How Differential Privacy Impacts Data Elicitation

JUBA ZIANI
Georgia Institute of Technology

Many studies and online platforms rely on data from individuals to learn properties of a population or train data-driven services and recommendation systems. However, this data is often sensitive or private, and individuals may be unwilling to share personal information. At the same time, the introduction of *Differential Privacy* has given us a formal and principled way to protect the privacy of individuals. Differential Privacy adds noise to the data or the learner's computation to "hide" the data of any specific individual, while still permitting to learn properties at the level of a population.

In this letter, we discuss our recent work on how one may use differential privacy to increase individuals' willingness to share their personal information. We are particularly interested in understanding the following trade-off: on the one hand, providing more privacy requires adding more noise, which leads to less accurate models; on the other hand, providing more privacy allows more individuals to share their personal data, leading to better models. The main challenges are two-fold: i) we may need to provide different levels of privacy protections to individuals with differing privacy preferences and ii) we aim to incentivize individuals to share their data without payments, but instead through the utility they obtain from the data-driven service offered by the learner.

Categories and Subject Descriptors: K.4.1 [**Computers and society**]: Public Policy Issues—*Privacy*; J.4 [**Social and Behavioral Sciences**]: Economics; F.0 [**Theory of Computation**]: General

General Terms: Algorithms, Economics, Security

Additional Key Words and Phrases: online platforms, data acquisition, privacy-aware agents, differential privacy, endogenous participation

1. MOTIVATION AND CONTEXT

Larger and larger amounts of data are collected about individuals and online users every year. Often, potentially highly sensitive data about individuals is collected and used to help train machine learning models and making better recommendations or decisions for these same individuals. In turn, privacy concerns and considerations immediately arise, and data privacy has effectively become one of the most major societal concerns and challenges of our time.

Providing privacy is far from a trivial challenge. The easiest way to obtain privacy may be to simply not collect or use data about individuals. At a high-level and perhaps in oversimplified terms, this is the approach behind the new European Union's General Data Protection Regulation or GDPR (see [E.U. 2016]). But, if possible, one may want to be able to still use and learn useful lessons from data, while simultaneously guaranteeing privacy to the individuals whose data we are using. For example, one could try to anonymize the data they collect in order to make it hard to identify any specific individual in the data.

However, many past, naive techniques to provide privacy have failed to do so in

Author's address: jziani3@gatech.edu

a meaningful way. This has been a consequence of the fact that many privacy techniques used in the past were derived from "intuition", and their guarantees were not properly formalized. In turn, these privacy protections failed when stressed by unplanned for and unexpected reconstruction attacks. As an example, data anonymization–which may intuitively sound like a reasonable way to provide privacy protections–has been shown to be vulnerable to very basic reconstruction attacks; a well-known example of this is how Latanya Sweeney uniquely identified the medical records of the governor of Massachusetts by noting that they were the only one having a certain combination of observable demographic attributes in the anonymized data [Sweeney 2000].

One of the most significant development in the area of privacy, in response to such previous privacy failures, was the introduction of *Differential Privacy* by [Dwork et al. 2006]. Differential Privacy essentially aims to give us the best of both worlds, by allowing us i) to still learn important lessons and make accurate inference from data while ii) making sure it is hard (in an information-theoretic sense) to recover accurate information about any specific individual in the data we feed to our machine learning models. The biggest strengths of differential privacy are that it provides *formal and provable* privacy guarantees, that *it does not need to anticipate and reason about specific types of reconstruction attacks*, and that *it allows for easy privacy accounting even in complex systems.*

Differential privacy relies on the addition of noise to the data, to the training process of the algorithm, or to the final estimate or model trained on said data. In turn, differential privacy exhibits a privacy-accuracy trade-off: intuitively, the more noise is added, the better the privacy guarantee of our algorithm is going to be, but the less accurate will the insights or predictions from this model be. A significant amount of effort in the space of differential privacy has been spent on developing algorithms and mechanisms that provide the best possible accuracy-privacy trade-off.

