THE UNIVERSITY OF HONG KONG
DEPARTMENT OF COMPUTER SCIENCE

Title:
On Optimality of Jury Selection Problem in Crowdsourcing

Yudian, Reynold, Silviu, Luyi

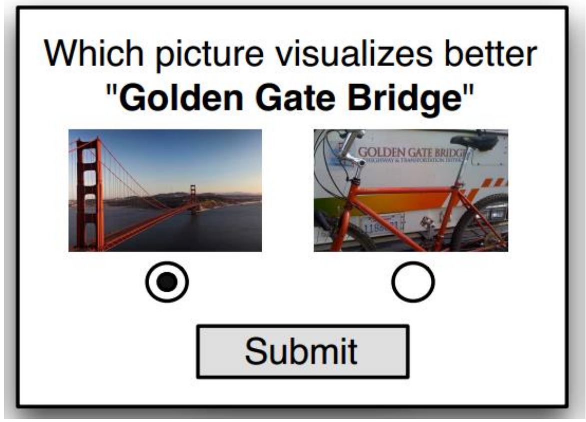
EDBT 2015

Outline

- Introduction (Crowdsourcing)
- Problem Definition (Jury Selection Problem)
- Our Solution (Optimality)
- Conclusion

Why do we need crowd ?

- Problems



- Possible Solutions

M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.

Crowdsourcing Definition

- Definition

  Coordinating a crowd to do micro-tasks that solve problems.

- Example

problems: entity resolution

Object
iPhone 2nd Gen
iPhone Two
iPhone 2
iPad Two
iPad 2
iPad 3rd Gen

An example micro-task :

Are they the same?

iPad 2 = iPad Two

○ YES ○ NO

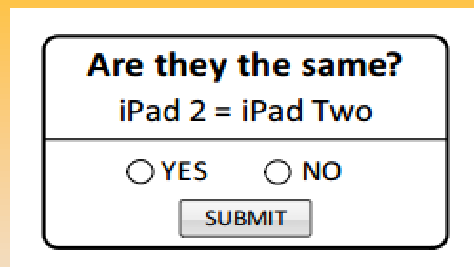
SUBMIT

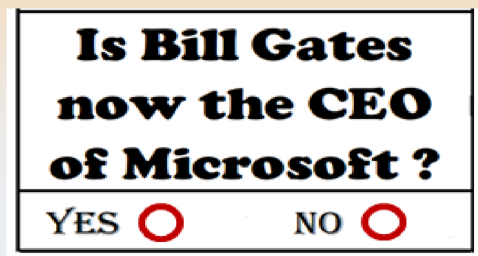
crowd

Amazon Mechanical Turk

- Requesters

- Micro-Tasks





- Workers

- Official Amazon Mechanical Blog (August, 2012)

more than ***500,000 workers*** *from* ***190 countries***

http://mechanicalturk.typepad.com/blog/2012/08/mechanical-turk-featured-on-aws-report.html

Outline

- Introduction (Crowdsourcing)
- Problem Definition (Jury Selection Problem)
- Our Solution (Optimality)
- Conclusion

Problem Intuition (Worker Selection)-VLDB 12

Given (1) a Task

(2) a fixed Budget B

(3) a set of workers

Worker Selection Problem:

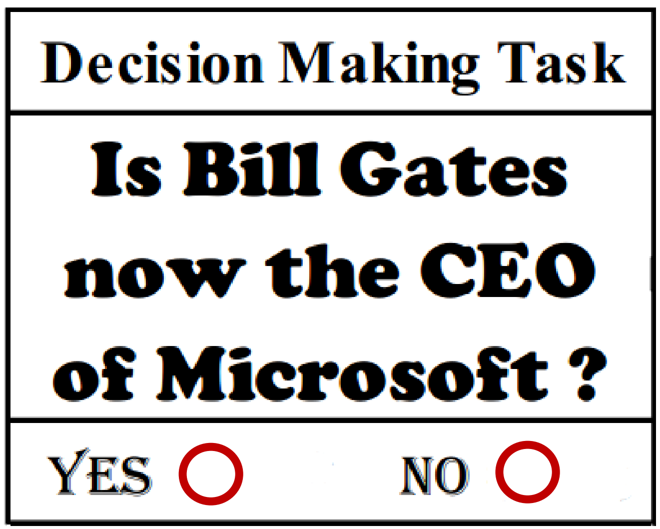
Choose a subset of workers, such that the task can be completed successfully (i.e., with high quality), in the most economical manner ?

