2015

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

John M. Abowd
Cornell University, jma7@cornell.edu

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

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

Abstract

This paper has been replaced with http://digitalcommons.ilr.cornell.edu/ldi/37.

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 and a Lindahl 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 Private Multiplicative Weights query release mechanism 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 General Social Survey and the Cornell National Social Survey indicate that the welfare losses from under-providing data accuracy while over-providing privacy protection can be substantial.

Keywords

Economics, Privacy, Population Statistics, Confidentiality, Public Goods, Abowd, Schmutte

Comments

This paper has been replaced with http://digitalcommons.ilr.cornell.edu/ldi/37.

Abowd acknowledges direct support from the U.S. Census Bureau and from NSF Grants BCS-0941226, TC-1012593 and SES-1131848.

Alternate location for replication archive: https://doi.org/10.5281/zenodo.290231

REVISITING THE ECONOMICS OF PRIVACY: POPULATION STATISTICS AND CONFIDENTIALITY PROTECTION AS PUBLIC GOODS

John M. Abowd
Department of Economics
Labor Dynamics Institute
Cornell University
john.abowd@cornell.edu

Ian M. Schmutte
Department of Economics
Terry College of Business
University of Georgia
schmutte@uga.edu

January 29, 2015

Abowd acknowledges direct support from the U.S. Census Bureau and from NSF Grants BCS-0941226, TC-1012593 and SES-1131848. 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. This paper was written while the first author was Distinguished Senior Research Fellow at the Census Bureau. 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. We acknowledge helpful comments from Ashwin Machanavajjhala, Bruce Spencer, Lars Vilhuber and Nellie Zhao. 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 Vickery-Clark-Groves and Lindahl mechanisms. 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 Private Multiplicative Weights query release mechanism 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 General Social Survey and the Cornell National Social 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

JEL classification: C40, C81, H41

1 Introduction

Like so many other ideas in information economics, George Stigler (1980) began the analysis of the economics of privacy, taking off as Posner (1981) noted 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 question of 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.

In its current form, the economics of privacy focuses on the economic value of information about the habits of consumers that are known to the curators of databases produced by intermediating commercial transactions on the Internet. 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.

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 un-

derstanding 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. Data-rights acquisition via a Lindahl mechanism improves upon the VCG auction but still does not attain the socially optimal data release and privacy protection. 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 both the Lindahl and VCG levels and the socially optimal privacy losses are greater than those generated by private data suppliers using either the Lindahl or VCG mechanisms.

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 (Manski 1993; 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 gen-

erate 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 data 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*. 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 predictive distribution of the confidential data given the released statistical summaries. This formalization is at the heart of the modern data privacy literature that emerged in computer science.

1.2 The Emergence of the Differential Privacy Paradigm

Cryptographers know how to protect secrets. The formal requirement for establishing the strength of an encryption algorithm is to publish the entire algorithm, including all of its input data requirements and their values, but not including the password or encryption key. Then, let other cryptographers try to break the code.

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 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 full encryption of the confidential data, and therefore useless for data publication. She proposed developing a privacy protection method that "captures the increased risk to one's privacy incurred by participat-

ing 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 economic 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 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 economic 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.[1]

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

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

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 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.

Publication of the results of queries from private databases or search logs has generated a series of papers in the electronic commerce literature. McSherry and Talwar (2007) focused on the question of how a trusted custodian of private data can efficiently compensate the data owners to permit the use of their data by an analyst seeking the answer to an additional collection of queries.

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 with the least-cost acquisition of data-use rights (privacy loss).

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 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. Al-

though 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. Privacy-preserving publication by statistical agencies treats all citizens as equally protected under the relevant confidentiality laws. Our paper is 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 either the VCG or Lindahl 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 from the General Social Survey and the Cornell National Social Survey to quantify the parameters of our solution to 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 as-

signing an identity or a sensitive attribute to a particular respondent, given the newly released data, to the prior odds, given all previously released data. Differential 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.

