Electronic Markets (2015) 25:109–123
DOI 10.1007/s12525-015-0188-8

SPECIAL THEME

Pricing private data

Vasilis Gkatzelis · Christina Aperjis · Bernardo A. Huberman

Received: 25 September 2013 / Accepted: 19 February 2015 / Published online: 20 March 2015

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.

Keywords Private data · Unbiased samples · Market design · Certainty equivalent · Pricing · Incentive compatible

JEL Classification D44 · D47 · D82

Responsible Editor: Rainer Böhme

The first two authors were working at HP Labs when this work took place.

V. Gkatzelis (✉)
Computer Science Department, Stanford University,
Palo Alto, Standford, CA, USA
e-mail: gkatz@cs.stanford.edu

C. Aperjis
Power Auctions, Washington, DC 20007-3591, USA
e-mail: caperjis@gmail.com

B. A. Huberman
HP Labs, Palo Alto, CA, USA
e-mail: bernardo.huberman@hp.com

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 (Haddadi et al. 2012). 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 company 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 Singer (2012).

One solution that would alleviate part of this controversy would be the creation of 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 in to this market if the price is high enough. This approach has the potential to make useful data repositories accessible to interested parties while respecting the preferences and privacy attitudes of individuals.

In this work, we present mechanisms that facilitate this exchange while satisfying a collection of desired properties.

We consider settings where each buyer is interested in a specific attribute of a representative subset of individuals with certain characteristics (and not in the private data of specific individuals). For example, a company that designs games might be interested in how much time Facebook users who are in their twenties spend playing games online. This can be achieved by giving each buyer access to an *unbiased sample* of a certain size, that is, to the values of this attribute 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 (Carrascal et al. 2013). Quite often one's privacy attitude is correlated with the value of the attribute that a buyer may be interested in (Huberman et al. 2005). In order to minimize this bias one needs to set the price high enough so that almost all the individuals choose to participate in the market. This implies that the cost of a truly unbiased sample can be extremely high even if just a few of the individuals are very concerned about their privacy. Even though we anticipate that this effect will be less pronounced for the types of queries that we consider, this is an intrinsic problem that all mechanisms aiming to elicit unbiased samples face. In order to alleviate this problem, our mechanisms provide the market maker with the ability to control the extent to which bias is introduced into the sample in exchange for decreased cost. Our goal is to minimize the expected value of the price that the buyers are asked to pay for the samples. This way, we can increase the buyers' interest in the market and hence the market's chance for adoption, while ensuring that the sellers are willing to participate in the market.

The individual sellers may also differ with respect to their attitude towards risk. A risk-averse individual prefers a guaranteed payment to a risky one with the same or even larger expected payment. One can take advantage of the risk aversion 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 (Aperjis and Huberman 2012).

In this paper we show that appropriately bundling the buyer demand can significantly decrease the price of unbiased samples. We first demonstrate how bundling the buyer requests can amplify the benefits from risk aversion described in Aperjis and Huberman (2012) by leveraging the fact that individuals tend to exhibit more risk aversion for higher payments (Holt and Laury 2002). More specifically, we identify the optimal way to bundle demand so as to minimize the expected payment to a risk averse seller. We then show that the same demand bundling technique also provides optimal worst-case guarantees in different settings. Throughout this paper, we take a prior-free approach and assume no knowledge of the distribution of the 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 the average of some value, the sum, or the weighted sum (Roth and Schoenebeck 2012; Ghosh and Roth 2011; Dandekar et al. 2012; Cummings et al. 2015). In contrast to that work, we consider a scenario where buyers pay for access to raw anonymized data instead of just an estimate for its statistical value. The selling of raw private data has been previously considered (e.g., (Riederer et al. 2011)) but not for unbiased samples, which is the focus of this paper.

A relevant line of work studies how to estimate certain statistics in the context of differential privacy (Ghosh and Roth 2011; Dandekar et al. 2012). The approach that the literature on differential privacy follows is to add unbiased noise to the information that is being sampled in an attempt to minimize the chance that any individual records can be identified. 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 can possibly render this mechanism very expensive for the buyer. Our approach avoids this problem by using an unbiased sampling technique; this technique induces small unbiased subsets of the data that a buyer can then use to compute statistics about them. In contrast to the differential privacy literature, our approach does not add any noise to the data.

More recently, Roth and Schoenebeck (2012) 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.

In the following section, we provide a more detailed description of the structure of this market, of our objectives, and of the barriers that we faced.

The market

Consider a data repository that contains information on n individuals (the sellers). For instance, this repository could contain information obtained by a company which provides its customers with access to streaming movies on the web. As the customers use this service, the company collects information about their viewing habits. A buyer is a third party interested in obtaining access to a representative sample of a subset of this data; in this example a buyer

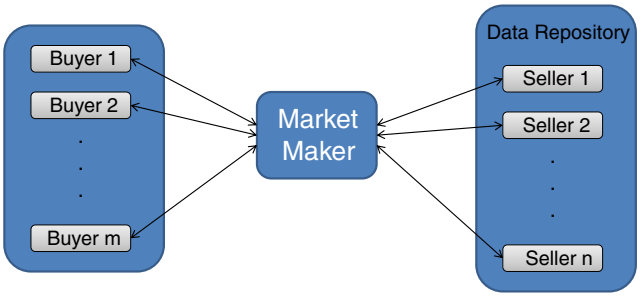


Fig. 1 The market-maker facilitates interactions between buyers and sellers

could be a TV channel that is interested in the amount of time that a certain demographic spends watching movies, or in the average rating of a movie by this demographic. A buyer like this would therefore be willing to pay in order to gain access to the corresponding information of an unbiased, and thus representative, sample of individuals from that demographic.

Such efforts from companies to gather information about their target group of customers, known as market research, is not a new trend. Traditionally, the acquisition of this type of information took place via polls and surveys, but these methods now seem inefficient in light of the unprecedented amount of relevant information that is being collected online. On the other hand, when a data-repository sells this information to the interested buyers without the permission of the users that this data pertains to, the privacy of these individuals is breached. The users may be reluctant to allow access due to privacy concerns, but they may be willing to reconsider if the potential rewards are high enough.

