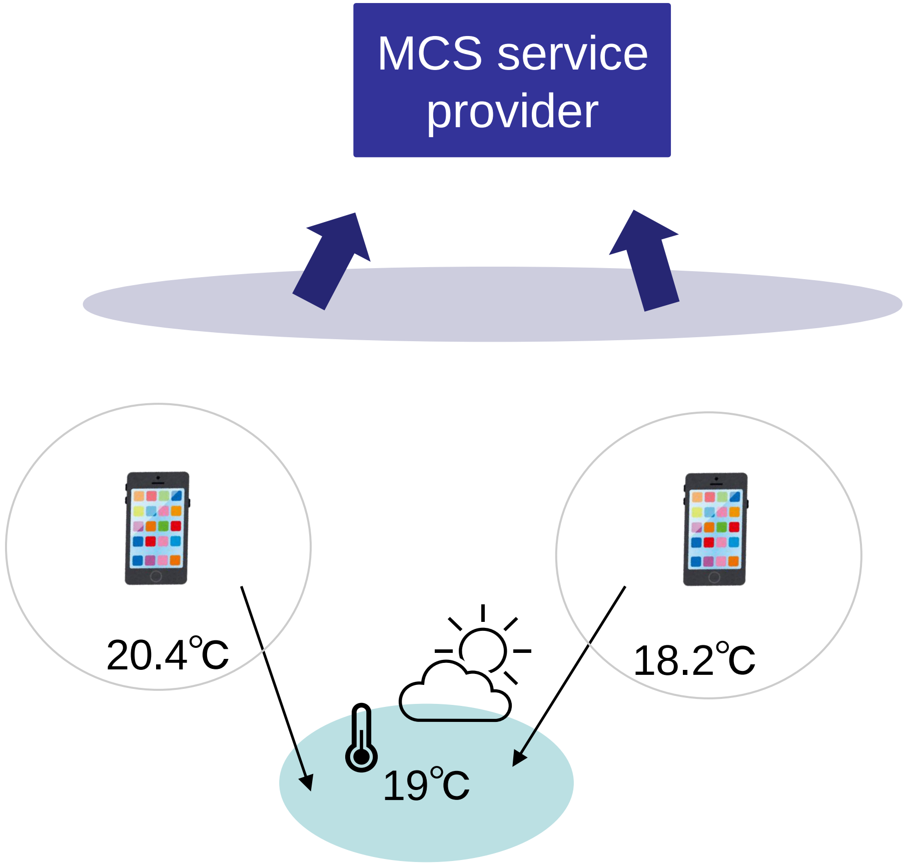
Optimum Worker Sampling in Crowdsensing with Multiple Areas

Chihiro Matsuura　Noriaki Kamiyama

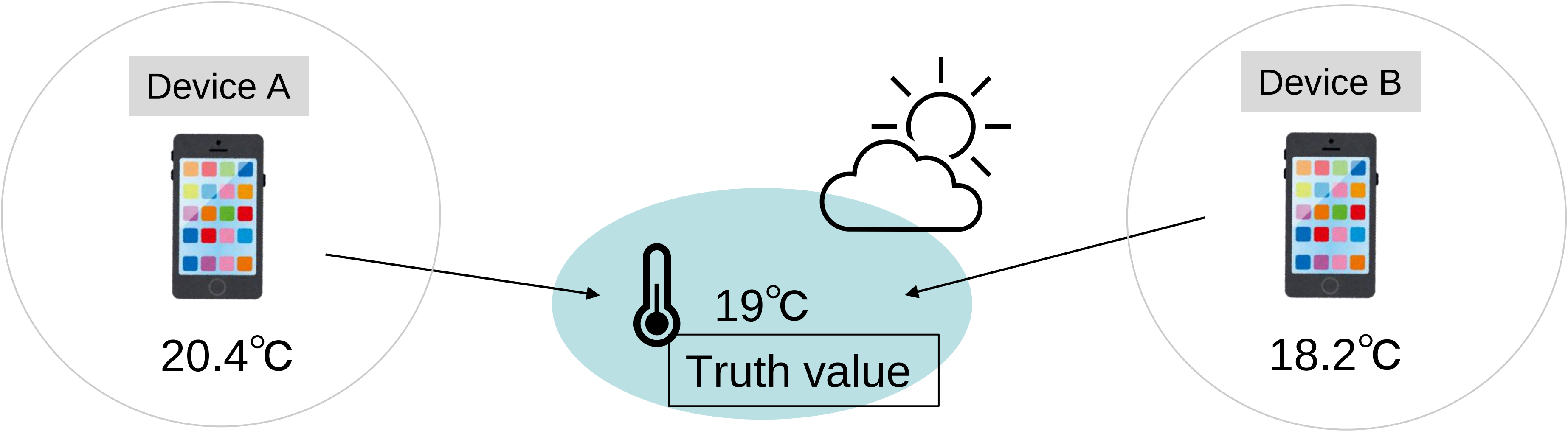
Ritsumeikan University
Graduate School of Information Science and Engineering

