CIS 800/002 The Algorithmic Foundations of Data Privacy November {22, 29}, 2011

Lecture 20+21

Lecturer: Aaron Roth *Scribe: Aaron Roth*

Differential Privacy and Mechanism Design

In this lecture, we'll give a brief introduction to mechanism design, and see how differentially private mechanisms can be used as a building block in building truthful mechanisms.

We'll start with the very simple example of a single item auction. Suppose an auctioneer has a single item for sale. Bidders $i \in [n]$ each have a private valuation for the item v_i. If a bidder i wins the item, but must pay p_i dollars, then his total utility is $u_i = v_i - p_i$. Bidders are rational, and will act to maximize their utility. As mechanism designers, we get to design the auction rule to achieve some goal. Suppose we want to maximize social welfare: to allocate the item to the person who wants it the most. Consider the following simple auction rule:

FirstPrice:

1. Each bidder i submits a bid b_i.
2. $i^* = \arg\max_{i \in [n]} b_i$
3. Allocate the item to bidder i^* in exchange for payment $p_{i^*} = b_{i^*}$

How should rational bidders behave? Its not clear...

Example 1 *Suppose $n = 2$ and $v_1 = 10$, $v_2 = 4$. Truthful bidding is not optimal: it results in 1 winning the item and getting utility $u_1 = 10 - 10 = 0$. If 2 bids 4, then 1 should bid $4 + \epsilon$... But 2 might also shade his bid.. What if they don't know each others value?*

As we see, it is not at all clear how agents will bid. Without knowing how the agents bid, we can't say anything about the performance of the mechanism! It allocates the item to the person who reported the highest value, but this might not be the person with the highest value...

SecondPrice:

1. Each bidder i submits a bid b_i.
2. $i^* = \arg\max_{i \in [n]} b_i$
3. $i_2^* = \arg\max_{i \in [n] \setminus \{i^*\}} b_i$
4. Allocate the item to bidder i^* in exchange for payment $p_{i^*} = b_{i_2^*}$

Unlike FirstPrice, it is always in each persons best interest to truthfully report $b_i = v_i$ in secondprice. Write b_{-i} to denote the set of all bids b_j for $j \neq i$.

Claim 1 *For all i, $b_{-i} \in \mathbf{R}^{n-1}$, $b_i' \in \mathbf{R}$:*

$$u_i(\mathit{SecondPrice}(b_i, b_{-i})) \geq u_i(\mathit{SecondPrice}(b_i', b_{-i}))$$

Proof By case analysis ■

We have just shown that SecondPrice is *truthful* or *incentive compatible.* It is always in each person's best interest to bid their true value. Therefore, because it allocates the item to the individual with the highest reported valuation, if agents are rational, it also succeeds in allocating the item with the highest true valuation, which was our objective.

Of course, we needed to introduce payments to make the above mechanism truthful. What do we do if, as in many situations (elections, public works, organ transplants, ...) payments are not allowed?

Lets now make some general definitions.

Definition 2 (The Environment) *An environment is determined by:*

1. *A set N of n players.*
2. *A set of types T_i for each player $i \in N$. (e.g. values v_i in the auction setting)*
3. *A finite set S of social alternatives (e.g. allocations in the auction setting)*
4. *A set of reactions R_i for each player $i \in N$.*
5. *A utility function $u_i : T_i \times S \times R_i \to [0, 1]$ for each agent i.*

We write T_{-i} for $\prod_{j \neq i} T_i$ and $t_{-i} \in T_{-i}$. Write $r_i(t, s, \hat{R}_i) \in \arg\max_{r \in \hat{R}_i} u_i(t, s, r)$ to denote i's optimal reaction to type t and alternative s among choices $\hat{R}_i \subseteq R_i$.

A direct revelation mechanism $\mathcal{M}$ defines a game which is played as follows:

1. Each player i reports a type $t'_i \in T_i$.
2. The mechanism chooses an alternative $s \in S$ and a subset of reactions for each player $\hat{R}_i \subseteq R_i$.
3. Each player chooses a reaction $r_i \in \hat{R}_i$ and experiences utility $u_i(t_i, s, r_i)$.

