The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Optimal Spot-Checking for Improving Evaluation Accuracy of Peer Grading Systems

Wanyuan Wang,[1†*] **Bo An,**[2‡] **Yichuan Jiang**[1,§]
[1]School of Computer Science and Engineering, Southeast University, China
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[†,§]{wywang, yjiang}@seu.edu.cn, [‡]boan@ntu.edu.sg

Abstract

Peer grading, allowing students/peers to evaluate others' assignments, offers a promising solution for scaling evaluation and learning to large-scale educational systems. A key challenge in peer grading is motivating peers to grade diligently. While existing spot-checking (SC) mechanisms can prevent peer collusion where peers coordinate to report the uninformative grade, they unrealistically assume that peers have the same grading reliability and cost. This paper studies the general *Opt*imal *S*pot-*C*hecking (OptSC) problem of determining the probability each assignment needs to be checked to maximize assignments' evaluation accuracy aggregated from peers, and takes into consideration 1) peers' heterogeneous characteristics, and 2) peers' strategic grading behaviors to maximize their own utility. We prove that the bilevel OptSC is NP-hard to solve. By exploiting peers' grading behaviors, we first formulate a single level relaxation to approximate OptSC. By further exploiting structural properties of the relaxed problem, we propose an efficient algorithm to that relaxation, which also gives a good approximation of the original OptSC. Extensive experiments on both synthetic and real datasets show significant advantages of the proposed algorithm over existing approaches.

Introduction

Peer grading, allowing students/peers to evaluate others' assignments, not only helps the instructor bring qualified feedbacks to classrooms but also helps students self-study using other peers' solutions (Raman and Joachims 2014; Caragiannis, Krimpas, and Voudouris 2015). Besides its direct application to educational systems (e.g., Coursera and EdX), peer grading is also useful in reputation and crowdsourcing systems where it is difficult to evaluate peers' contributions (Witkowski et al. 2013; Ho, Frongillo, and Chen 2016). One of the key challenges in peer grading is how to motivate students to grade assignments diligently (Sadler and Good 2006; Liu and Chen 2016; Shnayder et al. 2016). Well-known Peer Prediction (Miller, Resnick, and Zeckhauser 2005; Radanovic and Faltings 2015; Kong, Ligett, and Schoenebeck 2016) and Bayesian Truth Serum (Prelec 2004; Witkowski and Parkes 2012) mechanisms work by paying a reward to a peer if his (belief) report is predicted to be correct based on other peers' reports. However, most of these incentive mechanisms are vulnerable to peer collusion, where peers could have a motivation to report the uninformative grade by a pre-agreed grading rule (Gao, Wright, and Leyton-Brown 2016).

Spot-checking (SC) mechanisms can prevent the peer collusion issue (Jurca and Faltings 2005; Carbonara et al. 2015). In SC, the instructor checks some assignments by himself and offers a reward to a peer who grades diligently. Existing SC researches have shown that under the special setting where peers are homogeneous with the same grading reliability and cost, a simple SC mechanism, such as random (Wright, Thornton, and Leyton-Brown 2015) or uniform (Gao, Wright, and Leyton-Brown 2016), is efficient to motivate peers to be diligent. However, in practice, peers often have heterogeneous grading reliability and cost (Dasgupta and Ghosh 2013; Agarwal et al. 2017), for example, in an empirical online peer grading test (Kulkarni et al. 2013), peers with suitable backgrounds have 25% disagreements in average, and varied by peers, about 75% grading is completed in 9.5 minutes to 17.3 minutes. Under such a general setting, randomized SC mechanisms might perform poorly (as we show in this paper) on maximizing assignments' evaluation accuracy. The focus of this paper is finding the optimal SC mechanism to maximize assignments' evaluation accuracy in a practical setting with heterogeneous peers.

The first contribution of this paper is a general SC model for peer grading systems with strategic and heterogeneous peers. We assume that the instructor has a spot-checking budget, denoting the maximum number of assignments he is capable of checking. Given budget K, the instructor's objective is maximizing assignments' evaluation accuracy aggregated from peers, which can be formulated as a bilevel optimal SC (OptSC) problem. In the upper level, the instructor determines the probability each assignment needs to be checked, and in the lower level, peers are strategic that choose the optimal grading strategies to maximize their own utility. We prove it is NP-hard to solve OptSC. To address the NP-hardness, our second contribution is formulating a single level relaxation to approximate the bilevel OptSC. We show that, compared to OptSC, the relaxation loses very limited accuracy. By further exploiting the structural properties of the relaxed problem, our third contribution is to propose

[*]This work was done when the first author was a Research Fellow in Nanyang Technological University.

an efficient algorithm that achieves accuracy within nearly a constant factor with respect to the original OptSC. Finally, we conduct extensive experiments on both synthetic and real datasets to validate the advantages of the proposed algorithm over other existing approaches.

Related Work

Many peer grading systems (PGSs), such as peerScholar (Pare and Joordens 2008), Crowdgrader (de Alfaro and Shavlovsky 2014), Mechanical TA (Wright, Thornton, and Leyton-Brown 2015) and Peer Assessment (Shnayder and Parkes 2016) have been developed. Existing PGSs mainly include three phases:

Calibration, where peers' grading reliability and cost are calibrated and learned (Kulkarni et al. 2013). Those researches (Tran-Thanh et al. 2014; Xue et al. 2016; Liu and Chen 2016) of learning peers' reliability and cost are orthogonal to our work of optimizing SC mechanisms, and providing parameters input for peer modeling.

Grading, where students grade peer assignments according to pre-designed incentive mechanisms. Peer-predication and Bayesian truth serum (Prelec 2004) mechanisms, reward a peer if his (belief) report is predicted to be correct based on other peers' reports, have proven to be effective in eliciting truthful grades (Dasgupta and Ghosh 2013; Shnayder et al. 2016). However, they also provide a means for students to collude, such as grading by pre-agreed evaluation rule. SC (Jurca and Faltings 2005; Gao, Wright, and Leyton-Brown 2016) and audit mechanisms (Carbonara et al. 2015), can prevent such a peer collusion issue by allowing the instructor to check some assignments, and rewarding a peer if he is verified to grade diligently. However, existing SC mechanisms mainly focus on theoretic guarantee on truthfulness, do not address the optimization problem of maximizing assignments' evaluation accuracy with the general setting where the instructor has a limited spot-checking budget and peers are heterogeneous on grading reliability and cost.

Aggregation, where students' grades are aggregated to estimate assignments' true grades. Traditional aggregation methods, such as majority voting (Sheng, Provost, and Ipeirotis 2008), expectation-maximization (Whitehill et al. 2009), maximum a posteriori (MAP) (Ok et al. 2016) and belief propagation (Karger, Oh, and Shah 2011), mainly focus on finding aggregation rules to maximize the probability that the aggregated grade correctly predicts the underlying true value. Most of these works assume honest peers who always grade diligently. One exception is a recent paper that studies both elicitation and aggregation (Ho, Frongillo, and Chen 2016), however, a prior information about true value of each assignment is required.

In summary, compared with most related SC mechanism studies, which use simple random policy for homogeneous PGSs, we consider designing the optimal SC mechanism and the aggregation rule to maximize PGSs' reliability in a more practical and general setting where instructor has the budget constraint and peers are heterogeneous and strategic.

Model

In a typical peer grading system (PGS), there are n (≥ 2) peers/students I and n assignments J of these students. The true quality q_j of each assignment j is drawn from a set of possible categories Q. For ease of analysis, we use the binary grade criterion Q=$\{-1, 1\}$, which can be interpreted as categories bad (-1) and good (1). Let G=(I, J, E) denote the bipartite grading graph between peers and assignments (Dasgupta and Ghosh 2013). That is, $(i, j) \in E$ if peer i grades assignment j, and $(i, i) \notin E$ guarantees that each peer does not grade his own assignment. Let $I(j)$ denote peers who grade assignment j, and $J(i)$ denote assignments graded by peer i. $|I(j)|$=$|J(i)|$=l, where l is the load of peers.

