Pricing Private Data

Vasilis Gkatzelis*
New York University
New York, NY

Christina Aperjis
HP Labs
Palo Alto, CA

Bernardo A. Huberman
HP Labs
Palo Alto, CA

September 14, 2012

Abstract

We consider a market where buyers can access unbiased samples of private data by appropriately compensating the individuals to whom the data corresponds (the sellers) according to their privacy attitudes. We show how bundling the buyers' demand can decrease the price that buyers have to pay per data point, while ensuring that sellers are willing to participate. Our approach leverages the inherently randomized nature of sampling, along with the risk-averse attitude of sellers in order to discover the minimum price at which buyers can obtain unbiased samples. We take a prior-free approach and introduce a mechanism that incentivizes each individual to truthfully report his preferences in terms of different payment schemes. We then show that our mechanism provides optimal price guarantees in several settings.

*Work performed while an intern at HP Labs.

Electronic copy available at: http://ssrn.com/abstract=2146966

1 Introduction

As the great value of big data is being recognized and the cost of computer memory keeps dropping, the amount of personal information gathered about individual consumers has reached unprecedented levels. The economic value of this data is reflected in the success of many Internet companies, from search engines to social media sites and data repositories which routinely sell this information.

Still, large amounts of potentially useful private data cannot be accessed by interested parties due to privacy concerns [6]. In particular, a number of companies and entities gather lots of data about groups of individuals that would be very useful to third parties. For instance, a hospital may have information about individuals with a certain disease that a pharmaceutical or a researcher would like to know, or a cable provider may have information about the viewing habits of a certain demographic of interest to a TV channel. However, these entities are often reluctant to allow others to access such data because of the privacy concerns of the corresponding individuals. At the same time, individuals' data is being bought and sold by data brokers, such as Acxiom, often without the knowledge of the individuals that the data pertains to [14].[1]

One solution is to create a market for private data through which buyers can pay individuals (sellers) in exchange for obtaining access to their private data. Sellers can then opt to participate in this market if the price is high enough. In what follows we consider settings where a buyer is interested in getting access to the data of a representative subset of individuals with certain characteristics. This can be achieved by giving the buyer access to an *unbiased sample* of a certain size, that is, data for a subset of individuals who are chosen uniformly at random from the set of all individuals with the characteristics the buyer is interested in. Such a sample will typically be representative because of the Law of Large Numbers as long as a representative subset of individuals chooses to participate in the market.

Different individuals may have diverse privacy attitudes, and as a result they may be willing to participate in the market for different prices. Quite often one's privacy attitude is correlated with the value of certain attributes that a buyer may be interested in [9]. In order to minimize this bias, we set the price high enough so that almost all individuals choose to participate in the market. Our goal is to set the minimum such price so as to maximize the buyers' interest in the market and the amount of data that is traded.

[1] As has been publicly discussed, such trades are controversial [13].

Electronic copy available at: http://ssrn.com/abstract=2146966

Recently, Aperjis and Huberman showed how one can take advantage of the risk averse attitude of some sellers to set a price per data point that is lower than the price that the most privacy concerned sellers are willing to accept [2]. It is possible to further reduce the price per data point by bundling buyers' demand because individuals tend to exhibit more risk aversion for higher payments [8]. It is therefore of interest to design a market for private data that can reduce the price by properly bundling the demand.

In this paper we design such a market and identify the optimal way to bundle demand so that it minimizes the expected payment to a risk averse seller. We take a prior-free approach and assume no knowledge of the distribution of sellers' privacy and risk attitudes.

Markets for private data have been previously studied in the setting of a buyer interested in estimating a certain statistic property of a set of private data, such as its average [12], sum [5] or weighted sum [4]. In contrast to that work, we consider a scenario where buyers pay for access to the raw data instead of just an estimate for its statistical value. The selling of raw private data has been previously considered (e.g., [11]) but not for unbiased samples, which is the focus of this paper.

One line of existing work studies how to estimate certain statistics in the context of differential privacy [5, 4]. A drawback of the differential privacy approach is that in order to achieve a reasonably accurate estimate the buyer needs to use data from the majority of individuals in the subset of interest, which renders this mechanism very expensive for the buyer.[2] Our approach avoids this problem by selling access to small unbiased subsets of the data that a buyer can then use to compute statistics about them.

More recently, Roth and Schoenebeck [12] have shown how to estimate the average of a set of private values using the Horvitz-Thompson estimator. This approach assumes that the mechanism has access to the distribution from which the sellers' privacy attitudes are generated. In contrast, we take a prior-free approach with respect to sellers' privacy attitudes.

As in [2], we consider a market-maker who facilitates interactions between buyers and sellers by eliciting the privacy and risk preferences of sellers and by setting the price appropriately. In contrast to [2], we construct mechanisms that the market-maker can use to elicit this private information that (1) are dominant strategy truthful,[3] (2) are individually rational, and

[2]For instance, Ghosh and Roth (2011) consider the problem of getting an unbiased estimate of the sum of n bits (each of these bits corresponds to a seller). In order to have that $Pr[|\text{estimated sum} - \text{true sum}| \geq 0.08n] \leq 1/3$, the buyer needs to pay 95 percent of all sellers.

[3]Truthfulness here refers to reporting privacy and risk preferences; this information

(3) provide guarantees in terms of minimizing the price that buyers pay. For our formal results, we assume that the demand is fixed and known in advance by the market-maker.

Our mechanisms take as input a distribution that represents how sellers will be sampled. In our Baseline Mechanism, the market-maker asks each seller to report his *privacy "cost"*, that is, for what price he would be willing to allow n buyers to get access to his data. The market-maker then constructs a payment function that specifies what monetary payment a seller will receive depending on how many buyers get access to his private data. This payment is set high enough so that all sellers are willing to participate in the sampling.

Our second mechanism is the Certainty Equivalent (CE) Mechanism. As in the Baseline Mechanism, each seller is first asked to report his privacy cost. Then, each seller is also asked to report the guaranteed amount of money that he would view as equally desirable as participating in the Baseline Mechanism, i.e., his certainty equivalent for a lottery that corresponds to the Baseline Mechanism. For risk-averse sellers, the certainty equivalent is lower than the expected value of the lottery. The market-maker then uses the reported values to determine a payment function for each seller.

We first consider the case where each seller's privacy cost is linear in the number of buyers that get access to his data. We identify the distribution that minimizes the certainty equivalent of each risk averse seller. We call this the *ordering distribution* because it is uses a random ordering of sellers to determine a sample for each buyer. The ordering distribution corresponds to the optimal way of bundling buyers' demand in the sense that it minimizes the expected payment for which a risk averse seller is willing to participate in the market when the CE Mechanism is used.

We then consider the more general case where each seller's cost may be any concave function of the number of buyers. We show that the ordering distribution provides the best possible worst-case guarantee for the price of the Baseline Mechanism within the class of distributions that are not aware of the cost functions a priori. In particular, no cost-oblivious distribution can achieve an expected payment per seller that is less than two times the expected payment from a distribution that we can select when the costs are known in advance. We then show that the ordering distribution achieves an approximation factor of two compared to the case that costs are known in advance.

---

will be used to determine the price that a buyer has to pay. We assume that the data a buyer is interested in is accurately stored in some database.

The paper is structured as follows. Section 2 describes the basic structure and mechanics of the market. Section 3 presents our formal model. Section 4 considers the case where the costs to the sellers are linear in the number of buyers. In Section 5, we study the case of general cost functions. Section 6 concludes.

2 The Market

In this section, we describe how a market for private data can work and highlight the key aspects of our approach; the approach we take here builds on [2].

Consider a data repository that contains information on n individuals (the sellers). For instance, this could be information obtained by a company about its customers through its interactions with them. A buyer is a third party interested in obtaining access to a certain subset of the data repository such as the data that corresponds to a representative subset of individuals from a certain demographic, and who currently does not have the right to access this information. A buyer will be able to access information from an unbiased, and thus representative, sample of individuals if he appropriately compensates the corresponding individuals.

Similarly to [2], we envision that a market-maker facilitates the interactions between buyers and sellers, as shown schematically in Figure 1. The market-maker could be a data broker or an entity (e.g., cable provider, hospital) with data about groups of individuals that it is currently reluctant to sell access to because of privacy concerns. The market-maker elicits the following information from each side of the market:

(1) the demand of each buyer, i.e., what type of individuals he is interested in and the data of how many such individuals he wants to get access to, given a price per individual

(2) the privacy attitude of each seller, i.e., for what price he would be willing to allow a buyer to get access to his data

Suppose that the market-maker already has information on each seller's privacy attitude. Then, every time a potential buyer provides a request specifying what type of individuals he is interested in, the market-maker can quote the price that the buyer needs to pay per individual in order to obtain an unbiased sample. The buyer can then choose how many individuals' data to buy access to, thus trading off accuracy and cost: a larger sample will provide more accurate results, but at a higher cost. The buyer pays

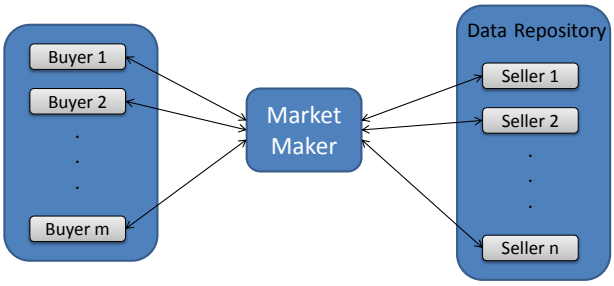


Figure 1: The market-maker facilitates interactions between buyers and sellers.

the market-maker the corresponding price and the market-maker returns an unbiased random sample of the requested size. Finally, the market-maker uses the buyer's payment to appropriately compensate the sellers, while keeping a cut for himself. However, for expository ease, we ignore the market-maker's cut in our formal analysis.