Mobile crowdsensing

- Mobile Crowdsensing（MCS）
  - Utilize mobile terminals as IoT devices
- Advantages of MCS
  - Low cost due to no need to build new infrastructure
  - Highly functional sensing as good as conventional IoT devices
  - High penetration rate and huge amount of data can be collected
- Examples of Application
  - Useful for weather forecasting, etc.



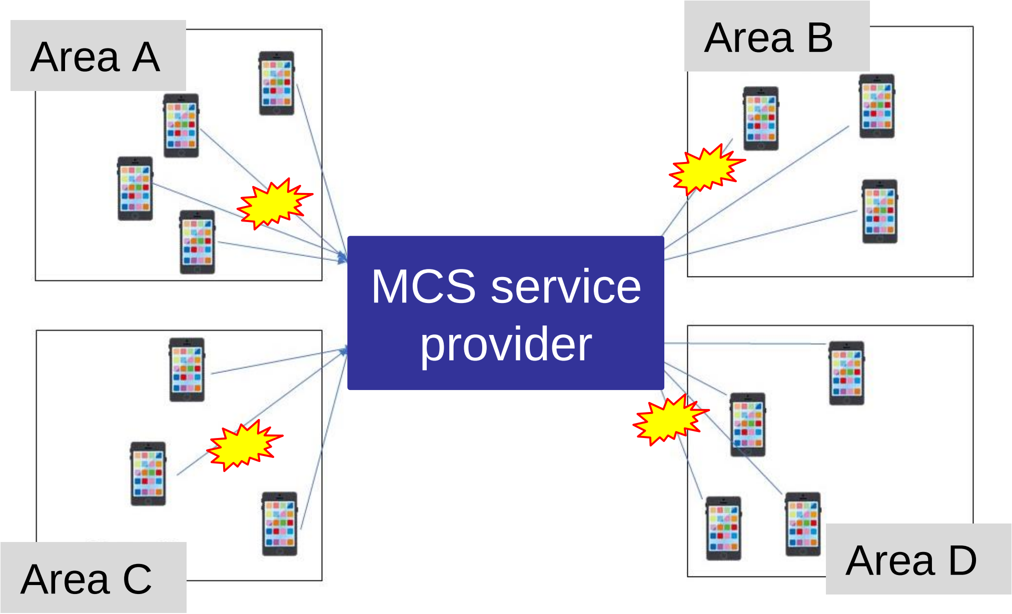
Problems in MCS



- Occurrence of errors in measurement
  - Incorrect data due to sensor failure or human error
  - Transmission of erroneous data by malicious workers
    → Occurrence of data poisoning attacks

Estimating measured values

- CRH (Conflict Resolution on Heterogeneous data) [1]: weighting worker’s reports to minimize estimation error by weighted averaging
- In MCS consisting multiple areas, proposed to optimize number of placed attackers in each area with maximizing estimation error [2]

