A Mean-Field Stackelberg Game Approach for Obfuscation Adoption in Empirical Risk Minimization

Jeffrey Pawlick and Quanyan Zhu

New York University Tandon School of Engineering

Supported in part by an NSF IGERT grant through the Center for Interdisciplinary Studies in Security and Privacy (CRISSP) at New York University

Tracking Online and in the Internet of Things

- Online behavior is captured by third-party trackers and fingerprinting technologies.
- Internet of things (IoT) devices capture behavioral data.
  - Accelerometers, heart rate sensors
  - Sleep trackers, food logs
- Machine learning algorithms reveal information about race and political party [Kosinski et *al*. 2013], mood and personality type [Peppet 2014].

Obfuscation Adoption

- Obfuscation:

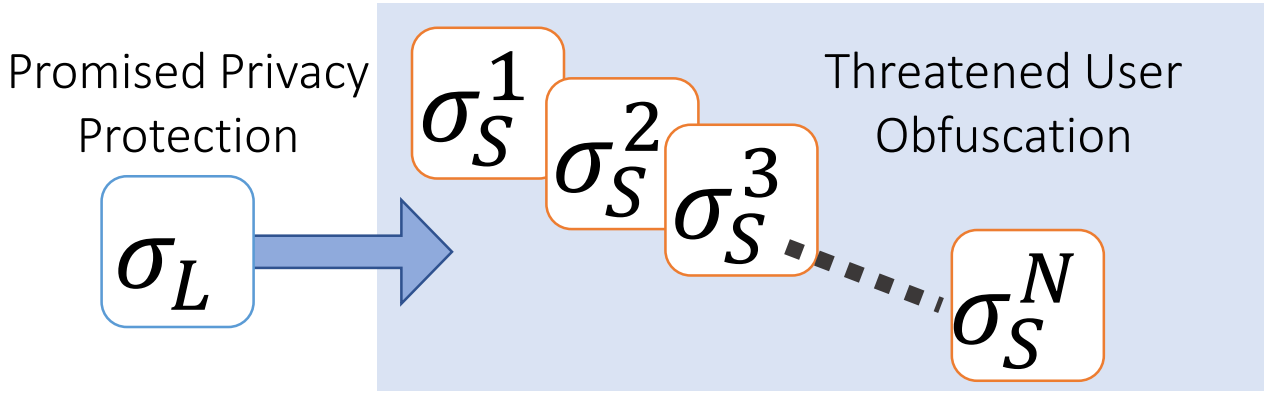
  "the deliberate addition of ambiguous, confusing, or misleading information to interfere with surveillance and data collection"

  [Brunton & Nissenbaum 2015].

- Examples:
  - *TrackMeNot* [Howe & Nissenbaum 2009]
  - *CacheCloak* [Meyerowitz & Choudhury 2009]
- Question: Can obfuscation adoption force machine learning agents to adopt privacy protection?

Modeling Obfuscation using Game Theory

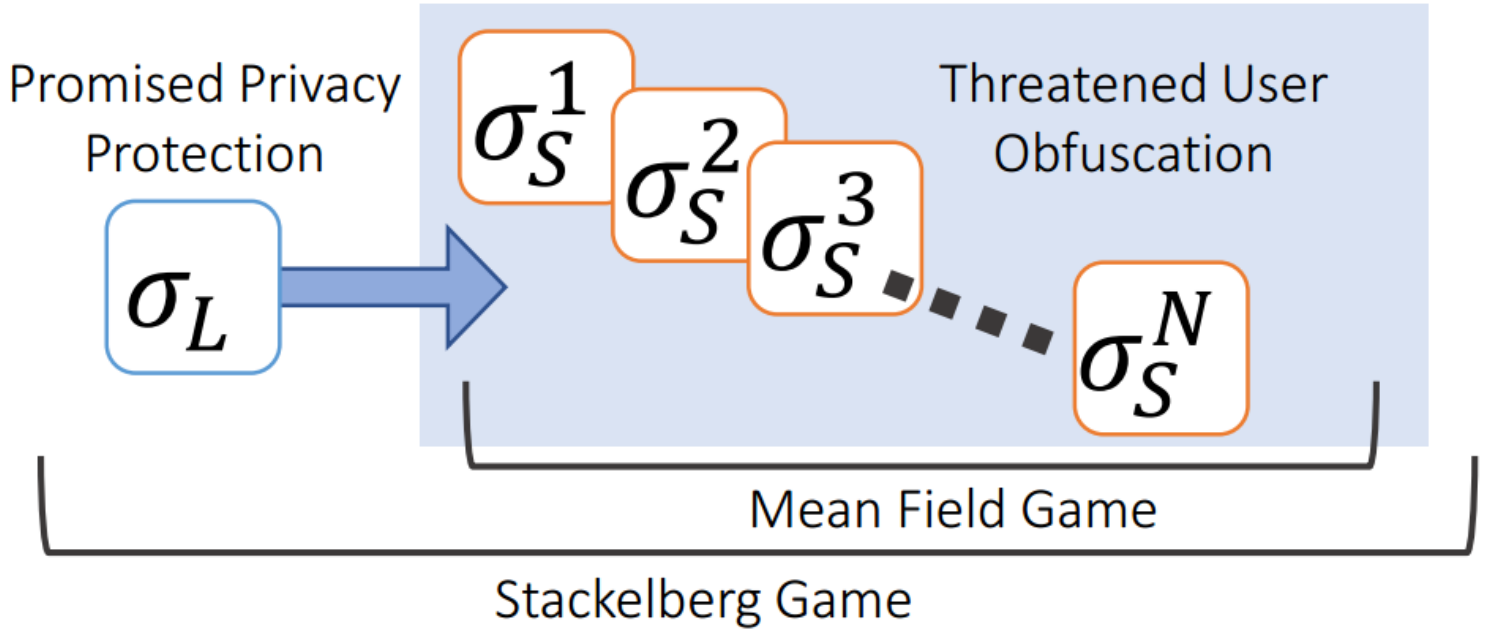
- Obfuscation is a strategic interaction between a machine learner and a set of users.
- Game theory studies strategic interactions between multiple rational agents.
- In equilibrium, each agent reacts optimally to the strategies of the other agents.



σ_L = standard deviation of learner protection

σ_S^i = standard deviation of user $i \in \{1, \ldots, N\}$ obfuscation

N+1 Player Game Theory Model



- Mean field game: each user must respond optimally to the *average behavior* of the other users.
- Stackelberg game: the learner can promise (or not promise) a level of privacy protection, and then the users react.

