Heinrich C. Mayr, Martin Pinzger (Hrsg.): INFORMATIK 2016,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2016 15

NoPhish: Evaluation of a web application that teaches people being aware of phishing attacks

Alexandra Kunz[1], Melanie Volkamer[1], Simon Stockhardt[1], Sven Palberg[2], Tessa Lottermann[2], Eric Piegert[2]

Abstract: Phishing has evolved to a serious cause of risk in our daily contact with the World Wide Web. Therefore, different extensions and plugins for web browsers were developed to detect phishing websites. To furthermore minimize the risk of falling for a phishing attack, the users themselves have to be educated. Therefore, the online game "NoPhish" has been developed, which explains the basics of phishing attacks and how to detect them efficiently. In the following study, the success rate of this online tool was measured. The goal was to determine which phishing strategies are effective in fooling users, which strategies can be practised well and which strategies are still effective in fooling users after having been taught by some educational material. The effectiveness of "NoPhish" in increasing users' security awareness and the ability of detecting phishing URLs could be proven. Furthermore, it could be determined which types of phishing should be drawn special attention to in future development of phishing education material.

Keywords: Usable security, security awareness, phishing education

1 Introduction

Phishing represents an attempt from so-called phishers to elicit confidential information of users by using faked websites. These phishers want to get access to private account information and passwords which are used for e.g. e-banking, social networks or online shops. A successful phishing attack can have disastrous consequences for the victims leading to financial losses and identity theft. Usually these phishers send fraudulent e-mails or chat messages with a link and the order to click on it. There is a multitude of different phishing attacks like spear phishing where phishers want to increase their success rate by sending e-mails to specific companies with individual matched content. Another type of phishing is called clone phishing where phishers clone a previously sent message and replace the legit content with malicious information like links or formulas. That phishing plays a major role in our daily life shows the statistic of the Anti-Phishing Working Group (APWG), identifying around 50,000 new phishing websites every month, with retail being the most targeted industry sector at the moment and payment services close behind, but also in other sectors such as social networking. More and more companies fear that users will lose confidence in electronic commerce. Therefore, an

[1] Technische Universität Darmstadt, Fachbereich Informatik, SECUSO, 64289 Darmstadt, kontakt@secuso.org
[2] Technische Universität Darmstadt, Fachbereich Informatik, SECUSO, 64289 Darmstadt, kontakt@tu-darmstadt.de

efficient protection against phishing is needed.

Ludl, McAllister, Kirda and Kruegel (2007) engage in the effectiveness of techniques to detect phishing sites in their research. Their results show a 90% rate of detection of phishing attempts of blacklist-based solutions. The method identifies phishing sites after they are launched and reported as suspicious. These solutions represent an efficient way of discovery. However, their research covers only extensions and plugins for web browsers. But to furthermore minimize the risk, the users themselves have to be educated.

Dhamija, Tygar and Hearst (2006) studied which phishing attacks succeed in fooling users and why. They tested 9 different types of phishing attacks like different types of spoofing or different types of requested information. 22 participants were shown 20 web sites and asked afterwards to determine which ones were fraudulent. Additionally, the users had to evaluate their confidence about their decisions. Generally, the users were quite confident about their answers. However, 7 participants had never heard of phishing before they participated the study. Furthermore, 13 participants never paid attention to "HTTPS" and further 5 participants mentioned that they even never take a look at the address bar at all. To sum up, the results of the study lead to the conclusion that educating users about phishing has to be taken very seriously. These alarmingly high values show even more the importance of user education such as in the research of Sheng, Magnien, Kumaraguru, Acquisti, Cranor, Hong and Nunge (2007) with their online-game „Anti-Phishing Phil". Their results demonstrate that games represent a highly efficient way of information transfer. Using a story-based approach with a phish character guiding through the game they provided a challenging, contextual and interactive game experience. Their results show the improved skills of the game users in detecting phishing-URLs compared to groups, which were not allowed to use the interactive game but instead read existing training material or had been tutored.

