Privacy Games

Yiling Chen, Or Sheffet, Salil Vadhan

Harvard University

Privacy Tools for Sharing Research Data

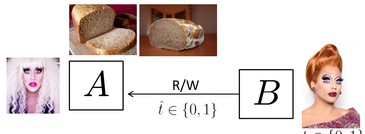
A National Science Foundation Secure and Trustworthy Cyberspace Project

Big Question:

Why and how do privacy concerns effect people's behavior?

Introduction

- Start with a basic game [NOS12]. (Wholewheat or Rye)



- What if B views her type as private?
  - Why?
  - How much does privacy worth?
- Due to many potential factors:
  - Adversary trying to learn her type
  - Future games and agents
  - Beliefs over future events
  - Randomness in mechanisms

- Vast literature [GR11, NST12, X13, CCKMV13, GLRS14] on privacy concerned agents, uses differential privacy:

$$\text{A mechanism } M \text{ is } \epsilon\text{-differentially private if } \forall D \sim D', \forall o$$
$$\Pr[M(D) = o] \le e^{\epsilon} \Pr[M(D') = o]$$

- Differentially private mechanisms have powerful guarantee

$$\mathrm{E}[\mathcal{U}_{\text{future}} | \text{ participate}] \approx_{(e^{-\epsilon}, e^{\epsilon})} \mathrm{E}[\mathcal{U}_{\text{future}} | \neg\text{participate}]$$

- In our case (n=1), B can play Randomized Response $\Pr[\hat{t} = t] = \frac{1+\epsilon/2}{2}$

$$\frac{\Pr[\hat{t} = 0 \mid t = 0]}{\Pr[\hat{t} = 0 \mid t = 1]} = \frac{1+\epsilon/2}{1-\epsilon/2} \le e^{\epsilon}$$

- Should B play Randomized Response, then her future utility changes by $(1 \pm \epsilon)$-factor

Related Work

- [GR11, NST12, X13, CCKMV13, GLRS14]: Avoid modeling privacy concerns using "hardwired" privacy loss (altering the agent's utility)

$$\mathcal{U} = \mathcal{U}_{\text{mechanism}} - \mathcal{U}_{\text{privacy}}$$
$$\left(\text{Upper-Bound}(\mathcal{U}_{\text{privacy}}) \propto \max_{D \sim D', o} \left\{ \ln\left(\frac{\Pr[M(D) = o]}{\Pr[M(D) = o]}\right) \right\}\right)$$

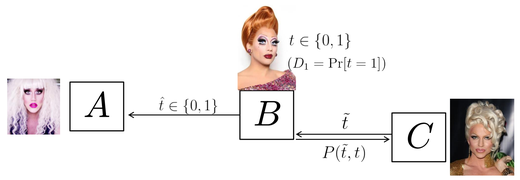
- Using an agent's intrinsic value of privacy v_i

- Make claims about privacy and truthfulness in
  - Voting
  - Facility location
  - Conducting a survey
- Give upper and lower bounds
  - When agents' valuations of privacy are correlated with type, then, worst-case settings cause the mechanism to have a non-useful output.

- Such a B agent does trade-off privacy for the sandwich.
  - Maximize utility by playing Randomized Response.
- We do not model B's privacy loss, we model her privacy concerns.

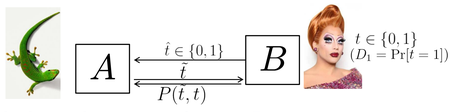
The Setting

We model B's privacy concerns in the form of a payment to an adversary's accusation.



Comments:

1. Bayes-Nash Equilibrium (not other solution concepts)
2. B and C are adversaries:
   1. $\arg\max_{\tilde{t}} P(\tilde{t}, t) = t$
   2. Zero sum game
3. We always assume $D_0 \ge D_1$
4. Let's ignore C...



- We *only* change the payments function P.

Types of Payment functions

1. Proper scoring rules
   - A reports her belief that B is of type 1 $(\tilde{t} \in [0,1])$
2. 0/1 payments
   - A makes an accusation as to B's type $(\tilde{t} \in \{0,1\})$
   - Gets 1 if she is correct, 0 if she is wrong
3. General payments & opting out
   - A makes an accusation as to B's type, or opts out $(\tilde{t} \in \{0,1,\perp\})$
   - Get 0 if she opts out, gets paid if she's right, pays if she's wrong

- Is there a connection to Randomized Response in each payment?

