Dealing with Sub-optimal Crowd Work: Implications of Current Quality Control Practices

Ujwal Gadiraju
L3S Research Center
Leibniz Universität Hannover
30167 Hannover, Germany
gadiraju@L3S.de

Neha Gupta
School of Computer Science
Jubliee Campus, University of
Nottingham, NG8 1BB, UK
psxng1@nottingham.ac.uk

To be presented at the CHI'16 Workshop, Crowd Dynamics: Exploring Conflicts and Contradictions in Crowdsourcing.
CHI'16, May 7–May 12, 2016, San Jose, California, USA.

Abstract
Paid crowdsourcing markets are on the rise and are projected to flourish in the near future. Several stakeholders benefit from the paradigm wherein a demand for human intelligence at scale is quickly met by a supply of readily available crowd workers. In this paper, we delve into the two fundamental parties that are involved in crowdsourcing - the *task requesters* and the *crowd workers*. We highlight important questions regarding quality control practices that need to be addressed to foster a sustainable online labour market. From a holistic standpoint, there are several challenges pertaining to dealing with quality of crowd work and the underlying working conditions of the workers, that remain unsolved. For instance, the requesters do not commonly consider or have access to the situational context that the workers are enveloped in, which is reflected in the typical quality control methods used to deal with crowd work. In some cases, workers have been found to be unreliable and are consequently penalized. In this paper, we discuss the implications of quality control practices followed by requesters in paid crowdsourcing markets, which also affect trust and reliability between the said groups.

Author Keywords
Crowdsourcing; Microtasks; Quality; Workers; Ethics

ACM Classification Keywords

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

Payment versus Quality of Work Conundrum: How Requesters Deal with Crowd Work

Over the last decade, crowdsourcing has been widely adopted, ubiquitously spanning a wide range of domains. Due to the power asymmetry in existing microtask crowdsourcing platforms such as Amazon's Mechanical Turk (AMT)[1] or CrowdFlower[2], requesters deal with quality control of crowd work in different ways. Some of these commonly adopted methods raise ethical concerns, and form the focus of this position paper.

Qualification Tests to Select a Reliable Crowd

Requesters adopt pre-screening mechanisms as a means to ascertain that crowd workers participating in their HITs are capable of providing high quality responses. While these pre-screening tests are typically short and do not require considerable amounts of time to complete successfully, requesters don't necessarily pay workers for completing such qualification tests. Should requesters pay workers for participating in pre-screening tests? To what extent does this depend on the length and effort required during the pre-screening phase?

Reject Work Due to Poor Task Design

On AMT requesters can reject work without paying the workers in case they believe that the quality of work is poor. An aspect that is not given due consideration is that poor or sub-optimal work can be a result of bad or flawed task design. Is a requester within her ethical rights to reject work without paying when the task design is poor? How can requesters share accountability of poor quality work? What if crowd workers could rate requesters' task design and clarity of instructions before requesters are allowed to deploy the HITs?

Reject Work, With or Without Paying the Workers

Since a requester has complete authority in adjudicating whether a piece of work is worthy of acceptance or rejection, the power asymmetry breeds distrust between workers and requesters (although workers can challenge requesters when their work is rejected). To avoid misjudging the threshold for "acceptable work" from the crowd, in the absence of transparent methods that help to gauge worker genuineness, some requesters pay all workers despite the quality of their work. In such cases, post-hoc filtering is typically adopted after paying all workers in order to prune responses for quality control. Is it fair on the part of the workers to accept full pay despite providing sub-optimal work? Is it fair on the part of the workers to have all their work rejected, with no pay, even when optimal work was provided? How can we make a more transparent system that supports filtering, and is fair?

How Workers Deal with Crowd Work

Previous research works [4, 8] have provided an account of the staggeringly different ethnographic contexts that crowd workers are embedded in. We also note that there is a difference in how workers behind sub-optimal work are typecast across different communities that deal with crowdsourcing.

Aspects that Hinder Crowd Workers

While contributing to tasks workers experience barriers such as language, technology and poor task design, which inhibit them from producing good quality work [7, 4]. In such cases, workers take measures such as completing a task partially or entirely before actually accepting it on the plat-

[1] https://www.mturk.com/mturk/
[2] http://www.crowdflower.com/

form, or requesting for support from their friends and family, and even requesters to help them complete the work [4, 8]. Where work is rejected for no fault of the workers, they make efforts to get in touch with either the platform or the requesters to state their case. This exercise does not always receive a positive response, but workers attempt to get a fair treatment, as the following vignette describes.

