CS6501: Topics in Learning and Game Theory
(Fall 2019)

How Can Classifiers Induce Right Efforts?

Instructor: Haifeng Xu

Outline

- Motivations and Model
- Examples and Results

Decisions and Incentives

Often today, ML is used to assist decisions about human beings

Decisions and Incentives

Often today, ML is used to assist decisions about human beings

- ➢ Education

WILEY

EDUCATION SERVICES

Services & Solutions Blog About Us

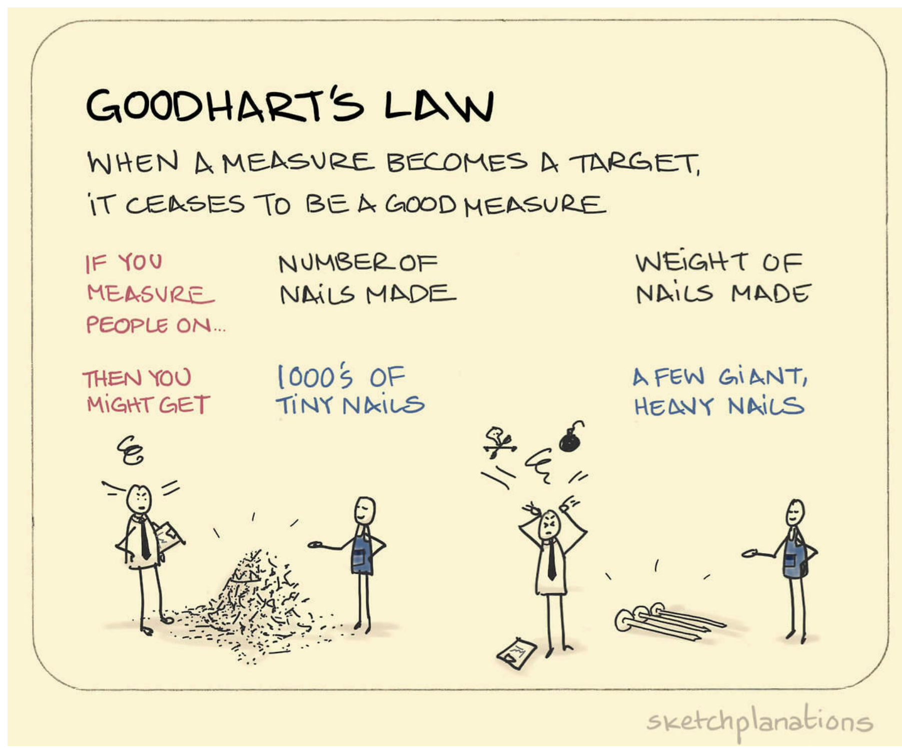
5 Ways Artificial Intelligence May Influence Higher Education Admissions & Retention

Artificial intelligence (AI) has officially entered the higher education realm, both hypothetically and in early practice. According to the report Artificial Intelligence Market in the US Education Sector, AI will grow at a compound annual rate of 47.7 percent from 2018 to 2022. Several technological and educational powerhouses will contribute to that growth as they commit substantial resources and personnel to develop digital platforms that use AI.

Decisions and Incentives

Often today, ML is used to assist decisions about human beings

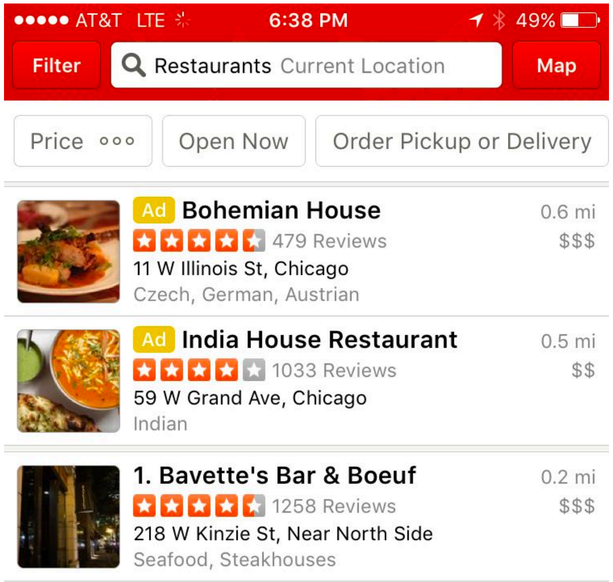
- Education
- When a measure becomes a target, gaming behaviors happen (Goodhart's Law)



Decisions and Incentives

Often today, ML is used to assist decisions about human beings

- ➢ Education
- ➢ When a measure becomes a target, gaming behaviors happen (Goodhart's Law)
- ➢ Many other applications: recommender systems, hiring, finance…
  - E.g., restaurants can game Yelp's ranking metric by pay for positive reviews or checkins



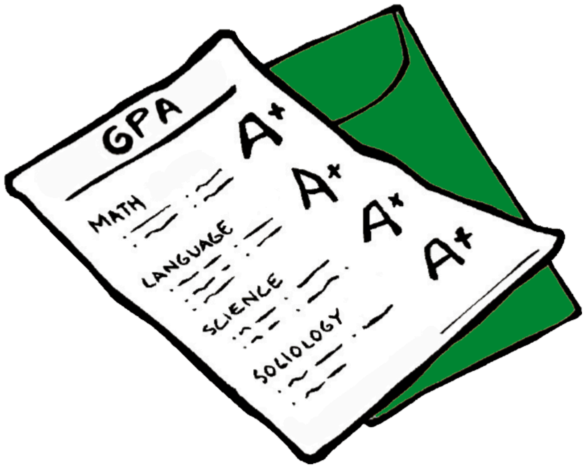
Decisions and Incentives

Often today, ML is used to assist decisions about human beings

- Education
- When a measure becomes a target, gaming behaviors happen (Goodhart's Law)
- Many other applications: recommender systems, hiring, finance…
  - E.g., restaurants can game Yelp's ranking metric by pay for positive reviews or checkins
- Particularly an issue when transparency is required

Education as a Running Example

Strategic Behaviors

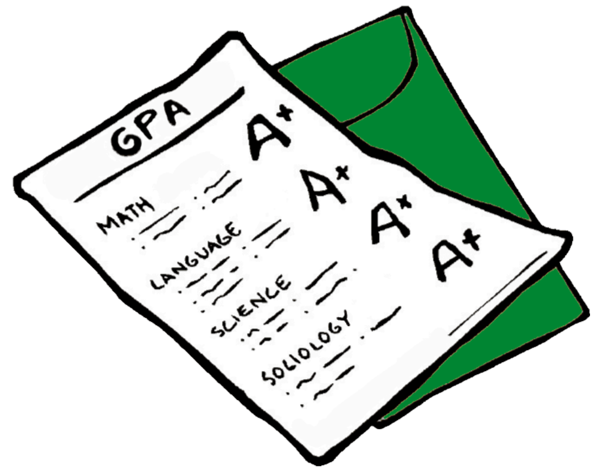


Goal/score
(determined by some measure)

Education as a Running Example

Strategic Behaviors

Desirable behavior ✓

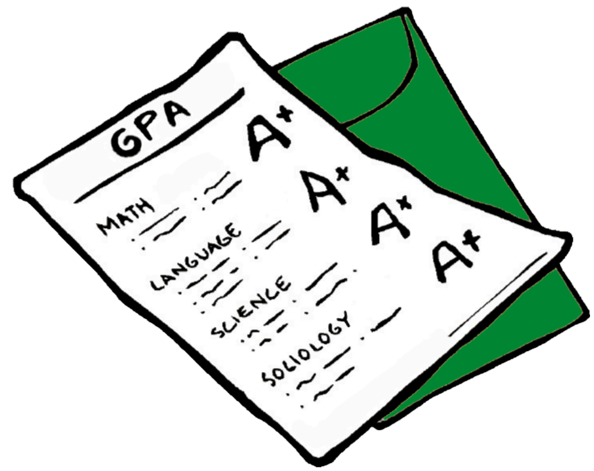


Goal/score
(determined by some measure)