Agents play so as to maximize their own utility. Note that since there is no further interaction after the 3rd step, rational agents will pick $r_i = r_i(t_i, s, \hat{R}_i)$, and so we can ignore this as a strategic step. Let $\mathcal{R}_i = 2^{R_i}$ and let $\mathcal{R} = \prod_{i=1}^{n} \mathcal{R}_i$. Then a mechanism is a randomized mapping $\mathcal{M} : T \to S \times \mathcal{R}$. We denote agents expected utilities for reporting a type t'_i when all other agents report type t'_{-i} as:

$$u_i(t_i, \mathcal{M}(t'_i, t'_{-i})) = \mathrm{E}_{s, \hat{R}_i \sim \mathcal{M}(t'_i, t'_{-i})}[u_i(t_i, s, r_i(t_i, s, \hat{R}_i))]$$

We want to design mechanisms that incentivize truthful reporting.

Definition 3 *A mechanism $\mathcal{M}$ is* dominant strategy truthful *if for all $i \in N$, $t_i \in T_i$ and $t'_{-i} \in T_{-i}$:*

$$u_i(t_i, \mathcal{M}(t_i, t'_{-i})) \geq u_i(t_i, \mathcal{M}(t'_i, t'_{-i}))$$

It is strictly dominant strategy truthful if for each t_i, there exists a t'_{-i} that makes the inequality strict.

i.e. given a truthful mechanism, no rational agent ever has any incentive to lie about his type. We can also define an approximate version of truthfulness:

Definition 4 *A mechanism $\mathcal{M}$ is ϵ-approximately* dominant strategy truthful *if for all $i \in N$, $t_i \in T_i$ and $t'_{-i} \in T_{-i}$:*

$$u_i(t_i, \mathcal{M}(t_i, t'_{-i})) \geq u_i(t_i, \mathcal{M}(t'_i, t'_{-i})) - \epsilon$$

For an approximately truthful mechanism, agents may have incentive to lie, but the incentive is only very small.

Note that it is very easy to design a mechanism that is truthful: we can merely ignore the reported types, and pick a random alternative $s \in S$. But as mechanism designers, we will also have some objective that we wish to optimize – e.g. in the auction example, we wished to allocate the item to the person who desired it the most. A common objective is to maximize the social welfare $F : T \times S \times R \to \mathbf{R}$ defined to be:

$$F(t, s, r) = \frac{1}{n} \sum_{i=1}^{n} u_i(t_i, s, r_i)$$

Definition 5 *A mechanism $\mathcal{M}$ α-approximates an objective F if for all t:*

$$\mathrm{E}_{s,\hat{R}\sim\mathcal{M}}[F(t,s,r(t,s,\hat{R}))] \geq \max_{t,s,r} F(t,s,r) - \alpha$$

Ok! Now we can set about designing mechanisms. We will focus on the social welfare objective, but nothing here will be specific to that. First lets consider *unrestricted* mechanisms that always output $\hat{R}_i = R_i$ for each player. As differential privacy aficionados, our first attempt at constructing a useful mechanism might be to select an alternative $s \in S$ from the exponential mechanism. Note that the social welfare objective is $1/n$-sensitive, and so the exponential mechanism $\mathcal{M}_\epsilon(t)$ selects each $s \in S$ with probability proportional to: $\exp(\epsilon n F(t,s,r(t,s)/2))$. This mechanism is ϵ-differentially private, but what can we say about its incentive properties?

Call a mechanism *non-imposing* if it always outputs $\hat{R}_i = R_i$. Such a mechanism is a mapping $\mathcal{M} : T \to S$.

Theorem 6 *For any $\epsilon \leq 1$, a non-imposing mechanism that is ϵ-differentially private is 2ϵ approximately truthful.*

Proof Fix any ϵ-differentially private mechanism $\mathcal{M} : T \to S$. For any i, t_i, t'_{-i} we have:

$$\begin{aligned}
u_i(t_i, \mathcal{M}(t_i, t'_{-i})) &= \mathrm{E}_{s\sim\mathcal{M}(t_i,t'_{-i})}[u_i(t_i,s)] \\
&= \sum_{s\in S} \Pr[\mathcal{M}(t_i,t'_{-i}) = s] u_i(t_i,s) \\
&\geq \sum_{s\in S} \exp(-\epsilon) Pr[\mathcal{M}(t'_i,t'_{-i}) = s] u_i(t_i,s) \\
&= \exp(-\epsilon) u_i(t'_i, \mathcal{M}(t_i,t'_{-i})) \\
&\geq u_i(t'_i, \mathcal{M}(t_i,t'_{-i})) - 2\epsilon
\end{aligned}$$

