An Incentive Mechanism to Elicit Truthful Opinions for Crowdsourced Multiple Choice Consensus Tasks

Siyuan Liu*, Chunyan Miao*, Yuan Liu*, Han Yu*, Jie Zhang† and Cyril Leung*
*Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly
†School of Computer Engineering
Nanyang Technological University, Singapore 639798
Email: {syliu, ascymiao, liu.yuan, han.yu, zhangj, cleung}@ntu.edu.sg

***Abstract*—Crowdsourcing is a rapidly growing technology to harness human intelligence to solve problems that are not suitable for automation. It is especially suitable for consensus tasks which collect opinions from human workers to gain insight into real-world phenomena. In these tasks, motivating workers to provide their truthful opinions is a challenging problem. Most existing incentive mechanisms proposed to address this problem assume that workers share a common prior. However, this assumption is not always valid in practice, resulting in that the existing approaches may discourage workers to provide their truthful opinions. In this paper, we propose a novel incentive mechanism – iMET – to elicit truthful opinions for crowdsourced multiple-choice consensus tasks. By incorporating the concepts of worker credibility and similarity, iMET rewards workers for providing truthful opinions without assuming a common prior. Through extensive simulations on the basis of a collected real-world dataset, iMET has been demonstrated to outperform another two widely used incentive mechanisms in eliciting truthful opinions, especially when diverse truthful opinions are held by workers.**

***Keywords*-incentive mechanism; truthful; crowdsourcing; consensus tasks;**

I. Introduction

In recent years, crowdsourcing has gained remarkable popularity and been widely applied to harness human intelligence in various applications such as image labeling [1], story telling [2], and product categorization [3]. Consensus tasks, which are designed to reveal the answers regarding a set of questions, are particularly suitable for crowdsourcing. An example of consensus tasks is the *Galaxy Zoo* project[1] which asks more than 100,000 workers (i.e., human participants who provide opinions to consensus tasks) to classify over one million galaxy images captured through Sloan Digital Sky Survey into different categories. After collecting the redundant opinions, the final answer for a consensus task can be computed according to an opinion aggregation rule (e.g., majority voting).

Eliciting truthful opinions from workers is important for the task requester to make sound decisions. Several incentive mechanisms have been proposed to reward workers for providing truthful opinions [4], [5], [6], [7], [8], [9], [10], [11], [12]. However, most of them assume that the workers share a common belief about the prior probability with respect to the given topic. This is not always true in practice, especially when subjective opinions are sought [13]. In this paper, we propose a novel ***i***ncentive ***m***echanism to ***e***licit ***t***ruthful opinions (iMET) for consensus tasks consisting of closed class questions [14] (i.e., multiple-choice questions where the workers are asked to choose their opinions from a finite and pre-defined list of options). To relax the assumption of workers holding a common prior, iMET employs the concepts of credibility and similarity to probabilistically model the difference among workers' hidden truthful opinions. More specifically, iMET compares a worker's provided opinions with other workers'. On the basis of other workers' credibility, the similarity between workers, and the opinion comparison results, iMET justly rewards the worker in providing opinions even if his/her truthful opinions differ from other workers'. Thus, iMET contributes to eliciting truthful opinions for crowdsourced multiple-choice consensus tasks by relaxing the common prior assumption.

We collect a real-world dataset consisting of human workers' truthful opinions and design simulations based on the collected dataset to compare iMET with another two widely used mechanisms. Through the evaluations, it has been demonstrated that iMET achieves significant advantages over the existing approaches in terms of eliciting truthful opinions under various conditions, especially when workers are holding different truthful opinions.

II. Related Work

The peer prediction method [5] is one of the most well known approaches to eliciting truthful opinions. It pays a worker for providing an opinion regarding an item (e.g., the quality of a product). The payment is calculated according to a proper scoring rule [15], [16] based on the worker and another randomly selected worker's opinions. The peer prediction method is proven to be incentive compatible[2] under

[1] http://zool.galaxyzoo.org/

[2] Incentive compatibility is the property of a mechanism such that workers or agents are rewarded for providing their truthful opinions [5].

the assumption that all the workers hold a common prior which is known to the mechanism. As an implementation of the peer prediction method, the side-payment mechanism [17] has been proposed to elicit truthful reputation opinions (i.e., ratings) in multi-agent based e-marketplaces under the assumption that all agents share the same prior.

Bayesian Truth Serum (BTS) [4] improves the peer prediction method in the sense that the commonly held prior does not need to be known to the mechanism. BTS rewards a worker through the combination of information and prediction scores. The scores are calculated based on the worker and another worker's reported opinions and the worker's prediction on the frequency of other workers reporting a specific opinion. BTS simplifies the peer prediction rule in that the mechanism does not require the knowledge of prior or posterior. However, it still assumes that all workers hold a common prior. Robust Bayesian Truth Serum (RBTS) [6] was proposed to improve BTS in the aspects of avoiding negative payments and relaxing the requirement of worker population size under the common prior assumption. Radanovic and Faltings [9] further extended RBTS to support non-binary opinions.

In an attempt to accommodate the situation where workers have diverse prior beliefs, the shadow peer prediction mechanism [8] was proposed. Before experiencing an item (e.g., a product), a worker first reports his/her private prior belief that other workers will receive items with good quality. A binary rating is then reported after the worker experiences the item. A shadow post belief [8] is calculated considering the reported prior belief and rating. Based on the quadratic scoring rule [15] and the rating provided by another randomly selected worker, the worker is rewarded according to his/her reported prior belief and calculated shadow post belief. The shadow peer prediction mechanism releases the common prior assumption depending on the existence of a temporal factor such that the worker can report prior belief and rating before and after experiencing the item. However, this mechanism is not suitable for crowdsourced consensus tasks because it is difficult for the workers to report their prior belief before being shown with the tasks.

