CIS 700 Differential Privacy in Game Theory and Mechanism Design January 17, 2014

Lecture 1

Lecturer: Aaron Roth *Scribe: Aaron Roth*

Intro to Differential Privacy for Game Theorists, and Digital Goods Auctions.

1 Introduction

This course will be concerned with a recently discovered set of connections between *differential privacy* and *game theory and mechanism design*. Differential Privacy is a privacy "solution concept" that has been studied over the past decade in the computer science literature: it is a constraint that one can impose on the design space for some algorithmic problem, that has meaningful guarantees limiting the harm that individuals face as a result of their data being used as part of a computation. For the most part, the study of differential privacy has been focused firmly in the realm of *algorithm design*, in which the data is assumed to have been faithfully gathered, and free of the incentive problems that must be dealt with in *mechanism design*.

Nevertheless, as we shall see, differential privacy has a natural game theoretic semantics that makes it interesting for a couple of reasons. Most basically, it gives us a meaningful way to bound how much an individual might be harmed through loss of "privacy", and so understanding how to control it allows us to attempt to study mechanism design in settings in which agents have preferences not just for *outcomes*, but also over *mechanisms* which might affect their privacy in different ways. More surprisingly, however, the tools developed to design differentially private algorithms turn out to be useful for designing mechanisms with desirable incentive properties in large markets, even in the absence of privacy concerns. In this class, we will explore both of these connections.

2 Differential Privacy

We will begin with the original definition of differential privacy, although we will quickly derive a modified but equivalent form that will be easier for us to think about. Differential privacy is a property of an *algorithm*, which may be randomized:

Definition 1 (Randomized Algorithm) *A randomized algorithm $\mathcal{M}$ with domain A and discrete range B is associated with a mapping $M : A \to \Delta(B)$. On input $a \in A$, the algorithm $\mathcal{M}$ outputs $\mathcal{M}(a) = b$ with probability $(M(a))_b$ for each $b \in B$. The probability space is over the coin flips of the algorithm $\mathcal{M}$.*

We will be interested in algorithms that take input consisting of n-tuples of elements x from some domain X: $x \in X^n$. Think of there being n people, each of whom can have as input any element in X, and a vector $x \in X^n$ is the collected inputs of all n people. We will be particularly interested in pairs of inputs that result from the unilateral deviation of a single person i. Here we will borrow notation from the game theory literature:

Definition 2 *Given a vector $x \in X^n$, we write $x_{-i} \in X^{n-1}$ to denote $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. Given $x'_i \in X$, we write $(x_{-i}, x'_i) \in X^n$ to denote the vector $(x_{-i}, x'_i) = (x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)$.*

We are now ready to define differential privacy.

Definition 3 (Differential Privacy [DMNS06]) *A randomized algorithm $M : X^n \to \mathcal{O}$ is ϵ-differentially private if for all i, for all $x \in X^n$ for all $x'_i \in X$, and for all outcome events $S \subseteq \mathcal{O}$:*

$$\Pr[M(x) \in S] \leq \exp(\epsilon) \Pr[M(x_{-i}, x'_i) \in S]$$

A couple of remarks are in order:

1. You should think about the differential privacy condition as *really* being the more intuitive $\Pr[M(x) \in S] \leq (1+\epsilon)\Pr[M(x_{-i}, x'_i) \in S]$: that no event can increase in probability by more than a $(1+\epsilon)$ multiplicative factor when a single individual changes their data. For $\epsilon \leq 1$, the conditions are approximately the same, because we have $(1+2\epsilon) \geq \exp(\epsilon) \geq (1+\epsilon)$. We use $\exp(\epsilon)$ because it will be mathematically more convenient.

2. The neighboring relation is symmetric, so we also have the condition: $\Pr[M(x) \in S] \geq \exp(-\epsilon)\Pr[M(x_{-i}, x'_i) \in S]$. Differential privacy promises that a single individual cannot through his choice of report make any event more *or less* likely by a factor of more than $\approx 1+\epsilon$.

There are lots of reasons why this is an excellent formalization of "privacy", but we will focus on one in this class: if a rational agent gets to choose whether a mechanism M is run on input x or on neighboring input x', then no matter what his utility function is, he will be roughly indifferent between the two choices if the mechanism is ϵ-differentially private.