We envision that a market-maker facilitates the interactions between buyers and sellers, as shown schematically in Fig. 1. The role of the market-maker could, for example, be played by the company that owns the data repository. The market-maker can set the price that a buyer needs to pay per individual seller in order to obtain access to an unbiased sample, while ensuring that a representative set of individuals choose to opt in to this market. The buyer pays the market-maker and the market-maker uses the buyer's payment to appropriately compensate the sellers, while keeping a cut for himself.

In this work, we approach this problem from the market-maker's perspective, aiming to minimize the expected payment that the sellers will receive, while ensuring that these sellers still choose to opt in to the market. Seeking to minimize this expected payment, one 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. Such an approach would not give an unbiased sample though, since the value of the attribute that the buyers are interested in will often be correlated with the corresponding seller's privacy attitude (Huberman et al. 2005). The requirement of an unbiased sample implies that each individual should be selected with the same probability, independently of how much he values privacy.[1]

Privacy attitudes Apart from the need to ensure that the samples provided to the buyers will be unbiased, the market-maker faces a much more important barrier: different individuals have different privacy attitudes (Huberman et al. 2005; Acquisti et al. 2013; Cvrcek et al. 2006; Hann et al. 2007), and the market-maker does not have a prior regarding the privacy attitudes of the sellers. 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 though, even those unconcerned about their privacy may pretend that they are if they expect that this will result in getting higher payments. Therefore, our goal is to design *truthful* mechanisms that the market-maker can use, i.e., mechanisms that incentivize the sellers to always report their true privacy related preferences. Also, to ensure that no seller regrets participating in the market, we restrict ourselves to *individually rational* mechanisms; that is, mechanisms that always reward each seller at least as much as the privacy cost that he suffered. Note that this private data is already stored and the individuals cannot change it, so the market maker does not need to worry about eliciting the true value of the data as well.

Example 1 Consider a company which has 1000 subscribers and knows the value of some attribute α for each one of them. Some third party is interested in buying access to values of attribute α from an unbiased sample of 100 out of the 1000 subscribers (without knowing which 100 subscribers participate in the sample). Each subscriber may be willing to be part of the sample for a different compensation, but nobody, except the particular subscriber, knows how high this compensation is. For simplicity, assume that 300 of the subscribers would not want to be part of the sample unless they receive an payment of at least $10 and, among the remaining 700 subscribers, 300 require at least $5, and 400 do not care about their privacy and would not really mind being in the sample even if they are not compensated. How can a market maker incentivize each subscriber to report the truth regarding his privacy attitude while choosing an unbiased sample and ensuring that every sampled subscriber is happy with his compensation?

[1] 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.

A Solution In light of the restriction to truthful, individually rational mechanisms, and given the fact that the market-maker does not have a prior regarding the sellers' privacy attitudes, the number of interesting solutions is very limited. In fact, no mechanism can avoid introducing some bias to the samples unless it assumes that the market-maker knows a price which is high enough to attract all the sellers. Nevertheless, even without this assumption, there is a natural mechanism that only introduces a negligible amount of bias. The mechanism first asks each seller to report the minimum price for which he would allow a buyer to access the value of his α attribute. Then, among the sellers that reported the maximum price, c^{max}, one of them is chosen uniformly at random and he is discarded from the set (in Example 1, c^{max} =\$10 and one of the 300 subscribers who would request this compensation is discarded). A random sample of sellers is then selected from the remaining set (in Example 1, 100 subscribers would be chosen randomly among the 999 non-discarded ones). Each sampled seller gets paid c^{max} in return, and the sample is sold to the interested buyer. Assuming that the initial set of sellers was large enough, and for most interesting statistics, the fact that just one seller was discarded based on his reported privacy attitude introduces only a negligible bias. Also, as we show, this mechanism is both truthful and individually rational. The mechanism described above, which we call the *Baseline Mechanism* (for more details see Section "Baseline mechanism for linear costs"), provides a first way to satisfy the truthfulness and individual rationality constraints, but it does not do much in terms of minimizing the expected payment that is offered to the sellers.

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 which 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. Similarly, other risk averse sellers may also be willing to replace the lottery with a less risky lottery that has a lower expected payment. We leverage this fact by designing mechanisms that allow each seller to replace the initial lottery with less risky ones that lead to a decrease in the expected payment; more importantly we also show how to bundle the buyers' requests in order to amplify the effect of risk aversion.

Example 2 Expanding on Example 1, assume that, among the 400 subscribers that do not care about being sampled, half of them are risk-averse. In particular, assume that if these bidders were provided with the option of receiving ¢70 irrespective of whether they are sampled or not, and the option of receiving \$10 if they are sampled (which happens with probability around 0.1) and \$0 otherwise, then they would prefer the first option. Note that a risk-neutral subscriber would prefer the latter option whose expected payment is \$1. Using this observation, instead of just using the Baseline mechanism described above, we could also offer to each subscriber the option of receiving a certain payment of ¢70 before the sample is chosen. This way, the risk-averse subscribers would be able to increase their happiness by opting to use this alternative compensation option, while the expected cost of the sample would drop from \$1000 (\$10 per sampled subscriber) to \$940. Each of the 200 risk-averse subscribers would receive ¢70, irrespective of whether he is sampled or not (a total cost of \$140), and each one of the remaining 800 subscribers would receive \$10 only if sampled, which happens with probability 0.1 each (a total expected cost of \$800).

Bundling buyers' requests It is known that individuals tend to exhibit more risk aversion for higher payments (Holt and Laury 2002). This suggests that, if there happen to exist such risk averse individuals in the market, it may be possible to further reduce the price per data point by bundling the buyers' demand. For instance, the market-maker may choose to sample sellers in a way such 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 effect we choose to correlate the samples of different buyers in order to induce lotteries with higher risk; as a result, in order to avoid this risk, the risk-averse sellers prefer risk-free alternatives, even if the expected payment is substantially smaller. Note that this approach does not assume or require that the majority of the sellers are risk averse. Having said that, the more such individuals exist, the more substantial the improvement on the expected payment. Also, 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.