III. The iMET Mechanism

Suppose a task requester is interested in knowing the answers of a consensus task which is in the form of a set of multiple-choice questions, and publish the task on a crowdsourcing platform such as Amazon Mechanical Turk[3] (AMT) or CrowdFlower[4] to solicit opinions from workers. Each worker holds truthful opinions to the questions, and may provide opinions that are different from his/her truthful opinions. When the task is over, each worker will receive a reward for providing an opinion to a question. More specifically, we take into consideration the *credibility* of a worker and the *similarity* of a pair of workers, and then reward a worker based on worker credibility and similarity, and other workers' opinions. The definitions of credibility and similarity are given as follows.

Definition 1: The **credibility** of a worker is the probability that he/she provides a truthful opinion to a question in the task.

Definition 2: The **similarity** between two workers is the probability that their truthful opinions to a question in a task are the same.

Denote the credibility of a worker i as c_i, and the similarity between a pair of workers i and j as s_{ij}. The proposed incentive mechanism is developed as follows.

A. Mechanism Development

Inspired by the side-payment mechanism [17], we reward a worker i with monetary compensation r_{ij}^q for providing an answer (i.e., opinion) to a question q by comparing his/her answer a_i^q with another worker j's ($j \neq i$) answer a_j^q. In a generalized form, i receives a reward of α when i's answer is the same as j's answer, and β when their answers are different. This generalized form is expressed as follows:

$$r_{ij}^q = \begin{cases} \alpha & \text{when} \quad a_i^q = a_j^q, \\ \beta & \text{when} \quad a_i^q \neq a_j^q. \end{cases} \tag{1}$$

We first calculate the probability that $a_i^q = a_j^q$ and $a_i^q \neq a_j^q$. Suppose each question q in the task has K options. Let $\mathcal{H}_i$ and $\mathcal{H}_j$ be the states in which i and j provide truthful answers, respectively; and let $\mathcal{D}_i$ and $\mathcal{D}_j$ be the states in which they provide untruthful answers, respectively. There exist four possible combinations (i.e., $\mathcal{H}_i$ and $\mathcal{H}_j$, $\mathcal{H}_i$ and $\mathcal{D}_j$, $\mathcal{D}_i$ and $\mathcal{H}_j$, $\mathcal{D}_i$ and $\mathcal{D}_j$). As the prime cheating behavior in current crowdsourcing platforms is to randomly select an option as the answer to a question [14], it indicates that workers generally have no special preference over the question options. Thus, in this paper, when a worker is providing an untruthful answer to a question, we assume that he/she randomly selects an option which is different from his/her truthful answer. It is necessary to differentiate a worker's cheating strategy and his/her behavior of providing untruthful answers. For example, suppose a worker's cheating strategy is to provide a random answer to a question, then there is $\frac{K-1}{K}$ probability that he/she is providing an untruthful answer, and $\frac{1}{K}$ probability that he/she is providing the truthful answer though unintentionally. The probability that the answers provided by i and j (i.e., a_i^q and a_j^q) are the same with regard to a question q for each of the four cases is as follows.

Firstly, when i and j both provide their truthful answers, the probability that their answers are the same is:

$$Pr(a_i^q = a_j^q | \mathcal{H}_i, \mathcal{H}_j) = s_{ij} \tag{2}$$

[3]https://www.mturk.com/mturk/welcome

[4]http://www.crowdflower.com/

Secondly, when i provides his/her truthful answer and j does not, there are two scenarios to consider. In the first scenario where i and j's truthful answers are the same, the answers provided by them cannot be the same since j is not providing the truthful answer. In the second scenario where i and j's truthful answers are different, there is a probability of $\frac{1}{K-1}$ that their provided answers are the same. For example, suppose there are $K = 3$ options (i.e., A, B and C), i's truthful answer is A, and j's truthful answer is B. As j is not providing the truthful answer, he/she will choose A or C as the answer with an equal probability. Therefore, there is $\frac{1}{3-1} = 0.5$ probability that j's answer is the same as i's answer. Summarizing the two scenarios, the probability that i and j's answers are the same when i provides his/her truthful answer and j does not is:

$$\begin{aligned} &Pr(a_i^q = a_j^q | \mathcal{H}_i, \mathcal{D}_j) \\ &= s_{ij} \times 0 + (1 - s_{ij}) \times \frac{1}{K-1} \\ &= \frac{1 - s_{ij}}{K-1}. \end{aligned} \tag{3}$$

Thirdly, similar to the above case, when i does not provide his/her truthful answer and j does, the probability that their answers are the same is:

$$Pr(a_i^q = a_j^q | \mathcal{D}_i, \mathcal{H}_j) = \frac{1 - s_{ij}}{K-1}. \tag{4}$$

Fourthly, when neither i nor j provides his/her truthful answer, there are also two scenarios to consider. In the first scenario where i and j's truthful answers are the same, the probability that their answers are the same is $\frac{K-1}{(K-1)^2} = \frac{1}{K-1}$. For example, suppose there are three options (i.e., A, B and C), and i and j's truthful answers are both A. As i is not providing the truthful answer, he/she randomly selects B or C as his/her answer with an equal probability. As j is not providing the truthful answer either, he/she also randomly selects B or C as his/her answer with an equal probability. The probability that i and j both provide B or C as the answer is $\frac{1}{4}$. Then, the probability that i and j's answers are the same is $2 \times \frac{1}{4} = \frac{1}{2}$.

