Conducting Usable Privacy & Security Studies with Amazon's Mechanical Turk

Patrick Gage Kelley
School of Computer Science
Carnegie Mellon University
pkelley@cs.cmu.edu

ABSTRACT

Being able to conduct human subjects experiments in a distributed method across the Internet is frequently desirable to support broad tests of usability. Until recently these experiments were commonly advertised in an ad-hoc fashion, using mailing lists, contest sites, and online bulletin boards. Recently Amazon's Mechanical Turk, a service where users can complete short tasks and receive automatic payment, has become prominent in the HCI community. We describe three different usable privacy and security experiments that were conducted through Mechanical Turk, highlighting both reasons for using Amazon's service as well as common pitfalls that we encountered.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Web-based interaction; H.1.2 [**Models and Principles**]: User/Machine Systems—*human factors*

General Terms

Human Factors, Security, Experimentation

Keywords

Mechanical Turk, crowdsourcing, experimental design, usability, privacy

1. INTRODUCTION

Amazon Mechanical Turk (MTurk) began as an internal service to allow employees to spot and mark duplicate items that Amazon offered for sale. In November 2005, the service was made public featuring a more general framework for similar tasks, branded Human Intelligence Tasks or HITs.

The concept of a Human Intelligence Task is directly tied to the Turk, the original Turk was a chess playing "machine," which was in fact an expert human chess player hidden inside of, and controlling the device. MTurk's FAQ states: "Amazon Mechanical Turk is based on the idea that there are still many things that human beings can do much more effectively than computers," just as humans could play chess much more competently than machines in the 1800's [1].

In this paper we present advice on using MTurk as a platform for conducting usable privacy and security (UPS) experiments. In these studies the tasks workers complete are surveys and interactive user studies, tasks that require real users.

The remainder of the paper will detail three of our own UPS experiments on MTurk, followed by an in-depth discussion bringing in related work as well as our own findings in four areas: demographics, avoiding gaming of the system, MTurk process, and experiment design.

2. CASE STUDIES

For each of the following case studies we describe a brief overview of the experiment performed by CyLab Usable Privacy and Security (CUPS) Lab members followed by a description of how the MTurk setup was specialized.

2.1 User's attitudes towards targeted advertising

We completed a series of two MTurk surveys to understand American participants' perceptions and attitudes towards behavioral and targeted advertising. We had approximately 600 participants complete the survey, with drop-out rates (participants who returned the HIT) between 37% and 42%. The survey-design was relatively straightforward and the demographics were in line with those that we discuss below. For more details see MacDonald et al. (forthcoming 2010)

2.2 Privacy labels

We have been refining the design of a "Nutrition Facts" inspired privacy policy format over the last three years. We used MTurk as a compliment to laboratory testing during our design iteration phase, as well as for our large-scale testing. For the design iterations we performed a series of four small-scale comparisons between two or three variants to quickly refine our design. We will speak more to the benefits of quick iteration in our final section on experiment design. For our large-scale test we wanted only native english speakers to test the label and posted a text requirement as well as an error-correction question as explained in our demographic section below. For more details on the study design and results see Kelley et al. [4]

Symposium on Usable Privacy and Security (SOUPS) 2010, July 14–16, 2010, Redmond, WA USA
.

2.3 Phishing Susceptibility

To test phishing susceptibility we created "an online survey that could not be designed in compliance with mere good-faith incentives." Here, with a difficulty in assessing correctness based exclusively on the survey results, we needed a way to verify that responses could be trusted. We created a two question screening task described in full in [3]. For a description of the full phishing survey it compliments see [10].

3. EXPERIMENTS WITH MTURK

3.1 Demographics

MTurk has now reached 200,000 users, the demographics of which have been surveyed at length. A survey from 2008 with over 1000 respondents found that Turkers were younger, had a lower income, were less likely to have children, and were more likely to be female [8].

In April 2010, a longitudinal report on MTurk demographics was published (including the above results), which reported that the male/female ratio had become nearly even, but more importantly showed an increasing number of respondents from India [9]. Turkers from India are likely to be even younger than their U.S. counterparts, have lower incomes, and be more dependent on the income they earn from MTurk. Across the last two years of data, self-reported average hourly wages tend to be just under $2.00/hour. Ross et al. [9] conclude with a description of a shifting common persona of a Turker, from "stay-at- home moms who want to supplement the household income; office workers Turking during their coffee breaks; college students making a few dollars while playing a game; and recession-hit Turkers doing what they can to make ends meet ... [to] young, highly-educated Indian males."