Example 3 The examples discussed above treat each third party interested in buying samples independently. Suppose there is a total of $m = 150$ interested buyers, such that 50 buyers want a sample of 50 sellers and 100 buyers want a sample with 200 sellers. In this case, repeatedly using the Baseline mechanism for each one of these buyers, would yield a total expected payment of \$250 for each subscriber (an expected payment of \$1 from each of the first type of buyers and \$2 from the second type). The first

observation is that, with such higher payments, the risk-aversion effects become significantly more pronounced. The second observation, which is the main focus of this paper, is that correlating the samples that these buyers receive can increase the effects of risk-aversion even further. If we choose a different sample independently for each buyer, then the probability that a subscriber never gets sampled, and hence receives no payment, is extremely small. If, on the other hand, we choose a random ordering of the subscribers once, and then each buyer seeking a sample of size d receives the attribute values of the first d subscribers according to the ordering, this probability remains high. In our example with the 150 buyers, for instance, the probability that a subscriber receives no payment would be 0.8, i.e., the probability that the subscriber would not be among the first 200 in the ordering. If the subscriber was among the first 100 in the ordering (with probability 0.1), he would be sampled 150 times, leading to a payment of $1500! Finally, subscribers that are in the top 200, but not top 100, positions in the ordering, would be sampled 100 times for a payment of $1000. As we show in this paper, a risk-averse subscriber facing this more extreme lottery would be much more likely to settle for a guaranteed payment which is significantly lower than the expected payment of $250. In particular, we show that correlating the samples in the fashion described above guarantees some highly desirable properties.

Paper Structure In Section "Model", we provide the formal definitions and the assumptions of our model. For the first set of results, in Section "Linear costs", we assume that the minimum payment that the sellers request is linear in the number of buyers gaining access to their data. In other words, if k buyers gain access to some seller's data, the minimum payment that the seller requests is exactly k times the minimum payment that he requests if just one buyer gains access. In Section "General costs" we also consider the non-linear case, for which, even adapting the idea of the Baseline Mechanism described above is not obvious (we address this issue in more detail in Section "General costs"). Nevertheless we show that an adaptation of the Baseline Mechanism, combined with the bundling method mentioned above provides an optimal worst-case guarantee with respect to the expected payment minimization objective. Finally, in Section "Discussion", we conclude this work with a discussion regarding the contributions and the limitations of our results.

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 he wishes to buy access to. We assume that the demand is price-insensitive; that is, buyer b wants to get an unbiased sample of d_b individuals irrespectively of the price.[2] Furthermore, for expository ease we assume that all buyers are interested in individuals with the same characteristics. We note, however, that our results can be directly applied to the general case where each buyer may define a large enough subset of sellers that he is interested in; one simple way to do this would be to use our mechanisms for each one of these subsets separately.

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 exactly 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 $c_i(\cdot)$ and $u_i(\cdot)$ are private information of seller i, so only he knows these functions. We make the following assumption:

Assumption 1 The utility of seller i for obtaining a monetary payment x while allowing k buyers to access his data is equal to $u_i(x - c_i(k))$.

The intuition behind this utility function is that, since the value of the cost $c_i(k)$ that the seller suffers is based on a monetary measure, the seller's utility depends on the difference between the payment that the seller receives and the privacy cost he incurs. In what follows, we refer to the difference $x - c_i(k)$ as the seller's *profit*. The utility of the seller is therefore a non-decreasing function of this profit. Formally, some seller i is risk-averse if his utility function $u_i(\cdot)$ is concave. Alternatively, a seller may be risk-neutral (which corresponds to a linear utility function) or risk-seeking (corresponding to a convex utility function). 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 (Mas-Colell et al. 1995).

Given any subset of sellers $N' \subseteq N$ and a set of sample size requests d_b, one from each buyer $b \in B$, there are many different ways in which one can generate a set of unbiased samples of the requested sizes using sellers in N'. More formally, one can think of sampling as a probability

[2] This implies that the value that buyer b sees in obtaining access to an unbiased sample of d_b individuals is higher than the price that he is asked to pay. Our mechanisms work more generally for settings where the demand does not change drastically with the price or the market-maker has a good estimate of the right range for the price. See (Gkatzelis et al. 2012) for details.

distribution over all possible outcomes. Let $\Psi(N')$ denote the set of all such distributions that produce the requested unbiased samples from N'; for notational simplicity, we suppress the dependence of Ψ on the buyers' requests. Given a distribution $\psi \in \Psi(N')$ representing the sampling, let $p_\psi(k)$ be the probability that the data of some seller is sold to exactly k buyers. Thus, p_ψ represents the distribution of the number of times that a seller will be sampled. Clearly, if $\psi \in \Psi(N')$, the fact that ψ is unbiased implies that the distribution p_ψ is the same for all sellers in N'.

If k buyers gain access to the data of seller i, this seller will be compensated with some payment, which we denote by $\pi_i(k)$. In contrast to the probability that k buyers gain access to one's data, the corresponding payment may vary across sellers. We restrict our attention to payment functions $\pi_i(\cdot)$ that are deterministic; that is, the value of $\pi_i(k)$ for every seller i and every $k \leq m$ is decided before the sampling begins and is independent of ψ. Then, the expected utility of seller i for a given distribution p_ψ and a payment function $\pi_i(\cdot)$ equals

$$\sum_{k=0}^{m} p_\psi(k) u_i(\pi_i(k) - c_i(k)).$$

Similarly, the expected total payment (over all sellers) is equal to

$$\sum_{i=1}^{n} \sum_{k=0}^{m} p_\psi(k) \pi_i(k), \tag{1}$$

Since we are assuming that the demand is fixed, the market-maker's objective is to minimize the value of Eq. 1. In our attempt to optimize this objective, we restrict ourselves to 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 *ex post individually rational*, that is, every seller experiences a non-negative utility at each possible outcome.

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 the certainty equivalent, which we define next, is crucial for some of our mechanisms.

Definition 3.1 The *certainty equivalent* of seller i for lottery (p_ψ, π_i), denoted $e_i(p_\psi, \pi_i)$, is the amount of profit for which seller i is indifferent between receiving the expected profit of (p_ψ, π_i) and receiving a certain profit of $e_i(p_\psi, \pi_i)$; that is,

$$u_i(e_i(p_\psi, \pi_i)) = \sum_{k=0}^{m} p_\psi(k) u_i(\pi_i(k) - c_i(k)).$$

We assume that buyers are risk-neutral. However, this assumption is not essential for our mechanisms as long as the market-maker is risk-neutral.

