The Output-Agreement Method Induces Honest Behavior in the Presence of Social Projection

ARTHUR CARVALHO, STANKO DIMITROV, and KATE LARSON
University of Waterloo

The output-agreement method has been successfully used to reward agents in a variety of crowdsourcing settings. This method consists of a simple payment function that randomly matches two agents’ reported information and rewards agreement. In this letter, we discuss how the output-agreement method might induce honest behavior when there exists social projection, i.e., when each agent believes that his private information is the most popular one amongst his peers.

Categories and Subject Descriptors: J.4 [Social and Behavioral Science]: Economics
General Terms: Economics, Human Factors, Theory
Additional Key Words and Phrases: Output-Agreement Method, Social Projection

1. INTRODUCTION

Recent years have seen an increased interest in crowdsourcing as a way of obtaining information from a large group of agents at a reduced cost. A simple payment structure that compares agents’ reported information and rewards agreements, known as the output-agreement method, has been successfully used to reward agents in many different crowdsourcing scenarios [von Ahn and Dabbish 2008]. In this letter, we discuss how social projection, which is a judgmental heuristic where agents assume that others will behave as they themselves do, may be the reason why the output-agreement method has been successfully used in practice.

Our argument is that when an agent believes that his private information is the most popular one, then honest reporting maximizes the chance of a random agreement. Formally, we show that when risk-neutral agents independently answer a multiple-choice question, the output-agreement method induces honest behavior when there exists social projection. The output-agreement method, as we consider in this letter, rewards a fixed amount of payoff units \( v_{\text{max}} \in \mathbb{R} \) if two reported answers agree with each other, and \( v_{\text{min}} \in \mathbb{R} \) otherwise, where \( v_{\text{max}} > v_{\text{min}} \).

Waggoner and Chen [2013] argued that the output-agreement method does not elicit honest answers. Instead, it elicits the correct answer according to the common knowledge among agents. We obtain a different result because, in contrast to Waggoner and Chen’s work, we make assumptions on the nature of agents’ information structure so as to model social projection.

2. RELATED WORK

An attractive feature of the output-agreement method is that it does not require a ground truth to score an agent’s report. Two well-known methods to induce honest reporting without assuming a ground truth are the Bayesian truth serum (BTS)
method [Prelec 2004] and the peer-prediction method [Miller et al. 2005].

Similar to our setting, the BTS method works on a single multiple-choice question with a finite number of answers. Each agent endorses the answer most likely to be correct and predicts the empirical distribution of the endorsed answers. Agents are evaluated by the accuracy of their predictions as well as based on how surprisingly common their reported answers are. Under the BTS scoring method, collective honest reporting is a Bayes-Nash equilibrium.

The BTS method has two major drawbacks. First, it requires the population of agents to be large. Second, besides reporting their answers, agents must also make predictions about how their peers will report their answers. While the artificial intelligence community has recently addressed the former issue [Witkowski and Parkes 2012], the latter issue is still an intrinsic requirement for using the BTS.

The peer-prediction method [Miller et al. 2005] does not share the drawbacks of the BTS method. In the setting of the peer-prediction method, a number of agents experience a product and rate its quality. A mechanism then collects the ratings and makes payments based on those ratings. The peer-prediction method makes use of the stochastic correlation between the signals the agents observe from the product to achieve a Bayes-Nash equilibrium where every agent reports honestly.

In spirit, the output-agreement method is a peer-prediction method. Radanovic and Faltings [2013] showed that a simpler version of the output-agreement method that rewards one payoff unit if two reported answers agree with each other, and zero otherwise, induces honest reporting under the so called self-dominant assumption. Such an assumption means that an agent believes that his answer is the most popular answer. We note in this letter that the self-dominant assumption is consistent with the well-established social-psychological phenomenon of social projection.

There is evidence that social projection exists intra-groups and, to a less degree, inter-groups [Robbins and Krueger 2005]. Some formal models have been proposed to model social projection, e.g., Brenner and Bilgin [2011] proposed a social projection model based on support theory, whereas Busemeyer and Pothos [2012] discussed a quantum model of social projection.

