The 39th IEEE International Conference on Distributed Computing Systems

Incentivizing the Workers for Truth Discovery in Crowdsourcing with Copiers

Lingyun Jiang, Xiaofu Niu, Jia Xu, Dejun Yang, Lijie Xu

School of Computer, Jiangsu Key Laboratory of Big Data Security & Intelligent Processing, Nanjing University of Posts & Telecommunications

Crowdsourcing Applications

video analysis

Smart Citizen

knowledge discovery

human-robot interaction

Knowledge Repositories Based on Crowdsourcing

Existence of copiers will invalidate most of the existing truth discovery methods since they consider that workers are independent of each other.

tasks \ workers	1	2	3	4	5
Stonebraker	MIT	Berkeley	MIT	**MIT**	**MT**
Dewitt	MSR	MSR	UWise	**UWisc**	**UWisc**
Bernstein	MSR	MSR	MSR	**MSR**	**MSR**
Carey	UCI	UCI	BEA	**BEA**	**BEA**
Halevy	Google	Google	UW	**UW**	**UW**

Incentive Mechanism for Crowdsourcing with Copiers (IMC^2)

P1: Given the conflicting values provided by crowdsourcing workers with copiers, how to estimate the true value?

Dependence and Accuracy based Truth Estimation (DATE)

P2: How to incentivize the strategic workers with high accuracy?

reverse auction

Crowdsourcing process

Challenges

Submitting the same data with others does not imply the copying behavior directly.

It is difficult to detect the copiers

Which one is the copier if any two workers submit the same data?

Need to compute the dependence in both directions

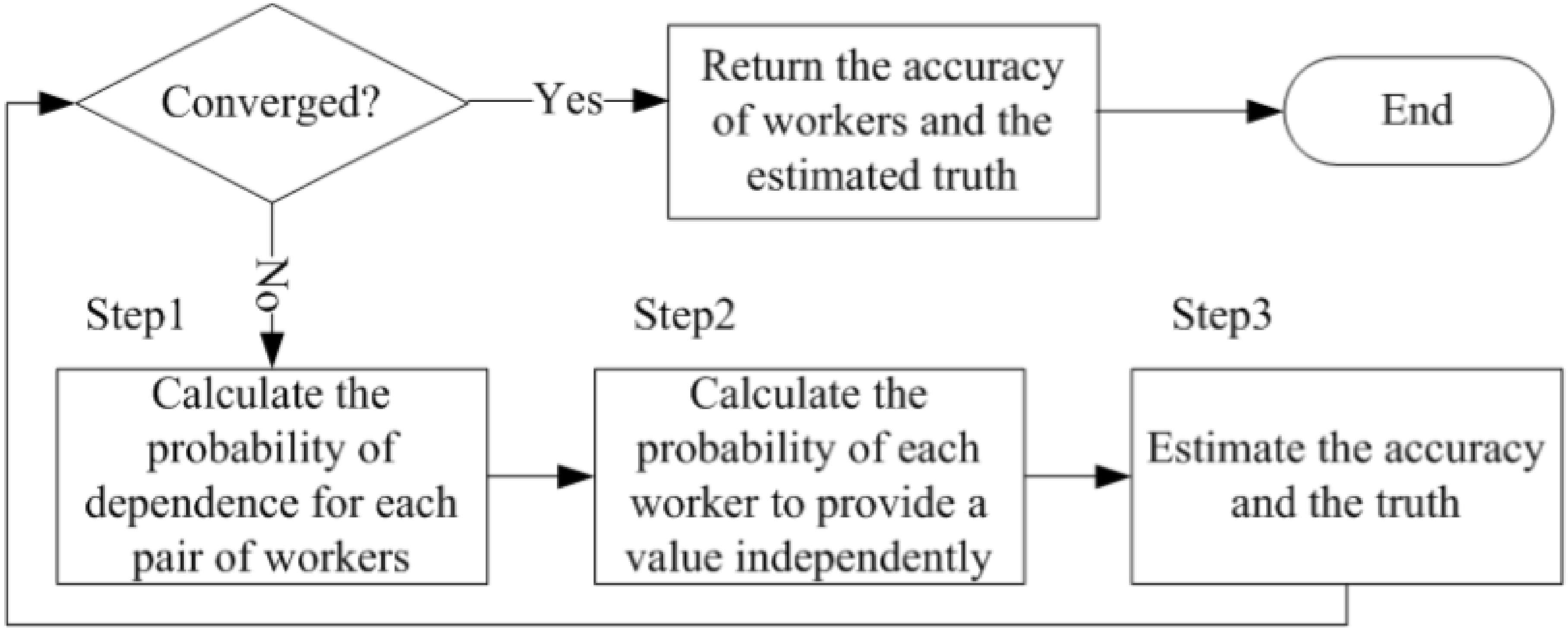
The copiers may contribute to the truth discovery by submitting the combination of the manual data and copied data.

