Lab Study or Field Study: the Problems that Researchers Face in Measuring the Effectiveness of Browser SSL Warning Designs

Rong Wang

Department of Computer Science, University of Auckland

rwan074@aucklanduni.ac.nz

***Abstract*—Secure Socket Layer (SSL) protocol have become a very commonly used protocol due to the demand of secure communication between two parties over the Internet. It is used in a wide range of applications to provide users services such as e-mail, Internet banking etc. Most web browsers have an active SSL warning that is displayed to the user when the browser could not successfully identify the service provider it is trying to connect to. This report will analyse some previous studies conducted to evaluate browser SSL warning effectiveness.**

***Index Terms*—Browser security, SSL warning, Laboratory study, Field study**

I. Introduction

THE Internet has become one of the primary source for people to gain access to information and services. Browsers have become one of the most widely used application that people use in order for people to interact with the Internet. One of the challenges that the browser vendors face is to carefully design their browser's security mechanisms in order to effectively protect their users from malicious third parties. With the world's Internet population reach the mark of three hundred million[1], any vulnerabilities in browser could potentially affect millions of users worldwide due to browser's widespread usage.

Most modern web browsers have what is so-called active SSL warning when a secure communication could not be established between the user and the other party, SSL warnings signal a potential Man-In-The-Middle attack is taking place. On contrast to passive security indicators (e.g. Google Chrome's lock icon on its URL bar), an active warning normally requires certain actions to be performed by the user in order to override it. Since the user is expected to make the final decision upon encountering such warnings, it is crucial that the warning is well designed to provide the user necessary and sufficient information in order for the user to make the correct choice. As SSL warnings

[1]Number taken from Internet World Stats www.internetworldstats.com

could also be false positive meaning that there is no actual threat present to the user but rather an unexpected error (misconfiguration in server etc.) with the service provider, it is than the browser's job to help the user to identify the situation.

One common problem that security researchers face in study of browser security is that the effectiveness of one specific security mechanism is hard to measure and quantify. How does one gather data? What sort of data does one need? Before we can analyse it. In this report, we will take a close look at three previous studies that focused on SSL warnings, but were conducted in very different ways. Two of these studies were laboratory studies and the other being a field study. All of these three studies measured the percentage of users in their sample that chosen to ignore the warning upon seeing one thus provided a base of comparison for us.

In this report, we will try to evaluate their methodologies as well as analyse their results. Hopefully this will provide some insights to anyone in the future that wishes to conduct studies in a similar area.

II. Overview

The three studies we will discuss in this report are:

- [SE09] referred to as the "**CMU study**" in this report conducted by Sunshine et al.
- [SH11] referred to as the "**UBC study**" in this report by Sotirakopoulos et al.
- [AF13] referred to as the "**FAC study**" (Firefox and Chrome study) by Akhawe and Felt.

The CMU study was a laboratory study conducted at Carnegie Mellon University in 2009 that investigated user behaviour towards the native SSL warnings implemented in Internet Explorer 7 (IE7), Firefox2 (FF2), Firefox (FF3) as well as two re-designed custom warnings. The UBC study was the other laboratory study, it was conducted at the University of British Columbia in 2011. As stated by the author of [SH11] (page 1), the purpose of this study was to "validating and extending" the CMU study. This study experimented with IE7 and FF3's native SSL warning as well as two of their re-designed custom warnings, one for each browser. The FAC study was conducted in a very different manner. Firstly, the data used in this was not collected in a laboratory environment, the researchers relied on a cross platform performance testing framework called Telemetry that is implemented in both FF and Google Chrome (GC) to collect data from endpoint users. This framework only work in the background and does not interact with users in any way, that is, it has no impact on user decisions but only record certain behaviours. Secondly, this study investigated Phishing and Malware warnings as well as SSL warning for both of these web browsers. Since this report focuses on SSL warning, we will ignore the finding of this study on Phishing and Malware warnings. Thirdly, this study was conducted this year (2013), due to the rapid development of web browsers, the data of this study was collected from FF23 and GC25. This may concern some readers when we compare the findings of this study on FF to the other two studies as they

