Folien

Example for data correlation

- exp: salaries of persons in the same company, interests of close friends, ...

Graphic

- a simple data acquisition model
- Analyst buys data from red agent (salary data)
- Red agent‘s data is correlated with blue agents‘ data
- Blue agents suffer from information leakage
- Question of blue agent: should he also contribute his data to at least get a payment?
- This indicates that information leakage leads to oversharing thus to inefficiency in data markets

---

System Model

- consider online platform: one analyst and many agents
- Each agent owns data and cost
- Agents are divided into groups of same correlation strength (i.e. group of friends)
- analyst aims to buy data from agents to perform an estimation task (estimate the mean of the agents’ data)
- analyst would like to purchase all the private data to obtain an unbiased estimator
- However, limited budget, analyst wisely select data in order to balance between the bias and variance of the estimator

How the mechanism works

- First, analyst presents price menu to the agents
- Menu consists of a payment rule $P(\cdot)$ and selection probability $A(\cdot)$, both of which depend on the reported cost $\tilde{c}$ of the agents
- Second, given the menu, an agent decides if she would like to join the platform.
- agent who decides to join the platform reports her cost, which determines the payment and a selection probability.
- An agent who joined the platform is selected with probability $A(.)$ to sell its data to the analyst for a payment $P(.)$
- All agents who joined the platform further receive a participation benefit of the platform

---

Utility Fct

- If the agent does not join the platform, only experiences privacy cost $g(c, \theta_i; \alpha_i)$ induced by information leakage due to data correlation.

- If agent joins the platform but is not selected to report her data, her utility is $h(c, \theta_i; \alpha_i) + w(\bar{\theta})$, where $w(\bar{\theta})$ is the participation benefit, and h() is the privacy cost including both the cost of interacting with the platform and the privacy cost of information leakage due to data correlation
- If the agent joins the platform and is selected to report her data, she suffers her overall cost c. Based on the agent's reported cost $\tilde{c}$, she receives a payment $P(\tilde{c})$, thus her utility is $P(\tilde{c}, \theta_i; \alpha_i) - c + w(\bar{\theta})$.
- N denotes set of agents who joined the platform

3 important aspects: privacy cost, participation rate/benefit, correlation strength

Participation rate (Theta)
- Ratio of number of agents who join platform vs total number of agents

Participation Benefit (w)
- non-negative value received by agents joining the platform
- Network-Effect: w() is increasing in average participation rate
- Assume w() is continuous

Correlation Strength (Alpha)
- divide agents into groups of same correlation strength

Assumption (Monoton and Bounded)
- Monoton in cost since high cost means high value of privacy
- Privacy cost increases as more agents join the platform or as correlation strength increases (since more information leakage)
- $g() \leq h()$ since joining leaks data through interaction with the platform (even if no data reported)
- $h() \leq c$ since h is just a fraction of overall cost c

Assumption (h is linear in c)
- is shown to be appropriate for this setting in prior work

Expected utility function
- to sum up agent has the expected utility function if he decides to join or not

---

now: MECHANISM DESIGN

- Talk about common desirable properties of mechanisms
- Truthfulness: guarantees that rational agents will report their true cost
- Expected budget constraint: not realistic to have unlimited budget

---

Finally time to state the analyst problem
- Aims to minimise linear combination of bias and variance by designing payment function P and allocation rule A subject to truthfulness and budgetary constraints
- key trade-off: bias and variance
- Bias: not every agent participates, estimator could be biased towards agents who participate
- Exp. For bias: analyst wants to estimate percentage of population with HIV (sensitive information),
- High-cost for agents with HIV -> won't participate
- Bias towards agents without HIV (underestimate %HIV)

- analyst wants to control participation rate to further adjust bias by design the mechanism

- Variance: randomness of mechanism and data-cost pair of the agents (unknown to the analyst)

Equilibrium Characterization
- Define equilibrium w.r.t. participation rate, since mechanism impacts the participation rate, which again impacts bias and variance, which are to be minimised
- Notion of equilibrium guarantees that no agent wants to alter her decision given the equilibrium participation rate profile THETA*
- (participation rate profile = vector of average participation rate of each group)
- binary variable d = decision of joining (1) or not (0)
- Continuous distribution f_i of cost for agents in group i

Now ready to present first Theorem
Payment Function
- For simplicity only write h,b,tau however these are functions of participation rate and correlation strength
- This theorem gives us payment function as function of the allocation rule

- Won‘t prove whole theorem due to time restrictions, see paper for more detail

---

Proof 1.0

- Goal: show that the utility function is maximised if c=c~
- K is parameter that does not depend on reported cost

- Derivative of expected utility
- Recall that $c \geq h(c)$ [Assumption bounded & monoton]
- Get 2 cases

Proof 2.0

- Other direction: truthfulness of mechanism implies 1&2
- Idea: „reverse engeniere“ payment function from utility function

- V is parameter depending on c_max and Theta
- in the paper it is shown that indeed v = Tau

- Reporting true cost maximises utility
- Since $c-h(c) \geq 0$
- yet to show: the payment rule induces an equilibrium participation profile equal to the desired profile

---

Properties of the Mechanism

- analyst decides on desired participation rate, plots payment function and allocation rule based on that
- Mechanism induces a threshold-based participation w.r.t. a threshold cost
- Higher information leakage leads to lower payment; to see this plot payment function without information leakage (-h)

Optimising the worst-case

- Joint distribution of data and cost
- Don‘t need to consider payment function anymore (function of A)

Optimal Allocation Rule

- Adversary: selects q to destroy the analyst‘s utility

- Analyst: selects A to improve utility
- Convex-concave optimisation: equilibrium equals saddle-point, can be analytically found (through KKT-conditions)
- Theoretical result: allocation rule can have two optimal structures depending on system parameters

---

Impacts of information leakage
- Data-correlation vs. Payment: payment of an agent is decreased in data correlation strength,
- since higher data correlation leads to more privacy loss leads to lower payment

- Participation rate vs payment:
- Higher participation rate leads to more privacy loss leads to lower payment

---

Conclusion
- privacy loss due to information leakage due to data correlation potentially leads to smaller payment for data and encourages more agents to contribute their data