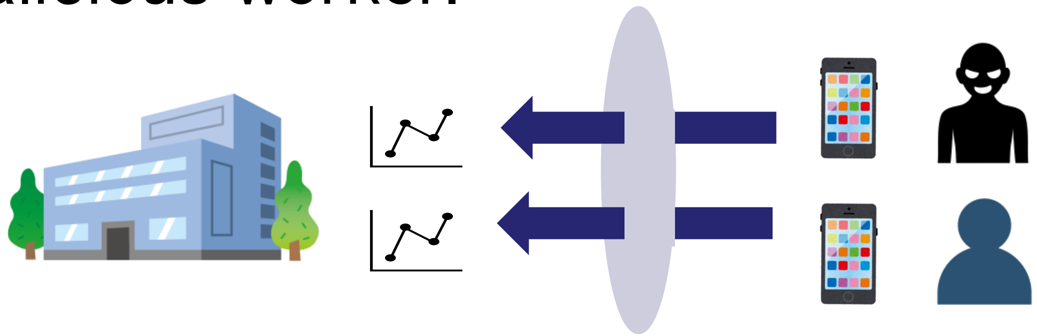


[1] Q. Li, et al., Conflicts to Harmony: A Framework for Resolving Conflicts in Heterogeneous Data by Truth Discovery, IEEE Trans. Know. Data Eng., 28 (8), Aug. 2016

[2] R. Fujimoto and N. Kamiyama, Poisoning Attacks in Crowdsensing Over Multiple Areas, IEEE GLOBECOM 2022

Purpose of this research

- Issues:
    - Existing studies devised sampling methods with only normal workers present.
    - Data poisoning attacks do not send outliers, making it difficult to identify the malicious worker.

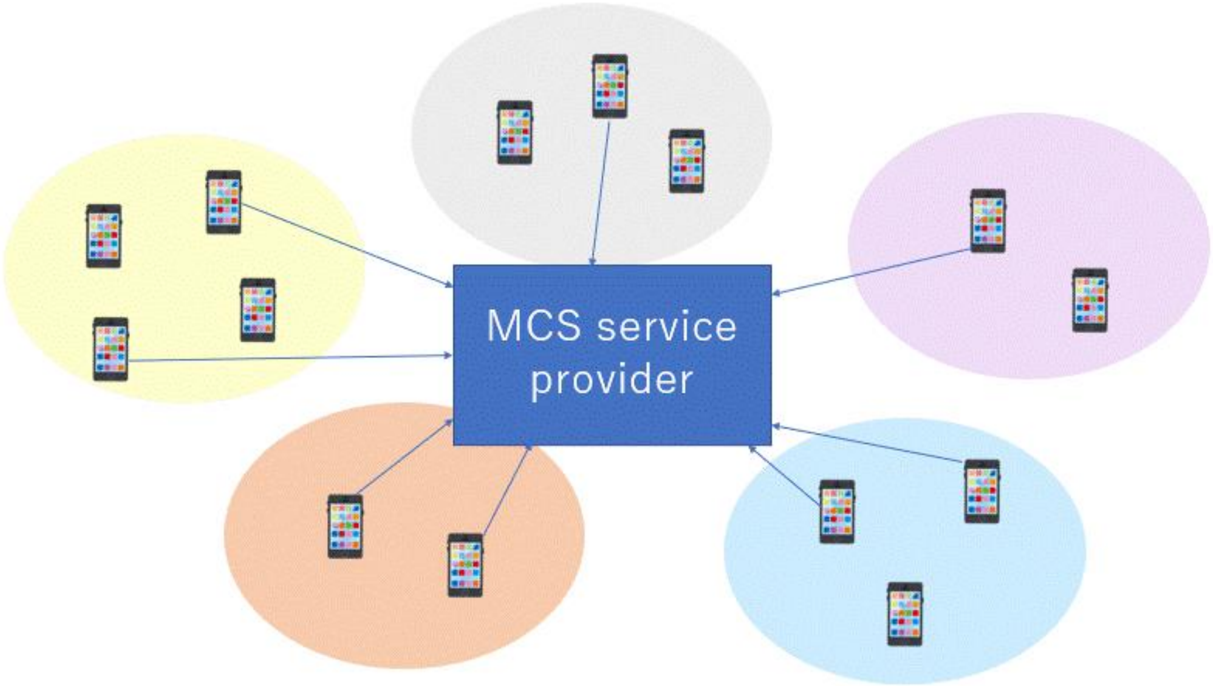


In Multiple area MCS with multiple malicious workers, proposes method of selecting optimal sample count for each area to minimize error in all areas

Multi-area MCS

- Data collection area consists of multiple K areas.
  - Collects sensing data from workers and estimates true value in each area
  - A certain number of malicious workers are included in each area.