Often, this is done assuming that the learner has access to a fixed data-set. I.e., given that the learner has already collected data, what is the best they can do in terms of balancing out privacy and accuracy given this static data-set? This is obviously a major and important building block when it comes to understanding the properties of differential privacy. However, in this research note, I want to argue that this is not the entire picture, as it does not take into account the *data collection process.*

The point of view that I will argue in the remainder of this research note builds up on the following observation: the promise of privacy guarantees and the way we use individuals' data to provide them with insights or services *changes these individuals' incentives to share and provide their personal data*, and in turn affect the data collection process and the data-sets we have access to. Providing more privacy means adding more noise, but can–perhaps counter-intuitively–also lead to better accuracy guarantees via the collection of more data. In our recent work [Cummings et al. 2023], we aim to take this entire pipeline—both data collection and data estimation with differential privacy—into account.

2. OUR MODEL

2.1 Motivation

Let us start with two high-level examples of the types of considerations we are interested in modeling.

Example 2.1 Online platform. *Consider an online platform like YouTube. YouTube collects data about users, and uses this data to train its recommendation system and to decide what videos to recommend to what user. Individuals with privacy concerns have the option to opt out from making an account or from sharing their data. They can still access YouTube's content if they decide to do so, but do not get personalized recommendations from the platform if they opt out of sharing their watch and search histories.*

Example 2.2 Medical study. *Consider a medical study about a rare disease. An individual may participate in a medical study in the hope that this will help develop new treatment options that the individual will directly benefit from. However, doing so may reveal personal and sensitive information about themselves; in fact, even just learning of their participation in the study would be a serious breach of privacy, since it would reveal that the individual has the rare disease.*

In both examples, there is a notion of trade-off between the *privacy cost* incurred from sharing one's personal data and the *benefit* from the resulting model, study, or statistic computed by the learner or platform. This benefit depends partially on how many agents choose to share their data; e.g., the more people share their data with YouTube, the better the platform's recommendation system will be. In such settings, we can hope to leverage differential privacy to lower agents' barriers to sharing their data and provide them with improved outcomes and services.

From a more technical point of view, the name of the game is the following: we want to carefully choose the amount of noise we add for privacy to trade-off i) incentivizing more participation and data sharing (leading to more accurate models absent privacy noise) and ii) model inaccuracies resulting from the addition of noise. This task is rendered more complicated by the fact that different individuals may have varying privacy attitudes and preferences. Our goal in [Cummings et al. 2023] is in fact to develop a better formal and theoretical understanding of this trade-off between privacy costs and resulting quality of service, and to design optimal data acquisition algorithms which respect the users' differing preferences.

2.2 A more formal description of our model

Let us now discuss the model in more details. In [Cummings et al. 2023], we study a platform that faces a population of n agents. Each agent has a private data point, and has a privacy parameter c_i that models how much agent i cares about their own privacy; the higher the value of c_i, the more stringent agent i's privacy requirements are.

Once the platform has collected data from the agents, it computes an unbiased estimator $\hat{\mu}$ of the mean of the data distribution. Here, we focus on this simple mean estimation task to keep the problem simple and tractable (but note that one may want to also study more complex machine learning tasks). Letting $S \in [n]$ be the set of agents that decide to join the platform, we assume that the platform's

estimator is a linear, weighted empirical mean:

$$\hat{\mu}(S, \vec{w}, \eta) = \sum_{i \in S} w_i d_i + Z(\eta),$$

where w_i is the weight assigned to the data of agent i and Z is a random variable drawn from appropriately chosen noise parameterized by η. The platform aims to optimize over both the choice of weights $\{w_i\}_{i \in S}$ and of noise parameter η, so as to minimize the variance of the estimator. The choice of weights and parameter η impacts the agents' incentives and decisions on whether to participate in the platform and share their data, in two ways:

(1) It affects the level of privacy experienced by each agent i. It is not hard to show this level of privacy is $w_i \eta$; intuitively, the less weight we give to agent i's data in the learner's computation, the more privacy i gets.
(2) It affects the accuracy of the platform's estimate, which in turn impacts the benefit obtained by an agent were they to join the platform.

The question is then the following: how do we design the w_i's and the parameter η optimally, in a way that gives us the most accurate estimate possible?