Claim 4 *An algorithm $M : X^n \rightarrow \mathcal{O}$ is ϵ-differentially private if and only if for every utility function $u_i : \mathcal{O} \rightarrow \mathbb{R}_{\geq 0}$ and for every $x \in X^n$ and $x'_i \in X$:*

$$\mathrm{E}_{o\sim M(x)}[u_i(o)] \leq \exp(\epsilon)\mathrm{E}_{o\sim M(x_{-i},x'_i)}[u_i(o)]$$

Proof First, suppose M is ϵ-differentially private. Then we have:

$$\begin{aligned}
\mathrm{E}_{o\sim M(x)}[u_i(o)] &= \sum_{o\in\mathcal{O}} u_i(o)\cdot\Pr[M(x)=o] \\
&\leq \sum_{o\in\mathcal{O}} u_i(o)\cdot\exp(\epsilon)\Pr[M(x_{-i},x'_i)=o] \\
&= \exp(\epsilon)\mathrm{E}_{o\sim M(x_{-i},x'_i)}[u_i(o)]
\end{aligned}$$

In the reverse direction, fix an event $S \subseteq \mathcal{O}$, and define u_i to be $u_i(o) = \begin{cases} 1, & \text{If } o \in S; \\ 0, & \text{otherwise.} \end{cases}$ We have for this utility function:

$$\mathrm{E}_{o\sim M(x)}[u_i(o)] \leq \exp(\epsilon)\mathrm{E}_{o\sim M(x_{-i},x'_i)}[u_i(o)]$$

But this is equivalent to:

$$\Pr[M(x) \in S] \leq \exp(\epsilon)\Pr[M(x_{-i},x'_i) \in S]$$

which is the differential privacy condition. ∎

So differential privacy promises, simultaneously for all n people i, no matter what their utility functions are, that they cannot substantially hurt their expected future utility[1] by changing their reported input to the algorithm x_i. This focus on unilateral deviation should seem familiar to you if you have studied game theory, and will form the core of the connection between the two fields.

3 Mechanism Design

Lets go through the basic definitions of mechanism design and then jump right into a problem we might want to solve. We will think of n rational (utility maximizing) agents i who have privately known *types*

[1] Caution: Note that in the utility theoretic definition, we do make the assumption that the utility function $u : \mathcal{O} \rightarrow \mathbb{R}_{\geq 0}$ takes on only non-negative values. If the utility function can take on negative values, then this interpretation of the guarantee is no longer true: in this case, the expected utility of a player can in principle jump from 0 to a nonzero value, which isn't bounded by any multiplicative factor.

$t_i \in \mathcal{T}$. A mechanism $\mathcal{M} : \mathcal{T}^n \to \mathcal{O}$ is a mapping between (reported) types of the n agents, and some outcome space $\mathcal{O}$. Agents have preferences over outcomes, which are encoded by their types: the utility that agent i gets from outcome $o \in \mathcal{O}$ is defined to be:

$$u_i(o) \equiv u(t_i, o)$$

where $u : \mathcal{T} \times \mathcal{O} \to [0, 1]$ is some utility function.

We want to design mechanisms that both choose a desirable social outcome (according to some objective function we might have) and also incentivize agents to report their true types (so that we know we are evaluating our objective function on the right data).

Definition 5 *A mechanism $\mathcal{M} : \mathcal{T}^n \to \mathcal{O}$ is (ϵ-approximately) dominant strategy truthful if for all $t \in \mathcal{T}^n$ and for all i and $t'_i \in \mathcal{T}$:*

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

That is, no player can gain more than ϵ utility by mis-representing their type. If $\mathcal{M}$ is randomized, then we take $u_i(\mathcal{M}(t))$ to mean $\mathrm{E}_{o \sim \mathcal{M}(t)}[u_i(o)]$ (i.e. we assume players are risk neutral.*)*

Approximate truthfulness will be most interesting in settings in which we can take the approximation parameter ϵ tending to 0 as the number of people n grows large. We will call this *asymptotic truthfulness.*

Note the striking similarity between the definitions of (approximate) dominant strategy truthfulness and differential privacy. We have the immediate observation:

Claim 6 ([MT07]) *If $\mathcal{M}$ is ϵ-differentially private for $\epsilon \leq 1$, then it is also ϵ-approximately dominant strategy truthful.*