In the second scenario where i and j's truthful answers are different, there is a probability of $\frac{K-2}{(K-1)^2}$ that their answers are the same. For example, given the same three options, suppose i's truthful answer is A and j's truthful answer is B. As i is not providing the truthful answer, he/she randomly selects B or C as his/her answer with an equal probability. As j is not providing the truthful answer either, he/she randomly selects A or C as his/her answer with an equal probability as well. Then, the probability that their answers are the same is $(3-2) \times \frac{1}{(3-1)^2} = \frac{1}{4}$. Summarizing the two scenarios, the probability that i and j's answers are the same when both of them are not providing their truthful answers is:

$$\begin{aligned} &Pr(a_i^q = a_j^q | \mathcal{D}_i, \mathcal{D}_j) \\ &= s_{ij} \times \frac{1}{K-1} + (1 - s_{ij}) \times \frac{K-2}{(K-1)^2} \\ &= \frac{K + s_{ij} - 2}{(K-1)^2}. \end{aligned} \tag{5}$$

Given i and j's credibility c_i and c_j, we have:

$$\begin{aligned} Pr(\mathcal{H}_i, \mathcal{H}_j) =& c_i c_j; \\ Pr(\mathcal{H}_i, \mathcal{D}_j) =& c_i(1 - c_j); \\ Pr(\mathcal{D}_i, \mathcal{H}_j) =& (1 - c_i)c_j; \\ Pr(\mathcal{D}_i, \mathcal{D}_j) =& (1 - c_i)(1 - c_j). \end{aligned} \tag{6}$$

The probability that i and j's answers are the same is:

$$\begin{aligned} Pr(a_i^q = a_j^q) =& Pr(\mathcal{H}_i, \mathcal{H}_j) \times Pr(a_i^q = a_j^q | \mathcal{H}_i, \mathcal{H}_j) \\ &+ Pr(\mathcal{H}_i, \mathcal{D}_j) \times Pr(a_i^q = a_j^q | \mathcal{H}_i, \mathcal{D}_j) \\ &+ Pr(\mathcal{D}_i, \mathcal{H}_j) \times Pr(a_i^q = a_j^q | \mathcal{D}_i, \mathcal{H}_j) \\ &+ Pr(\mathcal{D}_i, \mathcal{D}_j) \times Pr(a_i^q = a_j^q | \mathcal{D}_i, \mathcal{D}_j). \end{aligned} \tag{7}$$

The probability that i and j's answers are the different is:

$$\begin{aligned} Pr(a_i^q \neq a_j^q) =& Pr(\mathcal{H}_i, \mathcal{H}_j) \times Pr(a_i^q \neq a_j^q | \mathcal{H}_i, \mathcal{H}_j) \\ &+ Pr(\mathcal{H}_i, \mathcal{D}_j) \times Pr(a_i^q \neq a_j^q | \mathcal{H}_i, \mathcal{D}_j) \\ &+ Pr(\mathcal{D}_i, \mathcal{H}_j) \times Pr(a_i^q \neq a_j^q | \mathcal{D}_i, \mathcal{H}_j) \\ &+ Pr(\mathcal{D}_i, \mathcal{D}_j) \times Pr(a_i^q \neq a_j^q | \mathcal{D}_i, \mathcal{D}_j). \end{aligned} \tag{8}$$

Then, the expected value of r_{ij}^q is calculated as:

$$\mathbf{E}(r_{ij}^q) = Pr(a_i^q = a_j^q) \times \alpha + Pr(a_i^q \neq a_j^q) \times \beta. \tag{9}$$

Substituting Eqs. (2) – (8) into Eq. (9), we further have:

$$\begin{aligned} \mathbf{E}(r_{ij}^q) =& \alpha \big[c_i c_j s_{ij} + c_i(1 - c_j)\big(\frac{1 - s_{ij}}{K-1}\big) \\ &+ (1 - c_i)c_j\big(\frac{1 - s_{ij}}{K-1}\big) \\ &+ (1 - c_i)(1 - c_j)\big(\frac{K + s_{ij} - 2}{(K-1)^2}\big)\big] \\ &+ \beta \big[c_i c_j(1 - s_{ij}) + c_i(1 - c_j)\big(1 - \frac{1 - s_{ij}}{K-1}\big) \\ &+ (1 - c_i)c_j\big(1 - \frac{1 - s_{ij}}{K-1}\big) \\ &+ (1 - c_i)(1 - c_j)\big(1 - \frac{K + s_{ij} - 2}{(K-1)^2}\big)\big]. \end{aligned} \tag{10}$$

To incentivize a worker i to provide truthful answers, $\mathbf{E}(r_{ij}^q)$ should monotonically increase with his/her credibility c_i which is the probability of i providing truthful answers. Therefore, the first order derivation of Eq. (10) with respect to c_i should be greater than zero. It is formally expressed as follows:

$$\frac{\partial \mathbf{E}(r_{ij}^q)}{\partial c_i} = \frac{(Kc_j - 1)(Ks_{ij} - 1)(\alpha - \beta)}{(K-1)^2} > 0. \tag{11}$$

From Eq. (11), it can be derived that $\alpha > \beta$ if $(Kc_j - 1)(Ks_{ij} - 1) > 0$, and $\alpha < \beta$ if $(Kc_j - 1)(Ks_{ij} - 1) < 0$. It is impossible to find a universal solution of α and β to ensure Eq. (11) to be true given any c_j and s_{ij}. Therefore, we propose two rewarding schemes $\mathcal{M}_1$ and $\mathcal{M}_2$ as follows:

- $\mathcal{M}_1$ is triggered when $(Kc_j - 1)(Ks_{ij} - 1) > 0$, and

$$r_{ij}^q = \begin{cases} \alpha_1 & \text{when} \quad a_i^q = a_j^q, \\ \beta_1 & \text{when} \quad a_i^q \neq a_j^q, \end{cases} \tag{12}$$

  where $\alpha_1 > \beta_1$;
- $\mathcal{M}_2$ is triggered when $(Kc_j - 1)(Ks_{ij} - 1) < 0$, and

$$r_{ij}^q = \begin{cases} \alpha_2 & \text{when} \quad a_i^q = a_j^q, \\ \beta_2 & \text{when} \quad a_i^q \neq a_j^q, \end{cases} \tag{13}$$

  where $\alpha_2 < \beta_2$.