Our own studies saw similar demographic results, however in two cases we focused our recruitment on specific demographics. In the privacy label work we wanted native english speakers. On the HIT acceptance page we explicitly stated said "Only native english speakers are eligible." However, to better understand if this requirement was being followed, three pages into our survey, on the demographics page we then asked participants to specify their native language in a free form text field. 664 respondents (86.9%) reported english, with an additional 52 specified an Indian language and 48 specified other languages (6.8% & 6.3% respectively). Based on these results we saw that a simple text notice seems to have shifted the participant pool from approximately 56% (U.S. according to Ross) to 87%. Our work on attitudes towards targeted advertising used the built in country restriction without a check to confirm.

Overall, the demographics of the MTurk community are becoming better understood, and with focused demographic questions and built in or in survey screening, successful experiments can be conducted on targeted groups.

3.2 Avoiding gaming

Early research work involving the MTurk community, as well as many of the internet guides on creating MTurk HITs focus on avoiding cheaters and gaming of the system. In general, crowdsourcing systems seem prone to gaming as Kittur found in 2008 [5]. Markus Jakobsson describes a series of tricks for creating survey questions that lead to truthful responses due to the question construction [7]. As we mentioned above in our work on privacy labels, although the HIT page expressed a requirement for native english speakers, when we asked on the third page of our survey for participants to enter their native language in a free form text field, just over 12% of our participants entered something that was not English, showing that it was easier for that 12% to just be truthful when faced with a blank field.

Screening tasks are also a tool to fight survey gaming. In the phishing susceptibility work a two question qualification task was used. Participants who failed the the qualification received only $0.20 (but not a rejected HIT, more on that in MTurk process), but those who passed could go on to the full survey to earn $4.00. The two questions used were based on reading the text of an e-mail, with one simple task and one difficult task. 1,198 of 1,962 participants (61%) correctly answered both questions. 1,726 of 1,962 (88%) correctly answered the easy question. From the qualified participants we observed women were more likely to answer the difficult question correctly than men; older participants were more likely to answer the difficult question than younger participants. Additionally, professionals and students were more likely to answer the difficult question correctly compared to hourly workers, financial workers, and other occupations. Finally, time to task completion was not found to be a good indicator of qualification.

However, while methods exist to fight gaming, Kosara and Ziemkiewicz did not see similar gaming effects and believe that using the bonus reward system to financially incentivize correctness in combination with the ability for a requestor to reject a HIT seemed to effectively combat gaming [6].

3.3 MTurk process

In terms of actually creating MTurk HITs, much of the process is straight-forward and details are given in the Amazon MTurk Best Practices Guide [2].

One area that Jakobsson highlights of specific reference to UPS studies is "hiding your motives." For our privacy label work, our HIT description mentioned "information design" and "a policy" but did not mention privacy or security. By not explicitly discussing privacy or security we aim to avoid a selection bias among those accepting the HIT. In general, short descriptive information pages seem to work best in practice. In our targeted advertising work we designed the survey to begin with an essay question so as to discourage Turkers early on in the process, instead of putting more difficult questions late in the survey. This avoids later decisions to complete the task with a series of random answers.

A common concern both inside and outside of the research community is the payment Turkers receive as hourly rates can be quite low. We have had academic reviewers describe these low wages as researchers exploiting the MTurk community. And to some extent this may be an unfortunate truth; the initial demographic studies mentioned above were performed with a payment of only $0.01. In the privacy label work our effective hourly rate ranged from $2.45 to $3.54 across all runs. In both our phishing susceptibility and privacy label work we varied the pay rates in smaller pilot studies to find an affordable payment that seemed correct for the task.

To conduct large-scale tests HITs must frequently be run in several batches. As also reported by Kosara, HITs are accepted most frequently immediately after posting and slow down as they age. While more slots can be added to a currently existing HIT (preventing any Turker duplication) adding these additional slots does not refresh the age of the HIT. As a result for our larger tests, such as the privacy label test of 781 participants we ran six separate batches and then compared Turker IDs after completion. Warnings to not complete a HIT in a similar line of studies seem to be compelling for fear of having a HIT rejected.

