On the Role of Task Design in Crowdsourcing Campaigns

Carlo Bernaschina, Ilio Catallo, Piero Fraternali, and Davide Martinenghi
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Italy
first.last@polimi.it [*]

Abstract

Despite the success of crowdsourcing marketplaces, fully harnessing their massive workforce remains challenging. In this work we study the effect on crowdsourcing campaigns of different feedback and payment strategies. Our results reveal the joint effect of feedback and payment on the quality and quantity of the outcome.

Introduction

Lately, crowdsourcing has gained traction as a valuable method for the accomplishment of tasks that proved difficult to be solved by means of automatic algorithms. The effectiveness of such an approach is demonstrated by the thousands of tasks that are daily published on crowdsourcing marketplaces such as CrowdFlower.

Despite their success, it is still not completely clear how to fully leverage these platforms. Indeed, creating collaboration between strangers is really challenging, even in face-to-face situations. In the context of a crowdsourcing initiative, engaging individuals may become even more difficult due to the lack of direct interaction. As a result, crowd workers tend to provide low quality contributions. Strategies to mitigate the risk of receiving poor contributions include increased payments, the adoption of some quality assurance techniques, and the conscious use of cognitive biases aiming at increasing the significance of the task (Chandler and others 2013). In this regard, studies find mixed results on the effect of financial incentives. It has been shown (Rogstadius and others 2011) that increasing monetary rewards induces workers to perform more units of a task, but does not affect work quality. Paying more could even reduce quality after a certain point because it encourages speed (Heer and others 2010). However, the authors in (Yin and others 2013) reported that changing the magnitude of the reward over two consecutive tasks does affect the quality of the results, due to the existence of an anchoring effect. Similarly, the author in (Harris 2011) reported that when payments change according to the workers' performance, the quality of work increases as payment increases. As for feedback, it has been shown (Lee and others 2013) that the usage of feedback techniques can help workers in performing better.

Differently from the mentioned works, we investigate the *combined effect* of different payment and feedback strategies. We hypothesize that the joint effect of motivational and payment feedback, as well as of performance-based payments could affect both quality and quantity of the results. We validate our hypothesis on an image labeling task, in which crowd workers are prompted with a sequence of images, and required to indicate if each image can be considered relevant for the *cooking* topic.

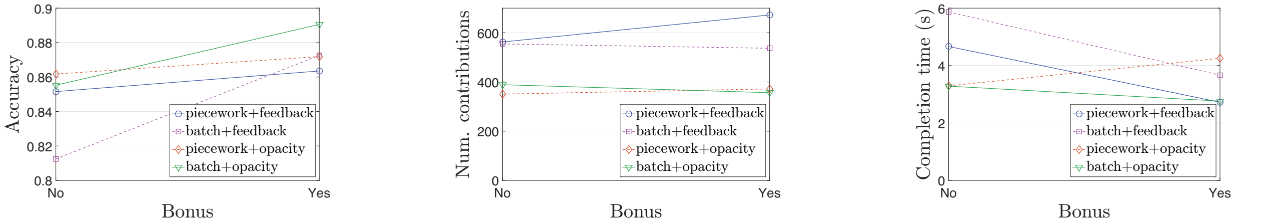
Study design

In this section, we first describe the adopted crowdsourcing task and the associated treatments. Second, we provide precise definitions of the dependent variables under study.

Treatments. We provided workers with a sequence of $1,000$ images, and asked them to indicate whether each such image could be considered relevant for the cooking topic. Participants were randomly partitioned into four treatments: *i) piecework+opacity*, whose execution conditions completely resemble those of a generic crowdsourcing marketplace; *ii) piecework+feedback*, in which workers are paid a fixed rate per task while being informed about the amount of money they have earned so far, as well as the number of tasks left to be completed; *iii) batch+opacity*, in which workers are paid in batches of images, and no payment is provided for partially completed batches; and *iv) batch+feedback*, in which workers are paid in batches of imaged while being provided with precise feedback on the state of execution of the task. Moreover, some participants were informed that we were going to grant an additional bonus for each correct answer they provided. We set the payment per contribution to 0.1c\$. The bonus payment was of 0.2c\$ per correct answer.

Dependent variables. We measured the effects of the above treatments on three dependent variables: *i) task accuracy*, which can be computed as a function of the difference between the gold standard and the provided label (each image comes with an expert-proven label); *ii) number of submitted contributions*, by which we mean the number of tasks each worker completed (workers were free to leave the experiment at any time); and *iii) completion time*, by which we mean the number of seconds spent to complete a single task.