Education as a Running Example

Strategic Behaviors

Undesirable behavior



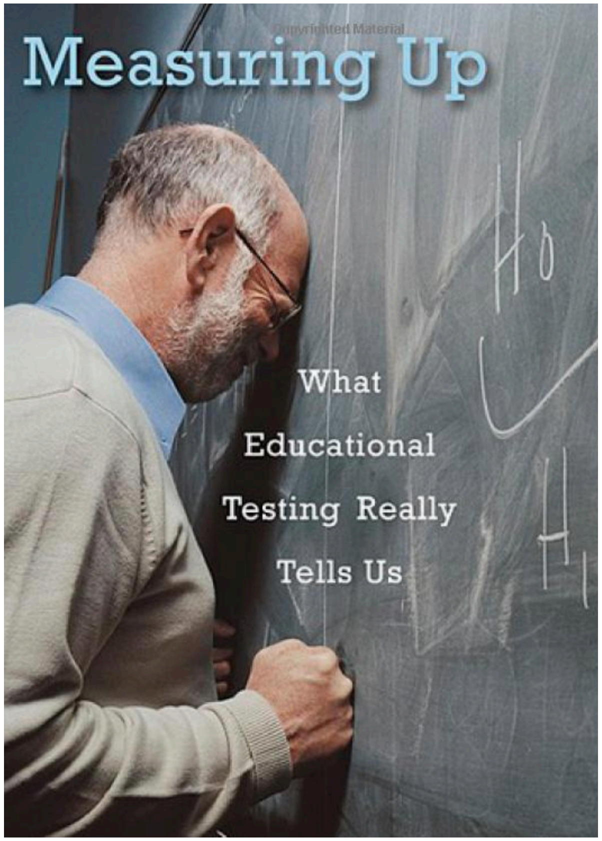
Goal/score
(determined by some measure)

Education as a Running Example

- Some strategic behaviors are desirable, and some are not

I think it's best to. . . distinguish between seven different types of test preparation: Working more effectively; Teaching more; Working harder; Reallocation; Alignment; Coaching; Cheating. The first three are what proponents of high-stakes testing want to see

-- Daniel M. Koretz, *Measuring up*



Education as a Running Example

- Some strategic behaviors are desirable, and some are not

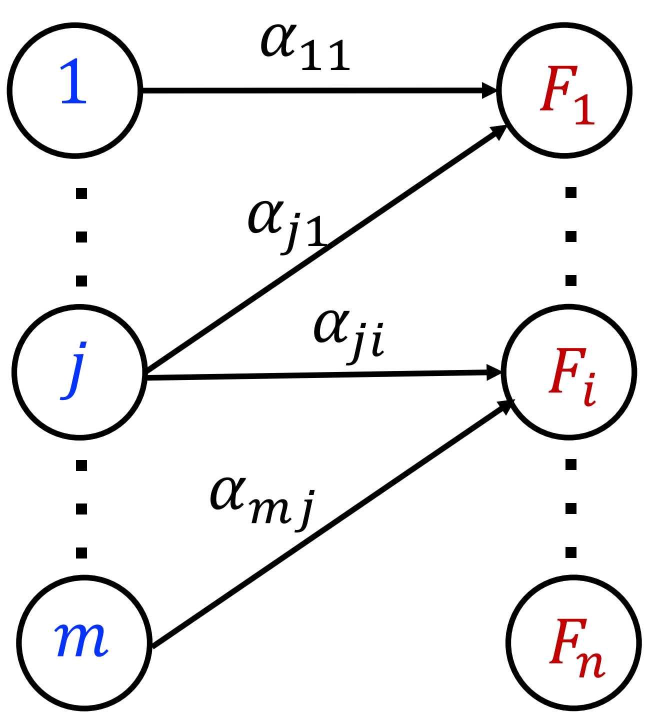
The Main Question

How to design decision rules to induce desirable strategic behaviors?

- Usually not possible to keep the rule confidential
- Should not simply use a rule that cannot be affected at all
- So, this requires careful design

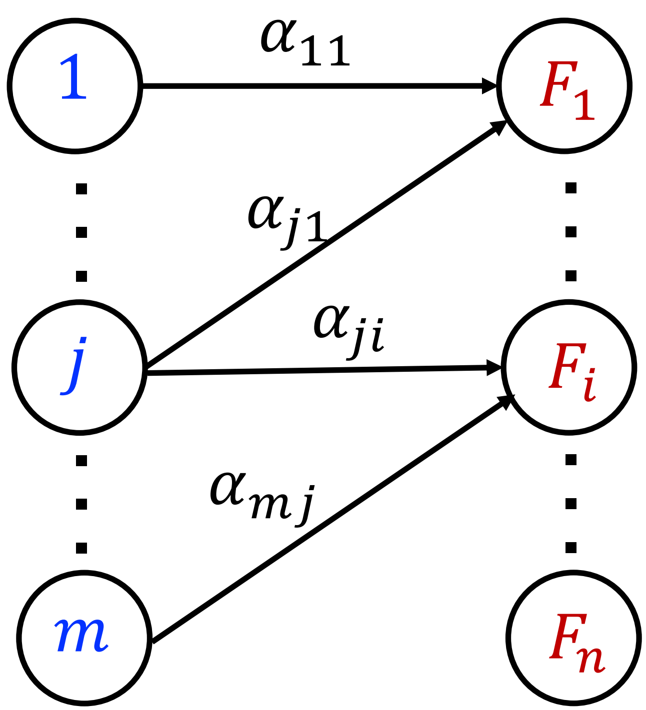
The Mathematical Model

- m available actions (e.g., study hard, cheating)
- n different features (e.g., HW grade, midterm grade)
- Each unit effort on action j results in $\alpha_{ji}(\geq 0)$ increase in feature i



A Game between Agent and Principal

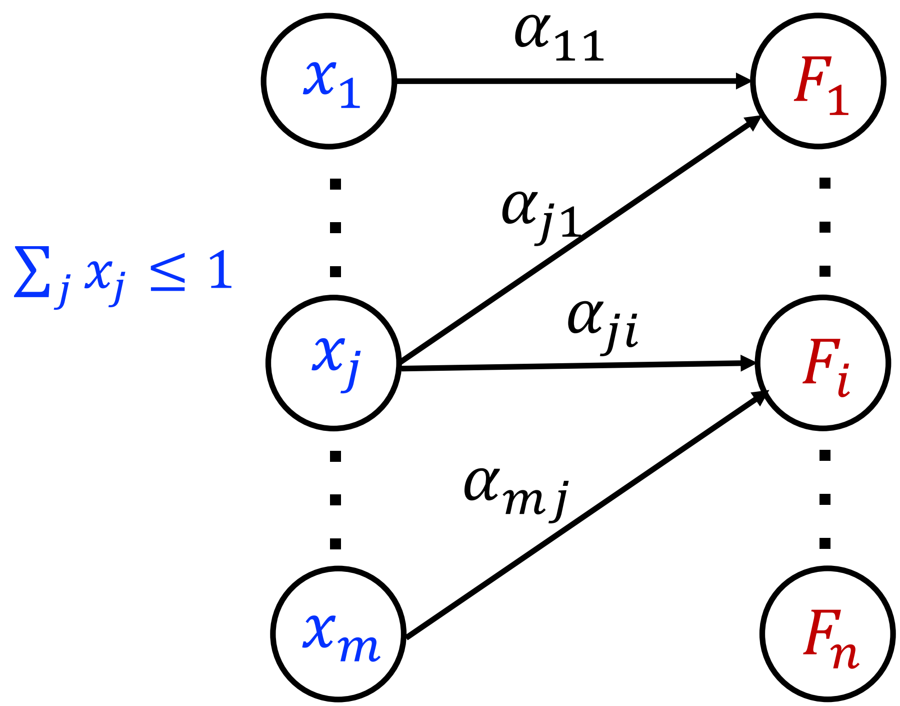
- Agent's action: allocation $(x_1, \cdots, x_m)$ of 1 unit of effort to actions



