Building Castles in Quicksand: Blueprint for a Crowdsourced Study

Arne Renkema-Padmos
CASED / TU Darmstadt
Darmstadt, Germany
arne.renkema-padmos@cased.de

Melanie Volkamer
CASED / TU Darmstadt
Darmstadt, Germany
melanie.volkamer@cased.de

Karen Renaud
School of Computing Science
University of Glasgow
Glasgow, UK
karen.renaud@glasgow.co.uk

CHI 2014, April 26 - May 01 2014, Toronto, ON, Canada

http://dx.doi.org/10.1145/2559206.2578861

Abstract

Finding participants for experiments has always been a challenge. As technology advanced, running experiments online became a viable way to carry out research that did not require anything more than a personal computer. The natural next step in this progression emerged as crowdsourcing became an option. We report on our experience of joining this new wave of practice, and the difficulties and challenges we encountered when crowdsourcing a study. This led us to re-evaluate the validity of crowdsourced research. We report our findings, and conclude with guidelines for crowdsourced experiments.

Author Keywords

Study; Crowdsourcing; Participants

ACM Classification Keywords

H.1.2 [User/Machine Systems]: Human Factors

General Terms

Design; Human Factors; Experiment

Sophie: Calcifer! Calcifer! Are you the one moving the castle?
Calcifer: Of course I am. No one else does any work around here.
— Howl's Moving Castle, Miyazaki Hayao

Introduction

Crowdsourcing platforms are increasingly popular in facilitating usable security research. In general, they provide an online service that allows requesters to post jobs for workers to complete. At SOUPS (2013), 6 of 15 papers used crowdsourcing. This trend is likely driven by speed, availability, and cost. Given its ubiquity and perceived advantages, we also wanted to exploit the potential of crowdsourcing based research.

Our study's focus was the possibility of encouraging security by emphasising salience through simplification, instead of requiring users to search for new security signals. We hypothesised that URL pruning would increase detection of phishing websites, based on theories of cognitive overload [26] and deceptive mimicry [7]. URL pruning was proposed in [18] and is partially implemented in iOS7 Safari. To evaluate the hypothesis, we adapted the experiment reported in [14] on URL highlighting to enable crowdsourced deployment.

We implemented our study using a SurveyMonkey survey. Our aim was to recruit Mechanical Turk (MTurk) participants via CrowdFlower. We report on the problems we identified during deployment, and what they mean in terms of validity of the results. Finally, we identify possible improvements to methodology for studies of this kind. Our contributions are:

- identifying bugs in a major crowdsourcing platform, and
- synthesising guidelines for replicable and valid crowdsourced studies in HCISec.

Experimental Setup

To study URL pruning (i.e. displaying phisher.com instead of amazon.com.phisher.com/login) we adapted Lin et al's lab study on URL highlighting [14] to the CrowdFlower crowdsourcing platform. Use of CrowdFlower is common where MTurk is geoblocked [17]. MTurk is accessible over CrowdFlower as a channel partner.

Lin et al's lab study was extended with a comparison group, we used multiple-choice and open questions instead of in-person interviews, and we showed screenshots instead of viewing a live web browser. Full URLs were tested against base domains. Correspondingly, there were two groups.

In our study, participants in both groups were asked for password disclosure intent on 16 randomised screenshots containing 8 modified (phishing) and 8 unchanged (authentic) URLs. They were asked to verbalise in open text fields what they based their decisions on. In the second part of the study, they were subsequently directed to the URL bar. Again, they were asked to provide a rating of the 16 screenshots and rationale for the rating. In a third part of the study, they were asked to provide the URL that they associated with well-known logos. Finally, we collected some demographic data. We performed a debrief afterwards. A survey code was provided at the end so that participants could verify completion and apply for payment from the platform.

Screenshots were made using Google Chrome 7 for Windows, a popular browser [25]. Target organisations and URLs were adapted for a US audience as Lin et al conducted their study in Canada. Screenshots were taken on a fresh browser over a US based proxy. Attack types (modified URLs) were slightly adapted from the previous

paper in order to to be more representative. Our cover story was unchanged: a study into the effectiveness of trust indicators in web browsers. Two surveys were chained to support within phase randomisation with SurveyMonkey.