3. OUR CONTRIBUTIONS

The first contribution of our paper is the model. There has been an extensive line of work on designing algorithms and mechanisms to elicit private data [Roth and Schoenebeck 2012; Chen et al. 2018; Chen and Zheng 2019; Acemoglu et al. 2019; Liao et al. 2022; Gkatzelis et al. 2015; Abernethy et al. 2015; Cai et al. 2015; Liu and Chen 2016; 2017; Liu et al. 2020; Chen et al. 2020; Perote and Perote-Pena 2003a; 2003b; Dekel et al. 2010; Meir and Rosenschein 2011; Meir et al. 2012; Fleischer and Lyu 2012; Nissim et al. 2012; Cummings et al. 2015; Cummings et al. 2021; Ghosh and Roth 2015; Ghosh et al. 2014; Cummings et al. 2015; Liao et al. 2020; Ghosh and Ligett 2013]. However, to the best of our knowledge, our work is the first one with the following unique combination of modeling considerations: i) the agents derive utility from the learner's model rather than from monetary payments, ii) the learner provides *personalized* differential privacy guarantees to the different agents, and iii) the agents' participation decisions are endogenous in that the quality of the model and the utility any given agent gets depend on the other agents' decisions. Closest to our work is the excellent follow-up work of [Fallah et al. 2022] that builds on an earlier version of our model and extends our results to strategic agents that may misreport their privacy preferences.

Second, we provide a decomposition of the problem into two simpler algorithmic components. The first component is that at fixed η, the optimization problem faced by the platform is in fact convex; hence, algorithmically, finding the optimal weights as a function of η is tractable. Then, the second component is simply that since η is a 1-dimensional, real number, it is relatively simple to optimize over.

That said, our results go beyond this algorithmic characterization; in fact, we provide a semi closed-form solution to the problem faced by the platform. Intuitively and informally, we should expect agents with higher c_i's (more stringent privacy requirements) should have lower weights w_i on their data so as to obtain more privacy. Formally, we show that the weights should be chosen as follows:

Informal Theorem 3.1. *The optimal weights are such that:*

(*1*) *Agents with the least stringent privacy preferences, i.e. with a cost c_i below a certain threshold, are pooled together and given the same weight.*

(*2*) *Agents with more stringent privacy preferences, i.e. with a cost c_i above this threshold, see their data being given a weight that decreases inversely proportionally to their privacy requirements.*

In particular, this structure is–perhaps surprisingly–a bit different from what we would have initially expected. On top of this, we believe our semi closed-form characterization to be interesting for several additional reasons:

(1) We show that the above form for the optimal solution holds in two variants of how we model the agents' privacy preferences and participation decisions (please see the manuscript [Cummings et al. 2023] for more details on these models). This form for the optimal solution is also the same one that we observed in [Chen et al. 2018]. This seems to imply that this is not a specificity of our current model, but possibly a deeper and more fundamental property of data elicitation in the presence of privacy considerations.

(2) Further, the structure of our optimal solution is that it is simple and interpretable. This is significant, as optimal mechanisms in many settings can look very complex, to a point that it is not believable that one may want to use such a solution in practice. Having access to simple algorithms and mechanisms that perform well give them practical value.

(3) Finally, because the optimal weights for a given value of η are fairly simple, they facilitate the search over the best possible value of the parameter η. In our work, we use the semi closed-form solution for the weights as a building block to optimize over η and to obtain semi closed-form expressions (that depend on the choice of integer threshold t) for η under some additional assumptions.

Beyond this characterization, we note that our results can be extended to the case in which the agents' privacy costs are unknown, and the learner has to elicit them from strategic agents that may misreport them when beneficial. While our paper touches briefly on such strategic considerations in one of the variants of our model, those are studied more carefully and extensively in the work of [Fallah et al. 2022].

REFERENCES

Abernethy, J., Chen, Y., Ho, C.-J., and Waggoner, B. 2015. Low-cost learning via active data procurement. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. 619–636.

Acemoglu, D., Makhdoumi, A., Malekian, A., and Ozdaglar, A. 2019. Too much data: Prices and inefficiencies in data markets. Tech. rep., National Bureau of Economic Research.

Cai, Y., Daskalakis, C., and Papadimitriou, C. 2015. Optimum statistical estimation with strategic data sources. In *Conference on Learning Theory*. PMLR, 280–296.

Chen, Y., Immorlica, N., Lucier, B., Syrgkanis, V., and Ziani, J. 2018. Optimal data acquisition for statistical estimation. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 27–44.

Chen, Y., Liu, Y., and Podimata, C. 2020. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems 33*, 15265–15276.

Chen, Y. and Zheng, S. 2019. Prior-free data acquisition for accurate statistical estimation. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. 659–677.