Proof This follows immediately from the definition, and the observation that:

$$\mathrm{E}_{o \sim M(t)}[u_i(o)] \geq \exp(-\epsilon)\mathrm{E}_{o \sim M(t_{-i}, t'_i)}[u_i(o)] \geq (1-\epsilon)\mathrm{E}_{o \sim M(t_{-i}, t'_i)}[u_i(o)] \geq \mathrm{E}_{o \sim M(t_{-i}, t'_i)}[u_i(o)] - \epsilon$$

where the first inequality is the differential privacy condition, the second follows from $\exp(-\epsilon) \geq 1 - \epsilon$, and the third follows from the fact that utilities are bounded in $[0, 1]$. ■

It is worth noting a couple of things about this connection between privacy and approximate truthfulness.

1. A reasonable objection is that this straightforward application of privacy makes not just truthful reporting an approximate dominant strategy – it makes *everything* an approximate dominant strategy. Why should people tell the truth in such cases? As we will see, however, a typical application of this approach will give a more nuanced guarantee, in which truthful reporting remains an approximate dominant strategy, but not everything does.

2. The differential privacy definition composes across player deviations: if 4 players change their reports, then the probability of any event can change by at most $\exp(4\epsilon)$. Hence, differential privacy automatically promises approximate *group strategyproofness* as well. No coalition of k players can improve their expected utility by more than a factor of $\exp(k\epsilon)$ by deviating from truthtelling behavior.

Regarding the first point above, we make the following observation. Suppose the utility-relevant event for each player i in fact comes from some other outcome space $\mathcal{O}'_i$, and players have utility functions $u_i : \mathcal{O}'_i \to [0, 1]$. Suppose moreover, that there is some function $f : \mathcal{O} \times \mathcal{T} \to \mathcal{O}'_i$ mapping both the outcome $o \in \mathcal{O}$ selected by $\mathcal{M}$ and agent i's reported type t'_i to an outcome $f(o, t'_i) \in \mathcal{O}'_i$. If we have that for every fixed o, $f(o, \cdot)$ makes truthful reporting a dominant strategy, then truthful reporting remains an 2ϵ-approximate dominant strategy to $\mathcal{M}$, but non-truthful reports may no longer be approximate dominant strategies.

Claim 7 *If $\mathcal{M} : \mathcal{T}^n \to \mathcal{O}$ is ϵ-differentially private and $f(o, \cdot) : \mathcal{T} \to \mathcal{O}'_i$ is dominant strategy truthful for every $o \in \mathcal{O}$, then truthful reporting is a ϵ-approximate dominant strategy for the mechanism $\mathcal{M}' : \mathcal{T}^n \to \mathcal{O}'$ that produces outcome $f(\mathcal{M}(t), t_i)$ for agent i.*

Proof

$$\mathrm{E}_{o\sim M(t)}[u_i(f(o,t_i))] \geq \exp(-\epsilon)\mathrm{E}_{o\sim M(t_{-i},t'_i)}u_i(f(o,t_i))] \geq \exp(-\epsilon)\mathrm{E}_{o\sim M(t_{-i},t'_i)}[u_i(f(o,t'_i))]$$

$$\geq (1-\epsilon)\mathrm{E}_{o\sim M(t_{-i},t'_i)}[u_i(f(o,t'_i))] \geq \mathrm{E}_{o\sim M(t_{-i},t'_i)}[u_i(f(o,t'_i))] - \epsilon$$

■

4 Digital Goods Auctions

A *digital goods auction* is one in which the auctioneer is selling copies of an identical good, with zero marginal cost of production, to n bidders. Think about Microsoft trying to auction off copies of Windows: there is no finite supply, since copying software is free. As such, its not interesting to consider *welfare maximizing* auctions in this setting, because in this case, you could just give the item to all bidders, and charge them nothing. Instead, we will be interested in revenue.

Formally, each bidder has a type $v_i \in [0,1]$ which represents his value for receiving a copy of the good. (We let the type space $\mathcal{T} = [0,1]$, and values $v_i \in \mathcal{T}$ are identified with bidder types). The outcome space $\mathcal{O} = 2^{[n]} \times [0,1]^n$ is the set of all subsets of bidders who might be selected to win an item, together with the set of all prices they might be charged: an outcome is a pair (S, p), where bidders $i \in S$ receive an item and must pay p_i. Bidders who i receive an item get utility $v_i - p_i$; bidders who do not receive an item get utility 0:

$$u_i(S,p) = \begin{cases} v_i - p_i, & i \in S; \\ 0, & \text{otherwise.} \end{cases}$$

As the mechanism designer, we are interested in the revenue obtained by the auction. For each outcome (S, p), the revenue is simply the sum of the prices collected: $\mathrm{Rev}(S,p) = \sum_{i \in S} p_i$.