- Considers problem of determining number of sample workers in each area when total number of sample workers is given N



CRH (Conflict Resolution on Heterogeneous data)

- Purpose
  - Infer true values from multiple measurements

- Outline
  - Sets low reliability for workers with large differences between true and measured values
  - Uses average of measured values weighted by reliability as estimate value

- Algorithms
  1. Initialize reliability w_{k} of each user k to 1
  2. Update per-user reliability w_{k} by (1)
  3. Using (2), estimate from measured values v_{k} and w_{k} of each worker k
  4. Iterate steps 2 and 3 until convergence of estimates and reliability

$$w_k = -\log \frac{(v_k - \tilde{v})^2}{\sum_{k \in N \cup A} (v_k - \tilde{v})^2} \qquad (1)$$

$$\tilde{v} = \frac{\sum_{k \in N \cup A} v_k w_k}{\sum_{k \in N \cup A} w_k} \qquad (2)$$

N: set of normal workers
A: set of malicious workers

DPA method (Data Poisoning Attack)

- Purpose
  - Update reported values of attack worker to maximize error

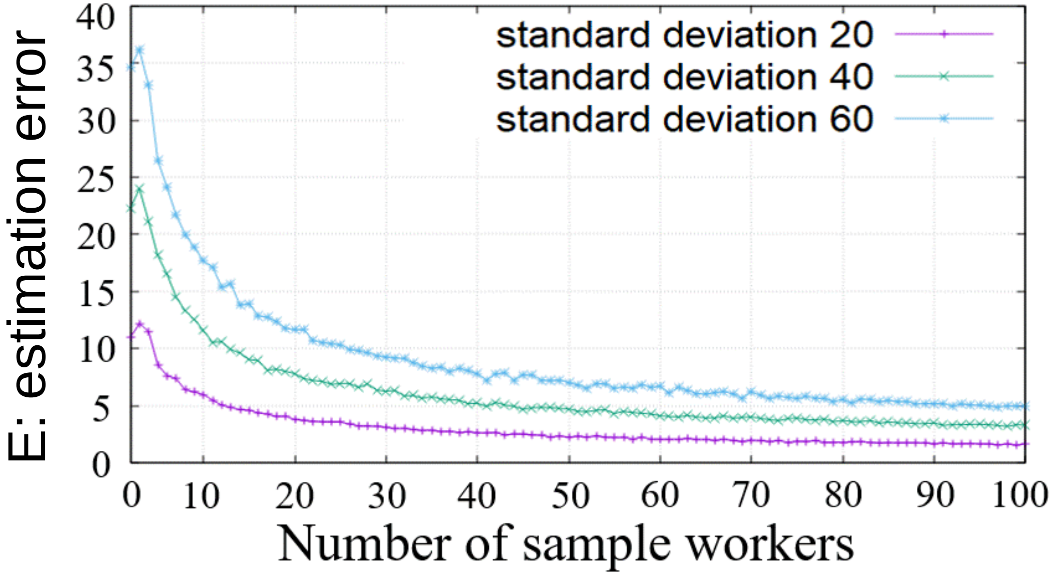
- Outline
  - Applicable only in single area
  - Alternate steps of updating CRH method estimates and reliability of each user, and calculating attacker's reported values

- Algorithms
  1. Initializes reported value v_k for each attack worker k
  2. Applies CRH method with normal workers only
  3. Applies CRH method for all workers including attackers
  4. For each attacker k, updates reported value v_k using (3)
  5. Iterate step 3 and 4 until v_k converges

$$v_k = v_k + 2 \times (\hat{v} - \tilde{v}) \times \frac{w_k}{\sum_{k \in N \cup A} w_k} \tag{3}$$

Accuracy of CRH method in single area

- Analyze effect of normal worker count on estimation error when applying CRH method to single area



- experimental conditions
  - Number of areas： 1
  - Number of normal workers: 2 ≦ n ≦ 100
  - Average of normal worker measurements： 50
  - Standard deviation of normal worker measurements: 20，40, 60

- The smaller the standard deviation of worker measurements, the higher the estimation accuracy.

