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
and
Ian M. Schmutte
Department of Economics, University of Georgia

February 6, 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 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 or the NSF. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme on Data Linkage and Anonymisation where some work on this paper was undertaken, 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 acknowledge helpful comments from Larry Blume, David Card, Cynthia Dwork, Stephen Fienberg, Ron Jarmin, Dan Kifer, Ashwin Machanavajjhala, Mallesh Pai, Jerry Reiter, Bruce Spencer, Lars Vilhuber and Nellie Zhao along with seminar and conference participants at the U.S. Census Bureau, Cornell University, CREST, 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 in the Digital Commons space of the Cornell Labor Dynamics Institute http://digitalcommons.ilr.cornell.edu/ldi/22/.

Abstract

We consider the problem of the public release of statistical information about a population–explicitly accounting for the public-good properties of both data accuracy and privacy loss. We first consider the implications of adding the public-good component to recently published models of private data publication under differential privacy guarantees using a Vickery-Clark-Groves mechanism. We show that data quality will be inefficiently under-supplied. Next, we develop a standard social planner's problem using the technology set implied by (ε, δ)-differential privacy with (α, β)-accuracy for the Multiplicative Weight Exponential Mechanism query release system to study the properties of optimal provision of data accuracy and privacy loss when both are public goods. Using the production possibilities frontier implied by this technology, explicitly parameterized interdependent preferences, and the social welfare function, we display properties of the solution to the social planner's problem. Our results directly quantify the optimal choice of data accuracy and privacy loss as functions of the technology and preference parameters. Some of these properties can be quantified using population statistics on marginal preferences and correlations between income, data accuracy preferences, and privacy loss preferences that are available from survey data. Our results show that government data custodians should publish more accurate statistics with weaker privacy guarantees than would occur with purely private data publishing. Our statistical results using the Federal Statistical System Public Opinion Survey combined with the American Community Survey and the National Health Interview Survey indicate that the welfare losses from under-providing data accuracy while over-providing privacy protection can be substantial.

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

1 Introduction

Like so many other ideas in information economics, George Stigler (1980) began the analysis of the economics of privacy. That analysis emerged alongside the observation by Posner (1981), drawn from contemporary legal analyses, of privacy as the right to conceal details about one's life from others, including the government. While most of Stigler's treatment addresses the origin of the demand for privacy by individuals, he identified the source of angst driving the public discussions in the 1970s by focusing squarely on the observation that: "[g]overnments (at all levels) are now collecting information of a quantity and in a personal detail unknown in history" (p. 623). And this more than a decade before the birth of the Internet. Stigler correctly predicted that the problem would be how to properly constrain the use of this information rather than how to defend against its acquisition in the first place.

As Acquisti and Varian (2005) note, 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. Acquisti et al. (2013) experimentally evaluate individuals' willingness-to-pay to protect otherwise public information and their willingness-to-accept payment for permitting the disclosure of otherwise private information. These experiments are explicitly set in the context of commercial enterprises that seek to acquire these private data as part of a mutually beneficial exchange with well-informed consumers. The prototypical example is online shopping. In the extensive literature that they review, the consumer's benefit from increased privacy is a direct consequence of the value of her private information to the counterparty in a commercial transaction. Specifically, they studied differences in consumer behavior when choosing between a $10 anonymous loyalty card and a

$12 identifiable card (transactions could be linked to the actual consumer). Acquisti et al. find that their experimental subjects displayed, for monetarily equivalent transactions: (1) unequal willingness-to-pay to protect private data versus willingness-to-accept payment to disclose the same private data and (2) order effects in their choices. Because of these endowment and order effects, they reject the normative conclusion that consumers value privacy very little based on their observed willingness to part with these data for very little compensation when shopping online. In this paper, we recognize such a behavioral effect by using explicit formulations of the payment systems and interdependent preferences to reason about the economic value of the privacy loss from statistical summaries.

In its current form, the economics of privacy focuses primarily on the economic value of information about the habits and characteristics of consumers that are known to the curators of databases produced by intermediating commercial transactions and social networks on the Internet. The information is known to these providers because it was either shared voluntarily or harvested without explicit notice during an interaction of the consumer with the Internet site. Acquisti et al. (2016) survey the informational asymmetries that result from this shared information. While they focus almost exclusively on the private value of the information, their survey also covers aspects of the externalities produced by these information exchanges. In the final section of their paper, they discuss some aspects of the scientific value of these databases, including those held by government agencies, noting that "[h]ow to balance researchers' and society's needs to access granular data with the need to protect individuals' records is a question that simultaneously involves economists and scholars in other disciplines, such as statisticians and computer scientists" (p. 43). That challenge is the core of our paper.

The Role of Statistical Agencies

What does the economics of privacy have to say about Stigler's Orwellian governmental databases? For agencies that enforce laws with criminal and civil penalties, the citizen/consumer's interest in concealing certain private information is apparent and amenable to study using the private valuation models we just introduced. But what would Stigler have said about the appropriate way to think about constraining the government's use of private personal information when that information is collected by an agency whose sole statutory purpose is to publish statistical summaries based on those personal data?

Stigler explicitly acknowledged the public-good nature of these publications, and, of course, he applied the Coase Theorem to make the following argument. The private information will be collected and disseminated efficiently if the property rights are fully assigned and the transactions costs of acquisition and dissemination are minimized. He recognized that dissemination was a very low marginal cost activity, even in 1980, and that using markets to control the re-use of the information after it had been acquired in a voluntary transaction between informed adults might remain very difficult. There is an important insight here for modeling statistical agencies. If one wishes to study their optimal use of private data, one must understand the derived demand for the statistical information those data convey to the citizens. In order to apply the Coase Theorem, one must understand both the social costs of the use of private information by agencies that collect it and the social benefits derived from its dissemination in statistical summaries. Whether or not there is a market failure to analyze, understanding efficient breaches of privacy requires modeling their full social cost and benefit.

In this paper we focus on the public-good properties of the statistical information disseminated by government agencies and the public-good properties of the

privacy protections they provide. We use techniques from economics, computer science, and statistics to make our arguments, but our main goal is to demonstrate that using methods from all three disciplines permits a more complete understanding of both the privacy protection technologies and the sources of the citizen/consumer's interest in accurate public data.

This is not a trivial proposition. Around the world, national statistical offices exist for the primary purpose of collecting and publishing data about their citizens and the businesses that operate within their jurisdictions. Since these are costly functions, and since most statistical agencies are prohibited from performing law enforcement functions using the data that they collect for statistical purposes, we need to model how the business of data provision directly relates to citizen demand for particular kinds of information. In our model, this demand arises because utility depends upon properties of the population that require statistical data to assess. This is not a new idea. Akerlof (1997) posited essentially the same interdependent preferences that we use when he hypothesized that utility might depend upon the deviation of the individual's choice from the average in the economy. How can one evaluate such preferences without data on the population averages? The literature that grew out of Akerlof's work took the existence of fully accurate population statistics as given, and assumed that they could be collected without any privacy loss.

Our consumers also display preference interdependence. Specifically, we assume that individuals care about their place in the income distribution and their relative health status within the population. They cannot evaluate these relative preferences without statistical information. They explicitly recognize that such data can have varying quality. If they acquire statistical information of known quality from a private provider who acquires data-use rights through a Vickery-

Clark-Groves (VCG) auction, the consumers won't buy accurate enough data because their private demand will not reflect the benefit that others in the population get from knowing that same information with given quality. We solve the complete social planner's problem when the accuracy of the published statistical data and the privacy loss from providing the confidential inputs are both public goods. We prove that the socially optimal data accuracy exceeds the VCG levels and the socially optimal privacy losses are greater than those generated by private data suppliers using the VCG mechanism.

