Volume 5, Issue 4, 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

On The Comparison of Exposure to Information Security Risk of Web Users by Analysing Their Web Browsing History

Dr. Pragya Singh, Vivek Singh, Pranay Singh, Vyom Srivastava
Cyber Laws and Information Security Division,
Indian Institute of Information Technology,
Allahabad, Uttar Pradesh India

Abstract— Web browsing history of a web users can be a good source of information to characterize them. If a user follow some pattern while browsing the web, it becomes easy to tell if he is exposed to information security risk or not. This paper suggests to calculate a metric (RiskScore) that can be used to compare two or more web users using some attributes of web browsing history.

Keywords— Information security risk, Web browsing history, Comparing Risk exposure, Correlation

I. INTRODUCTION

Among the various vectors used by cyber-criminal to compromise security of a victim, World Wide Web is the most prevalent. It has become a common knowledge that malicious or compromised web pages are used to infect user devices, steal credentials and perform other malicious operations. Every year large amount of fund is invested in research to automatically identify and reduce software vulnerabilities. A few researches have also shown that despite all this effort, the percentage of web pages that are malicious or that have been compromised to serve malicious content is steadily increasing[1,2,3] . The problem with these type of research is that they are not able to classify how much a particular user is exposed to information security risk. These researches have made their prediction for a large number of user base.
The ever increasing number of malicious websites does not always affect every type of users, there are some users who get affected more and there are some users who remain safe without being affected. This is because the users who are safe, are mostly those who are knowingly or unknowingly confine their web browsing activities to safe zones of the internet. In this case it is possible that users web browsing history can helps in deciding if the user is exposed to information security risk or not. This makes it possible to compare and rank information security risk exposure to various users by analysing and comparing their web browsing history. This type of classification is quite attractive in the field of cyber insurance [4], where user profiling is important for concise risk evaluation. In various cases it is needed to assess and compare users against other users in terms of risk exposure. Unfortunately very few number of researches have been made to counter these needs. Comparing user in terms of information security risk exposure can be helpful in various ways, especially in the industry's perspective.
A research at Symantec research labs France has shown that user web browsing attributes are effective to predict information security risk exposure to the concerned web user, Also it is possible to classify web users by categorizing risk exposed to them [5]. Despite the fact that correlation between user's web browsing behaviour and information security risk exposure associated with them can derived, there can be various other factors that affect the information security risk exposure to a particular user, like his Information technology awareness level, experience in a job related to computer science etc. These could play more important role in predicting the associated risk to a web user [6].
Some researchers have also tried to find solutions for these problems especially by performing field studies on computer usage of a limited set of subjects [7]. Some other researchers have tried to relate a few specific browsing behaviour like browsing pornographic sites and being exposed to risk [8]. But there is no study that is general enough to compare user against their security risk exposure.
In this paper we have tried to bring out the most effective attributes of web browsing history that are capable enough to help in comparison of users against their exposure to information security risk . We have filtered down already derived but very large in number, browsing history attributes by using the very famous 80: 20 rule of Pareto principle [9]. The attributes were already derived by a comprehensive study done using telemetry data collected by Symantec [5]. Using the data set they classified users into different categories, one who are safe and the others who are at risk. After classification of users into these two categories 44 major attributes were then derived out that were effective set of attributes in predicting the risk exposure to a user [5]. For this correlation between Risk and Various attributes were calculated.
The need to filter out these resources came into existence when it was realized that these 44 attribute were difficult and costly in terms of time and resource to process for a normal user or organization. To solve this problem filtering those attribute which has good correlation with associated risk were found out. We then also build a python based tool that can process those 9 (20 percent of 44) attributes and can provide a score for a user to compare it against others for ranking and other useful purposes. The software is based on python programming language and need users history as an input to give a risk score for that user.

This paper contains six sections, section 2 describes the variables and characteristics to profile history. Section 3 presents the method used to derive risk scores. Section 4 discuss about the limitations in calculation of risk scores. Section 5 is for discussion on future work and section 6 concludes the paper.

II. VARIABLES AND ATTRIBUTES FOR HISTORY PROFILING

The essence of this paper is the extracted attributes of user's web browsing history. Generally attributes can be a lot of things like values that shows how much a user surfs the web, how popular are the websites being visited, when the user is most active or does user visit similar type of websites or his visits are much diversified. There are very large number of attributes that can be thought of. An exhaustive list of attributes was found out and then 44 were found to have correlation with risk exposure. Among those 44 attributes the correlation coefficient varied from very weak to medium range. We in this paper have considered 35 out 44 attribute which have correlation coefficient more than 0.5. On those 35 attributes, we applied Pareto principle and found the top 7(20% of 35) which had substantial correlation coefficient value. The following paragraph will describe the various attributes that are considered in this paper.

