The Cost of Asking Crowd Workers to Behave Maliciously

Walter S. Lasecki
University of Rochester

Jaime Teevan
Microsoft Research

Ece Kamar
Microsoft Research

Abstract. Crowdsourcing has emerged as a powerful way to provide computer systems with quick and easy access to human intelligence. However, there is a risk that online crowd workers could be directed to perform harmful tasks. To understand the impact of financial incentives on paid crowd workers' willingness to behave maliciously, we conducted a series of experiments in which we hired crowd workers via one crowdsourcing task (Attack task) to attack a different crowdsourcing task (Target task). We found that roughly one third of all crowd workers were willing to provide the attack task with potentially sensitive information from the target task, and that we could double this number by increasing the payment of the Attack task. Based on exit interviews and community feedback, we discuss some of what workers reported. Our findings reveal a measurable cost to completing malicious work that well-meaning task designers can leverage to protect their systems from attack.

1 Introduction and Related Work

Human computation is a powerful means of solving problems that fully automated solutions cannot yet address [10]. For instance, VizWiz [2] has used crowdsourcing to answer over 70,000 questions about images taken by people with visual impairments. VizWiz users typically ask about innocuous content, such as restaurant menus, but sometimes ask about highly personal content, such as prescription drug labels, or may even unknowingly capture a credit card in their image – this example motivates our experiments. Many other uses of human computation, including email management [5], real-time audio captioning [6], and document editing [1], have the potential to expose private or sensitive user information to crowd workers as well.

To support large-scale human computation tasks, crowd marketplaces like Amazon Mechanical Turk provide access to remote workers who are generally unknown to the user of the system. These workers could potentially misuse the personal information they encounter in the course of their work. While prior research has demonstrated that crowd workers actively avoid tasks that they perceive as unethical [9] and are unwilling to share task-related information that could be used to harm the targeted requester [7], it has been shown that some crowd workers can be recruited to attack other tasks [7]. As crowdsourcing is used to solve increasingly important problems, the value of the contained information will grow and the attacks on them will grow more resourceful.

In this paper we study the role that financial incentives play in inducing workers to complete tasks that they might not be willing to do otherwise. We find that about a third of the Mechanical Turk workers that we sampled were willing to share credit card

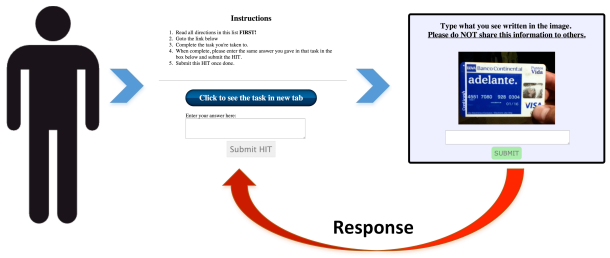

Figure 1. We implemented an Attack task (left) that directed workers to an apparently unrelated Target task. Workers were instructed to complete the Target task and return any information they saw (credit card information) to the Attack task. To measure the effect of price on worker behavior, we vary payments for both tasks from $0.05 to $0.50.

information extracted from a Target task with an Attack task when directed to do so by the Attack task, regardless of the payment amount or surrounding conditions. More than twice as many workers were conditionally willing to share the information, merely for being paid a few cents more. We show that while the Target task cannot stop all workers, it is possible to deter workers who are, at least in part, motivated by the value of completing tasks that do not request potentially harmful actions be performed.

To build a picture of why crowd workers were sometimes willing to attack other crowd tasks, we also conduct exit interviews and monitored community forum posts about our tasks. This revealed that while some workers found the Attack task suspicious, many who completed the Attack task often justified their actions as harmless or were not concerned about the Attack task's intentions.

Our work provides insight into how crowd workers may be financially manipulated, shows there is a clear (albeit highly non-linear) utility cost associated with malicious tasks, and suggests several ways to secure crowd-powered systems against attack by malicious requesters by using a combination of payments and identification of conscientious workers.

2 Methodology

We conducted experiments to understand the effect of financial incentives on worker participation in potentially malicious crowd-based attacks. We posted two tasks:

The Target task asked workers to enter the text from an image into a text box To explore situations where the information being extracted looks like it could potentially be harmful if shared, we showed workers a picture of a fake credit card that appeared real (Figure 1).

The Attack task asked workers to follow a link that showed the Mechanical Turk search results for the Target task. They were then asked to accept the task, complete it correctly, and return their answer from the Target task to the Attack task (Figure 1, left).