■

Of course we also know that the exponential mechanism is pretty accurate as well. Translating our utility theorem for the exponential mechanism:

Theorem 7 *$\mathcal{M}_\epsilon$ α-approximates the social welfare F for:*

$$\alpha = O\left(\frac{1}{\epsilon n}(\log|S|)\right)$$

So this is pretty good already! We have a generic method of designing approximately truthful mechanisms with good utility guarantees (if n is large enough) for arbitrary type spaces and objective functions! This is something that does not exist in the mechanism design literature for exact truthfulness. On the other hand, we might want to do better..

Recall, why did we want truthfulness in the first place? It was because we wanted to guarantee that optimizing over reported types was as good as optimizing over true types. But if the mechanism is only approximately truthful, how do we know that the reported types are the true types? Why shouldn't agents take their incentive to lie, however small? If we are to be satisfied with approximate truthfulness, we need further to assume that agents will not lie if given only a small incentive. Perhaps mis-reporting has some small external cost (guilt?), or perhaps the beneficial lie is difficult to find. In such cases, perhaps we can stop here. But in other cases, we would like an exactly truthful mechanism.

The idea will be to randomize between a differentially private mechanism with good social welfare properties (i.e. the exponential mechanism), and a strictly truthful mechanism which punishes false reporting but which has poor social welfare properties. If we mix correctly, then we will get an exactly truthful mechanism with reasonable social welfare guarantees.

Here is one such punishing mechanism which is simple, but not necessarily the best for a given problem:

Definition 8 *The commitment mechanism $M^P(t')$ selects $s \in S$ uniformly at random and sets $\hat{R}_i = \{r_i(t'_i, s, R_i)\}$. i.e. it picks a random outcome, and then forces everyone to react as if their reported type is their true type.*

Define the *gap* of an environment as:

$$\gamma = \min_{i, t_i \neq t'_i, t_{-i}} \max_{s \in S} \left(u_i(t_i, s, r_i(t_i, s, R_i)) - u_i(t_i, s, r_i(t'_i, s, R_i))\right)$$

i.e. γ is a lower bound over players and types of the worst-case cost (over s) of mis-reporting. Note that for each player, this worst-case is realized with probability at least $1/|S|$. Therefore we have the following simple observation:

Lemma 9 *For all i, t_i, t'_i, t_{-i}:*

$$u_i(t_i, \mathcal{M}^P(t_i, t_{-i})) \geq u_i(t_i, \mathcal{M}^P(t'_i, t_{-i})) + \frac{\gamma}{|S|}$$

Note that the commitment mechanism is strictly truthful: every individual has at least a $\frac{\gamma}{|S|}$ incentive not to lie.

This suggests a way to achieve an exactly truthful mechanism that also gets social welfare guarantees:

Definition 10 *The punishing exponential mechanism $\mathcal{M}^P_\epsilon(t)$ defined with parameter $0 \leq q \leq 1$ is:*

1. *With probability $(1-q)$ return $\mathcal{M}_\epsilon(t)$*
2. *With probability q return $\mathcal{M}^P(t)$.*

Observe that by linearity of expectation, we have for all t_i, t'_i, t_{-i}:

$$\begin{aligned}
u_i(t_i, \mathcal{M}^P_\epsilon(t_i, t_{-i})) &= (1-q) \cdot u_i(t_i, \mathcal{M}_\epsilon(t_i, t_{-i})) + q \cdot u_i(t_i, \mathcal{M}^P(t_i, t_{-i})) \\
&\geq (1-q)\left(u_i(t_i, \mathcal{M}_\epsilon(t'_i, t_{-i})) - 2\epsilon\right) + q\left(u_i(t_i, \mathcal{M}^P(t'_i, t_{-i})) + \frac{\gamma}{|S|}\right) \\
&= u_i(t_i, \mathcal{M}^P_\epsilon(t'_i, t_{-i})) - (1-q)2\epsilon + q\frac{\gamma}{|S|} \\
&= u_i(t_i, \mathcal{M}^P_\epsilon(t'_i, t_{-i})) - 2\epsilon + q\left(2\epsilon + \frac{\gamma}{|S|}\right)
\end{aligned}$$