To foreshadow what we develop below, we provide a concise summary of the antecedents to our work here. We adopt the differential privacy definitions in Blum et al. (2008) and Hardt and Rothblum (2010) explicitly, noting that these are also used by Hardt et al. (2010), Gupta et al. (2012) and Dwork and Roth (2014). Then, we consider normalized linear queries of the histogram representations of databases with N entries. N is always the population under study, not a sample. The global sensitivity of all such queries is $\frac{1}{N}$. We use the (α, β)-accurate query mechanism from Definition III.1 and the (ε, δ)-differential privacy Definition III.2 from Hardt and Rothblum as adapted by Gupta et al. (2012) and Dwork and Roth (2014). We consider only the maximally accurate allowable query; hence, our analysis sets their query round, k, at the allowable limit for the Private Multiplicative Weight (PWM) mechanism in their Figure 1, which is their parameter c. We implement the enhancements to PMW found in Gupta et al. (2012)–in particular, tighter accuracy bounds and an explicit limit on the number of query rounds required to reach the target accuracy for given privacy loss parameters. Readers familiar with the data privacy literature in computer science now have all the necessary information to put our contribution in context. For the benefit of economists and statisticians, the development below is self-contained with a road map to relate

our proofs to the relevant papers in the differential privacy literature.

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 all c queries with (α, β)-accuracy from the PWM 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: c, β, and δ 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.[2] 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

[2]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 $\chi = \{0, 1\} \times \{1, \ldots, 6\}$.

database D of type $i \in \chi$. We use the ℓ_1 norm:

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

Observe that $||x||_1 = N$, the number of records in the database. Given two histograms, x and y, $||x - y||_1$ measures the number of records that differ between x and y. We define *adjacent histograms* as those for which the ℓ_1 distance is at most 1.[3]

3.1.2 Queries

A *linear query* is a mapping $f : [-1, 1]^{|\chi|} \times \mathbb{Z}^{*|\chi|} \to \mathbb{Z}^*$ 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 $f : [-1, 1]^{|\chi|} \times \mathbb{Z}^{*|\chi|} \to [0, 1]$ 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

[3] 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.

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 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

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

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

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

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

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

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 innovative model of Ghosh and Roth (2011). Ghosh and Roth (GR, hereafter) show that differential privacy can be priced as a commodity using a formal auction model. They prove the existence of a mechanism that yields the lowest-cost method for

answering a database query with $(\varepsilon, 0)$-differential privacy and (α, β)-accuracy.[4]

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

[4]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 and the Lindahl mechanism (Mas-Colell et al. 1995) for the procurement of privacy protection by a profit-maximizing data curator acting as a price-discriminating monopsonist when buying data-use rights.

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 consumers use the published data for free.

The VCG mechanism implies a single price for each data-use right purchased. In the Lindahl mechanism, a single private data provider can perfectly price dis-

criminate when procuring data-use rights with their attendant privacy loss. As long as property rights over privacy exposure are sufficiently clear, the Lindahl private producer can 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 both cases data quality is under-produced compared to the social optimum and privacy protection is over-produced; i.e., there is too little privacy loss.

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.[5] 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. The preferences of consumer i are given by

$$v_i\left(y_i, \alpha, \varepsilon_i\right) = \ln y_i + p_\varepsilon \varepsilon_i - \gamma_i \varepsilon_i + \eta_i(1-\alpha) - p(1-\alpha). \tag{3}$$

[5]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.

Equation (3) implies that preferences are quasilinear in data quality, $1-\alpha$, privacy loss, ε_i, and log income, $\ln y_i$.[6] The term $p_\varepsilon \varepsilon_i$ represents the total payment an individual receives if her bit is used in an $(\varepsilon_i, 0)$-differentially private mechanism. p_ε is the common price per unit of privacy, to be determined by the model. The individual's marginal preferences for data accuracy (a "good") and privacy loss (a "bad"), $(\gamma_i, \eta_i) > 0$, 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. Individuals each consume one unit of the published statistic, which has information quality, I, defined in terms of (α, β)-accuracy, $I = (1-\alpha)$. The price of data quality, p, is also determined by the model.