- Next: Task and Worker

C. C. Cao, J. She, Y. Tong, and L. Chen. Whom to ask? jury selection for decision making tasks on micro-blog services. PVLDB, 5(11):1495–1506, 2012

Task : Decision Making Task

- Answers are "yes" and "no"
- One (unknown) ground truth

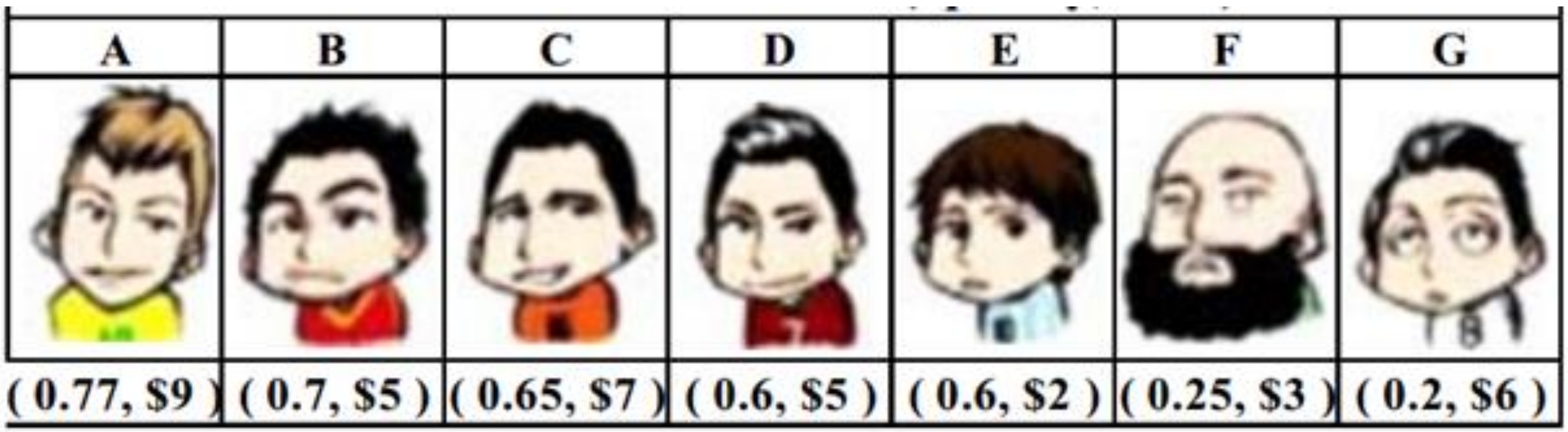


- Simplicity
- (Extensions) Multiple Choice Tasks

Yudian Zheng, Reynold Cheng, Silviu Maniu and Luyi Mo. On optimality of jury selection in crowdsourcing. In *International Conference on Extending Database Technology (EDBT), 2015*

Worker - (quality , cost)

- Each Worker: (quality, cost)   Ex: A (0.77, $9)



A	B	C	D	E	F	G
(0.77, $9)	(0.7, $5)	(0.65, $7)	(0.6, $5)	(0.6, $2)	(0.25, $3)	(0.2, $6)

- Jury: a subset of workers (Ex: {A,B,D})

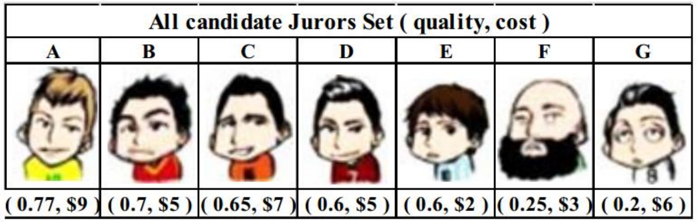
C. C. Cao, J. She, Y. Tong, and L. Chen. Whom to ask? jury selection for decision making tasks on micro-blog services. PVLDB, 5(11):1495–1506, 2012
X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu, and M. Zhang. Cdas: A crowdsourcing data analytics system. PVLDB, 5(10):1040–1051, 2012
P. Venetis and H. Garcia-Molina. Quality control for comparison microtasks. In CrowdKDD, 2012.

Jury Selection Problem

Task:

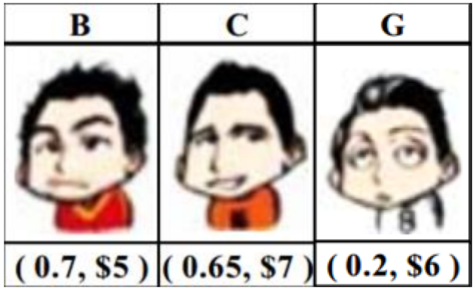
Decision Making Task
Is Bill Gates now the CEO of Microsoft ?
YES ○ NO ○

Budget: $20 Workers:



* Select a Jury (subset of workers) such that the Jury Quality is maximized in all Jury whose cost does not exceed the Budget.