Linear costs

In this section, we focus on the case when the sellers' cost functions are linear, and we refer to the value of $c_i(1)$ by c_i. Linearity implies that the cost function of seller i is of the form $c_i(k) = k \cdot c_i$, and hence c_i is the single parameter that characterizes his privacy attitude and that our mechanisms need to elicit. We begin by discussing the Baseline Mechanism, which we mentioned in Section "The market". In Section "General costs", we extend this Baseline Mechanism to a setting with general (not necessarily linear) cost functions.

Baseline mechanism for linear costs

The Baseline Mechanism begins by asking each seller i to report his parameter c_i; let c^{max} be the highest reported value. The mechanism then chooses some seller $\hat{j}$ uniformly at random among the ones that reported a parameter value equal to c^{max}, and it discards this seller. Finally, the mechanism uses some distribution $\psi \in \Psi(N \setminus \{\hat{j}\})$ in order to generate the requested samples excluding seller $\hat{j}$, and each time some seller is sampled, he receives a payment equal to c^{max}. The following Theorem shows two important properties of the Baseline Mechanism.

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

One natural concern about this mechanism is that, although Theorem 4.1refBLtruth shows that the mechanism is dominant strategy truthful, it is nevertheless not collusion-resistant: two sellers can agree that one of them will report an artificially high price in order to increase the payment c^{max} that the other seller might receive if he is sampled, in which case the two sellers can share this high payment. However, in order to prevent such a collusion among at most x sellers, the mechanism can simply discard the sellers that reported the top x prices and define c^{max} to be the x-th highest reported price instead. This variation of the mechanism provides a very natural tradeoff between collusion resistance and the introduction of bias to the samples. Alternatively, one could also choose the value of x in a randomized fashion in order to introduce more uncertainty and deter sellers from colluding.

Observe that there are many unbiased distributions in $\Psi(N \setminus \{\hat{j}\})$ that the mechanism can use. One choice 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, one 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. Since the payment for each sampled seller is always c^{max}, irrespective of the distribution ψ, the total payment of the market-maker is unaffected.

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 buyers 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 $u_i(k(c^{max} - c_i))$. Hence, if $c_i < c^{max}$, then seller i strictly prefers to be sampled as many times as possible.

One of the major drawbacks of the Baseline Mechanism is that prices may be high, and very high prices could deter buyers from participating in the market. We thus aim to decrease prices — while ensuring that sellers are still willing to participate — in order to increase the chance that buyers are willing to participate in the market. A useful implication of Definition 3.1 is that, given some distribution p_ψ, seller i is indifferent between being compensated according to some payment function $\pi_i(\cdot)$ and being compensated according to $\pi_i'(\cdot)$, where $\pi_i'(k) \equiv e_i(p_\psi, \pi_i) + c_i(k)$. If some seller is risk-averse, then his utility function is concave, and hence the certainty equivalent of a lottery is smaller than its expected value, i.e.

$$e_i(p_\psi, \pi_i) < \sum_k p_\psi(k)(\pi_i(k) - c_i(k)).$$

As a result, if some seller i is risk-averse, then the expected payment under $\pi_i'(\cdot)$ is smaller than the expected payment under $\pi_i(\cdot)$. In other words, a risk-averse seller would prefer a smaller payment in expectation that is appropriately distributed among the potential outcomes. Since we have assumed that buyers are risk-neutral, the seller's indifference between $\pi_i(\cdot)$ and $\pi_i'(\cdot)$ provides room for decreasing the price that a buyer has to pay. In the next section, we introduce the *Certainty Equivalent* mechanism that uses this fact in order to reduce the expected price that buyers will pay.

CE mechanism for linear costs

The Certainty Equivalent (CE) Mechanism, just like the Baseline Mechanism, uses some unbiased distribution $\psi \in \Psi(N \setminus \{\hat{j}\})$ in order to generate the samples, but, unlike the Baseline Mechanism, it essentially offers to each seller two different payment function options instead of just one. The first payment function, $\pi^{max}(\cdot)$, is equivalent to that of the Baseline Mechanism, i.e. $\pi^{max}(k) = c^{max} \cdot k$, while the second one, $\pi_i'(\cdot)$, is tailored to seller i's cost function and it guarantees seller i a certain risk-free profit, irrespective of the outcome of ψ.

Following the same steps as the Baseline Mechanism, the CE mechanism also asks each seller i to report his cost parameter $\hat{c}_i$ and it discards one of the sellers that reported the highest parameter value. In contrast to the Baseline mechanism though, in order to know which one of the two payment functions the seller would prefer, the mechanism presents each seller i with the lottery (p_ψ, π^{max}), and asks him to report his certainty equivalent $\hat{e}_i$ for this lottery (for a discussion regarding the incentives of the seller in reporting his certainty equivalent truthfully, see the detailed description of the mechanism following its formal definition). In order to attract a risk-averse seller to choose the risk-free payment function $\pi_i'(\cdot)$ instead of the lottery (p_ψ, π^{max}), this payment needs to be at least $\pi_i'(k) = \hat{e}_i + \hat{c}_i k$ for all k, thus ensuring that the seller will experience a profit of at least $\hat{e}_i$ no matter what the outcome of ψ will be. If we let

$$w \equiv \sum_k p_\psi(k)k$$

denote the expected number of times that a seller is sampled, and using (1), we get that the expected total payment for seller i if he was offered $\pi_i'(\cdot)$ would be $\hat{r}_i \equiv \hat{e}_i + \hat{c}_i w$. In other words, apart from the profit of $\hat{e}_i$ that the seller needs to experience, the payment needs to also cover the seller's expected costs whenever the seller is sampled. Of course, the CE mechanism only offers risk-free payment alternatives that lead to a lower expected total payment compared to $\pi^{max}(\cdot)$, i.e. $\hat{r}_i < c^{max} w$. As a result, these alternatives may only attract risk-averse sellers and the CE Mechanism is equivalent to the Baseline Mechanism for the rest. As a result, the expected payment of the CE Mechanism is always at most as much as that of the Baseline Mechanism, and the more pronounced the risk-aversion attitude of the sellers is, the smaller the expected payment of the CE Mechanism compared to that of the Baseline.

