IEICE TRANSACTIONS

on Information and Systems

VOL. E97-D NO. 9
SEPTEMBER 2014

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Preventing Participation of Insincere Workers in Crowdsourcing by Using Pay-for-Performance Payments

Shigeo MATSUBARA[†a)], *Member and* **Meile WANG**[†b)], *Nonmember*

SUMMARY We propose a method for finding an appropriate setting of a pay-per-performance payment system to prevent participation of insincere workers in crowdsourcing. Crowdsourcing enables fast and low-cost accomplishment of tasks; however, insincere workers prevent the task requester from obtaining high-quality results. Instead of a fixed payment system, the pay-per-performance payment system is promising for excluding insincere workers. However, it is difficult to learn what settings are better, and a naive payment setting may cause unsatisfactory outcomes. To overcome these drawbacks, we propose a method for calculating the expected payments for sincere and insincere workers, and then clarifying the conditions in the payment setting in which sincere workers are willing to choose a task, while insincere workers are not willing to choose the task. We evaluated the proposed method by conducting several experiments on tweet labeling tasks in Amazon Mechanical Turk. The results suggest that the pay-per-performance system is useful for preventing participation of insincere workers.

key words: *crowdsourcing, mechanism design*

1. Introduction

Crowdsourcing is a new problem-solving method that requests an undefined public to accomplish a task [1]. Human workers are involved in a problem-solving process in many situations. For example, consider that a requester wants to know what impression several alternatives of advertisement copy give to potential buyers in order to select the best advertisement copy consistent with the product's characteristics. For this task, it might be possible to apply natural language processing techniques but it is often difficult to obtain results of sufficient quality without focusing on the application domain. On the other hand, people can easily accomplish this kind of task even if they are not trained. However, this brings about another challenge of how to motivate workers to complete a task. Thus, applying multiagent system technologies, including mechanism design, is promising.

This provides a new challenge of how to prevent participation of insincere workers. Amazon Mechanical Turk (MTurk) is a crowdsourcing marketplace that enables requesters to publish a task and workers to get paid by accomplishing the task. The individual tasks are called Human Intelligence Tasks (HITs). If a requester publishes a HIT on MTurk, it will be posted on the HIT list. Workers can select a HIT from the list and submit the results. If the requester approves the submitted results, the requester pays the workers. It has been reported that a translation task can be performed quickly at low cost by using MTurk [2]. However, it has been pointed out that there is a significant number of insincere workers [3].

Insincere workers are those who complete a task but who deliberately do not pay attention to the task instructions and try to get paid with the least effort. In MTurk, payment will not be given if the requester does not approve the submitted answer. However, requesters often do not know what the correct answer is. In addition, workers are not likely to select a task if its requester has a high disapproval rating. This makes it difficult for requesters to disapprove submitted answers if they want to publish subsequent tasks. Thus, the approval system is not effective for excluding insincere workers. It is also possible that a requester will first impose a qualification test on workers and provide the main tasks only to the workers who pass the test. However, this causes another problem of ensuring a sufficient number of workers.

A countermeasure against insincere workers is to use redundancy. That is, a requester gives the same task to more workers, and then tries to obtain the correct answer by applying a voting method or machine learning techniques such as the expectation-maximization (EM) algorithm [3], [4]. However, even if the payments to individuals are small, redundancy increases the problem-solving cost. If the requester has many tasks, the overhead from redundancy becomes severe. In addition, if the task is a kind of poll, such as investigating the impression of a copy of a product, the requester needs to ask many workers to complete the task. If there are insincere workers, the obtained distribution of answers might be skewed, which leads to wrong conclusions. Therefore, it is preferable to develop a method for preventing participation of insincere workers in advance instead of relying on post processing. To solve this problem, we examined the pay-for-performance payment system.

As mentioned above, although there is an approval system in MTurk, many tasks can be virtually viewed using the fixed payment system. That is, if a worker submits an answer, the worker can receive payment, such as \$0.05. The requester can set a bonus in MTurk. For example, a bonus of \$0.01 will be paid if the worker completes an additional task such as answering a questionnaire. If a requester sets the fixed payment to quite a lower value and sets the bonus to a substantial amount, the pay-for-performance payment system can be implemented, i.e., a worker is paid a lower amount of money as the fixed payment and can receive a

Manuscript received December 17, 2013.
Manuscript revised April 28, 2014.

[†]The authors are with the Department of Social Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan.
a) E-mail: matsubar@i.kyoto-u.ac.jp
b) E-mail: wang@ai.soc.i.kyoto-u.ac.jp

DOI: 10.1587/transinf.2013EDP7441

bonus if he/she submits the correct answer. Here, we expect that if fixed-payment and the pay-for-performance payment tasks simultaneously exist, the amount of the expected payment for the pay-for-performance payment task becomes lower for insincere workers; therefore, insincere workers would prefer fixed-payment tasks.