Peers. Peers' grades are denoted by Z=$(z_{ij})_{i\in I, j\in J}$, where $z_{ij}\in\{-1, 1\}$ if $(i, j)\in E$ and z_{ij}=0 if $(i, j)\notin E$. Moreover, as required in many practical peer grading systems (de Alfaro and Shavlovsky 2014), peers should also provide detail comments on assignments. The observed grade mainly depends on a peer's reliability, which denotes the probability of grading an assignment correctly. A peer's reliability is an increasing function of the effort level he puts in grading. Let e_{ij} denote peer i's effort level on assignment j. For simplicity, we consider binary effort level, i.e., $e_{ij}\in\{0, 1\}$. Putting in full effort e_{ij}=1 incurs cost $c_{ij}(1) \in [0, 1]$, while putting in zero effort e_{ij}=0 incurs zero cost $c_{ij}(0)$=0. To simplify notations, in the following, $c_{ij}(1)$ is substituted by c_{ij}. A peer who puts in zero effort grades arbitrarily with reliability p_i^0=$\mathbb{P}(z_{ij}$=$q_j|e_{ij}$=$0)$=0.5 ($\mathbb{P}$ means the probability), denoting a random estimate (Witkowski et al. 2013). A peer who grades with full effort or diligently, produces his maximum reliability p_i^1=$\mathbb{P}(z_{ij}$=$q_j|e_{ij}$=$1)$>0.5.

To motivate peers to grade diligently, the SC mechanism is introduced. In SC, the instructor himself can check and grade some assignments. Given a peer-assignment pair $(i, j) \in E$, if peer i is checked with grading assignment j diligently, i will gain a reward $r_{ij}\in[0, 1]$; otherwise, if i is checked with putting zero effort on j, he will not receive any reward. On the other hand, if j is not checked by the instructor, i will not receive any reward (Liu and Chen 2016). Assume that assignment j will be spot-checked with probability $x_j\in[0, 1]$, peer i's expected utility $u_{ij}(e_{ij}, x_j)$ gained by putting in effort $e_{ij}\in\{0, 1\}$ on j is[1]

$$u_{ij}(e_{ij}, x_j) = e_{ij}(x_j r_{ij} - c_{ij}). \quad (1)$$

Considering the external reward and intrinsic grading cost, peer i's best strategy on grading assignment j is

$$e_{ij}^* = \arg\max_{e_{ij}\in\{0,1\}} u_{ij}(e_{ij}, x_j). \quad (2)$$

The Instructor. The instructor estimates unchecked assignments' quality by aggregating peers' grades. We adopt the widely used weighted majority voting (WMV) aggregation method (Sheng, Provost, and Ipeirotis 2008), which can guarantee accuracy performance. For an assignment j, its estimated value $\tilde{q}_j$ aggregated by WMV can be computed by

$$\tilde{q}_j = \begin{cases} 1, & \sum_{i\in I(j)} w_{ij} z_{ij} \geq 0; \\ -1, & \sum_{i\in I(j)} w_{ij} z_{ij} < 0. \end{cases} \quad (3)$$

[1]Our results can be extended to involving peer i's reliability $p_i^{e_{ij}}$ in his utility function, i.e., $u_{ij}(e_{ij}, x_j) = e_{ij}(p_i^{e_{ij}} x_j r_{ij} - c_{ij})$.

where w_{ij}=$2p_{ij}-1$ is the weight of peer i's grade on assignment j and $p_{ij}\in\{p_i^0, p_i^1\}$ is the reliability of peer i on grading assignment j. This design of WMV has two desirable properties: 1) the weight is proportional to peers' reliability and 2) if peer i grades arbitrarily with reliability p_i^0=0.5, his weight becomes zero, indicating that the arbitrary uninformative grade will be discarded in the final aggregation.

Given an assignment j and its peers' reliability profile $\mathbf{p}_j$=$(p_{ij})_{i\in I(j)}$, let $\mathbb{P}_e(j,\mathbf{p}_j)$=$\mathbb{P}(q_j\neq\tilde{q}_j,\mathbf{p}_j)$ denote j's exact error rate (i.e., the probability of returning the incorrect grade) under WMV, which can be computed by

$$\sum_{\mathcal{S}\subseteq I(j)}\left(\prod_{i\in\mathcal{S}}(1-p_{ij})\prod_{i\in I(j)\setminus\mathcal{S}}p_{ij}\right)\mathbb{1}_{\mathcal{X}(\mathcal{S},j)\geq\mathcal{X}(I(j)\setminus\mathcal{S},j)}.$$

The instructor considers all possible peer subsets $\mathcal{S}\subseteq I(j)$ who grade incorrectly and remaining peers $I(j)\setminus\mathcal{S}$ who grade correctly such that the aggregated grade is incorrect. The function $\mathbb{1}_{x\geq y}$ equals 1 if $x>y$, equals 0.5 if $x=y$, and equals 0 if $x<y$. The function $\mathcal{X}(\mathcal{S},j)=\sum_{i\in\mathcal{S}}(2p_{ij}-1)$, denoting the total weight of peers $\mathcal{S}$. We further define j's exact accuracy rate $\mathbb{P}_a(j,\mathbf{p}_j)$=$\mathbb{P}(q_j=\tilde{q}_j,\mathbf{p}_j)$ =$1-\mathbb{P}_e(j,\mathbf{p}_j)$.

Computing the exact error rate $\mathbb{P}_e(j,\mathbf{p}_j)$ requires considering 2^l (l=$|I(j)|$) peer combinations, which is intractable for large-scale PGSs peer where each assignment is graded by dozens of peers (Piech et al. 2013). Moreover, the structure of $\mathbb{P}_e(j,\mathbf{p}_j)$ is complex and hard to analyse. Alternatively, inspired by error rate analysis of crowd labelling (Li, Yu, and Zhou 2013), we apply a simple but meaningful upper bound error rate $\mathbb{P}_e^u(j,\boldsymbol{p}_j)$ of WMV to approximate $\mathbb{P}_e(j,\mathbf{p}_j)$:

$$\mathbb{P}_e^u(j,\boldsymbol{p}_j)=e^{-0.5\sum_{i\in I(j)}(2p_{ij}-1)^2}. \tag{4}$$

Proposition 1. *Given an assignment j and reliability profile $\mathbf{p}_j$=$\{p_{ij}\}_{i\in I(j)}$, we have $\mathbb{P}_e(j,\mathbf{p}_j)\leq\mathbb{P}_e^u(j,\mathbf{p}_j)$*[2].

Next, we show that the upper bound error rate deceases with peer reliability, which is consistent with the exact error rate measure. We first define relationship '$\succ$' between two reliability profiles $\mathbf{p}_j$ and $\mathbf{p}'_j$: $\mathbf{p}_j\succ\mathbf{p}'_j$, iff $\exists i\in I(j)$: $p_{ij}>p'_{ij}$ and $\forall k\in I(j)\setminus i, p_{kj}\geq p'_{kj}$.