Fig. 1. SSL warning page for FF3

are of a much earlier version. **Fig. 1.** is a picture of the SSL warning page of FF2 and **Fig. 2.** is the same page for FF23. Note that the text in the red box in Fig. 2. is only shown when the user click on the corresponding buttons, "**Technical Details**" and "**I Understand the Risks**". From these two pictures, we can see that even thou the actual wording used in these warnings are different. They both exhibit a similar overall design, they both have "Larry the passport officer" (the black policeman like drawing with yellow background) in the top left corner and the main warning message in bold in centre top. If there were any observed difference in these two SSL warning, it should be due to the wording rather than the overall design of these warnings.

III. Methodology Discussion

This section will describe methodologies used to gather data in these studies, we will try to not include too much details in this section but rather talk about the general procedure that these studies went through. We will talk about specifics in detail when we discuss them.

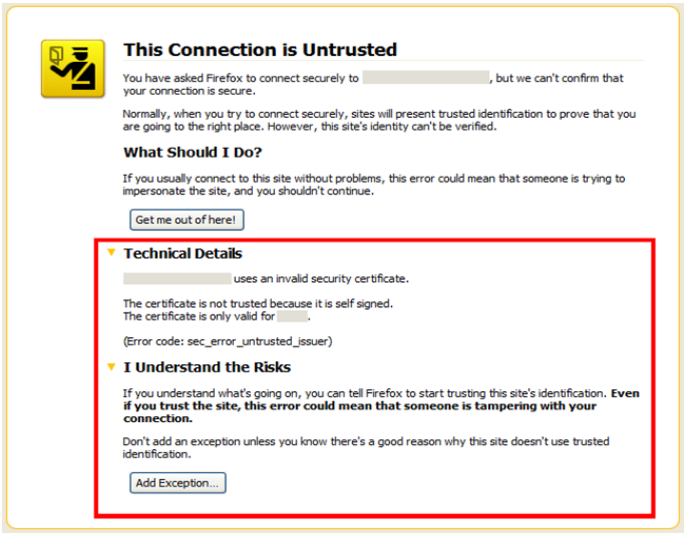

Fig. 2. SSL warning page for FF23

A. The CMU Study

Firstly, the researchers posted their study "on the experiment list of the Center for Behavioral Research at CMU and also hung posters around the CMU campus" [SE09] (page 8). All potential participants were then given an online screening survey and only those who passed the online survey (those who met the requirement of the study) were allowed to participate. The selection procedure stopped once they had 100 eligible participants.

The experiment was then conducted in a laboratory environment. Each participant was randomly assigned to one of the five conditions. Each condition consists of one particular web browser (FF2, FF3 or IE7) and one SSL warning design (native FF2, FF3 or IE7 SSL warning or one of the two custom designs). The participant was then asked to perform different tasks given by the instructor on site including access their Internet banking service as well as access the University's online library service. Various measurements were made during

Fig. 3. SSL warning page for FF2

their performance. FF2 had a very different looking (see **Fig. 3.**). Since this design is quite out of date comparing to modern SSL warning designs, we will ignore the results and findings that solely focused on FF2 in this report.

Finally, the participants were asked to complete an exit survey upon completion of their tasks.

B. The UBC Study

This study was designed in a similar way to the CMU study, participants from both of these studies had to go through three phases, namely they were, recruitment, actual laboratory experiment and an exit survey. The experiment and exit survey phase were done in a similar fashion. Both experiments were conducted in a laboratory environment, with instructors on site giving the participants similar tasks to complete while measuring some behavioural factors of the participants, specifically the percentage of individuals that ignored the SSL warnings upon seeing them. Both exit surveys were of similar style as well, we believe the researchers of this study deliberately did so as one of their study purpose was to validate the findings of the CMU study. However, researchers of this study designed the recruitment phases differently in order to avoid or reduce some factors that they believed were affecting the validity of the CMU study.