However, a payment pair must be chosen carefully. Consider the case where other requesters publish tasks with a fixed payment of \$0.05. If a requester publishes a similar task with a fixed payment of \$0.00 and a bonus of \$0.05, sincere workers as well as insincere workers will not be willing to do this task because payment is not guaranteed even for sincere workers who always obtain the correct answer. That is, sincere workers expect to obtain lower payment if they choose this pay-for-performance task compared to fixed-payment tasks. Therefore, a payment setting must be found that is not attractive to insincere workers but attractive to sincere workers.

We propose a method of calculating an appropriate payment pair of fixed payment and bonus in the pay-for-performance payment system. First, given the ratio of sincere workers in the worker pool, the accuracy ratios of answering the correct answers for sincere/insincere workers, we discuss finding a payment setting that is not attractive to insincere workers and attractive to sincere workers. Attractive means that a pay-for-performance task gives a larger expected payment to sincere workers compared to fixed-payment tasks. We evaluated the performance of the proposed method by publishing several tasks on MTurk. Some readers might think this is a simple calculation, but the probability of obtaining a bonus depends on other workers' submitted answers because the correct answer is determined by collective decision making, such as plurality voting. That is, the probability that workers can obtain a bonus is affected by several factors such as the number of workers, which means that simple calculation is not sufficient to solve the problem.

Section 2 describes related work. Section 3 explains our model of pay-for-performance payment systems in crowdsourcing and Sect. 4 describes our method for calculating the expected payment in the pay-for-performance system. Section 5 explains how to estimate the parameters in our model that conforms to the actual task request in MTurk. Section 6 describes the experimental design and explains the experimental results of evaluating the performance of the proposed method. Section 7 concludes the discussion.

2. Related Work

Related studies have focused on reward and monetary incentive in crowdsourcing [5]–[10], but these studies did not investigate the pay-for-performance system. Witkowski investigated how to motivate persons to report the true evaluation in reputation markets [11] but the applied area was different.

Pay-for-performance have been studied in economics [12]. Holmstrom gives a general framework for determining the optimal contract for the problem of moral hazard [13]. However, applying his study to our problem is not straightforward. As explained in Sect. 5, we have to consider the problem of parameter estimation, i.e., how to estimate the ratio of sincere workers and accuracy rates of sincere workers/insincere workers. Such issues cannot be covered in Holmstrom's study. In addition, it is not clear whether workers in MTurk are rational as the economic literature assume. We should examine whether the methods of incentivizing workers in the economic literatures hold for environments in crowdsourcing.

3. Model

We focus on labeling tasks for selecting the most suitable label for a given text or image among alternative labels. For example, workers are asked to answer a question whether the following sentence gives a positive or negative impression.

> You can find out anything on Facebook and Twitter.

This task may not have an expert who can always give the correct answer. In such a case, the correct answer, i.e., the ground truth, is determined by collective decision making such as plurality voting or Bayesian inference. We call such an alternative a "correct" answer that receives the most votes if a requester ask a sufficient number of sincere workers.

This definition is suitable for cases that the requester's objective is to obtain the alternative with the most votes. For example, determining the accented syllable in a word, and determining the relevance between given keywords and content of a web page are included in such tasks.

If insincere workers are included, we have to estimate the correct answer. In this paper, we use a simple EM algorithm to estimate the correct answer, although many variations exist [14], [15].

The discussions in this section and in Sect. 4 assume two alternatives, while the discussion in Sect. 6 assumes three alternatives. In addition, we assume that if there is more than one alternative with the most votes, an alternative is selected at random and is treated as the correct answer.

A worker is characterized as sincere or insincere. Insincere workers always randomly select an alternative without considering which alternative is appropriate. We assume that a sufficient number of workers exist in the worker pool. The ratio of sincere workers is represented as α. Accomplishing a task incurs the cost of c_S for a sincere worker, while c_I for an insincere worker ($c_S > c_I$). If workers belong to the same type, they incur the same cost for completing a task because we assume that completing a task does not require any expertise.

Whether a worker is sincere or insincere is related not to an ability of the worker but to an attitude of the worker for completing a task. Thus, it may happen that a worker accomplishes a task sincerely but the answer is wrong. Utilizing the pay-for-performance system has a risk to exclude such sincere workers. This problem can be mitigated if

there is a correspondence relation between the reported answer and the correct answer. For example, if it is known that a worker always chooses minority opinions for all the questions, its report helps to estimate the correct answer, although the detailed discussion is beyond the scope of this paper.

The accuracy rate, i.e., the probability of answering the correct answer, is represented as p for sincere workers and q for insincere workers. The distribution of answers might be different for different tasks. For example, in the task of selecting an appropriate label for a given sentence, the distribution that positive:negative = 0.9:0.1 is obtained for sentence X and positive:negative = 0.51:0.49 for sentence Y. $p > q$ and $p > 0.5$ are also assumed.