- For each Jury:



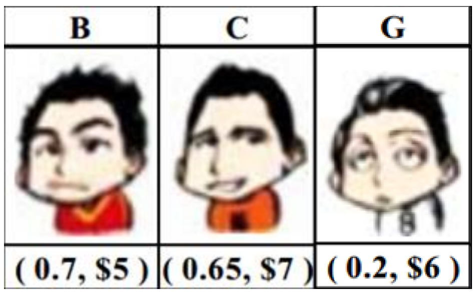
(1) Jury Cost: $5+$7+$6=$18

(2) Jury Quality: JQ ({0.7,0.65,0.2}),

Pr (correctly deriving a result based on workers' answers)

Jury Quality Computation (MV) – VLDB12

- Jury Quality for Majority Voting Strategy



- MV : return the answer which receives the highest votes

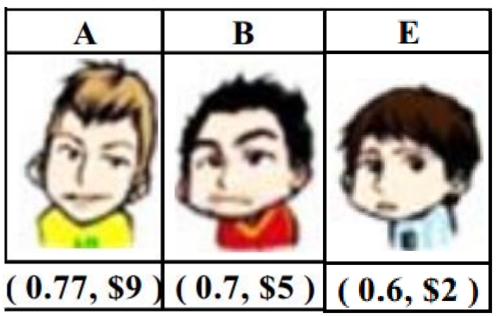
- Cost({$5,$7,$6})=18≤20

- JQ({0.7,0.65,0.2},MV)
  =0.7*0.65*0.8+0.7*0.35*0.2+0.3*0.65*0.2
  +0.7*0.65*0.2=54.3%

Optimal Jury Set- VLDB 12 (Is it optimal ?)

- Enumerating all Jury set satisfying budget constraint

optimal jury set



- Cost({$9,$5,$2})=16⩽20
- JQ({0.77,0.7,0.6},MV)=77.42%

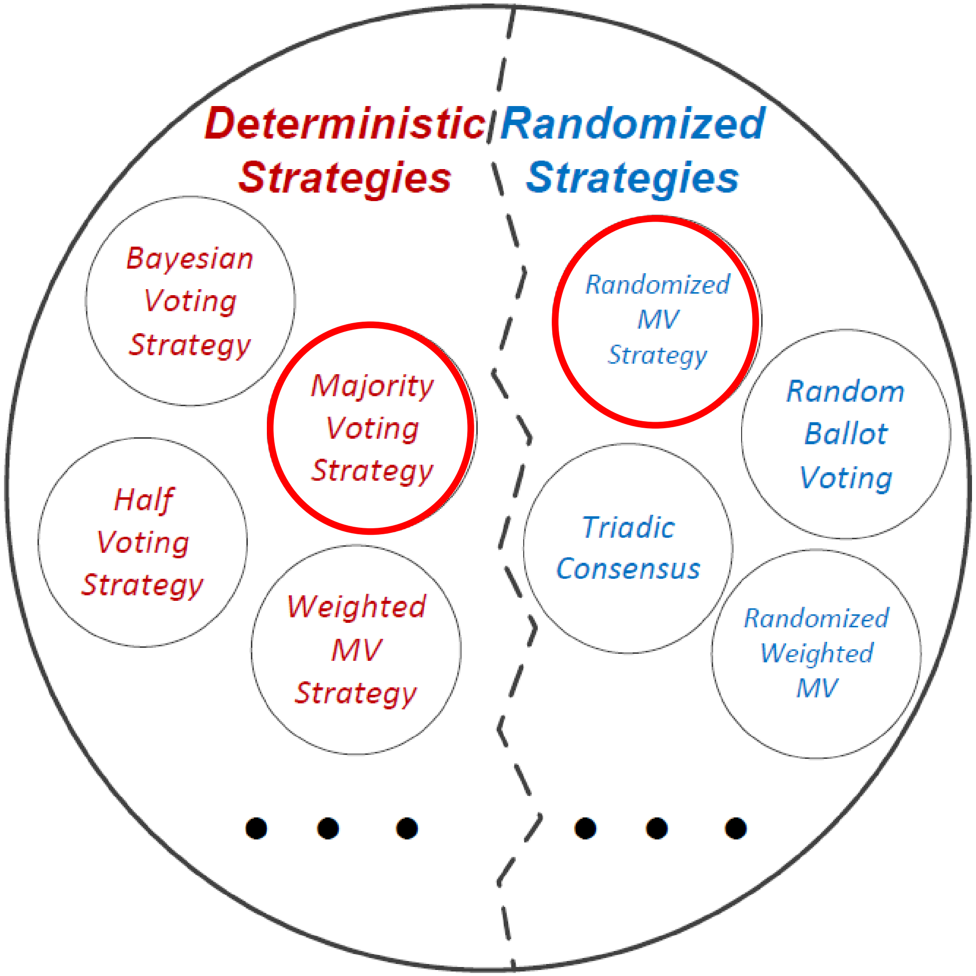
- Question: Is it optimal ?