Proposition 2. *For an assignment j and two peer reliability profiles $\mathbf{p}_j$ and $\mathbf{p}'_j$, where $\mathbf{p}_j\succ\mathbf{p}'_j$, $\mathbb{P}_e^u(j,\mathbf{p}_j)<\mathbb{P}_e^u(j,\mathbf{p}'_j)$.*

The Instructor's Objective. In practice, the instructor can only check a limited number of assignments, denoted as the spot-checking budget K. Given such budget K, the instructor's objective is to optimize the SC policy $\mathbf{x}$=$(x_j)_{j\in J}$ of determining each assignment j's checking probability x_j, with the aim of maximizing assignments' average evaluation accuracy. We formulate a bilevel optimization program for the OptSC problem as follows:

$$\max_{\mathbf{x}}\ \Phi(\mathbf{x})=\frac{\sum_{j\in J}\left(1-(1-x_j)e^{-0.5\sum_{i\in I(j)}(2p_i^{e_{ij}}-1)^2}\right)}{n}, \tag{5}$$

$$\textbf{s.t.}\ u_{ij}(e_{ij},x_j)\geq u_{ij}(e'_{ij},x_j),\forall i\in I(j),e'_{ij}\in\{0,1\}, \tag{6}$$

$$\sum_{j\in J}x_j\leq K, \tag{7}$$

$$\forall j\in J, x_j\in[0,1]. \tag{8}$$

In the upper level Eq.(5), for each assignment $j\in J$, $1-x_j$ is the probability of not checking j and $e^{-0.5\sum_{i\in I(j)}(2p_i^{e_{ij}}-1)^2}$ is j's upper bound error rate $\mathbb{P}_e^u(j,\mathbf{p}_j)$ if it is not checked, where $p_i^{e_{ij}}\in\{p_i^0,p_i^1\}$ is peer i's reliability on j. Then, the terms $(1-x_j)\mathbb{P}_e^u(j,\mathbf{p}_j)$ and $1-(1-x_j)\mathbb{P}_e^u(j,\mathbf{p}_j)$ are j's upper bound error rate and lower bound accuracy rate under $\mathbf{x}$, respectively. In the lower level Eq.(6), each peer i maximizes his utility u_{ij} by choosing the optimal strategy e_{ij} on grading j. Given an SC policy $\mathbf{x}$, we define assignments' total lower bound accuracy rate, $\Phi_n(\mathbf{x})$=$n\cdot\Phi(\mathbf{x})$.

Given an SC policy $\mathbf{x}$, each peer's best strategy can be uniquely determined for his monotone utility function. Thus, we can substitute $\mathbb{P}_e^u(j,\mathbf{p}_j)$ by $\mathbb{P}_e^u(j,\mathbf{x})$. In the following, for convenience, we substitute *upper bound error rate* and *lower bound accuracy rate* by error rate and accuracy rate.

Analysis and Algorithm

Section 4.1 shows the NP-hardness of OptSC. In Section 4.2, we propose an efficient algorithm to approximate OptSC and analyze its approximation ratio in Section 4.3.

Problem Complexity

We show that OptSC is NP-hard by reducing an arbitrary 0-1 *knapsack* decision problem (KDP) to an OptSC.

Theorem 1. *The OptSC is NP-hard.*

Proof. Given a set of items $\mathcal{I}$=$\{1,\ldots,n\}$, each with a cost $c_i\in\mathbb{Z}^+$ and a value $v_i\in\mathbb{Z}^+$, and the knapsack's capacity $C\in\mathbb{Z}^+$. Here, without loss of generality, we assume that $\max_{i\in\mathcal{I}}c_i<C$. A KDP asks that given $K\in\mathbb{Z}^+$, whether there exists a subset $\mathcal{S}\subseteq\mathcal{I}$ so that $\sum_{i\in\mathcal{S}}c_i\leq C$ and $\sum_{i\in\mathcal{S}}v_i\geq K$. For any KDP=$\langle\mathcal{I},C,K\rangle$, we construct the corresponding OptSC as follows: for each item $i\in I$, we create an assignment $j(i)$ and a peer i who grades $j(i)$. Each peer i's grading reliability is set as $p_i^0=0.5$ and

$$p_i^1=0.5((-2\ln(1-\frac{v_i/v_{max}}{1-c_i/C}))^{0.5}+1) \tag{9}$$

where v_{max}=$\frac{\max_{i\in\mathcal{I}}v_i/(1-c_i/C)}{1-e^{-0.5}}$. Let spot-checking budget be 1, and the cost and reward of peer-assignment pair $(i,j(i))$ be c_i and C, respectively. The construction can be done in polynomial time. We can show that the constructed OptSC has a spot-checking policy with average accuracy rate $\frac{K/v_{max}+1}{n}$ iff KDP=$\langle\mathcal{I},C,K\rangle$ has a solution, and the detailed proof for this "iff" conclusion is shown in the appendix. □

An Efficient Approximation Algorithm

We begin by presenting some notations and rules that are useful for approximation algorithm design.

[2]All omitted proofs are in the online appendix: http://www.ntu.edu.sg/home/boan/papers/AAAI18 Peer Grading Appendix.pdf.

Critical Checking Probability: θ_{ij} and η_j. For an assignment j with checking probability x_j, peer i putting in full effort e_{ij}=1 will gain utility $u_{ij}(1,x_j)$=$x_j r_{ij}-c_{ij}$ and putting in zero effort e_{ij}=0 will gain $u_{ij}(0,x_j)$=0. To elicit i to grade j diligently, the checking probability x_j should satisfy $x_j \geq \theta_{ij}$=c_{ij}/r_{ij} such that $u_{ij}(1,x_j)\geq 0$. Thus, we first define θ_{ij}=c_{ij}/r_{ij}, the critical checking probability of the assignment j with respect to peer i, above which i grades j diligently and under which i grades j arbitrarily. We next define η_j=$\max_{i\in I(j)} \theta_{ij}$, the critical checking probability of j with respect to peers $I(j)$. A diligent peer-assignment pair (i,j) denotes i grades j diligently.

***E*rror *R*ate *F*irst (ERF) Rule.** Given a PGS G=(I,J,E) and an SC policy $\mathbf{x}$, let $\mathbb{P}_e^u(j,\mathbf{x})$ denote the error rate of assignment j under $\mathbf{x}$. Now assume that there is extra tiny budget ϵ that cannot elicit any non-diligent peer-assignment pair to be diligent. It is optimal to maximize PGS's evaluation accuracy by allocating ϵ to the assignment j^* that has the largest error rate, i.e., j^*=$\arg\max_{j\in J} \mathbb{P}_e^u(j,\mathbf{x})$. We verify this rule by analysing the structure of Eq.(5). Let $\mathbf{x}^*$=$(x_{-j^*}, x_{j^*} = x_{j^*}+\epsilon)$ and $\mathbf{x}'$=$(x_{-j'}, x_{j'} = x_{j'}+\epsilon)$ ($x_{j^*}+\epsilon<1$, $x_{j'}+\epsilon<1$) denote policies of allocating ϵ to j^* and $j'(\neq j^*)$, where x_{-j} is the checking probabilities of all assignments except j under $\mathbf{x}$. The total accuracy difference between $\mathbf{x}^*$ and $\mathbf{x}'$ is $\Phi_n(\mathbf{x}^*)-\Phi_n(\mathbf{x}') = \frac{\partial\Phi_n(\mathbf{x})}{\partial x_{j^*}}\epsilon - \frac{\partial\Phi_n(\mathbf{x})}{\partial x_{j'}}\epsilon = \epsilon(\mathbb{P}_e^u(j^*,\mathbf{x})-\mathbb{P}_e^u(j',\mathbf{x}))\geq 0$.

Next, we present the approximation algorithm. The key idea behind our algorithm is that we first formulate a single level relaxation to approximate the bilevel OptSC. Then, we design an efficient approximation algorithm for the relaxed problem, which also offers performance guarantee for the original OptSC.