We consider a fixed-payment system and a pay-for-performance payment system as reward payment systems. The payment amount in the fixed-payment system is represented as r. The utility of a sincere worker can be represented as $u_S = r - c_S$ and $u_I = r - c_I$ for an insincere worker.

The pay-for-performance payment system is characterized as a payment pair of fixed payment and bonus. The former is represented as a and the latter is represented as b. The fixed payment will be paid regardless of correct/incorrect if a worker submits an answer, while b will be paid if the submitted answer is the alternative with the most votes. We do not say that the answer is correct but the submitted answer is an alternative with the most votes. The utility of workers can be represented as follows.

$$u_S = a + b \cdot f_S - c_S \quad ; sincere$$
$$u_I = a + b \cdot f_I - c_I \quad ; insincere,$$

where f_S represents the probability that a sincere worker can obtain b and f_I represents the probability that an insincere worker can obtain b. Note that f_S and f_I do not directly correspond to p and q. This is because which alternative become the alternative with the most votes depends on the other workers' answers. That is, as the ratio of insincere workers changes, the distribution of answers changes and the alternative with the most votes also changes. For example, if all the workers are insincere except one, each alternative has almost the same probability of becoming the alternative with the most votes. Thus, if the sincere worker always obtains the correct answer, he/she often fails to obtain b. The method for calculating f_S and f_I is described in Sect. 4.

We assume that similar types of tasks are published by other requesters at the same time. This assumption holds in MTurk. Workers face the problem of which task they should choose. We assume that workers choose one that gives a larger expected payment. Therefore, once f_S and f_I are obtained, a requester should find a payment pair of a and b that satisfy the following conditions.

$$a + b \cdot f_S - c_S \geq r - c_S \quad ; sincere$$
$$a + b \cdot f_I - c_I < r - c_I \quad ; insincere$$

These conditions are required to prevent participation of insincere workers. We do not intend to control the behaviors of all the requesters nor intend that a requester publishes two tasks, where one is in the fixed-payment system and the other is in the pay-for-performance payment system, but we consider a situation in which other requesters publish similar tasks with fixed payment.

4. Method for Calculating Expected Payment

This section explains method for calculating the amount of expected payments for sincere and insincere workers if the pay-for-performance payment system is used.

Consider a sincere worker choose a task in the case that an identical task is given to three workers. If the remaining two workers are sincere, the probability p_1^S that the sincere worker can obtain b is given as follows.

$$p_1^S = p(p^2 + 2p(1-p)) + (1-p)((1-p)^2 + 2(1-p)p)$$

The terms on the right side enclosed in p corresponds to the case in which the alternative with the most votes is the correct answer, while the terms on the right side enclosed in $1 - p$ corresponds to the case in which the alternative with the most votes is the incorrect answer. In the terms enclosed in p, term of p^2 corresponds to the case that all the answers of the three workers are correct. Term of $2p(1-p)$ corresponds to the case that two votes are the correct answer and a vote is the incorrect answer. In the terms enclosed in $1-p$ can be explained in the same manner. Terms of $(1-p)^2$, $2(1-p)p$ correspond to the cases that all the answers of the three workers are incorrect, and the case that two votes are the incorrect answer and a vote is the correct answer, respectively.

The probability p_2^S for the case of a sincere worker and an insincere worker, and the probability p_3^S for the case of no sincere worker and two insincere workers can be calculated in the same manner.

By multiplying the probability of occurrence of each case to the values of p_1^S, p_2^S, p_3^S, we can obtain the probability that a sincere worker obtains b, i.e., the value of f_S, as follows.

$$f_S = \alpha^2 p_1^S + 2\alpha(1-\alpha)p_2^S + (1-\alpha)^2 p_3^S$$

We can calculate the probability that an insincere worker can obtain b, f_I, in the same way. In addition, we can calculate the probability that a requester can obtain the correct answer.

Figure 1 shows how to calculate the probabilities of obtaining b for sincere workers, f_S, and that for insincere workers, f_I, in the case that an identical task is given to n workers. We omit the expression for the case that n is an even number due to page limitations.

From the above discussion, we can formalize the decision problem of the requester as the following optimization problem.

Objective function $\max_{a,b,n} v(p^R) - payment$
Constraints
$a + b \cdot f_S \geq r$; sincere
$a + b \cdot f_I < r$; insincere,

$$f_S = \sum_{k=0}^{n-1} {}_{n-1}C_k \alpha^k (1-\alpha)^{n-1-k} p^S(k, n-1-k)$$

$$p^S(k,l) = p\Big(\sum_{i=0}^{(n-1)/2} \sum_{j=0}^{i} {}_kC_j p^{k-j}(1-p)^j {}_lC_{i-j} q^{l-(i-j)}(1-q)^{i-j}\Big) + (1-p)\Big(\sum_{i=0}^{(n-1)/2} \sum_{j=0}^{i} {}_kC_j (1-p)^{k-j} p^j {}_lC_{i-j} (1-q)^{l-(i-j)} q^{i-j}\Big)$$