The mechanism then defines what the risk-free alternative payment $\pi_i'(\cdot)$ for each seller will be. More specifically, the mechanism decides what the expected payment $\bar{r}$ of this alternative payment will be, and then, for each bidder i, the alternative payment function appropriately distributes this expected payment in order to guarantee some risk-free profit. Formally, this means that $\pi_i'(k) = \bar{r} + \hat{c}_i(k - w)$, which implies that, for every k, the profit of the bidder equals

$$\pi_i'(k) - \hat{c}_i \cdot k = \bar{r} - \hat{c}_i \cdot w,$$

which is independent of k. Depending on whether $\hat{r}_i$ is smaller than $\bar{r}$ or not, the mechanism then knows whether i prefers this risk-free alternative payment or not, and chooses the preferred payment function on his behalf (Step 3 below).

Finally, the CE Mechanism produces unbiased samples from the set $N' = N \setminus \{\hat{j}\}$ 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 chosen on his behalf before the sampling. It is important to point out that the distribution ψ is not affected by the payment choices in any way; that is, how often a seller will be sampled is independent of the payment function that was chosen for him.

What follows is the sequence steps of the CE Mechanism:

(1) Ask every seller i to report the parameter c_i.
   - Denote the reported values by $\hat{c}_i$.
   - Define $c^{max} \equiv \max_i\{\hat{c}_i\}$ and some $\hat{j} \in \arg\max_i\{\hat{c}_i\}$.

(2) Ask every seller i to report his certainty equivalent for (p_ψ, π^{max}), where $\pi^{max}(k) \equiv c^{max} \cdot k$, and let $\hat{e}_i$ denote the reported value. An expected payment of $\hat{r}_i \equiv \hat{e}_i + \hat{c}_i w$ is needed to guarantee i a profit of $\hat{e}_i$.

(3) Let $\overline{r} \equiv \arg\max\{f(r)(c^{max} w - r)\}$, where $f(r) \equiv |\{j \in N \setminus \{\hat{j}\} : \hat{r}_i \leq r\}|$. For every seller i and for $k \leq m$, choose i's preferred payment function $\pi_i(\cdot)$:
   Set $\pi_i(k) \equiv \overline{r} + \hat{c}_i(k - w)$ if $\hat{r}_i < \overline{r}$, or
   Set $\pi_i(k) \equiv \pi^{max}(k)$ otherwise.

(4) Produce unbiased samples using some distribution $\psi \in \Psi(N \setminus \{\hat{j}\})$. Pay each seller i according to the $\pi_i(\cdot)$ that was chosen for him in Step 3:
   - 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 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 risk-free payment option with a lower expected payment. In choosing the value of $\overline{r}$, the expected payment of the risk-free alternative payments, the market-maker is faced with the following tradeoff: if the value of $\overline{r}$ is small, the benefit from the sellers that choose it will be greater, but the number of sellers it would attract might be smaller; similarly, setting $\overline{r}$ to be closer to $c^{max} w$ may increase the number interested sellers, yet the benefit from each one of these sellers will be smaller. In order to maximize the expected benefit, the market-maker then sets $\overline{r}$ equal to the value of r that maximizes $f(r)(c^{max} w - r)$, where $f(r) \equiv |\{j \in N \setminus \{\hat{j}\} : \hat{r}_i \leq r\}|$. The value of $f(r)$ corresponds to the number of sellers who would be interested in the risk-free payment alternative with an expected payment of r, and $(c^{max} w - r)$ corresponds to the expected benefit from each one of these sellers.

This approach can be used in a setting where the sellers are expected to be non-strategic when reporting their certainty equivalents in Step 2, but it does not guarantee truthful reporting when sellers are strategic in reporting their certainty equivalent value as well. More specifically, although the sellers prefer that the mechanism knows their true certainty equivalent when choosing between two payment functions on their behalf, one can come up with examples where some bidder might misreport his certainty equivalent in order to increase the value of $\overline{r}$. To guarantee truthful reporting, the market-maker can choose a value $\overline{r}_i$ for each seller i which instead maximizes $f_{-i}(r)(c^{max} w - r)$, where $f_{-i} \equiv |\{j \in N \setminus \{i, \hat{j}\} : \hat{r}_i \leq r\}|$ is determined by all sellers other than i. This way, $\overline{r}_i$ does not depend on the value $\hat{e}_i$ that 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.[3]

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

Theorem 4.2 *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.

It is useful to note that w, the expected number of times that a seller is sampled, does not depend on the choice of $\psi \in \Psi(N')$.

Lemma 4.3 *Let n' denote the number of sellers in N'. If $\psi \in \Psi(N')$, then*

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

Given a set of buyer requests, the expected payment generated by the CE Mechanism for these buyers is minimized

[3] In order to verify this fact, note that the value $\overline{r}$ which maximizes $f(r)(c^{max} w - r)$ will always correspond to the $\hat{r}_i$ value for some seller i. If this were not the case, then slightly decreasing the value of $\overline{r}$ would not affect $f(r)$, but it would increase $(c^{max} w - r)$, a contradiction. Hence, in order to avoid suboptimality, the value of r would in general be "controlled" by some seller i for whom $\overline{r} = \hat{r}_i$. This seller can, in general, increase $\overline{r}$ by lying in order to slightly increase $\hat{r}_i$.

when the certainty equivalent values of the risk-averse bidders are as small as possible. In the rest of this section, we show that we can actually identify the distribution that minimizes the certainty equivalent of every risk averse seller for the lottery $(p_\psi, \pi^{\max})$, over all possible distributions $\psi \in \Psi(N')$. This distribution $\psi^* \in \Psi(N')$ leads to a lottery $(p_{\psi^*}, \pi^{\max})$ that essentially maximizes the risk while the expected payment remains the same.

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

Definition 4.4 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.

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'.