A Game between Agent and Principal

- Agent's action: allocation $(x_1, \cdots, x_m)$ of 1 unit of effort to actions
  - Effort profile $x(> 0)$ decides feature values

$$F_i = f_i(\sum_j x_j \alpha_{ji}) \quad \text{(an increasing concave fnc)}$$

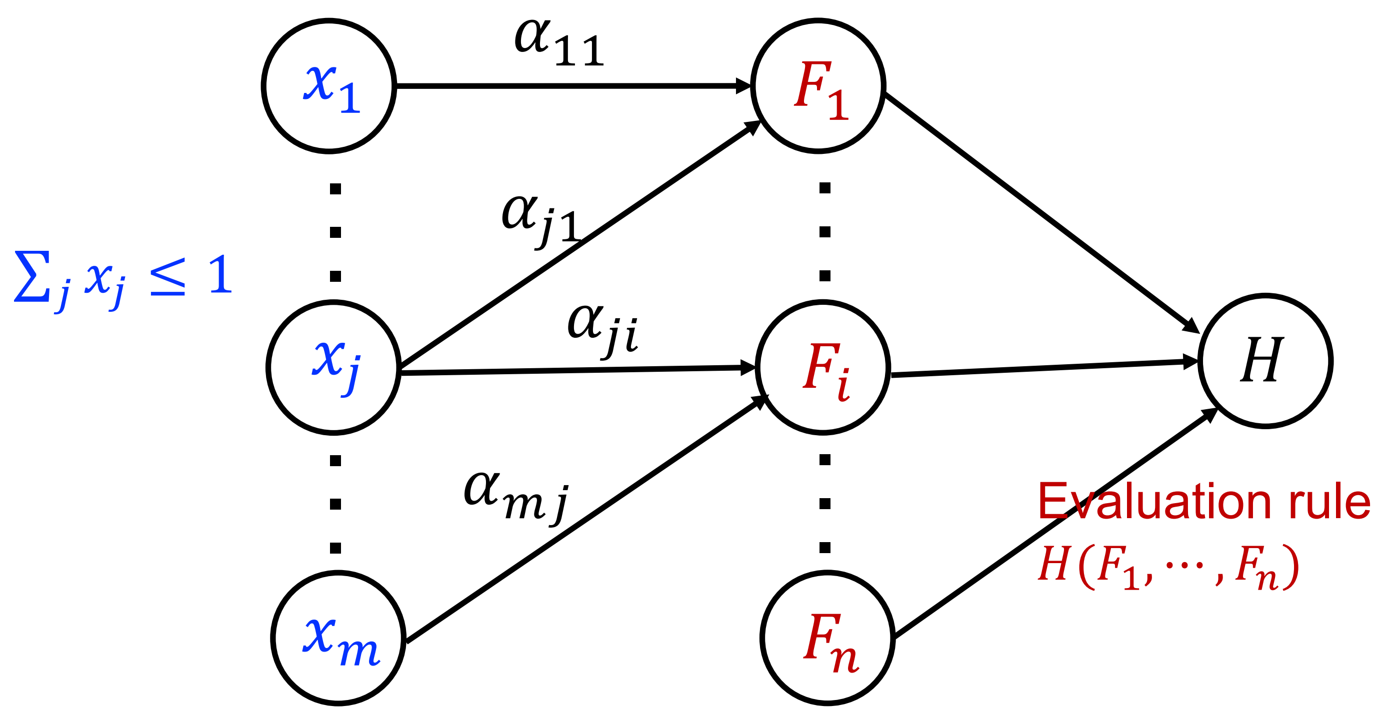


A Game between Agent and Principal

- Agent's action: allocation $(x_1, \cdots, x_m)$ of 1 unit of effort to actions
  - Effort profile $x (> 0)$ decides feature values

$$F_i = f_i(\sum_j x_j \alpha_{ji}) \quad \text{(an increasing concave fnc)}$$

- Principal's action: design the evaluation rule $H(F_1, \cdots, F_n)$
  - H is increasing in every feature, and publicly known (e.g., a grading rule)



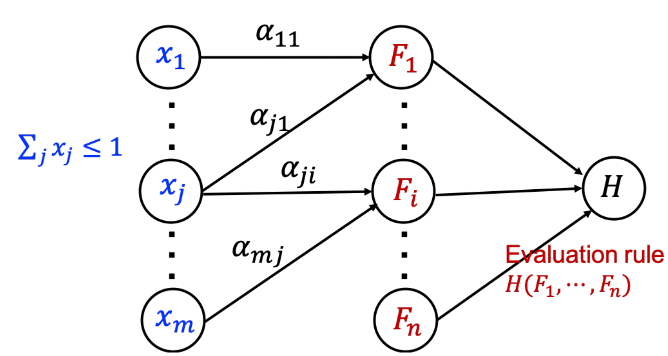
A Game between Agent and Principal

- Agent's action: allocation $(x_1, \cdots, x_m)$ of 1 unit of effort to actions
  - Effort profile $x (> 0)$ decides feature values

$$F_i = f_i(\sum_j x_j \alpha_{ji}) \quad \text{(an increasing concave fnc)}$$

- Principal's action: design the evaluation rule $H(F_1, \cdots, F_n)$
  - H is increasing in every feature, and publicly known (e.g., a grading rule)
- Principal has a desirable effort profile x^* (e.g., x^* = "work hard")
- Agent goal: choose x to maximize H

Q: Can the principal design H to induce her desirable x^*?

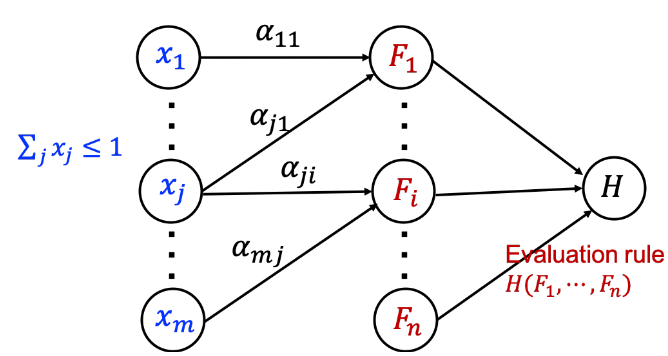


A Game between Agent and Principal

Q: Can the principal design H to induce her desirable x^*?

Relation to problems we studied before

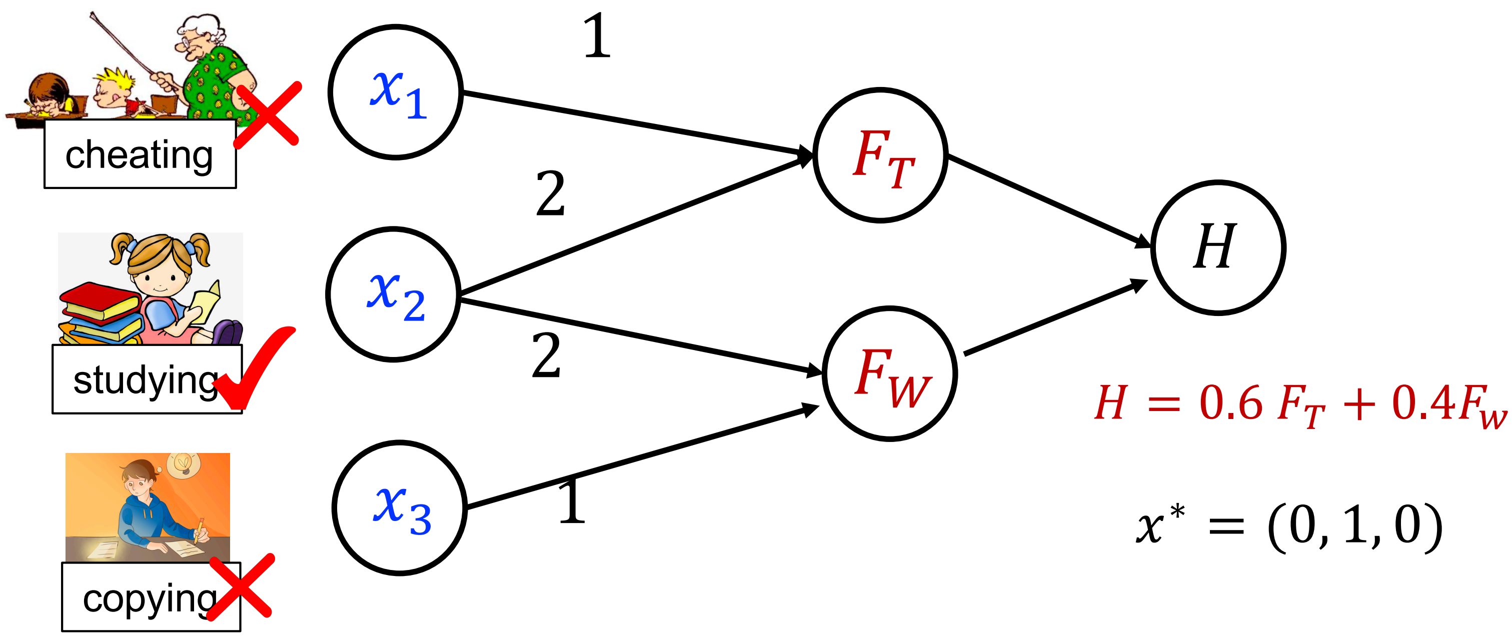
- This is a Stackelberg game
  - First, principal announces the evaluation rule H
  - Second, agent best responds to H by picking effort profile x
- This is a mechanism design problem
  - Want to design evaluation rule H to induce desirable response x^*
- More generally, this a *principal-agent mechanism design* problem
  - Rich literature in economics, explosive recent interest in EconCS