Furthermore, when $(Kc_j - 1)(Ks_{ij} - 1) = 0$, worker i should get the same reward under the above two rewarding schemes at the borderline points $c_j = \frac{1}{K}$ or $s_{ij} = \frac{1}{K}$ to ensure $\mathbf{E}(r_{ij}^q)$ to be continuous. Thus, we have the following equation:

$$\mathbf{E}(r_{ij}^q)|_{(\alpha_1,\beta_1)} = \mathbf{E}(r_{ij}^q)|_{(\alpha_2,\beta_2)}, \tag{14}$$

when $c_j = \frac{1}{K}$ or $s_{ij} = \frac{1}{K}$. Here, $\mathbf{E}(r_{ij}^q)|_{(\alpha_1,\beta_1)}$ and $\mathbf{E}(r_{ij}^q)|_{(\alpha_2,\beta_2)}$ are the expected value of r_{ij}^q when $\mathbf{E}(r_{ij}^q)$ is calculated using α_1 and β_1, and α_2 and β_2, respectively. By substituting $c_j = \frac{1}{K}$ or $s_{ij} = \frac{1}{K}$ into $\mathbf{E}(r_{ij}^q)|_{(\alpha_1,\beta_1)}$ and $\mathbf{E}(r_{ij}^q)|_{(\alpha_2,\beta_2)}$, they can be calculated as below:

$$\mathbf{E}(r_{ij}^q)|_{(\alpha_1,\beta_1)} = (\alpha_1 - \beta_1) \times \frac{1}{K} + \beta_1, \tag{15}$$

$$\mathbf{E}(r_{ij}^q)|_{(\alpha_2,\beta_2)} = (\alpha_2 - \beta_2) \times \frac{1}{K} + \beta_2. \tag{16}$$

To maintain the ex-post individual rationality[5] of the proposed mechanism, the simplest solution of β_1 and α_2 is $\beta_1 = 0$ and $\alpha_2 = 0$ as $\alpha_1 > \beta_1 \geq 0$ and $\beta_2 > \alpha_2 \geq 0$. Substituting $\beta_1 = 0$ and $\alpha_2 = 0$ into Eqs. (14) – (16), the relationship between α_1 and β_2 can be derived as:

$$\alpha_1 = (K - 1)\beta_2. \tag{17}$$

Setting α_1 as 1 (per unit of reward) and merging the borderline points to one side, r_{ij}^q can be determined as shown in Table I.

Table I
REWARD FOR WORKER i THROUGH COMPARISON WITH WORKER j

	$s_{ij} \geq \frac{1}{K}$		$s_{ij} < \frac{1}{K}$	
	$a_i^q = a_j^q$	$a_i^q \neq a_j^q$	$a_i^q = a_j^q$	$a_i^q \neq a_j^q$
$c_j \geq \frac{1}{K}$	$\alpha_1 = 1$	$\beta_1 = 0$	$\alpha_2 = 0$	$\beta_2 = \frac{1}{K-1}$
$c_j < \frac{1}{K}$	$\alpha_2 = 0$	$\beta_2 = \frac{1}{K-1}$	$\alpha_1 = 1$	$\beta_1 = 0$

According to Table I, worker i receives a reward of 1 when the answers for a question from workers i and j are the same under the conditions: 1) worker j's credibility is not less than $\frac{1}{K}$ (i.e., $c_j \geq \frac{1}{K}$) and the similarity between i and j is not less than $\frac{1}{K}$ (i.e., $s_{ij} \geq \frac{1}{K}$); or 2) $c_j < \frac{1}{K}$ and $s_{ij} < \frac{1}{K}$. On the other hand, worker i receives a reward of $\frac{1}{K-1}$ when the answers for a question from workers i and j are different under the conditions: 1) $c_j \geq \frac{1}{K}$ and $s_{ij} < \frac{1}{K}$; or 2) $c_j < \frac{1}{K}$ and $s_{ij} \geq \frac{1}{K}$.

Finally, with the reward r_{ij}^q determined for worker i given worker j, the reward r_i^q that worker i receives for providing an answer to question q is calculated through averaging r_{ij}^q with respect to all $j \in I^q - \{i\}$, where I^q is the set of workers providing answers to question q:

$$r_i^q = \frac{\sum_{j \in I^q, j \neq i} r_{ij}^q}{|I^q| - 1}. \tag{18}$$

Theoretically speaking, iMET is a dominant strategy incentive compatible mechanism[6]. Through the derivation of Table I, we know that the expected value of r_{ij}^q monotonically increases with c_i beyond the borderline points at $c_j = \frac{1}{K}$ or $s_{ij} = \frac{1}{K}$ (according to Eq. (11)). As r_i^q is the average of r_{ij}^q according to Eq. (18), the expected value of r_i^q also monotonically increases with c_i beyond the borderline points at $c_j = \frac{1}{K}$ or $s_{ij} = \frac{1}{K}$. Therefore, worker i can obtain the maximum reward by providing truthful answers. This analysis applies to all the workers. Therefore, providing truthful answers is a worker's dominant strategy, and iMET is a dominant strategy incentive compatible mechanism.