Relaxing OptSC. Given an SC policy $\mathbf{x}$, let $\mathcal{S}(\mathbf{x})$ be the set of diligent peer-assignment pairs (i,j) where peer i grades assignment j diligently, i.e., $\mathcal{S}(\mathbf{x})$=$\{(i,j)|x_j \geq \theta_{ij},(i,j)\in E\}$. Let $J(i,\mathcal{S}(\mathbf{x}))$=$\{j|\exists(i,j)\in\mathcal{S}(\mathbf{x})\}$ be assignments in $\mathcal{S}(\mathbf{x})$ that are graded diligently by peer i and $I(j,\mathcal{S}(\mathbf{x}))$=$\{i|\exists(i,j)\in\mathcal{S}(\mathbf{x})\}$ be peers in $\mathcal{S}(\mathbf{x})$ who grade assignment j diligently. We formulate a single level *P*eer-*A*ssignment-oriented relaxation OptSC_PA. This relaxation is a combinatorial optimization problem of finding the optimal diligent peer-assignment pair set $\mathcal{S}\subseteq E$ to maximize assignments' accuracy rate $\Phi^\tau(\mathcal{S})$, shown as follows.

$$\max_{\mathcal{S}\subseteq E} \Phi^\tau(\mathcal{S}) = \frac{\sum_{j\in J}\left(1-(1-x_j)e^{-0.5\sum_{i\in I(j)}(2p_{ij}-1)^2}\right)}{n}, \quad (10)$$

$$\textbf{s.t.} \qquad x_j = \max_{i\in I(j,\mathcal{S})}\theta_{ij}, \quad (11)$$

$$p_{ij} = p_i^1, \forall(i,j)\in\mathcal{S}; p_{ij}=p_i^0, \forall(i,j)\notin\mathcal{S}, \quad (12)$$

$$\sum_{j\in J} x_j \leq K. \quad (13)$$

In Eq.(10), $\mathcal{S}\subseteq E$ is the set of selected peer-assignment pairs. To elicit all peer-assignment pairs in $\mathcal{S}$ to be diligent, Eq.(11) proposes a critical checking policy x_j of checking each assignment j with the maximal critical checking probability with respect to peers $I(j,\mathcal{S})$. This critical checking policy guarantees that the original OptSC and the relaxation OptSC_PA achieve nearly the same accuracy rate.

Theorem 2. *Given $K\leq\sum_{j\in J}\eta_j$, let $\mathbf{y}$=$(y_j)_{j\in J}$ and Φ_{opt} be OptSC's optimal SC policy and accuracy rate. Let $\mathcal{S}$ and Φ^τ_{opt} be OptSC_PA's optimal peer-assignment pair set and accuracy rate. We have $\Phi_{opt}-\Phi^\tau_{opt}\leq\frac{1+n^{\mathbb{1}}}{n}$, where n is the number of peers and $n^{\mathbb{1}}$ is the number of assignments that are checked by probability 1 in $\mathbf{y}$.*

Proof. Under $\mathbf{y}$, we split all assignments J into three disjoint groups, $\mathcal{L}$, $\mathcal{H}$ and $\mathcal{F}$, where $\mathcal{L}$=$\{j|y_j=\max_{i\in I(j,\mathcal{S}(\mathbf{y}))}\theta_{ij}\}$ denotes assignments checked by the critical checking probability, $\mathcal{H}$=$\{j|y_j>\max_{i\in I(j,\mathcal{S}(\mathbf{y}))}\theta_{ij}, y_j\neq 1\}$ denotes assignments neither checked by the critical checking probability nor probability 1, and $\mathcal{F}$=$\{j|y_j=1, y_j\notin\mathcal{L}\}$ denotes assignments checked by probability 1, but not in $\mathcal{L}$, where $|\mathcal{F}|$=$n^{\mathbb{1}}$.

Case 1 [$n^{\mathbb{1}}$=0]: Let j^* be the assignment in $\mathcal{H}$ that has the largest error rate under $\mathbf{y}$, i.e., $\mathbb{P}_e^u(j^*,\mathbf{y})$=$\max_{j\in\mathcal{H}}\mathbb{P}_e^u(j,\mathbf{y})$. According to the ERF rule, we can improve $\mathbf{y}$ by transferring some budget from other assignments $j\in\mathcal{H}$ to j^* until the transferred budget can elicit certain peer in $I(j^*)\backslash I(j^*,\mathcal{S}(\mathbf{y}))$ to be diligent on j^*. This budget transfer will not decrease any assignment's error rate. We proceed this budget transfer process until one of the two scenarios happens: 1) all assignments in $\mathcal{H}$ are checked by the critical checking probability and 2) there is only one assignment in $\mathcal{H}$ that is not checked by the critical checking probability. Let $\mathbf{y}'$ be the final policy after this budget transfer, we have $\Phi(\mathbf{y}')\geq\Phi(\mathbf{y})$ in OptSC.

For scenario 1), we have $\Phi(\mathbf{y}')\leq\Phi^\tau(\mathcal{S})$ because $\mathcal{S}$ is the optimal critical checking policy. Thus, $\Phi_{opt}-\Phi^\tau_{opt}=\Phi(\mathbf{y})-\Phi^\tau(\mathcal{S})\leq\Phi(\mathbf{y}')-\Phi^\tau(\mathcal{S}))\leq 0\leq 1/n$.

For scenario 2), let j^* be the assignment that is not checked by the critical checking probability under $\mathbf{y}'$ and z_{j^*}=$y'_{j^*}-\max_{i\in I(j^*,\mathcal{S}(\mathbf{y}'))}\theta_{ij^*}\leq 1$ be the redundant non-critical checking budget on j^*. Removing z_{j^*} from j^* and defining the corresponding critical checking policy $\mathbf{y}''\geq(y'_{-j^*}, y_{j^*}''=y'_{j^*}-z_{j^*})$. Because z_{j^*} is the redundant non-critical checking budget, we have $\Phi(\mathbf{y}'')$=$\Phi(\mathbf{y}')-z_{j^*}/n$ in OptSC. Let $\Phi^\tau(\mathcal{S})$=$\Phi^\tau(\mathcal{S}(K))$ denote the optimal accuracy rate of OptSC_PA with budget K. Then, we have $\Phi^\tau(\mathcal{S}(K))\geq\Phi^\tau(\mathcal{S}(K-z_{j^*}))$. With budget $K-z_{j^*}$, the policy y'' is a critical checking policy and $\mathcal{S}(K-z_{j^*})$ is the optimal critical checking policy. Thus, we have $\Phi^\tau(\mathcal{S}(K-z_{j^*}))\geq\Phi(\mathbf{y}'')\geq\Phi(\mathbf{y}')-z_{j^*}/n\geq\Phi(\mathbf{y})-z_{j^*}/n$. Finally, we have $\Phi_{opt}-\Phi^\tau_{opt}=\Phi(\mathbf{y})-\Phi^\tau(\mathcal{S})\leq\Phi(\mathbf{y})-\Phi^\tau(\mathcal{S}(K-z_{j^*}))\leq z_{j^*}/n\leq 1/n$.

Case 2 [$n^{\mathbb{1}}>0$]: Besides j^*=$\arg\max_{j\in\mathcal{H}}\mathbb{P}_e^u(j,\mathbf{y})$, there are other $n^{\mathbb{1}}$ assignments in $\mathcal{F}$ that have higher accuracy rates in $\mathbf{y}$ for OptSC than those in $\mathcal{S}$ for OptSC_PA. Similar to Case 1), we can further derive that $\Phi_{opt}-\Phi^\tau_{opt}\leq\frac{1+n^{\mathbb{1}}}{n}$. The condition that the assignment $j\in\mathcal{F}$ checked by probability 1 in OptSC is harsh, where the error rate $\mathbb{P}_e^u(j,\mathbf{y})$ of j under full critical checking probability η_j must be larger than all assignments $j'\in\mathcal{L}$ that are checked by the partial critical checking probability $y_{j'}\leq\eta_j$. With budget $K\leq\sum_j\eta_j$, the scenario that an assignment is checked by probability 1 happens infrequently, and compared to n, $n^{\mathbb{1}}$ is very small. □

Properties of OptSC_PA. We observe that in OptSC_PA, the objective function Φ^τ satisfies monotone and submodu-

lar properties with respect to $\mathcal{S}$. Let U be a non-empty finite set and f be a function $f\colon 2^U \to \mathbb{R}$, where 2^U denotes the power set of U. The function f is ***monotone*** if $f(\mathcal{A}) \leq f(\mathcal{B})$ for all $\mathcal{A} \subseteq \mathcal{B} \subseteq U$ and ***submodular*** if $f(\mathcal{A} \cup s) - f(\mathcal{A}) \geq f(\mathcal{B} \cup s) - f(\mathcal{B})$ for all $\mathcal{A} \subseteq \mathcal{B} \subseteq U$ and $s \in U \backslash \mathcal{B}$. We first define a couple of quantities that will be useful in Theorem 3. Let $\mathbb{P}_e^u(j, I(j,\mathcal{S})) = e^{-0.5 \sum_{i \in I(j,\mathcal{S})} (2p_{ij}^1 - 1)^2}$ denote the error rate of assignment j when peers $I(j,\mathcal{S})$ grade diligently on j. For two disjoint sets $I(j,\mathcal{S}_1)$ and $I(j,\mathcal{S}_2)$, where $I(j,\mathcal{S}_1) \cap I(j,\mathcal{S}_2) = \emptyset$, we have $\mathbb{P}_e^u(j, I(j,\mathcal{S}_1) \cup I(j,\mathcal{S}_2)) = \mathbb{P}_e^u(j, I(j,\mathcal{S}_1)) \cdot \mathbb{P}_e^u(j, I(j,\mathcal{S}_2))$.