Our work is thus related to a burgeoning literature in public economics on the role of preference interdependence in the provision of public goods. 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). Preference interdependence is also important for explaining discrepancies between macroeconomic and microeconomic outcomes (Futagami and Shibata 1998) and for the design of public policy. Aronsson and Johansson-Stenman (2008) show that preference interdependence affects the optimal provision of public goods, but the direction is theoretically ambiguous. Their work also shows that preference interdependence will affect the optimal tax schedule–an aspect of the public goods problem we ignore in our formulation in order to focus on the optimal trade-off between privacy loss and data accuracy. We think that our use of preference interdependence to generate the demand for accurate statistical data is an important contribution to this literature.

1.1 Technologies for Privacy Protection

Like so many other ideas in the efficient operation of statistical agencies, Ivan Fellegi (1972) initiated the statistical analysis of data confidentiality. Fellegi understood that ensuring the confidentiality of individual data collected by the agency, an essential obligation, was most likely to be threatened by what he called "residual disclosure"–what would now be called a "subtraction attack" in computer science or a "complementary disclosure" in statistical disclosure limitation (SDL). This breach of privacy occurs when the statistical agency releases so much summary information that a user can deduce with certainty some of the private identities or attributes by subtracting one tabular summary from another. Fellegi established the properties of what became the workhorse of SDL–primary and complementary suppression of items in the published statistical tables. Risky items–ones that reveal a citizen's private data–are suppressed–not published in the public table–and just enough non-risky items are also suppressed so that the table is provably secure from a subtraction attack. Armed with this tool, statistical agencies around the world adopted this practice and a large literature, nicely summarized in Duncan et al. (2011), emerged with related techniques. The choice of primary suppressions is usually based on one of several risk measure (see, for example, Federal Committee on Statistical Methodology (2005)). The choice of complementary suppressions is inherently *ad hoc* in the sense that many sets of complementary suppressions meet the criteria for protecting the risky items but the methods provide limited guidance for choosing among them.

To help assess the trade-off between privacy loss and data quality, statisticians developed another important disclosure limitation tool that is immediately accessible to economists–the risk-utility ($R - U$) confidentiality map. The $R - U$ confidentiality map first appeared in Duncan and Fienberg (1999), who used it to

characterize three different SDL strategies for publishing tabular count data. They did not label the graph an $R-U$ confidentiality map. Duncan et al. (2001) named the $R-U$ confidentiality map. They used it to model the trade-off between the disclosure risk associated with a particular privacy protection method and the accuracy of the released statistical summaries, which they called "data utility." A full treatment can be found in Duncan et al. (2011, p. 125-135). Economists will instantly recognize the $R-U$ confidentiality map as the production possibilities frontier for the data publication technology when it is constrained by the requirement to protect against privacy loss. In this paper, we complete the formalization of this idea by deriving the exact PPF for our privacy-preserving publication technology as part of our public-goods model. In what follows, we will reserve the term "utility" for its usual role in economic theory.

It was another seminal contributor to the methodology of statistical agencies, though, who first posed the SDL problem in the form that has become the dominant methodology in computer science. Tore Dalenius (1977) hypothesized that it was insufficient for a statistical agency to protect against direct disclosures of the type studied by Fellegi. In Dalenius' model, the statistical agency also had to protect against providing so much information that a user could "determine the value" of a confidential item "more accurately than is possible without access to" the publicly released statistical summary (p. 433). This definition of a statistical privacy breach is now called *inferential disclosure*. In statistics, Duncan and Lambert (1986) completed the mathematical formalization of inferential disclosure by showing that the appropriate tool for studying such privacy losses was the posterior probability of the confidential item given the released statistical summaries. In cryptography, Goldwasser and Micali (1982; 1984) defined a semantically secure encryption as one in which the posterior probability of any cleartext message,

given the cyphertext, equals the prior probability of the same message.[1] Thus, bounding the posterior odds of an inferential disclosure, the conceptual analog of an unauthorized decryption, became the formalization at the heart of the modern data privacy literature.

1.2 The Emergence of the Differential Privacy Paradigm

Cryptographers know how to protect secrets. In the early 2000s, a group of cryptographers led by Cynthia Dwork (2006) and including Dwork et al. (2006) formalized the privacy protection associated with SDL in a model called ε-differential privacy. Using this framework, Dwork and Naor (2008) proved that it was impossible to deliver full protection against inferential disclosures because a privacy protection scheme that provably eliminated all such disclosures was equivalent to a semantically secure encryption of the confidential data, and therefore useless for data publication.[2] She proposed developing a privacy protection method that "captures the increased risk to one's privacy incurred by participating in a database" (p. 1), which she parameterized with $\varepsilon \geq 0$, where $\varepsilon = 0$ is full protection.

Dwork (2008, p. 3) foreshadowed our view that the differential privacy parameter is a public good when she wrote: "[t]he parameter ε ... is public. The choice of ε is essentially a social question." We begin our own analysis using the electronic commerce view of McSherry and Talwar (2007), which closely resembles the framework that grew out of Stigler's "incentive to conceal" notion of personal privacy. Data custodians may purchase data-use rights from individuals whose information was collected for legitimate but unrelated business purposes in order

[1] We thank Cynthia Dwork for drawing our attention to the work of Goldwasser and Micali.
[2] Evfimievski et al. (2003) prove a similar result using a related definition of privacy.

to compute and release an additional statistical summary that was not originally planned. The purchase is a private transaction between informed agents. However, a direct consequence of the electronic commerce privacy work, as proven by Ghosh and Roth (2011), is that privacy protection for this type of statistical data release has a public good character–it is non-rival (Mas-Colell et al. 1995, p. 359)–just as Dwork originally noted.

The amount of privacy an individual sacrifices by participating in an ε-differentially private mechanism neither exacerbates nor attenuates the expected sacrifice of privacy for any other individual in the database. The protection provided by differential privacy (our Definition 2, which is identical to the one found in Dwork and Roth (2014)) bounds the supremum across all individuals of the privacy loss–it is worst-case protection for the entire database. Thus, differential privacy is inherently non-rival. Any improvement in privacy protection is enjoyed by all entities in the database, and any reduction in privacy is suffered by all entities.

A subtle distinction emerges when considering the difference between voluntary and compulsory systems for participation in the database versus participation in the statistical summaries. Specifically, when an opt-in system is used for producing the summaries, all those who elect to participate get ε-differential privacy by construction of the payment system. Those who opt out get 0-differential privacy. In compulsory participation systems, all entities in the database get ε-differential privacy. In either case, all members of the population receive at least ε-differential privacy because $\varepsilon > 0$. For statistical agencies using population censuses and administrative record systems, participation in the database and in the statistical summaries is usually compulsory. Our analysis of the suboptimality of private provision permits opting out of the statistical summaries but not the database. Our analysis of optimal public provision assumes compulsory

participation in the both the database and the statistical summaries. The opt-in method, which is a private provider's only feasible technology, may produce biased summaries–a possibility that we do not analyze in this paper because it was already recognized by Ghosh and Roth (2011), who carefully defined the target level of accuracy to control self-selection bias.[3]

There were precursors to the differential privacy paradigm. Denning (1980) studied the security risks of releasing summaries based on samples from a confidential database. Agrawal and Srikant (2000) coined the phrase privacy-preserving datamining and analyzed some preliminary ways to accomplish it. Sweeney (2002) formalized the protection provided by SDL methods that guard against identity disclosure with a model known as k-anonymity. Machanavajjhala et al. (2007) formalized SDL methods that guard against attribute disclosure with a model known as ℓ-diversity. Evfimievski et al. (2003) explicitly modeled privacy breaches based on posterior predictive distributions in a formal setting very similar to differential privacy. But it is the differential privacy algorithms, and their explicit formalization of inferential disclosure protection, that have become the workhorse of the computer science data-privacy literature. We base much of our modeling on the methods in Dwork and Roth (2014). For economists, Heffetz and Ligett (2014) is a very accessible introduction.