Privacy attitudes. In general, the market-maker does not know the privacy attitude of each seller. Our goal is to design a mechanism that the market-maker can use to truthfully elicit this information from each seller. Different individuals have different privacy attitudes [9, 1, 3, 7]. For instance, some individuals may not be concerned about privacy and would allow a buyer to access their data in exchange for a few cents, whereas others may only consent if paid at least $10. Since all individuals prefer to be paid more, even those unconcerned about their privacy may pretend that they are if they expect that this will result in getting higher payments.

On the other hand, buyers are interested in obtaining unbiased samples without having to pay too much for them. Confronted with this problem, the market-maker may be tempted to use a mechanism along the lines of a reverse auction: ask each seller to report the minimum price for which he would allow a buyer to access his data, and then sell a buyer access to the data of the sellers who reported the lowest prices.[4] But this will not give an unbiased sample because the value of an attribute is often correlated with its corresponding privacy valuation [9]. The requirement of an unbiased sample implies that each individual should be selected with the same probability,

[4]By setting seller compensations appropriately, this mechanism will induce truthful reporting.

independently of how much he values privacy.[5]

We now describe a mechanism that truthfully elicits sellers' privacy attitudes while introducing negligible bias; this is the Baseline Mechanism of Section 4.1 for the case of one buyer. The mechanism first asks each seller to report the minimum price for which he would allow a buyer to access his data. Then, the seller with the maximum reported price, c^{max}, is discarded and a random sample of sellers from the remaining set is selected. A buyer gets access to the data of each sampled seller and each sampled seller gets paid c^{max} in return.

Risk aversion. Observe that with the aforementioned mechanism, a seller receives a high payment (c^{max}) if he is sampled and no payment otherwise. Consider a seller that is not concerned about privacy and would be willing to give access to his data if paid any positive amount. Then, for this seller the mechanism is equivalent to a lottery with gives him a high payment (c^{max}) with probability equal to the proportion of sellers that will be sampled and no payment otherwise. If the seller is risk-averse, he will prefer to get a certain payment with lower expected value rather than participate in the lottery. The market-maker can then offer this seller an appropriate certain payment instead of the lottery and thus reduce the expected payment of the buyer.

Other risk averse sellers may also be willing to replace the lottery with a less risky lottery that has a lower expected payment. The market-maker can then further reduce the expected price that the buyer has to pay for an unbiased sample. Some examples are discussed in [2]. In this paper we take a prior free approach and assume that the market-maker does not have any information about the sellers' risk attitudes. Nevertheless, we show that it is possible to design a mechanism that elicits the true risk attitude (in addition to the true privacy attitude) of each individual seller.

Bundling buyers' requests. Individuals tend to exhibit more risk aversion for higher payments [8]. This suggests that it may be possible to further reduce the price per data point by bundling the buyers' demand. For instance, the market-maker may sample sellers so that a seller's data is either accessed by a large number of buyers in return for a large payment or by no buyer in return for no payment. In order to do this, the market-maker needs to elicit for what price each would be willing to allow n buyers to get access to his data. We consider the case where the price in linear in n

[5]It is possible to produce an unbiased statistic from a biased sample (e.g., with the Horvitz-Thompson estimator), but here we take a prior free approach and do not restrict attention to a specific statistic. It is thus natural in our setting to aim for unbiased samples.

for each seller in Section 4. Then, in Section 5, we consider the non-linear case. Note that the bundling we do here is different than product bundling where several products are offered for sale as one combined product; here we bundle buyers' request, i.e., the demand.

In the following sections, we assume that demand is price-insensitive and known by the market-maker; this assumption allows us to focus on the seller side of the market and to study the optimal way to bundle buyers' demand. Furthermore, for expository ease we assume that all buyers are interested in individuals with the same characteristics. We note, however, that our results also hold in the general case where each buyer may be interested in sellers with different characteristics.

3 Model

Buyers. We use B to denote the set of m buyers that are interested in acquiring access to the data of representative sets of the sellers. Each buyer $b \in B$ reports his demand d_b, which represents the size of the unbiased sample that buyer b wishes to buy access to. We assume that demand is price-insensitive; that is, buyer b is interested in obtaining access to an unbiased sample of d_b sellers regardless of the price he has to pay per individual seller. In fact, our mechanisms work more generally for settings where demand does not change drastically with the price and/or the market-maker has a good estimate of the right range for the price. In Section 6, we discuss how we can relax this assumption.

Sellers. Let N denote the set of n sellers who are willing to sell access to their private data. Each seller $i \in N$ is characterized by two functions representing his privacy and risk attitudes. The privacy attitude of seller i is modeled with a non-decreasing cost function $c_i : \mathbb{N} \to \mathbb{R}$, where $c_i(k)$ represents the smallest payment for which seller i would allow k buyers to access his data. We assume that $c_i(0) = 0$. We model the risk attitude of seller i with a non-decreasing utility function $u_i : \mathbb{R} \to \mathbb{R}$ with $u_i(0) = 0$. Both c_i and u_i are private information of seller i, that is, only seller i knows these functions.

We make the following assumption:

Assumption 3.1 *The utility of seller i for obtaining a monetary payment x while allowing k buyers to access his data is equal to $u_i(x - c_i(k))$.*

When faced with randomness with respect to the payment or the number of buyers that will access his data, we assume that each seller aims to maximize the expected value of his utility [10].

Given a set of sellers N' and buyers' demand $\{d_b, b \in B\}$ let $\Psi(N')$ denote the set of all distributions that produce unbiased samples of sizes $\{d_b, b \in B\}$ from N'. In particular, a distribution $\psi \in \Psi(N')$ will produce a set of unbiased samples $\{s_b, b \in B\}$ from N', where s_b represents the set of d_b sellers to whose data buyer b will get access. Given a distribution ψ representing the sampling, let $p_\psi(k)$ be the probability that the data of a seller is sold to exactly k buyers. Thus, p_ψ represents the distribution of the number of times that a seller will be sampled. If $\psi \in \Psi(N')$, the distribution p_ψ is the same for all sellers in N'.

If k buyers get access to the data of seller i, he will be compensated with some payment which we denote by $\pi_i(k)$. In contrast to the probability that k buyers get access to one's data, the corresponding payment may vary across sellers. We assume that each $\pi_i(k)$ is deterministic. Then, the expected utility of seller i for a given distribution p_ψ and corresponding payments π_i is equal to $\sum_{k=0}^{m} p_\psi(k) u_i(\pi_i(k) - c_i(k))$.

Given distribution p_ψ and payments π_i for $i \in N$, the expected total payment (over all sellers) is equal to $\sum_{i=1}^{n} \sum_{k=0}^{m} p_\psi(k)\pi_i(k)$. Thus, the expected price per seller is given by

$$\frac{\sum_{i=1}^{n} \sum_{k=0}^{m} p_\psi(k)\pi_i(k)}{\sum_{b \in B} d_b}.$$

Because we are assuming that the demand is fixed, the market-maker aims to minimize the nominator, which is the expected total payment (over all sellers).

Suppose that we fix some distribution ψ and payments π_i. We write (p_ψ, π_i) to denote the lottery according to which the number of buyers that get access to one's data is drawn from the distribution p_ψ and seller i is compensated according to the payment function π_i. The concept of certainty equivalent, which we define next, is crucial for some of our mechanisms.

Definition Fix a distribution ψ and payments π_i. The *certainty equivalent* of (p_ψ, π_i) for seller i, denoted $e_i(p_\psi, \pi_i)$, is the amount of money for which seller i is indifferent between the lottery (p_ψ, π_i) and the certain amount $e_i(p_\psi, \pi_i)$; that is, $u_i(e_i(p_\psi, \pi_i)) = \sum_{k=0}^{m} p_\psi(k) u_i(\pi_i(k) - c_i(k))$.

An individual is risk-averse if he prefers a less risky lottery to a more risky one with the same expected payment. Experimental studies have shown that many people exhibit risk-averse behavior even for small payoffs, e.g., payments equal to a couple of dollars; however, the effect is more prominent when the lotteries involve high payments [8].

Formally, seller i is risk-averse if his utility function u_i is concave. Risk-aversion implies that the certainty equivalent of a lottery is smaller than its expected value. That is, given a lottery (p_ψ, π_i), if seller i is risk-averse, we have that $e_i(p_\psi, \pi_i) < \sum_k p_\psi(k)(\pi_i(k) - c_i(k))$. Alternatively, a seller may be risk-neutral (which corresponds to a linear utility function) or risk-seeking (which corresponds to a convex utility function).

Suppose that the number of buyers that will get access to the data of seller i will be drawn according to some distribution p_ψ. Fix a payment function π_i. Because of Assumption 3.1, seller i is indifferent between being compensated according to π_i and being compensated according to π_i', where $\pi_i'(k) \equiv e_i(p, \pi_i) + c_i(k)$. If seller i is risk-averse, then the expected payment under π_i' is smaller than the expected payment under π_i. We assume that buyers are risk-neutral; thus, even though in this example a risk-averse seller may be indifferent between the payments π_i and π_i', a buyer strictly prefers to compensate seller i according to π_i'.

In the following sections we introduce mechanisms that can take advantage of risk aversion to reduce the expected price that each buyer has to pay per seller, while ensuring that all sellers are willing to participate in the market. The price is reduced more when more sellers exhibit risk-averse behavior; however, our mechanisms do not require that all sellers are risk-averse.

We focus on mechanisms that are *dominant strategy truthful*, that is, for each seller it is a dominant strategy to report his true privacy and risk attitudes. Moreover, our mechanisms are *individually rational*, that is, for every seller his expected utility from participating is at least as large as his expected utility from not participating. In many cases we have the stronger property of *ex post individual rationality*, that is, every seller experiences a non-negative utility at each possible outcome.