Is it possible to provide a better solution for JSP, by replacing MV with another strategy?

Outline

- Introduction (Crowdsourcing)
- Problem Definition (Jury Selection Problem)
- Our Solution (Optimality)
- Conclusion

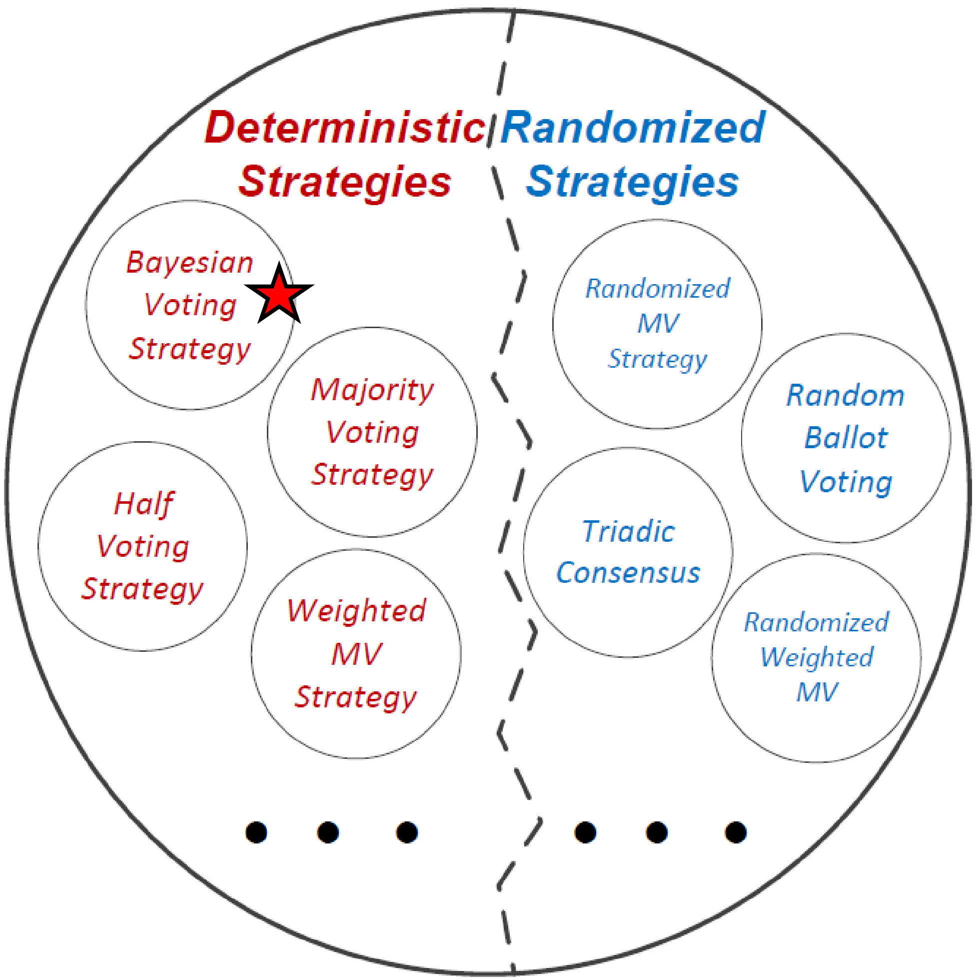
Classification of Voting Strategies



Based on whether the result is returned with degree of randomness, we can classify the voting strategies into two categories:
deterministic voting strategy (left part in the graph)
and
randomized voting strategy (right part in the graph).

Example:
{0,1,1} 0.7,0.6,0.2
Majority Voting (Deterministic):
return 1
Randomized Majority Voting (Randomized):
return 0 with probability 1/3
return 1 with probability 2/3

Existence of Optimal Voting Strategy



Given a Jury set J and a strategy S, the corresponding Jury Quality JQ(J,S) can be computed. An important question is:

Does there exists an optimal strategy S*, such that given a Jury set J, the JQ for this strategy is not lower than the JQ for any strategy (including all deterministic and randomized strategies) ?