We do not include any 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 these topics for future

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

work.

In what follows we present the GR results using our notation and definitions. See Appendix A.1 for a complete summary of the translation from their notation and definitions to ours.

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.1

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. Then, profits $P(I)$ are

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

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

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

where the solution is evaluated at I^{VCG}.[7] 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

[7] 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.

uniqueness of a solution is guaranteed.

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

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

$$\max_{I_i \geq 0} \eta_i \left(I_{-i} - I_i \right) - pI_i \tag{10}$$

where I_{-i} is the amount of data quality provided in the absence of any monetary payment from i. Consumer i's willingness to pay is non-negative if, and only if, $\eta_i \geq p$; that is, if the marginal utility from increasing I exceeds the price. If there exists at least one consumer for whom $\eta_i \geq p$, then the solution to equation (9) is attained by $I^{VCG} > 0$. We next show that there is only one such 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

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

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

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

Marginal cost is positive $\frac{dC^{VCG}(I^o)}{dI} > 0$. Therefore, data quality is under-provided by a competitive supplier when data quality is a public good. More succinctly, $I^{VCG} < I^0$. Therefore, privacy protection must be over-provided by equation (8).

4.5 The Price-discriminating Monopsonist Provider of Data Quality

Now consider the problem of a single private data provider who produces $\hat{s}$ with accuracy $(\alpha, 1/3)$ using the same technology as in equations (7) and (8). We now allow the producer to price-discriminate in the acquisition of data-use rights–that is, the private data-quality supplier is a price discriminating monopsonist in the market for data-use rights. This relaxes the assumptions of the GR VCG mechanism to allow for the unrealistic possibility that the data quality provider knows the population values of γ_i. GR acknowledge this theoretical possibility when discussing the individual rationality and dominant-strategy truthful requirements of their mechanism. They reject it as unrealistic, and we agree. We are considering this possibility to show that even when the private data-quality provider is allowed to acquire data-use rights with a lower cost strategy than the VCG mechanism, data quality will still be under-provided.

The producer must decide how many data-use rights (and associated privacy loss ε, the same value for all i) to purchase from each member of the database,

or, equivalently, how much to charge members of the database to opt out of participation in the mechanism for computing the statistic. (They cannot opt out of the database.) Let $\pi \in \{0,1\}^N$ be the participation vector. Using the Lindahl approach, let $p^L_{\varepsilon_i}$ be the price that satisfies, for each consumer i,

$$p^L_{\varepsilon_i} \leq \gamma_i, \text{with equality if } \pi_i = 1. \tag{12}$$

Equation (12) says that the Lindahl prices are those such that the choice of ε is exactly the value that each individual would optimally choose on her own. Even with our assumption of linear preferences, the Lindahl prices are unique for every consumer who participates in the mechanism for computing the statistic.

Given a target data quality of $I = (1-\alpha)$, the producer's cost minimization problem is the linear program

$$C^L(I) = \min_{\pi} \left(\sum_{i=1}^{N} \pi_i p^L_{\varepsilon_i} \right) \varepsilon \tag{13}$$

subject to

$$\sum_{i=1}^{N} \pi_i = N - \frac{(1-I)N}{1/2 + \ln 3} \text{ and } \varepsilon = \frac{1/2 + \ln 3}{(1-I)N}.$$

The solution is for the producer to set $\pi_i = 1$ for the H members of the database with the smallest $p^L_{\varepsilon_i}$ and $\pi_i = 0$, otherwise. Note that if

$$\frac{dC^L(I)}{dI} \leq \frac{dC^{VCG}(I)}{dI}$$

for all I, which will be proven in Theorem 1, then the Lindahl purchaser of data-

use rights will produce more data quality at any given price of data quality than the VCG purchaser.