1.3 Current Economic Uses of Differential Privacy

It isn't just statistical agencies that release data as a public good. The standard definition of a public good is that its use by one individual does not preclude its use by another–non-rivalry in consumption. Sometimes a second condition

[3]They nevertheless acknowledge that bias in the privately-provided summary statistics may still exists in their solution (Ghosh and Roth 2011, Remark 5.2).

is added that one person's use of the public good does not exclude another's use–non-exclusion in consumption. The second condition is not essential, and governments often expend resources to allow exclusive use of otherwise public data when they enforce patents and copyrights. It is easy to see how a statistical agency's publication of data on the distribution of income in the society, the cost of living, incidence of diseases, or national income accounts satisfies the non-rivalry and non-exclusivity conditions. It is perhaps less obvious, but equally true, that the release of statistics about users, searches, "likes," or purchases associated with businesses like Amazon, Facebook and Google also satisfies these conditions. In addition, the publication of a scientific article based on confidential information provided by a statistical agency or proprietary information provided by a business satisfies these conditions.

Our work builds on the very thorough analysis in Ghosh and Roth (2011), who study the specific problem of compensating a sample of individuals for the right to use their data to compute a statistic from a private database already containing those data–think: tabulations using Facebook friend networks. Each individual who agrees to sell her data-use right is included in the published statistic, which has a specific level of accuracy and is computed using an auction-determined level of differential privacy protection. Their central contribution is to characterize the properties of a VCG mechanism that achieves a specified query accuracy for the population statistic with the least-cost acquisition of data-use rights (privacy loss).[4]

We build on the Ghosh and Roth problem by allowing the privacy-preserving answer to the query to be a public good. This is clearly within the spirit of their

[4]The electronic commerce applications of differential privacy begin with the work of McSherry and Talwar (2007), which studied mechanism design using differential privacy. They showed that mechanisms designed using ε-differential privacy limit the players' incentive to lie because they bound the expected gain from any coalition of size t deviating from truth-telling by $(\exp(t\varepsilon) - 1)$.

work since they motivate their problem by modeling a data analyst who wishes to obtain the most accurate estimate of a statistic within the constraints of a grant budget. Most sponsored research is published in open-access scientific journals, making the statistic under study by Ghosh and Roth a classic public good. Although the scientist elicits data for the study, and subsequently publishes the results in an open journal, the individuals who sell their data-use rights to the scientist are presumed to get no utility from the published results in the Ghosh and Roth framework. We let the subjects care about the quality of the scientific paper. As noted above, we also consider whether it is reasonable to treat privacy loss itself as a fully-private good, especially since the Ghosh-Roth mechanism implies that privacy loss is non-rival. Privacy-preserving publication by statistical agencies treats all citizens as equally protected under the relevant confidentiality laws. Our paper is therefore the first use of the differential privacy paradigm to compare the economic implications of public and private provision of privacy-preserving statistical data in which both data quality and privacy loss are public goods.

1.4 Plan of This Paper

Section 2 provides a concise summary of the privacy and confidentiality models we use that is accessible to readers familiar with the computer science data-privacy literature. We also provide sufficient detail on the legal, economic and statistical underpinnings of our work so that readers can understand the relevance of our arguments. Section 3 lays out the formal definitions of databases, histogram representations, query release mechanisms, (ε, δ)-differential privacy, and (α, β)-accuracy. This section is self-contained and includes a brief restatement of the impossibility proof for eliminating inferential disclosures. Section 4 proves the result that data accuracy is under-provided and privacy loss is too low when a private

data supplier uses the VCG data-use rights acquisition mechanisms as compared to the social optimum implied by the full public-goods model. Section 5 develops an efficient technology for providing accurate public data and differentially private protection that admits a proper PPF. Using this technology and well-defined interdependent preferences we solve the social planner's problem for the optimal data accuracy and privacy loss. Section 6 uses data to quantify the parameters of the social planner's problem. We consider the publication of income distribution and relative health status statistics for the population. We quantify the welfare loss from suboptimal overprovision of privacy protection and underprovision of data accuracy. Section 7 concludes.

2 Background

2.1 Differential Privacy and Statistical Disclosure Limitation

We work with the concept of differential privacy introduced by Dwork (2006). To reduce confusion, we note that the SDL literature defines confidentiality protection as the effort to ensure that a respondent's exact identity or attributes are not disclosed in the published data. Computer scientists define data privacy as limits on the amount of information contained in the published data about any person in the population, respondent or not. The two literatures have much in common but the main point of commonality that we use here are definitions of inferential disclosure, due to Dalenius (1977), and differential privacy, due to Dwork (2006).

Inferential disclosure parameterizes the confidentiality protection afforded by a particular SDL method using the ratio of the posterior odds of correctly assigning an identity or a sensitive attribute to a particular respondent, given the newly released data, to the prior odds, given all previously released data. Differ-

ential privacy parameterizes the privacy protection in a data publishing system by bounding the same posterior odds ratio for all potential respondents in all potential configurations of the confidential data.

2.2 Statistical Data Releases and Privacy Protection Are Both Public Goods

Publishing statistical data, whether the output of a government agency or of a open scientific study, involves making statistical summaries of the information that has been collected from the population under study available for anyone to use. Consistent with this principle, we formalize publishing statistical data as applying a query release mechanism with given privacy and accuracy properties to a confidential database. Formally, in terms of the differential privacy model summarized in Section 2.1, the answers to a fixed set of queries with (α, β)-accuracy from the MWEM query release mechanism with (ε, δ)-differential privacy are published by the agency. Any individual may, therefore, use these statistics for any purpose. Hence, they are public goods because their use is both non-rival and non-exclusive.

We also assume that the ε parameter of the (ε, δ)-differential privacy guarantee is a public good. Such an assumption means that all citizens are protected by the same (ε, δ)-differential privacy parameters even though they may place different utility values on ε. This is our interpretation of the "equal protection under the law" confidentiality-protection constraint that most national statistical agencies must provide. See, for example, U.S. Code Title 13 and Title 44 for an explicit statement of this provision for the American data laws that govern the U.S. Census Bureau (U.S. Code 1954) and American statistical agencies in general (U.S. Code 2002).

In this equal-protection sense, privacy protection is non-exclusive in consumption in the same manner as access to legal recourse through the courts is non-exclusive–it is a right of citizenship. But unlike access to the courts, where there is rivalry in consumption because one party's litigation congests the access of an unrelated party's litigation, statutory privacy protection is non-rival when it is provided via differential privacy. The reason for the non-rivalry is that the differential privacy protection is "worst case" protection. If the query release mechanism's worst possible breach is limited by the differential privacy bounds, then every citizen's protection is increased or decreased when the bounds are tightened or loosened, respectively. Alice can have more privacy in this sense if and only if Bob also enjoys the same increment. There is no crowding out of one party's privacy protection when privacy protection is provided to another party.

In our formal setup, only the published-data accuracy parameter α and the privacy protection parameter ε are considered explicit objects of production and consumption. These are the formal public goods. We hold the other parameters of the data publication process constant. It is a subject for future work to make these choices endogenous.

3 Preliminaries

This section provides all formal definitions used in our application of differential privacy. The goal is to highlight the important 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).

3.1 Databases, Histograms and Queries

A statistical agency, or other data curator, is in possession of 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. That is, any row that appears in the database is an element of χ.[5] All variables are discrete and finite-valued. This does not impose a limitation, since continuous data are always given discrete, finite representations when recorded on censuses, surveys or administrative record systems.

3.1.1 Histograms

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

[5] For example, if the variables recorded in the database are a binary indicator for gender, $g \in \{0, 1\}$, and a categorical index for six different completed levels of schooling, $s \in \{1, \ldots, 6\}$, then $\chi = \{0, 1\} \times \{1, \ldots, 6\}$.