We now show 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 4.5 *If $g : \mathbb{N} \to \mathbb{R}$ is concave and non-decreasing function, then, among all $\psi \in \Psi(N')$, function $G(\psi) \equiv \sum_{k=0}^{m} p_\psi(k) g(k)$ is minimized at $\psi = \psi^*$.*

But, from Definition 3.1, we know that the following is true for the certainty equivalent $e_i(p_\psi, \pi^{max})$ of seller i facing lottery (p_ψ, π^{max}) is

$$u_i(e_i(p_\psi, \pi^{max})) = \sum_{k=0}^{m} p_\psi(k) u_i((c^{max} - c_i)k).$$

Theorem 4.5, shows that ψ^* minimizes this value when u_i is concave, and, since u_i is non-decreasing, it also minimizes the value of the certainty equivalent.

Corollary 4.5 *For any seller with a concave utility function u_i, among all distributions $\psi \in \Psi(N')$, the certainty equivalent $e_i(p_\psi, \pi^{\max})$ is minimized at $\psi = \psi^*$.*

Examples

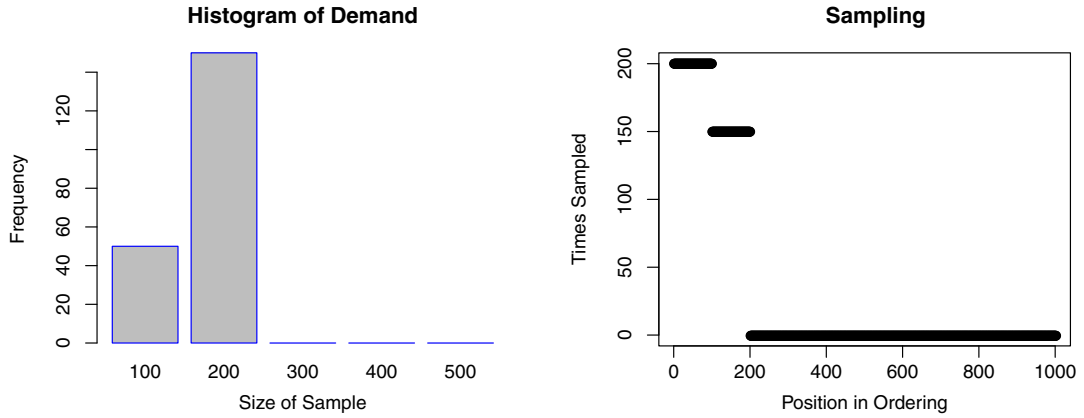
Suppose that out of $m = 200$ buyers, 50 have asked for samples with $d_b = 100$ sellers and 150 have asked for samples with $d_b = 200$ sellers. Furthermore, assume that when we remove the seller who reported that his cost is equal to c^{max}, we have 1000 sellers left: half of these sellers are risk-neutral 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.

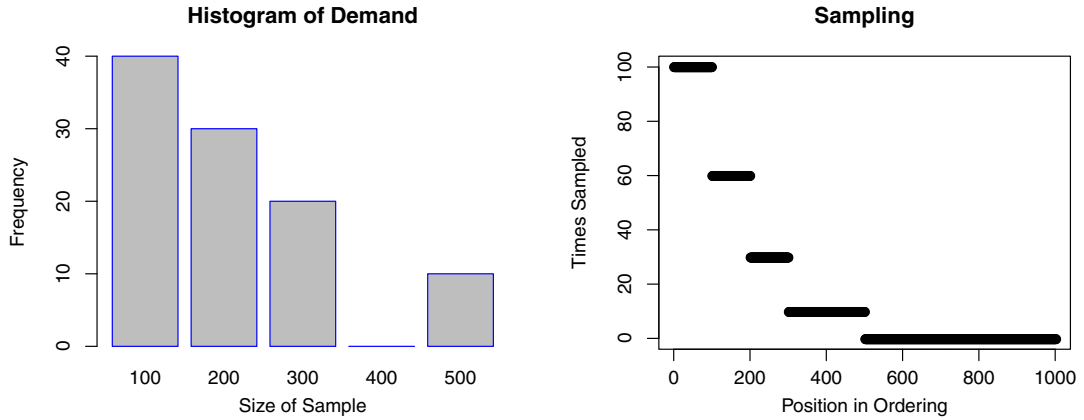
CE Mechanism without bundling To demonstrate the importance of the ordering distribution, we note that using a different distribution with the CE Mechanism may result in a price that is very close to the price of the Baseline Mechanism. In particular, if the samples of different buyers are independent,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

Estimating the certainty equivalent 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, so the market-maker needs to communicate this information to the sellers when requesting their certainty equivalent estimates. 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 (see, for example, the histograms in Figs. 2a and 2b). Alternatively, the market-maker can help a seller determine his certainty equivalent for the lottery (p_{ψ^*}, π^{max}) by showing 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 2a 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 2b 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 Fig. 2) and it will be relatively easy for a seller to determine his certainty equivalent.



(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 $d_b = 500$.

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

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 the Baseline Mechanism to a setting with general cost functions and show that the ordering distribution has good properties in this more general setting. In the full version (Gkatzelis et al. 2012), we also extend the CE Mechanism.

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 the marginal cost that the seller suffers by providing access to his data to $k+1$ buyers instead of k should not increase as k increases. Even more so when it is exactly the *same data* revealed to each one of the buyers.

We first note that for a special class of concave cost functions, the mechanisms of Section "Linear costs" can be applied with only minor modifications. In particular, this is the case if the privacy attitude of seller i can be represented by a function of the form $c_i(k) = c_i \cdot h(k)$, where c_i can be different for each seller, and h is a known 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 optimal for both the Baseline and the CE Mechanism.

In the more general case where sellers have arbitrary cost functions though, these mechanisms are not well defined: the sellers cannot be ordered based on their cost functions anymore; in particular, for two sellers i and j, it could be that $c_i(k) > c_j(k)$ and $c_i(k') < c_j(k')$ for $k \neq k'$.[4] Hence,

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

we consider the following generalization of the Baseline Mechanism for linear cost functions:

Baseline Mechanism for General Costs Each seller i is asked to report the values $c_i(k)$ for $k = 1, 2, ..., m$.[5] 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 sample 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 Gkatzelis et al. (2012) 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 paper we ignore this issue and assume for ease of exposition that the mechanism produces unbiased samples from the set of n sellers. As a result, the set of distributions we consider for the sampling is $\Psi(N)$.