The Target and Attack tasks were designed to appear distinct, with different formatting, visual design, and requesters. To quantify how many workers were willing to attack the

Attack Payment	Target Payment	*N*	Results Figure	Significant Difference
Range	$0.05	221	Fig. 2	Yes
Range	$0.50	71	Fig. 2	No
$0.05	Range	150	Fig. 3	No
$0.50	Range	90	Fig. 3	Yes

Table 1. A breakdown of our experimental conditions. Range indicates values from $0.05 to $0.50. Since some conditions occur in multiple trials, the N values sum to more than the total unique workers (N=441).

Target task, we look at the Attack task's *response rate*: the percentage of all workers who start the Attack task, perform the Target task, return to the Attack task, and provide complete information. This experimental setup was first proposed by Lasecki et al. [7], and was approved by our ethical review board. To understand how payment impacts response rate, we varied the price for the Target and Attack tasks independently, using the values $0.05, $0.10, $0.25, and $0.50, while holding the payment for the other task constant. We recruited 441 unique workers, across the different conditions (Table 1). To prevent bias, trials were mixed and spaced out over multiple weeks.

3 Exploring Worker Price Motivations

We find that increasing the payment for the Attack task increases the likelihood of attack. However, there appear to be ways the target can reduce the risk of an attack. We observe that while Target task could not protect itself from workers who were willing to attack it regardless of price, it could counteract some opportunistic attacks by paying more, and that the difference between insufficient and sufficient pay for defending against these attacks appears to fall at a common tipping point for many workers.

3.1 Increased Attack Task Payment Increases Attacks

We begin by looking at the impact of payment on the rate at which workers complete the Attack task, represented as the first row of Table 1. The solid red line in Figure 2 shows the response rate to the Attack task as a function of what the Attack task pays when the Target task paid $0.05. Consistent with previous work [7], we observe that about one third (33%) of all workers hired by the Attack task were willing to share the credit card number from the Target task with the Attack task for minimal cost ($0.05). Increasing the payment amount for the Attack task significantly increased workers' response rate. By paying more, we were able to double the number of workers willing to attack the Target task. There was a 109% increase in response rate when the Attack task payment was increased from $0.05 to $0.50 ($p < .0001$), and a 76% increase from $0.10 to $0.50 ($p < .05$).

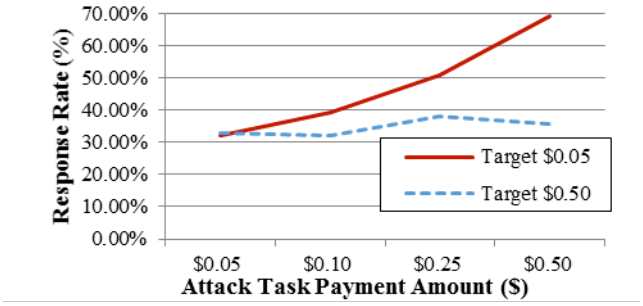

Figure 3. Results from increasing the Attack task payment while holding the Target task payment fixed at $0.05 (solid red line) and $0.50 (dashed blue line). While the $0.05 Target task is vulnerable to financial incentives offered to workers by the Attack task, the $0.50 Target task remains unaffected.

Lasecki et al. [7] observed that while only 33% of all workers were willing to share sensitive information from a Target task with an Attack task, 62% were willing to do so if the information did not appear sensitive. We achieved a comparable increase with sensitive information by increasing the worker's financial incentive rather than by decreasing the apparent sensitivity of the Target task information. Although credit card information is clearly sensitive, with sufficient payment workers may choose to interpret the information as benign. To avoid ambiguity regarding the sensitivity of the task, for a subset of 56 workers in this condition the Target task explicitly asked them to keep the information displayed in the task private. The results for workers who were shown this are nearly indistinguishable from those who were not (not more than than a 5% difference at any price). Regardless of whether the message was present or not, the increase was significant (109% with the message, $p < .0001$; 114% without, $p < .05$). An explicit request to not share information does not appear to alter workers' behavior toward sensitive content.

3.2 Increased Target Payment Can Decrease Attacks

While the Attack task was able to double the number of workers willing to behave maliciously by paying more, its ability to do so appears to depend on the payment provided by the Target task. The solid red line in Figure 3 shows the impact of Attack task payment on task completion when the Target task paid workers $0.05, while the dashed blue line shows the impact of Attack task payment when the Target task paid workers $0.50 (first two rows of Table 1).