Outline

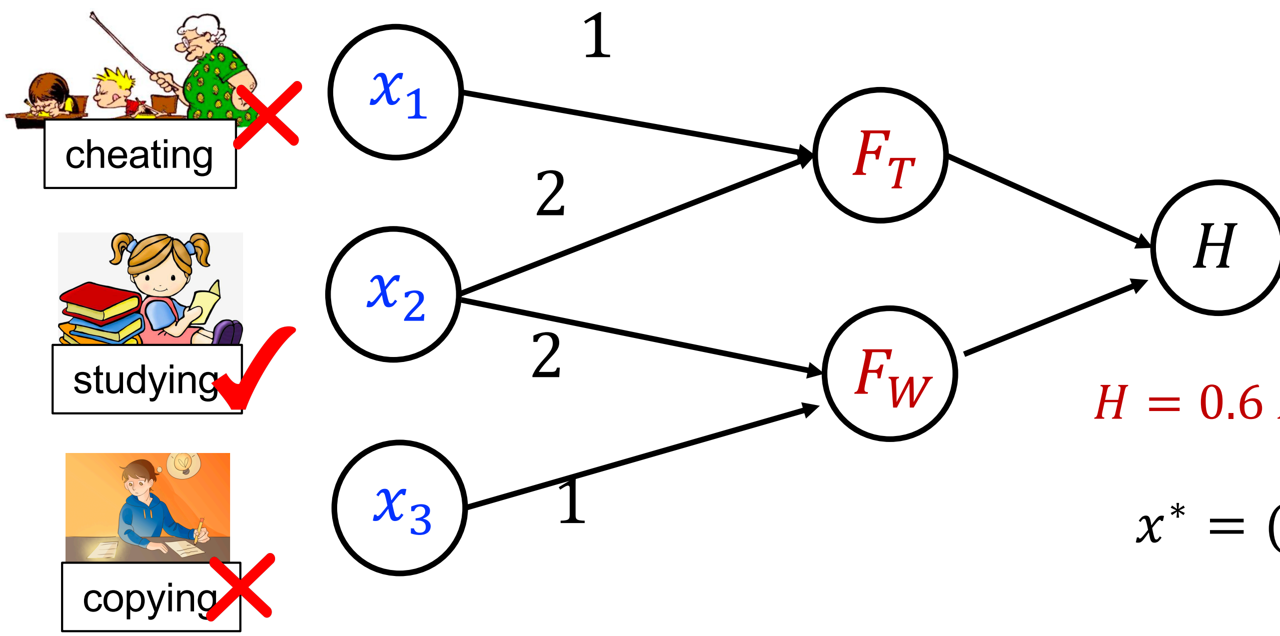
- Motivations and Model
- Examples and Results

Example: Classroom Setting



Q: Can the principal induce the desirable $x^* = (0,1,0)$?

Example: Classroom Setting



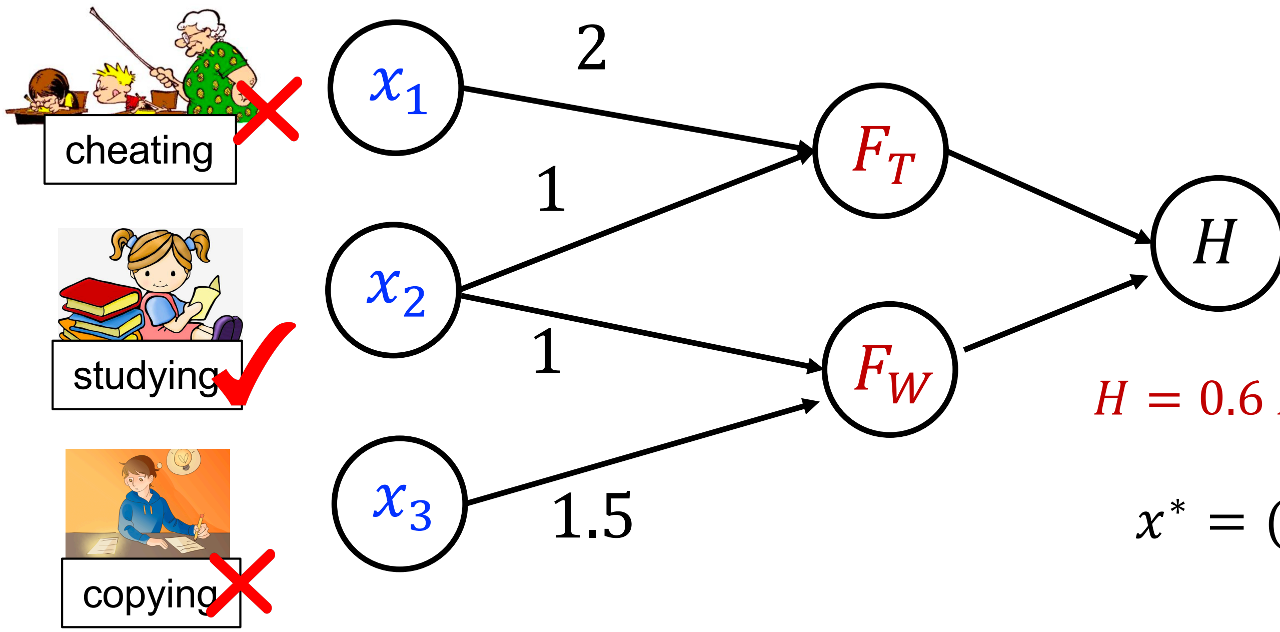
$$H = 0.6\,F_T + 0.4F_w$$

$$x^* = (0, 1, 0)$$

Q: Can the principal induce the desirable $x^* = (0,1,0)$?

- Ans: Yes
  - For any unit of effort on cheating or copying, agent would rather spend it on studying

Example: Classroom Setting

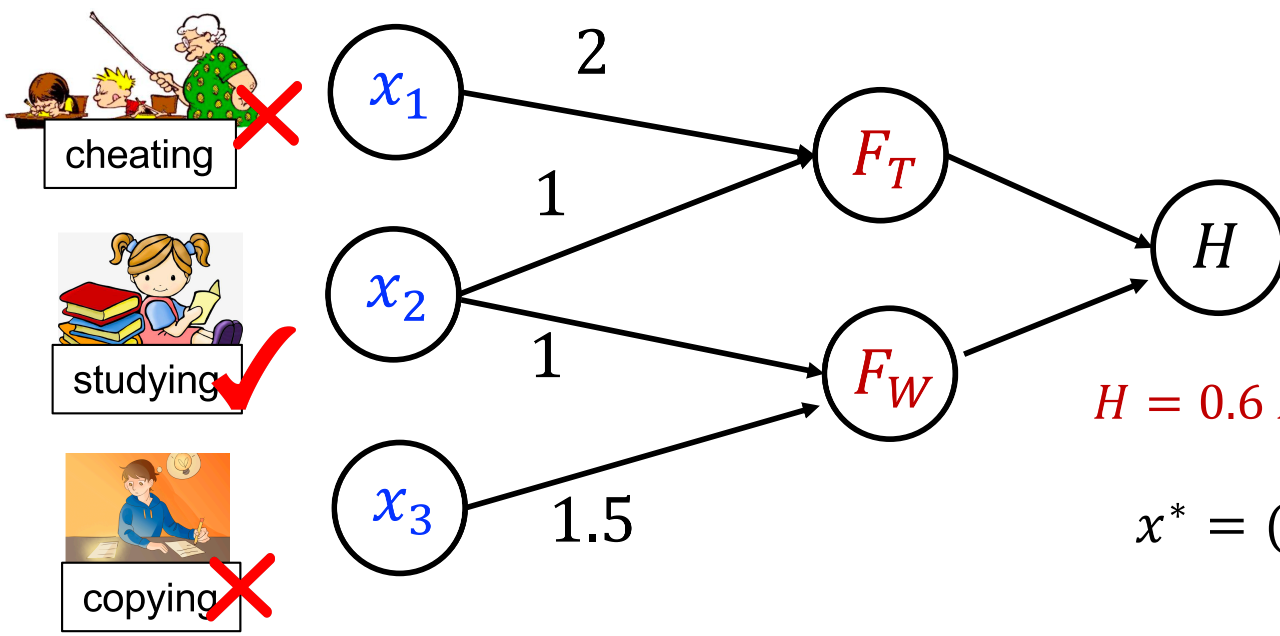


$H = 0.6\, F_T + 0.4 F_w$

$x^* = (0, 1, 0)$

Q: What about this setting?