Canova, Volkamer, Bergmann and Reinheimer (2014) had a similar approach for educating users by an interactive game. They developed the Android app "NoPhish" which tutores the users in detecting phishing-URLs. Due to the fact that people are not regularly confronted with phishing attacks, the authors analysed its effectiveness on users' knowledge retention. The results of their studies show that users of this app are more successful in detecting phishing-URLs, particular over a longer period.

Sheng, Holbrook, Kumaraguru, Cranor and Downs (2010) study the effectiveness of different educational materials to identify phishing webpages in their research. Participants of their study had to complete a role play to measure their susceptibility to fall for phishing attacks. On the basis of these results the participants were assigned to four different experimental groups with different educational material: a PhishGuru cartoon, Anti-Phishing Phil, popular web-based training materials and a combination of Anti-Phishing Phil plus a PhishGuru cartoon. The results of the authors' study show that educational material reduces the end-users' risk of supplying private information on phishing webpages by 40%.

A further work of Kumaraguru, Rhee, Acquisti, Cranor, Hong and Nunge (2007) is about

the comparison of three methods for improving the users' skills of detecting phishing attacks – two embedded training designs and another method consisting of simple email security notices. The embedded training methods consist of regularly sent phishing emails. If a user does not assess the mail as a phishing attack and clicks on a link, the user receives a warning. The first method presents information using text and graphics, the second method uses a comic strip format. The results of the authors' study demonstrate that the number of identified phishing emails decreased using the embedded training designs. The authors found out that the comic strip intervention was most effective while security notices were rather ineffective in teaching people about phishing attacks.

In the present study, the challenging and interactive web application of "NoPhish" was developed (based on Volkamer's "NoPhish" Android app) to educate users and increase their awareness and ability in detecting phishing URLs. The recruited participants had to do a pretest for checking their ability in detecting phishing attacks. Afterwards they ran through a training phase where they learnt the basic attacks of phishing and in a final step they had to test their learnt knowledge in a concluding posttest. The goal of the present study is to investigate 7 different types of phishing, more precisely to analyse which of them are effective. Another interesting fact is, which attacks can be practised well so that after some training time there is a high success rate in detecting phishing attacks. The last aspect is to investigate which phishing strategies are still effective in fooling users after they were educated by teaching materials.

2 Method

2.1 Participants

Participants were recruited from December 2015 until January 2016 via social networking sites like Facebook and via sending personal messages to circles of acquaintances. 65 users registered at the web application but 9 of them answered neither pretest nor posttest and 18 of them did not answer the posttest. Finally, 32 of 65 participants were left for analysing their data. The behaviour of 14 female participants and 18 male participants aged between 19 and 56 ($M = 28.54$, $SD = 10.79$) was analysed. They got no reward for participating the study. There were no special selection criteria or limitations to participate the study and the participants were not told about the purpose of the study.

2.2 Stimuli

The users were presented a total of 68 different stimuli, 28 items of the pretest and 40 stimuli of the education part. The same 28 items of the pretest were also presented in the posttest. The stimuli design and selection happened by the investigators. The stimuli were images of websites with corresponding web addresses of different types of phishing. Knowledge of the structure of web addresses could help identify phishing messages

(shown in Figure 1). The domain, the here called „who-section" (Wer-Bereich) is the most important part for detecting phishing URLs.

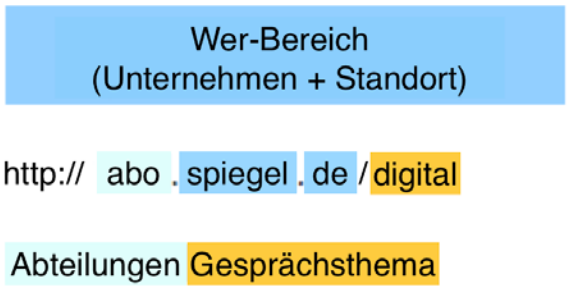

Fig. 1: Structure of a web address in "NoPhish"

In the study the following 7 types of phishing attacks were investigated:

- phishing type 1: URLs based on IP addresses (e.g. https://87.122.24.91/ws)
- phishing type 2: URLs where the who-section doesn't contain the company name (e.g. https://www.hfkjt.com)
- phishing type 3: URLs which contain the company name at the place of the department (Abteilung) (e.g. https://www.instagram.account.com)
- phishing type 4: URLs which contain the company name at the place of the topic of conversation (Gesprächsthema) (e.g. https://www.account.com/t-online.de/settings)
- phishing type 5: URLs where the who-section seems similar to the real URL but use an additional term (e.g. https://www.bahn-support.de)
- phishing type 6: URLs which contain typos (https://www.facebok.com)
- phishing type 7: URLs which contain similar looking letters and numbers like the real ones (e.g. https://www.clropbox.com).

The pretest has 14 faked items (2 items per phishing type) and 14 legitimate items. The same 28 items of the pretest are used for the posttest afterwards. The education part covers 8 different levels. Level 1 imparts knowledge of fundamental concepts to detect phishing attacks. Therefore, Level 1 doesn't contain any specific phishing addresses. Level 2 to Level 8 cover the above-named phishing types (1) to (7). Thereby each level highlights

one specific type of phishing, each with 6 corresponding URLs. The education part used only URLs without images of webpages. Every user got the same URLs in the same order.

2.3 Procedure

The experiment took about 45 minutes per person and was performed in 4 phases: (1) registration and personal data, (2) pretest, (3) education phase and (4) posttest.

Phase 1 – Registration and personal data. A first introduction about the experiment was shown to the participants on the welcome page of the web application. Furthermore, a definition of phishing and the following steps of the task were given. At first a pretest had to be done to classify the knowledge of every user, afterwards several levels to develop or strengthen the knowledge about phishing and finally a posttest for testing the learned skills. Furthermore, there were buttons to register or to login after a break. By clicking on the register-button the participant had to submit personal data like login credentials (e-mail-address and password), year of birth and gender. To provide a realistic experience the participants had to choose their frequently-used internet browser (Apple Safari, Google Chrome, Internet Explorer, Mozilla Firefox). According to the user's choice the images and address bars were personalised. In Mozilla Firefox, Internet Explorer and Google Chrome the important part of the web address is highlighted. This is not the case in Apple Safari. After submitting their data, the participants got to the pretest.

Phase 2 – Pretest. At the pretest every user was presented 28 images of a webpage with corresponding URLs. The participant's task was to identify if the shown webpage is a legit webpage or a phishing attack. Afterwards they had to state their confidence in their evaluation (on a scale of "very unsure", "unsure", "medium", "sure", "very sure"). After finishing the pretest the result was shown to the user (how many right answers they got).

Phase 3 – Education phase. After the pretest the training phase started. The participant got to an overview page of all levels, starting with a short introduction to the topic. After finishing the introduction, the task was to complete all levels. All levels are built up in the same way and cover one specific type of phishing. Every level consists of two parts, a theoretical part and a following practical task. The theoretical part explains the current phishing attack and how to detect it. After finishing the theoretical part, the user had to use his new learned skills in the practical task. An URL was presented and the participant had to identify if the URL is legit or not. After submitting the decision, the result was shown to the user. If it was a phishing URL, the following task was to highlight the who-section. Level 1 starts with the basics of a web address, the here called "who-section", the domain. Level 2 to Level 8 cover the different phishing attacks 1 – 7. At the beginning of every level the user got an overview of the so far learned phishing attacks. After finishing the practical task, the participants were shown how many questions they answered right and the following level was unlocked. The finished level was now locked so that the participants didn't have another chance to answer the questions again to get better results. After completing all levels there was an additional part, the concluding remarks which gave a short summary about phishing tricks which were not given attention to in the levels

of the web application.

Phase 4 – Posttest. After completing the education phase the participant finally had to pass the posttest. The proceeding was analogue to that in the pretest, as well the 28 items. Having finished the posttest, the participant was shown the performance for a last time.