B. Mechanism Implementation

As we do not know what workers' truthful opinions are, we have to estimate a worker's credibility and the similarity between a pair of workers based on the observed information to implement the proposed mechanism in reality.

We use the technique of *gold standard* (GS) questions [20] to estimate a worker's credibility. The technique of GS questions is commonly used in crowdsourcing to measure the quality of the opinions provided by the workers [20], [21], [22]. Two important features of GS questions are: 1) there are well-accepted consensus answers for the questions. It is thus reasonable to assume that the workers have the same truthful answers to the GS questions, which are known to the mechanism; and 2) the GS questions can be disguised to be undetectable from the questions to which the task requester is soliciting answers [22], which we refer to as *non-gold standard* (NGS) questions. GS and NGS questions differ in that workers may have different truthful answers for the NGS questions because of subjective beliefs. The combined GS and NGS questions as well as their alternative options are randomly sorted to be presented to each individual worker. The workers are assumed to behave consistently over the combined questions [14]. Suppose there are N^g GS

[5]A mechanism is ex-post individually rational when the utility of each player is non-negative [18].

[6]A mechanism is dominant strategy incentive compatible if truth-revelation is a dominant strategy equilibrium [19].

questions, and $N^{\neg g}$ NGS questions. The credibility c_i of a worker i can be estimated as below:

$$c_i \doteq \frac{N_i^g}{N^g}, \tag{19}$$

where N_i^g is the number of GS questions to which i provides the same answers as the answers known to the mechanism.

Similar to credibility, the similarity between two workers i and j has to be estimated based on the observed information. Denote γ as the probability that the answers from i and j are the same for an NGS question q. We have the following equation:

$$\begin{aligned}\gamma = & Pr(\mathcal{H}_i, \mathcal{H}_j) \times Pr(a_i^q = a_j^q | \mathcal{H}_i, \mathcal{H}_j) \\ & + Pr(\mathcal{H}_i, \mathcal{D}_j) \times Pr(a_i^q = a_j^q | \mathcal{H}_i, \mathcal{D}_j) \\ & + Pr(\mathcal{D}_i, \mathcal{H}_j) \times Pr(a_i^q = a_j^q | \mathcal{D}_i, \mathcal{H}_j) \\ & + Pr(\mathcal{D}_i, \mathcal{D}_j) \times Pr(a_i^q = a_j^q | \mathcal{D}_i, \mathcal{D}_j).\end{aligned} \tag{20}$$

Substituting Eqs (2) – (6) into Eq. (20), s_{ij} can be solved as:

$$s_{ij} = \begin{cases} (\gamma - \frac{1}{K})\frac{(K-1)^2}{(1-Kc_i)(1-Kc_j)} + \frac{1}{K} & c_i \neq \frac{1}{K} \wedge c_j \neq \frac{1}{K}, \\ \text{any} & c_i = \frac{1}{K} \vee c_j = \frac{1}{K}. \end{cases} \tag{21}$$

In the special case of $c_i = \frac{1}{K}$ or $c_j = \frac{1}{K}$, where i or j adopts the strategy of randomly selecting an option from the K available options for each question in the task, the probability of i and j giving the same answers for an NGS question is equal to $\frac{1}{K}$, no matter what s_{ij} is. It is worth pointing out that the similarity between a pair of workers is independent of their credibility. Eq. (21) is a theoretical measurement of s_{ij}, where γ is dependent on s_{ij}, c_i, and c_j.

As c_i and c_j can be estimated by Eq. (19), and γ can be estimated as $\frac{N_{ij}^{\neg g}}{N^{\neg g}}$ where $N_{ij}^{\neg g}$ is the number of NGS questions to which i and j provide the same answers, the similarity between i and j can be estimated as:

$$s_{ij} \doteq \begin{cases} (\frac{N_{ij}^{\neg g}}{N^{\neg g}} - \frac{1}{K})\frac{(K-1)^2}{(1-K\frac{N_i^g}{N^g})(1-K\frac{N_j^g}{N^g})} + \frac{1}{K} & \frac{N_i^g}{N^g} \neq \frac{1}{K} \wedge \frac{N_j^g}{N^g} \neq \frac{1}{K}, \\ \frac{1}{K} & \frac{N_i^g}{N^g} = \frac{1}{K} \vee \frac{N_j^g}{N^g} = \frac{1}{K}. \end{cases} \tag{22}$$

When $\frac{N_i^g}{N^g} = \frac{1}{K}$ or $\frac{N_j^g}{N^g} = \frac{1}{K}$, we set s_{ij} to be $\frac{1}{K}$ for simplicity. It is noted that similarity is actually defined over NGS questions, and thus measured based on the worker opinions for NGS questions. In experiments, we use Eqs. (19) and (22) to estimate worker credibility and similarity, respectively.

C. An Example

To help readers to better understand the proposed mechanism, we provide a walk-through example in this part. Suppose there are three workers involved in answering the same set of questions, and each question has $K = 2$ options. Workers 1 and 2 always provide their truthful answers, i.e., $c_1 = c_2 = 1$, and worker 3 adopts the strategy of selecting an option at random for each question, i.e., $c_3 = \frac{1}{K} = \frac{1}{2}$. Workers 1 and 2 hold different truthful answers for all questions, and worker 3 holds the same truthful answers as worker 1. According to Eq. (22), the similarity among the workers can be estimated as $s_{12} = s_{21} = 0$, $s_{13} = s_{31} = \frac{1}{K} = \frac{1}{2}$, and $s_{23} = s_{32} = \frac{1}{K} = \frac{1}{2}$. The expected rewards received by each worker for answering a question q are listed in Table II.