$$f_I = \sum_{k=0}^{n-1} {}_{n-1}C_k (1-\alpha)^k \alpha^{n-1-k} p^I(k, n-1-k)$$

$$p^I(k,l) = q\Big(\sum_{i=0}^{(n-1)/2} \sum_{j=0}^{i} {}_kC_j q^{k-j}(1-q)^j {}_lC_{i-j} p^{l-(i-j)}(1-p)^{i-j}\Big) + (1-q)\Big(\sum_{i=0}^{(n-1)/2} \sum_{j=0}^{i} {}_kC_j (1-q)^{k-j} q^j {}_lC_{i-j} (1-p)^{l-(i-j)} p^{i-j}\Big)$$

$$p^R = \sum_{k=0}^{n} {}_nC_k \alpha^k (1-\alpha)^{n-k}\Big(\sum_{i=0}^{(n-1)/2} \sum_{j=0}^{i} {}_kC_j p^{k-j}(1-p)^j {}_{n-k}C_{i-j} q^{n-k-(i-j)}(1-q)^{i-j}\Big)$$

Fig. 1 Probability calculation in requesting n workers (n is odd).

where p^R is the probability that the correct answer is obtained and function v represents the requester's value of obtaining the correct answer. Constraints are obtained from the discussion in Sect. 3, although the cost terms, c_S and c_I, are omitted from both sides because the cost is the same in the pay-for-performance payment system and in the fixed-payment system.

Constraints mean that a payment pair should be attractive for sincere workers but not for insincere workers. Insincere workers select an alternative at random and vote for it. Thus, the existence of insincere workers skews the distribution of answers, which makes difficult to know the correct answer. The left side of the expressions of the constraints corresponds to the expected payment in the pay-for-performance payments task, while the right side corresponds to the expected payment in the similar task with the fixed-payment payment system, where r is the payment amount in the fixed-payment task. The utility calculation includes the terms of cost, but the cost is the same in the pay-for-performance payment system and in the fixed-payment system. Thus, the cost terms, c_S and c_I, are omitted from both sides.

We assume that preventing participation of insincere workers can be achieved by motivating insincere workers to choose the fixed-payment tasks that provides a larger payoff compared to the pay-for-performance payment tasks. We implicitly assume that insincere workers can accurately calculate the expected payoff and there is a sufficient number of tasks in the task pool. That is, there are too many tasks for an insincere worker to select all. We assume that this assumption holds in MTurk.

So far, we assumed that the number of workers for a task n is given and focused on the discussion on how to set a payment pair a and b. The probability of obtaining the correct answer p^R depends on α, p, and q but does not directly depend on the payment setting. Thus, to solve the optimization problem mentioned above, we can ignore the term of $v(p^R)$ and focus on the term of payments. Although there are various payment settings that satisfy the above constraints, it seems better to set a to a significantly lower value, e.g., 0, to prevent participation of insincere workers. That is, the setting in which $a = 0$ and $b = r/f_S$ can maximize the objective function.

However, this setting may cause another problem. This gives the same expected payment to sincere workers as with the fixed-payment tasks, but sincere workers may fail to accurately calculate the expected payment. In the discussion on expected payment calculation, we assume that the parameter values of α, p, and q are common knowledge, but we may not be able to assume it in actual applications. For example, if a sincere worker underestimates the value of p, the setting of $a = 0$ and $b = r/f_S$ becomes less attractive for sincere workers compared to the fixed-payment task. To verify whether the above calculation is valid, we carried out experiments on MTurk. Before presenting the experimental results, we discuss the estimation of the parameters in the next section.

Worker collusion is an interesting and significant issue. Further investigation is necessary, but we can discuss its effect. The proposed method attempts to maximize the objective function under the constraint in which the expected payment in the pay-per-performance system is larger than or equal to that in the fixed-payment system. If almost all workers collude with each other and the method classifies these workers into sincere workers, it will set the payment pair in which worker obtains the payment close to that for the fixed-payment task. Thus, if colluding incurs an additional cost to workers such as communication costs, the pay-per-performance task is not preferable for insincere workers.

5. Parameter Estimation

The previous section introduced our method of calculating the expected payment. To perform the calculation, it is necessary to estimate α, p, and q. We propose the following procedure to obtain these values.

1. As a preliminary experiment, publish the task with fixed payment on MTurk.
2. Based on the results of preliminary experiments, calculate the mean values of α, p, and q.
3. Based on our method discussed in Sect. 4, calculate a

payment pair of a and b that satisfy the constraints.

The smallest task unit is called a HIT in MTurk. In a simple task such as sentence labeling, a HIT often consists of several sentences to be labeled instead of a single sentence. This is because the minimum payment is \$0.01 per unit task and handling HITs becomes cumbersome for both the requesters and workers if a unit task is too small. Therefore, We compose a HIT that contained ten sentences to be labeled. Whether it was correct or not could be determined for each of the ten sentences.