In Section 4 we consider the case of linear costs. More general cost functions are considered in Section 5. For each class of cost functions, we first introduce a Baseline Mechanism that is dominant strategy truthful and ex-post individually rational, and then consider a more involved mechanism that can further reduce the price by taking advantage of risk-aversion.

The mechanisms we introduce in the following sections lead to a set of *essentially* unbiased samples in the sense that instead of producing unbiased samples from the set of all sellers N, slightly smaller sets may be used where one or more sellers are excluded in a non-random fashion. Even though this might introduce some bias, the bias will be insignificant when the number of sellers n is large for most statistics of interest. Because we are interested in mechanisms that are dominant strategy truthful in a prior free setting,

this potential but insignificant bias cannot be avoided.

4 Linear Costs

In this section, we consider the class of linear cost functions and refer to the value of $c_i(1)$ by c_i. Thus, the cost function of seller i is $c_i(k) = k \cdot c_i$, and c_i is the single parameter that characterizes the privacy attitude of seller i. The mechanisms we consider aim to elicit the parameter c_i from each seller i.

4.1 Baseline Mechanism for Linear Costs

The Baseline Mechanism consists of the following steps:

(1) Ask every seller i to report the parameter c_i. Let c^{max} denote the highest reported value and $\hat{j}$ the index of the corresponding seller.

(2) For every buyer $b \in B$, produce an unbiased sample of size d_b from $N \setminus \{\hat{j}\}$ and pay each sampled seller c^{max}.

In other words, this mechanism discards the seller that reported the maximum cost and then produces unbiased samples from the set of remaining sellers. There are many ways in which this sampling can be done, that is, there are many distributions in $\Psi(N \setminus \{\hat{j}\})$. One way is to produce independent samples for each buyer b; i.e., for each buyer b, sample d_b sellers from $N \setminus \{\hat{j}\}$ uniformly at random. Alternatively, we could bundle the demand from some buyers together and then do the sampling; e.g., if $d_b = d_{b'}$ for buyers b and b' we could sample uniformly at random once (instead of twice) and give both buyers access to the same sample of sellers.

The price of the Baseline Mechanism for linear costs is c^{max} and does not depend on which distribution $\psi \in \Psi(N \setminus \{\hat{j}\})$ is used for the sampling. This is the minimum price that guarantees all sellers all willing to participate in a setting where the same payment function π applies to each seller, regardless of his risk attitude, and each $\pi(k)$ is deterministic.

More generally, each $\pi(k)$ could be a random variable. For instance, to produce a sample for buyer b, we could produce an unbiased sample of $d_b + 1$ sellers, remove the seller that reported the maximum cost among these sellers from the sample, and use his reported value to compensate the rest d_b sellers. This approach could lead to a lower price than the Baseline Mechanism, but at the same time would increase the bias. In what follows,

we assume that each $\pi_i(k)$ is deterministic, but note that some of our results can be applied more generally.

Observe that with the Baseline Mechanism, a seller may be sampled multiple times (i.e., for multiple buyers in B). The total payment to seller i is equal to the product of c^{max} and the number of buyes that obtain access to i's data. As a result, the utility that seller i derives depends on the number of times that he is sampled. In particular, if seller i is sampled k times he derives utility $u_i(k(c^{max} - c_i))$. If $c_i < c^{max}$, then seller i strictly prefers to be sampled more times.

We next show two important properties of the Baseline Mechanism. First, it is a dominant strategy for each seller to report his cost truthfully. Second, all sellers derive non-negative utility at every possible outcome. These properties imply that all sellers will participate and report their values truthfully.

Theorem 4.1 *The Baseline Mechanism for linear costs is dominant strategy truthful and ex-post individually rational.*

With the Baseline Mechanism, a seller gets a payment of c^{max} from every buyer that gets access to his data. Thus, the total expected payment to a seller is equal to c^{max} times the expected number of buyers that get access to his data. A risk-averse seller would prefer a smaller payment in expectation that is appropriately distributed between potential outcomes. In the following section, we introduce the *Certainty Equivalent (CE) Mechanism* that uses this fact to reduce the expected price that buyers have to pay.

4.2 CE Mechanism for Linear Costs

The Certainty Equivalent (CE) Mechanism takes as input a distribution $\psi \in \Psi(N')$ that produces unbiased samples $\{s_b, b \in B\}$ of sizes $\{d_b, b \in B\}$ from a set of sellers N'. Given ψ, we use $p_\psi(k)$ to denote the probability that the data of a seller in N' is sold to exactly k buyers. Moreover, let $w \equiv \sum_k p_\psi(k)k$ be the expected number of times that a given seller is sampled; the following lemma shows that this value is the same for any $\psi \in \Psi(N')$.

Lemma 4.2 *Let n' denote the number of sellers in N'. If $\psi \in \Psi(N')$, then $\sum_{k=0}^{m} k p_\psi(k) = \sum_{b \in B} d_b / n'$.*

Identically to the Baseline Mechanism, the first step of the CE Mechanism is to ask each seller to report his cost parameter. The second step

is to ask each seller to report his certainty equivalent for a lottery that corresponds to the Baseline Mechanism, where the samples are produced according to distribution ψ (which is the input to the mechanism) and each seller is paid c^{max}, i.e., the maximum reported cost, every time he is sampled.

Then, the CE Mechanism determines how each seller will be compensated as a function of the number of times that he is sampled (Step 3). A seller that reported a large certainty equivalent in Step 2 will be compensated as in the Baseline Mechanism; i.e., he will receive a payment equal to the number of times he is sampled times c^{max}. On the other hand, if a seller reported a small certainty equivalent in Step 2 and is sampled k times, he will receive more than $c^{max}k$ when k is small and less than $c^{max}k$ when k is large.

Finally, excluding the seller that reported the maximum cost, the CE Mechanism produces unbiased samples from the remaining set of sellers according to distribution ψ, as is the case with the Baseline Mechanism. However, in contrast to the Baseline Mechanism, each seller is paid according to the corresponding payment function that was determined in Step 3. If some sellers are risk-averse, the total expected payment (across all sellers) with the CE Mechanism is lower than the total expected payment with the Baseline Mechanism. As a result, the CE Mechanism achieves a lower price.

We formally describe the steps of the CE Mechanism for a given distribution $\psi \in \Psi(N')$ below:

(1) Ask every seller i to report the parameter c_i. We denote the reported values by $\hat{c}_i$. Define $c^{max} \equiv \max\{\hat{c}_i\}$ and $\hat{j} \equiv \arg\max\{\hat{c}_i\}$.[6]

(2) Define $\pi^{max}(k) \equiv c^{max} \cdot k$. Ask every seller i to report his certainty equivalent for (p_ψ, π^{max}). Let $\hat{e}_i$ denote the reported value.

(3) Define $\hat{r}_i \equiv \hat{e}_i + \hat{c}_i w$, $f_{-i}(r) \equiv |\{j \in N \setminus \{i, \hat{j}\} : \hat{r}_i \leq r\}|$ and $\overline{r}_i \equiv \arg\max\{f_{-i}(r)(c^{max}w - r)\}$. For every seller i: If $\hat{r}_i < \overline{r}_i$, set $\pi_i(k) \equiv \overline{r}_i + \hat{c}_i(k - w)$; otherwise set $\pi_i(k) \equiv \pi^{max}(k)$ for $k = 0, 1, ..., m$.

(4) Produce unbiased samples from $N' = N \setminus \{\hat{j}\}$ with the distribution ψ. Pay each seller i according to π_i; that is, if sampled exactly k times seller i receives a payment of $\pi_i(k)$.

We conclude the description of the CE Mechanism by explaining Step 3 in more detail. Observe that the expected payment to each seller from

[6] It is straightforward to deal with ties, but we assume that only one seller reports a value equal to c^{max} to simplify exposition.

lottery (p_ψ, π^{max}) is $c^{max}w$. The market-maker wishes to reduce this amount for some risk-averse sellers by offering them a less risky lottery that results in a lower expected payment.

Seller i has reported that he is indifferent between the lottery (p_ψ, π^{max}) and a fixed payment $\hat{e}_i$. Thus, by Assumption 3.1, he is also indifferent between the lottery (p_ψ, π^{max}) and the lottery $(p_\psi, \tilde{\pi})$ where $\tilde{\pi}(k) \equiv \hat{e}_i + \hat{c}_i k$; in the latter lottery the sampling is still done with distribution ψ but seller i always gets compensated by the fixed amount $\hat{e}_i$ plus his cost. The expected payment from lottery $(p_\psi, \tilde{\pi})$ is $\hat{r}_i \equiv \hat{e}_i + \hat{c}_i w$.

Seller i strictly prefers a lottery where he is sampled according to ψ and paid $r > \hat{r}_i$ in expectation if r is appropriately distributed; in particular, if paid $r + \hat{c}_i(k - w)$ when sampled k times. The market-maker can set a threshold r and give the payment π^{max} only to sellers with $\hat{r}_i > r$, while giving all other sellers an appropriately distributed expected payment of r. The lower the threshold r, the lower the expected payment to each seller that is below the threshold, but also the larger the number of sellers that are paid π^{max}.

The market-maker wishes to maximize the decrease in total expected payment (over all sellers) compared to a setting where all sellers are paid by π^{max}. The market-maker could then choose the threshold r that maximizes $f(r)(c^{max}w - r)$, where $f(r) \equiv |\{j \in N \setminus \{\hat{j}\} : \hat{r}_i \leq r\}|$ is the number of sellers below the threshold r. This approach can be used in a setting where sellers are expected to be non-strategic when reporting their certainty equivalents in Step 2, but does not guarantee truthful reporting when sellers are strategic.