$JQ(J, S^*) \geqslant JQ(J, S)$ for any S

We formally prove that the *Bayesian Voting Strategy (BV)* is the optimal strategy, i.e., $S^* = BV$.

*Proof of Optimality

To answer this question, let us reconsider Definition 3. Let $h(V) = \mathbb{E}[\mathbb{1}_{\{S(V)=0\}}]$. We have (i) $h(V) \in [0,1]$; and (ii) $\mathbb{E}[\mathbb{1}_{\{S(V)=1\}}] = 1 - h(V)$. Also, let $P_0(V) = \Pr(\mathbf{V} = V, \mathbf{t} = 0)$, and $P_1(V) = \Pr(\mathbf{V} = V, \mathbf{t} = 1)$. Hence, $JQ(J, S, \alpha)$ can be rewritten as

$$\begin{aligned} & \sum\nolimits_{V \in \Omega} [\, P_0(V) \cdot h(V) + P_1(V) \cdot (1 - h(V)) \,] \\ = & \sum\nolimits_{V \in \Omega} [\, h(V) \cdot (P_0(V) - P_1(V)) + P_1(V) \,] \end{aligned}$$

This gives us a hint to maximize $JQ(J, S, \alpha)$ and find the optimal voting strategy S^*. Let $h^*(V) = \mathbb{E}[\mathbb{1}_{\{S^*(V)=0\}}]$. It is observed that $P_1(V)$ is constant for a given V and $h(V) \in [0,1]$ for all S's (no matter it is a deterministic one or a randomized one). Thus, to optimize $JQ(J, S, \alpha)$, it is required that

1. if $P_0(V) - P_1(V) < 0$, $h^*(V) = 0$, and so, $S^*(V) = 1$;
2. if $P_0(V) - P_1(V) \geq 0$, $h^*(V) = 1$, and so, $S^*(V) = 0$.

Bayesian Voting Strategy

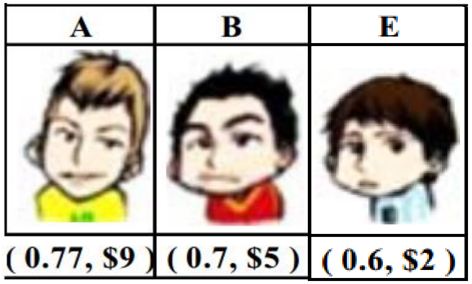
Example:
{0,1,1} 0.7,0.6,0.2

Majority Voting Strategy:
give 1 *vote for the supported answer*

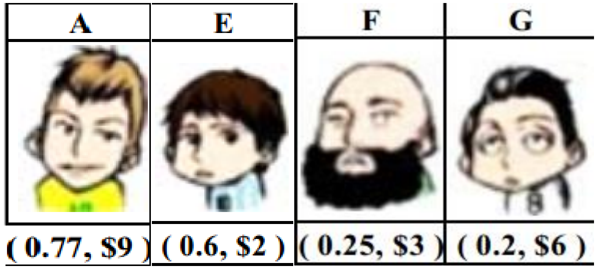
0: 1 (by worker 1)
1: 1 (by worker 2) + 1 (by worker 3) =2

Bayesian Voting Strategy (Deterministic Strategy):
give log[p/(1-p)] *vote for the supported answer*

0: log (0.7/0.3) = 0.8473
1: log (0.6/0.4) + log (0.2/0.8) = -0.981



JQ({0.77,0.7,0.6},MV)
=77.42%
JSP solution: {A, B, E}



JQ({0.77,0.6,0.25,0.2},BV)
=86.95%
JSP solution: {A, E, F, G}

JSP for BV : Complexity (1)(2)

1. Given Jury J, JQ computation for BV , or JQ(J,BV)

Recall that the JQ computation requires enumerating exponential number (w.r.t |J|) of states, i.e.,

$$|\{0,1\}| * |\{0,1\}|^{|J|} = 2^{|J|+1}$$

2. The number of Jury set satisfying Budget Constraint is

Exponential w.r.t. N, in the worst case 2^N

Complexity 1 of JSP

1. Given Jury J, JQ computation for BV , or JQ(J,BV)
Recall that the JQ computation requires enumerating exponential number (w.r.t |J|) of states, i.e.,

$$|\{0,1\}| * |\{0,1\}|^{|J|} = 2^{|J|+1}$$

- NP-hardness of JQ computation
- Polynomial Approximation Algorithm (with Pruning Technique)
- Bounded by 1% Error

*Q1: Computing JQ for BV is NP-hard

Partition Problem (NP-Complete Problem)

Input: $W=\{w_1, w_2, \cdots, w_n\}$, w_i is integer ($1 \leqslant i \leqslant n$)
Output: yes/no

Decide whether W can be partitioned into two disjoint multi-sets W_1 and W_2, such that the sum of elements in W_1 is equal to the sum of elements in W_2.