Empirical Risk Minimization (ERM)

- ERM is a machine learning method in which *L* estimates a predictor **f** by minimizing the empirical risk.
- Let $\{\boldsymbol{z}_i\}_{i\in S} = \{(\mathbf{x}_i, y_i)\}_{i\in S}$ denote the set of actual data vectors and labels.
- Let $\{\tilde{\boldsymbol{z}}_i\}_{i\in S} = \{(\mathbf{x}_i + \mathbf{v}_i + \mathbf{w}_i, y_i)\}_{i\in S}$ denote the data including learner and user noise.
- The perturbed predictor is given by

$$\boldsymbol{f}_d = \operatorname{argmin}_{\boldsymbol{f}\in\boldsymbol{F}}\{\rho R(\boldsymbol{f}) + \frac{1}{N}\sum_{i=1}^{N} l(\tilde{\boldsymbol{z}}_i, \boldsymbol{f})\}.$$

- For comparison, the classifier that minimizes the expected loss is given by

$$\boldsymbol{f}^* = \operatorname{argmin}_{\boldsymbol{f}\in\boldsymbol{F}}\mathbb{E}\{\rho R(\boldsymbol{f}) + l(\mathcal{Z}, \boldsymbol{f})\}.$$

Quantification of Accuracy: $\epsilon_g(\sigma_L, \bar{\sigma}_S^{-i}, \sigma_S^i)$

Definition 1. (ϵ_g-Accuracy) Let ϵ_g be a positive scalar. We say that $\boldsymbol{f}_d$ is ϵ_g-accurate if it satisfies

$$\mathbb{E}\{\rho R(\boldsymbol{f}_d) + l(\mathcal{Z}, \boldsymbol{f}_d)\} \leq \mathbb{E}\{\rho R(\boldsymbol{f}^*) + l(\mathcal{Z}, \boldsymbol{f}^*)\}.$$

Lemma 1. (Accuracy Level) The difference in expected loss between the perturbed classifier and the population-optimal classifier is on the order of

$$\epsilon_g\left(\sigma_L, \bar{\sigma}_S^{-i}, \sigma_S^i\right) \propto \frac{1}{\rho^2 N}\left(\sigma_L^2 + \frac{N-1}{N}\left(\bar{\sigma}_S^{-i}\right)^2 + \frac{1}{N}\left(\sigma_S^i\right)^2\right)$$

Quantification of Accuracy: $\epsilon_g(\sigma_L, \bar{\sigma}_S^{-i}, \sigma_S^i)$

Definition 1. (ϵ_g-Accuracy) Let ϵ_g be a positive scalar. We say that $\boldsymbol{f}_d$ is ϵ_g-accurate if it satisfies

$$\mathbb{E}\{\rho R(\boldsymbol{f}_d) + l(\mathcal{Z}, \boldsymbol{f}_d)\} \leq \mathbb{E}\{\rho R(\boldsymbol{f}^*) + l(\mathcal{Z}, \boldsymbol{f}^*)\}.$$

Lemma 1. (Accuracy Level) The difference in expected loss between the perturbed classifier and the population-optimal classifier is on the order of

$$\epsilon_g\left(\sigma_L, \bar{\sigma}_S^{-i}, \sigma_S^i\right) \propto \frac{1}{\rho^2 N}\left(\sigma_L^2 + \frac{N-1}{N}\left(\bar{\sigma}_S^{-i}\right)^2 + \frac{1}{N}\left(\sigma_S^i\right)^2\right)$$

Quantification of Privacy: $\epsilon_p(\sigma_L, \sigma_S^i)$

Definition 2. (ϵ_p-Privacy) An algorithm $\mathcal{A}(B)$ taking values in a set C provides (ϵ_p, δ)-differential privacy if, for all databases D and D' that differ in at most one entry, and for all $c \subseteq C$,

$$\mathbb{P}\{\mathcal{A}(D) \in c\} \leq \exp\{\epsilon_p\}\, \mathbb{P}\{\mathcal{A}(D') \in c\} + \delta$$

Lemma 2. (Privacy Level) The differential privacy level $\epsilon_p \in (0,1)$ is on the order of

$$\epsilon_g(\sigma_L, \sigma_S^i) \propto \left(\sigma_L^2 + (\sigma_S^i)^2\right)^{-1/2}$$

Quantification of Privacy: $\epsilon_p(\sigma_L, \sigma_S^i)$

Definition 2. (ϵ_p-Privacy) An algorithm $\mathcal{A}(B)$ taking values in a set C provides (ϵ_p, δ)-differential privacy if, for all databases D and D' that differ in at most one entry, and for all $c \subseteq C$,

$$\mathbb{P}\{\mathcal{A}(D) \in c\} \leq \exp\{\epsilon_p\}\, \mathbb{P}\{\mathcal{A}(D') \in c\} + \delta$$

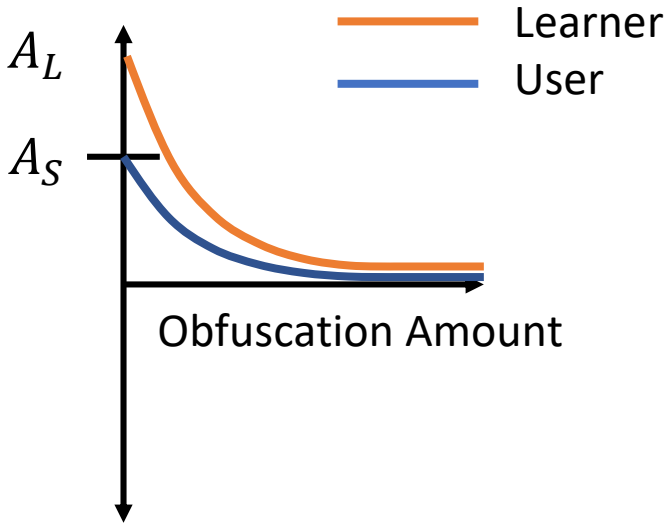
Lemma 2. (Privacy Level) The differential privacy level $\epsilon_p \in (0,1)$ is on the order of

$$\epsilon_g(\sigma_L, \sigma_S^i) \propto \left(\sigma_L^2 + (\sigma_S^i)^2\right)^{-1/2}$$

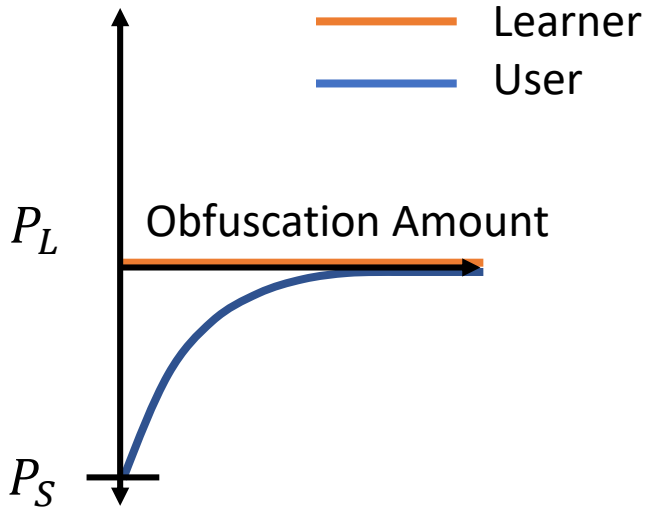
Modeling Utility Functions

$$U_L(\sigma_L, \bar{\sigma}_S) = A_L \exp\{-\epsilon_g(\sigma_L, \bar{\sigma}_S^{-i}, \sigma_S^i)\} - C_L \mathbf{1}_{\{\sigma_L > 0\}}$$