Informed consent was obtained via the "job" posting on the crowdsourcing website. The target group at CrowdFLower was participants from the MTurk platform. Participants visiting the link were randomly assigned to one of two groups (i.e. one of two SurveyMonkey surveys) based on a URL displayed via JavaScript.

Running the Study

The study commenced on 2 November 2013, and was monitored live by the first author. As the study was being monitored, a sudden spike in payouts was detected, indicating spamming of the study. There was a mismatch between the number of people claiming payout and the number of people completing the survey: participants were registering completion of the CrowdFlower "job" without actually filling in the SurveyMonkey survey. After contacting customer support, the source of the problem was identified. The study was halted by CrowdFlower customer support to prevent further fraudulent claims for non-completed jobs.

One problem was a bug in the crowdsourcing platform, causing the study to be listed in the wrong channels. We configured CrowdFlower such that only people from the MTurk channel ought to take part. Preliminary indicator of trouble were the absence of our study in the MTurk job listing, and a question from a user who did not know what a "Worker ID" was (an MTurk number). CrowdFlower customer support found that 'BitCoin' channels (bitcoinget, neodev, coinworker, tremorgames) were active. Another issue was missing verification of SurveyMonkey's completion token, due to outdated documentation from CrowdFlower.

Analysis

In reality, no MTurk members participated according to the CrowdFlower logs. After the survey was terminated, a post-mortem analysis identified various further problems.

After the experiment was terminated we decided that we could not analyse the data since the study had been seriously compromised and the resulting data was worthless. We did feel, however, that this was an opportunity to learn from the experience. It did not seem feasible merely to abandon crowdsourcing.

Many studies have been carried out this way and the fraud we experienced might be a new trend, or due to some flaw in the way we set up our study. Moreover, we felt it would be beneficial to other researchers if we captured the lessons we had learned so that other crowdsourcing studies could avoid the pitfalls and emerge with valid data. We argue that these issues are generalisable to other platforms, and are not restricted to CrowdFlower, as the technology on which crowdsourcing platforms are built is similar, all having the possibility of bugs.

Data from the CrowdFlower logs, including IP addresses and browser IDs, indicates one particular participant submitted 6 attempts. Others entered suspicious URLs like google.ro and bankofamerica.co.jp in the SurveyMonkey survey. This is suspicious because CrowdFlower was set to exclude non-U.S. participants, but the URLs are clearly international. Additionally, some people filled in all questions but not the confirmation code.

Submissions from the same IP address were found in both the control and intervention group, with many submissions coming from the same IP address subnet with the same browser ID (user agent string). This indicates the presence of bot activity. Looking at the style of answers in the surveys, there is clear evidence of bot activity.

The attentiveness question (a trick question to ensure participation was genuine) was not answered properly by many people. Only around 20% of participants correctly answered it, while many others appear to have answered the rest of the questions properly.

Discussion

The main cause of the failure of our study was that the CrowdFlower platform had bugs. The technical problems with the CrowdFlower platform should not come as a surprise, as almost all software has bugs. It may have been caused by a very large upgrade of the CrowdFlower framework, or by routine improvements. This problem reveals that there can be high-impact problems with crowdsourcing platforms. Such implementation problems can be difficult to detect if evaluation takes place on only one platform. This finding reveals the current reliance on Amazon Mechanical Turk as an unhealthy monoculture.

Moreover, it has to be acknowledged that if one platform cannot be trusted, then there is a need to carry out research on multiple platforms (similar to the approach particle physicists adopt). Although it can be said we have concrete doubts regarding one crowdsoucring platform, there is no fundamental reason that other crowdsourcing platforms will not have similar problems.

A large contributor, in the case of our study, was that the incorrect markup of the response code check enabled spammers to claim payment without actually taking part in the study. The reason for this problem was outdated documentation. Stability of a platform can be checked by running the same code in multiple studies. This hints at the importance of sharing code used in a study, so that other scientists can perform double checks and replicate studies.

We found other issues as well that impact on the appropriateness of crowdsourcing for the running of studies: There are indications that people answering the questions are not actually from the US, but using proxy servers in the US. Some of the open questions asking for the URL of Google got a response of google.ro (for Romania), and bankofamerica.co.jp (for Japan, although the domain doesn't actually exist).