Recruitment: One obvious problem during the recruitment procedure of the CMU study was that the sampling population was restricted to the University's students and staff. Furthermore, in the 100 eligible participants that actually went through the experiment, only 2 were non-student [SE09] (page 12). As in the words from the author of [SH11], "[University students as a group are mostly] Western, Educated, Industrialized, Rich and Democratic". Not only this is not a representative sample of the average users, this well characterized group may also have certain psychological trend that may have an impact on their decision making during the experiment hence reduce the accuracy even further if one tries to apply the finding of the CMU study to a more general population. Therefore, the researchers of the UBC study conducted the recruitment phase in a slightly different way in effort trying to reduce such effect.

"First we advertised the study ... around the UBC campus and the Vancouver community centers ... advertisements on Craigslist" [SH11] (page 4) (Craigslist is an online advertising site, see www.craigslist.org for more details). By doing so the researchers have successfully avoided having a primarily students based sample, their participants were more diverse in both age and occupations.

However, we believe that this was still not a good representative sample for the average user group. People who are more likely to visit community centers and one particular advertising site were the only two other groups other than students that could become potential participants for this study and people from these two groups may possess some other unknown characteristics that may affect their decision making process.

It is not hard to see that both samples in these two studies are biased. Only individuals belonged to certain social groups were the potential participants for both studies. Furthermore, the participants were not chosen at random from these social groups to participate in these studies neither, both of these studies requires the individuals to contact the researchers first in order to be listed as potential participants. This is known as self-selection bias, in which the participants' decision on participating in these studies maybe associated with some characteristics that may affect their decision making in the experiment and ultimately bias the results of the study.

This is one of the challenges that many researchers face now days. How do we obtain a good representative sample for the population of interest, the average users. There is no current solution to the problem. Even though we can increase the sampling group coverage by things such as advertising the study through a wider channel, the cost of having such a study is likely to very expensive or very time consuming. However, this does not necessarily mean that the results and findings of these studies are not useful at all. Even if we assume that the individuals participated in these studies have certain traits, these traits are less likely to be the main contributor to the gap difference observed in behaviour towards different SSL warnings designs and we can still use them to gain insights on how should we make better warning designs[2]. We will talk about this in more detail in the next section when we discuss the results of these studies.

C. The FAC Study

As mentioned before in the previous section, this study was designed very differently to the others. Through the usage of Telemetry framework the researchers were able to collect a huge amount of data, this study had more than sixteen million SSL warning impressions in total for FF23 and GC25. One warning impression is the recorded response that the one user performed (ignore or navigate away) upon encountering one SSL warning. Similar to the other studies, the researchers of this study calculated the percentage of warning impressions that were ignored by the user in their sample.

1) Sampling issues: The Telemetry framework only collects data from users who opt in their browser's data collection program meaning that this was not a random sample neither. It suffered the same draw back as the other studies from only being able to collect data from users who were

[2]The author of this report does not have any evidence to verify this point, no actual study nor data was conducted and collected for this report. The idea is that if the data collected in two studies were both from populations with certain similar traits, why would participants in these studies behave differently? Unless there are some other unknown factors, the way the warnings are designed in these studies seems the only logical contributor.

