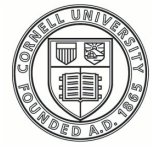
Cornell University
ILR School

Cornell University ILR School
DigitalCommons@ILR

Labor Dynamics Institute | Centers, Institutes, Programs

4-17-2017

Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods

John Abowd
Cornell University, John.Abowd@cornell.edu

Ian M. Schmutte
University of Georgia, schmutte@uga.edu

Follow this and additional works at: https://digitalcommons.ilr.cornell.edu/ldi
Thank you for downloading an article from DigitalCommons@ILR.
Support this valuable resource today!

This Article is brought to you for free and open access by the Centers, Institutes, Programs at DigitalCommons@ILR. It has been accepted for inclusion in Labor Dynamics Institute by an authorized administrator of DigitalCommons@ILR. For more information, please contact catherwood-dig@cornell.edu.

Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods

Abstract
We consider the problem of determining the optimal accuracy of public statistics when increased accuracy requires a loss of privacy. To formalize this allocation problem, we use tools from statistics and computer science to model the publication technology used by a public statistical agency. We derive the demand for accurate statistics from first principles to generate interdependent preferences that account for the public-good nature of both data accuracy and privacy loss. We first show data accuracy is inefficiently under-supplied by a private provider. Solving the appropriate social planner's problem produces an implementable publication strategy. We implement the socially optimal publication plan for statistics on income and health status using data from the American Community Survey, National Health Interview Survey, Federal Statistical System Public Opinion Survey and Cornell National Social Survey. Our analysis indicates that welfare losses from providing too much privacy protection and, therefore, too little accuracy can be substantial.

Comments
Abowd and Schmutte acknowledge the support of **Alfred P. Sloan Foundation Grant G-2015-13903** and **NSF Grant SES-1131848**. Abowd acknowledges direct support from the U.S. Census Bureau (before and during his appointment as Associate Director) and from **NSF Grants BCS- 0941226, TC-1012593**. Some of the research for this paper was conducted using the resources of the Social Science Gateway, which was partially supported by **NSF grant SES-0922005**. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the Census Bureau, NSF, or the Sloan Foundation. We also thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the Programme on Data Linkage and Anonymisation, supported by EPSRC grant no. EP/K032208/1. Abowd also acknowledges the Center for Labor Economics at UC Berkeley, where he was a visiting scholar when this work was initiated. We are grateful for helpful comments from Larry Blume, David Card, Michael Castro, Cynthia Dwork, John Eltinge, Stephen Fienberg, Mark Kutzbach, Ron Jarmin, Dan Kifer, Ashwin Machanavajjhala, Frank McSherry, Gerome Miklau, Kobbi Nissim, Mallesh Pai, Jerry Reiter, Eric Slud, Adam Smith, Bruce Spencer, Sara Sullivan, Lars Vilhuber and Nellie Zhao along with seminar and conference participants at the U.S. Census Bureau, Cornell University, CREST, George Mason University, Georgetown University, University of Washington Evans School of Public Policy, and the Society of Labor Economists. We thank Jennifer Childs and Casey Eggleston for providing data from the Federal Statistical System Public Opinion Survey conducted by the Census Bureau's Center for Survey Methodology. William Sexton provided excellent research assistance. No confidential data were used in this paper.

A complete archive of the data and programs used in this paper is available via http://doi.org/10.5281/zenodo.345385.

A previous version of this paper is http://digitalcommons.ilr.cornell.edu/ldi/22/. This version **supersedes** Document 22.

This article is available at DigitalCommons@ILR: https://digitalcommons.ilr.cornell.edu/ldi/37

Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods

John M. Abowd
U.S. Census Bureau and Department of Economics, Cornell University
john.maron.abowd@census.gov
Ian M. Schmutte
Department of Economics, University of Georgia
schmutte@uga.edu

April 17, 2017

Abowd and Schmutte acknowledge the support of Alfred P. Sloan Foundation Grant G-2015-13903 and NSF Grant SES-1131848. Abowd acknowledges direct support from the U.S. Census Bureau (before and during his appointment as Associate Director) and from NSF Grants BCS-0941226, TC-1012593. Some of the research for this paper was conducted using the resources of the Social Science Gateway, which was partially supported by NSF grant SES-0922005. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the Census Bureau, NSF, or the Sloan Foundation. We also thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the Programme on Data Linkage and Anonymisation, supported by EPSRC grant no. EP/K032208/1. Abowd also acknowledges the Center for Labor Economics at UC Berkeley, where he was a visiting scholar when this work was initiated. We are gratefule for helpful comments from Larry Blume, David Card, Michael Castro, Cynthia Dwork, John Eltinge, Stephen Fienberg, Mark Kutzbach, Ron Jarmin, Dan Kifer, Ashwin Machanavajjhala, Frank McSherry, Gerome Miklau, Kobbi Nissim, Mallesh Pai, Jerry Reiter, Eric Slud, Adam Smith, Bruce Spencer, Sara Sullivan, Lars Vilhuber and Nellie Zhao along with seminar and conference participants at the U.S. Census Bureau, Cornell University, CREST, George Mason University, Georgetown University, University of Washington Evans School of Public Policy, and the Society of Labor Economists. We thank Jennifer Childs and Casey Eggleston for providing data from the Federal Statistical System Public Opinion Survey conducted by the Census Bureau's Center for Survey Methodology. William Sexton provided excellent research assistance. No confidential data were used in this paper. A complete archive of the data and programs used in this paper is available via http://doi.org/10.5281/zenodo.345385.

Abstract

We consider the problem of determining the optimal accuracy of public statistics when increased accuracy requires a loss of privacy. To formalize this allocation problem, we use tools from statistics and computer science to model the publication technology used by a public statistical agency. We derive the demand for accurate statistics from first principles to generate interdependent preferences that account for the public-good nature of both data accuracy and privacy loss. We first show data accuracy is inefficiently under-supplied by a private provider. Solving the appropriate social planner's problem produces an implementable publication strategy. We implement the socially optimal publication plan for statistics on income and health status using data from the American Community Survey, National Health Interview Survey, Federal Statistical System Public Opinion Survey and Cornell National Social Survey. Our analysis indicates that welfare losses from providing too much privacy protection and, therefore, too little accuracy can be substantial.

Keywords: Demand for public statistics; Technology for statistical agencies; Optimal data accuracy; Optimal confidentiality protection

1 Introduction

This paper studies a novel, but important, problem regarding the optimal allocation of information stored in already-existing databases. We focus on the tradeoff between two competing uses of a finite amount of data. On one hand, the data can be used to produce statistical summaries that are germane to decision-makers. However, as is well-understood, publication of statistical summaries necessarily entails increased loss of privacy for those whose information is included in the data (Dinur and Nissim 2003). As more statistics are published, more privacy is lost.[1]

We contribute an economic framework for determining the optimal allocation of the confidential information in a database between accuracy of published statistics and privacy protection. The solution to this problem is based on the familiar intuition that the optimal choice should equate the marginal willingness to pay for increased statistical accuracy with the marginal technical rate at which accuracy can be increased by sacrificing privacy. To date, no such framework exists to help guide statistical agencies and private data custodians like Google or Facebook in their decisions about data access and data publication. Such an analysis requires an ability, first, to formalize the privacy loss associated with any specific data pub-

[1]There is a broad literature in statistical disclosure limitation (SDL) about the link between data accuracy and privacy. See Abowd and Schmutte (2015) for an overview aimed at economists. The privacy-preserving data publication results in computer science go much further. These results, known generically as database reconstruction theorems, differ materially from the re-identification attacks found in the SDL literature in that they use only the information contained in the data curator's own publications–not external information. Dwork et al. (2007) provide an exact bound on the fraction of randomly generated queries that can be answered exactly before exposing the entire database with near certainty. Muthukrishnan and Nikolov (2012) tighten the $O(\sqrt{N})$ lower bound on the cumulative noise required to limit database reconstruction. Kasiviswanathan et al. (2013) prove that the reconstruction attacks in Dinur and Nissim can be extended to M-estimators, and that they are polynomial-time algorithms. To drive the point home, Dwork et al. (2015) show that reconstruction attacks on summary statistics from genome-wide association studies are feasible using a single sample from the reference population.

lication, even when it could be combined with an arbitrary amount of other prior information. Second, it requires a formal description of, and ability to measure, preferences for data accuracy and data privacy. These are daunting challenges.[2]

To make progress, we draw on insights from statistical disclosure limitation (SDL) and recent developments on privacy-preserving data publication in computer science (Dwork 2008; Dwork and Roth 2014). Using these tools, we model the behavior of a public statistical agency, or data custodian, that operates under a dual mandate to produce accurate statistical summaries of the population, but also to protect the privacy of the entities on which it collects data. The data custodian in our model uses a *differentially private* publication mechanism, which characterizes privacy loss in terms of the maximum ex ante change in posterior inferences about a sensitive characteristic.[3] This characterization implies that the privacy guarantee is independent of other sources of information. Furthermore, given the publication technology we consider, the accuracy guarantee can only be improved by relaxing the privacy guarantee. When doing so, the data custodian must choose the maximum amount of privacy loss that can be tolerated, which implies a minimum acceptable level of accuracy. The result is familiar to economists: constrained optimization subject to a production possibilities frontier with two public goods (Samuelson 1954; Mas-Colell et al. 1995).

The key contribution of economics for this problem, and of our paper in particular, is a theoretical and empirical framework for selecting the data publication strategy that yields an optimal combination of statistical accuracy and data privacy. This depends on how the public values access to more accurate statistics

[2]Indeed, Chris Sims has argued that the related, and perhaps more challenging, task of cost-benefit analysis for collection and dissemination of official statistics is logically impossible (Eisen and Kaufman 1985, Chapter 3). Sims claims authorship for this chapter on his professional website http://www.princeton.edu/~sims/.

[3]Heffetz and Ligett (2014) provide an overview to differential privacy accessible to economists.

when doing so entails a loss of privacy. We therefore need a model of the demand for privacy and the demand for accuracy. Here, the existing literature provides less guidance. We propose a simple extension of the basic model of preference interdependence (Akerlof 1997). Utility depends on the consumer's relative position in the distribution of a characteristic of the population, income for example. The twist is that utility is therefore a function of the accuracy of the published statistics that characterize the distribution, as well as the loss of privacy from the existence of such publications.[4]

From this model we derive the willingness to pay for increased privacy in terms of foregone accuracy as a relatively simple function of the second moments in the joint distribution of observed outcomes (income and health status in our applications) and preferences for privacy and accuracy. In Section 5, we use data from two surveys to quantify the parameters of the social planner's problem. We illustrate the optimal strategy for publishing statistics on the distribution of income and body mass index (BMI), a health attribute. For the income application, we characterize the social willingness to pay using new data from the Federal Statistical System Public Opinion Survey (FSS POS), which contains income and proxy measures of preferences for privacy and data accuracy. For the health application, we use data from the Cornell National Social Survey, which also records self-reported health status and preferences for privacy and diagnostic accuracy. Using these statistics, we implement and simulate the data custodians' optimal

[4]Our extension of the standard model of interdependent preferences has some empirical support. It can be difficult to show that relative status affects individual behavior because models of interdependent preferences are not usually identified without restrictive assumptions (Postlewaite 1998; Luttmer 2005). However, Card, Mas, Moretti and Saez (2012) find that the behavior of university faculty responds to changes in the availability of information about their colleagues' salaries. Their findings show that not only are social comparisons a meaningful determinant of labor market behavior, but also highlight that the significance of social comparisons depends on the availability or quality of data on others' salaries.

data publication strategy. We quantify the welfare loss from suboptimal over-provision of privacy protection and under-provision of data accuracy.

Our goal is to draw the attention of researchers working on private-market applications of the economics of privacy and electronic commerce to the issues surrounding the appropriate use of confidential databases by national statistical offices to produce population statistics. The welfare analysis in this paper is novel and thorough, but also, as will be clear, limited. We need better models of the demand for privacy and accuracy, and better data on these preferences. Interest in these topics is not merely academic. At the U.S. Census Bureau, active research is underway to use formal privacy systems like differential privacy in producing official statistics (Abowd, Schmutte and Vilhuber 2017).[5] Similar research is underway in the private sector at Apple, Google, and Microsoft.[6]

Like so many other ideas in information economics, George Stigler (1980) initiated the modern economic analysis of privacy. While Stigler's focus was on the origins of the demand for privacy, he identified the source of angst driving public discussions in the 1970s with his observation that: "[g]overnments (at all levels) are now collecting information of a quantity and in a personal detail unknown in history" (p. 623). In the intervening decades, public concern over the proliferation, use, and possible abuses of governmental databases has persisted, and is now exacerbated by the omnipresent collection by private companies of records on our economic and social interactions. The potential for linking personal information across data sources means a breach of data security in one source could

[5]The U.S. Census Bureau was the first organization in the world to use a variant of differential privacy in a production setting (Machanavajjhala et al. 2008).

[6]Differential privacy was invented at Microsoft (Dwork et al. 2006). Google published the algorithms that allow anyone in the world to query the active websites of any Chrome browser user in a differentially private manner with a toolkit known as RAPPOR (Erlingsson et al. 2014). Apple has announced the use of differential privacy in iOS 10 (*Apple previews iOS 10, the biggest iOS release ever* 2016), but has not released details.

lead to financially or socially ruinous consequences through what Ohm (2010) calls a "database of ruin."

Stigler correctly observed that a key challenge was to properly constrain the use of this information rather than to obstruct its acquisition in the first place. Such obstruction is likely to be futile given the nature of technological change. More importantly, the information in governmental and private databases is of great value. Acquisti and Varian (2005) note, for example, that the privileged informational position of sellers in this market allows individual-level price discrimination on a massive basis. Consumers may have a strong interest in concealing the data that allow this price customization.

As the economy grows, so does the demand for official statistics published with greater frequency, in finer detail, and with more accuracy. At the same time, the public appetite, at least in the U.S., to fund data collection and to participate in surveys has waned. These trends have led to the rapid and increasing use of administrative data to augment the production of official statistics and the exploration of alternative data sources.[7] Our paper formalizes how the use of administrative data, while less burdensome to the public in pecuniary terms as well as the demand for their time, still burdens the citizenry through foregone privacy. In our formalization, that burden is explicit and measurable.