Limitations

There are various limitations of our study that have to be acknowledged:

- People may have been warned to look out for phishing, as CrowdFlower was, at the time, running an awareness campaign warning potential participants about fraudulent jobs. This shouldn't impact our findings too much, because we were very open about what we wanted to find in our job description.
- The exclusion criteria based on the attentiveness question may well have influenced the results. This attention check might have caused the data to be biased towards greater attentiveness. As such, it is unclear whether people ought to be excluded based on this test.

- We are not sure whether participants read the questions, understood them, and answered them honestly since we did not control the environment.
- Due to SurveyMonkey, there was a limitation with respect to image size, and, as such, compression artifacts were present, and scrolling in the browser was needed as the dimensions of the image were large.
- URL highlighting wasn't applied properly, as the complete domain name (our study), and not the base domain name (default setting) was highlighted in Chrome.
- In order to ensure that participants were actually engaging with the study, we could decide to exclude everyone who did not spot or understand the "golden question". Unfortunately this will also exclude genuine participators who missed the question due to a momentary lack of attention. This would probably skew the results. It is unclear whether and what kind of bias would result from selection methods that select respondents based on what is clearly a poor approximation of a real attentiveness test.
- The final window gave the participants the code they could use to claim payment. Unfortunately, some participants closed the final window accidentally, and thus could not claim payment, even though they had conscientiously participated in the experiment. It is unclear how such participants can be dealt with.
- Various papers have attempted to verify the representativeness of crowdsourced samples [6], and others have attempted to replicate findings from the lab [19]. These suffer from continuously changing demographics, as well as from the results of web-specific behaviours. Furthermore, it is unclear how representative crowdsourced samples are for studying higher level issues such as passwords and web browsers. It could reasonably be argued that such topics cannot be studied in an ecologically valid way by means of an online survey which does not replicate a real-life situation, and which does not require any level of genuine engagement from the participant. At a more general level, there is the issue of how representative crowdsourcing workers are with respect to the time they spend online.

Ethics

Besides methodological limitations, there are also ethical considerations that have to be taken into account, which will limit the types of studies that can be run. While running our study, a wide range of ethical questions were raised:

- Can informed consent and debriefing be feasibly obtained, and furthermore, can they be verified?
- Are cover stories advisable in online studies?
- How much demographic information can be acquired, and what are the data protection responsibilities of individuals carrying out studies?
- How much personal information can we store to distinguish bots from humans, and humans from humans? E.g. can we carry out deep Flash cookie mining? See, e.g. [13] for more privacy issue.
- Should we create a shared database of participants that have already participated in our study?
- Is it ethical to use recorded IP addresses to look up geolocation data via a 3rd party? Such information has not been authorised by the participants so do we have a right to obtain it for data verification purposes?
- Is it permissible to check participant activity and to try to verify whether proxies are being used?

- Can raw data be made available for replication purposes without explicit consent? What kind of consent is sufficient? Usually data is destroyed after analysis and only aggregated results made available in publications.

Research Guidelines

This section fits with the literature criticising the methodology of online studies. In the following subsection related work will be mentioned that covers the advantages and disadvantages, as well as recommendations for these kinds of studies, and studies in the broader area of usable security.

There are various sources for gaining advice to support study design and execution in crowdsourcing research. We analysed the following selection of literature to synthesise a set of guidelines: [9, 24, 20, 10, 23, 5, 8, 15, 17, 22]. The sources we consulted dealt with both crowdsourcing and more general methodological issues related to carrying out experiments. The sources were in general agreement with one another, although there was significant disagreement about the permissibility of deception.

Items not directly mentioned by the sources, but which are considered good general practice include:

- Record everything, all details, all assumptions, all decisions;
- Take screenshots as further evidence;
- Open up all your study details to the research community as far as possible, within the constraints imposed by ethical guidelines.

While the abovementioned are hardly novel, they take on a new importance in the era of crowdsourced research.

Aspects mentioned by the sources analysed include:

- The need for open data.
- The importance of open measures, including the code used in experiments. In science we need not just open data, but also open materials.
- Enable other scientists to check results (this supports a “trust but verify” culture).
- Don’t deceive participants, as it is done at the risk of participants losing trust in researchers.
- Ensure proper research ethics are adhered to.

