ProcessExplorer, Autoruns, SpyBHORemover, SpyDLLRemover, tshark, WinPrefetchView, WhatChanged; and custom Perl scripts developed for the purpose of the experiment. The scripts automated the execution of the tools and compiled statistical data on the system configuration, the environments in which the system was used, and the manner in which it was used.

In order to avoid biases in user behaviour and at the same time limit the liability of the university, the laptops were sold to the participants at an advantageous, below retail-market price, with laptops staying in their possession at the end of the study.

Regarding the anti-malware product, it was centrally managed on our server in a manner similar as is usually done for corporate installations to centralize distribution of signature file updates. All the AV clients installed on the laptops were thus sending relevant information to our server on any malware detection or suspected infection as they occurred.

4.2. Experimental protocol

Users were required to attend 5 in-person sessions: an initial session where they received their laptop and 4 monthly sessions where we collected the data and analyzed the computer.

During these monthly sessions, participants completed an on-line questionnaire about their computer usage and experience. The questionnaire was intended to assess the participant's experience with the AV product and gain insights about how the laptop was used. Meanwhile, the experimenter collected the local data compiled by the automated scripts.

The data compiled by our scripts included:

- The list of applications installed;
- The list of applications for which updates are available;
- The number of Web sites visited per day;
- The number of Web sites visited by categories per month;
- The number and type of files downloaded;
- The list of browser plug-ins installed;
- The number of different hosts to which the laptop communicated;
- The list of the different locations from which the laptop established connection to the Internet;
- The number of hours per day the laptop is connected to the Internet;
- The number of hours per day the laptop is on.

Diagnostics tools were also executed on the laptop to determine if an infection was suspected. If the AV product detected any malware over the course of the month, or if our diagnostics tools indicated that the laptop may be infected, we requested additional written consent from the participant to collect specific data, such as the browser history, the tshark log files (i.e. network traffic data), and the suspected file(s), in order to help us identify the means and the source of the infection.

In the last visit, participants completed an on-line exit survey about their experience during the study. The aim of this final survey was to identify activities or mindsets that may have unduly influenced the experimental results.

5. Results

5.1. Malware infections

We measured the number of detections by the AV product for each user as well as the number of missed detections by the AV. During the 4-month study, 380 files were detected as malicious by the AV product on 19 different user machines. Regarding the missed threats, our protocol for the identification and classification of missed detections [6] allowed us to detect 19 threats on 12 different machines. By the end of the study, 23 users out of 50 had at least one malware encounter or infection on their computer.

By combining the number of malware encounters reported by the AV and the number of missed detections, we classified users in two groups. The first group contains at-risk users, which are those that received at least one detection or infection. The second group, low-risk users, contains users who did not received any detections of infections during the experiment. As a result, 23 users were classified as being at-risk, and 27 were included in the low-risk group. In order to predict user's likelihood of being exposed to malware, we used the risk category as our dependent variable.

5.2. Features selection

We collected more than 50 different features and selected those that are more likely to predict user's risk of being infected by malware. As a result, we identified 12 features that we used to build our predicting model.

Social-demographic factors and characteristics. In previous work [7], we examined whether gender, age, employment/student status, work/study domain, and computer expertise had any relationship with likelihood of getting infected. We identified that users who reported having a high level of computer expertise were significantly more at risk of getting infected. Users were considered to have a high level of expertise only if they had already configured a home network, created a web page, and installed or re-installed an operating system on a computer. Overall, 18% of users were classified as computer experts for the purpose of our analysis.

While we did not find any significant relationship for the other factors, Table 1 suggests that age could be a risk factor, as shown by the variations between the total population and the at-risk group. We therefore decided to include the level of computer expertise and the age as features for our model.

Table 1. Proportion of users for each factor

Factor		Total	At-risk
Gender	Male	60%	61%
	Female	40%	39%
Age	18-24	38%	35%
	25-40	46%	61%
	41+	16%	4%
Status	Student	64%	70%
	Worker	30%	26%
	Unemployed	6%	4%
Field	Computer Science	26%	22%
	Natural Sciences	52%	48%
	Arts/Humanities	22%	30%
Computer	High	18%	30%
Expertise	Low	82%	70%

Web usage. We used the total number of hours to which the laptop was connected to the Internet as it is a good indicator of the time spent online by users. The underlying assumption being that more a computer is connected to Internet, higher are his chances of being exposed to malware. We also selected the total number of Web sites visited, as it has been identified as a contributing risk factor in our previous analysis [7]. Therefore, users who visit many Web sites are more likely to be exposed to malware by visiting a malicious Web site. Another indicator we selected is the number of files downloaded. As for Web sites, users who download a high number of files from the Internet increase their chance of downloading a malicious file. Table 2 shows the average difference for each factor for the low-risk population and the at-risk population. In almost all cases, the average of the at-risk population is twice the average of the safe population.