Table II
THE REWARDS FOR ANSWERING A SINGLE QUESTION q

i	1	2	3
$\mathbf{E}(r_{i1}^q)$	-	**1**	**0.5**
$\mathbf{E}(r_{i2}^q)$	1	-	0.5
$\mathbf{E}(r_{i3}^q)$	0.5	0.5	-
$\mathbf{E}(r_i^q)$	0.75	**0.75**	**0.5**

As examples, the expected rewards $\mathbf{E}(r_{21}^q)$ and $\mathbf{E}(r_{31}^q)$ for answering question q, which are highlighted in **bold** in Table II, are explained as follows.

- $\mathbf{E}(r_{21}^q)$ — reward for worker 2 by comparing with worker 1. Since $c_1 = 1$ and $s_{21} = 0$, the probability that they provide different answers is 1, and iMET rewards $r_{21}^q = \frac{1}{K-1} = 1$ to worker 2 according to Table I. Then, we have $\mathbf{E}(r_{21}^q) = 1 \times \frac{1}{K-1} = 1$.
- $\mathbf{E}(r_{31}^q)$ — reward for worker 3 by comparing with worker 1. The probability that they provide the same answers is $\frac{1}{K}$ as $c_3 = \frac{1}{K}$. Since $c_1 = 1$ and $s_{31} = \frac{1}{K}$, iMET rewards $r_{31}^q = 1$ to worker 3 according to Table I. Then, we have $\mathbf{E}(r_{31}^q) = \frac{1}{K} \times 1 = 0.5$.

Therefore, the expected reward for worker 2 who provides his/her truthful answers is higher than that of worker 3 who does not, no matter whether their truthful answers are the same or not, i.e., $\mathbf{E}(r_2^q) > \mathbf{E}(r_3^q) \geq 0$.

IV. EXPERIMENTAL EVALUATION

We conduct experiments to examine whether iMET can elicit truthful opinions under different realistic conditions based on real-world data. Though some datasets are available to study consensus tasks, they cannot be used in our experiments due to the lack of information about whether the workers provide truthful opinions. Thus, we collected a new dataset and conducted experiments based on it.

A. Experimental Settings

We collected questions from test.baidu.com[7] (a leading crowdsourcing platform in China) and distributed them to postgraduate students and faculty members (i.e., workers) in a university to solicit opinions. For each question, there are two different images. Workers are asked to provide their

[7]Please refer to http://test.baidu.com/crowdtest/eva/testerView/eva_id/5448/prePage for the entire question set.

truthful opinions on which image is clearer. An example question is shown in Figure 2. In total, we collected valid answers from 45 workers each answering 100 questions. As the number of the participants in our experiment is manageable and all of them are located in the same university, this allows us to conduct face-to-face interviews with each of the workers after he/she completed the tasks to verify if they are providing truthful answers. In the interviews, we asked them the specific reasons for their provided opinions with respect to some randomly selected questions. All participants stated during the interviews that they provided their truthful opinions for all the questions. Thus, this dataset is considered to include 45 sets of truthful answers regarding the 100 questions.

There are 10 questions to which all the workers provided the same answers. Thus there is no controversy with regard to these 10 questions, and they are used as GS questions[8]. Through analyzing the collected data, we obtained the similarity distribution as shown in Figure 1, by comparing the answers provided by each pair of workers (45×44 pairs in total). It can be observed that most of the similarity

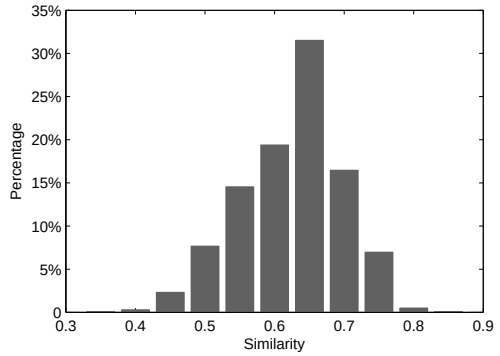


Figure 1. Similarity Distribution among Workers in Our Real-world Dataset

values fall between 0.3 and 0.8 with the mean value around 0.65. 10% of the similarity values fall below the mechanism borderline $\frac{1}{K} = 0.5$ ($K = 2$ as there are two options for each question). These observations suggest that there are differences among the workers' truthful opinions. For example, for the question shown in Figure 2, there are 28 workers reported that the image on the left is clearer than the one on the right, while the other 17 workers reported that the image on the right is clearer.

The real-world dataset provides us with an understanding about the characteristics of the truthful opinions from a practical perspective. In order to compare the proposed mechanism with other mechanisms under different conditions, we design simulation experiments based on the real-world dataset. We synthesize three simulated worker agent[9] populations, and each population includes 45 worker agents. The credibility settings of the worker agents are shown in Figure 3: 1) high-level credibility, where the average credibility $\overline{c}$ of worker agents is $\frac{3}{4}$; 2) medium-level credibility, where $\overline{c}$ is $\frac{1}{2}$; and 3) low-level credibility, where $\overline{c}$ is $\frac{1}{4}$. In Figure 3, the x-axis is the credibility value and y-axis is the number of the worker agents with each specific credibility value.