Our work builds directly on the electronic commerce literature by formally addressing the public-good nature of both the accuracy of published statistics and the protection of privacy in statistical databases. Ghosh and Roth (2011) and Hsu, Gaboardi, Haeberlen, Khanna, Narayan, Pierce and Roth (2014) model private provision of statistical summaries for use by a private analyst. The latter paper derives cost-minimizing mechanisms to compensate individuals for lost privacy

[7] A report from tge National Academies of Sciences, Engineering, and Medicine (2017) documents these trends, and considers some of the privacy issues we raise here.

when their data are used in a published statistic. Both papers, like the formal privacy literature more generally, fail to address the public-good nature of published data and privacy protection.

Our application is focused on the publication of official statistics, where it is natural to think of the accuracy of those publications as being both non-rival and non-excludable in consumption. Stigler noted that statistical summaries produced from government databases would be public goods. A recent review of the economics of privacy touches briefly on these public-good considerations (Acquisti et al. 2016). In Section 3, we show that when accuracy and privacy are both public goods, the former will be under-provided. A private producer using the correct technology may internalize the full cost of privacy protection, but not the full demand for data accuracy. This example shows why private provision of national statistics may fail, and also sheds light on why public statistical agencies may face incentives to over-weight privacy protection in official publications (National Academies of Sciences, Engineering, and Medicine 2017, Chapter 5).

That privacy protection is also a public good was foreshadowed by Dwork (2008, p. 3) when she wrote: "[t]he parameter ε ... is public. The choice of ε is essentially a social question." This observation has not previously made its way into models of the market for privacy rights. The form of privacy protection we use guarantees a provable limit on the information that can be disclosed about any individual in the population, now or in the future, using any feasible realization of the confidential data. As such, we argue that it satisfies a reasonable interpretation of the legal notion of equal protection under the law–all persons in the population represented by the confidential database receive the same confidentiality protection. In addition, the benefits from increased privacy protection for any individual in the population are automatically enjoyed by every other in-

dividual, whether that person's information is used or not–the privacy protection is therefore strictly non-rival.

The paper is organized as follows. Section 2 presents definitions and the basic model. Section 3 proves the under-provision of data accuracy when the privacy-preserving data publication system is run by a private firm, as in the classic applications in electronic commerce. Section 4 solves the appropriate social planner's problem for data accuracy and privacy protection when both are public goods. Section 5 presents the results of applying the full social optimum to the publication of income distribution and health statistics. Section 6 concludes.

2 Basic Model

We model the publication of statistics by a national agency from a confidential database for which it is the trusted custodian. The agency's publication method yields a technological frontier that describes the rate at which privacy must be sacrificed to increase the accuracy of published statistics. The optimal choice along this frontier depends upon the willingness of individuals to pay for increased accuracy with reduced privacy protection.

Our model assumes that the data have already been collected, subject to a fixed collection technology. We do not, therefore model the design of the data collection process, nor the monetary costs of collection.[8] By deliberately abstracting from these technical public-finance problems, we focus exclusively on the disutility that arises from the foregone privacy. The collected data are a resource designated for producing population statistics. Administrative data are collected in the process

[8]The design of many large-scale surveys is produced, in part, by minimizing the cost of estimating quantities for diverse subpopulations. Controlling the margin of error in these subpopulations is an important component of this survey-cost management strategy. Although design optimization may interact with overall data accuracy, we have abstracted from this consideration.

of managing public programs. Our analysis describes how the information embodied in both of these data sources should be optimally allocated between privacy protection and production of accurate statistics. In this paradigm, the social cost of data accuracy is measured in terms of the privacy loss when the agency publishes data, not when it collects those data.[9]

In addition to introducing our basic model, this section provides all formal definitions used in our application of differential privacy. We highlight tools that may be unfamiliar to economists and statisticians. Our summary draws on several sources to which we refer the reader who is interested in more details (Hardt and Rothblum 2010; Dwork and Roth 2014; Wasserman and Zhou 2010). Our notation follows Dwork and Roth (2014).

2.1 Databases, Histograms, and Queries

A data curator possesses a database, D. We model D as a table in which each row represents information for a single individual and each column represents a single characteristic to be measured. The database D contains N rows. The set χ describes all possible values the variables in the columns of the database can take. Therefore any row that appears in the database is an element of χ.[10] All variables

[9] Our conflation of the terms privacy and confidentiality is at odds with their conventional uses in official statistics. Former Census Bureau Director Kenneth Prewitt (2011), reflecting the position of many official statisticians, argues for using the term "don't ask" to reflect the privacy considerations associated with the design of the questionnaire, and "don't tell" for the promise to keep individuals' responses confidential in publications arising from the survey. While we agree with the ethical and public policy considerations embodied in this distinction, the conflation of privacy and confidentiality in the privacy-preserving data publication literature, and in this paper, reflects the view that the respondent has lost control of the data item once it is provided (e.g., in a survey), or provided it for a different purpose than the statistical agency's use (e.g., in administrative data). In this context, there is very little difference between privacy protection and confidentiality protection.

[10] In statistics, χ is the sample space. All structural zeros (combinations of values that are deemed impossible a priori) are removed from χ. For example, if the variables recorded in the database

are discrete and finite-valued, which is not restrictive since continuous data are always given discrete, finite representations when recorded on censuses, surveys, or administrative record systems.

2.1.1 Histogram Representation

For our analysis, we represent the database D by its unnormalized histogram $x \in \mathbb{Z}^{*|\chi|}$. The notation $|\chi|$ represents the cardinality of the set χ, and $\mathbb{Z}^*$ is the set of non-negative integers. Each entry in x, x_i, is the number of elements in the database D of type $i \in \chi$. We use the ℓ_1 norm:

$$||x||_1 = \sum_{i=1}^{|\chi|} |x_i| \,. \tag{1}$$

Observe that $||x||_1 = N$, the number of records in the database. Given two histograms, x and y, $||x - y||_1$ measures the number of records that differ between x and y. We define *adjacent histograms* as those for which the ℓ_1 distance is at most 1.[11]

2.1.2 Queries

A *linear query* is a mapping $f : [0,1]^{|\chi|} \times \mathbb{Z}^{*|\chi|} \to \mathbb{R}^*$ such that $f(m, x) = m^T x$ where $x \in \mathbb{Z}^{*|\chi|}$, $m \in [0,1]^{|\chi|}$, and $\mathbb{R}^*$ is the set of non-negative real numbers. A *counting query* is a special case in which m_i is restricted to take a value in $\{0, 1\}$. Counting

are a binary indicator for gender, $g \in \{0, 1\}$, and a categorical indicator for six different levels of program eligibility, $s \in \{1, \ldots, 6\}$, then $\chi = \{0, 1\} \times \{1, \ldots, 6\}$, and $|\chi| = 12$. If the pair (0,6) is impossible, then $\chi = \{0, 1\} \times \{1, \ldots, 5\} \cup (1, 6)$, and $|\chi| = 11$.

[11] If x is the histogram representation of D, y is the histogram representation of D', and D' is constructed from D by deleting exactly one row, then $||x - y||_1 = 1$. So, D and D' are adjacent databases and x and y are the adjacent histogram representations of D and D', respectively. Some caution is required when reviewing the related literature because definitions may be stated in terms of adjacent databases or adjacent histograms.

queries return the number of observations that satisfy particular conditions. They are the tool an analyst would use to calculate multidimensional margins from the contingency table representation of the database, which is precisely our histogram representation. A *normalized linear query* is a mapping $f : [0,1]^{|\chi|} \times \mathbb{Z}^{*|\chi|} \rightarrow [0,1]$ such that if $\tilde{f}$ is a linear query, then $f(m,x) = \tilde{f}(m,x)/||x||_1$.

We model queries about population proportions rather than counts. These correspond to the proportions from a saturated contingency table. To that end, we work with normalized linear queries unless otherwise specified. The use of normalization is not restrictive. It affects the functional form of privacy and accuracy bounds only through their dependence on the database size $||x||_1$. Any bound stated in terms of the unnormalized histograms and queries can be restated in terms of normalized histograms and queries.

2.2 Query Release Mechanisms, Privacy, and Accuracy

We model the data release mechanism as a randomized algorithm. The data curator operates an algorithm that provides answers to a set of k normalized linear queries drawn from the query set $\mathcal{Q}$.

Definition 1 (Query Release Mechanism) Let $\mathcal{Q}$ be a set of normalized linear queries, and $|\mathcal{Q}|$ the number of queries to be answered. A query release mechanism M is a random function $M : \mathbb{Z}^{*|\chi|} \times \mathcal{Q} \rightarrow [0,1]^{|\mathcal{Q}|}$ whose inputs are a histogram $x \in \mathbb{Z}^{*|\chi|}$ and a set of normalized linear queries $\mathcal{Q}$ with cardinality $|\mathcal{Q}|$. The mechanism output consists of responses to the $|\mathcal{Q}|$ queries. The probability of observing $B \subseteq [0,1]^{|\mathcal{Q}|}$ is $\Pr[M(x,\mathcal{Q}) \in B | x, \mathcal{Q}]$, the conditional probability, given x and $\mathcal{Q}$, that the published query answer is in $B \in \mathcal{B}$, where $\mathcal{B}$ are the measurable subsets of $[0,1]^{|\mathcal{Q}|}$.

Differential Privacy

Our definitions of differential privacy and accuracy for the query release mechanism follow Hardt and Rothblum (2010) and Dwork and Roth (2014).

Definition 2 (ε-differential privacy) Query release mechanism M satisfies ε-differential privacy if for $\varepsilon > 0$, for all $x, x' \in N_x$, all $\mathcal{Q}$, and for all $B \in \mathcal{B}$

$$\Pr\left[M(x, \mathcal{Q}) \in B | x, \mathcal{Q}\right] \leq e^{\varepsilon} \Pr\left[M(x', \mathcal{Q}) \in B | x', \mathcal{Q}\right],$$

where $N_x = \left\{(x, x') \; s.t. \; x, x' \in \mathbb{Z}^{*|\chi|} \text{ and } ||x - x'||_1 = 1\right\}$ is the set of all *adjacent histograms* of x, and where $\mathcal{B}$ are the measurable subsets of the query output space.

Accuracy

We can now define our measure of accuracy. For each query, $f_k \in \mathcal{Q}$, the query release mechanism returns an answer, a_k, that depends on the input database, the content of the query response, and the randomization induced by the query release mechanism.

Definition 3 ((α, β)-accuracy) Query release mechanism M satisfies (α, β)-accuracy if for $f_k \in \mathcal{Q}$ and a_k output from $M(x, \mathcal{Q})$,

$$\min_{1 \leq k \leq |\mathcal{Q}|} \left\{\Pr\left[|a_k - f_k(x)| \leq \alpha | x, \mathcal{Q}\right]\right\} \geq 1 - \beta.$$

This definition guarantees that the error in the answer provided by the mechanism is bounded above by α with probability $(1 - \beta)$ for the entire set of $|\mathcal{Q}|$ queries. The probabilities in the definition of (α, β)-accuracy are induced by the query release mechanism.

2.3 Interpretation

We now clarify the relationship between differential privacy and inferential disclosure. Without loss of generality, the histogram x can be treated as the realization of a random variable with $Multinomial(N, \pi)$ distribution, where the probabilities π are defined over χ, N is the number of records in the database, and the query set $\mathcal{Q}$ is given. Then x' is the same realization with exactly one record deleted, for an individual who contributed one count to cell x_ℓ of the histogram. Using Definition 2, consider the ratio that results from using the query release mechanism on these two adjacent histograms, conditional on x, x', and $\mathcal{Q}$.

Compute $\Pr[x\,|\pi, N, \mathcal{Q}]$ and $\Pr[x'\,|\pi, N-1, \mathcal{Q}]$ from the Multinomial assumption. A direct application of Bayes Theorem yields

$$e^{-\varepsilon} \leq \frac{\Pr[M(x,\mathcal{Q}) \in B\,|x,\mathcal{Q}]}{\Pr[M(x',\mathcal{Q}) \in B\,|x',\mathcal{Q}]} = \frac{\frac{\Pr[x|B,\pi,N,\mathcal{Q}]}{\Pr[x'|B,\pi,N-1,\mathcal{Q}]}}{\frac{\Pr[x|\pi,N,\mathcal{Q}]}{\Pr[x'|\pi,N-1,\mathcal{Q}]}} \leq e^{\varepsilon}, \tag{2}$$

where the left-hand side is the ε-differential privacy condition, the numerator of the right-hand-side is the posterior odds of the confidential database being x versus x' after B is released, and the denominator is the prior odds.[12] But the right-hand side simplifies to the odds of x_ℓ versus $x_\ell - 1$ given B, π, N, and $\mathcal{Q}$ relative to the odds of x_ℓ versus $x_\ell - 1$, given only π, N, and $\mathcal{Q}$. This odds ratio is the Bayes factor for updating any hypothesis about the deleted individual on publication of B. The Bayes factor for all individuals is bounded between $e^{-\varepsilon}$ and e^{ε} by the properties of differential privacy. By the symmetry of the definition of differential privacy, this result holds whether we add or remove an individual from the

[12] The presence of the lower bound $e^{-\varepsilon}$ occurs because x' is explicitly constructed by removing one row from the database D used to construct the histogram x. The direction of change in the posterior odds is not determined a priori. Increases and decreases are both bounded by the definition of differential privacy because x and x' are interchangeable.

population N.

The bound on the Bayes factor for all individuals in the population, and for all possible populations, states precisely the Duncan and Lambert (1986) formalization of inferential disclosure. Note, especially, it is impossible to prevent a disclosure in the sense considered by Dalenius (1977): "[i]f the release of the statistics S makes it possible to determine the value of [the confidential data item] more accurately than is possible without access to S, a disclosure has taken place..." (p.433). Dwork and Naor (2010) prove that inferential disclosure in this sense is impossible to prevent. In the language of cryptography, the trusted data curator must leak some information about the confidential data because the release of statistics that fully encrypt those data ($\varepsilon = 0$) would be worthless. In the language of economics, some risk of privacy breach is the marginal social cost of releasing any useful statistical information from the confidential database.[13] It should now be clear why we characterize differential privacy as worst-case privacy protection. Because the definition applies to all histograms of size N that could have been realized, all individuals who may or may not have been used to compute the statistics B, and all possible statistical releases in $\mathcal{B}$, the Bayes factor for all allowable configurations of population characteristics is bounded by e^{ε}. The worst privacy breach that can ever occur for any individual in any population described by χ is the logarithm of the bound on the Bayes factor in equation 2, namely ε.