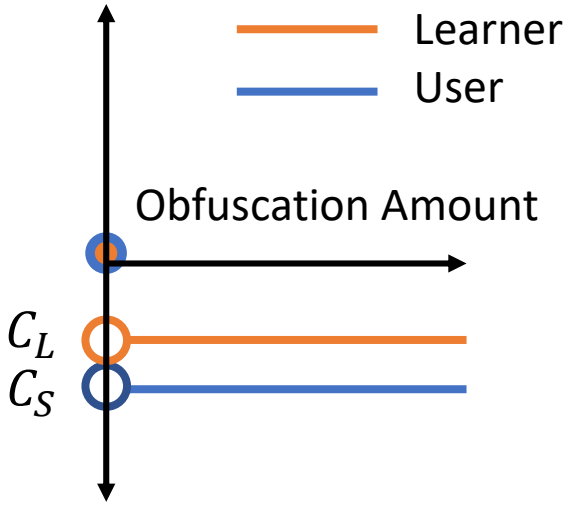
$$U_S^i(\sigma_L, \bar{\sigma}_S^{-i}, \sigma_S^i) = A_S^i \exp\{-\epsilon_g(\sigma_L, \bar{\sigma}_S^{-i}, \sigma_S^i)\} - P_S^i\left(1 - \exp\{-\epsilon_p(\sigma_L, \sigma_S^i)\}\right) - C_L \mathbf{1}_{\{\sigma_L > 0\}}$$



Accuracy Component

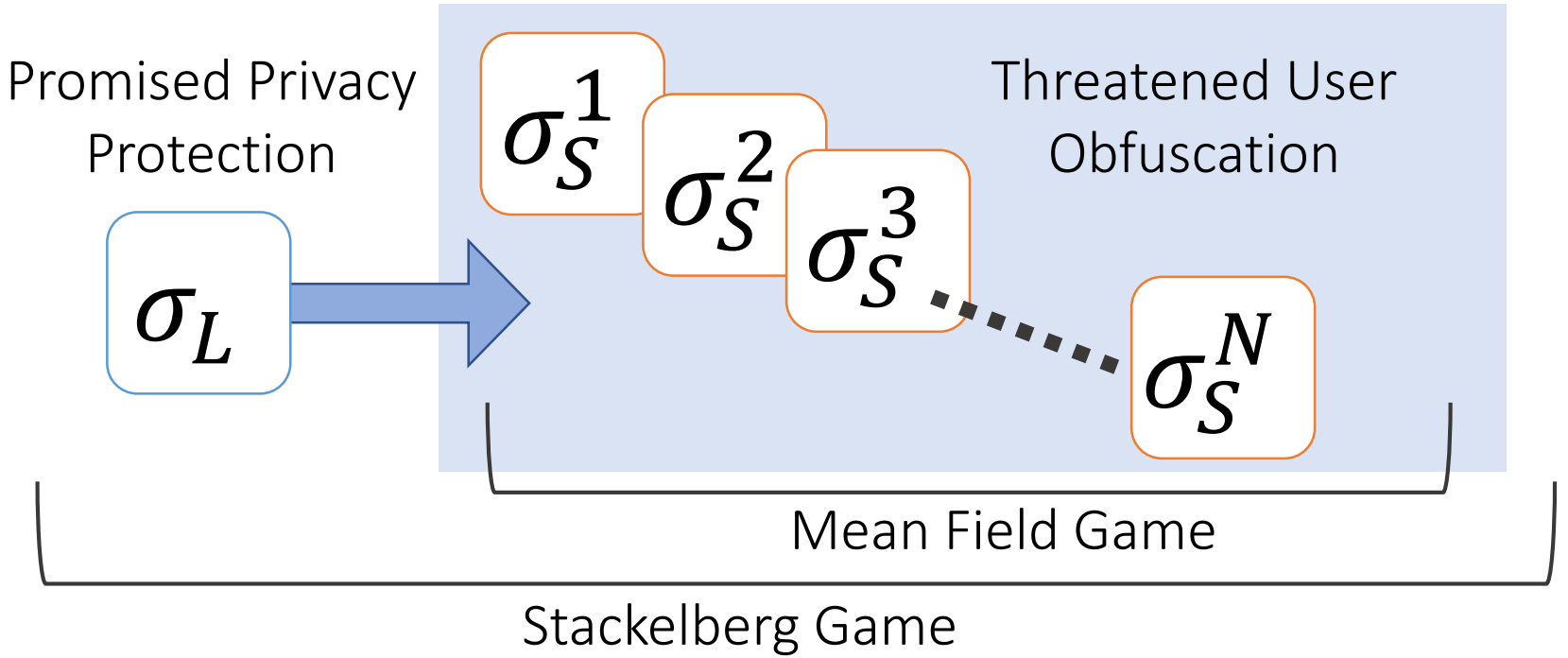


Privacy Component



Obfuscation Cost

Solution Proceeds Backwards in Time



Second-Stage Equilibrium: Mean-Field Game

- Define the best response of a user i to the average perturbations of users $-i$ by

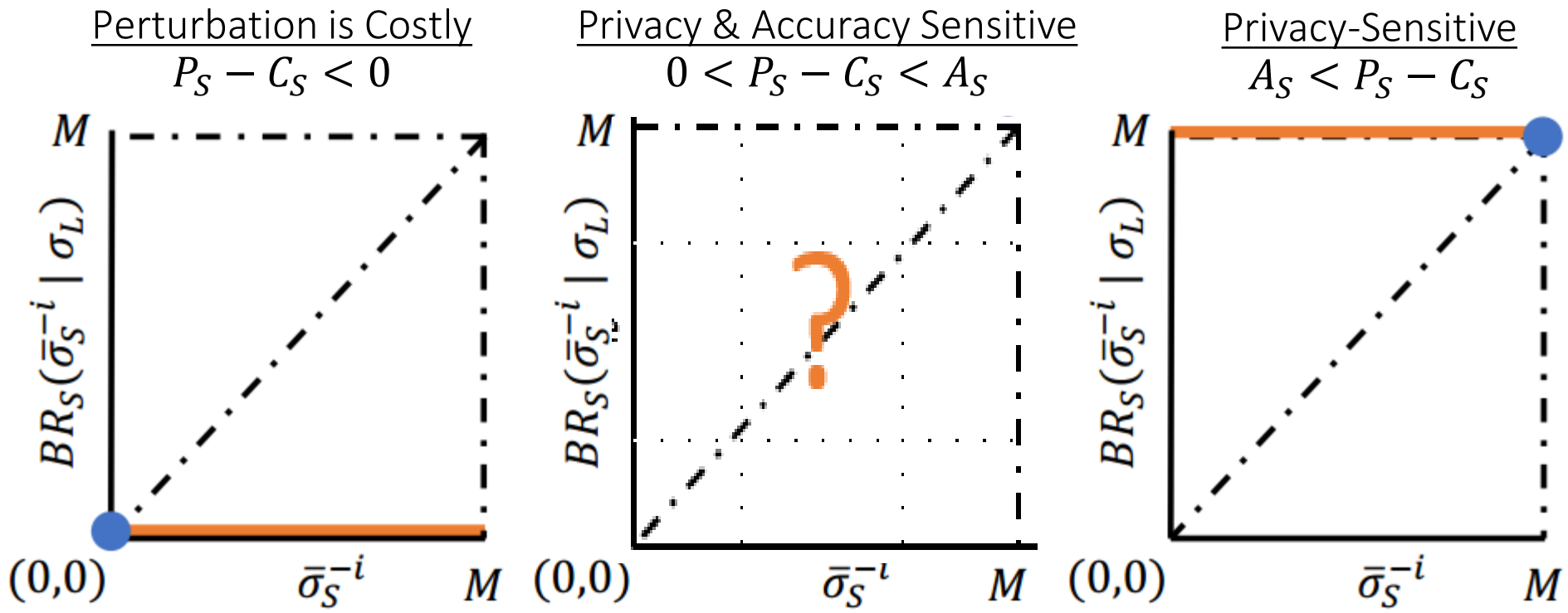
$$BR_S(\bar{\sigma}_S^{-i} \mid \sigma_L) = \arg \max_{\sigma_S^i \in \mathbb{R}_M} U_S^i\left(\sigma_L, \bar{\sigma}_S^{-i}, \sigma_S^i\right).$$