To guarantee that it is a dominant strategy for each seller to report his certainty equivalent truthfully, the market-maker maximizes $f_{-i}(r)(c^{max}w - r)$ when setting the threshold for seller i, where f_{-i} is determined by all sellers other than i. In this way, the threshold used for seller i does not depend on the certainty equivalent value he reported. We note that in some instances this approach sets different thresholds for different sellers and may result in a suboptimal solution where the total expected payment (over all sellers) is not minimized. However, this is something that cannot be avoided in general while guaranteeing truthful reporting.

We next show that the CE Mechanism has the desirable properties of dominant strategy truthfulness and individual rationality.

Theorem 4.3 *If every seller is either risk-averse or risk-neutral, the CE Mechanism for linear costs is dominant strategy truthful and ex-post individually rational. If some sellers are risk-seeking, a variation of the CE Mechanism for linear costs is dominant strategy truthful and individually*

rational.

The CE Mechanism takes as input a distribution $\psi \in \Psi(N')$ that for each buyer $b \in B$ produces an unbiased sample s_b of size d_b from the set N'. Since there are many such distributions, we are interested in the one that results in the lowest price for the buyers. This is in contrast to the Baseline Mechanism for linear costs, where the price is the same (and equal to c^{max}) for any distribution.

We identify the distribution that minimizes the certainty equivalent for $(p_\psi, \pi^{\max})$ for a risk averse seller over all distributions $\psi \in \Psi(N')$. By minimizing the certainty equivalents that risk-averse sellers report in Step 2 of the CE Mechanism, this distribution has the potential of minimizing the expected payment per seller and achieving an approximately optimal price.[7]

We next define the *ordering distribution*, which randomly orders the sellers once and then uses this ordering to determine the sample for each buyer. The earlier a seller is in the ordering the more samples he will be in.

Definition Given a set of sellers N', the *ordering distribution* $\psi^*(N')$ produces unbiased samples of sizes $\{d_b, b \in B\}$ as follows: First, order sellers in N' uniformly at random once. Then, for each buyer $b \in B$ return a sample that consists of the first d_b sellers in the ordering.

The ordering distribution produces unbiased samples from N'; thus, $\psi^*(N') \in \Psi(N')$.

Observe that if two buyers request samples of the same size, then the ordering distribution will give them the same sample of sellers. In this sense, the ordering distribution is bundling buyers' demand. In the extreme case that all buyers demand samples of the same size (i.e., $d_b = d_{b'}$ for all $b, b' \in B$), the ordering distribution will effectively produce one unbiased sample of sellers in N'. The following example illustrates how the ordering distribution works in the case that there is more variability in buyers' demand.

Example Suppose that N' consists of n' sellers. There are 200 buyers: 50 buyers choose $d_b = 100$ and 150 buyers choose $d_b = 200$. With the ordering distribution $\psi^*(N')$, seller i will be randomly assigned a position in $\{1, 2, ..., n'\}$. If seller i gets assigned a position in $\{1, 2, ..., 100\}$ (which occurs with probability $100/n'$), his data will be sold to all 200 buyers. If seller i gets assigned a position in $\{101, 102, ..., 200\}$ (which occurs with probability

[7] The price is minimized in settings that $f_{-i}(r)(c^{max}w - r)$ is maximized at the same r for all $i \in N$.

$100/n'$), his data will be sold to 150 buyers. For all other positions, his data will not be sold to any buyers.

We next show that for any concave non-decreasing function $g : \mathbb{N} \to \mathbb{R}$, the expected value $\sum_{k=0}^{m} p_\psi(k)g(k)$ is minimized at the ordering distribution among all distributions in $\Psi(N')$. This implies that the ordering distribution minimizes the certainty equivalent that a risk-averse seller will report in the CE mechanism. In other words, the ordering distribution minimizes the expected payment for which a risk-averse seller is willing to participate in the market, when the CE Mechanism is used.

Theorem 4.4 *If $g : \mathbb{N} \to \mathbb{R}$ is concave and non-decreasing, then $G(\psi) \equiv \sum_{k=0}^{m} p_\psi(k)g(k)$ is minimized at $\psi = \psi^*$ among all distributions $\psi \in \Psi(N')$.*

Corollary 4.5 *Set $c^{max} > c_i$ and $\pi^{max}(k) \equiv c^{max}k$. If u_i is concave, then the certainty equivalent $e_i(p_\psi, \pi^{\max})$ is minimized at $\psi = \psi^*$ among all distributions $\psi \in \Psi(N')$.*

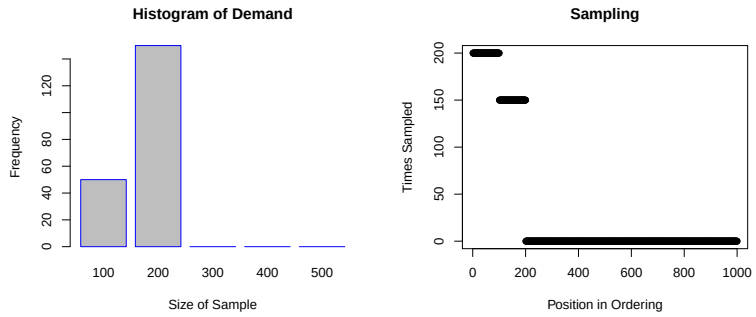
4.3 Examples

We have shown that the ordering distribution is optimal with respect to minimizing the certainty equivalent of a risk-averse seller for the lottery that corresponds to the Baseline Mechanism. In this section we illustrate how much the price that a buyer has to pay per seller in his sample may decrease when we use the ordering distribution with the CE Mechanism.

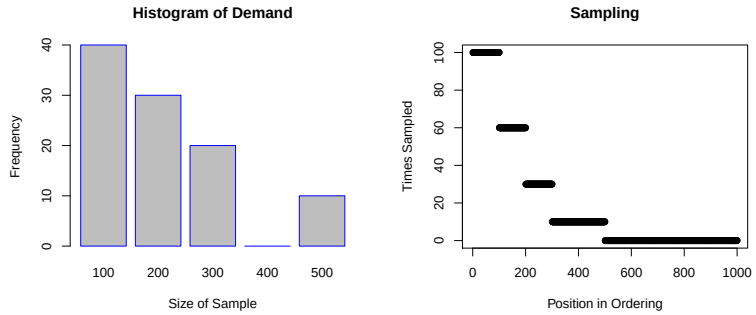
Baseline Mechanism. Suppose that $c^{max} = 10$. This means that with the Baseline Mechanism the price is equal to \$10; that is, every time a seller is sampled he is paid \$10. Equivalently, a buyer that gets access to a sample of d_b sellers pays $10d_b$ in total, or equivalently \$10 per seller.

Suppose that out of $m = 200$ buyers, 50 have asked for samples with $d_b = 100$ sellers and 150 have asked for samples with $d_b = 200$ sellers. Furthermore, assume that when we remove the seller who reported that his cost is equal to c^{max}, we have 1000 sellers left: half of these sellers are risk-neutral or risk-seeking and the other half that are risk-averse with $u_i(x) = 1 - e^{-0.002x}$ and $c_i = 0$.

CE Mechanism with ordering distribution. If the ordering distribution is used for the sampling, then the number of buyers that get access to the data of a specific seller may be 200, 100 or 0. Using the ordering distribution with the CE Mechanism yields a price of \$6.53 per seller, which is significantly lower than the price of the Baseline Mechanism.



(a) Demand histogram and sampling of ordering distribution when 50 buyers have chosen $d_b = 100$ and 150 buyers have chosen $d_b = 200$.



(b) Demand histogram and sampling of ordering distribution when 40 buyers have chosen $d_b = 100$; 30 have chosen $d_b = 200$; 20 have chosen $d_b = 300$; and 10 have chosen $d_b = 500$.

Figure 2: Two examples of buyer requests and the corresponding sampling of the ordering distribution.

CE Mechanism without bundling. To demonstrate the importance of the ordering distribution, we now show that using a different distribution with the CE Mechanism may result in a price that is very close to the price of the Baseline Mechanism. In particular, if the samples of different buyers are independent, the resulting price is \$9.96; that is, even though there is a decrease in price compared to the Baseline Mechanism because of risk-aversion, the effect is very small.

The certainty equivalent of seller i for the lottery (p_{ψ^*}, π^{max}) depends on the buyers' demand, the payment c^{max} and the number of sellers n. The market-maker can easily communicate the values c^{max} and n to the seller. One way of communicating the buyers' demand is through the histogram of sample sizes that buyers have requested.

Alternatively, the market-maker can help a seller determine his certainty equivalent for the lottery (p_{ψ^*}, π^{max}) by giving him a graph that represents how many times the seller's data will be sold as a function of his position in the ordering. Knowing that each position is equally likely and that he will be paid c^{max} every time he is sampled, the seller can determine his certainty equivalent. Figure 2(a) shows (i) the histogram of buyers' demand and (ii) the number of times sampled as a function of the position in the ordering for the example discussed above. Figure 2(b) shows the same plots for a different distribution of buyers' demand.

In practice, we expect that the market-maker will give buyers predefined sample size options to choose from. As a result, the set of distinct values of sample sizes representing buyers' requests will be small (similarly to the examples in Figure 2) and it will be relatively easy for a seller to determine his certainty equivalent.

5 General Costs

In the previous section we considered the case of linear cost functions and introduced two mechanisms, the Baseline Mechanism and the CE Mechanism, that the market-maker can use to facilitate interactions between buyers and sellers. In this section we extend these mechanisms to a setting with general cost functions and show that the ordering distribution has good properties in this more general setting.

Even though the mechanisms we introduce can be used for any cost functions, a specific class of interest is that of concave functions. We believe that concavity is a realistic property for a cost function c_i, because we expect that a seller's cost for providing access to exactly the *same data* k times

should not be more than k times the cost of allowing access once.