Example: Classroom Setting



$H = 0.6\, F_T + 0.4 F_W$

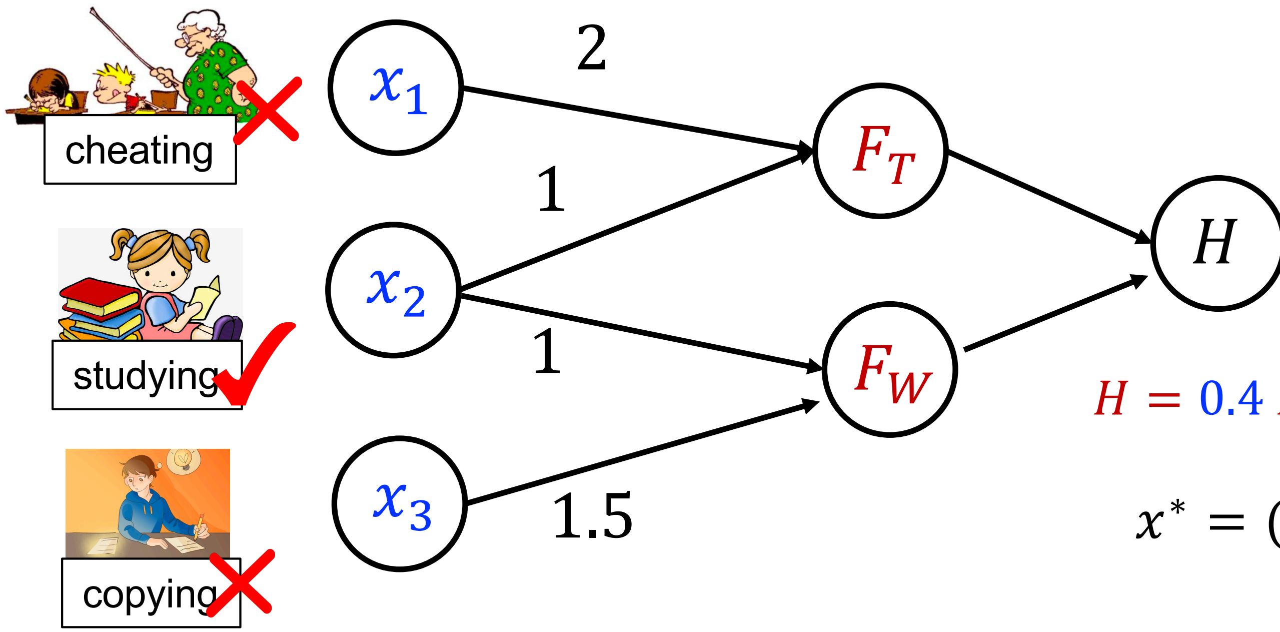
$x^* = (0, 1, 0)$

Q: What about this setting?

➢ Ans: No

- Spending 1 unit studying → H = 1
- Spending 1 unit on cheating → H = 1.2
- Problem: weight of exam is to large

Example: Classroom Setting

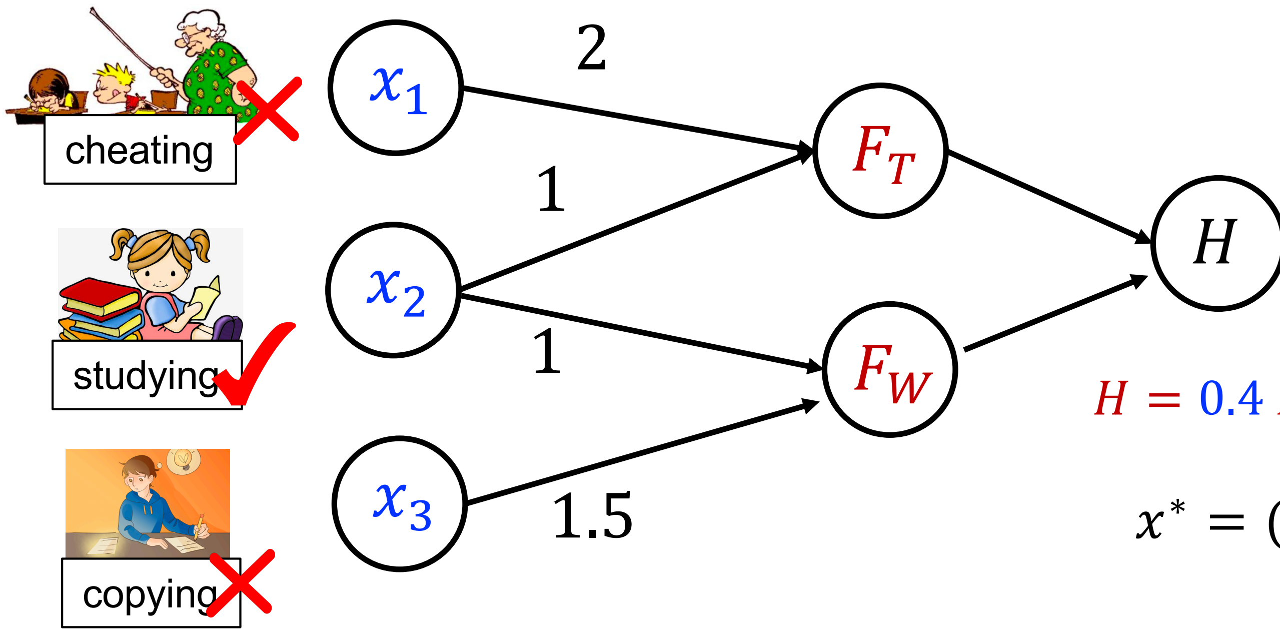


$H = 0.4\,F_T + 0.6F_w$

$x^* = (0, 1, 0)$

Q: What about changing H to our class's rule?

Example: Classroom Setting



$H = 0.4\, F_T + 0.6 F_w$

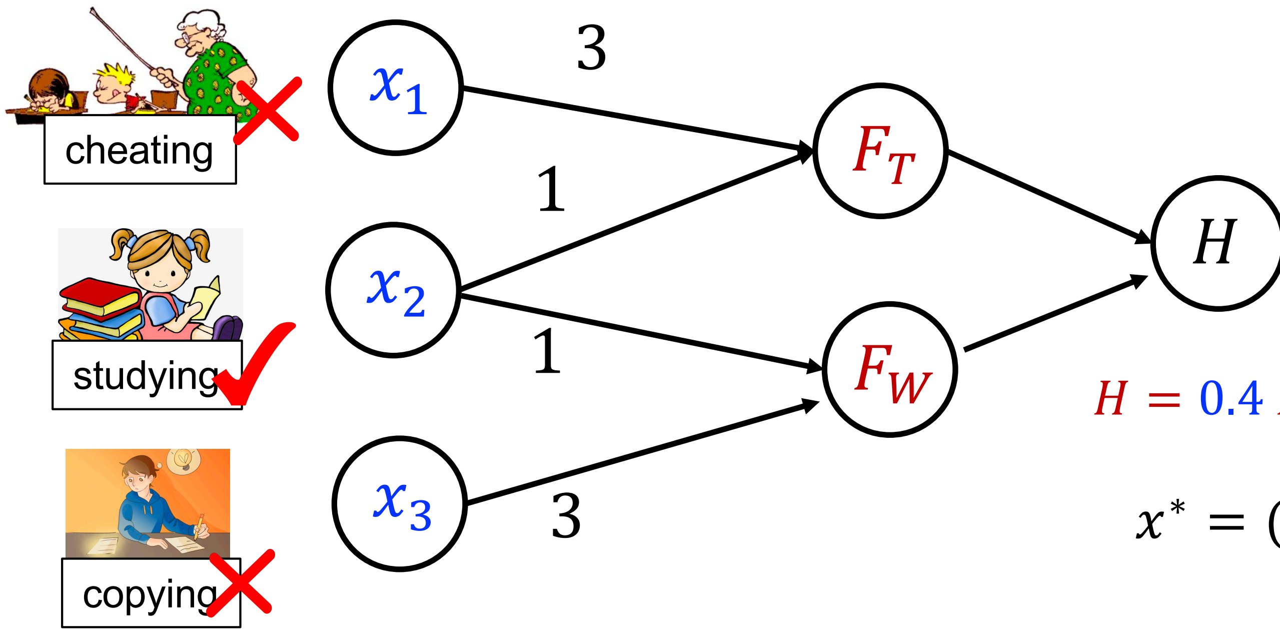
$x^* = (0, 1, 0)$

Q: What about changing H to our class's rule?