A simple index to evaluate the performance of each worker is to count the number of correct answers for the ten sentences. However, the task dealt with in this study may not have a strict correct answer, e.g., 1 + 1 is 2. For example, suppose that ten workers are asked to answer the impression of a sentence; Good or Bad. Assume that the result shows that Good:Bad = 9:1 for sentence X and Good:Bad = 6:4 for sentence Y. If worker 1 answered (X, Y) = (Good, Bad) and worker 2 answered (X, Y) = (Bad, Good), the number of correct answers is one for each worker. However, it is less likely that Bad would be selected for sentence X, compared to if Bad is selected for sentence Y. Therefore, it is not sufficient to determine the performance of a worker only by the number of the correct answers.

To address this problem, we calculate the probability of occurrence of the answer pattern. In the previous example, the probability of occurrence in which Good is selected for sentence X is 0.9 and the probability of occurrence in which Bad is selected for sentence Y is 0.4. Thus, the probability of occurrence in which (X, Y) = (Good, Bad) is calculated as $0.9 \times 0.4 = 0.36$. On the other hand, the probability of occurrence in which (X, Y) = (Bad, Good) is calculated as $0.1 \times 0.6 = 0.06$. Using this value enables us to rank workers 1 and 2.

The next step is to classify the workers as sincere or insincere. If this classification is obtained, p and q can be calculated. However, there is no basis to distinguish between sincere and insincere workers. After all, this problem can be reduced to whether the requester can find useful information in the submitted answers. However, such subjective judgment makes it difficult for other researchers to verify the results. Therefore, we place the criterion in which a worker is classified as sincere his/her performance is better than the result of completely random selection; otherwise, a worker is classified as insincere. In the previous example, if Good/Bad is randomly selected, the probabilities of occurrence in which (X, Y) = (Good, Bad) and (X, Y) = (Bad, Good) become $0.5 \times 0.5 = 0.25$ each. Thus, worker 1 is classified as sincere and worker 2 is classified as insincere.

Once α is obtained, the probability that a worker is sincere is calculated using the Bayes rule. If its value is greater than or equal to 0.5, the worker is determined as sincere; otherwise, the worker is insincere.

The next step to consider is how to determine the bonus payment. The accuracy rate takes continuous values. Therefore, it is not sufficient to determine the bonus payment based on whether a worker is sincere or insincere. This is a way to determine the bonus payment based on the probability of occurrence of the answer pattern, but this is difficult for workers to understand. Therefore, we calculate the bonus payment based on the number of correct answers. Consider that a HIT includes ten sentences. The performance of a worker does not take a binary value of good or bad, but takes step values from no correct answer to ten correct answers. Therefore, the bonus payment is calculated by multiplying the accuracy rate, which is given by dividing the number of correct answers by 10, by b. Because \$0.01 is the minimum unit of payment, a rounded value is used as the actual payment.

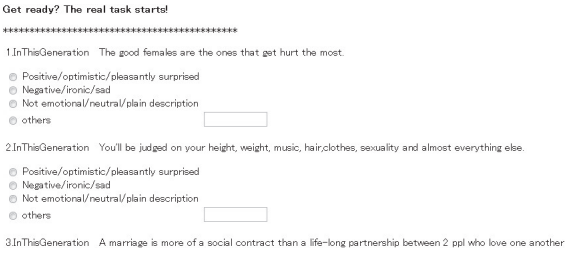


Fig. 2 Task presentation in MTurk.

6. Experiments

6.1 Preliminary Experimental Design

This section describes the design of the preliminary experiment. The purpose of the preliminary experiment was to identify the parameters for pay-for-performance payments, i.e., α, p, and q.

We chose a tweet labeling task for this experiment. The tweets were actual tweets collected from trending topics on Twitter.com. A worker selects a task from the task list in MTurk. In the task list, information such as the task name, name of requester, payment amount, is displayed. We told the workers that this task is a pay-for-performance payment task by posting that "Tweet labels (up to \$0.07 as performance bonus)" as the task name.

If a worker selects this task, then the actual task page is displayed. At the top of the task page, the content of the task is explained by giving examples. The ten tweets to be labeled are placed below the task instructions. A worker was asked to select an appropriate label for a tweet from the following three options (1) positive/optimistic/pleasantly surprised, (2) negative/ironic/sad, (3) not emotional/plain description. A questionnaire is placed at the bottom of the page to collect information on the worker such as country of residence, native language, and interest in trending topics. In addition, an alternative of "others" was provided. If the worker believed that other labels were more appropriate, he/she could choose an alternative of "others", and put an arbitrary label in the text field.

Twenty tweets were collected from each of the trending topics of "OnlyInSingapore" and "InThisGeneration."

Table 1 Results from preliminary experiment.