Theorem 3. *The objective function Φ^τ defined in Eq.(10) is monotone and submodular with respect to $\mathcal{S}$.*

Proof. **Monotone property**: Given a set of diligent peer-assignment pairs $\mathcal{S}$ and another peer-assignment pair $(i^*, j^*) \notin \mathcal{S}$, let $\mathcal{S}^* = \mathcal{S} \cup (i^*, j^*)$ and $\mathbf{x}^{\mathcal{S}} = (x_j^{\mathcal{S}})_{j \in J}$, $\mathbf{x}^{\mathcal{S}^*} = (x_j^{\mathcal{S}^*})_{j \in J}$ denote critical checking polices for $\mathcal{S}$ and $\mathcal{S}^*$, respectively. Then, we have 1) $I(j^*, \mathcal{S}^*) \backslash I(j^*, \mathcal{S}) = i^*$, 2) $\forall j \in J \backslash j^*$, $x_j^{\mathcal{S}} = x_j^{\mathcal{S}^*}$ and $\mathbb{P}_e^u(j, I(j,\mathcal{S})) = \mathbb{P}_e^u(j, I(j,\mathcal{S}^*))$, and 3) $x_{j^*}^{\mathcal{S}^*} \geq x_{j^*}^{\mathcal{S}}$. Finally, the difference between $\Phi_n(\mathcal{S}^*)$ and $\Phi_n(\mathcal{S})$ is $\Phi_n(\mathcal{S}^*) - \Phi_n(\mathcal{S}) = (1 - x_{j^*}^{\mathcal{S}}) \mathbb{P}_e^u(j^*, I(j^*,\mathcal{S})) - (1 - x_{j^*}^{\mathcal{S}^*}) \mathbb{P}_e^u(j^*, I(j^*,\mathcal{S}^*)) = \mathbb{P}_e^u(j^*, I(j^*,\mathcal{S})) \big((1 - x_{j^*}^{S}) - (1 - x_{j^*}^{\mathcal{S}^*}) \mathbb{P}_e^u(j^*, i^*) \big)$. Since $\mathbb{P}_e^u(j^*, I(j^*,\mathcal{S})) \geq 0$, we have $\Phi_n(\mathcal{S}^*) - \Phi_n(\mathcal{S}) \propto 1 - x_{j^*}^{S} - (1 - x_{j^*}^{\mathcal{S}^*}) \mathbb{P}_e^u(j^*, i^*) \geq x_{j^*}^{\mathcal{S}^*} - x_{j^*}^{S} \geq 0$. The operator $\propto$ means the positive relation.

Submodular property: Let $\mathcal{S}_1$ and $\mathcal{S}_2$ denote two diligent peer-assignment pair sets, where $\mathcal{S}_1 \subseteq \mathcal{S}_2$. For any diligent peer-assignment grading $(i^*, j^*) \notin \mathcal{S}_2$, $\mathcal{S}_1^* = \mathcal{S}_1 \cup (i^*, j^*)$ and $\mathcal{S}_2^* = \mathcal{S}_2 \cup (i^*, j^*)$, then, we have

$$
\begin{aligned}
&\Phi_n(\mathcal{S}_1^*) - \Phi_n(\mathcal{S}_1) - (\Phi_n(\mathcal{S}_2^*) - \Phi_n(\mathcal{S}_2)) \\
&= \mathbb{P}_e^u(j^*, I(j^*,\mathcal{S}_1)) \big((1 - x_{j^*}^{S_1}) - (1 - x_{j^*}^{\mathcal{S}_1^*}) \mathbb{P}_e^u(j^*, i^*) \big) \\
&\quad - \mathbb{P}_e^u(j^*, I(j^*,\mathcal{S}_2)) \big((1 - x_{j^*}^{S_2}) - (1 - x_{j^*}^{\mathcal{S}_2^*}) \mathbb{P}_e^u(j^*, i^*) \big) \\
&= \mathbb{P}_e^u(j^*, I(j^*,\mathcal{S}_1)) \big[(1 - x_{j^*}^{S_1}) - (1 - x_{j^*}^{\mathcal{S}_1^*}) \mathbb{P}_e^u(j^*, i^*) \\
&\quad - \mathbb{P}_e^u(j^*, I(j^*,\mathcal{S}_2 \backslash \mathcal{S}_1)) \big((1 - x_{j^*}^{S_2}) - (1 - x_{j^*}^{\mathcal{S}_2^*}) \mathbb{P}_e^u(j^*, i^*) \big) \big] \\
&\propto (1 - x_{j^*}^{S_1}) - (1 - x_{j^*}^{\mathcal{S}_1^*}) \mathbb{P}_e^u(j^*, i^*) \\
&\quad - \mathbb{P}_e^u(j^*, I(j^*,\mathcal{S}_2 \backslash \mathcal{S}_1)) \big((1 - x_{j^*}^{S_2}) - (1 - x_{j^*}^{\mathcal{S}_2^*}) \mathbb{P}_e^u(j^*, i^*) \big) \\
&\geq (1 - x_{j^*}^{S_1}) - (1 - x_{j^*}^{\mathcal{S}_1^*}) \mathbb{P}_e^u(j^*, i^*) \\
&\quad - \big((1 - x_{j^*}^{S_2}) - (1 - x_{j^*}^{\mathcal{S}_2^*}) \mathbb{P}_e^u(j^*, i^*) \big) \\
&= (x_{j^*}^{S_2} - x_{j^*}^{S_1}) - \mathbb{P}_e^u(j^*, i^*) (x_{j^*}^{\mathcal{S}_2^*} - x_{j^*}^{\mathcal{S}_1^*}) \\
&\geq (x_{j^*}^{S_2} - x_{j^*}^{S_1}) - (x_{j^*}^{\mathcal{S}_2^*} - x_{j^*}^{\mathcal{S}_1^*}) = (x_{j^*}^{\mathcal{S}_1^*} - x_{j^*}^{S_1}) - (x_{j^*}^{\mathcal{S}_2^*} - x_{j^*}^{S_2}) \geq 0.
\end{aligned}
$$

□

An Approximation Algorithm for OptSC_PA. Based on monotone and submodular properties of Φ^τ, we propose PASC, a *p*eer-*a*ssignment pair-based *SC* algorithm (i.e., Algorithm 1). Algorithm 1 mainly consists of two stages. **Stage 1–diligent grading elicitation:** In Lines 2-6, Algorithm 1 first finds one candidate SC policy by greedily eliciting the peer-assignment pair (i^*, j^*) that has the largest margin accuracy gain-cost ratio to be diligent, i.e.,

$$(i^*, j^*) = \arg\max_{(i,j) \in E} \frac{\Phi^\tau(\mathbf{x}_{-j}, x_j' = \theta_{ij}) - \Phi^\tau(\mathbf{x})}{x_j' - x_j} \qquad (14)$$

where $x_j' = \theta_{ij}$ is the critical probability of eliciting diligent grading of (i,j) and $x_j' - x_j$ is the budget necessary for this elicitation under the policy $\mathbf{x}$. In Lines 7-8, Algorithm 1 also finds another candidate policy $\mathbf{x}^*$ that only elicits the optimal peer-assignment pair (i^*, j^*) that has the largest accuracy rate gain under policy $\mathbf{x} = (0)_{j \in J}$ to be diligent. In Line 9, Algorithm 1 selects the policy from the two candidates $\mathbf{x}$ and $\mathbf{x}^*$ with larger accuracy rate. **Stage 2–remaining budget allocation:** In lines 10-11, if there is remaining budget $K^r = K - \sum_j x_j > 0$, Algorithm 1 iteratively allocates K^r to assignments that have the largest error rates.