self-selected to participate. Being much larger in sample size does not help resolve this problem neither. Larger samples are more likely to be less biased, but that is only the case when the sampling population is not heavily biased. Since the sampling population was not selected at random in this study, having a larger sample size is not likely to reduce its bias. Second point to notice is that all FF23 warning impressions were collected from pre-release channels [AF13] (page 15, table 7 shows no data on SSL warning impressions were collected in FF release channel) while GC25's sample were collected from pre-release as well as stable channels (stable channel is GC's equivalent channel to FF"s release channel). By default, two of FF's pre-release channel (there are three in total) participates in its Telemetry program, the author of this report is not sure which if any of GC's update channels have Telemetry enabled by default. Through collecting data from different types of channels for FF and GC, we could be getting our data from two very different sample populations as it is a logical assumption to make that the users of pre-release channels are more likely to be more technically experienced and therefore the technically more experienced users could be a potentially overrepresented group in the FF sample. And technical experienced users may have a different behaviour trend to average users.
Furthermore, due to the way that Telemetry is implemented in FF, the researchers could not identify individuals in their warning impressions, that is, they do not know if two warning impressions were generated by the same client or not. This could lead to a few number of individuals contribute a large number of clicked through warning impressions. Here is one possible scenario, a technician is fix a mis-configuration in his server that is causing an SSL warning, therefore he is constantly re-configuring the server as well as keep trying to access his server from another machine using FF. Another possible scenario is that the warning impressions was generated by a web crawler that is tasked to look for servers with SSL warnings (for web security research or some other reason). These types of individuals will likely to contribute to the click through rate a lot more than average users, and they could not be accounted for.

2) Laboratory effect: One advantage that this study had was that the behaviour recorded by Telemetry did not suffer from being subject to what is so-called "laboratory effect" like the other two studies did. Laboratory effect refers to the impact that the experiment environment had on how test participants' behaviour during the study. The author of [SH11] (page 8) stated that "we believe that there is a significant impact of the laboratory environment [on user behaviour]". On the exit survey of the UBC study, 33% of the participants who ignored the SSL warning claimed that the reason why they did so was because "It is a study" and another 13% responses were because "[the participants] wanted to complete the task" [SH11](table 4). The sense of safety that the study environment provided to the participants made them believe that their sensitive

personal information was safe. Another related phenomena was described by A. Patrick [AP07] as the "Task Focus" effect, in which "[during studies, the participants] take the tasks very seriously and are highly motivated to complete [them]". These two phenomenons together might have shifted the participants' behaviour dramatically in these studies. The researchers of the CMU study was aware of these effect, therefore they provided an alternative way to complete the tasks (e.g. to call the participant's bank rather than using their Internet banking service) in effort trying to reduce these effects. However, we believe that is not enough in this case here (and there is no evidence suggesting so). As reported by A. Patrick [AP07], sometimes the effect of task focus is so strong, it gets to an extend where the user "fail to notice or choose to ignore things ... around them". This study did not carry out any counter measures to address this problem (such as asking users related questions in the exit survey).

Since the Telemetry framework observe (or rather record) without interact with them, the users are not likely to be aware of the fact that they are being subject to certain study. Therefore, more realistic data can be collected this way.

The biggest drawback of this study as far as the author of this report believes is the fact that they were not able to obtain user feedbacks in any form. User feedbacks can provide a lot of insights about the decision making process that the user went through. The lack of user feedbacks caused them not being able to reach any definite conclusions. They can only try to interpret the data and behavioural features Telemetry recorded, but doing so without referring back to the user's behavioural insights might be inaccurate.

IV. Results and findings discussion

Before we can analyse the findings of these studies, we must ask, what is the ultimate purpose of browser SSL warnings? Is it to prevent the user from visiting the service when such an error occurs? Or is it something else? The author of this report believes that browser SSL warnings should behave as a "reference tool", it should provided all the necessary and sufficient information to the user in order for them to make the correct decision. Is there an actual attack taking place that could harm me? Or is it a just a machine error (server mis-configuration etc.) in the service provider I am trying to connect to. One could argue that it is the service provider's responsibility to make sure their hardware and software are working as intended and does not produce any errors, and the sole purpose of SSL warnings are to dis-encourage or even does not allow user to bypass at all time. This would be ideal in a "perfect" world where every machine as well as their administrators always behave as intended and does not introduce any errors. This is simply impossible with the technology we have today.

The author of [AF13] has pointed out that a large number of security experts in the industry nowadays believes that users can not be relied on making the correct security decisions and are "oblivious to security cues" [AF13](page 1). All three studies had evidence against this view, as

they have all shown that at least a proportion of users (out of those in their sample population at least) have paid attention to these warnings in these studies, but their actual behaviour is affect by a number of factors.