In this letter, we discuss a flexible social projection model that allows one to model the strength of the projection within the Bayesian learning framework. This modeling choice is desirable because it has been shown that the strength of the projection is context dependent, e.g., intra-group social projection seems to be stronger in groups that are artificially created in the laboratory than in groups that exist in the social world [Robbins and Krueger 2005].

Given our social projection model, we generalize the results by Radanovic and Faltings [2013] regarding the output-agreement method, i.e., we show that a function that rewards a fixed amount of payoff units \( v_{\text{max}} \) if two reported answers agree with each other, and \( v_{\text{min}} \) otherwise, for \( v_{\text{max}} > v_{\text{min}} \), induces honest reporting in the presence of social projection.

3. THE MODEL

We consider a multiple-choice question with a total of \( n \geq 2 \) exhaustive and mutually exclusive answers \( A_1, \ldots, A_n \). We assume that the population’s knowl-
We can write the above posterior as follows:

$$b = \text{probability that one of your peers will report the answer to the multiple-choice question, an agent's response to the query "what is the probability that an agent selected at random from the population has } A_k \text{ as the answer to the multiple-choice question.}$$

Each agent possesses a privately observed draw (signal) from $\Omega$. We refer to observed signals as honest answers. We denote the honest answer of an agent $i$ by $t_i \sim \Omega$, where $t_i \in \{A_1, \ldots, A_n\}$. Honest answers are independent, i.e., $P(t_i|t_j) = P(t_i)$. We say that agent $i$ is reporting honestly when his reported answer $r_i \in \{A_1, \ldots, A_n\}$ is equal to his honest answer, i.e., $r_i = t_i$.

A trusted entity is responsible for eliciting the answers and for rewarding the agents. Let $s_i$ be agent $i$'s reward after he reports $r_i$. We discuss how to compute $s_i$ in the next section. We make four major assumptions in our model:

1. **Autonomy**: agents cannot influence other agents’ answers, i.e., they do not know each other’s identity and they are not allowed to communicate with each other during the elicitation process.
2. **Risk Neutrality**: agents behave so as to maximize their expected rewards.
3. **Uninformative Dirichlet Priors**: each agent $i$ has a prior distribution over $\omega$. We assume that this prior is an uninformative Dirichlet distribution with hyperparameter $\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,n})$, i.e., $P(\omega|\alpha_i)$.
4. **Social Projection**: after observing $t_i$, every agent $i$ updates his belief by applying Bayes’ rule to his prior, i.e., $P(\omega|\alpha_i, t_i)$.

The first assumption means that agents work individually on the multiple-choice question. It describes how, for example, crowd workers traditionally solve tasks on the crowdsourcing platform Amazon Mechanical Turk. The second assumption means that agents are self-interested and no external incentives exist for each agent.

The third assumption means that all the elements of the hyperparameter $\alpha_i$ have the same value, i.e., $\alpha_{i,1} = \cdots = \alpha_{i,n} > 0$. In practice, this assumption implies that before an agent sees the underlying multiple-choice question and knowing only the number of possible answers $n$, the agent’s response to the query “what is the probability that one of your peers will report the answer $A_k$?” is $\frac{1}{n}$, for all $k \in \{1, \ldots, n\}$. That is, all the answers are equally likely a priori. Formally:

$$P(r_j = A_k|\alpha_i) = \mathbb{E}[\omega_k|\alpha_i] = \frac{\alpha_{i,k}}{\sum_{x=1}^{n} \alpha_{i,x}} = \frac{1}{n}$$

for $j \neq i$. The fourth assumption means that the posterior distributions are consistent with Bayesian updating. Moreover, this assumption implies that after solving the multiple-choice question, an agent’s response to the query “what is the probability that one of your peers will report the answer $A_k$?” is:

$$P(r_j = A_k|\alpha_i, t_i) = \mathbb{E}[\omega_k|\alpha_i, t_i] = \begin{cases} \frac{\alpha_{i,k} + 1}{\sum_{x=1}^{n} \alpha_{i,x}} & \text{if } t_i = A_k, \\ \frac{\alpha_{i,k}}{\sum_{x=1}^{n} \alpha_{i,x}} & \text{otherwise}. \end{cases}$$