Proposed method (1/3)

- Purpose
  - Considering upper limit N of total number of sample workers as constraint, optimally design number of sample workers $\mathrm{u_i}$ in each area i to minimize total error E in presence of malicious workers

- Optimization problems
  - Let $\tilde{v}$ be estimate calculated with normal workers only and $\hat{v}$ be the estimate calculated after inclusion of attacking workers, and denote error for each area as $|\tilde{v} - \hat{v}|$

    Objective function is formulated as in (1) (K: Number of areas)

$$\min E(u_1, u_2, \cdots, u_K) = \sum_{i=1}^{K} (\tilde{v}_i - \hat{v}_i)^2 \qquad (1)$$

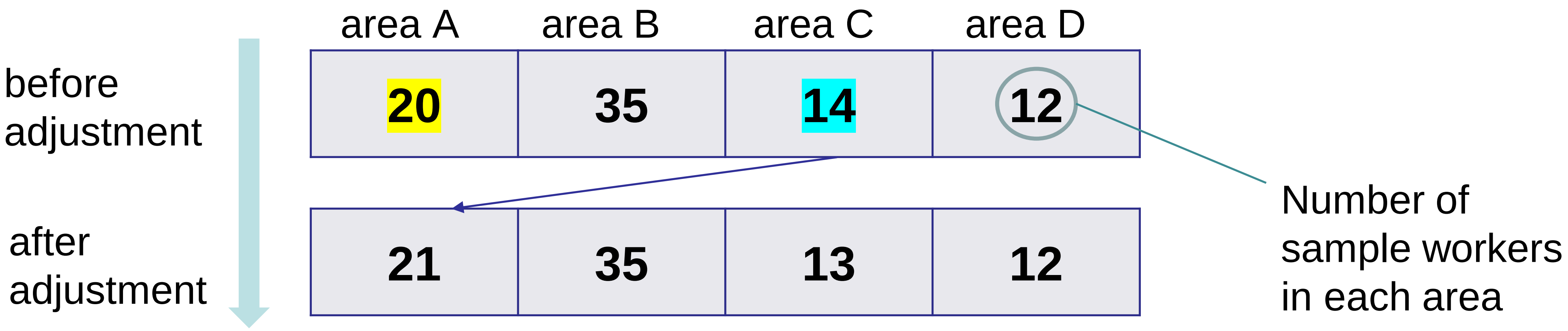
  - Constraints are shown in (2)

$$\sum_{i=1}^{K} u_i = N \qquad (2)$$

Proposed method (2/3)

Outline

1. Initializes number of sample workers in each area to $u_i = N/K$ and calculate total error E_{ini} at this time
2. Calculates average estimation error $e_{i,ui}$ for each sample worker count in each area and stored in table
3. Increments （decrements） number of sample workers in each area i and calculate decrease in estimation error e_{dec} (increase e_{inc})



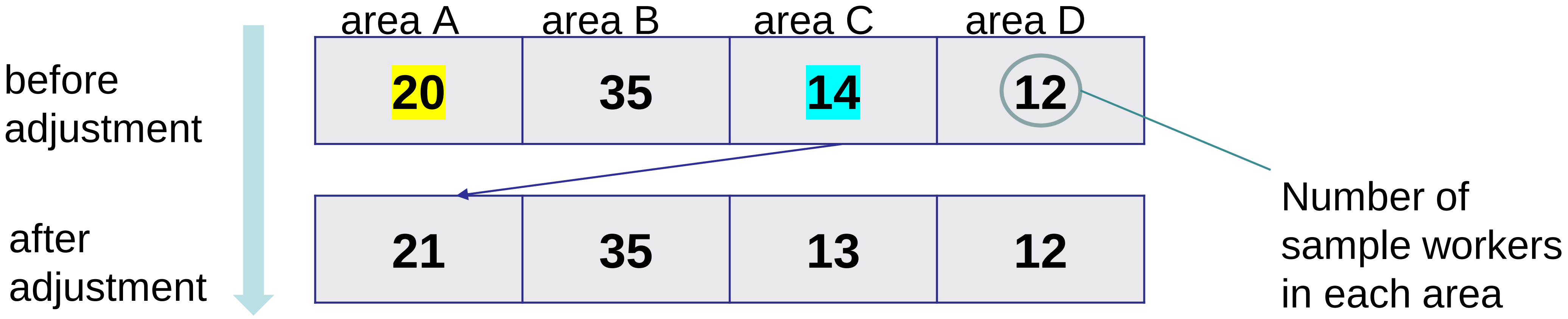
	area A	area B	area C	area D
before adjustment	20	35	14	12
after adjustment	21	35	13	12