- For simplicity, consider $A_S^i = A_S, P_S^i = P_S, C_S^i = C_S$ for all $i \in 1, \ldots, N$.
- Then the MFG requirement is that

$$\bar{\sigma}_S^* \in BR_S(\bar{\sigma}_S^* \mid \sigma_L).$$

Analysis: Mean Field Game

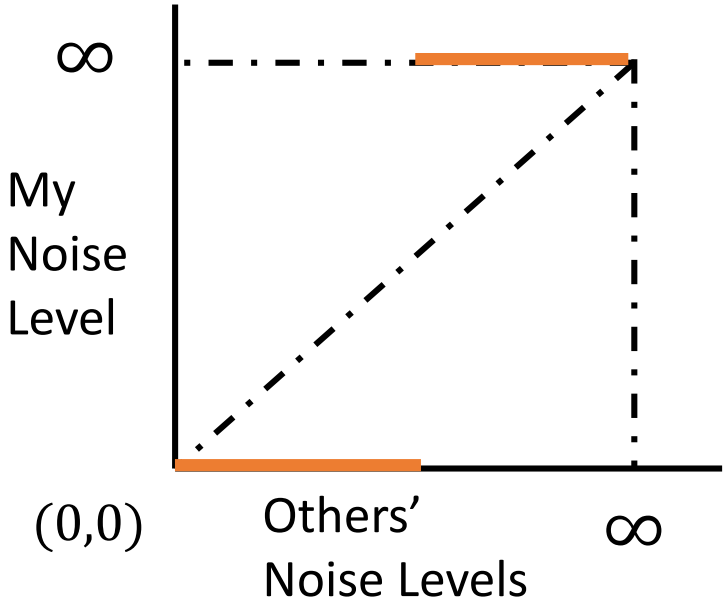
Lemma 3. If the learner does not perturb, then the users perturb either: 1) not at all or 2) as much as possible, depending on how much the field of other users perturb.



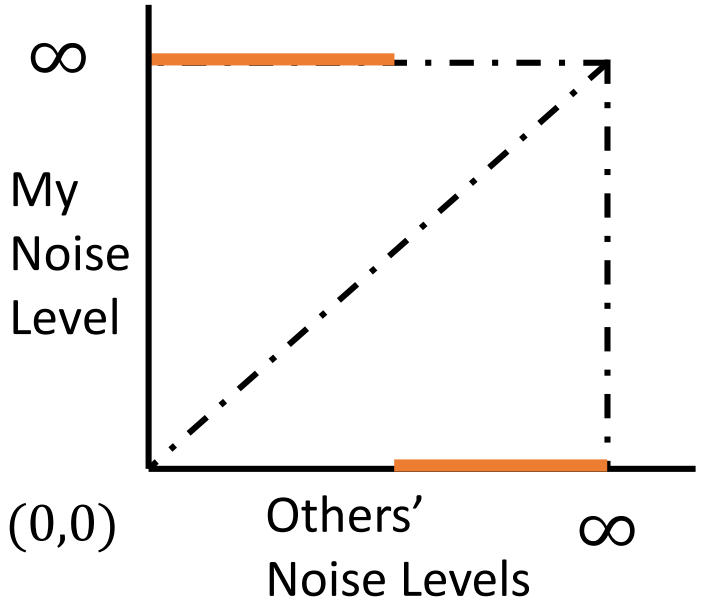
Analysis: Mean Field Game Best Response

- What is the best response in the middle region?

"But Frank jumped off a bridge…"



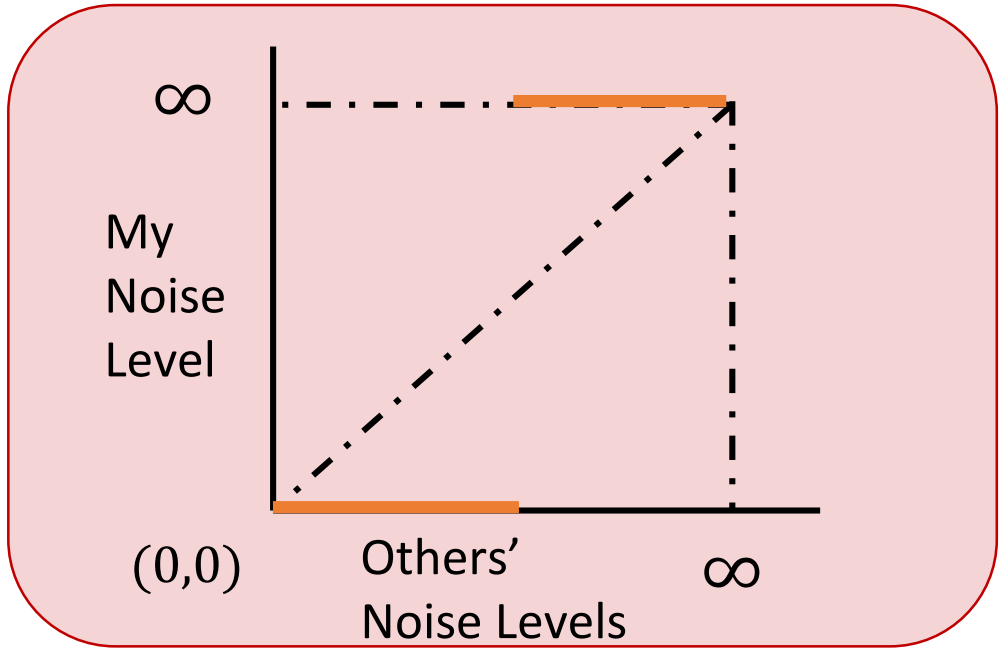
"Don't beat a dead horse"



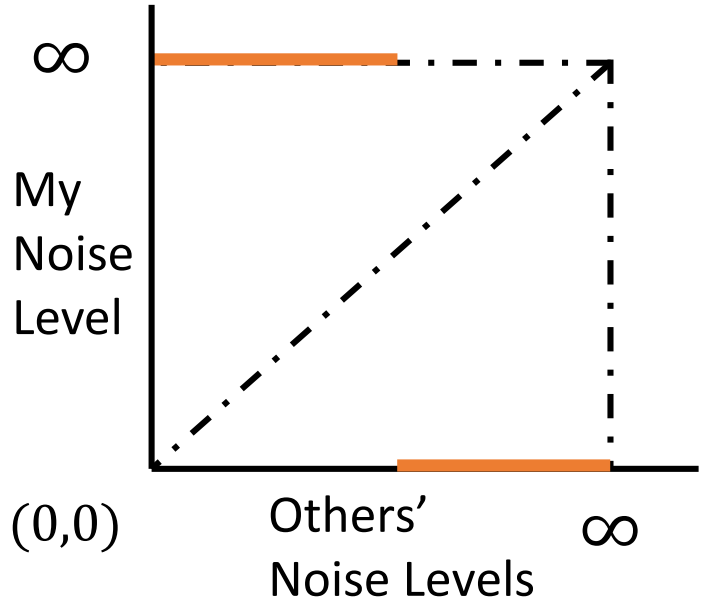
Analysis: Mean Field Game Best Response

- What is the best response in the middle region?