Reduction

Input: $W=\{w_1, w_2, \cdots, w_n\}$, w_i is integer ($1 \leqslant i \leqslant n$)

Construct $J=\{j_1, j_2, \cdots, j_n\}$ and $J' = \{j_1, j_2, \cdots, j_{n+1}\}$ based on W, then

(1) if $JQ(J', BV) > JQ(J, BV)$, then
the output for partition problem of W is "yes";

(2) if $JQ(J', BV) \leq JQ(J, BV)$, then
the output for partition problem of W is "no";

In order to prove the NP-hardness of computing JQ for BV, we can reduce the partition problem, a well-known NP-Complete Problem (also a decision problem) to the problem of computing JQ for BV.

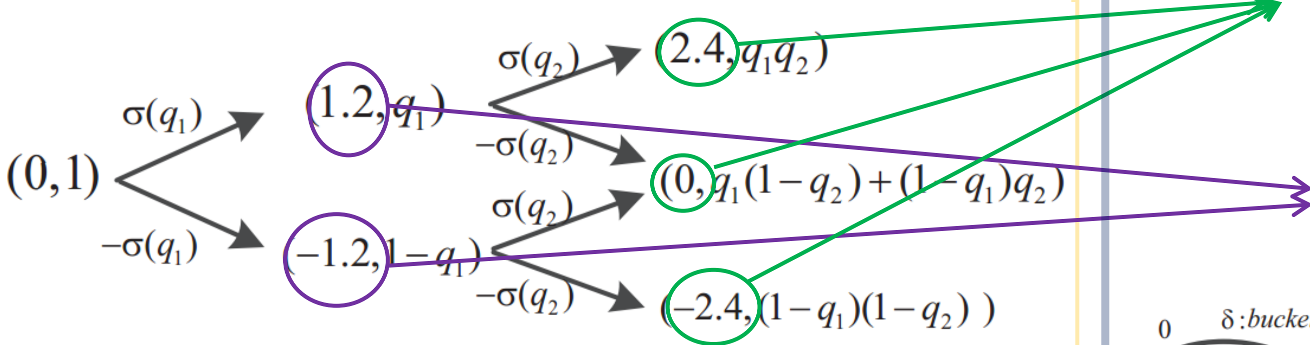
Since computing JQ for BV is not in NP (it is not a decision problem), then it is a NP-hard problem.

*Q1:Bucket-Based Approx. Alg. (Pruning)

Settings:

$$\sigma(q_1) = \sigma(q_2) = 1.2 \quad \sigma(q_i) = \log \frac{q_i}{1-q_i}$$

Compute JQ(J,BV):

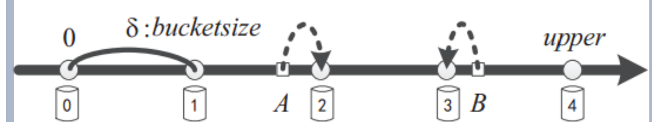


Real Computed JQ(J,BV):

$$q_1 q_2 + [q_1(1-q_2) + (1-q_1)q_2]/2$$

Approximations

$$\log \frac{0.99}{1-0.99} < 4.6$$



$$A = \log \frac{q_1}{1-q_1} \quad B = \log \frac{q_2}{1-q_2}$$

$numBuckets$

*Q1:Approximation Error Bound

Notations:

Let $\widehat{JQ}(J,BV)$ denote the estimated JQ of the approximation algorithm,
and $JQ(J,BV)$ denote the real JQ.

We can prove:

(1) $\widehat{JQ}(J,BV) \le JQ(J,BV)$ and

(2) $JQ(J,BV) - \widehat{JQ}(J,BV) < e^{\frac{5}{4 \cdot d}} - 1$

$d = \frac{numBuckets}{n}$

The time complexity of approximation algorithm is $\mathcal{O}(dn^3)$ and
if $d \ge 200$, the approximation error is bounded within 1%.

$numBuckets \ge 200n$

Real: 80%
Estimated: 79-80%

The polynomial algorithm will give within 1% approximation error bound.