We first note that for a special class of concave cost functions, the mechanisms of Section 4 can be applied with only minor modifications. In particular, this is the case if the privacy attitude of seller i is represented by the function $c_i(k) = c_i \cdot h(k)$, where h is an increasing concave function with $h(0) = 0$. Then, as in the case of linear costs, the market-maker can ask each seller to report his single parameter, c_i, and the payment π^{max} is determined by the maximum reported parameter. In this case, the ordering distribution is still optimal for the CE Mechanism.

In the more general case where sellers have arbitrary cost functions, three complications may arise compared to the linear case. First, sellers cannot be ordered based on their costs; in particular, it could be that $c_i(k) > c_j(k)$ and $c_i(k') < c_j(k')$ for $k \neq k'$.[8] Still, we can define the payment function $\pi^{max}(k) \equiv \max_i c_i(k)$ to guarantee that every seller is willing to sell access of his data to k buyers if paid $\pi^{max}(k)$.

Second, seller i may need to report m values, i.e., $\hat{c}_i(k)$ for $k = 1, 2, ..., m$, in order to communicate his cost function to the mechanism. However, we can significantly reduce the number of values that a seller needs to report if (1) the sampling is based on bundling buyers' demand (e.g., if we use the ordering distribution) and (2) there are relatively few different sample sizes requested by buyers, e.g., because buyers choose from predefined sample size options. Thus, even though for completeness our mechanisms ask each seller to report his whole cost function, this will not be necessary in many applications.

The third potential complication is that $\pi^{max}(k)$ may not be linear and, as a result, the total expected payment over all sellers and the resulting price per seller are affected by what distribution is used for the sampling — even for the Baseline Mechanism. In Section 5.1 we show that the ordering distribution minimizes the price per seller of the Baseline Mechanism when π^{max} is concave. Moreover, if each c_i is concave but π^{max} is not, the ordering distribution provides optimal price guarantees.

Then, in Section 5.2 we discuss how the CE Mechanism can be generalized for a setting with arbitrary cost functions, and show that this mechanism is dominant strategy truthful and individually rational.

[8] For instance, this is the case if $c_i(k) = k$ and $c_j(k) = 2\sqrt{k}$.

5.1 Baseline Mechanism for General Costs

Consider the following generalization of the Baseline Mechanism for linear cost functions (Section 4.1) to a setting with arbitrary costs. First, each seller i is asked to report the values $c_i(k)$ for $k = 1, 2, ..., m$. We denote the reported values by $\hat{c}_i(k)$ and set $\pi^{max}(k) \equiv \max_i \hat{c}_i(k)$. Then, for each buyer $b \in B$ the mechanism produces a of size d_b. A seller who is sampled exactly k times receives a payment of $\pi^{max}(k)$.

In the Baseline Mechanism for linear costs, the seller that reported the maximum cost was excluded from the sampling in order to guarantee truthfulness. With arbitrary costs, different sellers may correspond to the maximum cost for different values of k. In the case of arbitrary costs, the Baseline Mechanism is dominant strategy truthful if seller i with $\hat{c}_i(k) = \pi^{max}(k)$ is *not* sampled k times. Of course this introduces some bias, but the bias is negligible in many settings of interest, e.g., if the number of sellers is significantly larger than the maximum sample size requested by buyers.

In Section 5.2, we discuss how the sampling can be done in a way that guarantees truthfulness, focusing on a variation of the ordering distribution. However, for the purposes of this section we ignore this issue and assume for ease of exposition that the mechanism produces unbiased samples from the set of n sellers.

We are thus interested in the distribution $\psi \in \Psi(N)$ that minimizes the expected price of the Baseline Mechanism. Equivalently, we are interested in which distribution $\psi \in \Psi(N)$ minimizes the expected payment per seller when payments are given by π^{max}. In other words, we are interested in minimizing

$$\Pi(\psi) \equiv \sum_{k=0}^{m} p_\psi(k)\pi^{max}(k).$$

In the case of linear costs, π^{max} is linear and, as a result, $\Pi(\psi)$ obtains the same value for every $\psi \in \Psi(N)$; this follows from Lemma 4.2. With arbitrary cost functions, π^{max} may not be linear and the value of $\Pi(\psi)$ may be different for different $\psi \in \Psi(N)$. Theorem 4.4 implies that the ordering distribution ψ^* is optimal in the special case that π^{max} is concave. Note, however, that concavity of c_i for all sellers i does *not* imply concavity of π^{max}.

We now turn to the case that π^{max} is not concave. Even though we are interested in distributions that are oblivious of π^{max}, as a benchmark we consider the minimum value of $\Pi(\psi)$ that can be attained when we choose

a distribution $\psi \in \Psi(N)$ *knowing* π^{max}; denote this value by Π^{OPT}.[9] The following theorem shows that a distribution that is oblivious to the values of π^{max} cannot guarantee a better than 2 approximation factor of Π^{OPT}.

Theorem 5.1 *No distribution $\psi \in \Psi(N)$ that is oblivious to the payment function π^{max} can guarantee that $\Pi(\psi) \leq (2-\epsilon)\Pi^{OPT}$ for $\epsilon > 0$. This holds even if for each seller i the function c_i is concave.*

We have shown that a π^{max}-oblivious distribution cannot approximate Π^{OPT} within a factor better than 2. The following theorem shows that the ordering distribution ψ^* actually guarantees an approximation factor of 2. Thus, the ordering distribution achieves the best possible worst-case guarantee within the class of distributions that are not aware of the function π^{max} a priori.

Theorem 5.2 *If for each seller $i \in N$ the cost function c_i is concave, then*

$$\Pi(\psi^*) \leq 2\Pi^{OPT}.$$

5.2 CE Mechanism for General Costs

In the case of linear costs, the CE Mechanism takes as input a distribution $\psi \in \Psi(N')$ that produces unbiased samples $\{s_b, b \in B\}$ of sizes $\{d_b, b \in B\}$ from a set of sellers N'. In the case of arbitrary costs that we consider in this section, the CE Mechanism takes as input a distribution ψ that produces samples $\{s_b, b \in B\}$ of sizes $\{d_b, b \in B\}$ such that a seller i with $\hat{c}_i(k) = \pi^{max}(k)$ is *not* sampled k times, while introducing negligible bias.

Given such a distribution ψ, the CE Mechanism first asks each seller i to report his cost function c_i for the range $\{1, 2, ..., m\}$; we denote the reported values $\hat{c}_i(k)$ and define $\pi^{max}(k) \equiv \max_i \hat{c}_i(k)$. Then, the mechanism for general costs proceeds similarly to steps (2), (3) and (4) of the CE Mechanism for linear costs, presented in Section 4.2. In particular, each seller is asked to report his certainty equivalent for (p_ψ, π^{max}) (Step 2). Then, the mechanism uses the reported certainty equivalent values to determine the payment function π_i for each seller i to either be equal to π^{max} or such that $\pi_i(k) - \hat{c}_i(k)$ is a constant (Step 3); this step is further described below. Finally, the sampling occurs according to ψ and seller i is paid $\pi_i(k)$ if sampled k times (Step 4).

[9] Π^{OPT} is generally unattainable by a distribution that is oblivious to π^{max}. Moreover, it is unrealistic to assume that the mechanism will choose the distribution for the sampling as a function of the values that sellers report because then π^{max} cannot in general be elicited truthfully.

Step 3 of the CE Mechanism for linear costs can be adapted for the setting of general costs to determine the payment function π_i for each seller i whose reported values in Step 1 do not affect π^{max}, which is the case if $\hat{c}_i(k) \leq \max_{j \neq i} \hat{c}_j(k)$ for all k. In particular, let $\hat{r}_i \equiv \hat{e}_i + \sum_k p_\psi(k)\hat{c}_i(k)$, $f_{-i}(r) \equiv |\{j \in N \backslash \{i\} : \hat{r}_i \leq r\}|$ and $\overline{r}_i \equiv \arg\max\{f_{-i}(r)(\sum_k p_\psi(k)\pi^{max}(k) - r)\}$. Then, if $\hat{r}_i < \overline{r}_i$, set $\pi_i(k) \equiv \overline{r}_i + \hat{c}_i(k) - \sum_{k'} p_\psi(k)\hat{c}_i(k)$; otherwise set $\pi_i(k) \equiv \pi^{max}(k)$ for $k = 0, 1, ..., m$.

The aforementioned approach for determining the payment functions π_i does not guarantee truthful reporting from seller i if $\hat{c}_i(k) > \max_{j \neq i} \hat{c}_j(k)$ for some k. In particular, by reporting a higher value of $\hat{c}_i(k)$ he increases $\pi^{max}(k)$ which may increase the certainty equivalent values of other sellers for (p_ψ, π^{max}). This in turn may increase the threshold $\overline{r}_i$. As a result, seller i may be better off if assigned the payment that yields the same value $\pi_i(k) - \hat{c}_i(k)$ for all k. We next describe how to deal with this issue.

Consider a seller i for whom $\hat{c}_i(k') > \max_{j \neq i} \hat{c}_j(k')$ for some k'. Define $\pi^{max}_{-i}(k) \equiv \max_{j \neq i} \hat{c}_i(k)$. That is, π^{max}_{-i} is the payment function that the Baseline Mechanism would use if seller i did not participate. Suppose that the mechanism asks each seller $j \neq i$ to report his certainty equivalent for (p_ψ, π^{max}_{-i}), and then uses these reported values to determine the threshold $\overline{r}_i$ for agent i in Step 3. This guarantees that it is a dominant strategy for seller i to report $c_i(k)$ truthfully.

The following theorem shows that when sellers have general cost functions and the market-maker uses the CE Mechanism described above, sellers report their values truthfully.