Accuracy calculation method is needed for the copiers

Workers may take strategic behaviors by submitting dishonest bid prices to maximize their utilities.

Truthful auction mechanism

Dependence and Accuracy based Truth Estimation (DATE)



Step1: Calculate the Dependence Between the Workers

If the workers are independent

accuracy of worker i for task j with initial value ε

same true value $P_s^j = P(t_j \in T^s \mid i \perp i') = A_i^j \cdot A_{i'}^j$

same false value $P_f^j = P(t_j \in T^f \mid i \perp i') = num^j \cdot \frac{1 - A_i^j}{num^j} \cdot \frac{1 - A_{i'}^j}{num^j} = \frac{(1 - A_i^j) \cdot (1 - A_{i'}^j)}{num^j}$

different values $P_d^j = P(t_j \in T^d \mid i \perp i') = 1 - P_s^j - P_f^j$

conditional probability $P(\mathbf{D} \mid i \perp i') = \prod_{t_j \in T^s} P_s^j \cdot \prod_{t_j \in T^f} P_f^j \cdot \prod_{t_j \in T^d} P_d^j$

Step1: Calculate the Dependence Between the Workers

If the dependence is considered

$$P(t_j \in T^s \mid i \to i') = A_{i'}^j \cdot r + P_s^j \cdot (1-r)$$

the probability that a value provided by a copier is copied

$$P(t_j \in T^f \mid i \to i') = (1-A_{i'}^j) \cdot r + P_f^j \cdot (1-r),$$

$$P(t_j \in T^d \mid i \to i') = P_d^j \cdot (1-r)$$

$$\begin{aligned} & P(\mathbf{D} \mid i \to i') \\ = & \prod\nolimits_{t_j \in T^s} [A_{i'}^j \cdot r + P_s^j \cdot (1-r)] \\ & \cdot \prod\nolimits_{t_j \in T^f} [(1-A_{i'}^j) \cdot r + P_f^j \cdot (1-r)] \cdot \prod\nolimits_{t_j \in T^d} [P_d^j \cdot (1-r)] \end{aligned}$$

Step1: Calculate the Dependence Between the Workers

The probability of dependence

priori probability that two workers are dependent

$$P(i \to i' \mid \mathbf{D}) = \frac{P(\mathbf{D} \mid i \to i')P(i \to i')}{P(\mathbf{D} \mid i \to i')P(i \to i') + P(\mathbf{D} \mid i \perp i')P(i \perp i')}$$

$$= [1 + (\frac{1\text{-}\alpha}{\alpha}) \cdot \prod_{t_j \in T^s} \frac{P_s^j}{A_{i'}^j \cdot r + P_s^j \cdot (1-r)} \cdot \prod_{t_j \in T^f} \frac{P_f^j}{(1-A_{i'}^j) \cdot r + P_f^j \cdot (1-r)} \cdot (\frac{1}{1-r})^{|T^d|}]^{-1}$$

initial value $P(i \to i') = \alpha, P(i \perp i') = (1-\alpha), 0 < \alpha < 1$

Step2: Calculate the Probability of Providing the Value Independently

However, it is possible that a copier provides some of the values independently, and it will be inappropriate to ignore the contribution of these values.