"But Frank jumped off a bridge…"



"Don't beat a dead horse"

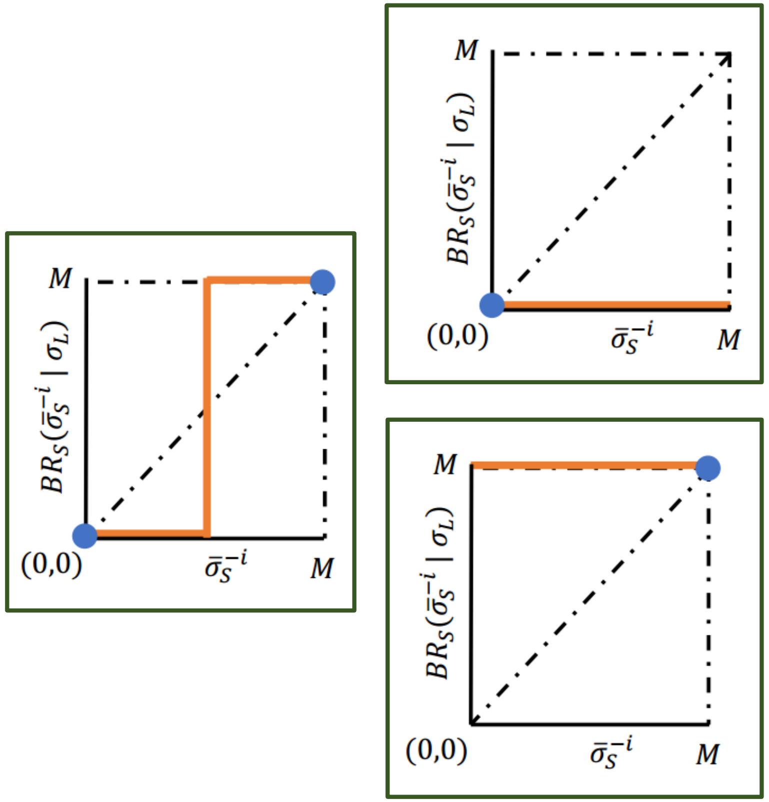


Analysis: Mean Field Game Equilibrium

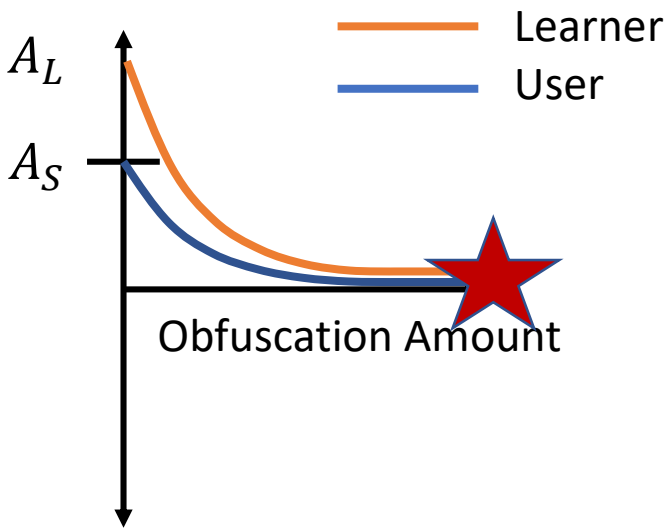
Theorem 1. (MFG Equilibrium) Given a promised privacy protection level σ_L^*, the MFG equilibrium is given by the symmetric strategies $\bar{\sigma}_S^* = \sigma_S^{1*} = \sigma_S^{2*} = \cdots = \sigma_S^{N*}$, where

$$\bar{\sigma}_S = \begin{cases} 0, & \text{if } \boldsymbol{P}(\sigma_L) < \boldsymbol{AC}(\sigma_L, M) < \boldsymbol{AC}(\sigma_L, 0) \\ \{0, M\}, & \text{if } \boldsymbol{AC}(\sigma_L, M) \leq \boldsymbol{P}(\sigma_L) \leq \boldsymbol{AC}(\sigma_L, 0) \\ M, & \text{if } \boldsymbol{AC}(\sigma_L, M) < \boldsymbol{AC}(\sigma_L, 0) < \boldsymbol{P}(\sigma_L) \end{cases}$$

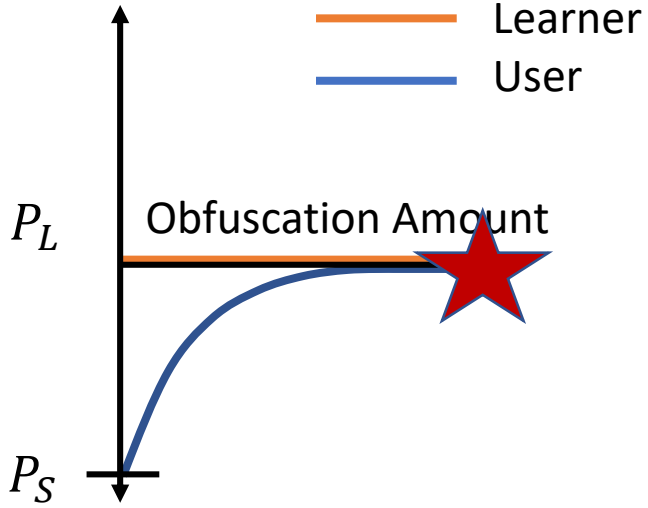
and M denotes a *maximal* level of perturbation.



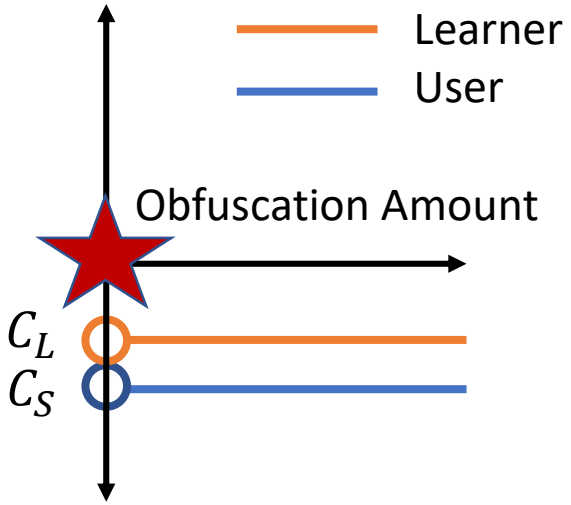
Tracker Receives Zero Utility if All Users Perturb