Theorem 5.3 *The CE Mechanism for general costs is dominant strategy truthful and individually rational.*

The CE Mechanism for general costs takes as input a distribution ψ that produces samples $\{s_b, b \in B\}$ of sizes $\{d_b, b \in B\}$ such that a seller i with $\hat{c}_i(k) = \pi^{max}(k)$ is *not* sampled k times. We now describe how this can be achieved with a variation of the ordering distribution. First, order sellers in N uniformly at random once. Observe that if for each buyer $b \in B$ we return a sample that consists of the first d_b sellers in the ordering, then the j-th seller in the ordering will be sampled $\sum_{b \in B} 1_{\{d_b \geq j\}}$ times.

To guarantee that a seller is not sampled k times if his reported value $\hat{c}_i(k)$ has determined the payment $\pi^{max}(k)$, we repeat the following process for the seller that is earliest in the ordering and has not yet been checked (starting with the first seller in the ordering). Suppose we are checking for the j-th seller in the ordering, whom we denote as seller ℓ. Let $k =$

$\sum_{b \in B} 1_{\{d_b \geq j\}}$. If $\hat{c}_\ell(k) > \max_{i \neq \ell} \hat{c}_i(k)$, then remove seller ℓ from the ordering and move everyone after him one position ahead. We repeat until all sellers have been checked and then for each buyer $b \in B$ we return a sample that consists of the first d_b sellers in the final ordering.

Similarly to the ordering distribution that we use in the case of linear costs, the distribution described above introduces some bias because certain sellers may be removed from the ordering. We note that it is not possible to completely eliminate the bias in a prior free setting while guaranteeing that all sellers will be truthful. Moreover, the bias will be very small if the number of sellers is much larger than the sample sizes that buyers request.

6 Conclusion

In this paper we studied a market for private data where buyers can obtain access to unbiased samples of private data by appropriately compensating the individuals to whom the data corresponds (the sellers). A market-maker facilitates the interactions between the two sides of the market. We focussed on how bundling the buyers' demand can decrease the price that buyers have to pay per individual, while ensuring that sellers are willing to participate.

Throughout the paper we took a prior-free approach and assumed no knowledge of the distribution of the seller's privacy and risk attitudes. We then constructed mechanisms that the market-maker can use to elicit sellers' privacy and risk attitudes truthfully, and showed that our mechanisms provide optimal price guarantees in several different settings.

To derive our formal results we assumed that the demand is price-insensitive and known by the market-maker. That is, buyer b is interested in obtaining access to an unbiased sample of d_b individuals regardless of the price he has to pay per individual. Given the distribution he will use for producing the samples, the market-maker elicits sellers' preferences with respect to two different pricing schemes: the first is risky, the second one is not but yields a lower expected payment. The sellers' choices determine the price. Since demand is assumed to be price-insensitive, each buyer b will still be willing to obtain an unbiased sample of size d_b for the derived price.

More generally, the size of the unbiased sample that a buyer may want to get access to could be a function of the price. In that case, we get a "cycle": for a fixed price the market-maker can learn the buyers' demand; on the other hand, for fixed demand the market-maker can use our bundling mechanisms to derive a good price for the buyers while ensuring that sellers are willing to participate in the market. If the derived price gives rise to

the same demand that the market-maker started with in order to derive the price (as in the case when demand is price-insensitive), then the market clears.

We note that there always exists a price at which the market clears, even if the demand if price-sensitive; for instance, this is the case for a price corresponding to our Baseline Mechanism. An open question is under what conditions, e.g., in terms of how demand depends on the price, a lower such price exists with the market-maker taking advantage of the risk aversion of some sellers. A related question is what processes the market-maker could use to converge to such a low price.

The "cycle" that arises in our market for private data distinguishes it from standard markets, where for a fixed price both demand and supply can be determined and the market clears if demand meets supply. Our setting is different because (1) buyers are interested in obtaining unbiased samples and, as a result, the market-maker needs to make sure that all sellers are willing to participate, and (2) the market-maker tries to take advantage of the inherently randomized nature of sampling and the risk aversion of some sellers to find a lower price (in expectation) per individual, rather than the one that the most privacy-concerned sellers require.

In this paper, we chose to "break the cycle" by assuming that demand is known and price-insensitive. In addition to price-insensitive settings, this is also a reasonable assumption for settings where demand does not change drastically with the price and/or the market-maker has a good estimate of the right range for the price (e.g., from past experience).

Alternatively, the market-maker could "break the cycle" by relying on sellers' beliefs about demand — instead of explicitly giving sellers information on demand as in Figure 2 — when eliciting the certainty equivalents. The mechanisms discussed in this paper would still work in this case. However, a potential drawback of relying on sellers' beliefs on demand is that the seller experience could be less simple.

Markets for private data such as the one we presented are quite realistic and in principle easy to implement. Given the great value of big data and the clamoring from the general public for a certain degree of control over its trading, it is not unreasonable to expect that such markets will become operational, thus benefitting both the sellers and buyers of big data.

References

[1] Alessandro Acquisti, Leslie John, and George Loewenstein. What is privacy worth? In *Twenty First Workshop on Information Systems and Economics (WISE)*, 2009.

[2] Christina Aperjis and Bernardo A. Huberman. A market for unbiased private data: Paying individuals according to their privacy attitudes. *First Monday*, 17(5), May 2012.

[3] Dan Cvrcek, Marek Kumpost, Vashek Matyas, and George Danezis. A study on the value of location privacy. In *Proceedings of Workshop on Privacy in the Electronic Society*, pages 109–118, 2006.

[4] Pranav Dandekar, Nadia Fawaz, and Stratis Ioannidis. Privacy auctions for inner product disclosures. *arXiv*, 1111.2885, 2011.

[5] Arpita Ghosh and Aaron Roth. Selling privacy at auction. In *ACM Conference on Electronic Commerce*, pages 199–208, 2011.

[6] Hamed Haddadi, Richard Mortier, and Steven Hand. Privacy analytics. *SIGCOMM Comput. Commun. Rev.*, 42(2):94–98, 2012.

[7] Il-Horn Hann, Kai-Lung Hui, Sang-Yong Tom Lee, and Ivan P.L. Png. Overcoming online information privacy concerns: An information-processing theory approach. *Journal of Management Information Systems*, 24(2):13–42, 2007.

[8] Charles A. Holt and Susan K. Laury. Risk aversion and incentive effects. *American Economic Review*, 92:1644–1655, 2002.

[9] Bernardo A. Huberman, Eytan Adar, and Leslie R. Fine. Valuating privacy. *Security Privacy, IEEE*, 3(5):22 – 25, 2005.

[10] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

[11] Christopher Riederer, Vijay Erramilli, Augustin Chaintreau, Pablo Rodriguez, and Balachander Krishnamurthy. For sale: Your data By: You. In *HotNets*, 2011.

[12] Aaron Roth and Grant Schoenebeck. Conducting truthful surveys, cheaply. In *ACM Conference on Electronic Commerce*, pages 826–843, 2012.

[13] Natasha Singer. Congress to examine data sellers. *The New York Times*, July 2012.

[14] Natasha Singer. You for sale: Mapping, and sharing, the consumer genome. *The New York Times*, June 2012.

Appendix: Proofs

Proof of Theorem 4.1: Consider some seller i and first suppose that $c_i < \max_{j \neq i} \hat{c}_j$. We observe that reporting any $\hat{c}_i < \max_{j \neq i} \hat{c}_j$ will not make a difference in the utility of seller i regardless of the outcome of the sampling; furthermore, he will derive a strictly positive utility every time seller i is sampled. On the other hand, if he reports $\hat{c}_i > \max_{j \neq i} \hat{c}_j$, then seller i will be excluded from the sampling and derive zero utility. The second case to consider is that $c_i > \max_{j \neq i} \hat{c}_j$. Then, by reporting $\hat{c}_i = c_i$, seller i's utility is equal to zero. However, there is no way of getting positive utility in this case. In particular, by reporting $\hat{c}_i < \max_{j \neq i} \hat{c}_j$, seller i will get negative utility whenever he is sampled.

Thus, reporting $\hat{c}_i \neq c_i$ can never increase the utility of seller i but in some circumstances may actually decrease it. This shows that truthful reporting is a dominant strategy for each seller. To show ex-post individual rationality, we observe that by reporting $\hat{c}_i = c_i$ seller i gets a positive utility whenever sampled and zero utility otherwise. ∎

Proof of Lemma 4.2: Let Z_i be a random variable that denotes the number of times seller i is sampled. We have that $\sum_{i=1}^{n} Z_i = \sum_{b \in B} d_b$ in order to meet the demand. Observe that the expected number of times that seller i is sampled under distribution ψ is $\mathbb{E}[Z_i] = \sum_{k=0}^{m} k p_\psi(k)$. Since $\psi \in \Psi$, this distribution produces unbiased samples, which implies that each seller is sampled the same expected number of times. Thus, summing over all sellers,

$$n' \sum_{k=0}^{m} k p_\psi(k) = \sum_{i=1}^{n} \mathbb{E}[Z_i] = \sum_{b \in B} d_b,$$

which concludes the proof.

Proof of Theorem 4.3: If the payment will be determined by the function π^{max}, Theorem 4.1 implies that it is a dominant strategy for seller i to report c_i truthfully and that we get ex-post individual rationality. For a seller i with $c_i < \max_{j \neq i} \hat{c}_j$, it is a dominant strategy to also report his certainty equivalent for (p_ψ, π^{max}) truthfully in Step (2), because his report $\hat{e}_i$ does not affect the threshold $\overline{r}_i$. Finally, if the payment is determined to be $\pi_i(k) \equiv \overline{r}_i + \hat{c}_i(k - w)$ and seller i has reported $\hat{c}_i$ truthfully, we have ex-post individual rationality because $\pi_i(k) - c_i k = \overline{r}_i - c_i w > \hat{r}_i - c_i w = \hat{e}_i > 0$.