Now, $\Pi(\psi) \equiv \sum_{k=0}^{m} p_\psi(k)\pi^{max}(k)$ is the expected payment per seller when payments are given by π^{max} and the sampling is according to ψ. In order to minimize the expected price of the Baseline Mechanism, we need to minimize $\Pi(\psi)$ over all $\psi \in \Psi(N)$. In the case of linear costs, π^{max} is linear and, by Lemma 4.3, $\Pi(\psi)$ obtains the same value for every $\psi \in \Psi(N)$. With arbitrary cost functions though, π^{max} may not be linear and the value of $\Pi(\psi)$ may be different for different $\psi \in \Psi(N)$. Therefore, our goal is to choose a distribution that minimizes this value. In the special case when π^{max} is concave, Theorem 4.5 implies that the ordering distribution ψ^* is optimal one. Note, however, that concavity of $c_i(\cdot)$ for all sellers i does *not* imply concavity of $\pi^{max} \equiv \max_i \hat{c}_i(k)$.

For the case when π^{max} is not concave, we are interested in distributions that are *oblivious* to π^{max} in the sense that they do not depend on the values $\pi^{max}(k)$.[6] 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}. We note that Π^{OPT} is generally unattainable by a distribution that is oblivious to π^{max}. 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}. The proof is provided in the appendix.

Theorem 5.1 *No $\psi \in \Psi(N)$ that is oblivious to π^{max} can guarantee $\Pi(\psi) \leq (2-\epsilon)\Pi^{OPT}$ for $\epsilon > 0$. This holds even if every seller's cost function 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. We provide a sketch of the proof in the Appendix; see (Gkatzelis et al. 2012) for the detailed proof.

Theorem 5.2 *If every seller's cost function is concave, then $\Pi(\psi^*) \leq 2\Pi^{OPT}$.*

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.

Discussion

In this paper we studied a market for private data where buyers can obtain access to unbiased samples of some private attribute value by appropriately compensating the individuals to whom this attribute values correspond (the sellers). A market-maker facilitates the interactions between the two sides of the market. We focused 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.

One important limitation of our approach is that it may require some non-trivial effort from the side of the sellers. More specifically, our mechanisms need to ask each seller several questions regarding both his privacy attitude and his attitude toward risk. As we prove, answering these questions accurately is to the seller's benefit, since this is how he can maximize his utility. In order to motivate the seller to participate though, we need to assume that the reward is high enough to cover for the additional cost that he suffers in order to accurately report his preferences. This may not be

[5] We can significantly reduce the number of values that a seller needs to report (1) if the sampling is based on bundling buyers' demand (e.g., if we use the ordering distribution) and (2) if there are relatively few different sample sizes requested by buyers, e.g., because buyers choose from predefined sample size options.

[6] 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.

the case initially if the demand is low, but it would be to the market-maker's best interest to subsidize this market until it gets adopted, which can then lead to significant benefits for all sides involved.

Finally, in order 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 is 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 seller, 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 Fig. 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. 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 benefiting both the sellers and the buyers of big data.

Appendix: A Proofs

We now provide the proofs of the results that were omitted from the main section.

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 4.3 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 4.2 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^{max}_{-i}), where $\pi^{max}_{-i} \equiv \max_{j\neq i} \hat{c}_j$, 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^{max}_{-i}).[7] □

Proof of Theorem 4.5 Let S denote the set of all subsets of B (i.e., the powerset of B), and let $S_b \subseteq S$ denote the set of all such subsets that include buyer 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 $b' < b$. What follows is described from the perspective of some arbitrary seller i. Given a distribution $\psi \in \Psi(N')$, let $q_\psi(s)$ denote the probability that seller i's data is sold to all buyers in s but nobody else.[8] Since q_ψ is a distribution over S, we have that $\sum_{s\in S} q_\psi(s) = 1$. Since $\psi \in \Psi(N')$, the probability that buyer b gets seller i in his sample must be equal to d_b/n', where n' is the number of sellers in N'. Equivalently, for each buyer b, $\sum_{s\in S_b} q_\psi(s) = d_b/n'$.

The ordering distribution $\psi^* = \psi^*(N')$ satisfies the following simple predicate: *If buyer b gets access to the data of seller i then so does buyer* $b-1$; equivalently, $q_{\psi^*}(s) = 0$ for every $s \in (S_b \setminus S_{b-1})$. We will show that $G(\psi)$ is minimized at ψ^* over all $\psi \in \Psi(N')$ using proof by contradiction. Assume that $G(\psi) < G(\psi^*)$ for some distribution $\psi \in \Psi(N')$ that does not satisfy the predicate. We will gradually modify distribution ψ until it satisfies the predicate without increasing the expected value G in the process (if g is *strictly* concave, then this modification leads to a decrease in G).

Let b be the first buyer in the ordering for which the predicate is not true, i.e., there exists some set s_A that contains b and does not contain $b-1$ ($s_A \in (S_b \setminus S_{b-1})$) such that $q_A \equiv q_\psi(s_A) > 0$. Since $d_{b-1} \geq d_b$, there must also exist some s_B that contains $b-1$ but not b, and occurs with some positive probability $q_B \equiv q_\psi(s_B) > 0$. 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 buyers 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 s_I and s_U. This leads to a new probability distribution $\hat{q}$ over S such that $\hat{q}(s_A) = q_A - q_{\min}$, $\hat{q}(s_B) = q_B - q_{\min}$, $\hat{q}(s_I) = q(s_I) + q_{\min}$, and $\hat{q}(s_U) = q(s_U) + q_{\min}$; for all other $s \in S$ we have $\hat{q}(s) = q_\psi(s)$. $\hat{q}$ corresponds to some distribution $\hat{\psi} \in \Psi(N')$.

We now show that $G(\hat{\psi}) \leq G(\psi)$. Let n_α be the number of buyers in s_α, and d the number of buyers in $s_A \setminus s_B$; thus, $n_A = n_I + d$, $n_U = n_B + d$, and $G(\psi') - G(\hat{\psi})$ is equal to

$$q_{\min} g(n_I + d) + q_{\min} g(n_B) - q_{\min} g(n_I) - q_{\min} g(n_B + d)$$
$$= q_{\min}[(g(n_I + d) - g(n_I)) - (g(n_B + d) - g(n_B))] \geq 0$$

The inequality holds because g is concave and $n_B > n_I$.