User level dependence → Value level dependence

It takes exponential time to enumerate all possible dependence for each value between all pairs of workers.

Step2: Calculate the Probability of Providing the Value Independently

A greedy algorithm

for each $t_j \in T$ **do**

  for each $v \in D^j$ **do**

    $i_0 \leftarrow \arg\max_{i:i,i' \in W_v^j, i \neq i'} P(i \to i' \mid \mathbf{D}) + P(i' \to i \mid \mathbf{D})$

    $\overline{W_v^j} \leftarrow \{i_0\}$;

    while $|\overline{W_v^j}| \neq |W_v^j|$ **do**

      $i_0 \leftarrow \arg\max_{i: i \in W_v^j \setminus \overline{W_v^j}, i' \in \overline{W_v^j}, i \neq i} P(i \to i' \mid \mathbf{D})$;

      $I_v^j(i_0) \leftarrow \prod_{i' \in \overline{W_v^j}} (1 - r \cdot P(i_0 \to i' \mid \mathbf{D}))$;

      $\overline{W_v^j} \leftarrow \overline{W_v^j} \cup \{i_0\}$;

    end while

  end for

end for

the probability of any worker i_0 who provides value v of task j independently

Step3: Estimate accuracy and truth

accuracy of value *v* of task *j*

accuracy of worker *i* for task *j*

estimated truth of task *j*

for each $t_j \in T$ **do**

  for each $v \in D^j$ **do**

$$P^j(v) \leftarrow \frac{\Pi_{i \in W_v^j} \frac{num^j \cdot A_i^j}{1 - A_i^j}}{\sum_{v' \in D^j} \Pi_{i \in W_{v'}^j} \frac{num^j \cdot A_i^j}{1 - A_i^j}}$$

  end for

  for each $i \in W, \ s.t. \ t_j \in T_i$ **do**

$$A_i^j \leftarrow \frac{\sum_{v \in D_i^j} P^j(v)}{|D_i^j|};$$

  end for

$$et^j \leftarrow \arg\max_{v \in D^j} \sum_{i \in W_v^j} A_i^j \cdot I_v^j(i);$$

end for

Repeat step1/2/3 until convergence

Reverse Auction

Social Optimization Accuracy Coverage (*SOAC*) problem:

Objective: Minimize $\sum_{i \in S} c_i \cdot x_i$

Subject to: $\sum_{i \in W} A_i^j \cdot x_i \geq \Theta^j, \forall t_j \in T$

$x_i \in \{0,1\}, \forall i \in W$

The SOAC problem is NP-hard!

Reverse Auction

unit cost for accuracy coverage

Step1: Winner selection

Step2: Payment determination

while $\sum_{t_j \in T} \Theta^{j}{}' \neq 0$ **do**

$i \leftarrow \arg\min_{k \in W \setminus S} \frac{b_k}{\sum_{t_j \in T_k} \min\{\Theta^{j}{}', A_k^j\}}$;

$S \leftarrow S \cup \{i\}$;

for each $t_j \in T_i$ **do**

$\Theta^{j}{}' \leftarrow \Theta^{j}{}' - \min\{\Theta^{j}{}', A_i^j\}$;

end for

end while

Reverse Auction Model

Step1: Winner selection

Step2: Payment determination

"critical payment" in Myerson's Theorem

for each $i \in S$ **do**

$W' \leftarrow W \setminus \{i\}, S' \leftarrow \varnothing, \Theta'' \leftarrow \Theta$;

while $\sum_{t_j \in T} \Theta^j{}'' \neq 0$ **do**

$i_k \leftarrow \arg\min_{k \in W' \setminus S'} \frac{b_k}{\sum_{t_j \in T_k} \min\{\Theta^j{}'', A_k^j\}}$;

$S' \leftarrow S' \cup \{i_k\}$;