By construction, the Lindahl solution satisfies the Pareto optimality criterion for data-use rights acquisition that

$$\sum_{i=1}^{N} \pi_i p_{\varepsilon_i}^{L} = \sum_{i=1}^{N} \pi_i \gamma_i. \tag{14}$$

Once again, the supplier implements the query response mechanism of equation (5) with $\left(\frac{1/2+\ln 3}{(1-I)N}, 0\right)$-differential privacy and $(1-I, \frac{1}{3})$-accuracy but pays each consumer differently for her data-use right. Notice that equation (14) describes the Pareto optimal privacy loss whether or not one acknowledges that the privacy protection afforded by ε is non-rival, only partially excludable, and, therefore, also a public good.

To implement the Lindahl solution, the data producer must be able to exclude the bits, b_i, of specific individuals when computing the statistic, and must have perfect knowledge of the every marginal disutility γ_i of increasing ε. When this information is not available, the producer can, and will, implement the first-best allocation by choosing a price through the VCG auction mechanism used by GR.

For readers familiar with the data privacy literature, we note that the statement that technology is given by equations (7) and (8) means that the data custodian allows the producer to purchase data-use rights with accompanying privacy loss of $\varepsilon = \frac{1/2+\ln 3}{(1-I)N}$ from $H(I)$ individuals for the sole purpose of computing $\hat{s}$ via the query response mechanism in equation (5) that is $\left(\frac{1/2+\ln 3}{(1-I)N}, 0\right)$-differentially

private and achieves $(1-I, \frac{1}{3})$-accuracy, which is exactly what Ghosh and Roth prove.

4.6 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 and (13) for the Lindahl 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 $I^{VCG} \leq I^L \leq I^0$, where I^0 is the Pareto optimal level of I solving equation (11), I^L is the privately-provided level when using the Lindahl mechanism to procure data-use rights and I^{VCG} is the privately-provided level when using the VCG procurement mechanism.

Proof. By construction, $F_\gamma(\gamma)$ is the distribution of Lindahl prices. 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

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

for $\ell = VCG, L$. Note that p_ε^ℓ is the disutility of privacy loss for the marginal participant in the VCG and Lindahl mechanisms, respectively. The total cost of producing $I = (1-\alpha)$ using the VCG mechanism is equation (7):

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

while the total cost of implementing the Lindahl mechanism is equation (13):

$$C^{L}(I) = \left(N \int_0^{Q\left(\frac{H(I)}{N}\right)} \gamma dF_\gamma(\gamma)\right) \varepsilon(I).$$

Using integration by parts and the properties of the quantile function,

$$\begin{aligned} C^L(I) &= \left[Q\left(\frac{H(I)}{N}\right) F_\gamma\left(Q\left(\frac{H(I)}{N}\right)\right) - \int_0^{Q\left(\frac{H(I)}{N}\right)} F_\gamma(\gamma)d\gamma\right] N\varepsilon(I) \\ &= \left[Q\left(\frac{H(I)}{N}\right) H(I) - N\int_0^{Q\left(\frac{H(I)}{N}\right)} F_\gamma(\gamma)d\gamma\right] \varepsilon(I). \end{aligned}$$

Differentiating with respect to I,

$$\frac{dC^L(I)}{dI} = \left[Q\left(\frac{H(I)}{N}\right) H(I) - N\int_0^{Q\left(\frac{H(I)}{N}\right)} F_\gamma(\gamma)d\gamma\right] \varepsilon'(I) + Q\left(\frac{H(I)}{N}\right) H'(I)\varepsilon(I).$$

The corresponding expression for $C^{VCG}(I)$ is

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

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 $Q'() > 0$ The results stated in the theorem follow by using the private equilibrium equation for the market price of I, which is p in equation (3),

