Measuring the Contribution of Novices in Penetration Testing

Rebecca Balebako, Akhil Shah, Kenneth Kuhn, Sherban Drulea, Christopher Skeels, Lara Schmidt
RAND Corporation
1776 Main Street
Santa Monica, CA 90407
{Balebako, ashah, kkuhn, sdrulea, cskeels, laras}@rand.org

1. INTRODUCTION

There is a large and growing demand for cyber security experts. Demand outstrips supply. The problem is particularly acute at Federal agencies that currently lack the flexibility to attract and retain top talent. At the same time, there are many people who have an understanding of cyber security and would value an opportunity to enter the field. We are interested in determining the feasibility of marshaling a mixed-expertise cyber cadre using technologies that enable an on demand surge of resources, both machine and human, for cyber operations.

It's important to consider the value novices and non-experts offer to a cyber operation. If novices offer no value, then it would likely be best to stick to the traditional model of contracting with small, specialized cybersecurity firms. If, on the other hand, novices can add value, crowd-based models may be an option.

2. OVERALL METHOD AND GOALS

In an attempt to quantify the value of a novice in a mixed-expertise crowd sourced cyber operation, we have chosen web-application penetration testing as an exemplar operation. Penetration testing involves looking for vulnerabilities in web applications. There exist both tools to help novices learn how to do penetration testing and relevant task lists written by thought leaders. We thus hypothesized that there may be benefits to having a mixed expertise crowd participate in a penetration testing campaign as well as an efficient way to split up the work of such a campaign. The recent growth of commercial entities in this space, for example BugCrowd^{TM}, encourages us[1].

The overall goal of this work is to examine two research questions:

1. What is the value of a non-expert in a crowd-sourced penetration testing campaign?
2. What instrument or scale can be used to measure participants' level of expertise?

To test the value that novices offer in penetration testing, we propose a problem-solving exercise in which a mixed expertise crowd is asked to find vulnerabilities in a system that we have created. The setup will allow an experimenter to create a design of experiment matrix and then to track, in real time, true and false positive findings by experiment participant. We can measure the quality of vulnerabilities reported, allowing us to quantify the value of each participant. In the poster, we will describe this method with the goal of eliciting feedback from the usable security community.

We needed a method for determining whether our participants were experts or novices, allowing us to compare expertise to their bug report scores. Novices here include people with limited prior cyber security work experience, people who would not typically work on cyber security at a Federal agency. We would like to know how many, and what type of, vulnerabilities different penetration testers will point out, and are particularly interested in the contributions of novices, as defined above. While previous research has identified some scales that can be used to measure participants' intentions to have good security behaviors [1] or some methods to identify experts in browser security [2], no work had been done to determine how to identify the level of expertise in penetration testing.

In this poster, we describe our method to test the value of non-experts in crowd-sourced penetration testing, and the results of a pre-study on developing an instrument to measure participants' level of expertise.

3. DESIGN OF INSTRUMENT

We developed an instrument to determine a participants' penetration-testing expertise. In a pilot study we examined only the instrument, and not the problem-solving component of the experiment. The on-line questionnaire was designed to be completed in 5-15 minutes, so that it can easily be included in the longer experiment without putting too heavy of a burden on participants. Our instrument has four components: a self-report survey of expertise, a self-report survey of previous experience, knowledge questions, and a scale previously designed to measure behavioral intention.

Our instrument is divided into five components: 1) Self-report of expertise, 2) Self-report of pen-testing experience, 3) Knowledge questions, 4) Security Behavior Intention Scale (SeBIS) [1], and 5) Demographics. Demographic questions covering items such as age and gender are not used to evaluate expertise, but are included to help us control for confounding variables.

[1] http://www.bugcrowd.com

Self-report of expertise questions allowed participants to evaluate their own level of expertise in security and in penetration tasting. Participants evaluated themselves using a 5-point Likert scale. Self-report of expertise in general can be inaccurate when participants do not have a good sense of range of expertise possible, or if the participants are overconfident or lack confidence. The section on self-report of experience included questions about security coursework, programming languages used, and types of systems pen-tested. Self-report of expertise can be inaccurate due to participants' memory, participants' desire to provide the right answer, or if the questions are not well understood or formulated.