Table 2. Average values of Web usage

Factor	Low-risk	At-risk
Time online	169.01	317.36
Number of web sites visited	11 637.11	28 340.62
Number of files downloaded	385.63	637.67

We also looked at the most used Web browser. Table 3 shows that the proportion of at-risk users for Chrome is almost twice as for the low-risk population. While these results do not imply that using Chrome is a risk factor, they suggest that Chrome's users could be more likely to adopt risky behaviors, and hence be exposed to malware.

Table 3. Most frequently used browser

Browser	Low-risk	At-risk
Internet Explorer	41%	19%
Firefox	33%	19%
Chrome	26%	62%

Categories of Web sites visited We also included Web sites categories that we have previously associated to a higher risk of malware infection [7]. Table 4 shows the average number of Web sites visited by each category for the low-risk and the at-risk population.

5.3. Prediction analysis

In order to predict user's likelihood of malware victimization, we used the risk category as our dependent variable, where 1 was associated to the at-risk group and 0 to the low-risk group. Our categorical inputs were the age, the level of computer expertise and the most used Web browser.

Table 5. Confusion matrix

	Training sample			Test sample			Validation sample		
	Low-risk	At-risk	All	Low-risk	At-risk	All	Low-risk	At-risk	All
Total (N)	15	14	29	5	4	9	6	3	9
Correct (N)	13	9	22	4	4	8	6	2	8
Incorrect (N)	2	5	7	1	0	1	0	1	1
Correct (%)	86.67	64.29	75.86	80	100	88.89	100	66.67	88.89
Incorrect (%)	13.33	35.71	24.14	20	0	11.11	0	33.33	11.11

Table 4. Average Web visits by category

Categories	Low-risk	At-risk
Peer-to-peer	6.4	26.3
Software Downloads	36.3	101.4
Streaming Media/MP3	301.5	1 141.5
Email	781.8	2 023.4
Social Networking	2 103.3	5 959.9
Pornography	22.0	317.8

The continuous outputs were the total number of hours connected to the Internet, the total number of Web sites visited, the total number of files downloaded and the total number of Web sites visited for the following categories: email, pornography, peer-to-peer, software downloads, social networking and streaming media/MP3.

We used the automated neural networks module in Statistica to build our predictive model using Multilayer perceptron (MLP) neural network, which used iterative training. To avoid the overfitting of the model, we divided our data in 3 samples. We used 60% of the sample for the training, 20% for the test and 20% for the validation. The first sample trains the model, the test sample verifies the performance of the network while being trained and the third sample, the validation set, performs a final validation to see how accurate is the network to predict new data. We trained up to 20 models and selected the one with the best predictive results, which gave us the model MLP 16-14-2. The selected MLP neural network has 16 inputs, 14 hidden units and 2 outputs. The training algorithm was Broyen-Fletcher-Goldfarb-Shanno (BFGS) and the best solution was found at training cycle 4. The network has an Identity activation for the hidden units and an Exponential activation function for the outputs.

Table 5 shows the confusion matrix for each sample. We can see that the training set has an accuracy of 75.86%, the test set of 88.89% and the validation set of 88.89%, which give us an overnall acuracy of 80.85%. Moreover, the confusion matrix allows us to see that the training sample was equally distributed between the two user groups. An unbalanced training set could have resulted in a high performance for the training, but a bad validation performance.

5.4. Model validation

In order to evaluate the performance of the model we used the Receiver operating characteristic (ROC) curve. It is a graphical representation of the performance of the model, where the sensitivity (true positives or recall) is function of the specifity (false positives or fall-out). The closer the ROC curve is to the upper left corner, the higher the overall accuracy of the model. If a model cannot discriminate between the two groups, the ROC curve will correspond to a diagonal line from the lower left corner to the upper right corner.