Complexity 2 of JSP

2. The number of Jury set satisfying Budget Constraint is

Exponential w.r.t. N, in the worst case 2^N

- NP-hardness of JSP
- Simulated Annealing Heuristic for general JSP

*Q2- NP-hardness

Combinatorial Optimization Problem

- Similar to Knapsack Problem, with the difference in the Objective Function

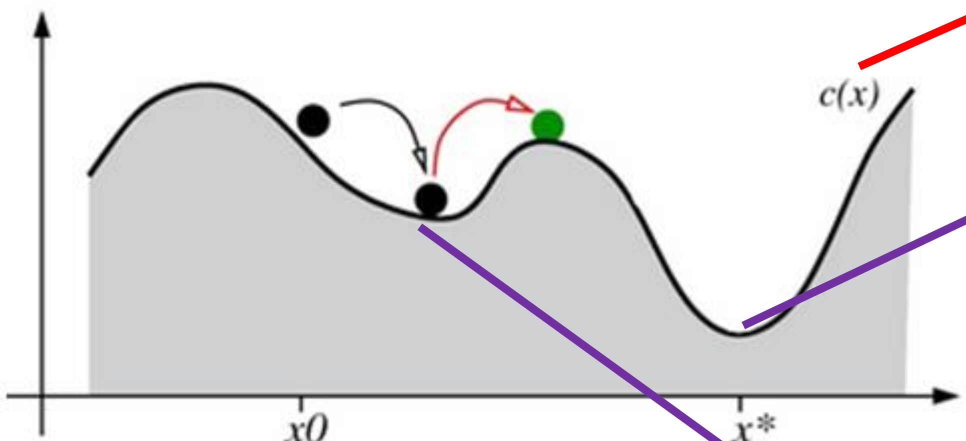
*NP-hard, intuitively as computing the JQ (Objective Function) is NP-hard

*Even though regarding it as an oracle, deriving the optimal solution is also NP-hard
=> N-th order knapsack problem

*Q2- Simulated Annealing Heuristic

Simulated Annealing Heuristic

- Heuristic solving combinatorial optimization problem
- avoid local minimum, probability of accepting a worse place



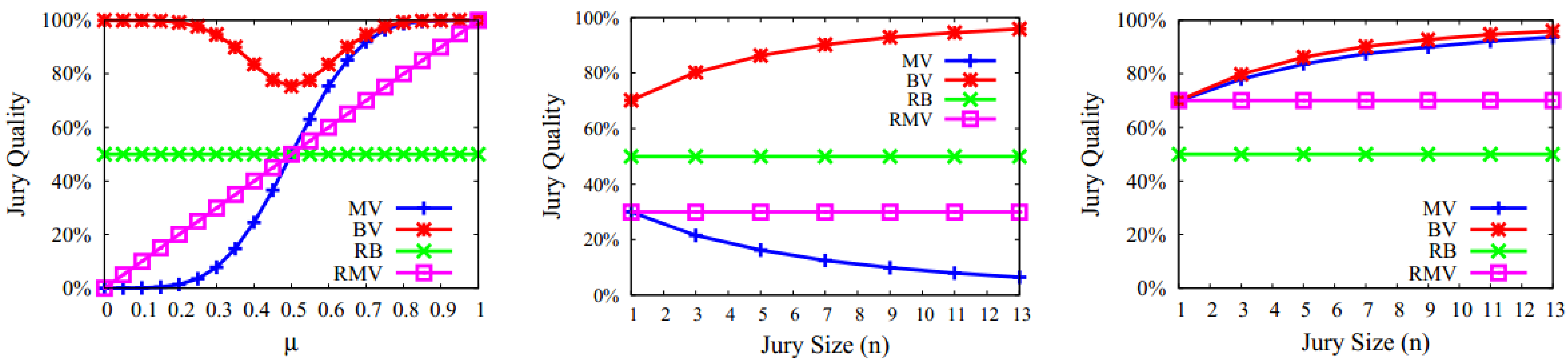
minimize the cost function: **c(x)**

global minimum

local minimum

starting point x0

*Simulated :Different Voting Strategies



(a) Varying μ (b) Varying n ($\mu = 0.3$) (c) Varying n ($\mu = 0.7$)