Figure 2. An Example Question

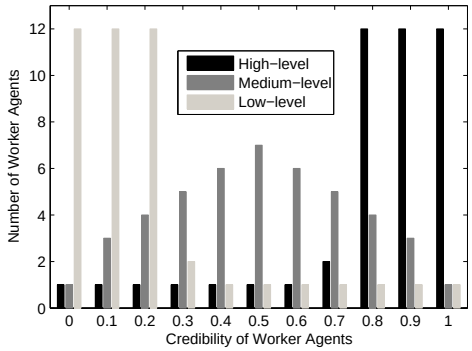


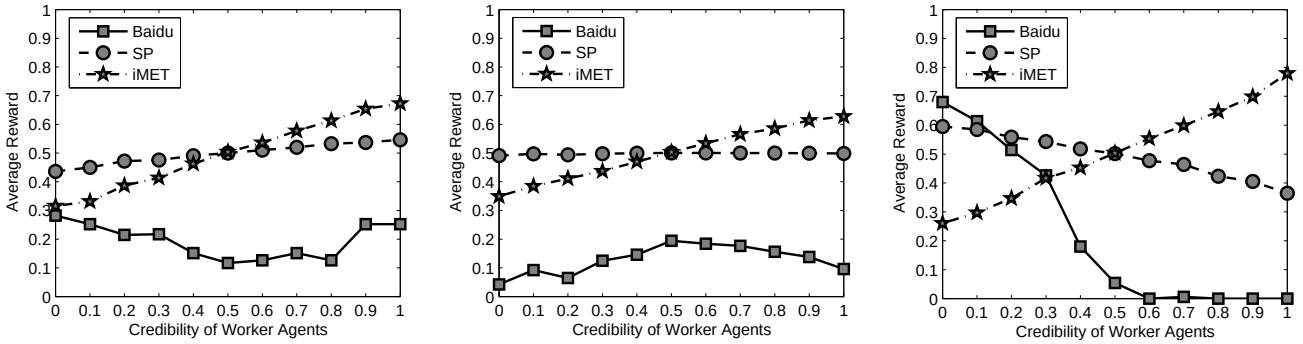
Figure 3. Credibility Settings in the Simulation

Each worker agent adopts a set of answers in the collected real-world dataset as its truthful answers, and is randomly assigned with a credibility value in each simulation which determines how it will answer the questions according to its truthful answers. For example, if a worker agent's credibility is 0.8 in a simulation, it will answer 80 randomly selected questions using its truthful answers, and select the other option as the answer for each of the remaining 20 questions.

In the experiments, we compare iMET with another two widely used incentive mechanisms – the static reward

[8]In reality, the GS questions can be composed from previous tasks or generated through programming [21].

[9]To clearly describe the experiments, the individuals in the real-word dataset are called "workers" and the corresponding simulated individuals are called "worker agents".



(a) High-Level (b) Medium-Level (c) Low-Level

Figure 4. Experimental Results under the Three Settings

mechanism used by test.baidu.com (Baidu), and the side-payment mechanism (SP) [17]:

- Baidu: It rewards a worker agent based on its accuracy rate (i.e., the percentage of a worker agent's answers being consistent with the majority opinions). Specifically, if more than 80% of a worker agent's answers are accurate, it receives a reward of 1 for each question; if $60\% - 80\%$ of the worker agent's answers are accurate, it receives a reward of 0.6 for each question; otherwise, it receives nothing;
- SP: It rewards a worker agent by comparing its answer to a question with another worker agent's answer to the same question. If their answers are the same, each worker agent receives a reward of 1; otherwise, they receive nothing.

To achieve high statistical accuracy, the results presented in this section are obtained by averaging from running the simulations under each setting for 100 times.

B. Experimental Results and Discussions

The average reward that a worker agent receives for answering a question is shown in Figure 4. It can be seen that the reward provided by iMET monotonically increases with worker credibility under all settings. Under the high-level credibility setting, the reward provided by SP also monotonically increases with worker credibility as shown in Figure 4(a). The Baidu mechanism results in a fluctuating curve, which sends mixed signals to workers, potentially causing confusions on whether they should provide truthful opinions to gain more rewards. Under the medium-level credibility setting as shown in Figure 4(b), SP provides almost the same rewards no matter what a worker agent's credibility is. The Baidu mechanism still results in a fluctuating curve. Under the low-level credibility setting as shown in Figure 4(c), Both SP and Baidu present opposite incentives by giving more rewards to worker agents with lower credibility.

In addition, we also compare the effectiveness of the three mechanisms in eliciting truthful opinions using the *average incentive strength* metric, Ψ, which is calculated as:

$$\Psi = \underset{c_i \neq c_j}{\text{avg}} \left\{ \frac{\overline{r}_i - \overline{r}_j}{c_i - c_j} \right\}, \tag{23}$$

where $\overline{r}_i$ is the average reward that a worker agent i receives for answering a question. A large and positive Ψ value suggests that a mechanism provides a strong average incentive in eliciting truthful opinions. A negative Ψ value means that a mechanism encourages worker agents to provide untruthful opinions.

Table III shows the Ψ values of the three mechanisms under the three credibility settings. It can be seen that iMET always achieve the highest positive average incentive strength, suggesting that the worker agents have the strongest incentive to provide truthful opinions using iMET. Under the high-level and medium-level credibility settings, Baidu and SP both achieve positive Ψ values, but smaller than that achieved by iMET. Under the low-level credibility setting, SP and Baidu both result in negative Ψ values, suggesting that they encourage worker agents to provide untruthful opinions.