We now turn to the seller i with $c_i > \max_{j \neq i} \hat{c}_j$. By reporting any value $\hat{c}_i > \max_{j \neq i} \hat{c}_j$, the seller will not be sampled and gets utility zero. By reporting $\hat{c}_i < \max_{j \neq i} \hat{c}_j < c_i$, seller i gets a negative utility if assigned

payment π^{max} in Step 3. On the other hand, if assigned the payment $\pi_i(k) \equiv \overline{r}_i + \hat{c}_i(k-w)$, seller i derives utility $u_i(\overline{r}_i - \hat{c}_i w - (c_i - \hat{c}_i)k)$ which may be positive for small values of k.

We now show that if seller i is risk-neutral or risk-averse, i.e., not risk-seeking, he derives negative utility in expectation. In particular, we have $\overline{r}_i - \hat{c}_i w - (c_i - \hat{c}_i)k < (c_i - \hat{c}_i)(w-k)$. Thus, $\sum_k p_\psi(k)(\overline{r}_i - \hat{c}_i w - (c_i - \hat{c}_i)k) < \sum_k p_\psi(k)(c_i - \hat{c}_i)(w-k) = 0$. And since u_i is concave or linear, $\sum_k p_\psi(k) u_i(\overline{r}_i - \hat{c}_i w - (c_i - \hat{c}_i)k) < \sum_k p_\psi(k) u_i(c_i - \hat{c}_i)(w-k) = 0$.

To conclude the proof, we consider the case that the seller i with $c_i > \max_{j \neq i} \hat{c}_j$ is risk-seeking. Then, it is plausible that seller i is better off reporting $\hat{c}_i < c_i$ in order to get utility $u_i(\overline{r}_i - \hat{c}_i w - (c_i - \hat{c}_i)k)$ which is positive for small values of k, but negative for large values. Even though such preferences are very unlikely, for the sake of completeness we describe how our CE mechanism can be extended to deal with this issue for the sake of completeness.

To avoid such situations, the mechanism can ask each seller $j \neq i$ to report his certainty equivalent for the lottery (p_ψ, π_{-i}^{max}), where $\pi_{-i}^{max} \equiv \max_{j \neq i} \hat{c}_j$, and use these values to determine the threshold $\overline{r}_i$ for seller i. Then, seller i will be included in the sampling only if $\hat{r}_i < \overline{r}_i$. This guarantees truthful reporting and individual rationality for seller i. Moreover, each seller $j \neq i$ has no reason to lie about his certainty equivalent for (p_ψ, π_{-i}^{max}).[10] ∎

Proof of Theorem 4.4: We use n' to denote the number of sellers in N'. What follows is described from the perspective of some arbitrary seller i. Let e_b denote the event that seller i is sampled by buyer b who requested a total of d_b sellers in his sample, and let $E = \{e_b \mid b \in B\}$ denote the set of all these events. Also, let S denote the set of all subsets of E (i.e., the powerset of E), and let $S_b \subseteq S$ denote the set of all such subsets that include event e_b. We sort all buyers in a non-increasing order of the sample sizes that they request, i.e. $d_b \geq d_{b'}$ if $b < b'$, and define $z_b \equiv d_b/n$.

Given a distribution $\psi \in \Psi(N')$, let $q_\psi(s)$ denote the probability that *exactly* the events in set s occur; that is, seller i's data is sold to all buyers in s but nobody else.[11] Since q_ψ is a distribution over the sets of S, we have

[10] A potential issue here is that seller j might not put in the effort needed for quantifying this certainty equivalent value (because his utility is not affected in any way by his report) and, as a result, not report the correct value. We can avoid this by not telling seller j which lottery each question corresponds to and/or by adding artificial questions about certainty equivalents of lotteries.

[11] Note that q_ψ is different than p_ψ which is a distribution over the number of times that seller i will be sampled.

that $\sum_{s \in S} q_\psi(s) = 1$. Since $\psi \in \Psi(N')$, the probability that the outcome s includes event e_b (or $s \in S_b$) must be exactly equal to z_b. Equivalently, for each buyer b, $\sum_{s \in S_b} q_\psi(s) = z_b$.

The ordering distribution ψ^* satisfies the following simple predicate: For every $b \in \{2, ..., |B|\}$, event e_b may take place only if event e_{b-1} also takes place; equivalently, $p_{\psi^*}(s) = 0$ for every outcome $s \in (S_b \setminus S_{b-1})$. Our goal is to show that $G(\psi)$ is minimized at ψ^* over all $\psi \in \Psi(N')$.

We use proof by contradiction. Assume that $G(\psi) < G(\psi^*)$ for some distribution $\psi \in \Psi(N')$ that does not satisfy the predicate defined above. Let b be the first buyer in the ordering for which this is not true, i.e., the probability distribution assigns positive probability to an outcome s_A that contains e_b but does not contain e_{b-1} ($s_A \in (S_b \setminus S_{b-1})$). We will show that gradually modifying distribution ψ until it satisfies the predicate can be achieved without any increase of the expected value G in the process (if g is *strictly* concave, then this modification actually leads to a decrease in G).

Let $q_A \equiv q(t_A) > 0$ be the probability of outcome s_A. Since $z_b \leq z_{b-1}$, there must also exist some outcome s_B that contains e_{b-1} but not e_b, and occurs with some positive probability $q_B \equiv q(s_B) > 0$. Define $q_{\min} \equiv \min(q_A, q_B) > 0$. Let s_I (resp., s_U) denote the outcome that contains exactly the *intersection* (resp., *union*) of the events contained in s_A and s_B. We modify ψ by removing probability mass $q_{\min}$ from s_A and s_B and moving it to outcomes s_I and s_U. This leads to a new probability distribution q' such that $q'(s_A) = q_A - q_{\min}$, $q'(s_B) = q_B - q_{\min}$, $q'(s_I) = q(s_I) + q_{\min}$, and $q'(s_U) = q(s_U) + q_{\min}$; for all other outcomes $s \in S$ we have $q'(s) = q(s)$. The modified distribution q' corresponds to some distribution $\psi' \in \Psi(N')$.

We now show that $G(\psi') \leq G(\psi)$. Let n_g denote the number of events contained in some outcome t_g, and d denote the number of events contained in t_A but not in t_B. Then, $n_A = n_I + d$ and $n_U = n_B + d$. We have that

$$\begin{aligned} G(\psi') - G(\psi) &= q_{\min} g(n_I + d) + q_{\min} g(n_B) - q_{\min} g(n_I) - q_{\min} g(n_B + d) \\ &= q_{\min}[(g(n_I + d) - g(n_I)) - (g(n_B + d) - g(n_B))] \\ &\geq 0 \end{aligned}$$

The inequality holds by the concavity of q and the fact that $n_B > n_I$.

This modification step can be repeated for this same pair of events e_i and e_{i-1} as long as the predicate is not satisfied. Note that, every time such a modification step takes place, either $q(s_A)$ or $q(s_B)$ becomes 0 and the probabilities of these outcomes are never raised again during the modifications steps for this same pair of events. This implies that we only need a finite

number of modifications before the induced lottery satisfies the predicate. We have shown that the expected value of g will not increase at any point during this process, and therefore G is minimized when the ordering lottery is used. ∎

Proof of Theorem 5.1: Consider two problem instances (A and B) with n sellers and $m = n^2 + n + 1$ buyers. In both instances, buyers' demand is the same: one buyer demands a sample of n (i.e., all of the sellers) and the remaining $n^2 + n$ buyers demand a sample of just a single seller. The two instances differ with respect to the sellers' cost functions and have different payment functions: $\pi_A^{max}(k) = \min\{k, n\}$ and $\pi_B^{max}(k) = \max\{k, n\}$; note that both can arise from concave c_i's.

Since we are interested in distributions that are oblivious to the payment function and the two instances differ *only* with respect to the payment functions, it suffices to show that if a distribution $\psi \in \Psi(N)$ gives a 2-approximation for instance A, then the same distribution ψ cannot give a better than $(2 - \epsilon)$-approximation for instance B for $\epsilon > 0$. More formally, we will show that if

$$\sum_{k=0}^{m} p_\psi(k)\pi_A^{max}(k) \leq 2\Pi_A^{\mathrm{OPT}}, \tag{1}$$

then

$$\sum_{k=0}^{m} p_\psi(k)\pi_B^{max}(k) > (2 - 6/n)\Pi_B^{\mathrm{OPT}}. \tag{2}$$

Setting $n > 6/\epsilon$ will then conclude the proof.

Since π_A^{max} is concave, ψ^* is optimal for instance A and

$$\Pi_A^{\mathrm{OPT}} = \sum_{k=0}^{m} p_{\psi^*}(k)\pi_A^{max}(k) = \frac{1}{n}\min\{n^2+n+1, n\} + \frac{n-1}{n}\min\{1, n\} = 2 - 1/n.$$

Thus, (1) implies that $\sum_{k=0}^{m} p_\psi(k)\pi_A^{max}(k) < 4$.

Then, $\sum_{k=0}^{n} p_\psi(k)\pi_A^{max}(k) = \sum_{k=0}^{n} k p_\psi(k) < 4$, which together with Lemma 4.2 implies

$$\sum_{k=n+1}^{m} k p_\psi(k) > n - 2. \tag{3}$$

Moreover, $\sum_{k=n+1}^{m} p_\psi(k)\pi_A^{max}(k) = n\sum_{k=n+1}^{m} p_\psi(k) < 4$, which implies that

$$\sum_{k=0}^{n} p_\psi(k) > 1 - 4/n. \tag{4}$$

Now consider instance B. The buyer that requested a sample of size n will be given access to the data of all sellers. To determine what samples other buyers will get, suppose we randomly split the set of n^2+n buyers demanding a single seller into n groups of $n+1$ buyers each. We label these groups $\{1, 2, ..., n\}$. We then assign seller i to the buyers of group i. This gives unbiased samples, because each buyer is equally likely to be in each group. Note that exactly n buyers get access to the data of a given seller, so $\Pi_B^{\text{OPT}} \leq n$.