Algorithm 1: Peer-Assignment Pair-Based Spot-Checking Algorithm PASC(G,K)

1 Initialize $\mathbf{x} = (0)_{j \in J}$, $\Omega = E$, $K^r = K$, $K^* = K$;
2 **while** $\Omega \neq \emptyset$ **do**
3 $\quad (i^*, j^*) = \arg\max_{(i,j) \in \Omega} \frac{\Phi^\tau(\mathbf{x}_{-j}, x_j' = \theta_{ij}) - \Phi^\tau(\mathbf{x})}{x_j' - x_j}$;
4 $\quad$ **if** $\theta_{i^* j^*} - x_{j^*} \leq K^r$ **then**
5 $\quad\quad x_{j^*} = \theta_{i^* j^*}$, $K^r = K^r - (\theta_{i^* j^*} - x_{j^*})$;
6 $\quad\quad \Omega = \Omega \setminus \{(i, j^*) \in \Omega | \theta_{ij^*} \leq x_{j^*}\}$;
7 $(i^*, j^*) = \arg\max_{(i,j) \in E} \Phi(0, \ldots, x_j = \theta_{ij}, \ldots, 0)$;
8 $\mathbf{x}^* = (0, \ldots, x_{j^*} = \theta_{i^* j^*}, \ldots, 0)$, $K^* = K^* - x_{j^*}$;
9 **If** $\Phi(\mathbf{x}^*) > \Phi(\mathbf{x})$, **then** $K^r = K^*$ and $\mathbf{x} = \mathbf{x}^*$;
10 **while** $K^r > 0$ **do**
11 $\quad j^* = \arg\max_{j \in J} \mathbb{P}_e^u(j, \mathbf{x})$, $\delta_b = \max\{1 - x_{j^*}, K^r\}$; $K^r = K^r - \delta_b$, $x_{j^*} = x_{j^*} + \delta_b$;
12 **Return** the SC policy $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

Approximation Ratio Analysis

In this section, we provide the approximation ratio Φ_{pasc}/Φ_{opt} of PASC, where Φ_{pasc} and Φ_{opt} are accuracy rates returned by PASC and optimum (OPT) of OptSC. Let $n^{\mathbb{1}}$ be the number of assignments that are checked by probability 1 in OPT, which has been discussed in Theorem 2.

Theorem 4. *If $K < \sum_j \eta_j$, $\frac{\Phi_{pasc}}{\Phi_{opt}} \geq \frac{K}{2(K + n^{\mathbb{1}} + 1)} (1 - \frac{1}{e})$.*

For $K \geq \sum_j \eta_j$, we first present an useful property of OPT.

Proposition 3. *Given the budget $K = K_1 + K_2$, we have $\boldsymbol{x}_{opt}(K) = \boldsymbol{x}_{opt}(K_1) + \boldsymbol{x}_{opt}(K_1 \oplus K_2)$, where $\boldsymbol{x}_{opt}(K)$ is the optimal SC policy with budget K and $\boldsymbol{x}_{opt}(K_1 \oplus K_2)$ is the optimal SC policy with budget K_2 under existing optimal SC policy $\boldsymbol{x}_{opt}(K_1)$.*

Theorem 5. *If $K \geq \sum_j \eta_j$, $\frac{\Phi_{pasc}}{\Phi_{opt}} \geq \frac{K}{K + n^{\mathbb{1}} + 1}$.*

Proof. We divide this setting into two sub-settings: 1) $K = \sum_j \eta_j$ and 2) $K > \sum_j \eta_j$.

Setting 1): $K=\sum_j \eta_j$. PASC can elicit all peers to be diligent, we have $\Phi^{\tau}_{opt}=\Phi_{pasc}$. According to Theorem 4, we have $\frac{\Phi^{\tau}_{opt}}{\Phi_{opt}}\geq\frac{K}{K+n^{\mathbb{1}}+1}$, which derives that $\frac{\Phi_{pasc}}{\Phi_{opt}}\geq\frac{K}{K+n^{\mathbb{1}}+1}$.

Setting 2): $K>\sum_j \eta_j$. Splitting K into two parts $K_1=\sum_j \eta_j$ and $K_2=K-K_1$. Based on Proposition 3, we have $\Phi_{opt}=\Phi_{opt}(K_1)+\Phi_{opt}(K_1\oplus K_2)$. Given the budget K_1, by Theorem 2, OPT can be divided into two cases:

Case 1) [$n^{\mathbb{1}}$=0]: In the case that each assignment $j\in J$ is checked by the critical checking probability η_j. With budget K_1, we have $\Phi_{pasc}(K_1)=\Phi_{opt}(K_1)$ by eliciting all peer-assignment pairs to be diligent. In the second stage with budget K_2 , we have $\Phi_{pasc}(K_1\oplus K_2)=\Phi_{opt}(K_1\oplus K_2)$ by the ERP rule. Thus, we have $\Phi_{opt}=\Phi_{opt}(K_1)+\Phi_{opt}(K_1\oplus K_2)=\Phi_{pasc}(K_1)+\Phi_{pasc}(K_1\oplus K_2)=\Phi_{pasc}$.

On the other hand, if one assignment j^* that has the largest error rate is not checked by the critical checking probability, i.e., $x_{opt}(j^*)>\eta_{j^*}$, and any other assignment $j\in J\backslash\{j^*\}$ is checked by the critical checking probability $x_j=\max_{i\in I(j,\mathcal{S}(\mathbf{x}_{opt}))}\theta_{ij}$. We further consider three cases according to the scale of the remaining budget K_2 in the second stage: 1) K_2 is tiny such that the assignment j^* cannot be checked by probability 1 in OPT. Then, we have that in OPT, only the assignment j^* has a larger checking probability than that in PASC, i.e., $x_{opt}(j^*)>x_{pasc}(j^*),\forall j\neq j^*, x_{opt}(j)\leq x_{pasc}(j)$. For this case, we have $\Phi_{opt}-\Phi_{pasc}\leq\frac{1}{n}$. 2) K_2 is moderate that can make j^* be checked by probability 1 in OPT, but by less than probability 1 in PASC. Here, we have that $\Phi_{pasc}(K_1\oplus K_2)\geq\Phi_{opt}(K_1\oplus K_2)$ because PASC allocates the whole K_2 to the assignment j^* that has the largest error. Thus, for this case, we have $\Phi_{opt}-\Phi_{pasc}=\Phi_{opt}(K_1)+\Phi_{opt}(K_1\oplus K_2)-(\Phi_{pasc}(K_1)+\Phi_{pasc}(K_1\oplus K_2))\leq 1/n$. 3) K_2 is large that can make j^* be checked by probability 1 in PASC, and OPT is checking another assignment j', where $x_{opt}(j')\geq\eta_{j'}$. For this case, we have that in OPT, only j' has a larger checking probability than that in PASC, i.e., $x_{opt}(x'_j)\geq x_{pasc}(x'_j),\forall j\neq j', x_{opt}(x_j)\leq x_{pasc}(x_j)$, which derives $\Phi_{opt}-\Phi_{pasc}\leq 1/n$. Other cases with larger budget can be reduced to above three cases.