1.[6]

3.1.2 Queries

A *linear query* is a mapping $f : [-1,1]^{|\chi|} \times \mathbb{Z}^{*|\chi|} \rightarrow \mathbb{Z}^*$ such that $f(m,x) = m^T x$ where $x \in \mathbb{Z}^{*|\chi|}$ and $m \in [-1,1]^{|\chi|}$. A *counting query* is a special case in which m_i is restricted to take a value in $\{0,1\}$. Counting queries return the number of observations that satisfy particular conditions. They are the tool an analyst would use to calculate multidimensional margins for the contingency table representation of the database. A *normalized linear query* is a mapping $f : [-1,1]^{|\chi|} \times \mathbb{Z}^{*|\chi|} \rightarrow [0,1]$ 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, or averages, rather than counts. These correspond to the proportions from a contingency table or the cell averages in a general summary table. To that end, we work with normalized linear queries unless otherwise specified. The use of normalization is not restrictive. It only affects the functional form of privacy and accuracy bounds via 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.

3.2 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 sequence of k normalized

[6] If x is the histogram representation of D, y is the histogram representation of D', and D' and 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 related literature because definitions may be stated in terms of adjacent databases or adjacent histograms.

linear queries drawn from the query space $\mathcal{F}$.

Definition 1 (Query Release Mechanism) Let $\mathcal{F}$ be a set of normalized linear queries with domain $[-1,1]^{|\chi|} \times \mathbb{Z}^{*|\chi|}$ and range $R \subseteq [0,1]$, and let k be the number of queries to be answered. A query release mechanism M is a random function $M : \mathbb{Z}^{*|\chi|} \times \mathcal{F}^k \to R^k$ whose inputs are a histogram $x \in \mathbb{Z}^{*|\chi|}$ and a set of k normalized linear queries $f = (f_1, \ldots, f_k) \in \mathcal{F}^k$. The probability of observing $B \subseteq R^k$ is $\Pr[M(x, (f_1, \ldots, f_k) \in B | X = x, F = f]$, where $\Pr[z \in B | X = x, F = f]$ is the conditional probability given $X = x$ and $F = f$ that the query output is in $B \in \mathcal{B}$, where $\mathcal{B}$ are the measurable subsets of R^k.

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, δ)-differential privacy) A query release mechanism M satisfies (ε, δ)-differential privacy if for $\varepsilon > 0$, $\delta > 0$, $\forall x, x' \in N_x$, and $\forall B \in \mathcal{B}$

$$\Pr[M(x, (f_1, \ldots, f_k)) \in B] \leq e^{\varepsilon} \Pr[M(x', (f_1, \ldots, f_k)) \in B] + \delta,$$

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

We now clarify the relationship between differential privacy and inferential disclosure. Our argument is a simplified version of Dwork (2006) that uses our definitions. Using Definition 2 consider the ratio that results from using the query release mechanism on two adjacent histograms x, x' conditional on the query se-

quence $f_1, \ldots, f_k$ and $\delta = 0$

$$\frac{\Pr\left[M(x, (f_1, \ldots, f_k)) \in B\right]}{\Pr\left[M(x', (f_1, \ldots, f_k)) \in B\right]} = \frac{\Pr\left[M(x, (f_1, \ldots, f_k) \in B | X = x, F = f\right]}{\Pr\left[M(x', (f_1, \ldots, f_k) \in B | X = x', F = f\right]}.$$

Without loss of generality the histograms x and x' can be treated as N samples from a discrete Multinomial distribution with probabilities π defined over χ, and holding the query sequence constant at $F = f$. We can compute $\Pr\left[X = x \,|\pi, N, F = f\right]$ and $\Pr\left[X = x' \,|\pi, N, F = f\right]$. A direct application of Bayes Theorem yields

$$\frac{\Pr\left[M(x, (f_1, \ldots, f_k) \in B | X = x, F = f\right]}{\Pr\left[M(x', (f_1, \ldots, f_k) \in B | X = x', F = f\right]} = \frac{\frac{\Pr[X=x|B,\pi,N,F=f]}{\Pr[X=x'|B,\pi,N,F=f]}}{\frac{\Pr[X=x|\pi,N,F=f]}{\Pr[X=x'|\pi,N,F=f]}} \leq e^{\varepsilon}, \qquad (2)$$

where 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, *i.e.*, the state of knowledge about x versus x' before B is released. As we noted in the introduction, this is precisely the Duncan and Lambert (1986) formalization of Dalenius (1977), although Duncan and Lambert's procedure is not based directly on the posterior odds ratio.

It should now be clear why we characterize differential privacy as worst-case privacy protection: it bounds the posterior odds ratio for inferential disclosure by e^{ε} over all possible publication outputs, B, considering every member of the population as potentially excluded from the database, N_x. It should also be clear why Dalenius' statement that "[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..." (Dalenius 1977, p. 433) 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 public-use 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. And in the language of statistical disclosure limitation, the $R - U$ confidentiality map must go through the origin–if there is no risk of privacy breach, there can also be no utility from the public-use statistics.

Accuracy

We can now define our measure of accuracy. The mechanism receives a sequence of normalized linear queries, $f_1, f_2, \ldots, f_k$ from $\mathcal{F}$, and returns, in real time, answers, $a_1 = M(x, (f_1))$, $a_2 = M(x, (f_1, f_2))$, ..., $a_k = M(x, (f_1, \ldots, f_k))$. These answers depend on the input database, the content of the query response, and the randomization induced by the query release mechanism.

Definition 3 ((α, β)-accuracy) A query release mechanism M satisfies (α, β)-accuracy for query sequence $\{f_1, f_2, \ldots, f_k\} \in \mathcal{F}^k$, $0 < \alpha \leq 1$, and $0 < \beta \leq 1$, if

$$\min_{1 \leq i \leq k} \{\Pr[|a_i - f_i(x)| \leq \alpha]\} \geq 1 - \beta.$$

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 sequence of k queries. The probabilities in the definition of (α, β)-accuracy are induced by the query release mechanism.

4 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 inno-

vative 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 $(\varepsilon, 0)$-differential privacy and (α, β)-accuracy.[7]

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, we make the accuracy of the statistic computed via the GR mechanism a public good whose demand is endogenous to our model. We show that private provision results in a suboptimally low level of accuracy and too little privacy loss. 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 mechanism delivers $(\varepsilon, 0)$-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 the public good–it summarizes the quality of the information that any consumer may access and use

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

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.

Our argument for suboptimal provision rests on two observations. First, the mechanism proposed by GR remains a minimum cost mechanism in our setting. Second, even if privacy loss were a partially excludable non-rival public good, accuracy would still be under-provided in the private market. These results follow from considering the use of the VCG mechanism by a private competitive data quality supplier for the procurement of privacy protection by a profit-maximizing data curator acting as a price-discriminating monopsonist when buying data-use rights.[8]

Suboptimality of 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 quality, a supplier using the VCG mechanism buys just enough privacy-loss rights to sell the data quality to the consumer with the highest data-quality valuation. All other

[8]The VCG mechanism implies a single price for each data-use right purchased. In theory, if the provider had access to a Lindahl mechanism (Mas-Colell et al. 1995), it could perfectly price discriminate when compensating consumers for their loss of privacy when procuring data-use rights. As long as property rights over privacy exposure are well-defined clear, the Lindahl private producer would internalize the full social cost of the required privacy reduction, but not the social benefit of increased data accuracy to the free-riding consumers who did not pay. In results available from the authors, we show that even in the Lindahl case data quality is under-produced compared to the social optimum and privacy protection is over-produced; i.e., there is too little privacy loss. Since the Lindahl mechanism is based on the unrealistic assumption that consumers' heterogeneous preferences for privacy are common knowledge, we do not focus on that case here.

consumers use the published data for free.