Sumita from Gujarat: *"The requester gives the link, but when you open the link it shows that the survey is already closed; but first we accept the task only then we can open the link. [..] Sometimes we don't get the completion code (at the end of a survey), then we cannot understand, we spent so much time – 1 hour, and wasted our time and we are not getting the code, then how are we going to submit the survey. [..] At that time, we have to return the HIT and write to the requester. Some requesters are very good and give immediate reply – 'sorry this happened, we will see to it', but some don't even bother to give us a reply whether it was their fault"*.

Finding the right means to accurately contextualize the 'intent' of the workers would go a long way towards finding a harmonious typology within which we can embed crowd work based on quality. In addition, the high variability of task types [2] pose challenges with regards to workers' familiarity with tasks.

Risks Crowd Workers Take
There are a number of aspects that requesters should consider regarding the context in which workers are embedded while contributing work in online labour markets [8, 5]. The work environments may not always be appropriate, and the devices that workers use to complete tasks may not be ergonomically suitable. Workers are potentially subject to psycho-social risks as pointed out in [5], since the availability of work from one hour to the next, from one day to another cannot be taken for granted. Moreover, the power asymmetry with regards to the reputation systems that dictate their access to fewer or more available tasks can render workers into a constant state of unrest. In many cases workers handle their daily chores in tandem with completing available work, subjecting them to distractions and disrupting concentration and their flow of work. These risks and issues generally remain invisible to the requesters, whose focus is usually on drawing out good work from the crowd. Considering the constraints despite which workers take part in several tasks, several questions emerge that need to be addressed in order to facilitate a crowdsourcing paradigm that is accountable from a holistic standpoint. Should a worker who is using poorer equipment to provide a high-quality of work be awarded a bonus (similar to corporates rewarding employees for working overtime or exhibiting extra efforts)? Should monetary incentives be a function of socio-economic aspects to an extent? Should requesters consider this in their task design?

Discussion
So what does this mean for requesters, researchers and designers who design tasks and tools to test the quality of work produced by the crowd? It is well known that a fraction of crowd workers intentionally provide ill-fitting responses with an aim to complete more work and attain quick pay [3]. While communities that predominantly deal with the algorithmic contexts of crowdsourcing (optimizing for parameters such as quality, time, etc.) freely use terms such as spammers [6], malicious workers [1] and so forth, other communities that focus on the human elements of crowdsourcing present cause for caution in typecasting sub-optimal work and workers behind it [9]; as do the workers themselves, as the following vignette from an AMT-related online forum suggests.

general65: *"I don't like it. Another idiot professor who thinks he knows what's best for the private market. This will only mean the government getting involved and regulating the requester's which in turn will end up in less pay for us. Someone please tell this idiot professor to stay in the classroom."*

However, it is important to note that since existing methods that detect work quality rely on the data that is produced without considering the circumstances of the workers, there is a need to distance the quality of work produced from the workers behind it.

Non-aggressive Filtering of Crowd Workers

Finding a reasonable way to deal with sub-optimal work in crowdsourcing labour markets is key for fostering trust between workers and requesters, and enhancing the reliability of crowd work. Could crowd work be typecasted based on quality into (i) Acceptable work (ii) Sub-optimal work, and (iii) Disruptive work? One way of doing this, would mean requesters adopting less-aggressive means to deal with sub-optimal work from crowd workers when there is little or no evidence of intentional disruptive work or malicious activity. Flagging workers who provide sub-optimal responses to scrutinize their work further implicitly, instead of rejecting their work without payment can be a way to achieve this.

REFERENCES

1. C. Eickhoff and A. P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.
2. U. Gadiraju, R. Kawase, and S. Dietze. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 218–223. ACM, 2014.
3. U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 1631–1640, 2015.
4. N. Gupta, D. Martin, B. V. Hanrahan, and J. O'Neill. Turk-life in india. In *Proceedings of the 18th International Conference on Supporting Group Work*, pages 1–11. ACM, 2014.
5. U. Huws. *Online labour exchanges or 'crowdsourcing': Implications for occupational safety and health*. European Agency for Safety and Health at Work (EU-OSHA), 2015.
6. G. Kazai, J. Kamps, and N. Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1941–1944. ACM, 2011.
7. S. Khanna, A. Ratan, J. Davis, and W. Thies. Evaluating and improving the usability of mechanical turk for low-income workers in india. In *Proceedings of the first ACM symposium on computing for development*, page 12. ACM, 2010.
8. D. Martin, B. V. Hanrahan, J. O'Neill, and N. Gupta. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 224–235. ACM, 2014.
9. N. Salehi, L. C. Irani, and M. S. Bernstein. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1621–1630. ACM, 2015.