Case 2) [$n^{\mathbb{1}}>0$]: Similar to Case 1) analysis, we can also derive that $\Phi_{opt}-\Phi_{pasc}\leq(1+n^{\mathbb{1}})/n$.

Combining the above conclusion that $\Phi_{opt}-\Phi_{pasc}\leq\frac{1+n^{\mathbb{1}}}{n}$, $\frac{\Phi_{pasc}}{\Phi_{opt}}\geq\frac{K}{K+n^{\mathbb{1}}+1}$ follows readily from Theorem 4. □

Uncertainty about Reliability and Cost

So far, we have addressed the optimal spot-checking problem with complete information where each peer i's diligent reliability p_i^1 and cost c_{ij} are known. We extend it to the incomplete information setting where $p_i^1\in[p_i^{1,min},p_i^{1,max}]$ and $c_{ij}\in[c_{ij}^{min},c_{ij}^{max}]$. Now the instructor's objective is to determine an SC policy, $\mathbf{x}$, that maximizes the accuracy rate $\Phi^*(\mathbf{x})$ over all of the possibilities that each p_i^1 and c_{ij} could be chosen from the defined intervals, formulated as follows:

$$\max_{\mathbf{x}}\ \Phi^*(\mathbf{x}), \tag{15}$$

$$\text{s.t.}\quad \Phi^*(\mathbf{x})=\min_{\mathbf{p}^1,\mathbf{c}}\frac{\sum_j\left(1-(1-x_j)e^{-0.5\sum_{i\in I(j)}(2p_i^{e_{ij}}-1)^2}\right)}{n},$$
$$p_i^{1,min}\leq p_i^1\leq p_i^{1,max}, c_{ij}^{min}\leq c_{ij}\leq c_{ij}^{max},\forall i\in I, j\in J$$
$$(6)-(8).$$

We can convert this incomplete information problem (**IP**) to the equivalent complete information problem (**CP**) defined in Eq.(5) with p_i^1=$p_i^{1,min}$ and c_{ij}=c_{ij}^{max}.

Theorem 6. *Let $\boldsymbol{x}^*$ be the optimal SC policy of **CP**. Then, $\boldsymbol{x}^*$ is also the optimal SC policy of **IP**.*

Experimental Evaluation

We experimentally verify the evaluation accuracy and scalability of the proposed algorithm on synthetic and real datasets. All computations are performed on a 64-bit PC with a dual-core 3.2 GHz CPU and 16 GB memory. All results are averaged over 500 instances.

Experiment on Synthetic Dataset

There are 1000 students and 1000 assignments. For each student i, his diligent reliability follows the Gaussian distribution $N(\mu,\delta^2)$, where μ=0.75 and δ=0.125. We allocate each assignment to l peers randomly. The cost c_{ij} and reward r_{ij} follow the Uniform distributions $U(0,1)$ and $U(c_{ij},1)$.

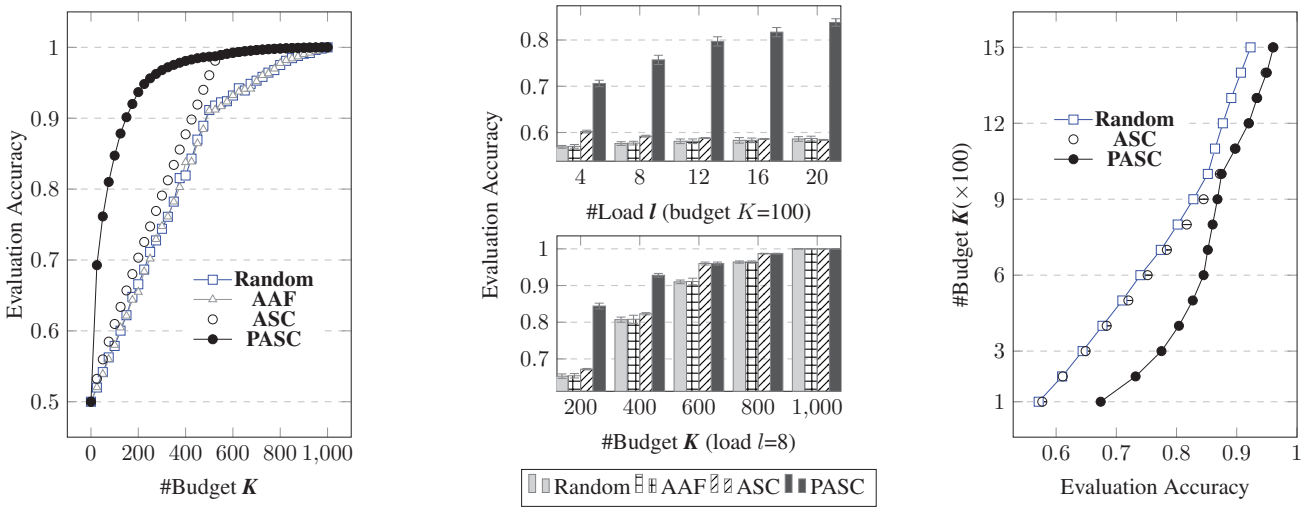
We compare our **PASC** algorithm with three algorithms:

- **Random**, where budget K is allocated to n assignments randomly, i.e., choose one assignment j, and allocate a random probability $x_j\leq 1$ to j.
- Assignment-oriented Spot-Checking algorithm (**ASC**), where budget is greedily allocated to the assignment that has the largest marginal gain-cost ratio.
- Assignment Allocation First (**AAF**), where we first partition these 1000 students into 250 groups, each group has the equal sized 4 members and has approximate group reliability. Each assignment is randomly allocated to $\frac{l}{4}$ groups of these 250 groups. After this assignment allocation, **Random** SC policy is exploited.

Accuracy Rate Evaluation: Table 1 shows the evaluation accuracy rate under various budgets K and loads l, from which we observe that 1) given a load and a budget, PASC has the largest accuracy rate, which is followed by ASC, AAF and Random. 2) Give a budget, accuracy rates of PASC, AAF and Random increase with load, while that of ASC decreases with load. This is because when each assignment j has a large load, more critical checking budget η_j is required. However, such incremental budget cannot proportionally improve accuracy due to the submodular property. 3) Given a load, these algorithms' accuracy rates increase with budget and the increment becomes smaller with the increase of budget. 4) With a pre-step assignment allocation, AAF performs slightly better than Random. This is because in Random, assignments are allocated to peers randomly, considering balancing peers reliability among assignments.

Budget K	100				300				500			
Load l	4	8	12	16	4	8	12	16	4	8	12	16
PASC	**0.770**	**0.832**	**0.874**	**0.900**	**0.915**	**0.956**	**0.976**	**0.986**	**0.956**	**0.982**	**0.993**	**0.997**
ASC	0.625	0.608	0.601	0.597	0.806	0.789	0.779	0.776	0.956	0.953	0.949	0.944
AAF	0.581	0.590	0.592	0.595	0.732	0.754	0.768	0.774	0.885	0.910	0.929	0.939
Random	0.577	0.584	0.587	0.589	0.729	0.752	0.765	0.772	0.881	0.907	0.925	0.934

Table 1: The evaluation accuracy on synthetic dataset. Each cell is statistically significant at 95% confidence level.



(a) Convergence on budget, load l=8. (b) Robustness (c) Real Dataset

Figure 1: (a) Convergence on Budget; (b) Robustness on Budget and Load; (c) Evaluation Accuracy on Real Dataset.

Convergence: Figure 1(a) shows the convergence of evaluation accuracy rate on budget, from which we observe that when the budget K is smaller than the critical budget K_c=$\sum_j \eta_j$, PASC performs much better than ASC, AAF and Random. When $K > K_c$, PASC and ASC produce the same accuracy rate and converge to the optimal accuracy. This is because when $K > K_c$, all peers can be elicited to be diligent in both PASC and ASC.