To calibrate our goals, we need to fix a price benchmark. Since we are not in a Bayesian setting (in which bidders valuations are drawn from a known prior distribution), there is no notion of a single, revenue-optimal truthful mechanism. A natural benchmark is that of the revenue of the best *fixed* price, given the realized bidder valuations[2]. Consider the revenue we can obtain if we offer every (rational) bidder a fixed price $p \in \mathbb{R}_{\geq 0}$. If the price that we offer is below their value, they will buy the item, contributing p to our total revenue. Otherwise, they will not buy the item, and they contribute 0 to our revenue. Hence, given bidder valuations $v \in [0,1]^n$, the revenue we can obtain with a fixed price $p \in \mathbb{R}_{\geq 0}$ is:

$$\mathrm{Fixed}(v,p) = p \cdot |\{i : v_i \geq p\}|$$

The revenue of the *best* fixed price is therefore:

$$\mathrm{OPT}(v) = \max_{p \in [0,1]} \mathrm{Fixed}(v,p)$$

We would like to design an auction that *in the worst case over bidder valuations v* obtains revenue that is close to $\mathrm{OPT}(v)$.

[2] There are various ways to justify this benchmark, but one is the following: for any fixed prior distribution from which bidders valuations are assumed to be drawn i.i.d., the Myerson revenue-optimal auction will end up charging each winning bidder some fixed price. What price this is depends on the distribution. If we can do almost as well as the revenue of the best fixed price in hindsight, even in the worst case over bidder valuations, then we are competitive with the Myerson optimal auction simultaniously for all priors, and we did it in a prior independent way!

4.1 A (Really) Simple Power Tool

Lets take a moment to introduce a really simple but powerful tool in differential privacy that we will have opportunity to meet again: the exponential mechanism.

The exponential mechanism $\mathcal{M}_E : X^n \to \mathcal{O}$ is instantiated with an arbitrary range $\mathcal{O}$ together with a *quality score* $q : X^n \times \mathcal{O} \to \mathbb{R}$, which maps input/output pairs to a number, which intuitively should be thought of as measuring how "good" an output o is, for input x. The mechanism outputs an element of $\mathcal{O}$ at random, giving exponentially weighted preference to those outputs with higher quality score.

A key parameter of the quality score is its *sensitivity*:

Definition 8 *The sensitivity of the quality score* $q : X^n \times \mathcal{O} \to \mathbb{R}$ *is defined to be:*

$$GS(q) = \max_{i\in[n], x\in X^n, x'_i \in X, o\in\mathcal{O}} |q(x,o) - q((x_{-i}, x'_i), o)|$$

It is the maximum absolute value change a single individual can affect on the quality score, in the worst case over outputs o.

Algorithm 1 The Exponential Mechanism

```
$\mathcal{M}_E(x; q, \mathcal{O}, \epsilon)$:
  **Output** $o \in \mathcal{O}$ with probability:
      $\frac{\exp\left(\frac{\epsilon \cdot q(x,o)}{2 \cdot GS(q)}\right)}{Z(x)}$
  where $Z(x) = \sum_{o\in\mathcal{O}} \exp\left(\frac{\epsilon \cdot q(x,o)}{2\cdot GS(q)}\right)$
```

The exponential mechanism will be a useful tool for us because of two basic theorems. First, the exponential mechanism is ϵ-differentially private:

Theorem 9 ([MT07]) *For every range* $\mathcal{O}$ *and quality score* q, $\mathcal{M}_E(\cdot; q, \mathcal{O}, \epsilon)$ *is* ϵ*-differentially private.*