HIT	#insincere	p	q
ITG1	5	0.627	0.420
ITG2	5	0.687	0.400
OIS1	5	0.620	0.320
OIS2	4	0.663	0.450

(ITG: InThisGeneration, OIS: OnlyInSingapore, #insincere: number of insincere workers (among 20 workers), p: accuracy rate of sincere workers, q: accuracy rate of insincere workers)

In each trending topic, the 20 tweets were divided into two groups of 10 tweets, and a HIT was composed based on each group of 10 tweets. Thus, there were four kinds of HITs.

The fixed payment was set to \$0.05. This means that a worker would be paid \$0.05 as long as he/she completed the HIT regardless of the correctness of the answers. We observed that \$0.05 is often used for similar tasks. In addition, MTurk provides a guideline that \$6 is suitable for an hour task. Thus, we chose \$0.05 as payment.

The tasks in the preliminary experiment were published on MTurk in February 2012. The duration of posting the HITs was set to seven days. From the questionnaire, we learned that 70% of workers lived in India, and other workers lived in the United States, United Kingdom, Canada, Bangladesh, and other countries.

6.2 Results of Preliminary Experiment

As mentioned above, the 20 tweets from two trending topics were divided into two groups for configuring a HIT. The classification of sincere and insincere workers was done according to the method described in the previous section.

Table 1 shows the obtained results in the preliminary experiments. From these results, the parameter values were determined as $\alpha = 0.7625$, $p = 0.6492$, $q = 0.3947$, and the following constraints were obtained to prevent participation of insincere workers when $n = 20$.

$$a + 0.6390b \geq 0.05$$
$$a + 0.3992b < 0.05$$

This calculation was done by using the formulas extended to the case in which the number of alternatives was three. If there are three alternatives, whether an alternative becomes a majority alternative depends on how the remaining two alternatives receive the votes. We assumed that the probabilities of selecting alternatives 2 and 3 are the same if alternative 1 is the correct answer. The formulas are omitted due to page limitations. The values of f_S and f_I depend on the number of requests per task. For example, $f_S = 0.5914$ and $f_I = 0.4345$ if $n = 5$, while $f_S = 0.6220$ and $f_I = 0.4186$ if $n = 10$.

It may be questioned as to why α obtained in the preliminary experiment was used for the main experiment as is. If insincere workers do not select the pay-for-performance tasks because the expected revenue is less than that for the fixed-payment tasks, α increases. However, it is difficult to evaluate this effect in advance. Thus, α obtained in the preliminary experiment is a safe estimation.

6.3 Experimental Design

We verified whether the proposed method of setting a payment pair of a and b works properly. In this experiment, we used the same set of tweets used in the preliminary experiment and published tweet labeling tasks on MTurk. The alternatives were the same as in the preliminary experiment. Choice is three: (1) positive/optimistic/pleasantly surprised, (2) negative/ironic/sad, (3) not emotional/plain. However, the "others" option was not provided because it was rarely selected in the preliminary experiment.

We used α obtained in the preliminary experiment for the evaluation experiments as is. This is because calculating the payments in the pay-for-performance payment system assuming a small α, it is safe in terms of attracting sincere workers and preventing participation of insincere workers.

Based on the data obtained in the preliminary experiments, the payment pair that maximizes the utility of the requester was $a = 0$ and $b = 0.0782$. However, MTurk allows us to use \$0.01 as a payment unit. Thus, it was expected that a payment pair of $a = 0$ and $b = 0.08$ would be best for the requester. In addition, if we set $a = 0.01$, $b = 0.0626$ was obtained by using the formulas. Thus, a payment pair of $a = 0.01$ and $b = 0.07$ was the second best for the requester. Based on these calculations, we selected the following ten payment settings to compare the results; (a, b)=(0, \$0.1), (0, \$0.09), (0, \$0.08), (0, \$0.07), (\$0.01, \$0.08), (\$0.01, \$0.07), (\$0.02, \$0.06), (\$0.02, \$0.05), (\$0.03, \$0.04), and (\$0.04,\$0.02).

The tweet labeling tasks were published on MTurk in July 2012. The period of posting the task was set to seven days.

6.4 Experimental Results

Table 2 lists the experimental results. The expected payments show those calculated using our method, where r_S represents the expected payment for a sincere worker and r_I represents the expected payment for an insincere worker.

Table 2 elucidates the following.

- If a is 0, it is likely to take a longer time to complete the task than in the case of other fixed payments. It may take a longer time; from 1.5 to 3 times. However, there is no problem in collecting 20 workers.
- In terms of the exclusion of insincere workers, the settings in which (a, b) = (\$0.01 \$0.07), (\$0.01 \$0.08), and (\$0.02 \$0.05) are superior. The number of insincere workers was 4.75 on average in the fixed-payment task, while it may be reduced to 1 to 1.5 in the pay-for-performance payment task.

  An interesting point is that the setting of $a = 0.01$ is more effective than that of $a = 0$ in excluding insincere workers. The hypothesis is as follows. First, it is easy to attract the attention of any worker if a payment

Table 2 Results of using pay-for-performance payment system.