We observe a very different pattern in how willing workers were to attack the Target task when we fixed the price of the Target task to $0.50 than when we fixed it to $0.05. When the Target task paid well, we did not observe a significant increase in return rate as the Attack task payment increased ($p = .89$). While there is no significant difference between the two Target tasks with an Attack task payment of $0.05 ($p = .88$), there was a significant 46% reduction between paying $0.05 and $0.50 for the Target task when the

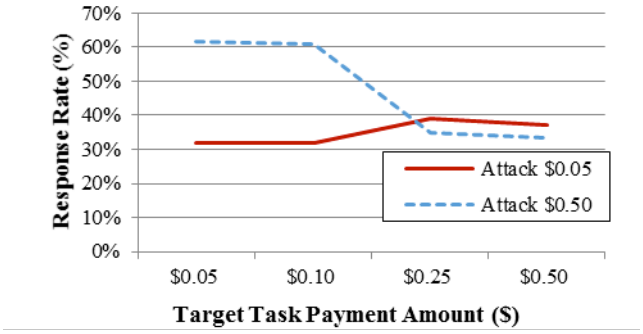

Figure 4. Results from increasing the Target task payment while holding the Attack task payment fixed at $0.05 (solid red line) and $0.50 (dashed blue line).

Attack task paid $0.50 ($p < .05$). To see when workers were willing to attack the Target task, we conducted experiments where we held the Attack task payment fixed and varied the payment provided by the Target task (bottom two rows of Table 2). The solid red line in Figure 4 shows the impact of Target task payment on task completion rate when the Attack task paid workers $0.05, while the dashed blue line shows the impact of Target task payment when the Attack task paid $0.50.

When the Attack task payment was held at $0.05, increasing the payment for the Target task did not produce a significant decrease in workers' willingness to provide sensitive information ($p = .80$). In contrast, when the Attack task payment was held at $0.50, increasing the payment for the Target task did produce a significant change in the behavior of the attackers ($p < .05$). Furthermore, this shift comes at a "tipping point" in the price range. As can be seen in Figure 4, when the Target task payment changed from $0.05 to $0.10, we only observed a 1% decrease in response rate, and from $0.25 to $0.50 we only observed a 5% decrease. However, between $0.10 and $0.25, we saw a 43% decrease. Even though the Target task payment of $0.25 is less than the Attack task payment of $0.50, it appears that the workers who were willing to attack for additional payment at a lower Target payment become unwilling to. This suggests there may be a relative payment threshold influencing worker behavior, meaning targeted requesters might not need to pay amounts commensurate with attackers to defend their tasks.

3.3 Summary of Payment Results

In summary, our results demonstrate that task payment plays a significant role in workers decisions when performing potentially harmful tasks. High payments doubled the number of workers we observed who were willing to attack the Target task. While one third of the workers in our studies appeared consistently willing to attack regardless of the surrounding circumstances, the Target task could prevent opportunistic workers from being financially recruited by explicitly requesting that information not be shared and

paying higher wages. Raising the pay for the Target task from $0.05 to just $0.25 entirely counteracted the effects of paying workers $0.50 for the Attack task, which suggests that targeted requesters might be able to defend against attacks for a lower price than attackers can make them more successful. However, since many workers were willing to return potentially harmful information despite any actions taken by the Target task, it is clear that a significant threat still exists.

4 Exploring Worker Perceptions

To get a better sense of the reasons behind workers' apparent willingness or unwillingness to participate in the Attack task, we collected feedback from a subset of workers via email-based exit interviews and community forums. This revealed that many workers found the Attack task suspicious, but those who completed the Attack task often justified their actions as harmless.

4.1 Methodology

We conducted exit interviews for a subset of the workers who performed our tasks. Participants were recruited from the set of workers who were willing to complete both the Target and Attack tasks. These workers were asked, upon the completion of the Attack task, to email the requester via existing Mechanical Turk mechanisms, and offered a $0.75 bonus for completing a survey and additional follow-up questions. The survey asked workers to describe their motivation for completing the task and if they had any hesitations, what they thought the relationship between the tasks were, and whether or not that had seen the instruction not to share any of the information they saw. These questions were all phrased so as to not reveal ourselves as both requesters before getting responses from workers. Eleven out of 25 workers who were shown the request to participate agreed to share their reactions with us, as well as one who contacted us unrelated to the interview.

We also reviewed the feedback related to the two tasks that was posted on Turkopticon [4], a popular Mechanical Turk requester rating site for workers, after our experiments. There were nine threads started about our tasks and six replies. Two of these were positive, two were negative, one was neutral, and the rest did not give ratings.