4.1 Model Setup

Following Ghosh and Roth (2011), each of the N private individuals possesses a single bit of information, b_i, that is already stored in a database maintained by a trusted curator.[9] For example, as in our first empirical application, this information could be the response to a single query about income of the form $b_i = 1$ if $y_i > y^*$ and $b_i = 0$ otherwise. Individuals each consume one unit of the published statistic, which has information quality I defined in terms of (α, β)-accuracy, that is $I = (1 - \alpha)$. Since I is a 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 quality, I, privacy loss, ε_i, and log income, $\ln y_i$.[10] The term $p_\varepsilon \varepsilon_i$ represents the total payment an indi-

[9] Trusted curator can have a variety of meanings. We mean that the database is held by an entity, governmental or private, whose legal authority to hold the data is not challenged and whose physical data security is adequate to prevent privacy breaches due to theft of the confidential data themselves. We do not model how the trusted curator got possession of the data, but we do restrict all publications based on these data to use statistics produced by a query release mechanism that meets the same privacy and confidentiality constraints. Therefore, no data user has privileged access for any query. These requirements closely mirror the statutory requirements of U.S. statistical agencies.

[10] 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

vidual receives if her bit is used in an $(\varepsilon_i, 0)$-differentially private mechanism. p_ε is the common price per unit of privacy, also to be determined by the model. 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 quality may be under-provided given its public-good properties. Violations of privacy might affect the goods market through targeted advertising and price discrimination as noted in Section 1. Accuracy of public statistics may also spill over to the goods market in important ways, in part by making firms more efficient, and thus able to produce and sell goods more cheaply. 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.

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 $\forall i$ in their Theorem 3.3. Income and accuracy are added to the Ghosh and Roth utility function because they are required for the arguments in this section. In Section 5 we develop a more complete model of the demand for accurate public-use statistics that includes interdependent preferences.

4.2 Cost of Producing Data Quality

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

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

with $(\alpha, 1/3)$-accuracy requires $\varepsilon_i = \varepsilon = \frac{1/2 + \ln 3}{\alpha N}$ for $H = N - \frac{\alpha N}{1/2 + \ln 3}$. In equation (5), the term $\text{Lap}(\sigma)$ represents a draw from the Laplace distribution with mean 0 and scale parameter σ.

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. They provide two mechanisms for implementing their VCG auction. We rely on their mechanism *MinCostAuction* and the properties they establish in Proposition 4.5. See Appendix A.2

We now derive the producer's problem of providing the statistic for a given level of data quality, which we denote by $I = (1 - \alpha)$. If p_ε is the payment per unit of privacy, 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. p_ε is 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. $H(I)$ 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

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

4.3 Private, Competitive Supply of Data Quality

Suppose a private profit-maximizing, price-taking, firm sells $\hat{s}$ with accuracy $(\alpha, 1/3)$, that is, with data quality $I = (1-\alpha)$ 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

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

where the solution is evaluated at I^{VCG}.[11] The price of data quality 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 number of people from whom data-use rights with privacy protection ε must be purchased. The second 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. As long as the cost function is strictly increasing and convex, the existence and uniqueness of a solution is guaranteed.

4.4 The Competitive Market for Data Quality When It Is a Public Good

At market price p_I, consumer i's willingness to pay for data quality 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 quality 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 by $I^{VCG} > 0$. We next show that there is only one such

[11] 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 postivie, terms in the second derivative. Sufficient conditions are that $Q()$ is the quantile function from the lognormal distribution (as we assume in Section 5) 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.

consumer.

It is a straightforward to verify that the consumers are playing a classic free-rider game (Mas-Colell et al. 1995, pp. 361-363) across N agents. In the competitive equilibrium, the only person willing to pay for the public good is the one with the maximum value of η_i. All others will purchase zero data quality but still consume the data quality purchased by this lone consumer. Specifically, the equilibrium price and data quality 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 quality, 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 quality will be under-provided by a competitive supplier when data quality 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).[12]

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

[12]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 statististic so that all other consumers may use it. That is why we have modeled this as a public good. And it is fully consistent with GR's scientist seeking data for a grant-supported publication.

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 $\left(\frac{1/2+\ln 3}{(1-I)N}, 0\right)$-differentially private and achieves $(1-I, \frac{1}{3})$-accuracy, which is exactly what Ghosh and Roth prove.

4.5 Proof of Suboptimality

Theorem 1 If preferences are given by equation (3), the query response mechanism satisfies equation (8) for $(\varepsilon, 0)$-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 4.3), 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. ■

5 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 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 describe the rate at which privacy must be sacrificed to increase accuracy of the published

statistics. The optimal choice along this frontier depends on the willingness of individuals to pay for increased accuracy with reduced privacy protection. 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.

Our model assumes the data have already been collected, and we do not model the monetary costs of collection. By deliberately abstracting from the public-finance problem, we focus on the costs that arise from the disutility of foregone privacy. Furthermore, in many settings, the costs of data collection are independent of data publication. For example, the U.S. government is constitutionally required to undertake the Decennial Census of Population. The data, having been collected, is a resource to be allocated toward producing population statistics. Similarly, administrative data are collected in the process of managing public programs. Our analysis describes how the information embedded in the collected data should be optimally allocated between privacy protection and production of accurate statistics. The social cost of data quality is measured in terms of the privacy loss when the agency publishes data, not when it collects those data.

5.1 Modeling Production Possibilities

We model a data custodian tasked with releasing public statistics calculated from a confidential database, D. The database contains a measurement, x_i, from each member of a population of size N. We follow the formal privacy literature in assuming x_i is a categorical, possibly ordinal, variable drawn from a domain $\rfloor\langle\rangle$. As in that literature, we observe this is without loss of generality since in practice, the set of acceptable values for continuous data is always finite. With some abuse of notation, we let x denote the histogram representation of D, so $\|x\|_1 = N$.

The custodian will publish the results of a set of linear queries, which generalizes the common practice of publishing contingency tables. To do so, the custodian operates a query release mechanism that is (ε, δ)-differentially private. Our goal is to determine where the custodian should set ε to optimally trade off privacy protection and accuracy of the published statistics. In what follows, we assume the custodian operates the Multiplicative Weights Exponential Mechanism (MWEM), the details of which we describe shortly. However, our analysis is generally valid for all differentially private mechanisms that yield a convex relationship between privacy loss and accuracy.[13]

5.1.1 The Multiplicative Weights Exponential Mechanism

The MWEM mechanism was introduced by Hardt et al. (2012). To operate the mechanism, the custodian chooses a subset, $\mathcal{Q}$, of feasible normalized linear queries $f \in \mathcal{F}$ to publish. The custodian also sets the privacy parameters (ε and, if possible, δ).

We summarize here the basic features of MWEM needed to understand our application. A more complete description appears in Appendix A.4. To understand MWEM, it is useful to first describe a simpler, but less efficient, algorithm: the Laplace Mechanism. One can think of the parameter ε as representing a fixed privacy 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

[13] One such mechanism is the Private Multiplicative Weights (PMW) mechanism, 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 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.

a random error drawn from the Laplace distribution with scale parameter $\frac{|\mathcal{Q}|}{\varepsilon}$. This approach works due to the additive composibility of (ε, δ)-differential privacy composes (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 unacceptably large.

MWEM economizes on expenditure of the privacy budget relative to the Laplace Mechanism as follows. The algorithm stores both the true data as well as synthetic data with the same structure that is not derived from the confidential data except according to the following procedure. For example, the synthetic data might be initialized with a uniform distribution across cells. At each round, the algorithm computes every query on the true data and the synthetic data. The query score is the absolute value of their difference. The algorithm selects a query at random with weight proportional to the query score, so that queries approximated poorly by the synthetic data are at higher risk of selection. The algorithm applies Laplace noise to the query applied to the true data, and then weights the entries in the synthetic database to match the noisy query response. In MWEM, the privacy budget is hence drawn down only for queries that are answered poorly. Upon completion, the custodian can publish answers to all queries, or the synthetic data, or both.