Understanding and perception

If we go back and take a look at **Fig. 3.**, it was not a good SSL warning mainly due to the wording used in this particular design. In the main body of the this warning message, its recommended action for users were "notify the site's webmaster about this problem" and "examine this site's certificate carefully" which is absurd when thinking from the average users' point of view. The average user does not have the necessary background knowledge nor the proper tools to do what is asked of them here, if not worse further confusing them. If we have a look at **Fig. 2.**, its successor from FF23 is doing a much better job. If we look at the message under the "Get me out" button, it says "If you usually connect to this site without problems, this error could mean that someone is trying to impersonate the site, and you shouldn't continue". This is a much wording than the ones used in FF2, it stated very clearly what is the potential threat here and it recommended an action the user can actually perform. To summarize, it provided or at least tried to explain to the user what could be potential happening (potential harm) as well as suggesting the user with one potential solution (leave the page).

The argument in the above paragraph leads to the idea that understanding of the situation and the level of risks perceived should play an important role in user's decision making process. However, both the CMU and UBC studies were not able to find any evidence supporting this idea. It might be due to the fact that the sample size in both of these studies were quite small so no statistically significant conclusions can be drawn. Or maybe the skewed sample in both studies biased towards some trends is more likely to ignore SSL warnings. For example, they might be of certain psychological trend that is less concerned about privacy in general and therefore less likely to pay attention to what is displayed in the warning.

The author of [AF13] also believed that providing sufficient information is crucial to users' decision making. In section 7.6 of [AF13], they author said that "[FF] places information about SSL errors under "Technical Details" and in the "Add Exception" dialog ... It is possible that moving this information into [FF] primary warning could reduce their click-through rates ...". But if we look at the message under "Technical Details" in **Fig. 2.** for example, average users probably won't understand the some of the terms used in there such as "[the certificate] is self signed" and since they do not understand what it means, it will not be likely to help them. As we have mentioned before, since this study lacked any form of feedbacks, some of its claims could not be verified.

User effort

One major design difference between FF23 (see **Fig. 2.**) and GC25 (see **Fig. 4.**) SSL warning designs