Number of sample workers in each area

Proposed method (3/3)

Outline

4. Increments number of sampled workers in area with largest decrease and decrements number of sampled workers in area with smallest increase
5. Iterates until change in total error $|E_{post} - E_{pre}|$ falls below threshold η, and calculates total error E_{conv} at this time



	area A	area B	area C	area D
before adjustment	20	35	14	12
after adjustment	21	35	13	12

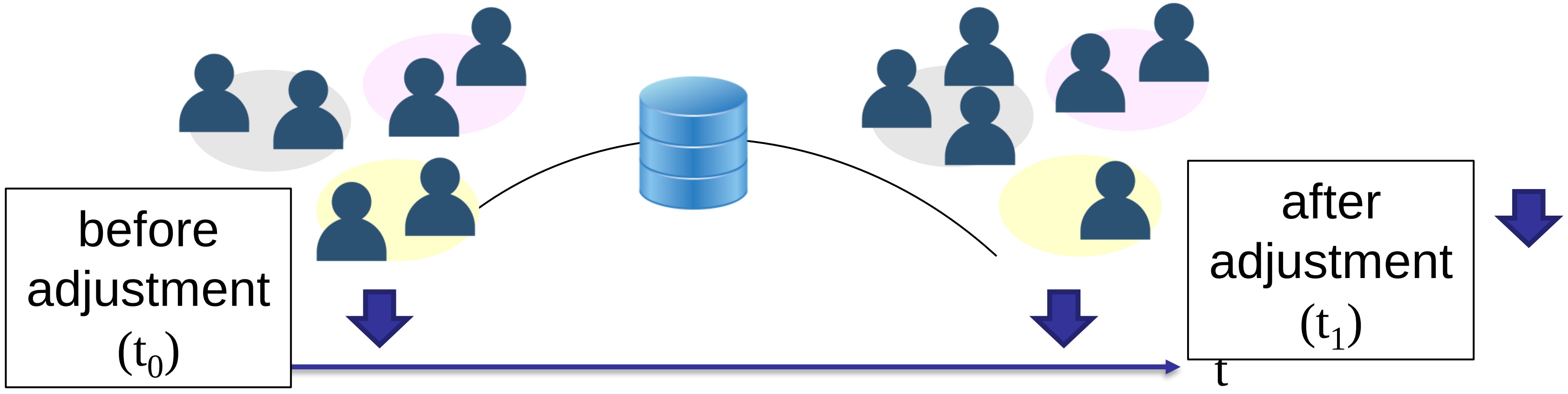
Number of sample workers in each area

Experimental conditions with only normal workers

- Numerical conditions

Symbol	Value
K	10
N	400
μ_i	50
σ_i	2, 7, 12, 17, 22, 27, 32, 37, 42, 47
η	10^{-5}

- Simulation conditions
  - Service provider computes estimates at time t_n
  - Define following two states with only normal workers
    - t_0 : before adjustment, t_1 : after adjustment