Therefore we have:

Theorem 11 *If $2\epsilon \leq \frac{q\gamma}{|S|}$ then $\mathcal{M}^P_\epsilon$ is strictly truthful.*

Note that we also have utility guarantees for this mechanism. Setting the parameter q so that we have a truthful mechanism:

$$\begin{aligned}
\mathrm{E}_{s, \hat{R} \sim \mathcal{M}^P_\epsilon}[F(t, s, r(t, s, \hat{R}))] &\geq (1-q) \cdot \mathrm{E}_{s, \hat{R} \sim \mathcal{M}_\epsilon}[F(t, s, r(t, s, \hat{R}))] \\
&= \left(1 - \frac{2\epsilon|S|}{\gamma}\right) \cdot \mathrm{E}_{s, \hat{R} \sim \mathcal{M}_\epsilon}[F(t, s, r(t, s, \hat{R}))] \\
&= \left(1 - \frac{2\epsilon|S|}{\gamma}\right) \cdot \left(\max_{t,s,r} F(t, s, r) - O\left(\frac{1}{\epsilon n}\log|S|\right)\right) \\
&\geq \max_{t,s,r} F(t, s, r) - \frac{2\epsilon|S|}{\gamma} - O\left(\frac{1}{\epsilon n}\log|S|\right)
\end{aligned}$$

Setting

$$\epsilon = O\left(\sqrt{\frac{\log|S|\gamma}{|S|n}}\right)$$

we find:

$$\mathrm{E}_{s,\hat{R}\sim\mathcal{M}_\epsilon^P}[F(t,s,r(t,s,\hat{R}))] \geq \max_{t,s,r} F(t,s,r) - O\left(\sqrt{\frac{|S|\log|S|}{\gamma n}}\right)$$

Note that in this calculation, we assume that $\epsilon \leq \gamma/(2|S|)$ so that the $q \leq 1$ and the mechanism is well defined. This is true for sufficiently large n. That is, we have shown:

Theorem 12 *For sufficiently large n, M_ϵ^P is α-approximates the social welfare F for:*

$$\alpha = O\left(\sqrt{\frac{|S|\log|S|}{\gamma n}}\right)$$

Note that this mechanism is truthful without the need for payments!

Lets now consider an application of this framework: the facility location game. Suppose that a city wants to build k hospitals to minimize the average distance between each citizen and their closest hospital. To simplify matters, we make the mild assumption that the city is built on a discretization of the unit line[1]. Formally, for all i let:

$$L(m) = \{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$$

denote the discrete unit line with step-size $1/m$. $|L(m)| = m+1$. Let $T_i = R_i = L(m)$ for all i and let $|S| = L(m)^k$. Define the utility of agent i to be:

$$u_i(t_i, s, r_i) = \begin{cases} -|t_i - r_i|, & \text{If } r_i \in s; \\ -1, & \text{otherwise.} \end{cases}$$

Note that $r_i(t_i, s)$ is here the closest facility $r_i \in s$.

We can instantiate Theorem 12. Note that in our case, we have: $|S| = (m+1)^k$ and $\gamma = 1/m$ (Because any two positions $t_i \neq t_i'$ differ by at least $1/m$). Hence, we have:

Theorem 13 *M_ϵ^P instantiated for the facility location game is strictly truthful and α-accurate for:*

$$\alpha = O\left(\sqrt{\frac{km(m+1)^k \log m}{n}}\right)$$

The exponential dependence on k can be removed by a more careful analysis of the punishment mechanism M^P.

Bibliographic Information This lecture is based on a paper of Nissim, Smorodinsky, and Tennenholtz, "Approximately Optimal Mechanism Design via Differential Privacy," 2010, which continues a line of research initiated by the work of McSherry and Talwar, "Mechanism Design via Differential Privacy", 2007.

[1] Note that if this is not the case, we can easily raze and then re-build the city.