[13] Dwork and Naor (2010) use the cryptographic language that perfect semantic security (Goldwasser and Micali 1982; 1984) is impossible in any privacy-preserving data publication system where the utility of the published statistics depends upon their accuracy. Perfectly semantically secure published data must be fully encrypted, and therefore have no accuracy when used in any subsequent analysis or decision. Perfect semantic security is equivalent to zero inferential disclosures. Evfimievski et al. (2003) developed a similar approach based on posterior probabilities. Kifer and Machanavajjhala (2011) clarified the role of the probability assumptions when using formal privacy methods by showing that the semantics associated with any particular Bayes factor bound depend upon the assumptions used to specify the joint distribution of all the rows in the database.

3 The Suboptimality of Private Provision

Using the differential privacy framework, we explicitly illustrate the potential for suboptimal private provision of public statistical data by adapting the very innovative model of Ghosh and Roth (2011). Ghosh and Roth (GR, hereafter) show that differential privacy can be priced as a commodity using a formal auction model. They prove the existence of a mechanism that yields the lowest-cost method for answering a database query with ε-differential privacy and (α, β)-accuracy.[14]

Their model takes the desired query accuracy as exogenous. The producer of the statistic purchases data-use rights from individuals whose data are already in the population database for the purpose of calculating a single statistic–the answer to one database counting query–that will then be published in a scientific paper. Funds for the purchase of the data-use rights come from a grant held by the scientist. GR assume that the statistical release is the private good of the purchaser of the data-use rights.

In this section, the accuracy of the statistic computed via the GR mechanism is a public good whose demand is endogenous. We show that private provision results in a suboptimally low level of accuracy and too much privacy protection. That is, we show that allowing the quality of the scientific research modeled in GR to matter to the population being studied results in an external benefit from the data publication that their model does not capture.

To model the demand for accuracy, we assume that the published statistical data deliver utility to the consumers from whom the rights to use the confidential inputs were purchased. The purchase of data-use rights takes the form of a payment to all consumers who agree to sell their data-use rights when the publication

[14] They prove their results for $\beta = 1/3$, but note that generalizing this is straightforward. See Dwork and Roth (2014, pp. 207-213) for this generalization.

mechanism delivers ε-differential privacy. The value of the published statistical data to all consumers, whether they sell their data-use rights or not, depends upon the accuracy of those data. Furthermore, this accuracy is a public good–it summarizes the quality of the information that any consumer may access and use without reducing its accuracy for some other consumer (it is non-rival), and no consumer can block another consumer's use (it is non-excludable). In plain English, the other scientists and general readers of the papers published in the GR world learn something too. They value what they learn. And they understand that what they learn is more useful if it is more accurate.

Suboptimal private provision of data accuracy is caused by the external benefit of data accuracy to all consumers that is not captured in the GR model. We formally model the demand for data accuracy. The demand for privacy protection, on the other hand, is derived from the private data publisher's cost-minimization problem. In the competitive equilibrium for privately-provided data accuracy, a supplier using a Vickrey-Clarke-Groves (VCG) mechanism buys just enough privacy-loss rights to sell the data accuracy to the consumer with the highest data-accuracy valuation. All other consumers use the published data for free.[15]

3.1 Model Setup

Following Ghosh and Roth (2011), each of N private individuals possesses a single bit of information, b_i, that is already stored in a database maintained by a trusted curator. In addition to their private information, each individual is endowed with income, y_i. Individuals each consume one unit of the published statistic, which has accuracy I defined in terms of (α, β)-accuracy, that is $I = (1 - \alpha)$. Since I is a

[15]Study of this case may be of special interest for some business-data collection for industries with a small number of dominant organizations.

public good, all consumers enjoy the benefits of I, but each consumer is charged the market price p_I, to be determined within the model, for her "share" of I, which we denote I_i, and the balance of the public good, which we denote $I^{\tilde{}i}$ is paid for by the other consumers. Thus, $I = I_i + I^{\tilde{}i}$ for all consumers.

The preferences of consumer i are given by the indirect utility function

$$v_i\left(y_i, \varepsilon_i, I_i, I^{\tilde{}i}\right) = \ln y_i + p_\varepsilon \varepsilon_i - \gamma_i \varepsilon_i + \eta_i \left(I_i + I^{\tilde{}i}\right) - p_I I_i. \tag{3}$$

Equation (3) implies that preferences are quasilinear in data accuracy, I, privacy loss, ε_i, and log income, $\ln y_i$.[16] Income and accuracy are added to the Ghosh and Roth utility function because they are required for the arguments in this section. In Section 4 we develop a more complete model of the demand for accurate public-use statistics. The term p_ε is the common price per unit of privacy, also to be determined by the model. The receipt $p_\varepsilon \varepsilon_i$ represents the total payment an individual receives if her bit is used in an ε-differentially private mechanism. The individual's marginal preferences for data accuracy (a "good") and privacy loss (a "bad," really an input here), $(\gamma_i, \eta_i) > 0$, are not known to the data provider, but their population distributions are public information. Therefore, the mechanism for procuring privacy has to be individually rational and dominant-strategy truthful.

We do not include any explicit interaction between the publication of statistical data and the market for private goods. This assumption is not without consequence, and we make it to facilitate exposition of our key point, which is that data accuracy may be under-provided due to its public-good properties. Viola-

[16] In this section, we keep the description of preferences for data accuracy and privacy protection as close as possible to the original Ghosh and Roth specification. They allow for the possibility that algorithms exist that can provide differential privacy protection that varies with i; hence ε_i appears in equation (3). They subsequently prove that $\varepsilon_i = \varepsilon$ for all i in their Theorem 3.3.

tions of privacy might affect the goods market through targeted advertising and price discrimination as noted in Section 1. The accuracy of public statistics may also spill over to the goods market by making firms more efficient. We reserve consideration of these topics for future work.

In what follows we present the GR results using our notation and definitions. See Appendix A.2 for a complete summary of the translation from their notation and definitions to ours.

3.2 Cost of Producing Data Accuracy

A supplier of statistical information wants to produce an (α, β)-accurate estimate, $\hat{s}$, of the population statistic

$$s = \frac{1}{N} \sum_{i=1}^{N} b_i \tag{4}$$

i.e., a normalized query estimating the proportion of individuals with the property encoded in b_i. Theorems 3.1 and 3.3 in GR prove that the estimator

$$\hat{s} = \frac{1}{N} \left[\sum_{i=1}^{H} b_i + \frac{\alpha N}{2} + \text{Lap}\left(\frac{1}{\varepsilon}\right) \right] \tag{5}$$

with $(\alpha, 1/3)$-accuracy requires a privacy loss of $\varepsilon_i = \varepsilon = \frac{1/2+\ln 3}{\alpha N}$ from $H = N - \frac{\alpha N}{1/2+\ln 3}$ members of the population. The term $\text{Lap}\left(\frac{1}{\varepsilon}\right)$ represents a draw from the Laplace distribution with mean 0 and scale parameter $\frac{1}{\varepsilon}$.

GR prove that purchasing the data-use rights from the H least privacy-loving members of the population; *i.e.*, those with the smallest γ_i, is the minimum-cost, envy-free implementation mechanism.[17] They provide two mechanisms for im-

[17] We note for completeness that the statistic $\hat{s}$, while computed on only H cases from the population of N, is evaluated relative to the population quantity s. GR use the same accuracy measure as we do; namely Definition 3 with a single query in the query set. Statisticians often use mean

plementing their VCG auction. We rely on their mechanism *MinCostAuction* and the properties given in their Proposition 4.5. See Appendix A.2 for additional details.

We now derive the producer's problem of providing the statistic for a given level of data accuracy, which we denote by $I = (1 - \alpha)$. If p_ε is the payment per unit of privacy loss, the total cost of production is $c(I) = p_\varepsilon H\varepsilon$, where the right-hand side terms can be defined in terms of I as follows. Using the arguments above, the producer must purchase from $H(I)$ consumers the right to use their data to compute $\hat{s}$. Then,

$$H(I) = N - \frac{(1-I)N}{1/2 + \ln 3}. \tag{6}$$

Under the VCG mechanism, the price of privacy loss must be $p_\varepsilon = Q\left(\frac{H(I)}{N}\right)$, where Q is the quantile function with respect to the population distribution of privacy preferences, F_γ. The lowest price at which the fraction $\frac{H(I)}{N}$ of consumers do better by selling the right to use their bit, b_i, with $\varepsilon(I)$ units of differential privacy is p_ε. $H(I)$ is increasing in I. The total cost of producing I is

$$C^{VCG}(I) = Q\left(\frac{H(I)}{N}\right) H(I)\varepsilon(I), \tag{7}$$

where the production technology derived by GR implies

$$\varepsilon(I) = \frac{1/2 + \ln 3}{(1-I)N}. \tag{8}$$

squared error instead of the absolute error embodied in this definition. Nevertheless, the statistic $\hat{s}$ trades-off bias and variance relative to the correct population statistic. The term $\alpha N/2$ is a bias correction.

3.3 Private, Competitive Supply of Data Accuracy

Suppose a private profit-maximizing, price-taking, firm sells $\hat{s}$ with accuracy $(\alpha, 1/3)$, that is, with data accuracy $I = (1-\alpha)$ at price p_I. Then, profits $P(I)$ are

$$P(I) = p_I I - C^{VCG}(I).$$

If it sells at all, it will produce I to satisfy the first-order condition $P'\left(I^{VCG}\right) = 0$ implying

$$p_I = Q\left(\frac{H(I)}{N}\right) H(I)\varepsilon'(I) + \left[Q\left(\frac{H(I)}{N}\right) + Q'\left(\frac{H(I)}{N}\right)\left(\frac{H(I)}{N}\right)\right] H'(I)\varepsilon(I) \quad (9)$$

where the solution is evaluated at I^{VCG}.[18] The price of data accuracy is equal to the marginal cost of increasing the amount of privacy protection–data-use rights–that must be purchased. There are two terms. The first term is the increment to marginal cost from increasing the amount each privacy-right seller must be paid because ε has been marginally increased, thus reducing privacy protection for all. The second term is the increment to marginal cost from increasing the number of people from whom data-use rights with privacy protection ε must be purchased. As long as the cost function is strictly increasing and convex, the existence and uniqueness of a solution is guaranteed.

[18]The second order condition is $P''\left(I^{VCG}\right) < 0$, or $\frac{d^2 C^{VCG}(I)}{dI^2} > 0$. The only term in the second derivative of $C^{VCG}(I)$ that is not unambiguously positive is $\frac{H(I)H'(I)^2\varepsilon(I)}{N^2} Q''\left(\frac{H(I)}{N}\right)$. We assume that this term is dominated by the other, always positive, terms in the second derivative. Sufficient conditions are that $Q()$ is the quantile function from the log-normal distribution (as we assume in Section 4) or the quantile function from a finite mixture of normals, and that $\frac{H(I)}{N}$ is sufficiently large; *e.g.*, large enough so that if $Q()$ is the quantile function from the $\ln N\left(\mu, \sigma^2\right)$ distribution, $Q^{*\prime\prime}\left(\frac{H(I)}{N}\right) + \sigma^2 Q^{*\prime}\left(\frac{H(I)}{N}\right)^2 \geq 0$, where $Q^*()$ is the standard normal quantile function.

3.4 Competitive Market Equilibrium

At market price p_I, consumer i's willingness to pay for data accuracy will be given by solving

$$\max_{I_i \geq 0} \eta_i \left(\tilde{I^i} + I_i \right) - p_I I_i \tag{10}$$

where $\tilde{I^i}$ is the amount of data accuracy provided from the payments by all other consumers, as noted above. Consumer i's willingness to pay is non-negative if, and only if, $\eta_i \geq p_I$; that is, if the marginal utility from increasing I exceeds the price. If there exists at least one consumer for whom $\eta_i \geq p_I$, then the solution to equation (9) is attained for $I^{VCG} > 0$.

We next show that there is only one such consumer. It is straightforward to verify that the consumers are playing a classic free-rider game (Mas-Colell et al. 1995, pp. 361-363). In the competitive equilibrium, the only person willing to pay for the public good is one with the maximum value of η_i. All others will purchase zero data accuracy but still consume the data accuracy purchased by this lone consumer. Specifically, the equilibrium price and data accuracy will satisfy

$$p_I = \bar{\eta} = \frac{dC^{VCG}\left(I^{VCG}\right)}{dI},$$

where $\bar{\eta}$ is the maximum value of η_i in the population–the taste for accuracy of the person who desires it the most. However, the Pareto optimal consumption of data accuracy, I^0, solves

$$\sum_{i=1}^{N} \eta_i = \frac{dC^{VCG}\left(I^0\right)}{dI}. \tag{11}$$

Marginal cost is positive, $\frac{dC^{VCG}\left(I^0\right)}{dI} > 0$, and $\sum_{i=1}^{N} \eta_i \geq \bar{\eta}$; therefore, data accuracy will be under-provided by a competitive supplier when data accuracy is a public good as long as marginal cost is increasing, which we prove below.

More succinctly, $I^{VCG} \leq I^0$. Therefore, privacy protection must be over-provided, $\varepsilon^{VCG} \leq \varepsilon^0$, by equation (8).[19]

For readers familiar with the data privacy literature, we note that the statement that technology is given by equations (7) and (8) means that the data custodian allows the producer to purchase data-use rights with accompanying privacy loss of $\varepsilon = \frac{1/2+\ln 3}{(1-I)N}$ from $H(I)$ individuals for the sole purpose of computing $\hat{s}$ via the query response mechanism in equation (5) that is $\frac{1/2+\ln 3}{(1-I)N}$-differentially private and achieves $(1-I, \frac{1}{3})$-accuracy, which is exactly what Ghosh and Roth prove.

3.5 Proof of Suboptimality

Theorem 1 If preferences are given by equation (3), the query response mechanism satisfies equation (8) for ε-differential privacy with $\left(1-I, \frac{1}{3}\right)$-accuracy, cost functions satisfy (7) for the VCG mechanism, the population distribution of γ is given by F_γ (bounded, absolutely continuous, everywhere differentiable, and with quantile function Q satisfying the conditions noted in Section 3.3), the population distribution of η has bounded support on $[0, \bar{\eta}]$, and the population in the database is represented as a continuum with measure function H (absolutely continuous, everywhere differentiable, and with total measure N) then $I^{VCG} \leq I^0$, where I^0 is the Pareto optimal level of I solving equation (11), and I^{VCG} is the privately-provided level when using the VCG procurement mechanism.

Proof. The proof appears in Appendix A.1. ■

[19]The reader is reminded that a smaller ε implies more privacy protection. It is also worth commenting that in the GR formulation the single consumer with positive willingness to pay is the entity running the VCG auction. That person is buying data-use rights from the other consumers, computing the statistic for publication, then releasing the statistic so that all other consumers may use it. That is why we have modeled this as a public good. It is fully consistent with GR's scientist seeking data for a grant-supported publication.