Accuracy Component



Privacy Component

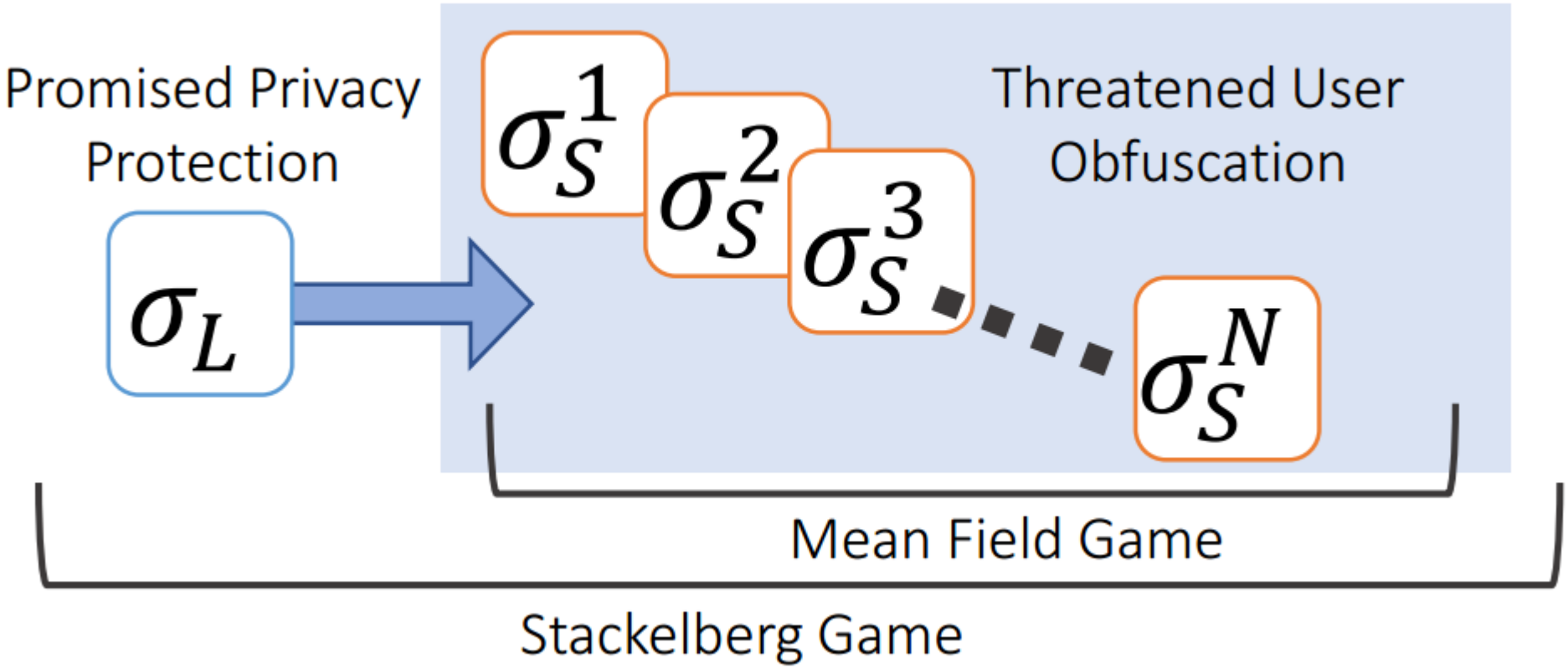


Obfuscation Cost

$$U_L(\sigma_L, \bar{\sigma}_S) = A_L \exp\{-\epsilon_g(\sigma_L, \bar{\sigma}_S^{-i}, \sigma_S^i)\} - C_L \mathbf{1}_{\{\sigma_L > 0\}}$$

$$U_S^i(\sigma_L, \bar{\sigma}_S^{-i}, \sigma_S^i) = A_S^i \exp\{-\epsilon_g(\sigma_L, \bar{\sigma}_S^{-i}, \sigma_S^i)\} - P_S^i\left(1 - \exp\{-\epsilon_p(\sigma_L, \sigma_S^i)\}\right) - C_L \mathbf{1}_{\{\sigma_L > 0\}}$$

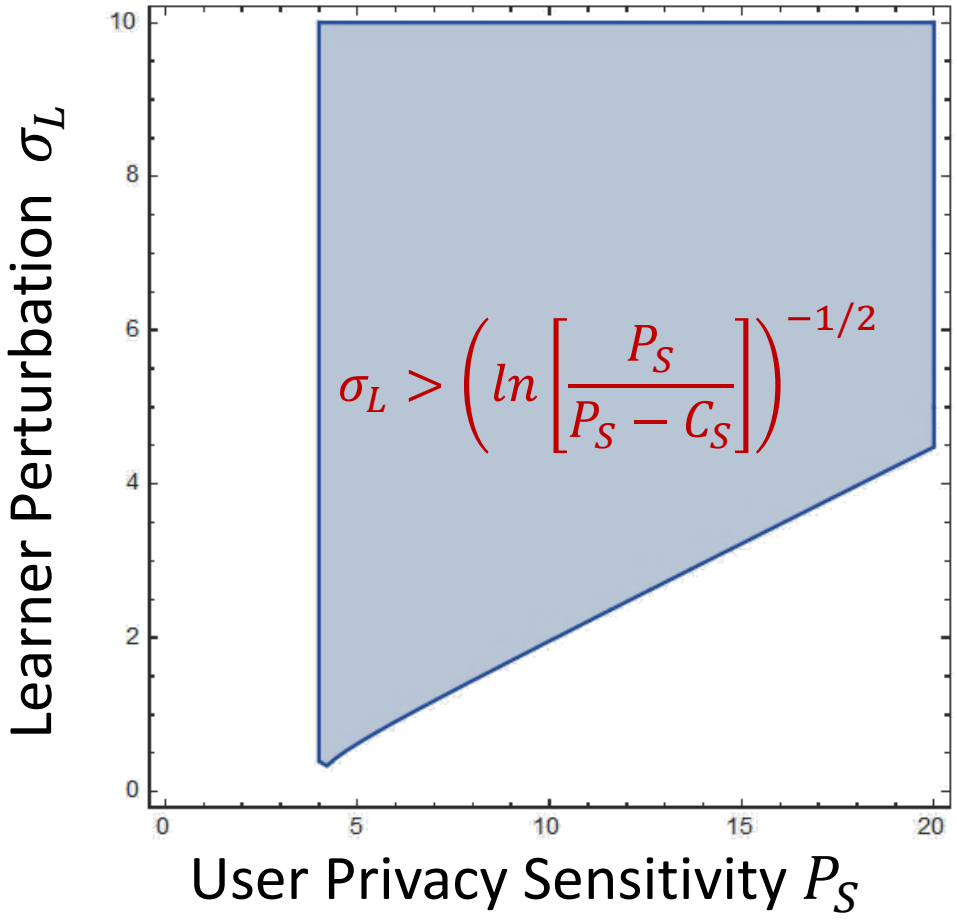
Mechanism: Learner Privacy Commitment



First-Stage Equilibrium: Stackelberg Game

- In differential privacy, a machine learner *promises* a limit on revealed information.
- Users then react to this limit, choosing whether to use the service.
- Therefore, L is a Stackelberg leader, and the users are together a Stackelberg follower who plays $\Gamma(\sigma_L)$, the user strategy *induced* by σ_L.
- Can L induce $\Gamma(\sigma_L) = 0$ by perturbing with a sufficient σ_L? (Otherwise, $\sigma_L^* = 0$.)

Can L induce $\Gamma(\sigma_L) = 0$?