Cummings, R., Elzayn, H., Gkatzelis, V., Pountorakis, E., and Ziani, J. 2023. Optimal data acquisition with privacy-aware agents. In *First IEEE Conference on Secure and Trustworthy Machine Learning*.

Cummings, R., Feldman, V., McMillan, A., and Talwar, K. 2021. Mean estimation with user-level privacy under data heterogeneity. In *NeurIPS 2021 Workshop Privacy in Machine Learning*.

Cummings, R., Ioannidis, S., and Ligett, K. 2015. Truthful linear regression. In *Conference on Learning Theory*. PMLR, 448–483.

Cummings, R., Ligett, K., Roth, A., Wu, Z. S., and Ziani, J. 2015. Accuracy for sale: Aggregating data with a variance constraint. In *Proceedings of the 2015 conference on innovations in theoretical computer science*. 317–324.

Dekel, O., Fischer, F., and Procaccia, A. D. 2010. Incentive compatible regression learning. *Journal of Computer and System Sciences 76,* 8, 759–777.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.

E.U. 2016. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Off. J. Eur. Union 119*, 1–88.

Fallah, A., Makhdoumi, A., Malekian, A., and Ozdaglar, A. 2022. Optimal and differentially private data acquisition: Central and local mechanisms. In *Proceedings of the 2022 ACM Conference on Economics and Computation*. 1141.

Fleischer, L. K. and Lyu, Y.-H. 2012. Approximately optimal auctions for selling privacy when costs are correlated with data. In *Proceedings of the 13th ACM conference on electronic commerce*. 568–585.

Ghosh, A. and Ligett, K. 2013. Privacy and coordination: Computing on databases with endogenous participation. In *Proceedings of the fourteenth ACM conference on Electronic commerce*. 543–560.

Ghosh, A., Ligett, K., Roth, A., and Schoenebeck, G. 2014. Buying private data without verification. In *Proceedings of the fifteenth ACM conference on Economics and computation*. 931–948.

Ghosh, A. and Roth, A. 2015. Selling privacy at auction. *Games and Economic Behavior 91*, 334–346.

Gkatzelis, V., Aperjis, C., and Huberman, B. A. 2015. Pricing private data. *Electronic Markets 25,* 2, 109–123.

Liao, G., Chen, X., and Huang, J. 2020. Social-aware privacy-preserving mechanism for correlated data. *IEEE/ACM Transactions on Networking 28,* 4, 1671–1683.

Liao, G., Su, Y., Ziani, J., Wierman, A., and Huang, J. 2022. The privacy paradox and optimal bias-variance trade-offs in data acquisition. *ACM SIGMETRICS Performance Evaluation Review 49,* 2, 6–8.

Liu, Y. and Chen, Y. 2016. Learning to incentivize: Eliciting effort via output agreement. *arXiv preprint arXiv:1604.04928*.

Liu, Y. and Chen, Y. 2017. Sequential peer prediction: Learning to elicit effort using posted prices. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Liu, Y., Wang, J., and Chen, Y. 2020. Surrogate scoring rules. In *Proceedings of the 21st ACM Conference on Economics and Computation*. 853–871.

Meir, R., Procaccia, A. D., and Rosenschein, J. S. 2012. Algorithms for strategyproof classification. *Artificial Intelligence 186*, 123–156.

Meir, R. and Rosenschein, J. S. 2011. Strategyproof classification. *ACM SIGecom Exchanges 10,* 3, 21–25.

Nissim, K., Orlandi, C., and Smorodinsky, R. 2012. Privacy-aware mechanism design. In *Proceedings of the 13th ACM Conference on Electronic Commerce*. 774–789.

Perote, J. and Perote-Pena, J. 2003a. The impossibility of strategy-proof clustering. *Economics Bulletin 4,* 23, 1–9.

Perote, J. and Perote-Pena, J. 2003b. The impossibility of strategy-proof clustering. *Economics Bulletin 4,* 23, 1–9.

Roth, A. and Schoenebeck, G. 2012. Conducting truthful surveys, cheaply. In *Proceedings of the 13th ACM Conference on Electronic Commerce*. 826–843.

Sweeney, L. 2000. Simple demographics often identify people uniquely. *Health (San Francisco) 671,* 2000, 1–34.