The strengths of this approach 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. This conserves on the privacy loss and controls the total error efficiently.

5.1.2 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 Let D be a confidential database with rows that are elements from the set χ with histogram x from population size $\|x\|_1 = N$. Let the set of all allowable normalized linear queries be $\mathcal{Q} \subseteq \mathcal{F}$ with cardinality $|\mathcal{Q}|$. Given $\varepsilon > 0$, the MWEM mechanism can deliver public answers to all queries in $\mathcal{Q}$ for that satisfy the following conditions:

1. Privacy: MWEM satisfies $(\varepsilon, 0)$-differential privacy;

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

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

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

Proof. ■

5.1.3 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. To relate our exposition to risk-return analysis, we define information quality as $I = (1 - \alpha)$ and characterize the PPF between it and differential privacy loss, ε, by a transformation function

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

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 quality 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

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

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

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

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.[14]

Figure 1 illustrates the PPF for our application to publication of statistics on the distribution of income, which we describe in detail in Section 6. We graph The PPF described by equation (16) with ε on the horizontal axis and I on the vertical axis. Because ε is a "bad" rather than a "good", 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 a diminishing marginal rate of transformation: it is increasingly costly, in terms of foregone privacy, to increase information quality.

We treat the parameters $(|\chi|, |\mathcal{Q}|, N)$ 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 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 which which to deliver the published statistics. The PPF determines the set of feasible pairs given the information budget. How to select the socially-optimal pair from this set is the problem we turn to next.

Before moving on, note that our framework could be extended to make the parameters governing the information budget endogenous. To obtain better privacy for a fixed level of desired accuracy, the custodian could, for example, limit the set

[14] 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; the accuracy bound is messy, but the closed form is straightforward to derive. See Appendix A.4.

of queries to publish. If we were to assume further data collection were possible, increasing the size of the database, N, could also shift the PPF.

5.2 The Optimal Level of Privacy

Given the data publication technology described above, the data custodian must ultimately choose a level of privacy protection, and with it, a guaranteed level of accuracy in the published statistics. From the perspective of our model, this is equivalent to choosing a target level of statistical accuracy then setting the minimum feasible amount of privacy loss under the publication technology. 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 privacy protection. When information quality and privacy protection are public goods, the solution is not obtained through market pricing. We therefore ask in this subsection what level of privacy a utilitarian social planner would choose to deliver. As already discussed, the answer depends on the average willingness to pay for data quality in terms of foregone privacy.

5.2.1 Preferences

We assume data are collected from all members of the population, and all members of the population may use the published statistics. Every person also consumes a set of pure private goods. Our formulation allows for arbitrary heterogeneity across individuals in preferences for privacy loss and the accuracy of published statistics. In doing so, we allow for the empirically relevant possibility that one group of people cares primarily about privacy, while getting little utility from consuming the data, while another set cares primarily about data quality.

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, which are common to all individuals in the population. 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 is data describing the distribution of income or the distribution of a health indicator.

5.2.2 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).

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 information quality. Intuitively, it measures the willingness-to-pay for increased privacy in terms of foregone data quality.

6 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 and data from the National Health Interview Survey (NHIS) to simulate publication of data 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 5.

To find the optimal levels of data quality and privacy loss, we specify a model in which individual preferences depend on others' outcomes. It is motivated by models of interdependent preferences (Pollak 1976; Card et al. 2012; Akerlof 1997; Alessie and Kapteyn 1991). For example, 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 distribution of preferences for data quality

and privacy, along with income, and health status. In what follows, we first define preferences in general terms and derive the solution to the social planner's problem. We then characterize the optimal solution to the data publication problem in each of our two applications, using data from opinion surveys to estimate the willingness of the social planner to pay for decreased privacy loss with reduced accuracy.

6.1 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 $\sum_{\ell=1}^{L} \xi_\ell = 1$.[15] 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.[16]

[15] 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.

[16] 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 quasiconcave in (ε, I), which is trivially true for equation (22), as long as

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

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

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

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

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

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

and

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

$(\gamma_i, \eta_i) > 0$ for all i, since it is linear in (ε, I). Hence, equation (19) is a well-specified social welfare function, quasiconcave in (ε, I), and the social planner's problem is well-specified since equation (16) is quasiconcave in (ε, I).

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

6.2.1 Publication Technology

We assume 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 population with reported income according to the ACS. To generate the database, we use the Bayes bootstrap to draw N records from the 2010–2014 files ACS using with 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?". There are $|\mathcal{Q}| = 318,003$ such queries. The custodian operates the MWEM mechanism to publish statistics from this database.

6.2.2 Measuring Preferences

To measure the marginal rate of substitution in the social planner's problem, we draw on data from the Federal Statistical System Public Opinion Survey (FSS POS). Our goal is to empirically quantify the distribution of the indirect utility function parameters.[17]

The FSS POS is a national public opinion survey conducted in conjunction with Gallup daily tracking surveys. 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 peoples privacy, or generally respect peoples privacy?"
- `FS7`, which records responses on a five-category Likert 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:[18]

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

[18] For many respondents, income is missing, and the data exhibit moderate levels of non-response 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 equivalent if we instead drop the missing cases.

- 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 $\text{Corr}\left[\eta_i, \ln y_i\right] = 0.040\,(\pm 0.003)$

To compute the MRS based on Equation (23), and using the estimated correlations 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 quality on the same scale as variation in utility that arises from the interaction with relative income.[19]

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.

6.2.3 Solution

Figure 1 illustrates the solution to the social planner's problem when the statistical agency operates the MWEM algorithm, as operationalized by Hardt et al. (2012).

[19] 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.

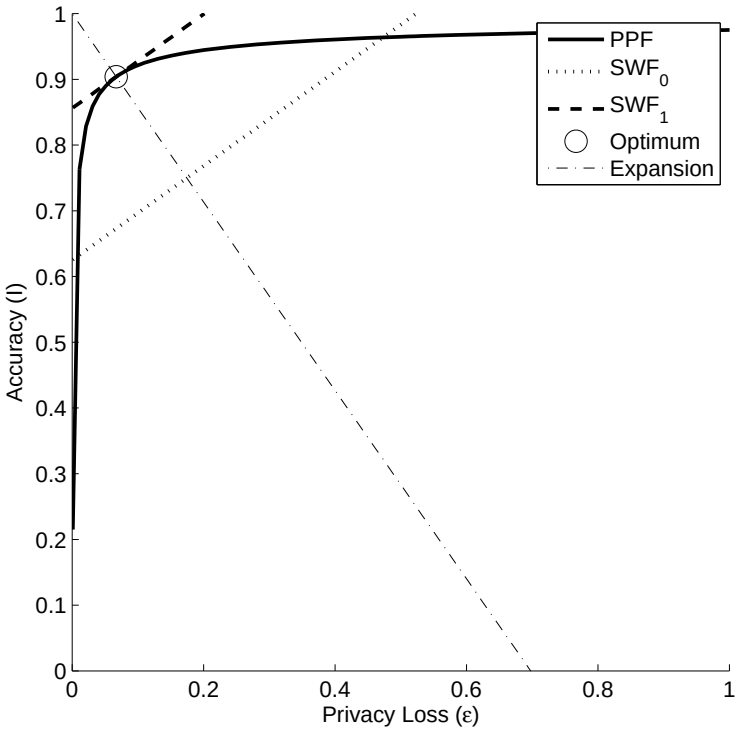

Figure 1: Solution to the Social Planner's Problem

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

Evaluated at the point where the $MRT = 1.040$, 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 5.2, 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 could be offset, on average, with an income transfer of approximately 1.3 percent of national income distributed evenly across the population.