Items specific to crowdsourcing are:

- Always check that your jobs are posted on the correct channel. Monitor the experiment continuously.
- Appropriate research designs for crowdsourced experiments are those where: the population is assumed to not be naïve and where basic cognitive processes are the focus of the study. This way they can be considered generalisable to wider populations.

Our experiences support these views. We can extend these with the following pointers:

- Taking screenshots allowed us to recheck our settings. Anyone running automated experiments should capture and communicate all settings.
- Because many settings are possible, and as this information is very tedious and detailed, a standards-based machine-readable format may be appropriate. It could enable settings to be communicated clearly, version controlled, and dependency checked.
- Study settings should be archived, e.g. through an online repository or attached to a paper.

- Studies can serve as a template for non-scientists to test their own work.
- Use multiple (independent) platforms (note that this may also help to check whether findings transfer across populations).
- Include questions to check that people come from where your supplier claims they originated from.
- Require all versions of the experiment to be published including the results.

For all these recommendations, a feasibility analysis is in order. E.g. see the failure of NSPW's policy in encouraging methodological rigour in papers [16]. Switching cold-turkey will most likely not work, and, as such, a gradual adoption roadmap is in order. The following question is thus of utmost importance: "What are the economic and technical incentives to do any of this?"

Conclusions

Crowdsourced studies have their supporters, e.g. [15], and detractors, e.g. [4]. In the end it is just a research tool with advantages and disadvantages. Bickering about fitness for "research" does not get us anywhere; looking at the pros and cons and periodically re-evaluating these does.

Crowdsourcing is an addition to, not a replacement of, traditional research methods for general research topics. It can help gather additional input alongside more traditional methods. Furthermore, it can help get research results where there would be none due to budgetary constraints.

Research methods differ by research area and topic, and the parsimonious nature of security research doesn't help in bringing clarity to methodological discussions.

While we have provided preliminary research guidelines, there are still many unanswered questions, including:

- Can experiments designed by scientists be turned into point-and-click templates that can be reused by others, especially within and across crowdsourcing platforms, necessitating interoperability?
- What tools are currently used for running experiments online, and do they support the proposed research guidelines?
- What crowdsourced research has been done so far, and how does it hold up to the research principles?
- What are the policies of conferences and journals with regards to research requirements, and how can these be changed for the better?
- Do we have an ethical responsibility for involving certain classes of users?
- What insights might be gained from looking at research from a security angle? E.g. are ideas such as assumed brokenness and 'multi-channel studies' relevant? [11, 12]
- How is and should access to research platforms be arranged? E.g. what is the impact of geoblocks on reproducibility?
- How many studies and conferences apply the guidelines presented?
- What types of experiments are a good fit for crowdsourcing studies?
- What are the possibilities and pitfalls for performing interviews through crowdsourced platforms?
- When to use and when not to use crowdsourcing for running studies?
- How can geoblocked services and the need for reproducibility be combined? Can we describe how others can get access?

When seeking answers, keep in mind that other disciplines have been struggling for a long time with these problems. We may just be rediscovering them. While we shouldn't blindly follow other disciplines, we should look at what overlap there is while being aware of our own context. There is much inspiration to be gained from topics such a the politics of big data and genomics, and the operation of large equipment by particle physicists.

Acknowledgements

The authors would like to thank the customer support of CrowdFlower for their timely and expert help in dealing with our issues. We would also like to thank Saul Greenberg for answering questions regarding details of the Lin et al study.

Open Access Research Data

The materials and primary data of the study are available online [21].

Materials included in support of open research are:

- genuine and phishing URL derivation
- screenshots of the websites
- surveys posted on SurveyMonkey
- detailed CrowdFlower settings
- JavaScript code for randomisation
- anonymised CrowdFlower results
- CrowdFlower support emails and documentation

For those that also want to pursue open research, see the Budapest Open Access Initiative[1], the Open Knowledge Foundation [2], and the Public Library of Science [3].

References

[1] Budapest Open Access Initiative. http://www.budapestopenaccessinitiative.org/.

[2] Open Knowledge Foundation. http://okfn.org/.

[3] Public Library of Science. http://www.plos.org/.