How much a user is active on the web can be found out by total number of hits (*hits)* , how diverse a user's activity are on the web (*distinct_Hits),* how many distinct urls have been requested (*distinct_url* how much a user is active on an average day(*hits_Per_Day),* What time of the day a user is more active, this includes total number of hits in business hours(*hits_business),* Number of hits in the evening(*hits_evening),* Number of hits in the night (*hits_per_night).* Also how diversified are the websites visited by the user can be found out by counting the number of hostnames(*hostnames),* Number of different Top level Domains (*tlds)* , number of distinct domain names (*domains).* Number of non 'com','net' and org top level domains (no_comnetorg_tlds). Similarly there were various other attributes which we have not listed to space limitation. The top seven attributes and their correlation coefficient are presented in the *table 1*.

Table 1: Shows selected attributes and there co-relation coefficient with Information Security Risk

Attributes	Correlation Coefficient
hits	0.34
distinct_urls	0.39
domains	0.44
hostnames	0.43
TLDs	0.41
no_comnetorg_tlds	0.41
hits_per_day	0.32

III. METHODS AND ANALYSIS

The essence of this paper is to help in comparison of information security risk exposure among various user. For this we have formulated a mechanism to calculate a risk score which does not signify any real or absolute meaning for risk but is used to compare two users. The user with higher risk score might have higher risk exposure to information security risk and the user with lower scores might have lower risk exposure to information security risk. The method to calculating risk score involves all of the seven attributes and each of the attribute have weightage in the order of their correlation coefficient values.

$Attribute_i \in$ {hits, dictinct_urls, domains, hostnames, TLDs, no_comnetorg_tlds, hits_per_day}

Example: $Attribute_1$ = hits, $Attribute_2$ = distinct_urls, etc.

$RiskScore\ [X]_i$ = RiskScore contribution by attribute 'i' to a user X.

RiskScore[X] = Total RiskScore for a User X for all attributes.

SOC = Sum of all correlation coefficient

= 0.34 + 0.39 + 0.44 + 0.43 + 0.41 + 0.41 + 0.32

= 2.74

$VOA\ [X]_i$ = Value Of Attribute 'i' for user X

$CCOA_i$ = Correlation Coefficient Of Attribute 'i'

$$(RiskScore\ [X]_i) = \frac{(VOA[X]_i) * (CCOA_i)}{SOC}$$

$$RiskScore\ [X] = \sum_{i=1}^{i=7} \log_{10} RiskScore[X]_i$$

Example: We have used web history of two users and found out various values for various attributes to test our hypothesis. The following values were extracted using the python based tool that we built.

User 1:

hits	110880
distinct_urls	74660
domains	3088
hostnames	4538
TLDs	125
no_comnetorg_tlds	37417
hits_per_day	356

User 2:

hits	3295
distinct_urls	1819
domains	166
hostnames	204
TLDs	24
no_comnetorg_tlds	612
hits_per_day	143

SOC = 2.74

RiskScore [User 1]$_1$ = VOA [User 1]$_1$ * CCOA$_1$ / SOC
= 110880 * 0.34 / 2.74
= 13758

RiskScore [User 1]$_2$ = 10626
RiskScore [User 1]$_3$ = 495
RiskScore [User 1]$_4$ = 712
RiskScore [User 1]$_5$ = 18
RiskScore [User 1]$_6$ = 5598
RiskScore [User 1]$_7$ = 41

RiskScore [User 1] = $\log_{10}13758 + \log_{10}10626 + \log_{10}495 + \log_{10}712 + \log_{10}18 + \log_{10}5598 + \log_{10}41$
= 20.328

RiskScore [User 2]$_1$ = VOA [User 1]$_1$ * CCOA$_1$ / SOC
= 3295 * 0.34 / 2.74
= 408

RiskScore [User 2]$_2$ = 258
RiskScore [User 2]$_3$ = 26
RiskScore [User 2]$_4$ = 32
RiskScore [User 2]$_5$ = 4
RiskScore [User 2]$_6$ = 92
RiskScore [User 2]$_7$ = 17

RiskScore [User 2] = $\log_{10}408 + \log_{10}258 + \log_{10}26 + \log_{10}32 + \log_{10}4 + \log_{10}92 + \log_{10}17$
= 11.738