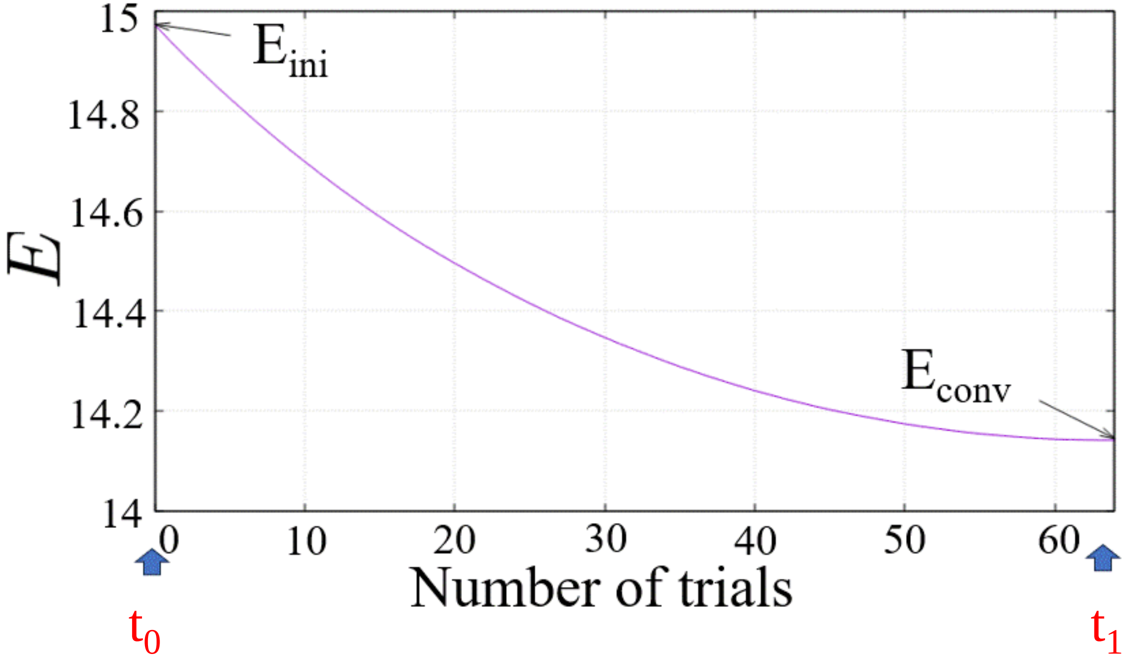


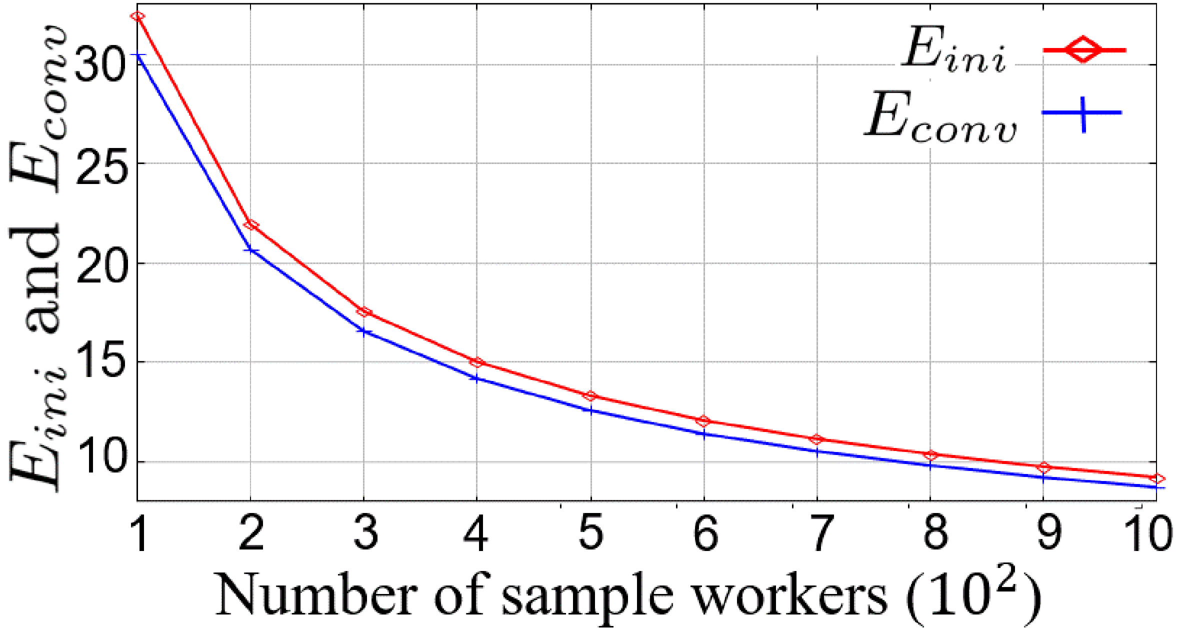
Time at which service provider calculates estimated value

Evaluation results with only normal workers

E_{ini} : total estimated error before adjustment

E_{conv}: total estimated error after adjustment





- E_{conv}, after applying proposed method, is lower than E_{ini}, confirming effect of suppressing total estimation error.
- More workers are placed in areas where standard deviation of worker measurements is greater.