4.2 Interview Results

A majority of workers reported that they found the task setup suspicious. Six of the survey respondents and nine forum posts explicitly discussed the suspicious nature of the tasks. Some of these suspicious workers took action to help other workers avoid the task or support the Target task. Four forum posts warned other workers that the Attack task seemed like a scam and should be avoided, and two flagged the task and claimed to have reported it to Amazon for Terms of Service violations. One survey participant also told us they reported it to Mechanical Turk. Interestingly, the worker still completed the Attack task. We were never contacted by Amazon regarding either task.

Other workers appeared unconcerned about the task setup. The positive feedback on the forum, for example, addressed payment speed and the simplicity of the tasks and ignored their relationship. Six of the 11 survey participants reported that they did not think much about the purpose of the Attack task. This is despite the fact that the Target task included an explicit request not to share information from the task. Of the six survey participants who reported remembering the task instructions clearly, five of them recalled seeing the explicit request not to share information from the task without prompting.

Some workers appeared to overlook the oddity of the two tasks until after they were completed. In response to a warning on Turkopticon about the Attack task by one worker, another worker replied, "You know what, I had the same issue today. […I] didn't pay it much mind when it happened." Three survey participants stated that they just assumed the two tasks were somehow related – a belief was echoed by the direct email we received.

In summary, while the Attack task appeared suspicious to many workers, only a few workers were willing to take action against it by posting to the forum, emailing the task owner, or reporting it to Amazon Turk. Most workers ignored their concerns. Moreover, since price still had a significant effect on the results even with randomized trials, it seems that workers' willingness to overlook problems may be affected by the payment level.

5 Conclusion

This paper explored the effects of task price on crowd workers' willingness to complete potentially harmful information extraction tasks. We found that while many workers were influenced by price, this influence was not uniform. Our results suggest that many workers do not want to attack another task as long as the payment for the task being targeted is not dwarfed by the pay offered by the attacker. In these cases, workers may decide not to respond even when this decision results in less pay across the two tasks. This does not appear to be true for one third of the workers we studied however. These workers may have overlooked the potential harm that could be caused by Attack tasks, or simply sought to maximize their earnings regardless of the context (implying low or no utility cost of malicious tasks for them).

Our findings can help task designers better account for the effect of worker payments on security in future crowd-powered systems. For example, tasks may be able to take advantage of the fact that one third of all workers were unwilling to attack the Target task despite high Attack task payments. It may be possible to identify a subset of these as ethical workers for tasks that require good behavior. This could be done using an *Ethical Gold* test, similar to the gold standard tests that are used to control quality in crowdsourced tasks. An Ethical Gold test may, for example, ask a worker to attack another task, and then only move forward with workers who refuse to. The willingness of some workers to speak up and report suspicious tasks to the platform and other workers in the community suggests that there also exist workers willing to go above and beyond to prevent these types of attacks, even if it is not in their best immediate financial interest.

References

1. Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. Soylent: a word processor with a crowd inside. In *UIST 2010*.
2. Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., Miller, R., Tatarowicz, A., White, B., White, S., and Yeh, T. Vizwiz: nearly real-time answers to visual questions. In *UIST 2010*.
3. Ipeirotis, P. Prisoner's dilemma and mechanical turk. http://bit.ly/Xo1WF5, 2009.
4. Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *CHI 2013*.
5. Kokkalis, N., Köhn, T., Pfeiffer, C., Chornyi, D., Bernstein, M.S. and Klemmer, S. EmailValet: Managing email overload through private, accountable crowdsourcing. In *CSCW 2013*.
6. Lasecki, W. S., Miller, C. D., Sadilek, A., Abumoussa, A., Borrello, D., Kushalnagar, R., and Bigham, J. P. Real-time captioning by groups of non-experts. In *UIST 2012*.
7. Lasecki, W. S., Teevan, J., and Kamar, E. Information extraction and manipulation threats in crowd-powered systems. In *CSCW 2014*.
8. Mason, W., and Watts, D.J. Financial incentives and the "performance of the crowd." In *ACM SIGKDD Explorations Newsletter*, 11(2), 2009.
9. Teodoro, R., Ozturk, P., Naaman, M., Mason, W., and Lindqvist, J. The motivations and experiences of the on-demand mobile workforce. In *CSCW 2014*.
10. von Ahn, L. Human computation. In *Design Automation Conference 2009*.