Dear Editors, Dear Reviewers,

We are thankful for your work and the suggestions regarding our paper, as well as the opportunity to present an extension of our work. We addressed all reviewer comments to improve the quality of our work and of course added new insights that were not part of the original work. While the original work featured Certainty Factors and Bayesian Networks as vehicles to balance the costs of quality control and the danger of fraud, we now look at deep learning techniques in comparison.

Please find below a list of major comments and changes implemented in the current journal version of the article:

- We added a paragraph in the Introduction to address the concerns regarding the use of our method as a way to discriminate against good workers from countries or any other involuntary inherent features of the worker, which are not based on their actions. We stressed that the proposed method is no way intended to discriminate potentially good workers, but rather to help to assess the potential quality of the worker by dynamically adjusting the number of gold questions required given the collected data. That means in particular that no worker is generally excluded from a crowdsourced task, but that at the beginning of a set of tasks in some instances more quality controls are instantiated which however quickly are reduced in case of non-fraudulent behavior.
- Regarding the features selected to profile workers, we agree that they are indeed quite limited and represent one of the main challenges in our work. Having more informative attributes could benefit our approach, but as we describe in detail in Section 5.1, currently platforms do not offer such attributes (nor will they offer them in the near future in compliance with data protection acts). Thus, we emphasize in Section 5.1 the scope and limitations of the features we used to learn and test our models.
- In Section 5.1 we now also provide more details about the way we created the different datasets with various spammer rates. In short, using a large worker pool, we select at random spammers and no spammers until we reach the desired ratio, e.g., Reliable75 means that we have randomly selected 75% of non-spammers and 25% of spammers from the total pool. The actual number that corresponds to these percentages is then calculated considering that we fixed the size of each test dataset to 200 instances.
- In Section 5.2.2 we now explicitly clarify how we calculated the priors of the attributes different that country namely ‘channel’, ‘started at’ and ‘trust’.
- Sections 4.3 and 5.4 include the main new material mentioned above, namely the ‘Siamese architecture’ and its corresponding experimental and evaluation setting. We then compare all deep learning results to the Bayesian/Certainty Factor approach and draw new conclusions. This content has not been published anywhere and serves as a good extension setting our initial approach into a broader perspective.
- Finally, we are sharing all the data we used to train and test our models to allow for a maximum of reproducibility.

Kind Regards, the Authors