Robustness: The instructor might have imprecise estimates of p_i^1 and c_{ij}, where $p_i^1 \in [\tilde{p}_i^1 - \delta_p, \tilde{p}_i^1 + \delta_p]$ and $c_{ij} \in [\tilde{c}_{ij} - \delta_c, \tilde{c}_{ij}^1 + \delta_c]$, $\tilde{p}_i^1$ and $\tilde{c}_{ij}$ are observed values. δ_p and δ_c are noise parameters, where δ_p=$\tilde{p}_i^1/10$ and δ_c=$\tilde{c}_{ij}/10$. We compare algorithm's worst-case accuracy rate in uncertain settings defined in Eq.(15). From Figure 1(b), we observe that PASC produces the largest accuracy rate.

Experiment on Real Dataset

Dataset: TREC[3] is a collection of topic-document relevance judgements labelled by workers on AMT. This dataset's data structure is similar to PGS's, where each worker (i.e., peer) is asked to judge whether a topic-document (i.e., assignment) is relevant (i.e., good) or not (i.e., bad). This dataset contains 1,977 judgements collected from 763 workers.

Experiment Setup: We first use l_{tra} training tasks to calibrate worker reliability. For each worker i who judges l_i^{cor} correct labels among l_{tra} tasks, his reliability is estimated as p_i^1=l_i^{cor}/l_{tra}. We model worker's cost and reward in a similar way with that in Section 12. We compute that workers' average reliability 0.89 and variance 0.16. Under the SC mechanism, a worker-task pair (i, j) that is elicited to be diligent, we directly use i's label in the dataset as i's judgement; otherwise, i reports a random judgement on j. Finally, we use the WMV to aggregate the estimated judgement.

[3]https://sites.google.com/site/treccrowd/

Accuracy Rate Evaluation: Figure 1(c) shows the evaluation accuracy in real dataset (in real dataset, the task allocation has been determined, thus AAF is unnecessary), from which we observe that 1) PASC performs the best on improving evaluation accuracy. 2) For PASC and ASC, before the critical budget point ($K \leq$1000), their increment rates drop with budget, while exceeding the critical budget, their increment rates goes up again. In real dataset, peers have high average reliability accuracy ($\sim$0.89). Even with limited budget, these high reliable peers can be elicited to be diligent, leading to a high base accuracy. When budget becomes moderate, such incremental budget can only improve limited accuracy due to the submodular property and the high base accuracy. Finally, when the budget becomes so large that it exceeds the critical budget, all peers will be diligent. The remaining budget will be allocated to the assignment that has the largest error rate, thereby improving the increment rate again.

Conclusion

This paper studies the problem OptSC of optimal spot-checking assignments to maximize assignments evaluation accuracy in general PGSs. The NP-hardness complexity of OptSC is analysed. A combinational optimization problem OptSC_PA is proposed to approximate OptSC. The monotone and submodular properties of OptSC_PA are exploited, and an efficient SC approximation algorithm is proposed. Experimental results show that on both syntectic and real datasets, the proposed algorithm achieves higher evaluation accuracy than other benchmark algorithms.

Acknowledgments

This research is supported by NRF2015 NCR-NCR003-004, the National Natural Science Foundation of China (No. 61472079, and No.61170164), and the Natural Science Foundation of Jiangsu Province of China (No. BK20171363).

References

Agarwal, A.; Mandal, D.; Parkes, D. C.; and Shah, N. 2017. Peer prediction with heterogeneous users. In *EC'17*, 81–98.

Caragiannis, I.; Krimpas, G. A.; and Voudouris, A. A. 2015. Aggregating partial rankings with applications to peer grading in massive online open courses. In *AAMAS'15*, 675–683.

Carbonara, A.; Datta, A.; Sinha, A.; and Zick, Y. 2015. Incentivizing peer grading in moocs: An audit game approach. In *IJCAI'15*, 497–503.

Dasgupta, A., and Ghosh, A. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *WWW'13*, 319–330.

de Alfaro, L., and Shavlovsky, M. 2014. Crowdgrader: A tool for crowdsourcing the evaluation of homework assignments. In *SIGCSE'14*, 415–420.

Gao, X. A.; Wright, J. R.; and Leyton-Brown, K. 2016. Incentivizing evaluation via limited access to ground truth: Peer-prediction makes things worse. In *CoRR:abs/1606.07042*.

Ho, C.-J.; Frongillo, R.; and Chen, Y. 2016. Eliciting categorical data for optimal aggregation. In *NIPS'16*. 2450–2458.

Jurca, R., and Faltings, B. 2005. Enforcing truthful strategies in incentive compatible reputation mechanisms. In *WINE'05*, 268–277.

Karger, D. R.; Oh, S.; and Shah, D. 2011. Iterative learning for reliable crowdsourcing systems. In *NIPS'11*, 1953–1961.

Kong, Y.; Ligett, K.; and Schoenebeck, G. 2016. Putting peer prediction under the micro(economic)scope and making truth-telling focal. In *WINE'16*, 251–264.

Kulkarni, C.; Wei, K. P.; Le, H.; Chia, D.; Papadopoulos, K.; Cheng, J.; Koller, D.; and Klemmer, S. R. 2013. Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact.* 20(6):33:1–33:31.

Li, H.; Yu, B.; and Zhou, D. 2013. Error rate analysis of labeling by crowdsourcing. In *MLMC@ICML'13*.

Liu, Y., and Chen, Y. 2016. Learning to incentivize: Eliciting effort via output agreement. In *IJCAI'16*, 3782–3788.

Miller, N.; Resnick, P.; and Zeckhauser, R. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51(9):1359–1373.

Ok, J.; Oh, S.; Shin, J.; and Yi, Y. 2016. Optimality of belief propagation for crowdsourced classification. In *ICML'16*, 535–544.

Pare, D., and Joordens, S. 2008. Peering into large lectures: examining peer and expert mark agreement using peerscholar, an online peer assessment tool. *Journal of Computer Assisted Learning* 24(6):526–540.

Piech, C.; Huang, J.; Chen, Z.; Do, C.; Ng, A.; and Koller, D. 2013. Tuned models of peer assessment in moocs. In *EDM'13*, 153–160.

Prelec, D. 2004. A bayesian truth serum for subjective data. *Science* 306(5695):462–466.

Radanovic, G., and Faltings, B. 2015. Incentives for subjective evaluations with private beliefs. In *AAAI'15*, 1014–1020.

Raman, K., and Joachims, T. 2014. Methods for ordinal peer grading. In *KDD'14*, 1037–1046.

Sadler, P. M., and Good, E. 2006. The impact of self-and peer-grading on student learning. *Educational assessment* 11(1):1–31.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *KDD'08*, 614–622.

Shnayder, V., and Parkes, D. C. 2016. Practical peer prediction for peer assessment. In *HCOMP'16*, 199–208.

Shnayder, V.; Agarwal, A.; Frongillo, R.; and Parkes, D. C. 2016. Informed truthfulness in multi-task peer prediction. In *EC'16*, 179–196.

Tran-Thanh, L.; Stein, S.; Rogers, A.; and Jennings, N. R. 2014. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence* 214:89 – 111.

Whitehill, J.; Ruvolo, P.; Wu, T.; Bergsma, J.; and Movellan, J. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS'09*, 2035–2043.

Witkowski, J., and Parkes, D. C. 2012. A robust bayesian truth serum for small populations. In *AAAI'12*, 1492–1498.

Witkowski, J.; Bachrach, Y.; Key, P.; and Parkes, D. C. 2013. Dwelling on the negative: Incentivizing effort in peer prediction. In *HCOMP'13*, 190–197.

Wright, J. R.; Thornton, C.; and Leyton-Brown, K. 2015. Mechanical TA: Partially automated high-stakes peer grading. In *SIGCSE'15*, 96–101.

Xue, Y.; Davies, I.; Fink, D.; Wood, C.; and Gomes, C. P. 2016. Avicaching: A two stage game for bias reduction in citizen science. In *AAMAS'16*, 776–785.