4 The Optimal Provision of Accuracy and Privacy

Having shown that both data quality and privacy loss have public-good properties when modeled using private supplier markets, we now formalize the problem of choosing their optimal levels. We invoke the classic public goods model of Samuelson (1954) as explicated in Mas-Colell et al. (1995, pp. 359-361) to solve for the Pareto optimal quantities of each public good.

4.1 Modeling Production Possibilities

We model a data custodian tasked with releasing public statistics calculated from a confidential database, D. The database contains a measurement from a domain χ for each member of a population of size N. As in the formal privacy literature, we assume the domain is finite. In practice, the set of acceptable values for continuous data is always finite. As in Section 2, x is the histogram representation of D.

The custodian publishes the results of a set of linear queries using an ε-differentially private mechanism. This formalization generalizes the conventional task of publishing contingency tables from underlying microdata. Our goal is to determine how the custodian should set the privacy-loss parameter, ε, to optimally allocate data information between privacy protection and accuracy of the published statistics. For concreteness, in what follows, we assume the custodian operates the Multiplicative Weights Exponential Mechanism (MWEM) introduced by Hardt, Ligett and McSherry (2012). However, our analysis is generally valid for all differentially private mechanisms that yield a convex relationship between privacy loss and accuracy.[20]

[20]One such mechanism is the Private Multiplicative Weights (PMW) mechanism, due to Hardt and Rothblum (2010), which is very similar to MWEM, but for a setting in which users address queries to the underlying database interactively. The theoretical accuracy guarantee of PMW is

4.1.1 The Multiplicative Weights Exponential Mechanism

We summarize here the basic features of MWEM. A more complete description appears in Appendix A.4. To operate MWEM on database x, the custodian chooses a set, $\mathcal{Q}$, of feasible normalized linear queries to publish, and sets the privacy-loss parameter ε.

To understand MWEM, it is useful to first consider a simpler, but less efficient, algorithm: the Laplace mechanism. One can think of the parameter ε as representing a fixed privacy-loss budget to be allocated across answers to various queries. The simplest approach is to calculate the answer to each query using the true data. The custodian can guarantee ε-differential privacy by publishing the true answer plus a random error drawn from the Laplace distribution with scale parameter $\frac{|\mathcal{Q}|}{\varepsilon}$. This approach works because the LaPlace mechanism for ε-differential privacy composes additively (for a proof see Dwork and Roth (2014, pp. 49-51)). When the set of queries is large, or the extent of privacy loss is low, the amount of noise added by the Laplace mechanism is correspondingly large.

MWEM economizes on the expenditure of the privacy-loss budget relative to the Laplace mechanism as follows. The algorithm stores the true data histogram and a synthetic histogram of the same size that is derived from the confidential data according to the following procedure. The synthetic histogram can be initialized with a uniform distribution across cells, and the weights mentioned below can be initialized at unity. For each of a finite number of rounds, the algorithm computes every query on the true and the synthetic histograms. Each query's score is the absolute value of the difference between the true answer and the answer from the synthetic histogram. The algorithm selects a query at random with

qualitatively similar to MWEM. We prefer MWEM for this analysis because the interactive setting envisioned by PMW is a less common form of data publication for public statistical agencies and also because, as far as we know, there is no practical implementation of PMW.

probability proportional to the query score, so that queries approximated poorly by the synthetic data are at higher risk of selection. The algorithm adds Laplace noise to the selected query's true answer. Then, the algorithm updates its weights so that the entries in the synthetic database match the noisy query responses. In MWEM, the privacy budget is drawn down only for queries that are answered poorly in the synthetic data. Upon completion, the custodian can publish answers to all queries, or the synthetic data, or both.

The strengths of MWEM relative to the Laplace mechanism are twofold. First, the approximation to the true histogram minimizes error, given the queries already answered. Second, the algorithm only adds noise when the approximate (*i.e.*, already public) answer is sufficiently far from the truth. Doing so conserves on privacy loss and controls the total error efficiently.

4.1.2 The Feasible Trade-off between Privacy Loss and Accuracy

The MWEM algorithm delivers an increasing and convex relationship between privacy loss and accuracy. That is, to increase accuracy, it is necessary to increase privacy loss, and there are diminishing returns to increasing privacy loss in obtaining increased accuracy. MWEM therefore provides the basis for a well-defined production possibilities frontier.

Theorem 2 Given histogram x with $\|x\|_1 = N$, query set $\mathcal{Q}$, accuracy failure rate $0 \leq \beta \leq 1$, and privacy-loss parameter $\varepsilon > 0$, MWEM satisfies the following conditions:

1. *Privacy*: MWEM satisfies ε-differential privacy;

2. *Accuracy*: MWEM satisfies (α, β)-accuracy, with

$$\alpha = \frac{K(\beta, |\chi|, |\mathcal{Q}|, N)}{\varepsilon^b}. \tag{12}$$

Furthermore, the constant K is decreasing in N and increasing in $|\chi|$ and $|\mathcal{Q}|$. For MWEM, the parameter $b = \frac{1}{3}$.

Proof. The proof appears in Appendix A.1. ■

4.1.3 The Production Possibilities Frontier

We show here that the accuracy guarantee obtained in Theorem 2 has a direct interpretation as a production possibilities frontier (PPF). The key accuracy parameter is α, which measures the worst-case deviation on a single query. Higher values of α correspond to lower accuracy.

We define data accuracy as $I = (1 - \alpha)$, and characterize the PPF between I and differential privacy loss, ε, by a transformation function

$$G(\varepsilon, I) \equiv I - \left[1 - \frac{K(\beta, |\chi|, |\mathcal{Q}|, N)}{\varepsilon^b}\right] \tag{13}$$

where the functional form of K is given in the proof of Theorem 2. All feasible pairs (ε, I) are contained in the transformation set

$$Y = \{(\varepsilon, I) \,|\varepsilon > 0, 0 < I < 1 \text{ s.t. } G(\varepsilon, I) \leq 0\}. \tag{14}$$

The PPF is the boundary of the transformation function defined as

$$PPF(\varepsilon, I) = \{(\varepsilon, I) \,|\varepsilon > 0, 0 < I < 1 \text{ s.t. } G(\varepsilon, I) = 0\}. \tag{15}$$

Equation (15) specifies the maximum information accuracy that can be published for a given value of privacy loss.

Solving for I as a function of ε, the data publication problem using the MWEM query release mechanism produces the production possibilities frontier

$$I\left(\varepsilon;|\chi|,|\mathcal{Q}|,N\right)=\left[1-\frac{K(\beta,|\chi|,|\mathcal{Q}|,N)}{\varepsilon^{b}}\right]. \tag{16}$$

The marginal social cost of increasing data accuracy I in terms of foregone privacy protection ε–the marginal rate of transformation–is

$$MRT\left(\varepsilon,I\right)\equiv\frac{dI}{d\varepsilon}=-\frac{\partial G/\partial\varepsilon}{\partial G/\partial I}=\frac{bK(\beta,|\chi|,|\mathcal{Q}|,N)}{\varepsilon^{b+1}}, \tag{17}$$

where the marginal rate of transformation is positive because privacy loss is a public bad. Application of the implicit function theorem yields the sign change in the middle equality.[21]

Figure 1 illustrates the PPF for our application to the publication of statistics on the distribution of income, which we describe in detail in Section 5. We graph the PPF described by equation (16) with ε on the horizontal axis and I on the vertical axis. Because ε is a public bad, the PPF is similar to the efficient risk-return frontier used in financial economics as well as the offer curve used in hedonic wage theory. The PPF separates feasible (ε, I) pairs, which are on and below the PPF, from infeasible pairs, which are above the PPF. The PPF also exhibits diminishing marginal rate of transformation: it is increasingly costly, in terms of foregone privacy, to increase information accuracy.

[21] As the proof of Theorem 2 shows, the equation that defines the transformation set is continuously differentiable with respect to both ε and I. This fact is not obvious from the text of Hardt et al. (2012), which introduced MWEM. In their presentation the relevant accuracy bound is reported using big-O notation. They did so for convenience as the exact bound is messy, but straightforward to derive.

We treat the parameters $(\beta, |\chi|, |\mathcal{Q}|, N)$ that determine K as outside the choice problem facing the data custodian. Doing so is not without consequence, as these parameters affect the location of the PPF. We think of them as determining the size of the "information budget" at the custodian's disposal. Our model envisions a custodian in possession of a fixed database and a charge to publish a fixed set of queries (contingency tables). Given these constraints, the custodian must choose the levels of privacy and accuracy to deliver the published statistics. The PPF determines the set of feasible pairs given the information budget.

4.2 The Optimal Levels of Accuracy and Privacy

Given the data publication technology, the data custodian must choose a level of privacy protection and a guaranteed level of accuracy in the published statistics. In practice, the data custodian's choice may depend on a host of legal, economic, and political considerations. Our goal is to characterize the optimal level of data accuracy and privacy protection. When data accuracy and privacy protection are public goods, the solution is not obtained through market pricing. We therefore ask in this subsection what levels of accuracy and privacy a utilitarian social planner would choose to deliver.

4.2.1 Preferences

We assume a closed system in the sense that data are collected from all members of the population, and all members of the population may benefit from use of the published statistics. Every person also consumes a set of pure private goods, the prices of which are exogenous. Our formulation allows for arbitrary heterogeneity in preferences for privacy loss and for the accuracy of published statistics. In doing so, we allow for the empirically relevant possibility that one group of peo-

ple cares primarily about privacy, while getting little utility from consuming the data, while another set cares primarily about data accuracy.

The indirect utility function, v_i, for each individual is

$$v_i\left(y_i, \varepsilon, I, x, p\right) = \max_q u_i\left(q, \varepsilon, I, x\right) \text{ s.t. } q^T p \leq y_i \tag{18}$$

where q is the bundle of L private goods chosen by individual i at prices p. The direct utility function $u_i\left(q, I, \varepsilon, x\right)$, also depends upon the privacy-loss public bad, ε, the data-accuracy public good, I, and on the data collected from all other individuals, which we represent here by the histogram vector, x. In our applications, x will contain data describing the distribution of income or the distribution of body-mass index.[22]

4.2.2 The Social Planner's Problem

We adopt the utilitarian linear aggregation form of the social welfare function

$$SWF\left(\varepsilon, I, v, y, x, p\right) = \sum_{i=1}^{N} v_i\left(y_i, \varepsilon, I, x, p\right) \tag{19}$$

where v and y are vectors of N indirect utilities and incomes, respectively. The social planner's problem is

$$\max_{\varepsilon, I} SWF\left(\varepsilon, I, v, y, x, p\right) \tag{20}$$

subject to the set of production possibilities characterized by Equation (16).

[22]We have abstracted from the problem of multivariate characteristics in x. That is, one might consider other demographic variables like race or birth date to be among the sensitive characteristics that the privacy-preserving publication system needs to protect. One might also note that individuals could have heterogeneous preferences about the sensitivity of these characteristics.

Assuming the indirect utility functions are differentiable, the conditions that characterize the welfare-maximizing levels of ε and I subject to the feasibility constraint are

$$\frac{\frac{\partial G\left(\varepsilon^0, I^0\right)}{\partial \varepsilon}}{\frac{\partial G(\varepsilon^0, I^0)}{\partial I}} = \frac{\frac{\partial}{\partial \varepsilon} \sum_{i=1}^{N} v_i \left(y_i, \varepsilon^0, I^0, x, p\right)}{\frac{\partial}{\partial I} \sum_{i=1}^{N} v_i \left(y_i, \varepsilon^0, I^0, x, p\right)} \tag{21}$$

and $PPF\left(\varepsilon^0, I^0\right)$. The left-hand side of equation (21) is the marginal rate of transformation from the production possibilities frontier while the right-hand side is the marginal rate of substitution between privacy loss and data accuracy. See Appendix A.3 for the technical details.

5 Applications

We conduct two empirical exercises to illustrate the normative content of our model. Our goal is to show how these methods can provide guidance to data providers about the optimal rate at which to trade off privacy loss for statistical accuracy. We present results for two applications where privacy loss and data accuracy are both highly salient: (1) publication of income distribution statistics; (2) publication of relative health status statistics. We use data from the American Community Survey (ACS) to simulate publication of detailed statistics on the income distribution. We use data from the National Health Interview Survey (NHIS) to simulate publication of statistics on the distribution of body-mass index (BMI). In each case, we characterize the PPF by specifying parameters the data custodian will use with the MWEM algorithm, as described in Section 4.

To find the optimal levels of data accuracy and privacy loss, we use a specific utility function to implement the interdependent preferences studied in Section 4 (Pollak 1976; Akerlof 1997; Card et al. 2012). In our first application, individuals care about the quality of income statistics because they want to know their

relative standing in the income distribution. The model yields a closed-form solution for willingness to pay that depends on the joint distribution of preferences for data accuracy and privacy, along with income, and health status. We use data from opinion surveys to estimate the willingness of the social planner to pay for decreased privacy loss with reduced accuracy.

5.1 The Specification of Preferences

For clarity, we focus here on the publication of income statistics. Our application to health statistics uses an identical specification up to relabeling. We assume each individual cares about her position in the income distribution. We also assume heterogeneity in individual tastes for privacy loss and data accuracy. A specification of the indirect utility function that captures the required features is

$$\begin{aligned} v\left(y_i, \varepsilon, I, \tilde{y^i}, p\right) &= -\sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln y_i \\ &\quad -\gamma_i \left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right) \varepsilon \\ &\quad +\eta_i \left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right) I \end{aligned} \tag{22}$$

where $(\gamma_i, \eta_i) > 0$ for all $i = 1, \ldots, N$, $\xi_\ell > 0$ for all $\ell = 1, \ldots, L$ and $\sum_{\ell=1}^{L} \xi_\ell = 1$.[23] The term $\left(\ln y_i - \mathrm{E}\left[\ln y_i\right]\right)$ represents the deviation of individual i's log income from the population mean.[24]

[23] In equation (22) and what follows, expectation, variance, and covariance operators are with respect to the joint distribution of $\ln y_i$, γ_i and η_i in the population of N individuals.