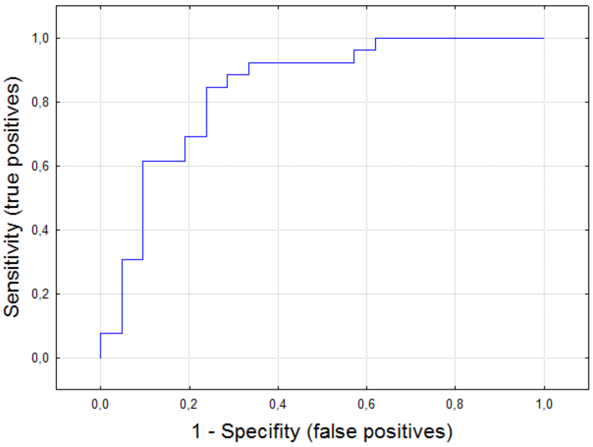
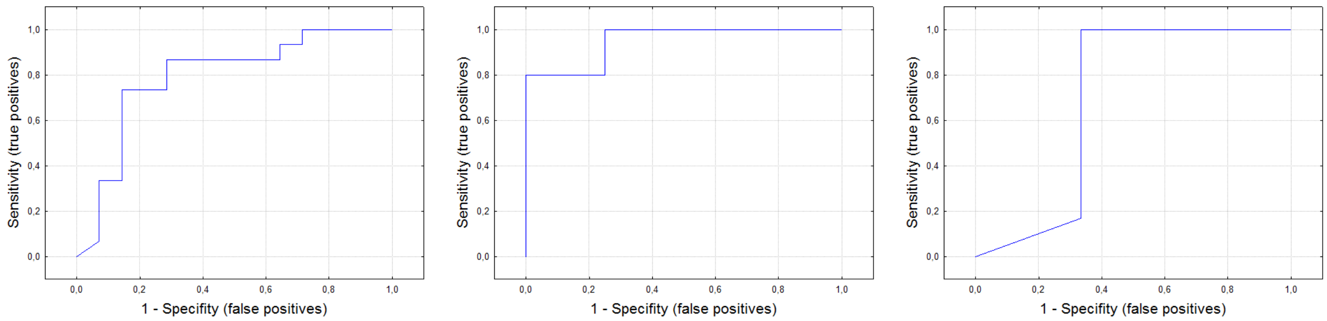


Figure 1. ROC Curve for the entire sample

Figure 1 shows the ROC curve for the entire sample. We can see that to achieve a sensitivity over 80%, a specifity of 25% will be required. Figure 2(a), Figure 2(b) and Figure 2(c) show the ROC curve for the training sample, test sample and validation sample, respectively. In comparison, we obtained similar results for the ROC curve of the training sample. However, for the test sample, the specifity would be lower in order to achieve a sensitivity of 80% while it would be over 30% for the validation sample.

We also used the area under the ROC curve to evaluate the performance of the model. When the model is not able to distinguish between to two groups the area under the ROC will be equal to 0.5 and when there is a perfect sep-



(a) Training sample

(b) Test sample

(c) Validation sample

Figure 2. ROC Curves by sample

aration between the two groups, the area equals 1 and the ROC curve will reach the upper left corner of the plot. The area under the ROC curve computed by Statistica is 0.835 with a threshold of 0.724, which means that we have a good model at predicting at-risk users.

6. Discussion

The results we presented are subjects to certain limitations. The number of malware infections could have been underestimated as our protocol [6] may have missed to identify malicious files that were not detected by the AV product. As a result, the number of at-risk users may be underestimated as we did not have ground truth. Another limitation is the size of our sample. We only had 27 users for the training sample, 9 users for test sample and 9 users for the validation sample. Moreover, our model is limited to a selected number of features, as we did not analyzed all the features we collected. Another potential limitation is the bias that might have been introduced by the fact the users knew they were taking part of a computer security experiment. However, we believe that this potential experimental bias did not significantly affect our results, as only 3 users out of 50 admitted changing their web usage for personal or security reasons.

In future work, we intend to expand our risk factors analysis, as we have collected more than 50 features, such as the type of files downloaded, the network diversity, user's expertise in computer security, level of security awareness, etc. Even though we have been able to build a good prediction model, user-based empirical data are not always available and in some cases, one can only rely on self-reported data. Hence, our next step is to develop a similar model using only self-reported data by users, as we collected both real-usage data and self-reported data.