➢ Ans: Yes

- Spending 1 unit studying → H = 1
- Shifting any amount of effort to copying or cheating only decreases H
- Whether we can induce x^* does depends on our design of H

Example: Classroom Setting

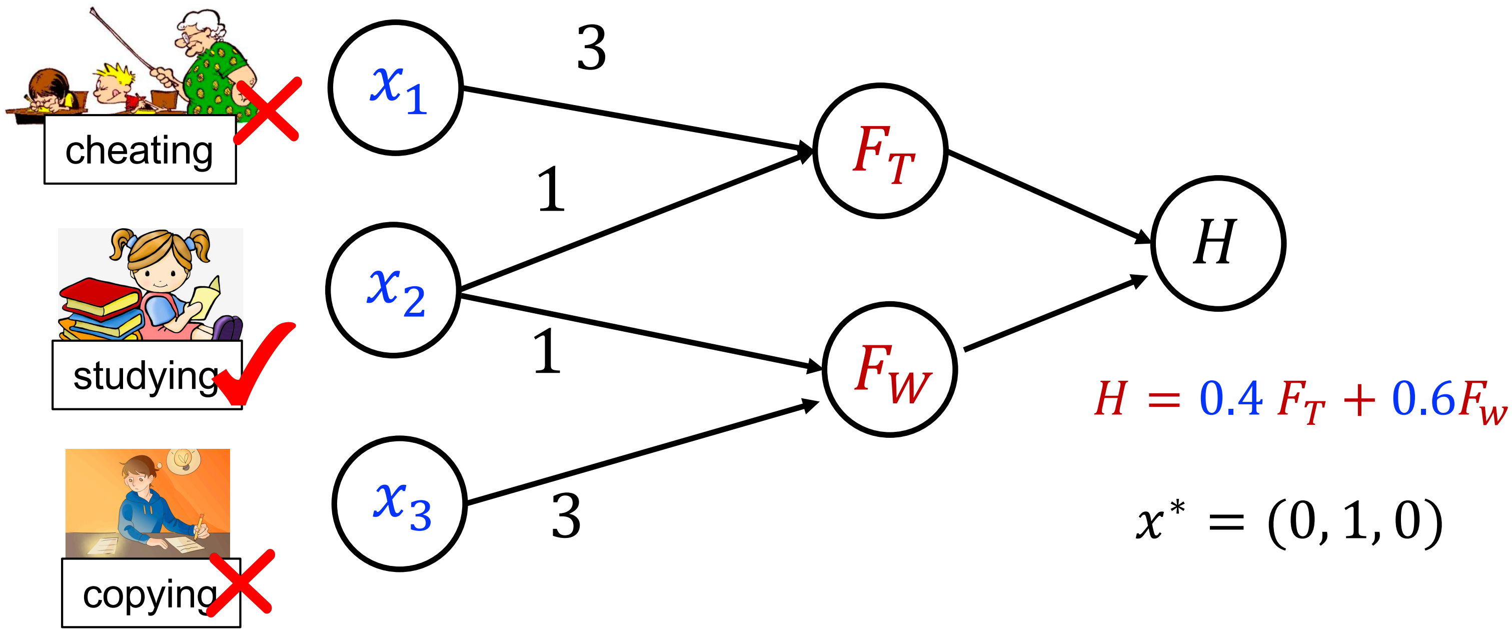


$$H = 0.4\,F_T + 0.6F_w$$

$$x^* = (0, 1, 0)$$

Q: What about these effort transition values?

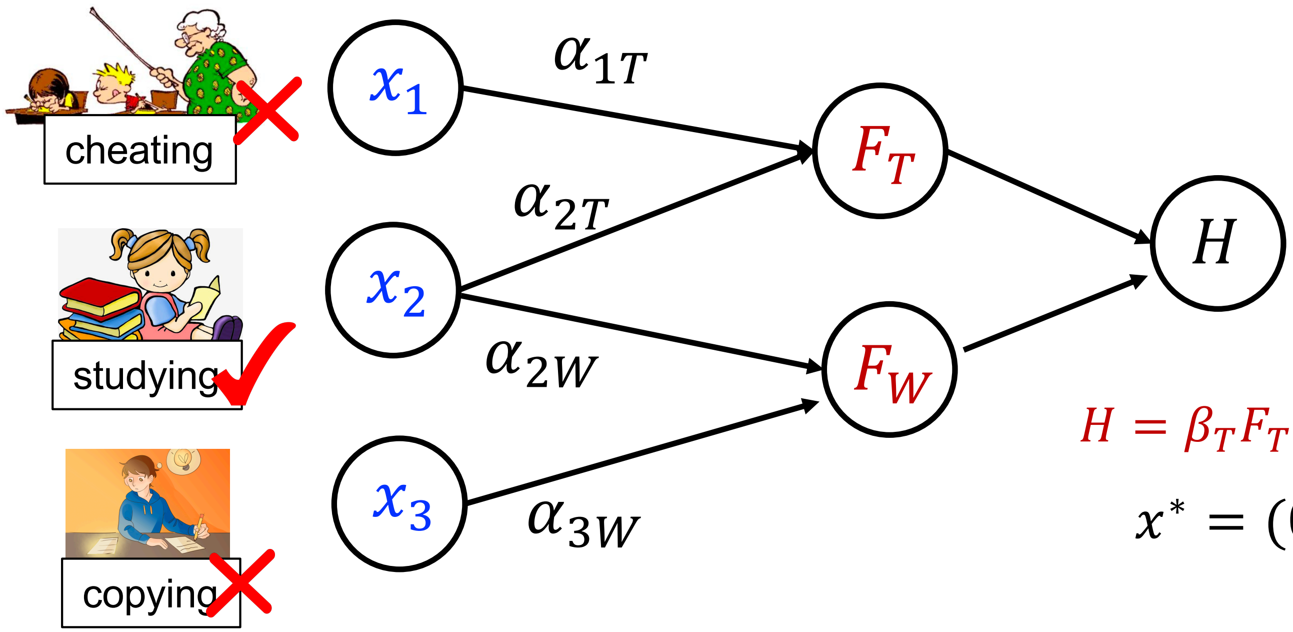
Example: Classroom Setting



Q: What about these effort transition values?