[*] This work is partly funded by the EC's FP7 "Smart H2O" project, and the EU and Regione Lombardia's "Proactive" project



(a) Mean accuracy

(b) Mean number of contributions

(c) Mean completion time (in seconds)

Figure 1: Experiment results throughout the 8 treatments

Experimental evaluation

The experimental campaign has been executed on Champagne (Bernaschina and others 2015), with CrowdFlower as the recruitment platform. We ran our experiment for one week, collecting contributions from 253 workers. For a more detailed comment please refer to the full version of this work at http://bit.ly/1fk5lgY.

Task accuracy. Figure 1a reports the mean plot for each treatment. As shown, regardless of the presence of a feedback strategy, task accuracy benefits from the presence of performance-based rewards (bonus). When batch payments are considered, the average accuracy improves from 0.81 to 0.87 in the absence of feedback, and from 0.86 to 0.89 when feedback is present. When changing the payment strategy to *piecework*, the improvements are still present, although marginal. Conversely, the use of feedback appears to worsen the average accuracy. Indeed, for both the considered payment strategies, the attained task accuracy value is lower when feedback is present. Namely, the decrease ranges from a minimum of 0.01 in the case of piecework, to a more marked value of 0.04 in the case of batch payments. This could be due to the fact that providing precise information about how many images are left can in fact incentivize quantity rather than quality. A 3-way ANOVA test confirmed the effect of bonuses on task quality ($p < 0.02$, $\alpha = 0.05$). Moreover, according to the result of the test ($p = 0.052$) we should reject the hypothesis of a negative impact of feedback on task quality. Still, with a p-value so close to 0.05 one could still question the practical significance of the test.

Number of submitted contributions. Figure 1b depicts the variation of the average number of contributions as the treatment changes. As shown, the number of contributions clearly increases with the presence of feedback, independently of other factors. The improvement is particularly remarkable ($+85\%$) when the requester opts for a fixed rate per task (piecework) together with a performance-based bonus. Indeed, a 3-way ANOVA test confirms the statistical significance of the feedback effect ($p < 0.001$). Moreover, the test does not reveal any specific interaction effects.

Completion time. Figure 1c shows that different treatments lead to different average completion times. A 3-way ANOVA test revealed that none of the studied factors has a statistically significant effect on the completion time.

Conclusions

We conducted an experimental evaluation on the role of task design in crowdsourcing campaigns. In particular, we investigated the joint effect of different payment methods and feedback strategies on a labeling task. Our analysis shows that these two factors can effectively be used in combination, as they affect both quality and quantity of the results, with no effect on the time required for the task completion.Namely, we saw that feedback impacts on the number of submitted contributions, whereas performance-based payments can be adopted for improving the outcome quality. Moreover, there was not enough evidence to confirm a negative influence of feedback on task quality. We deem the combined adoption of feedback and performance-based payments to represent a convenient trade-off for task requesters, who will sacrifice a marginal percentage of accuracy in exchange for a huge improvement in the number of submitted tasks.

As a future direction, it may be interesting to validate our results using a workforce composed of thousands, rather than hundreds, of workers. The possible independence of our findings w.r.t. the dimension of the workforce would justify the use of the discusssed payment and feedback techniques for crowdsourcing campaigns of any size. Our evaluation may further be extended by considering other crowdsourcing platforms, as we cannot exclude that a change of platform could affect the nature of our findings. Finally, we plan to extend the proposed study to other task types, such as multiclass classification or segmentation tasks.

References

Bernaschina, C., et al. 2015. Champagne: A web tool for the execution of crowdsourcing campaigns. In *WWW15*, 171–174.

Chandler, D., et al. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *J. of Economic Behavior & Organization* 90:123–133.

Harris, C. 2011. You're hired! an examination of crowdsourcing incentive models in human resource tasks. In *CSDM*, 15–18.

Heer, J., et al. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *CHI*, 203–212.

Lee, T. Y., et al. 2013. Experiments on motivational feedback for crowdsourced workers. In *ICWSM*.

Rogstadius, J., et al. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *ICWSM*, 321–328.

Yin, M., et al. 2013. The effects of performance-based financial incentives in online labor markets. In *AAAI*, 1191–1197.