Proof We'll prove privacy using the first form of the differential privacy definition we saw. Fix an arbitrary $x \in X^n, x'_i \in X$ and consider any output $o \in \mathcal{O}$. We have:

$$\begin{aligned}
\frac{\Pr[\mathcal{M}_E(x; q, \mathcal{O}, \epsilon) = o]}{\Pr[\mathcal{M}_E((x_{-i}, x'_i); q, \mathcal{O}, \epsilon) = o]} &= \frac{Z((x_{-i}, x'_i))}{Z(x)} \cdot \frac{\exp\left(\frac{\epsilon\cdot q(x,o)}{2\cdot GS(q)}\right)}{\exp\left(\frac{\epsilon\cdot q((x_{-i},x'_i),o)}{2\cdot GS(q)}\right)} \\
&= \frac{Z((x_{-i}, x'_i))}{Z(x)} \cdot \exp\left(\frac{\epsilon \cdot (q(x,o) - q((x_{-i}, x'_i), o))}{2 \cdot GS(q)}\right) \\
&\leq \frac{Z((x_{-i}, x'_i))}{Z(x)} \cdot \exp\left(\frac{\epsilon \cdot GS(q)}{2 \cdot GS(q)}\right) \\
&= \frac{Z((x_{-i}, x'_i))}{Z(x)} \cdot \exp\left(\frac{\epsilon}{2}\right) \\
&\leq \exp(\epsilon)
\end{aligned}$$

■

Next, we show that the exponential mechanism actually outputs an element with pretty high quality score (most of the time).

Theorem 10 ([MT07]) *Define* $OPT_{q,\mathcal{O}}(x) = \max_{o^* \in \mathcal{O}} q(x, o^*)$ *Let* $o = \mathcal{M}_E(x; q, \mathcal{O}, \epsilon)$*. Then:*

$$\Pr[q(x,o) \leq OPT_{q,\mathcal{O}}(x) - \frac{2 \cdot GS(q)}{\epsilon}(\log|\mathcal{O}| + t)] \leq \exp(-t)$$

Proof For any quantity t:

$$\begin{aligned}
\Pr[q(x,o) \leq t] &\leq \frac{\Pr[q(x,o) \leq t]}{\Pr[q(x,o) = \mathrm{OPT}_{q,\mathcal{O}}(x)]} \\
&\leq \frac{|\mathcal{O}| \exp\left(\frac{\epsilon t}{2 \cdot GS(q)}\right)}{\exp\left(\frac{\epsilon \mathrm{OPT}_{q,\mathcal{O}}(x)}{2 \cdot GS(q)}\right)} \\
&= |\mathcal{O}| \cdot \exp\left(\frac{\epsilon(t - \mathrm{OPT}_{q,\mathcal{O}}(x))}{2 \cdot GS(q)}\right)
\end{aligned}$$

Plugging in $t = \mathrm{OPT}_{q,\mathcal{O}}(x) - \frac{2 \cdot GS(q)}{\epsilon}(\log|\mathcal{O}| + t)$ we get:

$$\begin{aligned}
\Pr[q(x,o) \leq t] &\leq |\mathcal{O}| \exp\left(-(\log|\mathcal{O}| + t)\right) \\
&= \exp(-t)
\end{aligned}$$

■

4.2 Using the exponential mechanism to pick a price

We now turn our attention back to digital goods auctions. Recall that we want to compete with a "best fixed price" benchmark, and we will do so in a relatively straightforward way:

1. We will pick a price $\hat{p} \in [0, 1]$ from the exponential mechanism to maximize $\mathrm{Fixed}(v, p)$.
2. We will then sell to each bidder i at this price $\hat{p}$ if their reported value $v_i \geq p$. That is, for each bidder i: $i \in S$ and $p_i = \hat{p}$ iff $f(\hat{p}, v_i) \equiv \mathbf{1}_{v_i \geq \hat{p}} = 1$.
3. Since for every fixed price $\hat{p}$, reporting ones true value is a dominant strategy, and $\hat{p}$ is chosen in an ϵ-differentially private way, the whole mechanism will be ϵ-approximately dominant strategy truthful.

A couple of details to work out: first, if we are going to use $\mathrm{Fixed}(v, p)$ as the quality score in our exponential mechanism, we need to understand its sensitivity. But this is straightforward: Recall $\mathrm{Fixed}(v, p) = p \cdot |\{i : v_i \geq p\}|$. For every fixed price p, changing a single valuation v_i can change the size of the set $\{i : v_i \geq p\}$ by at most 1. Since $p \in [0, 1]$, that means $GS(\mathrm{Fixed}) = 1$.