- Ans: No, regardless of what H you choose
  - For whatever (x_1, x_2, x_3), $(x_1 + \frac{x_2}{2}, 0, x_3 + \frac{x_2}{2})$ is better for agent
  - There are cases where x^* just cannot be induced regardless of H

Example: Classroom Setting

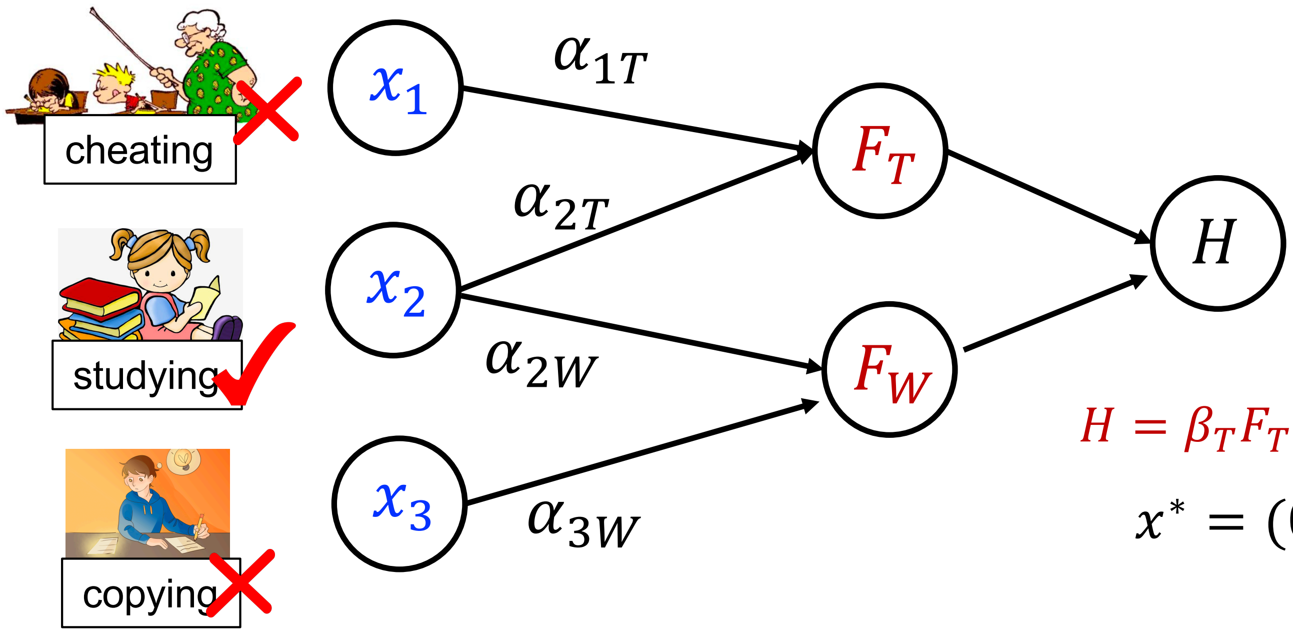


$H = \beta_T F_T + \beta_W F_w$

$x^* = (0, 1, 0)$

Q: In general, when would it be impossible to induce x^*?

Example: Classroom Setting



$H = \beta_T F_T + \beta_W F_w$

$x^* = (0, 1, 0)$

Q: In general, when would it be impossible to induce x^*?

- With $B = 1$ effort on studying, we get $(F_T, F_W) = (\alpha_{2T}, \alpha_{2W})$
- If $\exists$ (x_1, x_2, x_3) such that: (1) $x_1 + x_2 + x_3 < 1$; but (2) $x_1\alpha_{1T} + x_2\alpha_{2T} \geq \alpha_{2T}$ and $x_2\alpha_{2W} + x_3\alpha_{3W} \geq \alpha_{2W}$, then cannot induce effort on studying
  - This condition does not depend on H

Which Effort Profile Can Be Incentivized, and How?

- Let's focus on the special case $x^* = e_j$ for some j
- Previous argument shows a necessary condition

There is no $(x_1, \cdots, x_m) \geq 0$ such that:

1. $\sum_j x_j < 1$
2. $x \cdot \alpha \geq \alpha(j,\cdot)$

Note: x here is a row vector

Which Effort Profile Can Be Incentivized, and How?

- Let's focus on the special case $x^* = e_j$ for some j
- Previous argument shows a necessary condition

Define $\kappa_j \coloneqq \min_x \sum_j x_j$ subject to (1) $x \cdot \alpha \geq \alpha(j,\cdot)$; (2) $x \geq 0$. A necessary condition is $\kappa_j \geq 1$.

There is no $(x_1, \cdots, x_m) \geq 0$ such that:

1. $\sum_j x_j < 1$
2. $x \cdot \alpha \geq \alpha(j,\cdot)$

Note: x here is a row vector

Which Effort Profile Can Be Incentivized, and How?

- Let's focus on the special case $x^* = e_j$ for some j
- Previous argument shows a necessary condition

Define $\kappa_j := \min_x \sum_j x_j$ subject to (1) $x \cdot \alpha \geq \alpha(j,\cdot)$; (2) $x \geq 0$. A necessary condition is $\kappa_j \geq 1$.

Note: $\kappa_j \leq 1$ always because $x = e_j$ is feasible

Which Effort Profile Can Be Incentivized, and How?

- Let's focus on the special case $x^* = e_j$ for some j
- Previous argument shows a necessary condition

Define $\kappa_j := \min_x \sum_j x_j$ subject to (1) $x \cdot \alpha \geq \alpha(j,\cdot)$; (2) $x \geq 0$. A necessary condition is $\kappa_j = 1$.

Note: $\kappa_j \leq 1$ always because $x = e_j$ is feasible

Which Effort Profile Can Be Incentivized, and How?

- Let's focus on the special case $x^* = e_j$ for some j
- Previous argument shows a necessary condition

Define $\kappa_j \coloneqq \min_x \sum_j x_j$ subject to (1) $x \cdot \alpha \geq \alpha(j,\cdot)$; (2) $x \geq 0$. A necessary condition is $\kappa_j = 1$.

Theorem: (1) There is a way to incentivize e_j if and only if $\kappa_j = 1$. (2) Whenever e_j can be incentivized, there is a linear H of form $H = \sum_i \beta_i F_i$ that incentivizes e_j.

Which Effort Profile Can Be Incentivized, and How?

- Let's focus on the special case $x^* = e_j$ for some j
- Previous argument shows a necessary condition

Define $\kappa_j \coloneqq \min_x \sum_j x_j$ subject to (1) $x \cdot \alpha \geq \alpha(j,\cdot)$; (2) $x \geq 0$. A necessary condition is $\kappa_j = 1$.

Theorem: (1) There is a way to incentivize e_j if and only if $\kappa_j = 1$. (2) Whenever e_j can be incentivized, there is a linear H of form $H = \sum_i \beta_i F_i$ that incentivizes e_j.

Proof

- We know if $\kappa_j < 1$, we cannot incentivize e_j, so $\kappa_j = 1$ is necessary
- To prove sufficiency, we construct a linear H that indeed induce e_j when $\kappa_j = 1$

Linear H That Induces e_j

- Consider $H = \sum_i \beta_i F_i$, agent's optimization problem

$$\max_{x \in \Delta_m} H = \sum_i \beta_i \cdot f_i\left(\sum_k x_k \alpha_{ki} \right)$$

Value of feature i

Linear H That Induces e_j

- Consider $H = \sum_i \beta_i F_i$, agent's optimization problem

$$\max_{x \in \Delta_m} H = \sum_i \beta_i \cdot f_i\left(\sum_k x_k \alpha_{ki} \right)$$

- When would the optimal solution be $x^* = e_j$?
  - Ans: when $\frac{\partial H}{\partial x_j}|_{x=x^*} \geq \frac{\partial H}{\partial x_{j'}}|_{x=x^*}$ for all j' (verify it after class)
  - Spell the derivatives out:

$$\sum_i \beta_i \cdot \alpha_{ji} \cdot f_i'\left(\sum_k x_k^* \alpha_{ki}\right) \geq \sum_i \beta_i \cdot \alpha_{j'i} \cdot f_i'\left(\sum_k x_k^* \alpha_{ki}\right), \quad \forall j' \qquad \text{Eq.(1)}$$

Linear H That Induces e_j

- Consider $H = \sum_i \beta_i F_i$, agent's optimization problem

$$\max_{x \in \Delta_m} H = \sum_i \beta_i \cdot f_i\left(\sum_k x_k \alpha_{ki} \right)$$

- When would the optimal solution be $x^* = e_j$?
  - Ans: when $\frac{\partial H}{\partial x_j}|_{x=x^*} \geq \frac{\partial H}{\partial x_{j'}}|_{x=x^*}$ for all j' (verify it after class)
  - Spell the derivatives out:

$$\sum_i \beta_i \cdot \alpha_{ji} \cdot f_i'\left(\sum_k x_k^* \alpha_{ki}\right) \geq \sum_i \beta_i \cdot \alpha_{j'i} \cdot f_i'\left(\sum_k x_k^* \alpha_{ki}\right), \quad \forall j' \quad \text{Eq.(1)}$$

Q: Given $\tau_j = 1$, do there exist $\beta \neq 0$ so that Eq. (1) holds?