Although our model did not achieved a high prediction rate, it is a first attempt at developing an extensive user-based risk prediction model from real-usage data combined with social-demographic factors. From an end-user point of view and for key deciders in Information Technology (IT) management, understanding risk factors will help them make better evidence-based decisions on what countermeasures and deployment strategies will be a more effective of their resources. Beyond choosing which countermeasure to adopt, this will include determining and addressing risky behaviour, for example by designing adequate user training and awareness sessions that is appropriately targeted to the audience, or by adopting restrictive policies.

One sector for which we anticipate the development of predicting models is the cyber-insurance industry. Traditionally, the insurance sector has been shy to cover IT-related risk, due to a lack of data allowing the measurement of overall risk exposure and effectiveness of various risk-mitigating strategies. Hence, we believe it is essential to collect and analyze large amount of real-usage data if we want to better understand risk factors as well as the real effectiveness of security countermeasures. To this end larger scale computer security clinical trials should be conducted in order to obtain data on user demographics, user behavior, and concurrent malware exposure.

7. Conclusion

We presented the results of the first attempt at developing a risk prediction model based on both social-demographic and behavioral factors. Using data collected from 50 users over a 4-month period, we selected 12 features to build a predictive model using neural networks. Despite the limited size of the sample, our model achieved an accuracy up to 80% at predicting users' likelihood of getting infected by malware. Moreover, our results suggest that the at-risk user is between 25-40 years old, has a high self-reported level of computer expertise and uses Chrome as his main Web browser. We also confirmed previous findings that visiting many Web sites, as well as downloading many files from the Internet, may increase the risk of malware victimization.

More generally, our work provides evidence that user demographics and high-level behavior information is sufficient for doing basic risk modeling of users. Further studies are required, however, in order to increase the scope and quality of this model before it can be applied in a real-world context.

References

[1] J. Blackbird and B. Pfeifer. The global impact of anti-malware protection state on infection rates. In *Proc. Virus Bull. Conf*, 2013.

[2] D. Canali, L. Bilge, and D. Balzarotti. On the effectiveness of risk prediction based on users browsing behavior. In *Proceedings of the 9th ACM symposium on Information, computer and communications security*, pages 171–182. ACM, 2014.

[3] Y. Carlinet, L. M, H. Dbar, and Y. Gourhant. Analysis of computer infection risk factors based on customer network usage. In *Second International Conference on Emerging Security Information, Systems and Technologies (SECURWARE'08)*, pages 317–325., 2008.

[4] F. Lalonde-Levesque, C. Davis, J. Fernandez, S. Chiasson, and A. Somayaji. Methodology for a field study of anti-malware software. In *Workshop on Usable Security (USEC)*, pages 80–85. LNCS, 2012.

[5] F. Lalonde-Levesque, C. Davis, J. Fernandez, and A. Somayaji. Evaluating antivirus products with field studies. In *22th Virus Bulletin International Conference*, pages 87–94, September 2012.

[6] F. Lalonde Lévesque and J. M. Fernandez. Computer security clinical trials: Lessons learned from a 4-month pilot study. In *USENIX Workshop on Cyber Security Experimentation and Test (CSET)*. ACM, 2014.

[7] F. Lalonde Levesque, J. Nsiempba, J. M. Fernandez, S. Chiasson, and A. Somayaji. A clinical study of risk factors related to malware infections. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 97–108. ACM, 2013.

[8] M. Lee. Who's next? identifying risks factors for subjects of targeted attacks. In *Proc. Virus Bull. Conf*, pages 301–306, 2012.

[9] K. Michael, A. McNamee, and M. Michael. The emerging ethics of humancentric gps tracking and monitoring. In *Mobile Business, 2006. ICMB'06. International Conference on*, pages 34–34. IEEE, 2006.

[10] G. R. Milne, L. I. Labrecque, and C. Cromer. Toward an understanding of the online consumer's risky behavior and protection practices. *Journal of Consumer Affairs*, 43:449–473, 2009.

[11] F. T. Ngo and R. Paternoster. Cybercrime victimization: An examination of individual and situational level factors. *International Journal of Cyber Criminology*, 5(1):773–793, 2011.

[12] A. Somayaji, Y. Li, H. Inoue, J. Fernandez, and R. Ford. Evaluating security products with clinical trials. In *USENIX Workshop on Cyber Security Experimentation and Test (CSET)*, 2009.