$p_i \leftarrow \max\{p_i, \frac{\sum_{t_j \in T_i} \min\{\Theta^j{}'', A_i^j\}}{\sum_{t_j \in T_{i_k}} \min\{\Theta^j{}'', A_{i_k}^j\}} b_{i_k}\}$;

for each $t_j \in T_{i_k}$ **do**

$\Theta^j{}'' \leftarrow \Theta^j{}'' - \min\{\Theta^j{}'', A_{i_k}^j\}$;

end for

end while

end for

Theoretical Analysis

Lemma 1. ***IMC^2 is computationally efficient***

Truth Discovery： $O(\varphi n^2 m \, max_{j=1,2,\ldots,m}\{num^j\})$ Reverse Auction： $O(n^3 m)$

Lemma 2. ***IMC^2 is individually rational.***

Each winner will have a nonnegative utility while bidding its true cost.

Lemma 3. ***IMC^2 is truthful***

No worker can improve its utility by submitting a false cost, no matter what others submit.

Lemma 4. ***IMC^2 can approximate the optimal solution within a factor of $2\varepsilon H_\Omega$, where*** $\Omega = \frac{1}{\Delta v}\sum_{t_j \in T}\Theta^j$ **and** $\varepsilon = \max A_i^j \cdot |T_i| \cdot b_i, i \in W, t_j \in T$, $H_\Omega = 1 + \frac{1}{2} + \ldots + \frac{1}{\Omega}$.

Performance Evaluation for Truth discovery

Bench Mark Algorithms

MV (Majority Voting)

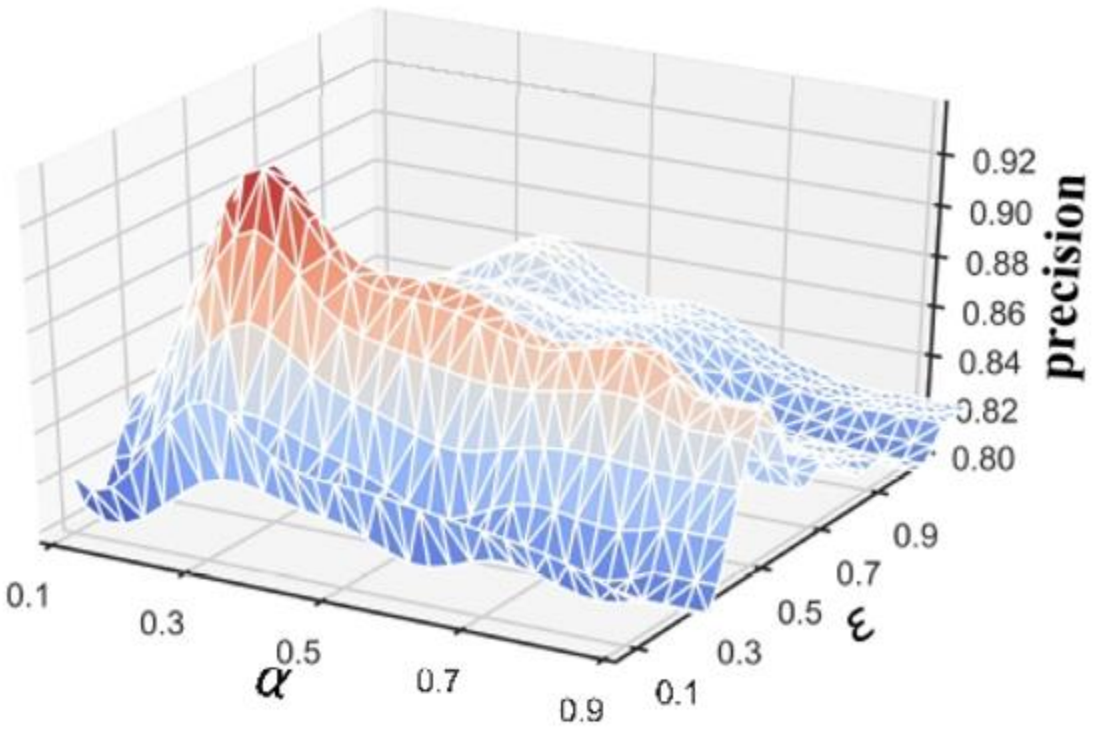
ED (Enumerate all workers' Dependence)