3 Results

The data was analysed to determine if there are any differences in the test scores between the two times of pretest and posttest caused by the treatment. Because there are two times of measurements, multivariate techniques were used which allow multiple dependent variables. A within-subject multivariate analysis of variance (MANOVA) was done with 7 independent variables and 2 dependent variables. Pretest and posttest score were measured as dependent variables and phishing types 1 – 7 were measured as independent variables. Comparing the test score means in Figure 2, phishing type 5 and 6 got the worst results in the pretest and also in the posttest but with the fact that phishing type 6 documents a major increase than phishing type 5.

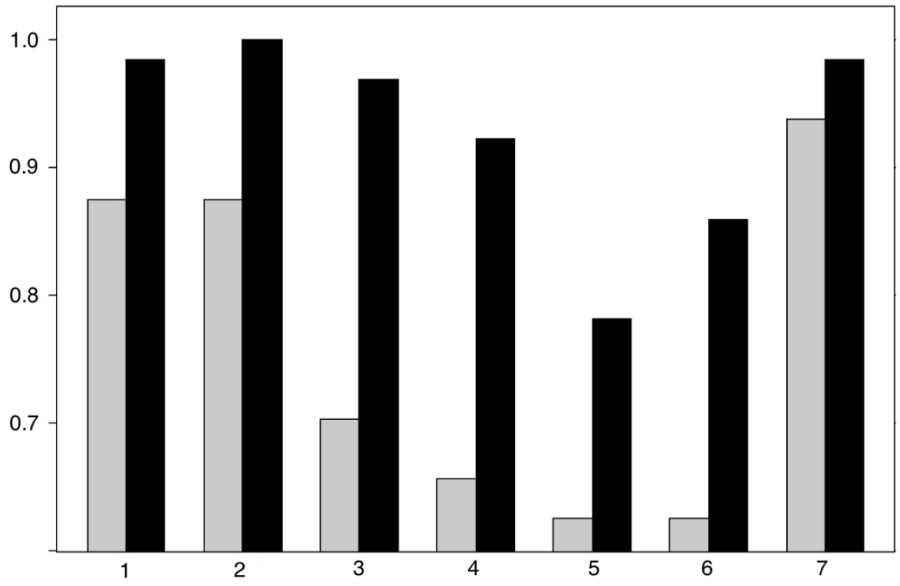

Fig. 2: Test score means of pretest (grey) and posttest (black), split up in the 7 types of phishing

Phishing type 7 reached the highest result in the pretest while phishing type 2 got the highest result in the posttest closely followed by phishing types 1 and 7. A total mean of all pretest scores was reached by a value of .74 (only test scores in a range between 0 and

1 are possible) and a total mean of all posttest scores was reached by a value of .93. Within the MANOVA the 4 following multivariate tests were done: (1) Wilks's lambda distribution, (2) Pillai's Trace Test, (3) the Lawley-Hotelling trace and (4) Roy's largest root. All of the 4 tests are based on the two matrices of "sum of squares" and "cross product". The 4 tests got significant with $F(7, 56) = 3.47$ and $p = .0037$ on a 99% level of significance. Conducting a series of follow-up ANOVAs of the dependent variables, split up in the 7 categories of phishing attacks, the values showed 6 of 7 significant results (Table 1). All of the results were significant except from phishing type 7.

phishing type	p-value	significance
1	.042	significant
2	.0071	significant
3	.000027	significant
4	.0029	significant
5	.0055	significant
6	.012	significant
7	.25	not significant

Tab. 1: Results of significance of ANOVAs

Following the principal analysis, a Welch two sample t-test was performed for testing the mean scores of pretest and posttest (of the particular phishing types) for significance. All tests were done on a 95% level of significance.

phishing type	p-value	significance	pretest score	posttest score
1	.04	significant	.88	.98
2	.009	significant	.88	1
3	.00005	significant	.70	.97
4	.003	significant	.66	.92
5	.006	significant	.50	.78
6	.01	significant	.62	.86
7	.30	not significant	.94	.98

Tab. 2: Results of Welch two sample t-test

In summary, 6 significant results were found (phishing types 1 – 6) and 1 non-significant result (phishing type 7). Furthermore, the correlations between the dependent variables were investigated, split up in the 7 categories of phishing attacks. The lowest correlation was shown between type 1 and type 6 with $r = .017$ and between type 1 and type 7 with $r = .034$. On the other hand, the highest correlation was shown between type 2 and type 3 with $r = .66$, between type 3 and type 4 with $r = .65$ and between type 4 and type 5 with $r = .52$. In general, only positive correlations appeared, varying around a mean of $r = .42$.