[24] In Appendix A.3, we verify that the vector v of indirect utility functions is homogeneous of degree zero in (p, y), strictly increasing in y, non-increasing in p, quasiconvex in (p, y), and continuous in (p, y). Therefore, $v\left(y_i, I, \varepsilon, \tilde{y^i}, p\right)$ is a well-specified indirect utility function in this economy with relative income entering every utility function with the same functional form provided equation (22) is quasiconvex in (ε, I), which is trivially true for equation (22), as long as $(\gamma_i, \eta_i) > 0$ for all i, since it is linear in (ε, I). Hence, equation (19) is a well-specified social welfare function, quasiconvex in (ε, I), and the social planner's problem is well-specified since equation (16) is

Equation (22) is motivated by Akerlof (1997), and subsequent work on public-good provision with interdependent preferences as described in Aronsson and Johansson-Stenman (2008) and the references therein. If we assume $I = 1$ and $\varepsilon = 0$, then our indirect utility function is consistent with the prior literature, which assumes that the income distribution is known by everyone with perfect accuracy and without disutility from privacy loss.

Substitution of equation (22) into equation (21) yields

$$\frac{\frac{\partial G\left(\varepsilon^0, I^0\right)}{\partial \varepsilon}}{\frac{\partial G(\varepsilon^0, I^0)}{\partial I}} = \frac{\frac{\partial}{\partial \varepsilon}\sum_{i=1}^{N} v_i\left(y_i, \varepsilon^0, I^0, \tilde{y^i}, p\right)}{\frac{\partial}{\partial I}\sum_{i=1}^{N} v_i\left(y_i, \varepsilon^0, I^0, \tilde{y^i}, p\right)} \tag{23}$$

$$\begin{aligned} \frac{bK(|\chi|, |\mathcal{Q}|, N)}{(\varepsilon^0)^{b+1}} &= \frac{\sum_{i=1}^{N} \gamma_i \left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right)}{\sum_{i=1}^{N} \eta_i \left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right)} \\ &= \frac{\mathrm{E}\left[\gamma_i\right] + \mathrm{Cov}\left[\gamma_i, \ln y_i\right]}{\mathrm{E}\left[\eta_i\right] + \mathrm{Cov}\left[\eta_i, \ln y_i\right]} \end{aligned} \tag{24}$$

Note that a sign change occurs on both sides of equation (24) because we are modeling one public good, I, and one public bad, ε. The full solution is

$$I^0\left(.\right) = 1 - \left\{\frac{1}{b} K(|\chi|, |\mathcal{Q}|, N)^{1/b} \frac{\mathrm{E}\left[\gamma_i\right] + \mathrm{Cov}\left[\gamma_i, \ln y_i\right]}{\mathrm{E}\left[\eta_i\right] + \mathrm{Cov}\left[\eta_i, \ln y_i\right]}\right\}^{b/(b+1)} \tag{25}$$

and

$$\varepsilon^0\left(.\right) = \left\{bK(|\chi|, |\mathcal{Q}|, N) \frac{\mathrm{E}\left[\eta_i\right] + \mathrm{Cov}\left[\eta_i, \ln y_i\right]}{\mathrm{E}\left[\gamma_i\right] + \mathrm{Cov}\left[\gamma_i, \ln y_i\right]}\right\}^{1/(b+1)}. \tag{26}$$

5.2 Application 1: Publication of Income Statistics

To illustrate our method as it applies to the publication of income statistics, we first describe production possibilities applied to income data from the ACS, derive the marginal rate of transformation, and then estimate willingness to pay from the

quasiconcave in (ε, I).

FSS POS. We then solve the social planner's problem to derive the optimal level of privacy loss and data accuracy.

5.2.1 Publication Technology

The data custodian is in possession of a database with the exact income for all eligible members of the U.S. population. To illustrate the feasibility of our approach, we construct a population-scale database of incomes from the 5-year ACS files for 2010–2014. Specifically, we generate a database with $N = 197,040,596$ records, which is the size of the 2012 adult population ages 18 to 64, inclusive, estimated in the ACS. To generate the database, we use the Bayes bootstrap to draw N records from the 2010–2014 ACS files using expected probability proportional to the sampling weights. The details of data preparation and analysis appear in Appendix A.5 and the associated code archive.

To simulate publication of the income distribution, we group income into 797 evenly-spaced bins, which is the size of the data domain, $|\chi|$. The bin sizes and labels are non-private. The set of queries to be answered consists of all interval queries; that is, all queries of the form "how many records fall between bin a and bin b, inclusive?". There are $|\mathcal{Q}| = 318,003$ such queries.[25] The custodian operates MWEM to publish statistics from this database.

5.2.2 Measuring Preferences

To estimate the marginal rate of substitution in the social planner's problem, we use the FSS POS. Our goal is to empirically quantify the distribution of the indirect utility function parameters.[26]

[25] Calculated as the number of ways to choose one or two bins from the full set of 797 bins.

[26] For more details of the the FSS POS see Childs et al. (2012) and Childs et al. (2015). See also Appendix A.5.

The FSS POS is a national public opinion survey conducted in conjunction with Gallup Daily Tracking Poll. From it, we use the following questions:

- `FS11`, which records responses on a five-category Likert scale measuring agreement with the following statement: "People can trust federal statistical agencies to keep information about them confidential."
- `FS14`, which records binary responses to the following question: "Would you say that federal statistical agencies often invade people's privacy, or generally respect people's privacy?"
- `FS7`, which records responses on a five-category Likert scale measuring the extent of agreement with the following statement: "Policy makers need federal statistics to make good decisions about things like federal funding."
- Family income, recorded in five categories.

We use `FS11` and `FS14` as proxy measures of the latent preference for privacy γ_i and `FS7` as a proxy measure of the latent preference for accuracy η_i. We compute the polychoric correlations between each preference measure and income:[27]

- Based on `FS7`, we find $\text{Corr}\left[\gamma_i, \ln y_i\right] = 0.082\,(\pm 0.003)$
- Based on `FS14`, we find $\text{Corr}\left[\gamma_i, \ln y_i\right] = 0.083\,(\pm 0.003)$
- Based on `FS11`, we find $\text{Corr}\left[\eta_i, \ln y_i\right] = 0.040\,(\pm 0.003)$

[27]For many respondents, income is missing, and the data exhibit moderate levels of nonresponse on the opinion variables. The preceding estimates are based on a complete data analysis in which the missing values are multiply imputed 500 times conditional on the observed data, and we account for the imputation uncertainty by combining the within and between implicate variance. The results are qualitatively similar if we drop missing cases.

To compute the MRS based on Equation (23) using the correlations reported above, we need additional modeling assumptions. Specifically, we assume log income and the latent preference parameters, η and γ are normally distributed, and that η and γ have unit variances. The data are informative about $\text{Corr}\left[\gamma_i, \ln y_i\right]$ and $\text{Corr}\left[\eta_i, \ln y_i\right]$. To pin down the location, we assume $\text{E}\left[\gamma_i\right] = \text{E}\left[\eta_i\right] = \sigma_{\ln y}$. This assumption puts variation in utility that arises from the direct valuation of privacy loss and data accuracy on the same scale as variation in utility that arises from the interaction with relative income.[28]

Invoking these assumptions, we have

$$\frac{\text{E}\left[\gamma_i\right] + \text{Cov}\left[\gamma_i, \ln y_i\right]}{\text{E}\left[\eta_i\right] + \text{Cov}\left[\eta_i, \ln y_i\right]} = \frac{1 + \text{Corr}\left[\gamma_i, \ln y_i\right]}{1 + \text{Corr}\left[\eta_i, \ln y_i\right]} \tag{27}$$

Substituting the polychoric correlations obtained from the FSS POS data, we estimate the $MRS = 1.040$. At the social optimum, a one-unit increment in privacy loss must be compensated by a 1.040 unit increase in data accuracy.

5.2.3 Solution

Figure 1 illustrates the solution to the social planner's problem when the statistical agency operates the MWEM algorithm. The social welfare function is based on the indirect utility function in equation (22). The solid line represents the production possibilities frontier under MWEM given the parameterization based on ACS data. The dashed lines are contour plots of the social welfare function (19) at representative non-optimal (SWF_0) and optimal (SWF_1) attainable levels of social

[28] We recognize that our assumptions on $\text{E}\left[\gamma_i\right]$ and $\text{E}\left[\eta_i\right]$ are somewhat arbitrary. We could, for example, also assume that individuals only care about ε and I through the relative income channel, in which case the terms involving $\text{E}\left[\gamma_i\right]$ and $\text{E}\left[\eta_i\right]$ would drop from the utility function. The implied MRS would be considerably different with those modeling assumptions. These considerations highlight the need for much better models and data on the demand for privacy and statistical accuracy.

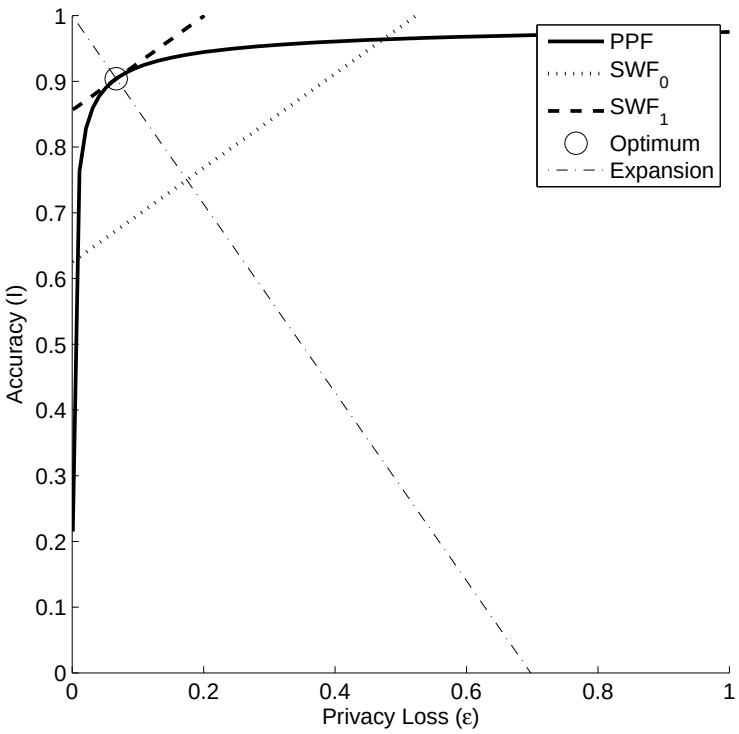

Figure 1: Solution to the Social Planner's Problem

welfare. The expansion path is the straight line that intersects the horizontal axis.

Evaluated at the point where the $MRT = 1.040$, we find the optimal accuracy and privacy are $I^0 = 0.862$ and $\varepsilon^0 = 0.042$. We can also evaluate the welfare cost of choosing suboptimally low privacy loss at the expense of data accuracy. This is the relevant scenario in the case of private provision since, as we showed in Section 4.2, the costs of privacy loss are internalized but the benefits of data accuracy are not. Choosing a point with $\varepsilon = 0.021$, which is equivalent to a 50 percent decrease in privacy loss, the corresponding value of I on the PPF, $I = 0.826$, results in an expected change in utility of -0.013 per person. This is equivalent to a loss of 1.3 percent of national income.

5.2.4 Simulations

The theoretical accuracy guarantee says that the worst-case query is answered to within 0.174 of it's true value. This bound is informative, but allows a considerable amount of noise. If the distribution of incomes were uniform, each entry would be on the order 0.001 in the normalized histogram. Our analysis is based on the worst case guarantee, which is the reliability of the method across all possible datasets and realizations of the randomized mechanism. In practice, the MWEM algorithm can outperform this worst-case bound, as shown by Hardt et al. (2012) and subsequently by Schmutte (2016).

Using our population data from the ACS, we run the MWEM algorithm 30 times using the optimal parameter configuration. The maximum error across all queries, averaged across the 30 implementations, is 0.0014, which is on the same order as the uniform histogram, and considerably lower than the worst-case guarantee. These results indicate that, beyond offering a framework for reasoning about optimal privacy protection, MWEM may be a practical method for publishing data; at least in this relatively simple context. Finally, note that when we cut ε by half to $\varepsilon = 0.021$, as in the policy counterfactual considered above, the average worst-case empirical error doubles to 0.002.

5.3 Application 2: Publication of Health Status Statistics

Our analysis of the publication of health statistics parallels the preceding analysis of income statistics. We use the same model for interdependent preferences, assuming that individuals care about their relative health status rather than relative income. These specifications yield an expression for the social willingness to pay for reduced privacy loss that depends on the correlation of health status with preferences for privacy and accuracy. We estimate these quantities using

data from the Cornell National Social Survey (CNSS) and use them to compute the socially optimal levels of privacy loss and data accuracy.

5.3.1 Publication Technology

We assume the data custodian is in possession of a database with the body-mass index (BMI) for all members of the U.S. population. To illustrate the feasibility of the mechanism, we construct a population-scale database based on BMI measured from the 2015 National Health Interview Survey (NHIS). Specifically, we generate a database with the distribution of BMI as collected in the NHIS of size $N = 242,977,154$. This is the size of the population, as reported from the ACS public-use tables, for all individuals age 18 and older not residing in group quarters, which is the universe for which BMI is collected in the NHIS. To generate this synthetic population database, we draw N BMI observations from the 2015 NHIS using the Bayes bootstrap with expected probability proportional to their sampling weights. The details of data preparation and analysis appear in Appendix A.5 and the associated code archive.

To simulate publication of the income distribution, we group BMI into $|\chi| = 800$ evenly-spaced bins. The bin sizes and labels are non-private. The set of queries to be answered consists of all interval queries; that is, all queries of the form "how many records fall between bin a and bin b, inclusive?". There are $|\mathcal{Q}| = 320,400$ such queries. The custodian operates MWEM to publish statistics from this database.

5.3.2 Measuring Preferences

Our model for preferences is identical to Equation 22 except we substitute the latent health status, $\ln h_i$, for income $\ln y_i$ in the terms involving ε and I. Making

the same distributional assumptions, it follows that we can estimate willingness to pay by

$$WTP = \frac{1 + \text{Cov}\left[\gamma_i, \ln h_i\right]}{1 + \text{Cov}\left[\eta_i, \ln h_i\right]}. \tag{28}$$

We measure the joint distribution of preferences for privacy, accuracy, and health status, using data from the Cornell National Social Survey (CNSS) from 2011, 2012, and 2013. The CNSS is a nationally representative cross-sectional telephone survey of 1,000 adults each year. The survey collects basic household and individual information, including income. In 2011, 2012, and 2013, the CNSS includes questions that elicit subjective health status along with attitudes toward the privacy of personal health information and the value of accurate health statistics.[29] We use the following questions from the CNSS:

- `JAq6`, "In general, how would you rate your overall health?" measured on a five-category scale;
- "If medical information could be shared electronically between the places where a patient receives medical care, how do you think that would:"
    1. `JAq4@b`, "…affect the privacy and security of medical information?" measured on a on a five-category scale (proxy for privacy preferences, γ).
    2. `JAq4@a`, "…affect the quality of medical care?" measured as a on a five-category scale (proxy for accuracy preferences, η).