NC (No Copier)： Consider all workers are independent

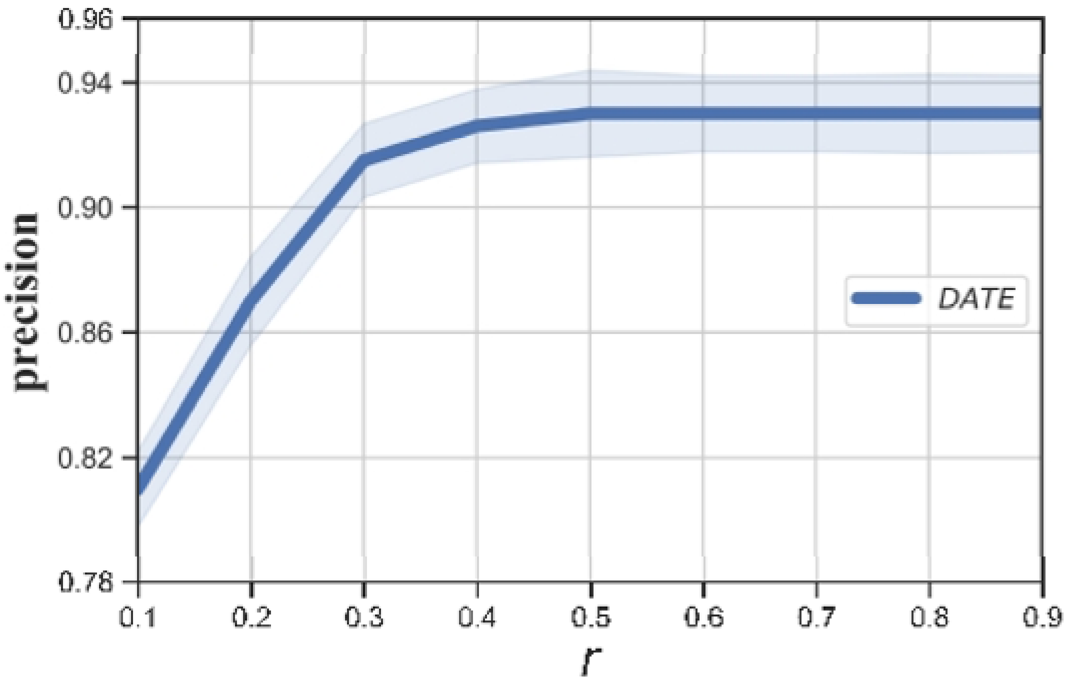
Dataset： Qatar Living Forum

It includes 300 questions, 120 workers and 6000 comments. Each comment can be annotated as *"Good"*, *"Bad"* or *"Other"*.

A. Impact of parameters



(a) Precision versus ε , α



(b) Precision versus r

ε: initial accuracy of any worker for any task

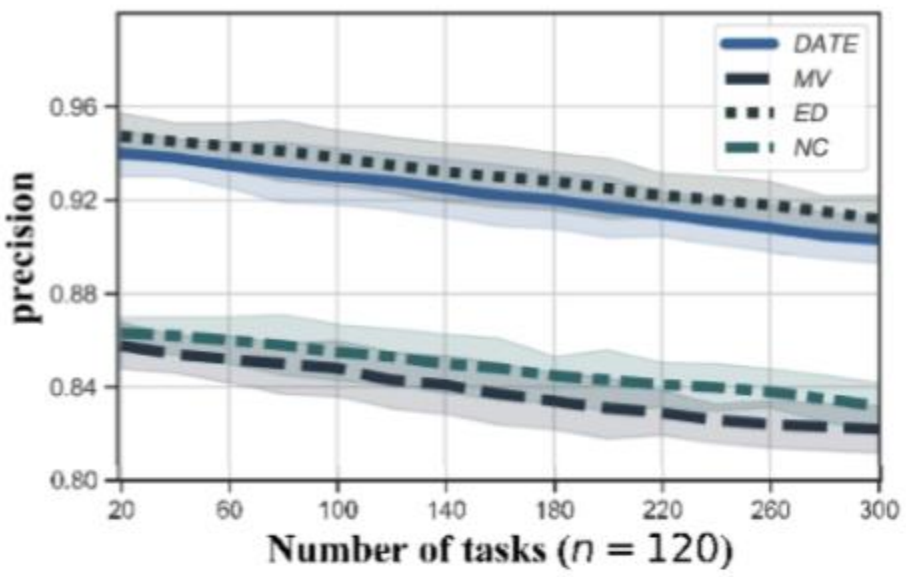
α: initial probability that any two workers are dependent