Payment settings		Expected payments		InThisGeneration				OnlyInSingapore			
a	b	r_S	r_I	Duration	#insincere	r_S	r_I	Duration	#insincere	r_S	r_I
0	0.07	0.0447	0.0279	1d07h08m	4	0.0385	0.0228	1d08h28m	0	0.0382	N.A.
0	0.08	0.0511	0.0319	17h53m	3	0.0532	0.0293	17h39m	3	0.0489	0.0267
0	0.09	0.0575	0.0359	1d02h47m	4	0.0574	0.0203	1d06h47m	2	0.0535	0.0450
0	0.1	0.0639	0.0399	21h51m	4	0.0756	0.0275	21h49m	1	0.0574	0.0400
0.01	0.07	0.0547	0.0379	18h37m	1	0.0531	0.0380	21h13m	2	0.0508	0.0380
0.01	0.08	0.0611	0.0419	15h15m	1	0.0576	0.0420	17h31m	1	0.0639	0.0340
0.02	0.05	0.0519	0.0400	17h00m	1	0.0518	0.0350	13h08m	1	0.0479	0.0500
0.02	0.06	0.0583	0.0440	17h16m	3	0.0602	0.0420	19h29m	2	0.0570	0.0440
0.03	0.04	0.0556	0.0460	22h37m	3	0.0533	0.0473	22h38m	2	0.0529	0.0420
0.04	0.02	0.0528	0.0480	18h59m	0	0.0530	N.A.	16h13m	2	0.0514	0.0470

(Duration: duration of completing all tasks, #insincere: number of insincere workers (among 20 workers), r_S: payment obtained by sincere worker, r_I: payment obtained by insincere worker)

larger than \$0.05 is displayed regardless if it is a or b. Second, if the a is 0, workers will believe it is not likely to happen that the requester will pays nothing. On the other hand, consider the case in which a is paid, even if it is a very small amount, e.g., \$0.01. In this case, a requester will likely pay no b if the performance is actually bad. A detailed examination of this hypothesis will be included in our future work.

- If we say that the proposed method can successfully predict the outcome when both the predicted and actual values are larger than \$0.05 or smaller than \$0.05, i.e., the predicted and actual values will attract sincere/insincere workers compared to the fixed-payment tasks. The method can successfully predict the outcome except in the following three cases. (a, b) = (\$0, \$0.08) and (\$0.02, \$0.05) for sincere workers and (a, b) = (\$0.02, \$0.05) for insincere workers in "OnlyInSingapore."
  However, the proposed method cannot accurately predict the outcome if we consider the value of rewards. A reason is the method of classifying workers as sincere or insincere workers. For example, if a worker submits three correct answers out of ten questions, the worker is often classified as a sincere worker. In the case of (a, b) = (\$0.02,\$0.05) in "OnlyInSingapore," the payment amounts were reversed for sincere workers and insincere workers. Finding a more appropriate method of classifying workers is also for future work.

To verify whether the decrease of the number of insincere workers does not result from a chance, we conducted statistical hypothesis tests on the experiment data. Here, the data of "InThisGeneration" and the data of "OnlyInSingapore" are dealt with together to investigate the topic-independent properties of the payment systems. First, we conducted Fisher's exact tests in terms of the number of insincere workers. The results show that the number of insincere workers in the payment settings of (\$0.01, \$0.08), (\$0.02, \$0.05), and (\$0.04, \$0.02) in the pay-for-performance payment system are significantly different from that in the fixed payment system ($p = .048$, $p = .025$, $p = .025$, respectively). However, there is no significant difference among the payment settings in the pay-per-performance payment system.

Second, we conducted independent t-tests for the payment settings of (\$0.01, \$0.08), (\$0.02, \$0.05), and (\$0.04, \$0.02) in the pay-for-performance payment system and the fixed payment system. Here, the dependent variables are the probability of occurrence for each answer set. For example, when the obtained distribution of answers are (positive, negative, not emotional) = (0.8, 0.1, 0.1) for question 1 and (positive, negative, not emotional) = (0.6, 0.3, 0.1) for question 2, if a worker answers positive for question 1 and negative for question 2, the probability of occurrence is calculated as $0.8 \times 0.3 = 0.24$. The t-tests show that $t(78) = -3.248$, $p = .002$ for (\$0.01, \$0.08), $t(78) = 0.434$, $p = .665$ for (\$0.02, \$0.05), $t(78) = -2.451$, $p = .016$ for (\$0.04, \$0.02). In the cases of (\$0.01, \$0.08) and (\$0.04, \$0.02), the probability of occurrence is significantly different from that of the fixed payment system. A possible reason about the differences among these payment settings is that if the number of sincere workers increases, it does not necessarily cause the convergence to a single alternative. For example, consider a tweet of "people giving you the death-stare upon seating the reserve seat." First, the word of "death" may catch the attention of workers and cause a feeling of negative. However, by reading the entire sentence it may cause another feeling of positive.