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

Hence, $I^{VCG} \leq I^L \leq I^0$, where the final inequality follows from equation (11), $\sum_{i=1}^N \eta_i > \bar{\eta}$, and the conditions on Q that imply $\frac{d^2C^{VCG}(I)}{dI^2} > 0$ and $\frac{d^2C^L(I)}{dI^2} > 0$ for sufficiently large $\frac{H(I)}{N}$. ■

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 the optimal level of privacy loss and data quality when both are explicitly public goods. We derive the technological frontier for publication of statistical data provided by national statistical agencies, who are assumed to use a confidential database for which they are the trusted custodian. Once we know the technological frontier, the optimal publication strategy depends on the willingness to pay for increased accuracy with reduced privacy protection. We use Samuelson (1954) as explicated in Mas-Colell et al. (1995, p. 360-61) to solve for the Pareto optimal quantities of each public good.

5.1 Production Possibilities for Privacy Protection and Data Publication

The custodian releases public statistics derived from a confidential database using the responses to normalized linear queries submitted by a representative analyst. The custodian is constrained to operate a query release mechanism that is (ε, δ)-differentially private. The data custodian releases these query answers sequentially as they are submitted by the representative analyst. The analyst is therefore able to submit additional queries for information after observing the answers to all previous queries.[8]

5.1.1 The Private Multiplicative Weights Mechanism

Let the data custodian use the Private Multiplicative Weights (PMW) query response mechanism introduced by Hardt and Rothblum (2010). Our presentation follows the definitions in Gupta et al. (2012). The custodian chooses the total number of normalized linear queries $f_t \in \mathcal{F}$ to allow $(t = 1, \ldots, k)$ generating the set $\mathcal{Q} \subseteq \mathcal{F}^k$ from which all allowable query sequences must be drawn. The custodian also sets the privacy parameters (ε and δ) and the parameters governing query accuracy (α and β) that define the rate of privacy loss under the mechanism.

We summarize the basic features of the PMW algorithm. A more complete description appears in Appendix A.3, but is not necessary to understand our ap-

[8]The concept of a representative analyst is without loss of generality. In the context of public-use statistics provided by government agencies, any analyst is in possession of all output from the release mechanism. By considering a representative analyst, we model the worst-case scenario of a fully-informed analyst who seeks to compromise privacy through targeted queries.

plication. The PMW algorithm calculates the candidate (ε, δ)-differentially private query response using the confidential data plus Laplace noise with scale parameter that depends upon $\alpha, \varepsilon,$ and δ at each step. Next, it produces a synthetic approximation to the confidential database using only the (ε, δ)-differentially private responses already released. The analyst calculates answers to all queries based on the current synthetic version of the database. If any privacy budget remains, an additional query is posed up to the limit k, and the (ε, δ)-differentially private response to the new query, if released, is used to update the synthetic database. At each round of operation t, the query release mechanism holds a noisy approximation x^*_{t-1} of the true histogram x that is based upon the previously released (ε, δ)-differentially private responses to the queries $f_1, \ldots, f_{t-1}$. In response to the next query f_t, the algorithm computes an (ε, δ)-differentially private response using the confidential data x as its input; then, the algorithm compares that answer to the answer based on x^*_{t-1}, which must be differentially private as a consequence of the construction of x^*_{t-1} because (ε, δ)-differential privacy composes (for a proof see Dwork and Roth (2014, pp. 49-51)). If the answer to query f_t using the approximate histogram x^*_{t-1} is close enough to the differentially private answer using the true histogram, then the mechanism returns the answer using only the public approximate histogram (the synthetic database). If not, it returns the differentially private answer to f_t and updates the approximation to the true histogram to x^*_t. When the algorithm stops, the (ε, δ)-differentially private responses to all k queries can be calculated from the final synthetic database with (α, β)-accuracy.

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

We show that the PMW algorithm establishes a convex and decreasing relationship between privacy loss and accuracy. It can therefore provide the basis for a well-defined production possibilities frontier that characterizes feasible trade-offs between privacy loss and accuracy.