r: initial probability that a value provided by a copier is copied

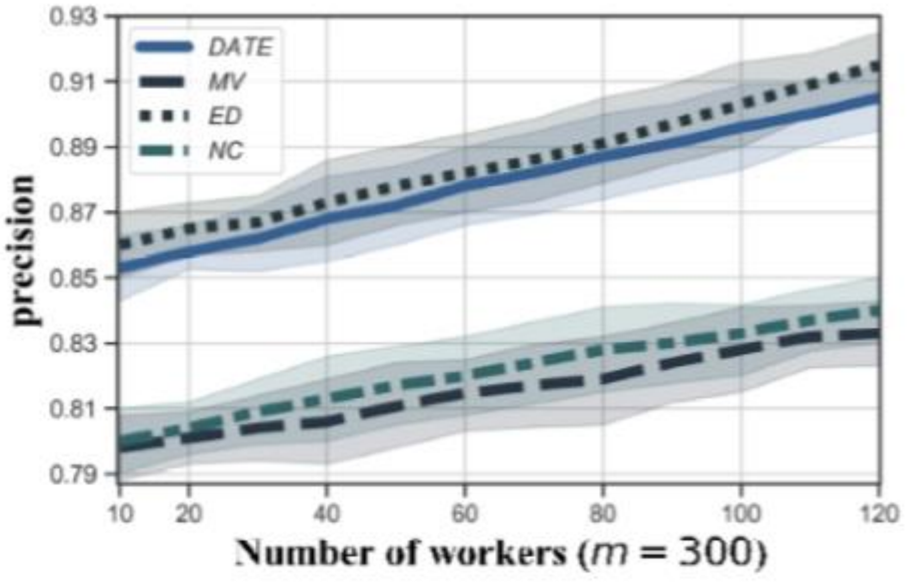
$$\textbf{Precision} = \frac{\sum_{t_j \epsilon T} g(et^j = et^{*j})}{|T|}$$