Bayes Nash Equilibrium 101

- Bayesian Game:
  - Type spaces for A and B: $\Gamma_A \times \Gamma_B$
  - A prior distribution Π over $\Gamma_A \times \Gamma_B$
  - Set of actions S_A, S_B
  - Utility functions: $u_A, u_B : \Gamma_A \times \Gamma_B \times S_A \times S_B \to R$
- Strategies:
  - $\sigma_A : \Gamma_A \to \triangle(S_A)$, $\sigma_B : \Gamma_B \to \triangle(S_B)$
- (σ_A^*, σ_B^*) is BNE if

$\forall t \in \Gamma_A$, $\mathrm{E}[u_A(T_A, T_B, \sigma_A^*, \sigma_B^*) \mid T_A = t] \ge \mathrm{E}[u_A(T_A, T_B, \sigma_A', \sigma_B^*) \mid T_A = t]$

$\forall t \in \Gamma_B$, $\mathrm{E}[u_B(T_A, T_B, \sigma_A^*, \sigma_B^*) \mid T_B = t] \ge \mathrm{E}[u_B(T_A, T_B, \sigma_A^*, \sigma_B') \mid T_B = t]$

Scoring Rules 101

- A well-known way to convert beliefs into payments.
- Setting:
  - A random variable $X \in \{0,1\}$ that we are trying to predict.
  - An expert knows the answer. But she's not necessarily truthful...
- The expert will report $x \in [0,1]$. We suggest payments (f_0, f_1) s.t.
  - If X=0, we will pay $f_0(x)$
  - If X=1, we will pay $f_1(x)$
- For proper scoring rules: if $\mathrm{E}[X] = \Pr[X = 1] = \mu$

$$\mu = \arg\max_x \{\mathrm{E}_{b \leftarrow X}[f_b(x)]\}$$

- We say scoring rule is *symmetric* if $f_0(1\text{-}x) = f_1(x)$

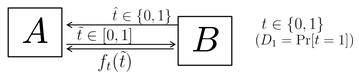
Notation

- We always denote B's strategy as

$$p = \Pr[\hat{t} = 0 \mid t = 0]$$
$$q = \Pr[\hat{t} = 1 \mid t = 1]$$

- We denote A's strategy as x when she sees the 0-signal, and y when she sees the 1-signal.

1. Proper Scoring Rules Payments



- Given that B plays (p,q), then A's best response is to report the Bayesian posterior belief

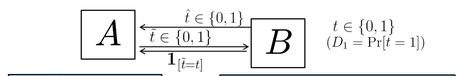
$$x^*(p,q) = \Pr[t = 1 \mid \hat{t} = 0]$$
$$y^*(p,q) = \Pr[t = 1 \mid \hat{t} = 1]$$

Thm: Assuming (f_0, f_1) is symmetric (and ρ_0=ρ_1 not too large) then at the BNE of the game B plays (p^*, q^*) s.t.

$$1 - x^*(p^*, q^*) = \Pr[t = 0 \mid \hat{t} = 0] = \frac{1}{2} + \epsilon(p^*, q^*)$$
$$y^*(p^*, q^*) = \Pr[t = 1 \mid \hat{t} = 1] = \frac{1}{2} + \epsilon(p^*, q^*)$$

- Not randomized response per-se! (Assuming $D_0 \neq D_1$, $p^* \neq q^*$)

2. 0/1-Payments



$u_A(\tilde{t}) = \mathbf{1}_{[\tilde{t}=t]}$ $\quad$ $u_B(\hat{t}) = \rho_t \mathbf{1}_{[\hat{t}=t]} - \mathbf{1}_{[\hat{t}=t]}$

- When the coupon valuation ρ_t is known and fixed:

Thm: In a BNE of the full game –
- When $\rho_0, \rho_1 \ge 1$
  Both types play deterministically (with signal=type)
- When $\rho_0, \rho_1 < 1$
  In *any* BNE of the game – $\exists$ type of B plays deterministically

- When the coupon valuation ρ_t is sampled from a continuous distribution A's strategy leads to a threshold phenomena in B's behavior

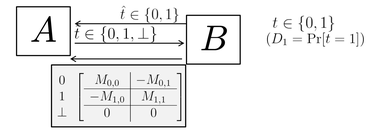
$$x = \Pr[\hat{t} = 0 \mid \hat{t} = 0]$$
$$y = \Pr[\hat{t} = 1 \mid \hat{t} = 1]$$

  - A's strategy translates to some $T (= x+y-1)$ s.t.
    - Type-t B agent with $\rho > T$ always plays truthfully $(\hat{t} = t)$
    - Type-t B agent with $\rho < T$ always lies $(\hat{t} = 1 - t)$
  - Note: As A doesn't know B's valuation – A believes B is playing *randomized response*.