Once again, we compute the polychoric correlations between the ordinal measures:

[29] For the CNSS (Cornell Institute for Social and Economic Research and Survey Research Institute n.d.), see https://www.sri.cornell.edu/sri/cnss.cfm.

- $\text{Corr}\left[\gamma_i, \ln h_i\right] = 0.015\ (\pm 0.021)$
- $\text{Corr}\left[\eta_i, \ln h_i\right] = 0.076\ (\pm 0.022)$

Concern about the privacy of health status is negligibly correlated with health status. Concern for the quality of medical information, is more positively correlated with health status. Making the relevant substitutions implies that at the social optimum

$$MRT\left(\varepsilon^0, I^0\right) = 0.94,$$

which implies that a one-unit increase in privacy loss must be compensated with a 0.94 increase in data accuracy. The estimated shadow price of reduced privacy loss is, therefore, lower in the context of health data than in the context of publishing income statistics.

5.3.3 Solution

Evaluated at the point where the $MRT = 0.94$, we find that the optimal accuracy and privacy loss are $I^0 = 0.872$ and $\varepsilon^0 = 0.0451$, respectively. Once again, we evaluate the welfare cost of choosing suboptimally low privacy loss at the expense of data accuracy. Choosing a point with $\varepsilon = 0.0226$, which is equivalent to a 50 percent decrease in privacy loss, the corresponding value of I on the PPF, $I = 0.839$, results in an expected change in utility of -0.013 per person.

5.3.4 Simulations

Using our synthetic population data based on the NHIS, we run the MWEM algorithm 30 times using the optimal parameter configuration. The maximum error across all queries, averaged across the 30 implementations, is 0.0015. When we cut ε by half to $\varepsilon = 0.226$, as in the policy experiment above, the average worst-case

empirical error rises to 0.002. As was the case with income statistics, these simulations show that the optimal choice for MWEM may yield a practical publication strategy in the context of publishing an indicator of health status.

5.4 Discussion

Our analysis suggests how data providers can combine information about their publication technology with data on the value of privacy and data accuracy to guide decision-making. We note, however, that the preference data from the surveys are not ideally suited to our applications. Obtaining our results using the available data requires a number of ancillary assumptions. We make careful note of these assumptions, and why they are needed. Progress on the questions identified by this paper will require better information on individual and social preferences for privacy and for data accuracy. We defer further speculation on these measurement issues to the conclusion.

One might also suppose that a straightforward combination of the indirect utility functions that generated demand for income distribution and health statistics should lead to a model in which the statistical agency provides both types of data to the population. Indeed, it is a rare government whose statistical agencies publish only one characteristic of the population. We do not develop that model here.

If the statistical agency wished to publish both the income distribution data with accuracy $I_y^0 = 0.862$ and the health status statistics with accuracy $I_h^0 = 0.872$, which are the two optimal values derived above, then the level of privacy protection would be the sum of $\varepsilon_y^0 = 0.042$ (income distribution) and $\varepsilon_h^0 = 0.045$ (health statistics). We have introduced the subscripts y and h to distinguish the solutions to the two problems. By the composability of ε-differential privacy, the actual privacy protection afforded by this publication strategy is $\varepsilon_{yh} = 0.087$. There

is no proof in our work (or anywhere else that we know) that the combination $I_y^0 = 0.862$ and $I_h^0 = 0.872$ with $\varepsilon_{yh} = 0.087$ is optimal in any sense. All of the proposed publications must be considered simultaneously in order to get the correct optimum. This is feasible for the technology we have adopted, which can handle the economies of scope implied by the composability of differential privacy, but we have not done these calculations.

6 Conclusion

This paper provides the first comprehensive synthesis of the economics of privacy with the statistical disclosure limitation and privacy-preserving data publication literatures. We develop a complete model of the technology associated with data publication constrained by privacy protection. Both the accuracy of the published data and the level of formal privacy protection are public goods. We solve the full social planning problem with interdependent preferences, which are necessary in order to generate demand for the output of government statistical agencies. The PPF is directly derived from very recent technology for ε-differential privacy with (α, β)-accuracy.

We compute the welfare loss associated with suboptimally providing too much privacy protection and too little accuracy. Income distribution statistics are provided when individuals care about their income relative to the population distribution. Reducing privacy loss in the published data by half relative to the social optimum and commensurately reducing data accuracy results in a utility loss of 0.013 log points – approximately equivalent to a 1.3% decrease in national income. A comparable welfare loss arises in our application to publication of health statistics, when the privacy loss is reduced by half.

A major barrier to research in this area is the lack of data on preferences for privacy and data accuracy. Self-reported attitudes toward privacy are increasingly collected in opinion surveys, but more information is needed on the price people attach to privacy loss; particularly as regards the sort of inferential disclosures considered in this paper. Data on the individual and social benefits of population statistics is even more scarce. New research is required, including carefully designed controlled experiments that identify the components of utility, such as relative income, that can only be assessed with statistical data on the relevant comparison population. Such experiments have already informed the role of relative income in the study of subjective well-being (Luttmer 2005; Clark et al. 2008) and the acquisition of private data for commercial use (Acquisti et al. 2013).

The concept of differential privacy allows a natural interpretation of privacy protection as a commodity over which individuals might have preferences. In many important contexts, privacy protection and data accuracy are not purely private commodities. When both are public goods, the market allocations might not be optimal. We show that it is feasible, at least in principle, to determine the optimal trade-off between privacy protection and data accuracy when the public-good aspects are important. We also use another feature of differential privacy, composability, to show that even though relatively accurate statistics can be released for a single population characteristic such as income distribution or relative health status, each statistic requires its own budget. If an agency is releasing data on many detailed characteristics of the population, a small total privacy-loss budget may not allow many of the statistics to be released with accuracy comparable to the accuracy shown in our applications. This is an important warning for the Information Age.

References

Abowd, J. M. and Schmutte, I. M. (2015). Economic analysis and statistical disclosure limitation, *Brookings Papers on Economic Activity* pp. 221–267. Spring.

Abowd, J. M. and Schmutte, I. M. (2017). Replication archive for: Revisiting the economics of privacy: Population statistics and confidentiality protection as public goods. Supported by the U.S. Census Bureau, the Sloan Foundation, and NSF Grants BCS-0941226, TC-1012593 and SES-1131848.

Abowd, J. M., Schmutte, I. M. and Vilhuber, L. (eds) (2017). *Proceedings from the 2016 NSF-Sloan Workshop on Practical Privacy*, Labor Dynamics Institute, Cornell University.

Acquisti, A., John, L. K. and Loewenstein, G. (2013). What Is Privacy Worth?, *Journal of Legal Studies* **42**(2): 249–274.

Acquisti, A., Taylor, C. and Wagman, L. (2016). The Economics of Privacy, *Journal of Economic Literature* **54**(2): 442–492.

Acquisti, A. and Varian, H. R. (2005). Conditioning prices on purchase history, *Marketing Science* **24**(3): 367–381.

Akerlof, G. A. (1997). Social distance and social decisions, *Econometrica* **65**(5): 1005–1027.

Apple previews iOS 10, the biggest iOS release ever (2016). Cited on March 5, 2017.
URL: *http://www.apple.com/newsroom/2016/06/apple-previews-ios-10-biggest-ios-release-ever.html*

Aronsson, T. and Johansson-Stenman, O. (2008). When the Joneses' Consumption Hurts: Optimal Public Good Provision and Nonlinear Income Taxation, *Journal of Public Economics* **92**(5-6): 986–997.

Buuren, S. V., Brand, J. P., Groothuis-Oudshoorn, C. G. and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation, *Journal of Statistical Computation and Simulation* **76**(12): 1049–1064.

Card, D., Mas, A., Moretti, E. and Saez, E. (2012). Inequality at work: the effect of peer salaries on job satisfaction, *American Economic Review* **102**(6): 2981–3003.

Childs, J. H., Willson, S., Martinez, S. W., Rasmussen, L. and Wroblewski, M. (2012). Development of the Federal Statistical System Public Opinion Survey, *JSM Proceedings Survey Research Methods Section*.

Childs, J., King, R. and Fobia, A. (2015). Confidence in U.S. federal statistical agencies, *Survey Practice* **8**(5).

Clark, A. E., Frijters, P. and Shields, M. A. (2008). Relative income, happiness, and utility: An explanation for the Easterlin paradox and other puzzles, *Journal of Economic Literature* **46**(1): 95–144.

Cornell Institute for Social and Economic Research and Survey Research Institute (n.d.). Cornell national social survey (cnss) integrated (beta version), Online.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control, *Statistik Tidskrift* **15**: 429–444.

Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy, *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on*

Principles of Database Systems, PODS '03, ACM, New York, NY, USA, pp. 202–210.

Duncan, G. and Lambert, D. (1986). Disclosure-limited data dissemination, *Journal of the American Statistical Association* **81**(393): 10–18.

Dwork, C. (2008). Differential privacy: a survey of results, *Theory and Applications of Models of Computation* pp. 1–19.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis, *Proceedings of the Third conference on Theory of Cryptography*, TCC'06, Springer-Verlag, Berlin, Heidelberg, pp. 265–284.

Dwork, C., McSherry, F. and Talwar, K. (2007). The price of privacy and the limits of LP decoding, *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing STOC '07*, ACM Digital Library, pp. 85–94.

Dwork, C. and Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy, *Journal of Privacy and Confidentiality* **2**(1): 93–107.

Dwork, C. and Roth, A. (2014). *The Algorithmic Foundations of Differential Privacy*, now publishers, Inc. Also published as "Foundations and Trends in Theoretical Computer Science" Vol. 9, Nos. 3–4 (2014) 211-407.

Dwork, C., Smith, A., Steinke, T., Ullman, J. and Vadhan, S. (2015). Robust traceability from trace amounts, *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS '15)*, ACM Digital Library, pp. 650–669.

Eisen, M. and Kaufman, G. M. (1985). *Natural Gas Data Needs in a Changing Regulatory Environment*, Committee on National Statistics, National Academies Press.

Erlingsson, Ú., Pihur, V. and Korolova, A. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response, *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14* pp. 1054–1067.

Evfimievski, A., Gehrke, J. and Srikant, R. (2003). Limiting privacy breaches in privacy preserving data mining, *SIGMOD Principles of Database Systems PODS '03*, ACM Digital Library, pp. 211–222.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*, Chapman & Hall/CRC Texts in Statistical Science, third edn, Taylor & Francis.

Ghosh, A. and Roth, A. (2011). Selling privacy at auction, *Proceedings of the 12th ACM conference on Electronic commerce*, EC '11, ACM, New York, NY, USA, pp. 199–208.

Goldwasser, S. and Micali, S. (1982). Probabilistic encryption & how to play mental poker keeping secret all partial information, *STOC '82 Proceedings of the fourteenth annual ACM symposium on Theory of computing* pp. 365–377.

Goldwasser, S. and Micali, S. (1984). Probabilistic encryption, *Journal of Computer and System Sciences* **28**(2): 270–299.

Hardt, M., Ligett, K. and McSherry, F. (2012). A Simple and Practical Algorithm for Differentially Private Data Release., *Nips* pp. 1–9.

Hardt, M. and Rothblum, G. N. (2010). A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis, *2010 IEEE 51st Annual Symposium on Foundations of Computer Science* pp. 61–70.

Heffetz, O. and Ligett, K. (2014). Privacy and data-based research, *Journal of Economic Perspectives* **28**(2): 75–98. Spring.

Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B. C. and Roth, A. (2014). Differential Privacy: An Economic Method for Choosing Epsilon, *2014 IEEE 27th Computer Security Foundations Symposium*, pp. 398–410.

Kasiviswanathan, S. P., Rudelson, M. and Smith, A. (2013). The power of linear reconstruction attacks, *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms SODA '13*, ACM Digital Library, pp. 1415–1433.
URL: *https://arxiv.org/abs/1210.2381v1*

Kifer, D. and Machanavajjhala, A. (2011). No free lunch in data privacy, *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, ACM Digital Library, New York, NY, USA, pp. 193–204.

Luttmer, E. F. P. (2005). Neighbors as negatives: relative earnings and well-being, *The Quarterly Journal of Economics* **120**(3): 963–1002.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J. and Vilhuber, L. (2008). Privacy: Theory meets practice on the map, *Proceedings - International Conference on Data Engineering* pp. 277–286.

Mas-Colell, A., Whinston, M. and Green, J. (1995). *Microeconomic theory*, Oxford student edition, Oxford University Press.

Minnesota Population Center and State Health Access Data Assistance Center (2016). Integrated Health Interview Series: Version 6.21, *[computer file]*, University of Minnesota.
URL: *http://ihis.us/*

Muthukrishnan, S. and Nikolov, A. (2012). Optimal private halfspace counting via discrepancy, *Proceedings of the forty-fourth annual ACM Symposium on Theory of Computing STOC '12*, ACM Digital Library, pp. 1285–1292.

National Academies of Sciences, Engineering, and Medicine (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*, Committee on National Statistics, National Academies Press, Washington, DC.

Ohm, P. (2010). Broken promises of privacy: responding to the surprising failure of anonymization, *UCLA Law Review* **57**: 1701.

Pollak, R. A. (1976). Interdependent preferences, *The American Economic Review* **66**(3): 309–320.

Postlewaite, A. (1998). The Social Basis of Interdependent Preferences, *European Economic Review* **42**(3-5): 779–800.

Prewitt, K. (2011). Why It Matters to Distinguish Between Privacy & Confidentiality, *Journal of Privacy and Confidentiality* **3**(2): 41–47.

Rubin, D. B. (1981). The Bayesian Bootstrap, *The Annals of Statistics* **9**(1): 130–134.

Samuelson, P. A. (1954). The pure theory of public expenditure, *Review of Economics and Statistics* **37**: 387–389.

Schmutte, I. M. (2016). Differentially Private Release of Data on Wage and Job Mobility, *Statistical Journal of the IAOS* **32**(1): 81–92.

Stigler, G. J. (1980). An introduction to privacy in economics and politics, *Journal of Legal Studies* **9**(4): 623–644.