4 Discussion

The task of this research was to determine the efficiency of the online tool developed before. The question is whether the measurements of the test scores change over time. Non-constant measurements of the test scores would indicate influence of the training material. A MANOVA analysis of the data showed that there is a significant difference between the multivariate measurements of test scores at the 99% level of significance. Looking at the differences of individual measurements between the different types of phishing, type 1 – 6 got significant and type 7 got not significant. An overall analysis of combined pretest and posttest scores of each phishing attack showed that phishing types 1, 2 and 7 were the most often correctly detected attacks. Phishing types 5 and 6 were the phishing attacks where the most mistakes were done. Comparing the test score means, phishing types 5 and 6 got the worst results in the pretest and posttest but with the fact that phishing type 6 documents a major increase in the posttest than phishing type 5. Phishing type 7 reached the highest result in the pretest while phishing type 2 got the highest result in the posttest closely followed by phishing types 1 and 7.

The above listed results show significantly changes in the different test scores which indicate influence of the training material. Posttest scores and the results of the belonging t-tests proof that users make significantly less mistakes when they have to decide whether a website is faked or not than before using the online tool. Furthermore, it was the goal to analyse which phishing strategies are effective in fooling users and which are still effective after educating users with teaching material. Results show that especially phishing types 5 (URLs where the who-section seems similar to the real URL but use an additional term e.g. https://www.bahn-support.de) and 6 (URLs which contain typos https://www.facebok.com) were difficult to identify as phishing attacks. Regarding the collected data in total, the participants performed worst at phishing types 5 and 6, even in the posttest. This indicates that theses types have to be considered seriously because they present the weakest points in users' attention. With regard to the research of Dhamija et. al (2006), nearly the same outcomes for phishing type 5 were found. About half of the presented URLs could not be identified as phishing attempts. A possible explanation for the bad performance regarding type 5 could be that users do not really know the exact URLs of their visited websites because they probably do not pay much attention to the important parts like the address bar, demonstrated in the study of Dhamija et. al (2006). A possible explanation for the bad performance at type 6 could be that URLs with typos look too similar to the real ones so that detecting the difference is rather unlikely by only having a short look at the URL. So, in future development of phishing education material special attention has to be drawn to these types of phishing. On the other hand, the posttest score means show high detecting results in phishing attacks 1, 2 and 7 which indicates that these types can be practised very well. Due to the fact that phishing type 7 got not significant in the univariate analysis and the t-test, it can be argued that type 7 (URLs which contain similar looking letters and numbers like the real ones e.g. https://www.clropbox.com) has a high success rate in detecting phishing URLs but additional training has no effect on the result. Because of the high scores in the pretest and also in the posttest, there is evidence

that this phishing attack is easy to discover and that this one is no effective technique to fool users. For now, there is no explanation why people are able to detect URLs with similar looking letters but on the other hand are not able to detect typos effectively. Interestingly, the study of Dhamija et. al (2006) shows completely opposed results in detecting phishing URLs with similar looking letters. With a success rate of about 9%, their results are in high contrast to this research with a success rate of about 94%. A possible explanation could be the different presentations of the URLs and corresponding websites. In this research only static images of websites were used, where Dhamija et. al (2006) presented websites with full functionality. This functionality, which made it possible for the participants to explore the website, its corresponding subpages and especially the changing might have influenced the participants' focus with regard to the important part (the address bar) for detecting phishing URLs. So, in summary only phishing types 1 and 2 can be practised well for detecting phishing URLs. The goal of this online tool is to educate users' security awareness and ability in detecting phishing URLs. Based on the total pretest score mean of all questions with .74 and posttest score mean of .93 a significant improvement can be proven. According to the research of Sheng et. al (2007) it could be confirmed that games represent a highly efficient way of information transfer.