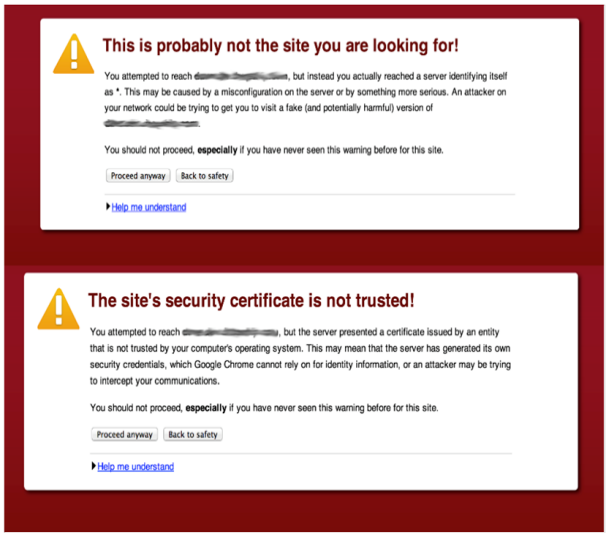

Fig. 4. SSL warning page for FF2

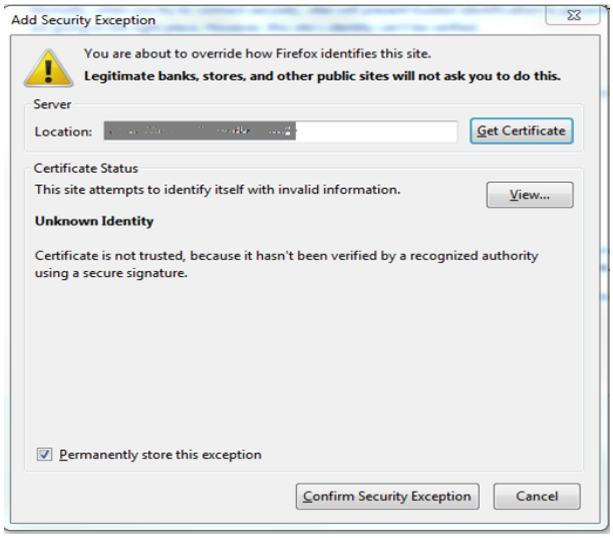

Fig. 5. SSL warning page for FF2

was that the user would only need to click on the "Proceed anyway" button in order to ignore GC's warning. With FF23, the user would have to first click the button "I Understand the Risks" then click the "Add Exception" button that can be seen at the bottom of **Fig. 2.** and finally click on the "Confirm Security Exception" button on the pop-up window as shown in **Fig. 5.**. To summarize, GC has one single page warning page which had the only one button the user has to click in order to bypass but FF requires the user to navigate through their warning message including clicking on three different buttons in two different windows in order to be able to ignore it. The author of this report believes this could be one of the major reasons why FF had a lower clickthrough rate than the other designs (IE7 had a similar design to GC). The author of [AF13] does not share the same point of view, in section 7.4 of [AF13], the author stated that "[their] data suggests that the amount of effort . . . does not always have a large impact on user behaviour". But their data failed to measure the amount of users who wished to ignore the warning but did not know how as it did not record the percentage of users who continued through all three clicks upon making the first click. The author of [SE09] stated that "FF3 users (FF3 had a similar SSL warning design FF23) may have been prevented from visiting the website because they did not know how to override warnings" [SE09]. (section 4.2.4) The evidence they had for this claim was that "Seven of the 14 participants who did not understand the FF3 warning called the bank." [SE09] (section 4.2.4). A moment of thought would reveal that their reason was logical, the user who did not understand the warning message is more likely to perceive less or no potential threat therefore should be more willingly to bypass such warning, but only a half of them did so. So the warning design being difficult to override is a potential reason.

V. Conclusion

As we have discussed in previous sections, all three studies have some designing flaws that makes the generalization from their sample population to the average users difficult or inaccurate. Even thou we can still use some of their findings or results to gain some insight on how should browser SSL warning to be designed, it is often the researcher's aim to be able to identify some behavioural insights toward the average users. So, is there any other study design model we could adapt here? Both authors of [AF13] and [SH11] have proposed a similar model that is a hybrid of field and laboratory study. For example, imagining a researcher is trying to investigating the effectiveness of one particular design of a SSL warning. The researcher installs the particular implementation on particular participants' computer as well as some performance measuring tools (with the participants' consent of course). And once the measuring tool has gathered enough information, the researcher can then study the measurements made as well as getting feedbacks from the particular participant. This might not be an easy task, the cost of such studies could be expensive, takes relatively long time before enough data is gathered and may have research ethical issues. Furthermore, this model still does not solve the problem of how could we obtain a good representative sample. Will the results from such study be good enough to extend to the average users? Probably not. But at least it provides us a new way of getting data that is not affected by the laboratory effect as well as getting feedbacks from study participants, which we have learned could provide useful insights.

References

[AF13] D Akhawe, A.P. Felt, Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness, Proceedings of the 22th USENIX Security Symposium, 2013

[SH11] A Sotirakopoulos, K Hawkey, K Beznosov, On the Challenges in Usable Security Lab Studies: Lessons Learned from Replicating a Study on SSL Warnings, Proceedings of the Seventh Symposium on Usable Privacy and Security. ACM, 2011

[SE09] J Sunshine, S Egelman, H Almuhimedi, N Atri, L.F. Cranor, Crying Wolf: An Empirical Study of SSL Warning Effectiveness, USENIX Security Symposium, 2009

[AP07] Andrew Patrick Commentary on Research on New Security Indicators 2007, http://www.andrewpatrick.ca/essays/commentary-on-research-on-new-security-indicators