Payment is not actually applied till a requestor approves that a Turker adequately completed the task. While Turkers can return tasks without penalty such as when faced with the discouragingly long essay-style questions above, rejections count towards their overall accept/reject ratio. A ratio that by default is set to 90 or 95% when creating a new HIT.

3.4 Experiment design

Kittur et al. describe three challenges to MTurk as a platform for research, one of which is demographic unknowns, which we have already covered [5]. The other two involve designing tasks for successful responses through MTurk. The first involves creating adequately sized, very short tasks which can be distributed across hundreds or thousands of users. Secondly, Kittur recommends that MTurk "is best suited for tasks in which there is a bona fide answer."

In terms of UPS experiments, framing the studies to fit MTurk's model may require a new way of thinking about testing users. While surveys about attitudes and perceptions transfer quite easily, traditional laboratory tests such as role-playing, think-alouds, and focus groups may be more difficult to translate. To assist with this process many experiments after being conducted in the lab on a small number of participants can be rewritten as a series of survey questions or an accuracy-based quiz which can be hosted online.

One additional concern to consider is creating a study design that will be approved by your institutions Institutional Review Board (IRB). At Carnegie Mellon, we have not had issues specifically pertaining to the use of MTurk for user studies.

In each of the above case studies we used external sites to actually host our experiments. Sites such as `surveymonkey.com` or `surveygizmo.com` can be linked to MTurk through the use of unique identifiers, or more simply by requiring participants to enter a short piece of information in both the survey and the MTurk HIT such as a completion code, a timestamp given to the participant at the completion of the survey, or even the last four digits of their telephone number. We have also used our own web application to capture more information such as the amount of time spent reading specific pages, specific button clicks, the amount of scrolling conducted, and we have future plans to record keypresses through javascript to better instrument the surveys we conduct through MTurk.

Finally, while there are certainly a number of limitations that have been discussed above, MTurk should be seen as a tool in the standard toolbox of usability researchers. The ability to run multiple variants of a study at extremely low costs and fast turn-around times make MTurk an incredibly useful proving ground. Each of our studies was first tested with 15-50 users over just the course of a day or two, leading to immediate results and the ability to refine before a complete experiment was launched. MTurk allows efficient digital piloting in a way that was previously unavailable.

4. ACKNOWLEDGMENTS

I would like to thank Lorrie Cranor, Julie Downs, Mandy Holbrook, Aleecia McDonald, Steve Sheng, and all of the other members of the CUPS Lab for their help in preparing this survey. Additionally I gratefully acknowledge support from National Science Foundation grant number 0524189 entitled "Supporting Trust Decisions" and from the CyLab at Carnegie Mellon under grant DAAD19-02-1-0389 from the Army Research Office.

5. REFERENCES

[1] Amazon Mechanical Turk `https://www.mturk.com`.

[2] Amazon Mechanical Turk. Best practices guide, 2010 `http://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf`.

[3] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. Are your participants gaming the system?: screening mechanical turk workers. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pages 2399–2402, New York, NY, USA, 2010. ACM.

[4] P. G. Kelley, L. Cesca, J. Bresee, and L. F. Cranor. Standardizing privacy notices: an online study of the nutrition label approach. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pages 1573–1582, New York, NY, USA, 2010. ACM.

[5] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 453–456, New York, NY, USA, 2008. ACM.

[6] R. Kosara and C. Ziemkiewicz. Do mechanical turks dream of square pie charts? In *BELIV '10: BEyond time and errors: novel evaLuation methods for Information Visualization*, pages 373–382, New York, NY, USA, 2010. ACM.

[7] Markus Jakobsson. Experimenting on mechanical turk: 5 how tos, July 2009 `http://blogs.parc.com/blog/2009/07/experimenting-on-mechanical-turk-5-how-tos/`.

[8] P. Ipeirotis. Turker demographics vs. internet demographics, 2009 `http://behind-the-enemy-lines.blogspot.com/2009/03/turker-demographics-vs-internet.html`.

[9] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI EA '10: Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, pages 2863–2872, New York, NY, USA, 2010. ACM.

[10] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pages 373–382, New York, NY, USA, 2010. ACM.