$\varepsilon = 0.5, \alpha = 0.2, r = 0.4$

B. Precision



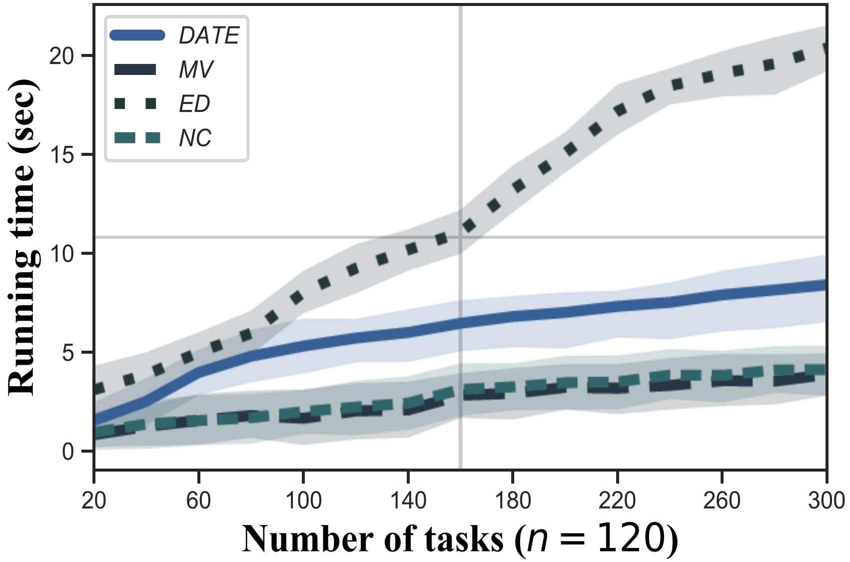
(a) Precision versus tasks

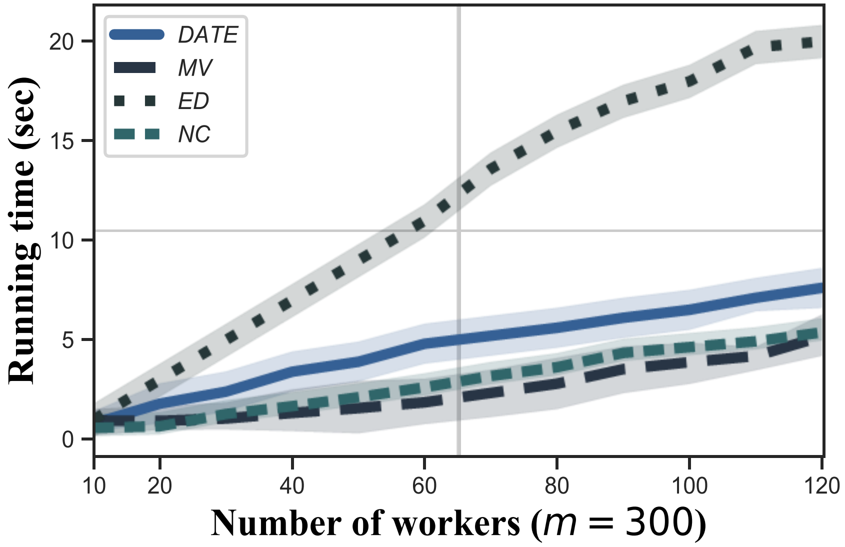


(b) Precision versus workers

DATE can obtain higher precisions (more than 0.85 in all cases) than those of *MV* and *NC* (with average improvement 8.4% and 7.4%, respectively).

B. Running time





For the setting *n*=120, *m*=300, our *DATE* only takes 42.6% of running time comparing with *ED*.

Performance Evaluation for Reverse Auction

Bench Mark Algorithms

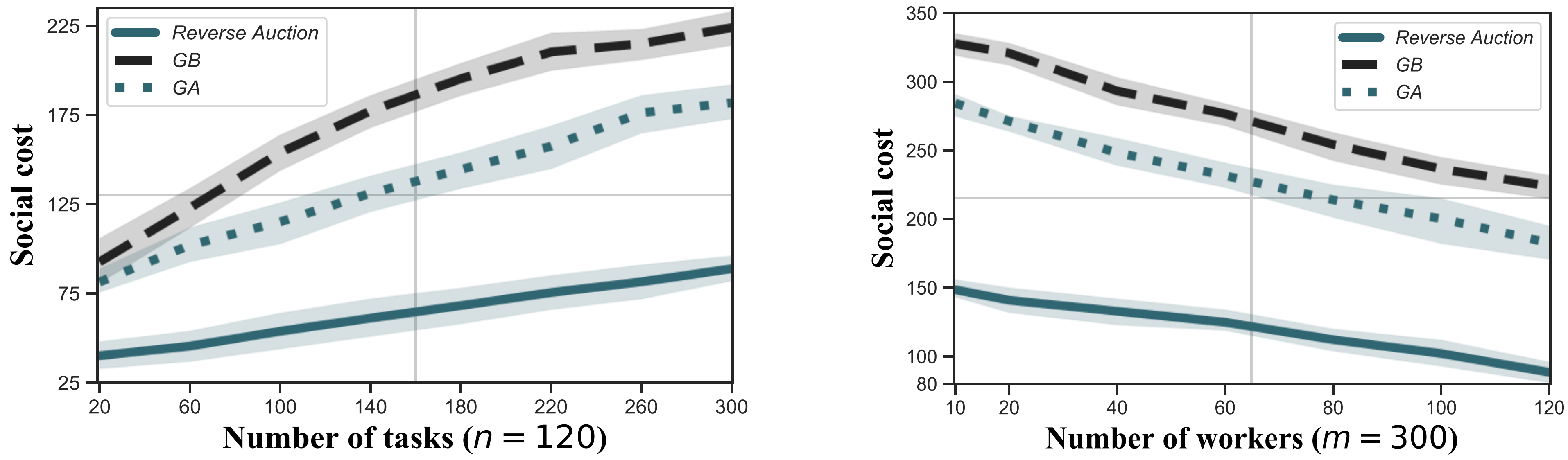
GA (Greedy Accuracy): Each time, *GA* selects the worker with the highest accuracy, and pays the critical value to the winners.