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 parameters $\varepsilon > 0$, $0 < \delta < 1$ and $0 < \beta < 1$, there exist query release mechanisms $M(x)$ including the PMW mechanism that can interactively answer sequences of queries $f_t \in \mathcal{Q}$ for $t = 1, \ldots, |\mathcal{Q}|$ that satisfy the following conditions:

1. Privacy: $M(x)$ satisfies (ε, δ)-differential privacy;

2. Accuracy: $M(x)$ satisfies (α, β)-accuracy, with

$$\alpha = \frac{K(\delta, \beta, |\chi|, |\mathcal{Q}|, N)}{\varepsilon^b} \tag{15}$$

where $b \leq \frac{1}{2}$ Furthermore, K is decreasing in N, δ, and β, and increasing in $|\chi|$ and $|\mathcal{Q}|$.

Proof. We begin with results from Gupta et al. (2012), who define two algorithms that satisfy the conditions of the theorem: the Median Mechanism (MM) and Private Multiplicative Weights (PMW) algorithms. In Gupta *et al.,* Theorems 3 and 4 prove the privacy and accuracy claims in our theorem up to the functional form for the accuracy bound. We now prove the functional form of the accuracy bound for the PMW mechanism is $\alpha = \frac{K(\delta,\beta,|\chi|,|\mathcal{Q}|,N)}{\varepsilon^b}$, with $b = 1/2$ and

$$K(\delta, \beta, |\chi|, |\mathcal{Q}|, N) = \frac{8\sqrt{3}(\ln|\chi|)^{1/4}\sqrt{\ln\left(\frac{4}{\delta}\right)\ln\left(\frac{|\mathcal{Q}|}{\beta}\right)}}{N^{1/2}}. \tag{16}$$

The discussion following Gupta *et al.* Theorem 4 shows that any $B(\alpha)$ iterative database construction (IDC) algorithm is (α, β)-accurate as long as

$$\alpha = \frac{96\sqrt{B(\alpha)}\ln(4/\delta)\ln(k/\beta)}{\varepsilon}. \tag{17}$$

Furthermore, PMW is a $B(\alpha) = 4\ln|\chi|/\alpha^2$ IDC algorithm. The result is established by substituting $4\ln|\chi|/\alpha^2$ for $B(\alpha)$ in equation (17), solving the equation for α, and dividing by N. The accuracy bound reported by Gupta et al. assumes the histogram and queries are unnormalized. Division by N addresses our application to normalized, rather than unnormalized, linear queries. Finally, consistent with Gupta *et al.*, we set $k = |\mathcal{Q}|$, the cardinality of the set of allowable queries. ■

5.1.3 The Production Possibilities Frontier

The production possibilities frontier (PPF) relating information quality $I = (1-\alpha)$ and differential privacy loss ε is defined by a transformation function

$$G(\varepsilon, I) \equiv I - \left[1 - \frac{K(\delta, \beta, |\chi|, |\mathcal{Q}|, N)}{\varepsilon^b}\right] \tag{18}$$

where Theorem 2 gives the functional form. All feasible pairs (ε, I) are contained in the transformation set

$$Y = \{(\varepsilon, I) \,|\varepsilon > 0, 0 < I < 1 \text{ s.t. } G(\varepsilon, I) \leq 0\}. \tag{19}$$

The PPF is the boundary of the transformation function defined as

$$PPF(\varepsilon, I) = \{(\varepsilon, I) \,|\varepsilon > 0, 0 < I < 1 \text{ s.t. } G(\varepsilon, I) = 0\}. \tag{20}$$

Equation (20) specifies the maximum data quality that can be published for a given value of privacy loss.