- Under the assumption $D_0 > D_1$, in the BNE it holds that x=1 (A always guesses B is of type 0 give the 0-signal) while y is such that

$$\Pr[\rho < y^*] = D_1$$

3. General Payments with Opting Out



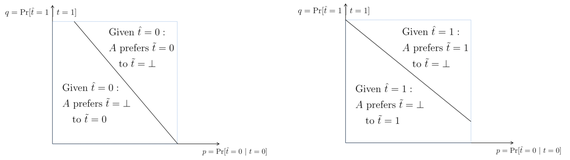
- Parameters are set so that without a signal, A always prefers to opt out

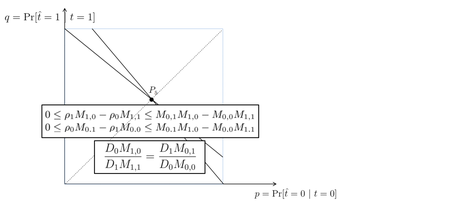
$$(D_0 M_{0,0} - D_1 M_{0,1} < 0, \quad D_1 M_{1,1} - D_0 M_{1,0} < 0)$$

- Thm: The BNE of the game are characterized as follows

Case No.	Condition	A's Strategy (always: $x_1 = y_0 = 0$)	B's strategy
1	$\rho_0 \ge M_{0,0} + M_{1,0}$ and $\rho_1 \ge M_{0,1} + M_{1,1}$	$(x_0, y_1) = (1, 1)$	$(1, 1)$
2	$\rho_0 \le M_{0,0}$ and $\frac{\rho_0}{\rho_1} \le \frac{M_{0,0}}{M_{0,1}}$	$(x_0, y_1) = (\frac{\rho_0}{M_{0,0}}, 0)$	$P_1 = (0, 1)$
3	$0 \le \rho_0 - M_{0,0} \le M_{1,0}$ $\rho_1 M_{1,0} - \rho_0 M_{1,1} \ge M_{0,1} M_{1,0} - M_{0,0} M_{1,1}$	$(x_0, y_1) = (1, \frac{\rho_0 - M_{0,0}}{M_{1,0}})$	P_2
4	$\rho_1 \le M_{1,1}$ and $\frac{\rho_0}{\rho_1} \ge \frac{M_{1,0}}{M_{1,1}}$	$(x_0, y_1) = (0, \frac{\rho_1}{M_{1,1}})$	$P_3 = (1, 0)$
5	$0 \le \rho_1 - M_{1,1} \le M_{0,1}$ $\rho_0 M_{0,1} - \rho_1 M_{0,0} \ge M_{0,1} M_{1,0} - M_{0,0} M_{1,1}$	$(x_0, y_1) = (\frac{\rho_1 - M_{1,1}}{M_{0,1}}, 1)$	P_4
6	$0 \le \rho_1 M_{1,0} - \rho_0 M_{1,1} \le M_{0,1} M_{1,0} - M_{0,0} M_{1,1}$ $0 \le \rho_0 M_{0,1} - \rho_1 M_{0,0} \le M_{0,1} M_{1,0} - M_{0,0} M_{1,1}$	See below	P_5

Table 1. The various conditions under which we characterize the BNEs of the Game. We use the notation $P_2 = (1 - \frac{D_0 M_{0,0}}{D_1 M_{0,1}}, 1)$, $P_4 = (1, 1 - \frac{D_0 M_{0,0}}{D_1 M_{0,1}})$, and $P_5 = \left(\frac{D_0 D_1 M_{0,1} M_{1,0} - D_1^2 M_{0,1} M_{1,1}}{D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1}}, \frac{D_0 D_1 M_{0,1} M_{1,0} - D_0^2 M_{0,0} M_{1,1}}{D_0 D_1 M_{0,1} M_{1,0} - D_0 D_1 M_{0,0} M_{1,1}}\right)$. The point P_5 lines at the intersection between two specific lines, and points P_2 and P_4 are the intersection points of each of those lines with the (q = 1)-line and (p = 1)-line resp. In case 6, the strategy of A is given by $(x_0, y_1) = \frac{1}{M_{1,0} M_{0,1} - M_{0,0} M_{1,1}} (-M_{1,1}\rho_0 + M_{1,0}\rho_1, M_{0,1}\rho_0 - M_{0,0}\rho_1)$.





- But even when B plays randomized response, she still doesn't trade-off the value of the coupon to the payments to the adversary
- Instead, mall changes to ρ_0, ρ_1 may change the BNE strategy sharply.

Future Directions

- Even when B's behavior ¼ Randomized Response: we don't get the elegance of hard-wired privacy loss way-of-thinking.
  - Is differential-privacy "wrong?" (i.e. not suited for how people think of privacy?)
    - Is it because worst-case guarantees are the "wrong" approach?
- Analyzing other scenarios
  - The standard model with a trusted data curator
- Is privacy related to "risk aversion?"

Acknowledgments

We would like to thanks Kobbi Nissim for many helpful discussions.