GB (Greedy Bid): Each time, *GB* selects the worker with the lowest bid, and follows the Vickrey Auction payment rule.

Dataset: eBay auction dataset

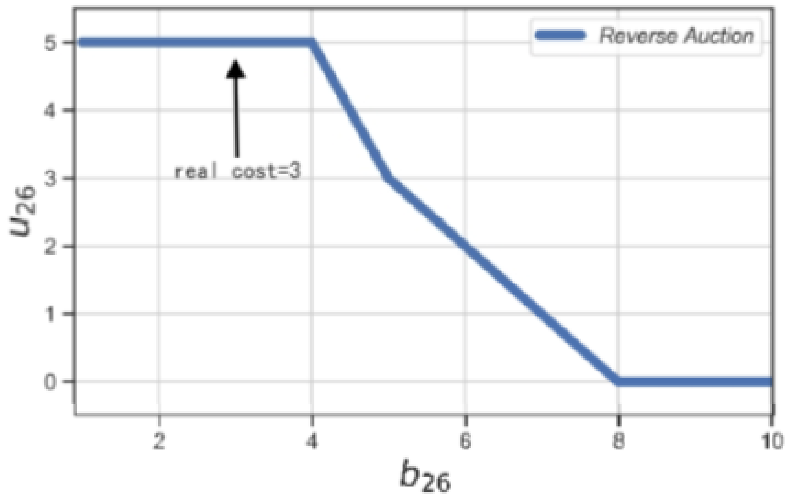
It contains 5017 bid prices for *Palm Pilot M515 PDA* from *eBay* buyers

C. Social cost

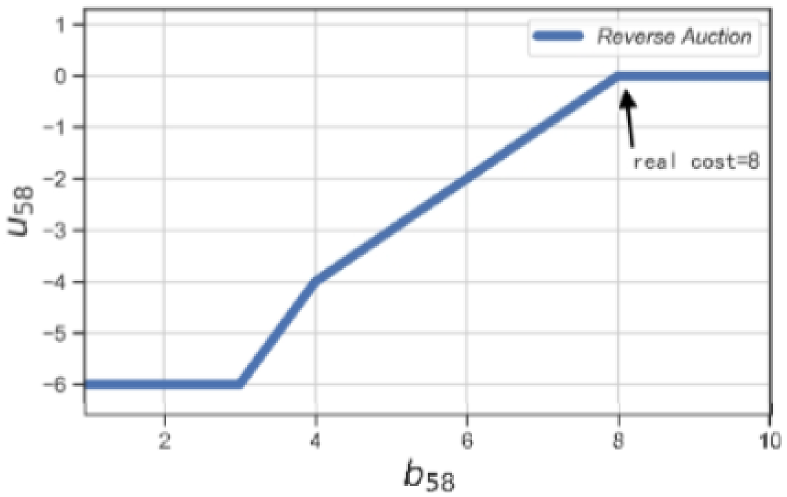


The *Reverse Auction* can obtain the lowest social cost comparing with *GA* and *GB* (with average decrease 40.2% and 59.4%, respectively).

D. Truthfulness



(a) Utility of user with ID=26 (winner)



(b) Utility of user with ID=58 (loser)

The users cannot improve their payoff by submit false cost.

Thank you!

Q & A

xujia@njupt.edu.cn