U.S. Census Bureau (2016). 2010-2014 ACS 5-year public use microdata samples (PUMS), *[computer file]*, U.S. Census Bureau's American Community Survey Office.
URL: *http://www2.census.gov/programs-surveys/acs/data/pums/2014/5-Year/*

U.S. Census Bureau, Population Division (2013). Annual Estimates of the Resident Population for Selected Age Groups by Sex for the United States, States, Counties, and Puerto Rico Commonwealth and Municipios: April 1, 2010 to July 1, 2012, *Technical report*, U.S. Census Bureau, Population Division.

U.S. Census Bureau, Population Division (2015). 2015 American Community Survey 1-Year Estimates, *Technical report*, U.S. Census Bureau, Population Division.

Wasserman, L. and Zhou, S. (2010). A Statistical Framework for Differential Privacy, *Journal of the American Statistical Association* **105**(489): 375–389. ISSN: 0162-1459, 1537-274X.

APPENDIX

A.1 Proofs Omitted from the Text

Proof of Theorem 1 Proof. Given a target error bound α, corresponding to data accuracy level $I = (1-\alpha)$, the private producer must procure data-use rights from the respondents in the confidential data with $\varepsilon(I)$ units of privacy protection from a measure of $H(I)$ individuals. Define

$$p_{\varepsilon}^{VCG} = Q\left(\frac{H(I)}{N}\right).$$

Note that p_{ε}^{VCG} is the disutility of privacy loss for the marginal participant in the VCG mechanism. The total cost of producing $I = (1-\alpha)$ using the VCG mechanism is equation (7):

$$C^{VCG}(I) = Q\left(\frac{H(I)}{N}\right) H(I)\varepsilon(I).$$

Differentiating with respect to I

$$\frac{dC^{VCG}(I)}{dI} = Q\left(\frac{H(I)}{N}\right) H(I)\varepsilon'(I) + \left[Q\left(\frac{H(I)}{N}\right) + Q'\left(\frac{H(I)}{N}\right)\frac{H(I)}{N}\right] H'(I)\varepsilon(I).$$

$0 < \frac{dC^{VCG}(I)}{dI}$ for all I, since $\varepsilon'(I) > 0$, $H'(I) > 0$, and $Q'() > 0$. The result in the theorem follows by using the equilibrium price for the private market sale of I, which is p_I in equation (3),

$$p_I = \bar{\eta} = \frac{dC^{VCG}(I^{VCG})}{dI}.$$

Hence, $I^{VCG} \leq I^0$, since $\sum_{i=1}^{N} \eta_i \geq \bar{\eta}$, and the conditions on Q that imply $\frac{d^2 C^{VCG}(I)}{dI^2} \geq 0$. ■

Proof of Theorem 2 Proof. Hardt, Ligett and McSherry (2012), henceforth HLM, prove that MWEM satisfies ε-differential privacy with (α, β)-accuracy using definitions based on unnormalized histograms and queries. They give an exact accuracy bound for MWEM in their Theorem 2.2. Their reported bound for unnormalized queries is $2N\sqrt{\frac{\log|\chi|}{T}} + \frac{10T\log|\mathcal{Q}|}{\varepsilon}$, where T is the number of iterations of MWEM. We rescale the error bound by the database size N to account for our normalization. Converting to normalized queries yields accuracy bound $\alpha = 2\sqrt{\frac{\log|\chi|}{T}} + \frac{10T\log|\mathcal{Q}|}{N\varepsilon}$. HLM note that the optimal number of iterations is the value of T that minimizes the bound. The optimal value is easily found to be $T = \left(\frac{\varepsilon N\sqrt{\log|\chi|}}{10\log|\mathcal{Q}|}\right)^{2/3}$. Our result follows by substituting T into the expression for the exact α bound. Note that substituting $I = 1 - \alpha$ yields an exact, differentiable expression for I with respect to ε as described in equation 16.

■

A.2 Translation of the Ghosh-Roth Model in Section 3 to Our Notation

In this appendix we show that the results in our Section 3, based on the definitions in the text using database histograms and normalized queries, are equivalent to the results in Ghosh and Roth (2011). In what follows, definitions and theorems tagged GR refer to the original Ghosh and Roth (GR, hereafter) paper. Untagged definitions and theorems refer to our results in the text.

GR model a database $D \in \{0,1\}^n$ where there is a single bit, b_i, taking values in $\{0,1\}$ for a population of individuals $i = 1, \ldots, n$. In GR-Definition 2.1, they define

a query release mechanism $A(D)$, a randomized algorithm that maps $\{0,1\}^n \to \mathbb{R}$, as ε_i-differentially private if for all measurable subsets S of $\mathbb{R}$ and for any pair of databases D and $D^{(i)}$ such that $H\left(D, D^{(i)}\right) = 1$

$$\frac{\Pr\left[A(D) \in S\right]}{\Pr\left[A\left(D^{(i)}\right) \in S\right]} \leq e^{\varepsilon_i}$$

where $H\left(D, D^{(i)}\right)$ is the Hamming distance between D and $D^{(i)}$.

Notice that this is not the standard definition of ε-differential privacy, which they take from Dwork et al. (2006), because a "worst-case" extremum is not included. The parameter ε_i is specific to individual i. The amount of privacy loss algorithm A permits for individual i, whose bit b_i is the one that is toggled in $D^{(i)}$, is potentially different from the privacy loss allowed for individual $j \neq i$, whose privacy loss may be $\varepsilon_j > \varepsilon_i$ from the same algorithm. In this case, individual j could also achieve ε_j-differentially privacy if the parameter ε_i were substituted for ε_j. To refine this definition so that it also corresponds to an extremum with respect to each individual, GR-Definition 2.1 adds the condition that algorithm A is ε_i-*minimally differentially private* with respect to individual i if

$$\varepsilon_i = \arg\inf_{\varepsilon} \left\{ \frac{\Pr\left[A(D) \in S\right]}{\Pr\left[A\left(D^{(i)}\right) \in S\right]} \leq e^{\varepsilon} \right\},$$

which means that for individual i, the level of differential privacy afforded by the algorithm $A(D)$ is the smallest value of ε for which algorithm A achieves ε-differential privacy for individual i. In GR ε_i-differentially private always means ε_i-minimally differentially private.

GR-Fact 1, stated without proof, but see Dwork and Roth (2014, p. 42-43) for a proof, says that ε_i-minimal differential privacy composes. That is, if algorithm $A(D)$ is ε_i-minimally differentially private, $T \subset \{1, \ldots, n\}$, and $D, D^{(T)} \in \{0,1\}^n$

with $H\left(D, D^{(T)}\right) = |T|$, then

$$\frac{\Pr\left[A\left(D\right) \in S\right]}{\Pr\left[A\left(D^{(T)}\right) \in S\right]} \leq e^{\left\{\sum_{i \in T} \varepsilon_i\right\}},$$

where $D^{(T)}$ differs from D only on the indices in T.

In the population, the statistic of interest is an unnormalized query

$$s = \sum_{i=1}^{n} b_i.$$

The ε_i-minimally differentially private algorithm $A\left(D\right)$ delivers an output $\hat{s}$ that is a noisy estimate of s, where the noise is induced by randomness in the query release mechanism embedded in A. Each individual in the population when offered a payment $p_i > 0$ in exchange for the privacy loss $\varepsilon_i > 0$ computes an individual privacy cost equal to $\upsilon_i \varepsilon_i$, where $\upsilon_i > 0$, $p \equiv (p_1, \ldots, p_n) \in \mathbb{R}_+^n$, and $\upsilon \equiv (\upsilon_1, \ldots, \upsilon_n) \in \mathbb{R}_+^n$.

GR define a mechanism M as a function that maps $\mathbb{R}_+^n \times \{0,1\}^n \to \mathbb{R} \times \mathbb{R}_+^n$ using an algorithm $A\left(D\right)$ that is $\varepsilon_i\left(\upsilon\right)$-minimally differentially private to deliver a query response $\hat{s} \in \mathbb{R}$ and a vector of payments $p\left(\upsilon\right) \in \mathbb{R}_+^n$. GR-Definition 2.4 defines individually rational mechanisms. GR-Definition 2.5 defines dominant-strategy truthful mechanisms. An individually rational, dominant-strategy truthful mechanism M provides individual i with utility $p_i\left(\upsilon\right) - \upsilon_i \varepsilon_i\left(\upsilon\right) \geq 0$ and $p_i\left(\upsilon\right) - \upsilon_i \varepsilon_i\left(\upsilon\right) \geq p_i\left(\upsilon^{\tilde{}i}, \upsilon_i'\right) - \upsilon_i \varepsilon_i\left(\upsilon^{\tilde{}i}, \upsilon_i'\right)$ for all $\upsilon_i' \in \mathbb{R}_+^n$, where $\upsilon^{\tilde{}i}$ is the vector υ with element υ_i removed.

GR define $\left(k, \frac{1}{3}\right)$-accuracy in GR-Definition 2.6 using the deviation $|\hat{s} - s|$ from the output $\hat{s}$ produced by algorithm $A\left(D\right)$ using mechanism M as

$$\Pr\left[|\hat{s} - s| \leq k\right] \geq \left(1 - \frac{1}{3}\right)$$

where we have reversed the direction of the inequalities and taken the complementary probability to show that this is the unnormalized version of our Definition 3 for a query sequence of length 1. GR also define the normalized query accuracy level as α, which is identical to our usage in Definition 3.

GR-Theorem 3.1 uses the GR definitions of ε_i-minimal differential privacy, $\left(k, \frac{1}{3}\right)$-accuracy, and GR-Fact 1 composition to establish that any differentially private mechanism M that is $\left(\frac{\alpha n}{4}, \frac{1}{3}\right)$-accurate must purchase privacy loss of at least $\varepsilon_i \geq \frac{1}{\alpha n}$ from at least $H \geq (1-\alpha)\, n$ individuals in the population. GR-Theorem 3.3 establishes the existence of a differentially private mechanism that is $\left(\left(\frac{1}{2} + \ln 3\right) \alpha n, \frac{1}{3}\right)$-accurate and selects a set of individuals $H \subset \{1, \ldots, n\}$ with $\varepsilon_i = \frac{1}{\alpha n}$ for all $i \in H$ and $|H| = (1-\alpha)\, n$.

In order to understand the implications of GR-Theorems 3.1 and 3.3 and our arguments about the public-good properties of differential privacy, consider the application of GR-Definition 2.3 ($\mathrm{Lap}\,(\sigma)$ noise addition) to construct an ε-differentially private response to the counting query based on GR-Theorem 3.3 with $|H| = (1-\alpha)\, n$ and the indices ordered such that $H = \{1, \ldots, |H|\}$. The resulting answer from the query response mechanism is

$$\hat{s} = \sum_{i=1}^{H} b_i + \frac{\alpha n}{2} + \mathrm{Lap}\left(\frac{1}{\varepsilon}\right),$$

which is the counting query version of equation (5) in the text. Because of GR-Theorem 3.3, we can use a common $\varepsilon = \frac{1}{\alpha n}$ in equation (5).

If this were not true, then we would have to consider query release mechanisms that had different values of ε for each individual in the population and therefore the value that enters equation (5) would be much more complicated. To ensure that each individual in H received ε_i-minimally differential privacy, the

algorithm would have to use the smallest ε_i that was produced for any individual. In addition, the FairQuery and MinCostAuction algorithms described next would not work because they depend upon being able to order the cost functions $\upsilon_i \varepsilon_i$ by υ_i, which is not possible unless ε_i is a constant or υ_i and ε_i are perfectly positively correlated. Effectively, GR-Theorem 3.3 proves that achieving (α, β)-accuracy with ε-differential privacy requires a mechanism in which everyone who sells a data-use right gets the best protection (minimum ε_i over all $i \in H$) offered to anyone in the analysis sample. If a change in the algorithm's parameters results in a lower minimum ε_i, everyone who opts to use the new parameterization receives this improvement. In addition, we argue in the text that when such mechanisms are used by a government agency they are also non-excludable because exclusion from the database violates equal protection provisions of the laws that govern these agencies.

Next, GR analyze algorithms that achieve $O(an)$-accuracy by purchasing exactly $\frac{1}{\alpha n}$ units of privacy loss from exactly $(1-\alpha)\, n$ individuals. Their algorithms *FairQuery* and *MinCostAuction* have the same basic structure:

- Sort the individuals in increasing order of their privacy cost, $\upsilon_1 \leq \upsilon_2 \leq \ldots \leq \upsilon_n$.
- Find the cut-off value υ_k that either exhausts a budget constraint (FairQuery) or meets an accuracy constraint (MinCostAuction).
- Assign the set $H = \{1, \ldots, k\}$.
- Calculate the statistic $\hat{s}$ using a differentially private algorithm that adds Laplace noise with just enough dispersion to achieve the required differential privacy for the privacy loss purchased from the members of H.

- Pay all members of H the same amount, a function of υ_{k+1}; pay all others nothing.

To complete the summary of GR, we note that GR-Theorem 4.1 establishes that FairQuery is dominant-strategy truthful and individually rational. GR-Proposition 4.4 establishes that FairQuery maximizes accuracy for a given total privacy purchase budget in the class of all dominant-strategy truthful, individually rational, envy-free, fixed-purchase mechanisms. GR-Proposition 4.5 proves that their algorithm MinCostAuction is a VCG mechanism that is dominant-strategy truthful, individually rational and $O(\alpha n)$-accurate. GR-Theorem 4.6 provides a lower bound on the total cost of purchasing k units of privacy of $k\upsilon_{k+1}$ GR-Theorem 5.1 establishes that for $\upsilon \in \mathbb{R}^n_+$, no individually rational mechanism can protect the privacy of valuations υ with $\left(k, \frac{1}{3}\right)$-accuracy for $k < \frac{n}{2}$.

In our application of GR, we use N as the total population. Our γ_i is identical to the GR υ_i. We define the query as a normalized query, which means that query accuracy is defined in terms of α instead of k; hence, our implementation of the VCG mechanism achieves $\left(\alpha, \frac{1}{3}\right)$-accuracy rather than $\left(\alpha N, \frac{1}{3}\right)$-accuracy. We define the individual amount of privacy loss in the same manner as GR.

A.3 Properties of the Indirect Utility Function in Section 4

We specify the indirect utility function for a given consumer as

$$v_i\left(y_i, \varepsilon, I, \tilde{y^i}, p\right) = -\sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln y_i - \gamma_i \left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right) \varepsilon + \eta_i \left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right) I$$

where $(\gamma_i, \eta_i) > 0$, $\xi_\ell > 0$, $\sum_{\ell=1}^{L} \xi_\ell = 1$ and $\mathrm{E}\left[\ln y_i\right] = \frac{1}{N}\sum_{i=1}^{N} \ln y_i$. To establish that this is an indirect utility function for a rational preference relation, we prove that

the vector v is homogeneous of degree zero in (p, y), nonincreasing in p, strictly increasing in y, quasiconvex in (p, y), and continuous in (p, y).