The knowledge section of this survey was divided into two sections: 5 multiple-choice questions, and 3 open-ended text questions about system attacks. Good expertise questions require identifying knowledge that is needed and formulating the questions in a way to evaluate the correctness of the answers. We used multiple-choice questions that were based on those used in certification for penetration testing, such as Certified Penetration Testing Engineer (CPTE). Questions included recognizing protocols that use encryption. However, certifications remain controversial, and not all domain-experts feel they represent expertise. Furthermore, our list of questions was very short; this limits our ability to assess the full scope of any participants' knowledge.

4. RESULTS FROM PRE-TESTING THE EXPERTISE INSTRUMENT

We piloted an instrument to measure penetration testing expertise. Our goal was not to find a binary result (is or is not an expert), but rather to obtain data that would help us place a participant within the range of expertise possible.

We paid Amazon MTurkers one dollar to participate. All other participation was voluntary. We tried to incentivize participation by offering a score at the end of the survey, with a brief description of what it meant. It is not clear whether this was successful.

We piloted the instrument questions on three groups of participants that were recruited based on their level of expertise. We recruited three groups of participants: Amazon MTurkers, whom we expected to skew towards novice (n=95), participants in on-line security forums whom we expected to have some domain knowledge, (n=17), and members of a security mailing list at a prominent university who we expected to skew towards expert (n=6). Our mturker participant group was 47% female, 17% of the forum participants were female, and none of the security researchers were female.

We did find different levels of self-reported penetration expertise in each recruitment group. In most components of the instrument, we found that recruitment groups responded in predictably different ways, in that the novices did not do as well as the experts or mid-level participants. In particular, self-report of penetration testing expertise and correct responses to open text knowledge questions were more often correct from the researchers. For example, on mturkers got none of the open-text answers correct on average (median=.5), forum participants correctly responded to 2.4 (median = 2), while the researchers responded to 2.67 (median =3) (ANOVA, p<.001).

We found some components of our instrument were not well differentiated between participants. For example, the SeBIS scale is not intended to measure security expertise, and we find that it does not (ANOVA, p=.15). We can remove this scale from future aspects of the study.

When including all participant groups, we found there was a significant correlation between the open-text knowledge questions and the multiple-choice knowledge questions (ANOVA p<.001). However, when examining the MTurk recruitment group only, we did not find a correlation. We hypothesize that the few MTurkers who answered the questions correctly were looking up the answers on-line (despite instructions not to) or guessing.

The results on self-report of experience were heavily skewed by participant group; while some forum participants and a few MTurkers reported some programming language knowledge, the security researchers were much more likely to have taken several security courses, and more likely to report having pen-tested a system. This may be a good metric for identifying those at the expert end of the spectrum, but may not be a good metric for identifying potential new pen-testers or good problem solvers who are nonetheless novices.

Overall, the results suggest that the two best components for measuring pen-testing expertise are the open-text questions and the self-report of pen-testing expertise.

5. SUMMARY

Substantial benefits may accrue to entities who are able to marshal a mixed-expertise force of for security testing. We have chosen penetrating testing as an exemplar of a cyber operation. We have developed a tool that allows a user to set up an experiment through which the user can track the activities of volunteers that participate in the experiment. We have also developed an instrument that a user can employ to evaluate the expertise level of volunteers. The tool and survey, taken together, should allow us to ascertain the value added by novices participating in a specific cyber operation.

6. REFERENCES

[1] Camp, J., Kelley, T ., Rajivan, P. 2015. Instrument for Measuring Computing and Security Expertise. Technical Report. Indiana University.

[2] Egelman, S. and Peer, E. 2015. Scaling the security wall: developing a Security Behavior Intentions Scale (SeBIS). In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). ACM Press, New York, NY, 2873-2882. DOI= http://doi.acm.org/10.1145/2702123.2702249