6.2.4 Simulations

One may be curious about the practicality of such methods. 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 the 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. This indicates 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 experiment above, the average worst-case error doubles to 0.002.

6.3 Example 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, only we assume individuals care about their relative health status rather than relative income. This yields 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 quality.

6.3.1 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 of 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 NHIS as a random sample. To generate database, we draw N BMI observations from the 2015 NHIS using the Bayes bootstrap with 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?". There are $|\mathcal{Q}| = 320,400$ such queries. The custodian operates the MWEM mechanism to publish statistics from this database.

6.3.2 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 use 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.[20] We use the following questions from the CNSS:

- `JAq6`, "In general, how would you rate your overall health?" measured as a Likert scale with five categories;
- "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?"

[20] 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.

measured as a Likert scale with five categories (proxy for privacy preferences, γ).

2. `JAq4@a`, "...affect the quality of medical care?" measured as a Likert scale with five categories (proxy for accuracy preferences, η).

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

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

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 quality. The estimated shadow price of reduced privacy loss is, therefore, lower in the context of health data, lower than in the context of publishing income statistics. That is, people are more willing to forgo privacy for increased accuracy in the context of health information than in the context of income.

6.3.3 Solution

Evaluated at the point where the $MRT = 0.94$, find the optimal accuracy and privacy are $I^0 = 0.872$ and $\varepsilon^0 = 0.0451$ Once again, we evaluate the welfare cost of choosing suboptimally low privacy loss at the expense of data accuracy. Choosing

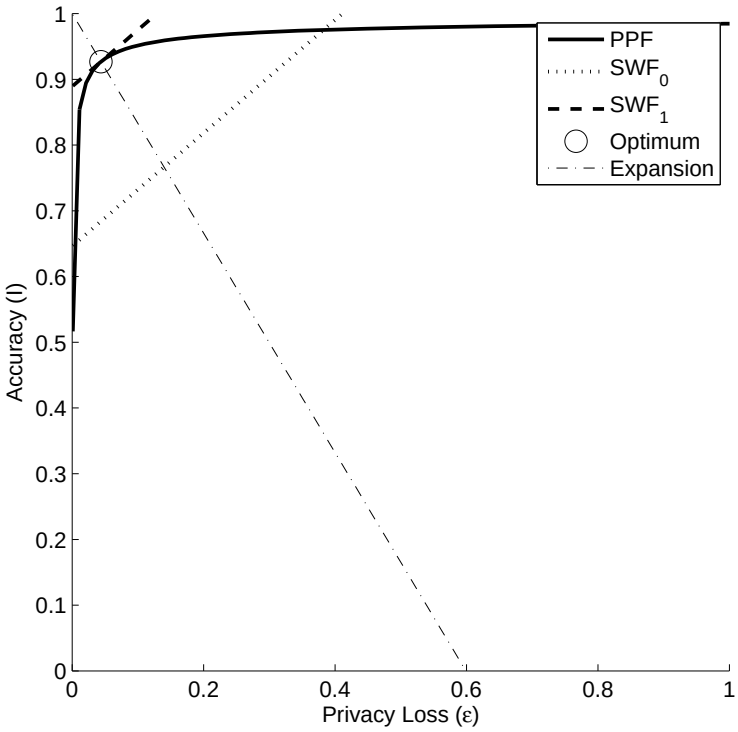

Figure 2: The Social Planner's Problem: Health Statistics

a point with $\varepsilon = 0.226$, which is equivalent to a 50 percent decrease in privacy loss, the corresponding value of I on the PPF, $I = 839$, results in an expected change in utility of -0.013 per person. This could be offset, with an income transfer of approximately 1.3 percent of national income distributed evenly across the population.

6.3.4 Simulations

Using the population data from 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 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.

6.4 Discussion

Our analysis suggests how data providers can combine information about their publication technology with data on the value of privacy and data quality 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 more and better information on individual and social preferences for privacy and for data quality. 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.

Instead, to illustrate a problematic consequence of this technology, we use results on the composability of $(\varepsilon, 0)$-differential privacy to reason about expanding the set of published queries. Intuitively, differential privacy loss is additively composable across independent data releases. [21]

[21] This analysis can be conducted for the more general (ε, δ)-differential privacy. To address composability formally we would need to define the concept of $k-$fold adaptive composition, with appropriate parameterization to illustrate the consequences of composability for (ε, δ)-differential privacy. These details are not necessary here, as we can use the more straightforward composability results for $(\varepsilon, 0)$-differential privacy to make our point.

If the statistical agency wished to publish both the income distribution data with accuracy $I_y^0 = 0.904$ and the health status statistics with accuracy $I_h^0 = 0.927$, which are the two optimal values derived above, then the level of privacy protection would be the sum of $\varepsilon_y^0 = 0.067$ (income distribution) and $\varepsilon_h^0 = 0.043$ (health statistics). We have added the subscripts y and h to distinguish the solutions to the two problems. By the composability of $(\varepsilon, 0)$-differential privacy, the actual privacy protection afforded by this publication strategy is $\varepsilon_{yh} = 0.11$. There is no proof in our work (or anywhere else that we know) that the combination $I_y^0 = 0.904$ and $I_h^0 = 0.927$ with $\varepsilon_{yh} = 0.110$ 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.

7 Conclusion

This paper provides the first comprehensive synthesis of the economics of privacy with the statistical disclosure limitation and privacy-preserving data analysis literatures. We develop a complete model of the technology associated with data publication constrained by privacy protection. Both the quality of the published data and the level of the 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 the most recent technology for (ε, δ)-differential privacy with (α, β)-accuracy. The statistical agency publishes using a Multiplicative Weights Exponential Mechanism query release system.

Consumers demand the statistics supplied by the government agency because of their interdependent preferences. They want to know where they fit in the income distribution and the distribution of relative health status. Thus, they are better off when they have more accurate estimates of those distributions, which can only be provided by inducing citizens to allow their data to be used in statistical tabulations. All consumers/citizens are provided (ε, δ)-differential privacy with the same values of the parameters due to worst-case protection afforded by this publication technology. All consumers/citizens use the same (α, β)-accurate statistical tabulations to assess their utility.

The solution to the social planning problem that optimally provides both public goods–data accuracy and privacy protection–delivers more data accuracy, but less privacy protection, than the VCG mechanism for private-provision of data. The reason is that the VCG mechanism for procuring data-use rights ignores the public-good nature of the statistics that are published after a citizen sells the right to use her private data in those publications. The VCG mechanism also does not account for the public good provided by the differential privacy protection, which is extended to the entire population even if only some citizens would have sold their data-use rights to the agency. The full social planner's problem compels all consumers to allow their data to be used in the published tabulations but guarantees privacy protect by restricting all publications to be based on the output of an efficient query release mechanism–one that produces maximally accurate statistics with the socially optimal differential privacy protection.

We compute the welfare loss associated with suboptimally providing too much privacy protection and too little accuracy. For the income distribution statistics, which are demanded when individuals care about their income relative to the population distribution, decreasing accuracy by three log points (3 percent) rel-

ative to the social optimum and commensurately increasing privacy protection decreases total utility by 0.008 log points. For the relative health statistics, the welfare loss from the same experiment is comparable. Both calculations show that over-provision of privacy protection is harmful to the citizens when the demand for the statistical products of the agencies is derived from interdependent preferences.

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 disclosure 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 relatively new 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 this is true, 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 privacy budget will not allow any 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

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.

Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining, *Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00* **29**(2): 439–450.