RiskScore [User 1] = 20.32
RiskScore [User 2] = 11.73

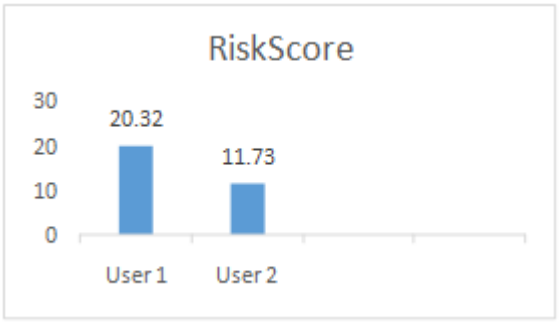

Figure 1: Comparison among User1 and User2 against their risk score

Figure 1 shows that user 2 is more exposed to information security risk. Similarly if we can calculate values of Risk Score for various user we might be able to rank them according to exposure to information security risk. This can also be considered as input while assessing the security posture of any organization or assessing the change in user behaviour in the course of time.

IV. LIMITATIONS

The essence of this paper is the exhaustive list of attributes extracted to check for the possible correlation with information security risk. It could be the case that the attributes gathered are not exhaustive. Also we have considered Pareto principle to select the most effective attributes (Top 20 percent attributes) which can have a diluting effect on the result. Also as our analysis rely on correlation and it is known that correlation is not a sufficient condition for causation as there are possibly some hidden variable which get ignored. We in this paper have assumed that we have all the possible effective attributes covered. Also sometime extracting even these attributes are not possible due to certain not so common system configuration of user. In cases like this, our paper may not be as effective as it is solely based on processing the attributes of web browsing history. In this paper we have not taken care of various context based attributes like user awareness level to information security or how much security measure has already been taken in hardening security of the user's system.

V. DISCUSSION AND FUTURE WORK

Lots of efforts have been made by researchers around the globe to analyse information security risk but very few have tried to compare exposure to information security risk of individual users. In this paper a method have been proposed to compare two or more individual users against their exposure to information security risk by analysing their web browsing history. The method has the capabilities to be used in various scenarios just by attaining a little context based information, for e.g. the method can be used by large Internet Service Providers to compare risk exposure to their customer by clustering them according to their geographical location and then comparing using our method. This will allow the Internet Service provider to gain an insight about their customer. With a little modification this paper can be used by firms to compare exposure to information security risk among various department or same department at different point of time. It is also helpful in assessing improvement in user awareness level after being subjected to information security training by calculating risk scores at different point of time and analysing them. The work can be more concise if more situation based attributes are taken into consideration.

VI. CONCLUSION

Think of a situation where one has been given a task to motivate employees of a firm regarding information security awareness and it has been planned to give performance appraisals based on their current awareness level, assessing their exposure to information security risk against each other might be very handy in helping to choose the deserving candidates to get the benefits. From such a rudimentary task to a task where security assessment of the firm is going on, this paper can be quite helpful in giving a perspective to see how employees of a firm have improved in the course of time. Being able to compare exposure to information security risk at different point of time and different set of web users, might provide a way to get an insight of people who are asset of any organization and recommendation (eg. need of information security training) of various types can be made. Comparison based on the processing of a few attributes of web browsing history is supposed to be very cost effective in terms of time and resource but has a potential to give a whiff of exposure to information security risk among various users or cluster of users which can be very valuable information at times. This paper has tried to give a mechanism to achieve that.

REFERENCES

[1] "Kaspersky Security Bulletin 2012. The overall statistics for ..." 7 Feb. 2015 <http://securelist.com/analysis/kaspersky-security-bulletin/36703/kaspersky-security-bulletin-2012-the-overall-statistics-for-2012/>

[2] "Internet Security Threat Report - Symantec." 2013. 7 Feb. 2015 <http://www.symantec.com/security_response/publications/threatreport.jsp>

[3] "Websense® 2013 Threat Report - Websense.com." 7 Feb. 2015 <http://www.websense.com/content/websense-2013-threat-report.aspx>

[4] Böhme, Rainer, and Galina Schwartz. "Modeling Cyber-Insurance: Towards a Unifying Framework." WEIS 7 Jun. 2010.

[5] Canali, Davide, Leyla Bilge, and Davide Balzarotti. "On the effectiveness of risk prediction based on users browsing behavior." Proceedings of the 9th ACM symposium on Information, computer and communications security 4 Jun. 2014: 171-182.

[6] Zhou, Yajin et al. "Hey, You, Get Off of My Market: Detecting Malicious Apps in Official and Alternative Android Markets." NDSS Feb. 2012.

[7] Lalonde Levesque, Fanny et al. "A clinical study of risk factors related to malware infections." Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security 4 Nov. 2013: 97-108.

[8] Wondracek, G. "Is the Internet for Porn? An Insight into the Online Adult ..." 2010. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.165.3360>

[9] Koch, Richard. The 80/20 principle: the secret to achieving more with less. Crown Business, 2011.