To prove that $v_i\left(y_i, I, \phi, y, p\right)$ is homogeneous of degree zero in (p, y), note that for all $\lambda > 0$

$$\begin{aligned} v_i\left(\lambda y_i, \varepsilon, I, \lambda y^{\tilde{}i}, \lambda p\right) &= -\sum_{\ell=1}^{L} \xi_\ell \ln\left(\lambda p_\ell\right) + \ln\left(\lambda y_i\right) - \gamma_i\left(1 + \ln\left(\lambda y_i\right) - \mathrm{E}\left[\ln \lambda y_i\right]\right)\varepsilon \\ &\quad +\eta_i\left(1 + \ln\left(\lambda y_i\right) - \mathrm{E}\left[\ln \lambda y_i\right]\right) I \\ &= -\sum_{\ell=1}^{L} \xi_\ell \ln \lambda - \sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln \lambda + \ln y_i \\ &\quad -\gamma_i\left(1 + \ln \lambda + \ln y_i - \mathrm{E}\left[\ln \lambda\right] - \mathrm{E}\left[\ln y_i\right]\right)\varepsilon \\ &\quad +\eta_i\left(1 + \ln \lambda + \ln y_i - \mathrm{E}\left[\ln \lambda\right] - \mathrm{E}\left[\ln y_i\right]\right) I \\ &= -\sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln y_i - \gamma_i\left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right)\varepsilon \\ &\quad +\eta_i\left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right) I \\ &= v_i\left(y_i, \varepsilon, I, y^{\tilde{}i}, p\right) \end{aligned} \tag{29}$$

since $\sum \xi_\ell = 1$ and $\ln \lambda = \mathrm{E}\left[\ln \lambda\right]$. Since homogeneity of degree zero holds for every v_i, it holds for v.

For all $\lambda > 1$,

$$\begin{aligned} v_i\left(y_i, \varepsilon, I, y^{\tilde{}i}, \lambda p\right) &= -\ln \lambda - \sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln y_i - \gamma_i\left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right)\varepsilon \\ &\quad +\eta_i\left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right) I \\ &< -\sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln y_i - \gamma_i\left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right)\varepsilon \\ &\quad +\eta_i\left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right) I \\ &= v_i\left(y_i, \varepsilon, I, y^{\tilde{}i}, p\right) \end{aligned}$$

since $\lambda > 1$, $\xi_\ell > 0$ for all ℓ and $\sum \xi_\ell = 1$. Therefore, v is nonincreasing in p.

For all $\lambda > 1$,

$$\begin{aligned} v_i\left(\lambda y_i, \varepsilon, I, \lambda y^{\tilde{}i}, p\right) &= -\sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln\left(\lambda y_i\right) - \gamma_i \left(1 + \ln\left(\lambda y_i\right) - \mathrm{E}\left[\ln \lambda y_i\right]\right) \varepsilon \\ &\quad + \eta_i \left(1 + \ln\left(\lambda y_i\right) - \mathrm{E}\left[\ln \lambda y_i\right]\right) I \\ &= -\sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln \lambda + \ln y_i \\ &\quad - \gamma_i \left(1 + \ln \lambda + \ln y_i - \mathrm{E}\left[\ln \lambda\right] - \mathrm{E}\left[\ln y_i\right]\right) \varepsilon \\ &\quad + \eta_i \left(1 + \ln \lambda + \ln y_i - \mathrm{E}\left[\ln \lambda\right] - \mathrm{E}\left[\ln y_i\right]\right) I \\ &> v_i\left(y_i, \varepsilon, I, y^{\tilde{}i}, p\right) \end{aligned}$$

since $\lambda > 1$ and $\ln \lambda = \mathrm{E}\left[\ln \lambda\right]$. Therefore, v is strictly increasing in y.

To prove quasiconvexity in (p, y), since we already showed $v_i\left(y_i, \varepsilon, I, y^{\tilde{}i}, p\right)$ is homogeneous of degree zero in (p, y), it suffices to show that the set $\left\{p \in \mathbb{R}^L_{++} : v_i(p, y) \leq \bar{v}\right\}$ is convex. But this follows simply from the concavity of the logarithmic function.

Continuity in (p, y) follows from the continuity of $\ln(x)$. Therefore, v is a vector of proper indirect utility functions.

A.4 The Multiplicative Weights Exponential Mechanism Algorithm

We provide a complete description of the MWEM mechanism based on the presentation in HLM.

To maintain consistency with the presentation in Sections 2 and 4, we present the MWEM algorithm using an unnormalized histogram to represent both the confidential and synthetic databases, and normalized linear queries operating on both the confidential and synthetic databases. This represents a departure from

the original presentation by HLM, which they give using an unnormalized histogram and unnormalized queries. All symbols in the algorithm described below have the same meaning as in our main text.

Algorithm *Multiplicative Weights Exponential Mechanism*

Input: An unnormalized histogram, x, from a database whose elements have cardinality $|\chi|$; number of records in the original database, $||x||_1 = N$; differential privacy parameter $\varepsilon > 0$; a number, T, of iterations; and a list of allowable normalized linear queries $\mathcal{Q}$ with cardinality $|\mathcal{Q}|$. Each normalized linear query, $f(x) \equiv \frac{1}{N} m^T x$ where $m \in [-1, 1]^N$.

1. Set the Laplace scale parameter: $\sigma = 2T/\varepsilon$.
2. Initialize the synthetic database: $\tilde{x}_0 = \frac{N}{|\chi|} u_{|\chi|}$, where $u_{|\chi|}$ is the unit vector of length $|\chi|$.
3. Initialize a probability distribution over $\mathcal{Q}$: $p_0 = \frac{1}{|\mathcal{Q}|} u_{|\mathcal{Q}|}$.
4. **for** $t \leftarrow 1$ **to** T
5. **for** each $f \in \mathcal{Q}$
6. Define score $s(x, f) \leftarrow |N(f(\tilde{x}_{i-1}) - f(x)|$.
7. Define $r(f) \leftarrow \exp(\varepsilon \times s(x, f)/4T)$.
8. **end for**
9. Update: $p_t \leftarrow [r(f)]_{f \in \mathcal{Q}}$.
10. Normalize: $p_t \leftarrow \frac{p_t}{||p_t||_1}$.
11. Sample f_t from $\mathcal{Q}$ given probability distribution p_t over $\mathcal{Q}$ (This is the exponential mechanism component.)
12. Sample A_t from $\mathrm{Lap}(\sigma)$.
13. Compute the noisy answer to f_t using the original database, $\hat{a}_t \leftarrow f_t(x) + A_t$. (This is the Laplace mechanism component.)
14. Compute the answer to f_t using the synthetic database, $\tilde{a}_t \leftarrow f_t(\tilde{x}_{[t-1]})$.

15. Compute the difference between the noisy and synthetic answers: $d_t \leftarrow \hat{a}_t - \tilde{a}_t$.
16. update mechanism: expend some of the privacy budget to update the synthetic data.
17. **for** $i \leftarrow 1$ **to** $|\chi|$
18. Update: $y_t[i] \leftarrow \tilde{x}_{t-1}[i] \times \exp(f_t(i) \times d_t/2)$.
19. Normalize: $\tilde{x}_t[i] \leftarrow N \times \frac{y_t[i]}{\sum_i y_t[i]}$.
20. **end for**
21. **end for**
22. Output: $\tilde{x} \leftarrow Avg_{t<T}\tilde{x}_t$

Here we highlight the key ideas as they relate directly to the notation we use in our analysis. HLM establish that the MWEM algorithm is ε-differentially private (their Theorem 2.1). In each of the T iterations, both the exponential mechanism and the Laplace mechanism are parametrized by $\varepsilon/2T$. Composition therefore implies ε-differential privacy. HLM state an error bound for MWEM in their Theorem 2.2. Their reported exact bound for unnormalized queries is $2N\sqrt{\frac{\log|\chi|}{T}} + \frac{10T\log|\mathcal{Q}|}{\varepsilon}$. We rescale the error bound by database size to account for the normalization. Converting to normalized queries gives an exact bound on α of $2\sqrt{\frac{\log|\chi|}{T}} + \frac{10T\log|\mathcal{Q}|}{N\varepsilon}$. HLM note that the optimal number of iterations is the value of T that minimizes the α bound. The optimal value is $T = \left(\frac{\varepsilon N\sqrt{\log|\chi|}}{10\log|\mathcal{Q}|}\right)^{2/3}$.

In practice, the basic algorithm requires some adjustment to give acceptable performance. None of these adjustments affect the privacy or accuarcy guarantees. HLM suggest such adjustments in their Sections 2.3.1 and 2.3.2. In particular, within each iteration the update rule may be applied to all previously sampled queries, multiple times, which can improve the fit of the synthetic database to the full query set without additional privacy loss. We include these variations in our

own experiment. The exact implementation details are reported in the replication archive that accompanies this article (Abowd and Schmutte 2017).

A.5 Details of Data Sources and Data Preparation

We use data from several different sources in our empirical applications. These data and the code used to prepare them are available in a replication archive (Abowd and Schmutte 2017).

A.5.1 Preparation of Synthetic Population Data on Income

To simulate publication of income statistics from a population database, we use the Bayesian bootstrap (Rubin 1981) to construct a synthetic population from the five-year ACS 5-year Public-Use Microdata Samples (ACS PUMS) for the years 2010–2014 (U.S. Census Bureau 2016). Our process is as follows:

- From the raw ACS files, retain the variables `ADJINC` (adjustment income factor), `PWGTP` (person weight), `AGEP` (age), `PINCP` (total income).
- Retain records for individuals with `AGEP` reported between 18 and 64, inclusive.
- Let n_{yr} be the number of records for 18–64 year-old individuals in each year. So n_{2010} is the number of such records from 2010, for example.
- We want to generate a synthetic population with size equal to the estimated population of 18–64 year old individuals in 2012, $N = 197,040,596$, reported by Census from the full ACS (U.S. Census Bureau, Population Division 2013).
- Let $n_{acs} = n_{2010} + n_{2011} + n_{2012} + n_{2013} + n_{2014}$

- Let w_i be the person-weight attached to individual i, reported as PWGTP.
- Let x_i be the data record of individual i.
- Let $yr(i)$ be the year associated with record i.
- Define α as a vector of length $\sum(n_{yr}) = n_{acs}$ whose entries are $(\tilde{w}_i * n_{yr(i)})$, where $\tilde{w}_i = w_i / \sum(w_i, \text{all years})$.
- Sample θ from the $Dirichlet(\alpha)$ distribution.
- Draw from the $Multinomial(N, \theta)$ distribution. This gives the list of observations from the survey data to retain in the population-level database.
- For the population database, convert reported income to 2012 dollars and then group income into bins with width \$2,000. This choice of width is arbitrary, and, importantly, non-private.
- As implemented, there are 797 bins for income. Bin 0 for income < -16000, Bin 796 for income $\geq$ 1,574,000, Bins 1 through 795 are in 2000 dollar increments. 2000 was selected arbitrarily. The upper and lower bins were set to match the bottom- and top-coding limits in the ACS PUMS.

A.5.2 Preparation of Synthetic Population Data on Health Status

We simulate publication of the population distribution of body-mass index (BMI) using the Bayesian bootstrap to construct a synthetic population from the National Health Interview Survey for the year 2015 (Minnesota Population Center and State Health Access Data Assistance Center 2016). Our process is as follows:

- From the raw data retain the variables `year`, `sampleweight`, and `bmi`. Retain only records for from 2015 and remove records that are out of universe or missing BMI.
- Let n_{nhis} be the number of records in NHIS year 2015 with reported BMI.
- The sampling frame for the BMI variable in NHIS is designed to be representative of the civilian non-institutional population age 18 or above in 2015. The size of that population is estimated to be $N = 242,977,154$ according to the Census 1-year ACS estimates for 2015 (U.S. Census Bureau, Population Division 2015).
- Let w_i be the person-weight for individual i (`sampleweight`).
- Let x_i be the data record of individual i.
- Define α as a vector of length n_{nhis} whose entries are $(\tilde{w}_i * n_n his)$, where $\tilde{w}_i = w_i / sum(w_i, \text{observed samples only})$. This rescales the weights to account for dropped observations with missing `bmi`.
- Sample θ from the $Dirichlet(\alpha)$ distribution.
- Draw from the $Multinomial(N, \theta)$ distribution. This gives the list of observations from the survey data to retain in the population-level database.
- We bin BMI into 800 bins of equal length between the allowed lower and upper limits of the `bmi`.

A.5.3 The Federal Statistical System Public Opinion Survey (FSS POS)

The FSS POS is a national public opinion survey conducted in conjunction with Gallup daily tracking surveys. It was developed by the Census Bureau in collab-

oration with other statistical agencies with the primary goal of measuring public trust in official statistics. After preliminary testing, the FSS POS was initially rolled out as a set of 25 questions added to the Gallup Overnight Tracking Poll. Between 2012 and 2015 there have been three waves of the FSS POS, with multiple rotations.

The set of variables collected in the FSS POS component change within and across each wave. However, the questions we focus on in this paper, as described in Section 5.2.2, are asked in every interview. In addition to the FSS POS component, the data include Gallup Daily Questions routinely asked of all respondents. Of these, we use categorical variables measuring gender, race, ethnicity, age (in nine bins), and monthly income (in 12 bins).

Not all respondents answer all questions. Rather than ignore the missing data, we multiply impute the missing values under a model that conditions on the observed values. Because the missing data pattern is not monotone, we use the fully conditional specification (FCS) method (Buuren et al. 2006). We specify a logistic regression model for each variable conditioning on the remaining variables, allowing the questions on privacy and accuracy attitudes to be fully interacted with each other and with income. We produce 500 implicates. The reported correlations and standard errors are computed using combining formulas that account for the within- and between-implicate variance (Gelman et al. 2013). As we note in the text, the results are qualitatively similar if we simply drop those cases with missing data values.

All variables used in the analysis are shown in the main text.

A.5.4 The Cornell National Social Survey

We use raw data from the the Cornell National Social Survey(CNSS) obtained from the CNSS integrated data application, obtained via [http://www.ciser.cornell.edu/beta/cnss/] by selecting all variables for all years (Cornell Institute for Social and Economic Research and Survey Research Institute n.d.). All variables used in the analysis are shown in the main text.