Second, we have to decide what set $\mathcal{O}$ to pick prices from. As defined, $\mathcal{O}$ must be a finite set[3]. Let us consider a discretized range $\mathcal{O} = \{\alpha, 2\alpha, \ldots, 1\}$, so that $|\mathcal{O}| = 1/\alpha$. We have a tradeoff to consider. On the one hand, taking a finer discretization (i.e. a smaller value of α) will allow the revenue of the best fixed price in our discrete set $\mathcal{O}$ to more closely approach $\mathrm{OPT}(v)$. On the other hand, the exponential mechanism does not guarantee us that it will find the *best* fixed price in $\mathcal{O}$ – but merely one that achieves revenue that is an additive $O\left(\frac{\log|\mathcal{O}|}{\epsilon}\right)$ of the best – so we pay a penalty for taking a discretization that is too fine.

Observe that for all v:

$$\max_{p \in \mathcal{O}} \mathrm{Fixed}(v, p) \geq \mathrm{OPT}(v) - \alpha n.$$

[3] It is possible to define and analyze the exponential mechanism when it has a continuous range, but I prefer finite things and I am writing these lecture notes, so there.

To see why this is, consider the optimal price $p^* \in [0,1]$ that achieves revenue $\mathrm{OPT}(v)$. There is some price $\hat{p} \in \mathcal{O}$ such that $p^* - \alpha \le \hat{p} \le p^*$. If we instead use price $\hat{p}$, the *number of people* who purchase the good $|\{i : v_i \ge \hat{p}\}|$ can only increase, and from each such person, we lose at most α in revenue.

Putting this all together, for any vector of valuations v, our mechanism will with high probability achieve revenue:

$$\mathrm{Rev} \ge \mathrm{OPT}(v) - \alpha n - O\left(\frac{\log 1/\alpha}{\epsilon}\right)$$

If we choose $\alpha = 1/n$, then this becomes:

$$\mathrm{Rev} \ge \mathrm{OPT}(v) - O\left(\frac{\log n}{\epsilon}\right)$$

Putting it all together, we have proven:

Theorem 11 ([MT07]) *There is an ϵ-approximately dominant strategy truthful auction that in the worst case over bidder valuations, guarantees revenue at least $OPT(v) - O\left(\frac{\log n}{\epsilon}\right)$ with probability at least* 99%*.*

Remark Note that in a "large economy" we should expect $\mathrm{OPT}(v)$ to grow with n. If $\mathrm{OPT}(v)$ grows even *mildly* with n (at a rate faster than $\log n$), then this theorem is guaranteeing that we get all but a diminishing fraction of the optimal revenue. More realistically, we might expect $\mathrm{OPT}(v)$ to grow linearly with n. In this case, we could take (say) $\epsilon = 1/\sqrt{n}$, and not only would the approximation to the optimal revenue become exact as n grows large, but the truthfulness guarantee would become exact as n grows large. This is an *asymptotically truthful mechanism* that achieves *asymptotically optimal* revenue in large economies, all without requiring access to a prior on bidder valuations. Whats more, because of the differential privacy guarantee, the incentive properties are robust to collusion.

What I want to emphasize most, however, is the *simplicity* of the mechanism. Once we developed the necessary machinery, designing this mechanism was trivial. Whats more, it naturally generalizes to multi-parameter domains, in which designing exactly truthful mechanisms can be difficult.

Finally, its worth noting the best exactly truthful mechanism known for this setting. [BBHM05] give an exactly dominant strategy truthful mechanism that achieves revenue $\mathrm{OPT}(v) - O\left(\sqrt{n}\right)$. What we have done in this lecture is traded exact truthfulness for asymptotic truthfulness, but we have gained an exponential factor in the additive loss we incur on top of our revenue benchmark!

Bibliographic Information The definition of Differential Privacy is from Dwork, McSherry, Nissim, and Smith 2006 [DMNS06]. I first heard it defined in its utility-theoretic form in a talk by Kobbi Nissim. Its application to digital goods auctions is from "Mechanism Design via Differential Privacy" by McSherry and Talwar, 2007 [MT07]. This was the first paper that suggested that differential privacy might be a useful tool in game theory. [MT07] also introduced the exponential mechanism, which quickly became one of the foundational building blocks in the differential privacy toolkit.

References

[BBHM05] M-F Balcan, Avrim Blum, Jason D Hartline, and Yishay Mansour. Mechanism design via machine learning. In *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on*, pages 605–614. IEEE, 2005.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC '06*, pages 265–284, 2006.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103, 2007.