We can repeat the modification step for this same pair of buyers b and $b-1$ as long as the predicate is not satisfied. After every modification, either $q(s_A)$ or $q(s_B)$ becomes 0 and the probabilities of these sets are never raised again during the modification steps for this same pair of buyers. Thus,

[7] 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.

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

only a finite number of modifications is needed until the induced lottery satisfies the predicate. Since the expected value G does not increase at any point during this process, we conclude that $G(\psi)$ is minimized at ψ^*. □

Proof of Theorem 5.1 Consider two instances (A and B) with n sellers and $m \equiv n^2 + n + 1$ 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 one 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

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

then

$$\sum_{k=0}^{m} p_\psi(k)\pi_B^{max}(k) > (2 - 6/n)\Pi_B^{\text{OPT}}. \tag{3}$$

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

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

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

Thus, (2) 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 4.3, implies

$$\sum_{k=n+1}^{m} k p_\psi(k) > n - 2. \tag{4}$$

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

$$\sum_{k=0}^{n} p_\psi(k) > 1 - 4/n. \tag{5}$$

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 (4) and (5) imply (3). 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, (5) implies that $\sum_{k=0}^{n} p_\psi(k)\pi_B^{max}(k) > n-4$. On the other hand, (4) 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 (3). □

Proof Sketch for Theorem 5.2 Let $\psi \in \Psi(N)$ be the distribution that achieves Π^{OPT}; for notational simplicity, we write $p \equiv p_\psi$, $p_o \equiv p_{\psi^*}$, and $\pi \equiv \pi^{max}$ for the remainder of the proof. We wish to show that

$$\sum_{k=0}^{m} p_o(k)\pi(k) \leq 2\Pi^{\text{OPT}} = 2\sum_{k=0}^{m} p(k)\pi(k). \tag{6}$$

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

We 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 $T = \{P(k) | (p(k) > 0) \vee (p_o(k) > 0)\}$ 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, (6) can be rewritten as follows:

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

Let $T_A = \{t_r \in T | P^{-1}(t_r) > P_o^{-1}(t_r)\}$ and $T_B = T \setminus T_A$. By definition of the set T_A, and using the fact that $\pi(\cdot)$ is an increasing function, it is easy to see that for any $t_r \in T_A$ we have $\pi(P^{-1}(t_r)) \geq \pi(P_o^{-1}(t_r))$. Therefore

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

In order to prove the theorem (for a complete proof see (Gkatzelis et al. 2012)) we only need to show the following inequality; summing it with the previous one proves the theorem since $T_A \cup T_B = T$.

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

Proving this inequality is significantly more demanding, but the main intuition behind why it holds is that, even though π 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. To verify that this is true, for some k, let i be the seller for which $c_i(k) = \pi(k)$. Then, for $k' \leq k$,

$$\pi(k) = c_i(k) \leq \frac{k}{k'} c_i(k') \leq \frac{k}{k'} \pi(k')$$

$$\Rightarrow \frac{\pi(k)}{k} \leq \frac{\pi(k')}{k'},$$

where the first inequality holds because of the concavity of $c_i(\cdot)$ and the second holds by definition of $\pi(k') = \max_i c_i(k')$.

We use this to upper bound the payment from distribution p_o due to $t_r \in T_B$ intervals; we show that for every expected buyer of p_o in the interval $(t_l, t_r]$ with $t_r \in T_B$, there exists some other interval $(\bar{t}_l, \bar{t}_r]$ with $\bar{t}_r \leq t_r$ and the same expected number of buyers for p. Since $\bar{t}_r \leq t_r$, the former interval leads to a smaller payment per buyer. □

References

Acquisti, A., John, L.K., & Loewenstein, G. (2013). What is privacy worth? *The Journal of Legal Studies*, *42*(2), 249–274.

Aperjis, C., & Huberman, B.A. (2012). A market for unbiased private data: Paying individuals according to their privacy attitudes. *First Monday*, *17*(5).

Carrascal, J.P., Riederer, C., Erramilli, V., Cherubini, M., & De Oliveira, R. (2013). Your browsing behavior for a Big Mac: economics of personal information online. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013* (pp. 189–200).

Cummings, R., Ligett, K., Roth, A., Zhiwei Steven, W., & Ziani, J. (2015). Accuracy for sale: Aggregating data with a variance constraint. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, ITCS 2015, Rehovot, Israel, January 11-13, 2015* (pp. 317–324).

Cvrcek, D., Kumpost, M., Matyas, V., & Danezis, G. (2006). A study on the value of location privacy. In *Proceedings of Workshop on Privacy in the Electronic Society* (pp. 109–118).

Dandekar, P., Fawaz, N., & Ioannidis, S. (2012). Privacy auctions for recommender systems. In *WINE* (pp. 309–322).

Ghosh, A., & Roth, A. (2011). Selling privacy at auction. In *ACM Conference on Electronic Commerce* (pp. 199–208).

Gkatzelis, V., Aperjis, C., & Huberman, B.A. (2012). Pricing private data. SSRN eLibrary.

Haddadi, H., Mortier, R., & Hand, S. (2012). Privacy analytics. *SIGCOMM Computer Communications Review*, *42*(2), 94–98.

Hann, I.-H., Hui, K.-L., Lee, S.-Y.T., & Png, I.P.L. (2007). Overcoming online information privacy concerns: An information-processing theory approach. *Journal of Management Information Systems*, *24*(2), 13–42.

Holt, C.A., & Laury, S.K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*, 1644–1655.

Huberman, B.A., Adar, E., & Fine, L.R. (2005). Valuating privacy. *IEEE Security Privacy*, *3*(5), 22–25.

Mas-Colell, A., Whinston, M.D., & Green, J.R. (1995). *Microeconomic Theory*: Oxford University Press.

Riederer, C., Erramilli, V., Chaintreau, A., Krishnamurthy, B., & Rodriguez, P. (2011). For sale: your data: by : you. In *Tenth ACM Workshop on Hot Topics in Networks, HOTNETS* (p. 13).

Roth, A., & Schoenebeck, G. (2012). Conducting truthful surveys, cheaply. In *ACM Conference on Electronic Commerce* (pp. 826–843).

Singer, N. (2012). *You for sale: Mapping, and sharing, the consumer genome*: The New York Times.