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

Alessie, R. and Kapteyn, A. (1991). Habit formation, interdependent preferences and demographic effects in the almost ideal demand system, *The Economic Journal* **101**(406): 404–419.

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.

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.

Denning, D. (1980). Secure statistical databases with random sample queries, *ACM Transactions on Database Systems* **5**(3): 291–315.

Duncan, G., Fienberg, S., Krishnan, R., Padman, R. and Roehrig, S. (2001). Disclosure limitation methods and information loss for tabular data, *in* P. Doyle, J. Lane, J. Theeuwes and L. Zayatz (eds), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier, pp. 135–166.

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

Duncan, G. T., Elliot, M. and Salazar-González, J.-J. (2011). *Statistical confidentiality principles and practice*, Statistics for Social and Behavioral Sciences, Springer New York.

Duncan, G. T. and Fienberg, S. E. (1999). Obtaining information while preserving privacy: a markov perturbation method for tabular data, *Statistical Data Protection (SDP '98)*, Eurostat, pp. 351–362.

Dwork, C. (2006). Differential privacy, *Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP)*, pp. 1–12.

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. and Naor, M. (2008). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy, *Journal of Privacy and Confidentiality* (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.

Evfimievski, A., Gehrke, J. and Srikant, R. (2003). Limiting privacy breaches in

privacy preserving data mining, *ACM SIGMOD Principles of Database Systems (PODS)*, pp. 211–222.

Federal Committee on Statistical Methodology (2005). Report on statistical disclosure limitation methodology, *Technical report*, Statistical and Science Policy, Office of Information and Regulatory Affairs, Office of Management and Budget.

Fellegi, I. P. (1972). On the question of statistical confidentiality, *Journal of the American Statistical Association* **67**(337): 7–18.

Futagami, K. and Shibata, A. (1998). Keeping one step ahead of the Joneses: status, the distribution of wealth, and long run growth, *Journal of Economic Behavior and Organization* **36**(1): 109–126.

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.

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., Gehrke, J. and Venkitasubramaniam, M. (2007). L-diversity: privacy beyond k-anonymity, *ACM Transactions on Knowledge Discovery from Data* **1**(1).

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

McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy, *48th Annual IEEE Symposium on Foundations of Computer Science 2007 (FOCS '07)*, pp. 94–103.

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

Posner, R. A. (1981). The economics of privacy, *The American economic review* pp. 405–409.

Postlewaite, A. (1998). The social basis of interdependent preferences, *European Economic Review* **42**(3-5): 779–800.

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

Schmutte, I. M. (2016). Differentially private publication of data on wages 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.

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* **10**(5): 571–588.

U.S. Code (1954). USC: title 13 - Census.

U.S. Code (2002). Confidential Information Protection and Statistical Efficency Act.

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 accuracy α, corresponding to data quality level $I = (1-\alpha)$, the private producer must procure 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

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

Comparison of the preceding marginal cost expressions establishes that $0 < \frac{dC^{L}(I)}{dI} \leq \frac{dC^{VCG}(I)}{dI}$ for all I, since $N\int_0^{Q\left(\frac{H(I)}{N}\right)} F_{\gamma}(\gamma)d\gamma > 0$, $H'(I) > 0$, and $Q'\left(\right) > 0$ The result in the theorem follows by using the private equilibrium equation for the market price 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 $\frac{d^2 C^{VCG}(I)}{dI^2} \geq 0$. ■

A.2 Translation of the Ghosh-Roth Model in Section 4 to Our Notation

In this appendix we show that the results in our Section 4, 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

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

which means that for individual i, the level of differential privacy afforded by the algorithm $A\left(D\right)$ 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 $A\left(D\right)$ 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 $v_i \varepsilon_i$, where $v_i > 0$, where $p \equiv (p_1, \ldots, p_n) \in \mathbb{R}_+^n$ and $v \equiv$

$(v_1, \ldots, v_n) \in \mathbb{R}^n_+$.

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 $A(D)$ that is $\varepsilon_i(v)$-minimally differentially private to deliver a query response $\hat{s} \in \mathbb{R}$ and a vector of payments $p(v) \in \mathbb{R}^n_+$. 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 $p_i(v) - v_i\varepsilon_i(v) \geq 0$ and $p_i(v) - v_i\varepsilon_i(v) \geq p_i\left(v^{\tilde{}i}, v'_i\right) - v_i\varepsilon_i\left(v^{\tilde{}i}, v'_i\right)$ for all $v'_i \in \mathbb{R}^n_+$, where $v^{\tilde{}i}$ is the vector v with element v_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 $A(D)$ 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 $\left(\frac{1}{2} + \ln 3\right)\alpha n$-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 ap-

plication of GR-Definition 2.3 ($\text{Lap}\,(\sigma)$ 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 $H = \{1, \ldots, |H|\}$. Assume, as we do in Theorem 1 and as GR do in their proof of GR-Theorem 3.3, that n is sufficiently large that we can ignore the difference between $(1-\alpha)\,n$ and $\text{ceil}\,((1-\alpha)\,n)$. The resulting answer from the query response mechanism is

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

which is 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 the algorithm produced. 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 modification of the algorithm results in a lower minimum ε_i, everyone who opts into the new algorithm 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 v_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 5

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 $\mathrm{E}\left[\ln y_i\right] = \frac{1}{N}\sum_{i=1}^{N} y_i$. 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 $\lambda > 0$

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

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 nondecreasing in p.

For all $\lambda > 1$

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

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), consider (p, y) and (p', y') such that $v_i\left(y_i, \varepsilon, I, y^{\tilde{}i}, p\right) \leq \bar{v}$ and $v_i\left(y_i', \varepsilon, I, y^{\tilde{}i\prime}, p'\right) \leq \bar{v}$ for all i. For any $\lambda \in [0, 1]$ let $(p'', y'') = \lambda\left(p, y\right) + \left(1 - \lambda\right)\left(p', y'\right)$. Then,

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

by the concavity of $\ln\left(x\right)$.

Continuity in (p, y) follows from the continuity of $\ln\left(x\right)$. 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 Hardt, Ligett and McSherry (2012), henceforth HLM.

To maintain consistency with the presentation in Sections 3 and 5, 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; a list of allowable normalized linear queries $\mathcal{Q} \subseteq \mathcal{F}$ 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 d_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 $(\varepsilon, 0)$-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 $(\varepsilon, 0)$-differential privacy. HLM state an error bound for MWEM in their Theorem 2.2.Their reported bound for unnormalized queries is $2N\sqrt{\frac{\log|\chi|}{T}} + \frac{10T\log|\mathcal{Q}|}{\varepsilon}$. We simply rescale the error bound by database size to account for the normalization. Converting to normalized queries gives $2\sqrt{\frac{\log|\chi|}{T}} + \frac{10T\log|\mathcal{Q}|}{N\varepsilon}$. 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 $T = \left(\frac{\varepsilon N \sqrt{\log |\chi|}}{10 \log |\mathcal{Q}|}\right)^{2/3}$.

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 our code archive, which is permanently archived in the Digital Commons space of the Cornell Labor Dynamics Institute http://digitalcommons.ilr.cornell.edu/ldi/22/.

A.5 Data Sources

We use raw data from the the Cornell National Social Survey(CNSS). The input data files sources are:

- American Community Survey (ACS)
- National Center for Health Statistics (NCHS)
- Federal Statistical System Public Opinion Survey
- Cornell National Social Survey: obtained from the CNSS integrated data application http://www.ciser.cornell.edu/beta/cnss/ by selecting all variables for all years. The original variable names include the "@" symbol, which is not recognized in Stata. The analysis is conducted on an edited version of the file also available in the public archive of this paper.

A complete archive of the data and programs used to produce the empirical results in this paper is available in the Digital Commons space of the Cornell Labor Dynamics Institute http://digitalcommons.ilr.cornell.edu/ldi/22/.