Table III
THE AVERAGE INCENTIVE STRENGTH (Ψ)

Credibility Settings	Baidu	SP	iMET
High-level	0.1065	0.0979	**0.3642**
Medium-Level	0.1086	0.0064	**0.2953**
Low-level	-0.8913	-0.1931	**0.4727**

According to the results shown in Figure 4 and Table III, it can be observed that iMET is the only mechanism consistently providing incentives to worker agents to report truthful opinions under all experimental settings. Its advantage over existing approaches is most significant in the low-level credibility setting, where the majority of the worker agents do not provide their truthful opinions. Though SP also provides more rewards to worker agents with higher credibility in the high-level credibility setting, the average incentive strength is not as strong as iMET due to its failure to account for the

existence of the different truthful opinions held by worker agents.

It is interesting to note that SP provides opposite incentives in the low-level credibility setting. The reason is that SP has two equilibria, and majority worker agents providing untruthful opinions will cause the strategy of providing untruthful opinions to become more beneficial than that of providing truthful opinions. Therefore, opposite incentives can be observed for SP when a majority proportion of the worker agents report untruthful opinions. The Baidu mechanism does not work under all three credibility settings. As Baidu adopts the majority rule, the existence of different truthful opinions results in the situation in which the calculated accuracy rate of a worker agent (i.e., the percentage of the questions to which the answers provided by a worker agent are the same as majority worker agents' answers) may not positively correlate with whether the worker agent provides truthful opinions. Thus, Baidu cannot always provide more rewards to the worker agents with higher credibility.

V. Conclusions

In this paper, we propose a novel incentive mechanism to elicit truthful opinions – iMET – for crowdsourced consensus tasks in the form of multi-choice questions. It advances the state-of-the-art in the following two aspects: 1) it relaxes the assumption that all the workers hold a common prior; and 2) it provides stronger incentives for workers to share their truthful opinions compared to existing mechanisms. In this way, iMET is more suitable in practical crowdsourcing systems where workers come from diverse backgrounds and the assumption of a common prior belief is not valid. Experimental results based on real-world data demonstrate that iMET can effectively elicit truthful opinions and significantly outperforms related works when diverse truthful opinions are held by workers.

Acknowledgment

This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office.

References

[1] L. V. Ahn and L. Dabbish, "Designing games with a purpose," *Communications of the ACM*, vol. 51, no. 8, pp. 58–67, 2008.

[2] B. Li, S. Lee-Urban, G. Johnston, and M. O. Riedl, "Story generation with crowdsourced plot graphs," in *AAAI 2013*.

[3] P. G. Ipeirotis, "Analyzing the amazon mechanical turk marketplace," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 17, no. 2, pp. 16–21, 2010.

[4] D. Prelec, "A bayesian truth serum for subjective data," *Science*, vol. 306, no. 5695, pp. 462–466, 2004.

[5] N. Miller, P. Resnick, and R. Zeckhauser, "Eliciting informative feedback: The peer-prediction method," *Management Science*, vol. 51, no. 9, pp. 1359–1373, 2005.

[6] J. Witkowski and D. C. Parkes, "A robust bayesian truth serum for small populations," in *AAAI 2012*.

[7] E. Kamar and E. Horvitz, "Incentives for truthful reporting in crowdsourcing," in *AAMAS 2012*, pp. 1329–1330.

[8] J. Witkowski and D. C. Parkes, "Peer prediction without a common prior," in *EC 2012*, pp. 964–981.

[9] G. Radanovic and B. Faltings, "A robust bayesian truth serum for non-binary signals," in *AAAI 2013*.

[10] H. Yu, C. Miao, Z. Shen, C. Leung, Y. Chen, and Q. Yang, "Efficient task sub-delegation for crowdsourcing," in *AAAI 2015*, pp. 1305–1311.

[11] H. Yu, Z. Shen, C. Miao, and B. An, "A reputation-aware decision-making approach for improving the efficiency of crowdsourcing systems," in *AAMAS 2013*.

[12] ——, "Challenges and opportunities for trust management in crowdsourcing," in *IAT 2012*, pp. 486–493.

[13] J. Witkowski, Y. Bachrach, P. Key, and D. C. Parkes, "Dwelling on the negative: Incentivizing effort in peer prediction," in *HCOMP 2013*.

[14] G. Sautter and K. Böhm, "High-throughput crowdsourcing mechanisms for complex tasks," *Social Network Analysis and Mining*, vol. 3, no. 4, pp. 873–888, 2013.

[15] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.

[16] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, pp. 359–378, 2007.

[17] R. Jurca and B. Faltings, "Towards incentive-compatible reputation management," *Trust, Reputation and Security: Theories and Practice, Lecture Notes in AI*, vol. 2631, pp. 138–147, 2003.

[18] J. V. Neumann and O. Morgenstern, *Theory of Games and Economic Behaviour*. Princeton University Press, 1944.

[19] W. Vickrey, "Counterspeculation, auctions, and competitive sealed tenders," *The Journal of Finance*, vol. 16, no. 1, pp. 8–37, 1961.

[20] D. Oleson, A. Sorokin, G. Laughlin, V. Hester, J. Le, and L. Biewald, "Programmatic gold: Targeted and scalable quality assurance in crowdsourcing," in *AAAI Workshop on Human Computation*, 2011, pp. 43–48.

[21] S. Komarov, K. Reinecke, and K. Z. Gajos, "Crowdsourcing performance evaluations of user interfaces," in *SIGCHI 2013*, pp. 207–216.

[22] P. Venetis and H. Garcia-Molina, "Quality control for comparision microtasks," in *Workshop on Crowdsourcing and Data Mining*, 2012, pp. 15–21.