7. Conclusions

We proposed a method for setting a payment pair of the fixed payment and bonus in the pay-for-performance payment system to prevent participation of insincere workers in crowdsourcing. For efficient problem solving in crowdsourcing, it is necessary to prevent participation of insincere workers. First, we proposed how to find a payment pair of fixed payment and bonus in the pay-for-performance payment system that can attract sincere workers but exclude insincere workers. Second, we proposed how to obtain the parameter values of the ratio of sincere workers, accuracy rate of sincere workers, and accuracy rate of insincere workers, which are used as input for the proposed calculation method. Third, we published several tasks on MTurk and evaluated the proposed method. The experimental results suggest the

following. (1) If we set the fixed payment to zero, it will take a slightly longer time to complete the task but rapid accomplishment of tasks is still possible. (2) Setting the fixed payment to a small positive value is more effective than setting it to zero for preventing participation of insincere workers. (3) The method can predict the tendency of whether the payment setting attracts sincere/insincere workers, but it is not as accurate in predicting the payment amount.

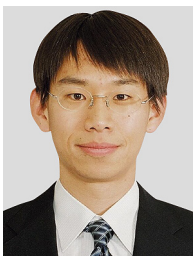
Finally, we assumed that a correct answer is determined by a majority vote. This can be extended to the case in which the first and second majority alternatives are treated as correct answers if there is any difference from random answering. On the other hand, a requester may want to obtain the answer distribution. In this case, it seems difficult to distinguish a sincere worker from an insincere worker if the sincere worker has minority opinions for all the questions. Developing a mechanism able to deal with such a situation is one of our future work.

Acknowledgments

This research was partially supported by a Grant-in-Aid for Scientific Research (S) (24220002, 2012-2016) from Japan Society for the Promotion of Science (JSPS).

References

[1] J. Howe, Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business, Crown Publishing Group, 2008.
[2] O.F. Zaidan and C. Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL '11, pp.1220–1229, 2011.
[3] P.G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," Proc. ACM SIGKDD Workshop on Human Computation, HCOMP '10, pp.64–67, 2010.
[4] A.P. Dawid and A.M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," J. Royal Statistical Society Series C Applied Statistics, vol.28, no.1, pp.20–28, 1979.
[5] E. Huang, H. Zhang, D.C. Parkes, K.Z. Gajos, and Y. Chen, "Toward automatic task design: A progress report," Proc. ACM SIGKDD Workshop on Human Computation, HCOMP '10, pp.77–85, 2010.
[6] W. Mason and D.J. Watts, "Financial incentives and the "performance of crowds"," SIGKDD Explor. Newsl., vol.11, no.2, pp.100–108, May 2010.
[7] J.J. Horton and L.B. Chilton, "The labor economics of paid crowdsourcing," Proc. 11th ACM Conference on Electronic Commerce, EC '10, pp.209–218, 2010.
[8] A.D. Shaw, J.J. Horton, and D.L. Chen, "Designing incentives for inexpert human raters," Proc. ACM 2011 Conference on Computer Supported Cooperative Work, CSCW '11, pp.275–284, 2011.
[9] A. Azaria, Y. Aumann, and S. Kraus, "Automated strategies for determining rewards for human work," Proc. 26th AAAI Conference on Artificial Intelligence, AAAI '12, 2012.
[10] A.G. Parameswaran, H. Garcia-Molina, H. Park, N. Polyzotis, A. Ramesh, and J. Widom, "Crowdscreen: algorithms for filtering data with humans," Proc. 2012 ACM SIGMOD International Conference on Management of Data, pp.361–372, 2012.
[11] J. Witkowski and D.C. Parkes, "Peer prediction without a common prior," Proc. 13th ACM Conference on Electronic Commerce, EC '12, pp.964–981, 2012.
[12] B. Salanié, The Economics of Contracts, MIT Press, 1997.
[13] B. Holmstrom, "Moral hazard in teams," The Bell Journal of Economics, vol.3, no.2, pp.324–340, 1982.
[14] J. Whitehill, P. Ruvolo, T. fan Wu, J. Bergsma, and J. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in Advances in Neural Information Processing Systems 22, ed. Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, and A. Culotta, pp.2035–2043, 2009.
[15] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in Advances in Neural Information Processing Systems 23, ed. J. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, pp.2424–2432, 2010.



Shigeo Matsubara is an associate professor of Department of Social Informatics, Kyoto University. From 1992 to 2006, he was a research scientist of NTT Communication Science Laboratories, NTT. He received his Ph.D. degree in Informatics from Kyoto University. During 2002-2003, he was a visiting researcher at University of California, Berkeley. He was also an adviser of NICT Language Grid project from 2006 to 2007. His research focuses on multi-agent systems and information economics. He has published in Artificial Intelligence Journal and other academic journals.

Meile Wang was a master course student of Department of Social Informatics, Kyoto University.