[4] Adar, E. Why I hate Mechanical Turk research (and workshops). In Proceedings of the CHI 2011 Workshop on Crowdsourcing and Human Computation (May 2011).

[5] Amazon Mechanical Turk. Requester Best Practices Guide. Amazon Web Services, June 2011.

[6] Buhrmester, M., Kwang, T., and Gosling, S. D. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science 6, 1 (2011), 3–5.

[7] Gambetta, D. Deceptive mimicry in humans. In Perspectives on imitation: From neuroscience to social science, S. Hurley and N. Chater, Eds. MIT Press, Cambridge, 2005, 221–241.

[8] Horton, J. J., Rand, D. G., and Zeckhauser, R. J. The online laboratory: Conducting experiments in a real labor market. Experimental Economics 14, 3 (2011), 399–425.

[9] Kitchenham, B., Pfleeger, S., Pickard, L., Jones, P., Hoaglin, D., El Emam, K., and Rosenberg, J. Preliminary guidelines for empirical research in software engineering. Software Engineering, IEEE Transactions on 28, 8 (2002), 721–734.

[10] Kittur, A., Chi, E. H., and Suh, B. Crowdsourcing user studies with mechanical turk. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08, ACM (2008), 453–456.

[11] Lasecki, W., Kamar, E., and Teevan, J. Raising an army: Attacking crowd systems. CrowdConf (2013).

[12] Lasecki, W., Teevan, J., and Kamar, E. Information extraction and manipulation threats in crowd-powered systems. In Proceedings of the 2014 ACM Conference on Computer Supported Cooperative Work (CSCW 2014), Baltimore (2014).

[13] Lease, M., Hullman, J., Bigham, J. P., Bernstein, M., Kim, J., Lasecki, W., Bakhshi, S., Mitra, T., and Miller, R. Mechanical turk is not anonymous. Social Science Research Network (2013).

[14] Lin, E., Greenberg, S., Trotter, E., Ma, D., and Aycock, J. Does domain highlighting help people identify phishing sites? In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11, ACM (New York, NY, USA, 2011), 2075–2084.

[15] Mason, W., and Suri, S. Conducting behavioral research on Amazon's Mechanical Turk. Behavior Research Methods 44, 1 (2012), 1–23.

[16] Maxion, R. A., Longstaff, T. A., and McHugh, J. Why is there no science in cyber science?: A panel discussion at NSPW 2010. In Proceedings of the 2010 Workshop on New Security Paradigms, NSPW '10, ACM (New York, NY, USA, 2010), 1–6.

[17] Office of Research Ethics. Human participant research guidelines: Use of crowdsourcing services. University of Waterloo, July 2013.

[18] Padmos, A. A case of sesame seeds: Growing and nurturing credentials in the face of mimicry, September 2011. Available at http://digirep.rhul.ac.uk/items/328c3d8b-3695-bfea-03f0-b651ac709211/1/.

[19] Paolacci, G., Chandler, J., and Ipeirotis, P. G. Running experiments on Amazon Mechanical Turk. Judgment and Decision Making 5, 5 (August 2010), 411–419.

[20] Peisert, S., and Bishop, M. How to design computer security experiments. In Fifth World Conference on Information Security Education, L. Futcher and R. Dodge, Eds., vol. 237 of IFIP — International Federation for Information Processing. Springer, 2007, 141–148.

[21] Renkema-Padmos, A., Volkamer, M., and Renaud, K. Building castles in quicksand: Blueprint for a crowdsourced study (materials). figshare (2014). http://dx.doi.org/10.6084/m9.figshare.938239.

[22] Schechter, S. Common pitfalls in writing about security and privacy human subjects experiments, and how to avoid them. Microsoft, January 2013.

[23] Schechter, S. Experimenting on Mechanical Turk: 5 How Tos. Microsoft, July 2009.

[24] Skitka, L. J., and Sargis, E. G. The Internet as psychological laboratory. Annual Review of Psychology 57 (2006), 529–555.

[25] StatCounter Global Stats. Top 5 desktop, tablet, and console browsers in the United States from Dec 2012 to Dec 2013, Jan. 2014. http://gs.statcounter.com/#browser-US-monthly-201212-201312.

[26] Sweller, J. Cognitive load during problem solving: Effects on learning. Cognitive Science 12, 2 (1988), 257–285.