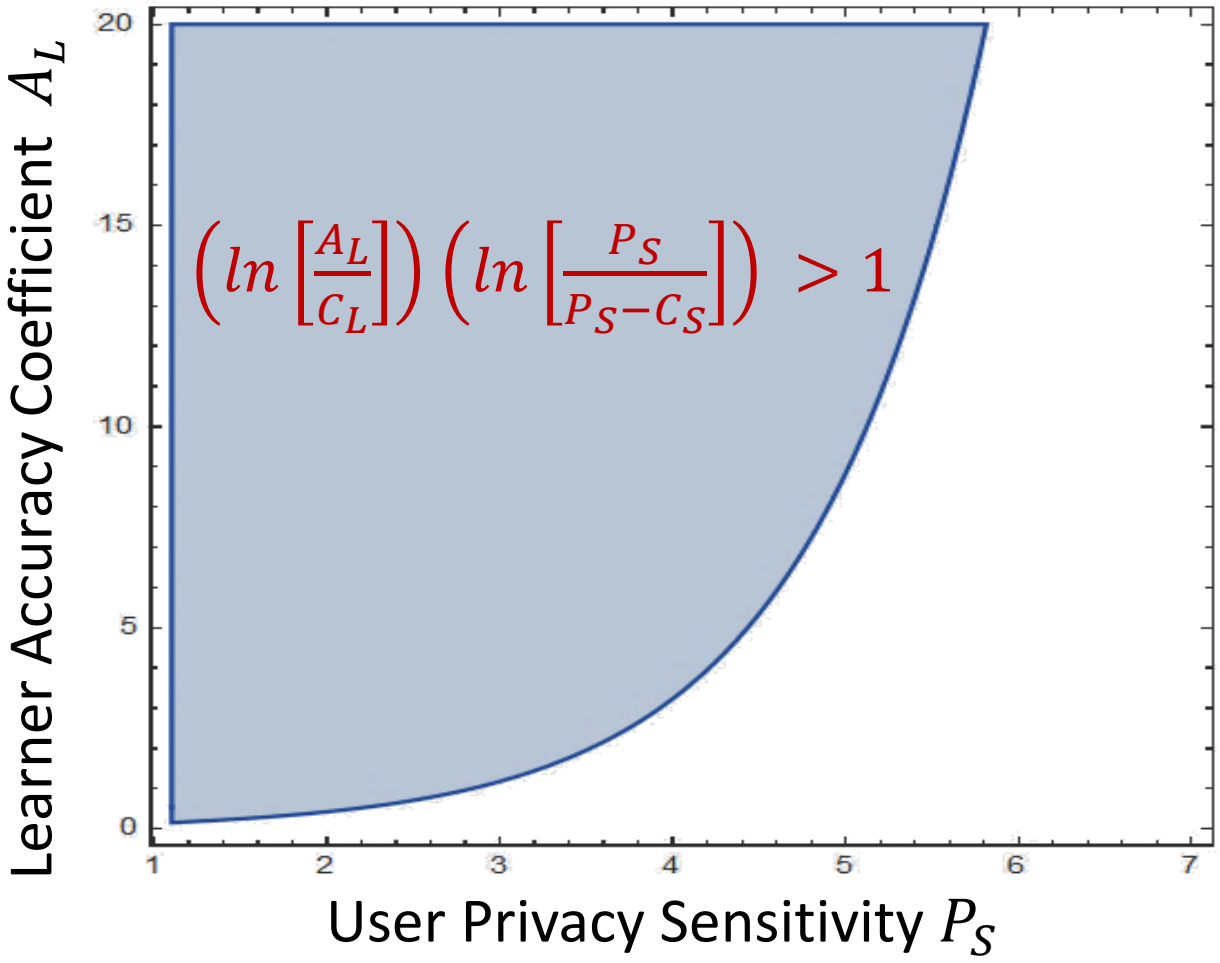
- The optimal user perturbation is

$$\bar{\sigma}_S = \Gamma(\sigma_L) = \begin{cases} M, & \text{if } \sigma_L < \left(ln\left[\frac{P_S}{P_S - C_S}\right]\right)^{-1/2} \\ 0, & if\, \sigma_L > \left(ln\left[\frac{P_S}{P_S - C_S}\right]\right)^{-1/2} \end{cases},$$

  where M is a large perturbation upper-bound.

- Yes, the learner can induce zero perturbation from the learners by promising sufficient protection.
- But is this too costly?

Is Privacy Protection Incentive-Compatible for Tracker?



The optimality equation for L is

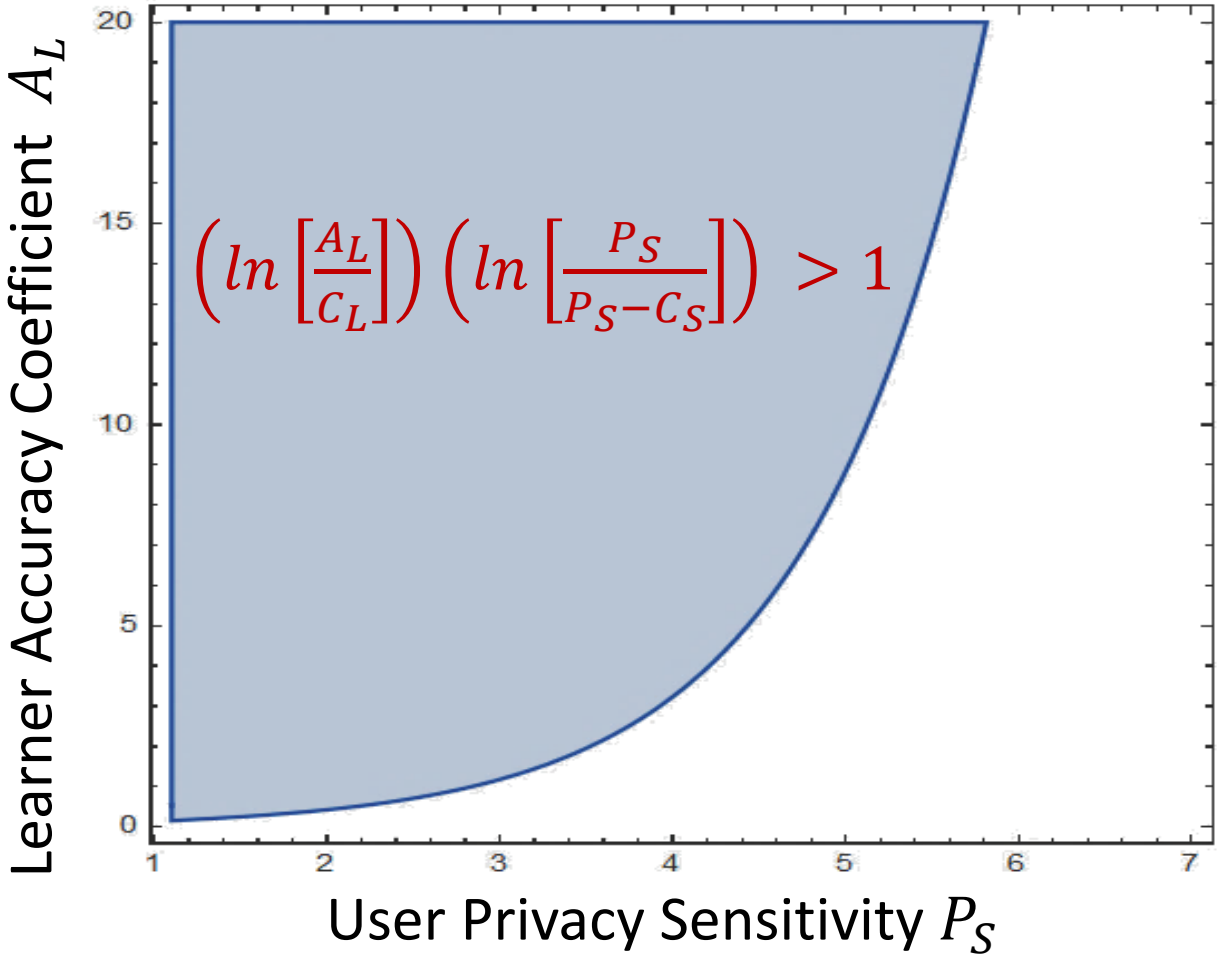
$$\sigma_L^* \in \arg\max_{\sigma_L \in \mathbb{R}_M} U_L(\sigma_L, \Gamma(\sigma_L)).$$

- The optimal σ_L^* is

$$\sigma_L^* = \begin{cases} 0, & \text{if } \frac{1}{\rho^2 N} > \ln\left\{\frac{A_L}{C_L}\right\}\ln\left\{\frac{P_S}{P_S - C_S}\right\} \\ \widehat{\tau}, & \text{if } \frac{1}{\rho^2 N} < \ln\left\{\frac{A_L}{C_L}\right\}\ln\left\{\frac{P_S}{P_S - C_S}\right\} \end{cases},$$

where $\hat{\tau} = \left(ln\left[\frac{P_S}{P_S - C_S}\right]\right)^{-1/2}$

Is Privacy Protection Incentive-Compatible for Tracker?