To conclude the proof, we show that (3) and (4) imply (2). First note that in order to satisfy the demand of the buyer who requests n sellers, every seller will be sampled at least once and paid at least $\max\{1, n\} = n$. Then, (4) implies that $\sum_{k=0}^{n} p_\psi(k)\pi_B^{max}(k) > n-4$. On the other hand, (3) implies that $\sum_{k=n+1}^{m} p_\psi(k)\pi_B^{max}(k) > n-2$. Summing these two inequalities, we conclude that $\sum_{k=0}^{m} p_\psi(k)\pi_B^{max}(k) > 2n-6$, which together with $\Pi_B^{\text{OPT}} \leq n$ implies (2). ∎

Proof of Theorem 5.2: Let $\psi \in \Psi(N)$ be the distribution that achieves Π^{OPT}, that is $\Pi^{\text{OPT}} = \sum_{k=0}^{m} p_\psi(k)\pi^{max}(k)$. For simplicity, we write $p \equiv p_\psi$ and $p_o \equiv p_{\psi^*}$ for the remainder of the proof. We wish to show that

$$\sum_{k=0}^{m} p_o(k)\pi^{max}(k) \;\leq\; 2\sum_{k=0}^{m} p(k)\pi^{max}(k). \tag{5}$$

Let $P(k) = \sum_{k' \leq k} p(k')$ denote the cumulative distribution function of p, and $P^{-1}(t) = \inf\{k \mid P(k) \geq t\}$ denote its generalized inverse distribution function; the functions $P_o(\cdot)$ and $P_o^{-1}(\cdot)$ are defined similarly for p_o.

We now wish to decompose the $[0, 1]$ interval into subintervals (t_l, t_r) for which we know that, for any two values $t, t' \in (t_l, t_r)$, we have $P^{-1}(t) = P^{-1}(t')$ and $P_o^{-1}(t) = P_o^{-1}(t')$. In order to do so, we let $T = \{P(k) \mid (p(k) > 0) \vee (p_o(k) > 0)\}$ be the set of distinct values in $[0, 1]$ that either $P(\cdot)$ or $P_o(\cdot)$ takes. If $0 < t_1 < t_2 < ... < t_{|T|} = 1$ is an ordering of the values in T and $t_0 = 0$, then we let $I = \{(t_0, t_1], (t_1, t_2], ..., (t_{|T|-1}, 1]\}$, which is a set of intervals that satisfy the property that we wanted. Given this property, (5) can be rewritten as follows:

$$\sum_{t_r \in T} (t_r - t_{r-1})\pi^{max}(P_o^{-1}(t_r)) \leq 2\sum_{t_r \in T} (t_r - t_{r-1})\pi^{max}(P^{-1}(t_r)). \tag{6}$$

Now, let $T_A = \{t_r \in T \mid P^{-1}(t_r) > P_o^{-1}(t_r)\}$ and $T_B = T \setminus T_A$. By definition of the set T_A, and using the fact that $\pi^{max}(\cdot)$ is an increasing

function, it is easy to see that for any $t_r \in T_A$ we have $\pi^{max}(P^{-1}(t_r)) \geq \pi^{max}(P_o^{-1}(t_r))$. We can therefore conclude that

$$\begin{aligned}\sum_{t_r \in T_A} (t_r - t_{r-1})\pi^{max}(P_o^{-1}(t_r)) &\leq \textstyle\sum_{t_r \in T_A} (t_r - t_{r-1})\pi^{max}(P^{-1}(t_r)) \\ &\leq \textstyle\sum_{t_r \in T} (t_r - t_{r-1})\pi^{max}(P^{-1}(t_r)). \quad (7)\end{aligned}$$

Let i be the seller for which $c_i(k) = \pi^{max}(k)$. Then, for $k' \leq k$,

$$\pi^{max}(k) \;=\; c_i(k) \;\leq\; \frac{k}{k'} c_i(k') \;\leq \frac{k}{k'} \pi^{max}(k') \quad \Rightarrow \quad \frac{\pi^{max}(k)}{k} \leq \frac{\pi^{max}(k')}{k'}, \tag{8}$$

where the first inequality holds because of the concavity of $c_i(\cdot)$ and the second inequality holds by definition of $\pi^{max}(k') = \max_i c_i(k')$. Thus, even though π^{max} may not be concave, the payment that a seller receives per buyer is decreasing in the total number of buyers that get access to his data. Also, note that the expected number of samples of p is equal to the expected number of samples of p_o (by Lemma 4.2) and equal to the area below the corresponding inverse cumulative distribution functions in the interval $[0, 1]$, or

$$\sum_s p(s)s \;=\; \sum_{t_r \in T} (t_r - t_{r-1})P^{-1}(t_r) \;=\; \sum_{t_r \in T} (t_r - t_{r-1})P_o^{-1}(t_r).$$

Finally, note that for any $t_r^* \in T$ the distribution p_o satisfies the property that the area below $P^{-1}(\cdot)$ in the interval $(0, t_r^*]$ is not smaller than the corresponding area below $P_L^{-1}(\cdot)$, i.e.

$$\sum_{t_r \in T_{\leq t_r^*}} (t_r - t_{r-1})P^{-1}(t_r) \geq \sum_{t_r \in T_{\leq t_r^*}} (t_r - t_{r-1})P_o^{-1}(t_r), \tag{9}$$

where $T_{\leq t_r^*} \equiv \{t_r \in T \mid t_r \leq t_r^*\}$. To verify this fact, assume that there exists some $t_r^* \in T$ for which this is not true. Then, the expected number of samples for p in $(0, t_r^*]$ is less than the expected number of samples for p_o in the same interval. Each one of these samples corresponds to some buyer, so there exists some buyer to whom p assigns a smaller probability of being assigned a seller in the outcomes of interval $(0, t_r^*]$ than p_o does. Since the total probability of that buyer being allocated a given seller has to be the same for p and p_o, this means that p assigns a higher probability of allocating a seller to this buyer for outcomes in the interval $(t_r^*, 1]$. This is a contradiction though since, by definition of p_o, any buyer to whom it assigns positive probability for outcomes of the interval $(0, t_r^*]$, will be assigned a

seller in *every* outcome of the $(t_r^*, 1]$ interval, and therefore p cannot assign higher probability than that to this buyer.

We now show that we can upper bound the payment from distribution p_o for each interval $(t_l, t_r]$ with $t_r \in T_B$. We present an iterative process that maps each such interval to a different interval $(\bar{t}_l, \bar{t}_r]$ with $\bar{t}_r \leq t_r$. In this iterative process we use a variable $\bar{t}_l$ which is initially set to 0.

While T_B is not empty, let t_r be its smallest value. Inequality (9) implies that there exists some other value $\bar{t}_r \leq t_r$ for which the area below $P^{-1}(\cdot)$ in the interval $(\bar{t}_l, \bar{t}_r]$ [12] is equal to $(t_r - t_{r-1})P_L^{-1}(t_r)$. For every $t \in (\bar{t}_l, \bar{t}_r]$ we know that

$$P_L^{-1}(t_r) \geq P^{-1}(\bar{t}_r) \geq P^{-1}(t).$$

The first inequality is true by definition of the set T_B and both inequalities use the fact that $P^{-1}(\cdot)$ is an increasing function. Using Inequality (8), we then have

$$\frac{\pi^{max}(P_L^{-1}(t_r))}{P_L^{-1}(t_r)} \leq \frac{\pi^{max}(P^{-1}(\bar{t}_r))}{P^{-1}(\bar{t}_r)} \leq \frac{\pi^{max}(P^{-1}(t))}{P^{-1}(t)}.$$

Since $p(k)\pi^{max}(k) = p(k)k\pi^{max}(k)/k$, using these latest inequalities we can deduce that the cost $(t_r - t_{r-1})\pi^{max}(P_L^{-1}(t_r))$ of p_o for the interval $(t_{r-1}, t_r]$ will be at most the cost of p for the interval $(\bar{t}_l, \bar{t}_r]$. We set $\bar{t}_l = \bar{t}_r$, and we repeat this step.

Note that, the implication of Inequality (9) that we use in the first step will continue to hold after every step since we have essentially removed the same area from the left end of both distributions. We have now shown that the expected payment of $\mathfrak{L}_3$ corresponding to the intervals of I_B will be upper bounded by the expected payment of the optimal lottery, or

$$\sum_{t_r \in T_B} (t_r - t_{r-1})\pi^{max}(P_o^{-1}(t_r)) \leq \sum_{t_r \in T} (t_r - t_{r-1})\pi^{max}(P^{-1}(t_r)).$$

Summing this inequality with (7) proves (6), which concludes the proof. ∎

Proof of Theorem 5.3: Similar arguments to the ones of Theorem 4.1 show that if the payment to seller i is given by π^{max}, then (1) truthful reporting is a dominant strategy and (2) we have ex-post individual rationality. Moreover, the threshold $\overline{r}_i$ for seller i in Step 3 does not depend on the values $\hat{c}_i$, $\hat{e}_i$ that he reports. As a result, it is a dominant strategy to also

[12] Note that this is not equal to $(\bar{t}_r - \bar{t}_l)P^{-1}(\bar{t}_r)$ since we do not know that for any two values $t, t' \in (\bar{t}_l, \bar{t}_r)$, $P^{-1}(t) = P^{-1}(t')$.

report the certainty equivalent in Step 2 truthfully. This is true even for a seller i for whom $c_i(k) > \max_{j \neq i} \hat{c}(k)$ for some k. To conclude the proof observe that each seller i derives non-negative expected utility when assigned payment π^{max} and his expected utility may only increase if assigned the other payment option. Thus, the mechanism is individually rational. ∎