- Eq (1) is also a set of linear constraints on β
- Ans: yes, through an elegant duality argument

Choosing the β

- Goal: $\sum_i \beta_i \cdot \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki}) \geq \sum_i \beta_i \cdot \alpha_{j'i} \cdot f_i'(\sum_k x_k^* \alpha_{ki}), \quad \forall j'$
- Let $A_{j,i} = \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki})$ which is a constant (x^* is given)
  - Let $A(j,\cdot)$ denotes the j'th row
- Need to check the linear system

$\exists \beta \neq 0$ such that

$$[A(j,\cdot)] \cdot \beta^T \geq [A(j',\cdot)] \cdot \beta^T, \forall j'$$

$$\beta \geq 0$$

Choosing the β

- Goal: $\sum_i \beta_i \cdot \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki}) \geq \sum_i \beta_i \cdot \alpha_{j'i} \cdot f_i'(\sum_k x_k^* \alpha_{ki}), \quad \forall j'$
- Let $A_{j,i} = \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki})$ which is a constant (x^* is given)
  - Let $A(j,\cdot)$ denotes the j'th row
- Need to check the linear system

$$\max_\beta \; [A(j,\cdot)] \cdot \beta^T$$
$$\text{s.t.} \quad \mathbf{1} \geq A \cdot \beta^T, \forall k$$
$$\beta \geq 0$$

$\Leftrightarrow$

$\exists \beta \neq 0$ such that

$$[A(j,\cdot)] \cdot \beta^T \geq [A(j',\cdot)] \cdot \beta^T, \forall j'$$
$$\beta \geq 0$$

obtains opt ≥ 1

Choosing the β

- Goal: $\sum_i \beta_i \cdot \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki}) \geq \sum_i \beta_i \cdot \alpha_{j'i} \cdot f_i'(\sum_k x_k^* \alpha_{ki}), \quad \forall j'$
- Let $A_{j,i} = \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki})$ which is a constant (x^* is given)
  - Let $A(j,\cdot)$ denotes the j'th row
- Need to check the linear system

$$\max_{\beta} \ [A(j,\cdot)] \cdot \beta^T$$
$$\text{s.t.} \ \ \mathbf{1} \geq A \cdot \beta^T, \forall k$$
$$\beta \geq 0$$

Dual LP $\Longrightarrow$

$$\min_{y} \ \mathbf{1} \cdot y^T$$
$$\text{s.t.} \ \ y \cdot A \geq A(j,:)$$
$$y \geq 0$$

obtains opt ≥ 1

Choosing the β

- Goal: $\sum_i \beta_i \cdot \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki}) \geq \sum_i \beta_i \cdot \alpha_{j'i} \cdot f_i'(\sum_k x_k^* \alpha_{ki})$, $\forall j'$
- Let $A_{j,i} = \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki})$ which is a constant (x^* is given)
  - Let $A(j,\cdot)$ denotes the j'th row
- Need to check the linear system

$$\max_{\beta} \ [A(j,\cdot)] \cdot \beta^T$$
$$\text{s.t.} \ \mathbf{1} \geq A \cdot \beta^T, \forall k$$
$$\beta \geq 0$$

obtains opt ≥ 1

Dual LP $\Longrightarrow$

$$\min_{y} \ \mathbf{1} \cdot y^T$$
$$\text{s.t.} \ y \cdot A \geq A(j,:)$$
$$y \geq 0$$

- The constraint is
  $\sum_k y_k \alpha_{ki} \cdot f_i' \geq \alpha_{ji} \cdot f_i'$, $\forall i$

i.e., $\sum_k y_k \alpha_{ki} \geq \alpha_{ji}$, $\forall i$

Choosing the β

- Goal: $\sum_i \beta_i \cdot \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki}) \geq \sum_i \beta_i \cdot \alpha_{j'i} \cdot f_i'(\sum_k x_k^* \alpha_{ki}), \quad \forall j'$
- Let $A_{j,i} = \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki})$ which is a constant (x^* is given)
  - Let $A(j,\cdot)$ denotes the j'th row
- Need to check the linear system

$$\max_{\beta} \ [A(j,\cdot)] \cdot \beta^T$$
$$\text{s.t.} \quad \mathbf{1} \geq A \cdot \beta^T, \forall k$$
$$\beta \geq 0$$

Dual LP ⟹

$$\min_{y} \ \mathbf{1} \cdot y^T$$
$$\text{s.t.} \quad y \cdot A \geq A(j,:)$$
$$y \geq 0$$

obtains opt ≥ 1

- The constraint is
  $\sum_k y_k \alpha_{ki} \cdot f_i' \geq \alpha_{ji} \cdot f_i', \ \forall i$

  i.e., $\sum_k y_k \alpha_{ki} \geq \alpha_{ji}, \ \forall i$
- Dual opt is exactly the def of $\kappa_j (= 1)$

Choosing the β

- Goal: $\sum_i \beta_i \cdot \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki}) \geq \sum_i \beta_i \cdot \alpha_{j'i} \cdot f_i'(\sum_k x_k^* \alpha_{ki}), \quad \forall j'$
- Let $A_{j,i} = \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki})$ which is a constant (x^* is given)
  - Let $A(j,\cdot)$ denotes the j'th row
- Need to check the linear system

$$\max_{\beta} \ [A(j,\cdot)] \cdot \beta^T$$
$$\text{s.t.} \quad \mathbf{1} \geq A \cdot \beta^T, \forall k$$
$$\beta \geq 0$$

Dual LP ⟹

$$\min_{y} \ \mathbf{1} \cdot y^T$$
$$\text{s.t.} \quad y \cdot A \geq A(j,:)$$
$$y \geq 0$$

obtains opt ≥ 1

- The constraint is
  $\sum_k y_k \alpha_{ki} \cdot f_i' \geq \alpha_{ji} \cdot f_i', \ \forall i$

  i.e., $\sum_k y_k \alpha_{ki} \geq \alpha_{ji}, \ \forall i$

- Dual opt is exactly the def of $\kappa_j (= 1)$

⟸

Primal opt $= 1$
β can be easily constructed

Choosing the β

- Goal: $\sum_i \beta_i \cdot \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki}) \geq \sum_i \beta_i \cdot \alpha_{j'i} \cdot f_i'(\sum_k x_k^* \alpha_{ki}), \quad \forall j'$
- Let $A_{j,i} = \alpha_{ji} \cdot f_i'(\sum_k x_k^* \alpha_{ki})$ which is a constant (x^* is given)
  - Let $A(j,\cdot)$ denotes the j'th row
- Need to check the linear system

$$\max_{\beta} \ [A(j,\cdot)] \cdot \beta^T \quad \text{s.t. } \mathbf{1} \geq A \cdot \beta^T, \forall k \quad \beta \geq 0$$

Dual LP ⟹

$$\min_{y} \ \mathbf{1} \cdot y^T \quad \text{s.t. } y \cdot A \geq A(j,:) \quad y \geq 0$$

obtains opt ≥ 1

- The constraint is
  $\sum_k y_k \alpha_{ki} \cdot f_i' \geq \alpha_{ji} \cdot f_i', \ \forall i$

  i.e., $\sum_k y_k \alpha_{ki} \geq \alpha_{ji}, \ \forall i$

- Dual opt is exactly the def of $\kappa_j (= 1)$

⟸ Primal opt $= 1$ ✓
β can be easily constructed

General x^*

- Similar conclusion holds with similar proof
- It turns out that the condition depends on S^*, the support of x^*

Theorem: (1) There is a way to incentivize x^* if and only if $\kappa_{S^*} = 1$ for some suitably defined κ_{S^*}. (2) Whenever x^* can be incentivized, there is a linear H that incentivizes x^*.

Optimization Version of the Problem

- Previously, principal has a single x^* to induce
  - Some of x^* can be incentivized, and some cannot
- A natural optimization version of the problem
  - Among all incentivizable x^*, how can principal incentivize the "best" one
  - Assume a utility function $g(x)$ over x

Optimization Version of the Problem

- Previously, principal has a single x^* to induce
  - Some of x^* can be incentivized, and some cannot
- A natural optimization version of the problem
  - Among all incentivizable x^*, how can principal incentivize the "best" one
  - Assume a utility function $g(x)$ over x
- Problem: maximize $g(x)$ subject to x is incentivizable

Theorem: The above problem is NP-hard, even when g is concave.

Open question:

- What kind of g can be optimized? Linear?
- What kind effort transition graph makes the problem more tractable?

Thank You

Haifeng Xu
University of Virginia
hx4ad@virginia.edu