Using the result of Theorem 2, and solving for I as a function of ε, the data publication problem using the PMW query release mechanism produces the production possibilities frontier

$$I(\varepsilon; \delta, \beta, |\chi|, |\mathcal{Q}|, N) = \left[1 - \frac{K(\delta, \beta, |\chi|, |\mathcal{Q}|, N)}{\varepsilon^b}\right]. \tag{21}$$

The PPF described by equation (21) can be graphed with ε on the horizontal axis

and I on the vertical axis. Figure 1 shows the PPF. Notice that it is always in the positive quadrant; however, because ε is a bad rather than a good, the PPF has properties similar to the efficient risk-return frontier used in financial economics. The PPF separates feasible (ε, I) pairs, which are on and below the PPF, from infeasible pairs, which are above the PPF. The PPF is concave from below: a chord connecting any two points on the PPF is entirely in the set defined by equation (19). The PPF asymptotically approaches the origin. The marginal social cost of increasing data accuracy I in terms of foregone privacy protection ε–the marginal rate of transformation–is

$$MRT\left(\varepsilon, I\right) \equiv \frac{\partial G/\partial \varepsilon}{\partial G/\partial I} = \frac{bK(\delta, \beta, |\chi|, |\mathcal{Q}|, N)}{\varepsilon^{b+1}} = \frac{dI}{d\varepsilon}, \tag{22}$$

where the marginal rate of transformation is the slope of the PPF, and not minus the slope, because privacy loss is a public bad.

We take the parameters $(\delta, \beta, |\chi|, |\mathcal{Q}|, N)$ that determine K to be outside the scope of the choice problem for this paper. Doing so is not without consequence, as these parameters also affect the PPF. We note that the data provider can, in principle, change the set of allowable queries $\mathcal{Q}$ to modify the PPF. We are treating the resource costs associated with $(\delta, \beta, |\chi|, |\mathcal{Q}|, N)$ as fixed, therefore, in order to focus on the trade-off between privacy loss and data accuracy as summarized byε and I. Note, however, that the population size N is fixed.

Before we continue, we interpret equation (21) using the concepts from differential privacy and statistical disclosure limitation. Equation (21) is a proper production possibilities frontier because it is based on the best available algo-

rithmic bound on the accuracy parameters (α, β), and hence on I, that delivers (ε, δ)-differential privacy. When we implement this PPF using the PMW algorithm, we achieve this bound. Hence, according to the definition in equation (20), our PPF delivers the privacy loss, data accuracy pairs that are on the boundary of the transformation function as described by the best available technology that can deliver at least (ε, δ)-differential privacy. If a newer algorithm achieves a better accuracy bound for the same privacy loss, then equation (21) can be modified to reflect that newer technology. In terms of statistical disclosure limitation, equation (21) is the $R - U$ confidentiality map. Any data publication protected using ε as the disclosure risk measure that does not deliver data quality on the boundary of the transformation set, *i.e.*, on equation (21) is inefficient. Any data publication strategy that claims to produce an accuracy, privacy loss pair above the PPF is infeasible–it must involve some violation of the assumptions of Theorem 2 or a computational error.

5.2 Preferences

Define the indirect utility function, v_i, for each individual as

$$v_i\left(y_i, \varepsilon, I, y^{\tilde{}i}, p\right) = \max_q u_i\left(q, \varepsilon, I, y^{\tilde{}i}\right) \text{ s.t. } q^T p \leq y_i \tag{23}$$

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 $u_i\left(q, I, \varepsilon, y^{\tilde{}i}\right)$, also depends upon the privacy-loss public bad, ε, the data-accuracy public good,

I, and on the vector of all other incomes in the population $y^{\tilde{}i}$, which we discuss in more detail below.