A fact that also has to be mentioned is the decrease of participating users. 65 users signed up at our online tool but only 32 of them performed the training and the tests until the end. A possible explanation for this huge decrease is probably that people only deal with something when they are really interested in or when they are forced to. Due to the fact that one run through the tool including pretest, training phase and posttest took about 45 minutes probably a lot of a participants quit the game because they did not want to put so much time in it and unfortunately neither in further education. Another aspect which has to be considered are the correlations between different types of phishing. There were strong correlations between phishing types 2 and 3, 3 and 4, 4 and 5 which indicates that these types of phishing could influence each other, so that maybe people could be able to identify phishing URLs without the need of additional different training material.

Due to the fact that people are not regularly confronted with phishing attacks, Canova et. al (2014) analysed the effectiveness on users' knowledge retention. The results of their studies show that users of the "NoPhish" Android app are more successful in detecting phishing-URLs, particular over a longer period. Another interesting aspect for further research could be analysing the tool's effectiveness on user's knowledge retention in particular for every of the above named types of phishing. For future work, it could be also interesting to what extent the conceptual design turns out to be suitable for different types of user groups. Within the study, only the age of the participants was collected. Improved results in the posttest could be found in the age group between 50 and 60 (5 usable results) like in the age group between 20 and 40 (27 usable results). Due to insufficient available data, there cannot be provided any information regarding how far the training presents an adequate method for specific age groups.

An aspect of this study which wasn't considered are the different types of internet browsers. According to the user's choice the images and address bars were personalised. In Mozilla Firefox, Internet Explorer and Google Chrome the important part of the web address is highlighted. This is not the case in Apple Safari. It was not considered which consequence the user's choice had. By using Apple Safari, the important part is not highlighted therefore identifying the who-section was an additional challenge for these users. These difficulties could have influenced the results and could have produced more low values in the graphics. For further research, this has to be taken into account and possibly researched in an additional study. Another aspect, which was not considered, was the existing knowledge of the users. Beforehand, the participants were not asked about their knowledge of phishing. Due to this fact, test scores could have turned out better than they really are. For now, it is not possible to act on the assumption that some kind of phishing attacks is easy to detect because the high results could be also explained by the users' knowledge. But the most important point, regarding the alarmingly high statistics of APWG, is still minimizing the risk of falling for phishing attacks. Therefore, the users themselves have to be educated. The goal of this web application was to increase users' security awareness and the ability of detecting phishing URLs. This could be significantly achieved.

References

[Ant15] Anti-Phishing Working Group (2015). Phishing Activity Trends Report. 4th Quarter 2014. Access at https://docs.apwg.org/reports/apwg_trends_report_q4_2014.pdf, 25.03.2016

[Ca14] Canova, G., Volkamer, M., Bergmann, C., & Borza, R. (2014). NoPhish: an anti-phishing education app. In Security and Trust Management (pp. 188-192). Springer International Publishing.

[DTH06] Dhamija, R., Tygar, J. D., & Hearst, M. (2006, April). Why phishing works. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 581-590). ACM.

[DCF07] Dodge, R. C., Carver, C., & Ferguson, A. J. (2007). Phishing for user security awareness. Computers & Security, 26(1), 73-80.

[ECH08] Egelman, S., Cranor, L. F., & Hong, J. (2008, April). You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1065-1074). ACM.

[Ku09] Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M. A., & Pham, T. (2009, July). School of phish: a real-world evaluation of anti-phishing training. In Proceedings of the 5th Symposium on Usable Privacy and Security (p. 3). ACM.

[Sh07] Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2007, July). Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In Proceedings of the 3rd symposium on Usable privacy and security (pp. 88-99). ACM.