The optimality equation for L is

$$\sigma_L^* \in \arg\max_{\sigma_L \in \mathbb{R}_M} U_L(\sigma_L, \Gamma(\sigma_L)).$$

- The optimal σ_L^* is

$$\sigma_L^* = \begin{cases} 0, & \text{if } \frac{1}{\rho^2 N} > \ln\left\{\frac{A_L}{C_L}\right\}\ln\left\{\frac{P_S}{P_S - C_S}\right\} \\ \widehat{\tau}, & \text{if } \frac{1}{\rho^2 N} < \ln\left\{\frac{A_L}{C_L}\right\}\ln\left\{\frac{P_S}{P_S - C_S}\right\} \end{cases},$$

where $\hat{\tau} = \left(ln\left[\frac{P_S}{P_S - C_S}\right]\right)^{-1/2}$

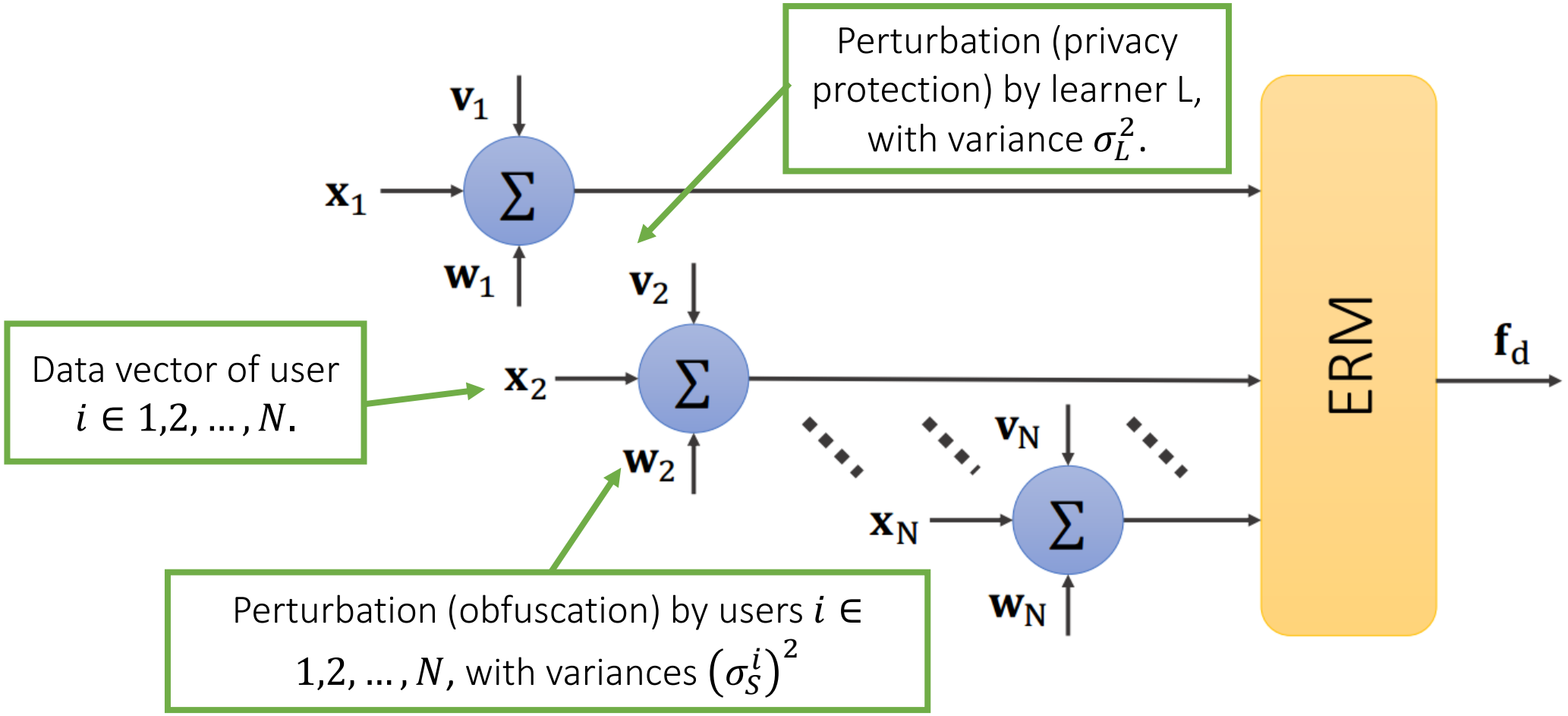
Summary of Equilibrium Results

#	Parameter Regime	$\bar{\sigma}_S^*$	σ_L^*	Significance
Status Quo	$P_S - C_S < A_S$	0	0	Users prefer accuracy to privacy. They do not obfuscate their data.
Market Breakdown	$P_S - C_S > A_S \cap \frac{1}{\rho^2 N} > \ln\left\{\frac{A_L}{C_L}\right\} \ln\left\{\frac{P_S}{P_S - C_S}\right\}$	M	0	Users prefer privacy, so they heavily obfuscate. The learner cannot do anything. The data market collapses.
Controlled Privacy	$P_S - C_S > A_S \cap \frac{1}{\rho^2 N} < \ln\left\{\frac{A_L}{C_L}\right\} \ln\left\{\frac{P_S}{P_S - C_S}\right\}$	0	$\hat{\tau}$	Users threaten to heavily obfuscate, but the learner avoids this by committing to a level of privacy protection.

Future work can consider detection of obfuscation and analyze the impact of other forms of cost functions. We can also estimate cost functions from existing applications.

Backup Slides

Data Flow Model



Bi-Level Game Equilibrium Definition

Definition 3. (Perfect Bayesian Nash Equilibrium). A perfect Bayesian Nash equilibrium (PBNE) of the overall game is $(\sigma_L^*; \sigma_S^{1*}, \sigma_S^{2*}, \dots, \sigma_S^{N*})$ such that $\bar{\sigma}_S^* = \sigma_S^{1*} = \sigma_S^{2*} = \cdots = \sigma_S^{N*}$, and

$$\bar{\sigma}_S^* = \Gamma(\sigma_L^*) = BR_S(\bar{\sigma}_S^* \mid \sigma_L^*),$$
$$\sigma_L^* \in \arg\max_{\sigma_L \in \mathbb{R}_M} U_L(\sigma_L, \Gamma(\sigma_L)).$$