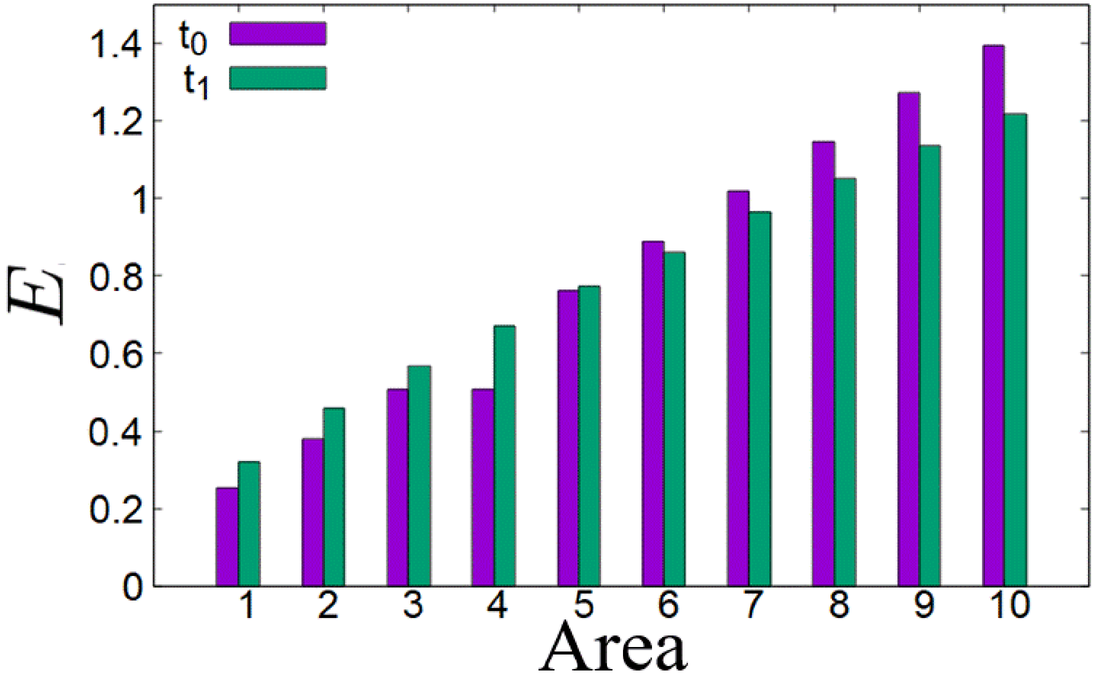
Experimental conditions with normal and malicious workers

- Numerical conditions

Symbol	Definition	Value
K	Number of areas	10
N	Total number of sample workers	400
μ_i	Average worker-reported value for each area i	50
σ_i	Standard deviation of worker-reported values for each area i	2, 3, 4, 5, 6, 7, 8, 9, 10, 11
m	Initial reported value of malicious workers	50
p	Percentage of malicious workers in population	0.05
η	Threshold used to determine difference in total error	10^{-5}

- Simulation conditions
  - Service provider computes estimates at time t_n
  - Define following two states with normal and malicious workers
    - t_0 : before adjustment, t_1 : after adjustment

Evaluation results with normal and malicious workers



- Confirming decrease in total error from time t_0 to time t_1 even when malicious workers exist
- Proposed method is effective even when malicious workers are mixed, although service providers have been sampling under conditions where collected data are indistinguishable

Summary

- Proposed method of setting optimal number of sample workers for each area with aim of minimizing total estimation error
  - With condition that total number of workers in sample is fixed
  - Even when malicious workers using DPA method exist, possible to avoid degradation of estimation accuracy

- Future work
  - Build a more secure MCS system by protecting worker data