Randomly generate 10 workers with quality $\mathcal{N}(\mu, 0.1^2)$

MV: Majority Voting

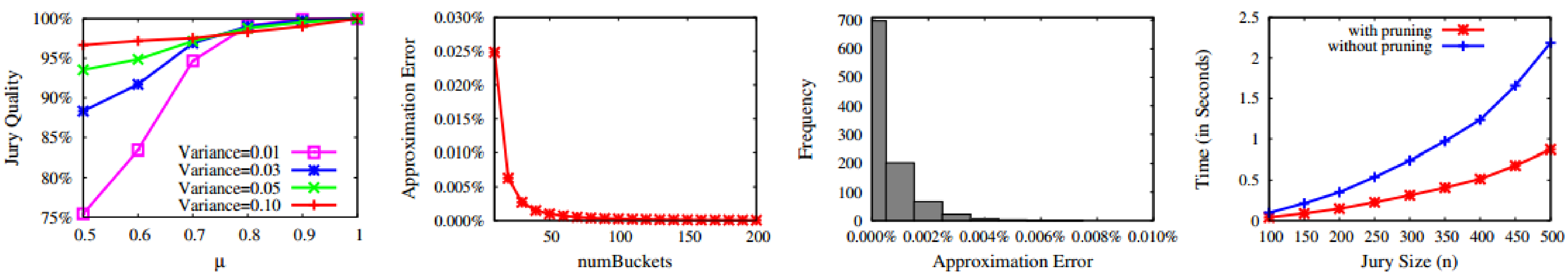
BV: Bayesian Voting

RB: Random Ballot Voting (Randomly returns 0 or 1)

RMV: Randomized Majority Voting

*Simulated : Proposed Approx. Algorithm

Observe the effect of our proposed approximation algorithms



(a) Varying μ and σ^2 (b) Varying $numBuckets$ (c) Approximation Error (d) Varying n

(a) effect with the change of mean and variance

(b) vary the bucket number

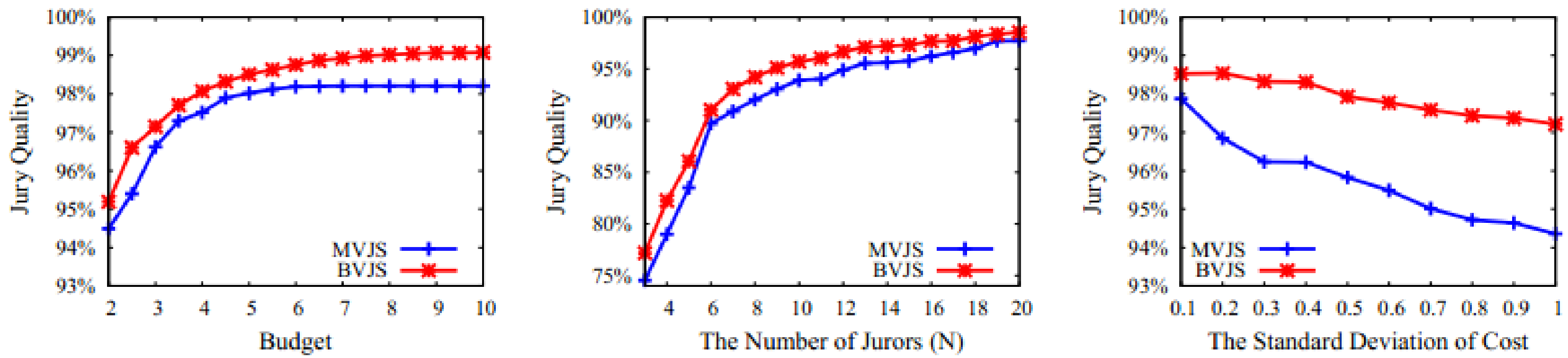
(c) approximation error bound

(d) pruning techniques

Real: End-to-End System Comparison

Collect Data from AMT:
600 questions, each question answered by 20 workers

Known Ground Truth -> workers' qualities



(b) Varying B (c) Varying N (d) Varying $\widehat{\sigma}$

C. C. Cao, J. She, Y. Tong, and L. Chen. Whom to ask? jury selection for decision making tasks on micro-blog services. PVLDB, 5(11):1495–1506, 2012

Outline

- Introduction (Crowdsourcing)
- Problem Definition (Jury Selection Problem)
- Our Solution (Optimality)
- Conclusion

System (Optimal Jury Selection System)

Decision Making Task
Is Bill Gates now the CEO of Microsoft ?
YES O   NO O

All candidate Jurors Set (quality, cost)

A	B	C	D	E	F	G
(0.77, $9)	(0.7, $5)	(0.65, $7)	(0.6, $5)	(0.6, $2)	(0.25, $3)	(0.2, $6)

Budget-Quality Table

Budget-Quality Table			
Budget	Optimal Jury Set	Quality	Required
5	{ E, F }	75%	5
10	{ F, G }	80%	9
15	{ B, F, G }	84.5%	14
20	{ A, E, F, G }	86.95%	20

Trade-Off

Budget 14

B	F	G
(0.7, $5)	(0.25, $3)	(0.2, $6)

Thank you !

Contact Info:
Yudian Zheng
DataBase Group
Computer Science Department
The University of Hong Kong