We can write the above posterior as follows:

$$P(r_j = A_k|\alpha_i, t_i) = \mathbb{E}[\omega_k|\alpha_i, t_i] = \begin{cases} \frac{1}{n} + \frac{y_i}{n} & \text{if } t_i = A_k, \\ \frac{y_i}{n-1} & \text{otherwise}. \end{cases} \quad (1)$$
for $0 < y_i = \frac{n-1}{n+\alpha_{i,k} n^2} < 1$. The above equation models social projection, which is the tendency to expect similarities between oneself and others [Robbins and Krueger 2005]. In other words, an agent believes that his honest answer is the most popular answer amongst his peers. With this perspective, the value of $y_i$ and, consequently, the elements $\alpha_{i,1}, \ldots, \alpha_{i,n}$ of the hyperparameter $\alpha_i$, determine the strength of the social projection, where a small (respectively, high) value for $y_i$ implies a weak (respectively, strong) social projection. It is important to note that we make no assumptions regarding the value of $y_i$, which means that the magnitude of the social projection can be different for different agents.

4. THE OUTPUT-AGREEMENT METHOD

Recall that $s_i$ is the reward agent $i$ receives after he reports $r_i$. Consider that $s_i$ is determined according to the output-agreement method, which is defined as follows:

$$s_i = \tau(r_i, r_j) = \begin{cases} v_{\text{max}} & \text{if } r_i = r_j, \\ v_{\text{min}} & \text{otherwise} \end{cases}$$

where $v_{\text{max}} > v_{\text{min}}$ and $j \neq i$. That is, agent $i$ receives the maximum payment $v_{\text{max}}$ if and only if he reports an answer equal to the answer reported by another agent $j$ randomly selected from the population of agents. Given the autonomy assumption, agent $j$ cannot influence agent $i$’s reported answer, and vice versa. Further, given the assumption that agents are risk neutral, agent $i$ behaves so as to maximize his expected reward, which is equal to:

$$\mathbb{E}[\tau(r_i, r_j)] = v_{\text{max}} \times P(r_j = r_i | \alpha_i, t_i) + \sum_{A_x \neq r_i} v_{\text{min}} \times P(r_j = A_x | \alpha_i, t_i)$$

Given that $v_{\text{max}}$ and $v_{\text{min}}$ are fixed, agent $i$ maximizes the above equation by moving as much probability mass towards $v_{\text{max}}$ as possible. The posterior in (1) implies that $P(r_j = t_i | \alpha_i, t_i) > P(r_j = A_x | \alpha_i, t_i)$, for all answers $A_x \neq t_i$. Consequently, agent $i$ strictly maximizes his expected reward if and only if he is honest, i.e., when he reports $r_i = t_i$. The above result means that each agent determines, according to his posterior belief, the answer most likely to be reported by a random peer. This answer turns out to be the agent’s honest answer in the presence of social projection.

As the literature on psychological projection has shown, social projection serves as an egocentric heuristic for inductive reasoning [Robbins and Krueger 2005]. Consequently, social projection might explain why the output-agreement method has been so effective in crowdsourcing settings. It would be exciting to validate or invalidate this hypothesis by determining the extent to which social projection exists in crowdsourcing settings.

To conclude, we expect the connection between social projection and honest reporting under the output-agreement method to be of empirical and theoretical value, and to open new research directions. Some relevant questions include:

—How can one induce social projection and, consequently, incentivize honest reporting under the output-agreement method?
—How to effectively measure and model social projection?
—Which variables affect the strength of social projection?
—Is there a correlation between the strength of social projection and expertise?
REFERENCES