5.3 The Social Planner's Problem

We adopt the linear aggregation form of the social welfare function

$$SWF\left(\varepsilon, I, v, y, p\right) = \sum_{i=1}^{N} v_i\left(y_i, \varepsilon, I, y^{\tilde{}i}, p\right) \tag{24}$$

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

$$\max_{\varepsilon, I} SWF\left(\varepsilon, I, v, y, p\right) \tag{25}$$

subject to equation (21). The PPF is clearly differentiable. Assuming the indirect utility functions are differentiable, the optimality conditions are

$$\frac{\frac{\partial G\left(\varepsilon^0, I^0\right)}{\partial \varepsilon}}{\frac{\partial G(\varepsilon^0, I^0)}{\partial I}} = \frac{\frac{\partial}{\partial \varepsilon}\sum_{i=1}^{N} v_i\left(y_i, \varepsilon^0, I^0, y^{\tilde{}i}, p\right)}{\frac{\partial}{\partial I}\sum_{i=1}^{N} v_i\left(y_i, \varepsilon^0, I^0, y^{\tilde{}i}, p\right)} \tag{26}$$

and $PPF\left(\varepsilon^0, I^0\right)$. The left-hand side of equation (26) is the marginal rate of transformation from the production possibilities frontier while the right-hand side is the marginal rate of substitution from the social welfare function.

5.3.1 Solving the Planner's Problem

We now specialize the indirect utility function to reflect individual preferences for privacy loss and data accuracy. We seed a functional form that directly motivates the consumer's demand for data accuracy. Consequently, we focus on indirect utility functions in which each individual cares about her place in the income distribution, a fact that she cannot know without statistical data on that distribution. The accuracy of these statistical data increases utility. The privacy loss from these statistical data reduces utility. If accurate data could be acquired at no privacy cost, individual utility would be an increasing function of relative income. It is worth noting that other models of interdependent preferences, such as Pollak (1976) and Alessie and Kapteyn (1991), also produce functional forms that require population statistics, which they are implicitly assuming can be acquired with full accuracy and no privacy loss.

We also want the consumers to display heterogeneity in their marginal tastes for privacy loss and data accuracy so that we can make direct comparisons to the problem we addressed in Section 4. An indirect utility function that captures the required features is

$$\begin{aligned} v\left(y_i, \varepsilon, I, \tilde{y^i}, p\right) &= -\sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln y_i \\ &\quad -\gamma_i \left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right) \varepsilon \\ &\quad +\eta_i \left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right) I \end{aligned} \tag{27}$$

where $(\gamma_i, \eta_i) > 0$ for all $i = 1, \ldots, N$, $\xi_\ell > 0$ for all $\ell = 1, \ldots, L$ and $\sum_{\ell=1}^{L} \xi_\ell = 1$. In

equation (27) 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. The term $(\ln y_i - \mathrm{E}[\ln y_i])$ represents the deviation of an individual's log income from the population mean. The indirect utility function defined by equation (27) thus directly reflects the assumption that individuals get utility from their relative position in the income distribution, as well as directly from their own income.

It is easy to see the antecedents of equation (27) in the specification used by Akerlof (1997), who assumed that utility depended on the levels of individual and average income instead of logarithms, and the subsequent work on the provision of public goods with interdependent preferences that his models sparked (see Aronsson and Johansson-Stenman (2008) and the references therein). We also note that these authors implicitly assume that the public statistics that enter their utility functions are measured with perfect accuracy and without privacy loss. We avoid the Manski (1993) reflection problem by using explicit empirical analogues of γ_i and η_i, and thus different covariations with income, in our empirical Section 6.

In Appendix A.2, 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 (27) is quasiconcave in (ε, I), which is trivially true for equation (27), as long as $(\gamma_i, \eta_i) >$

0 for all i, since it is linear in (ε, I). Hence, equation (24) is a well-specified social welfare function, quasiconcave in (ε, I), and the social planner's problem is well-specified since equation (21) is quasiconcave in (ε, I).

Substitution of equation (27) into equation (26) yields

$$\begin{aligned} \frac{\frac{\partial G\left(\varepsilon^0, I^0\right)}{\partial \varepsilon}}{\frac{\partial G(\varepsilon^0, I^0)}{\partial I}} &= \frac{\frac{\partial}{\partial \varepsilon} \sum_{i=1}^{N} v_i \left(y_i, \varepsilon^0, I^0, \tilde{y^i}, p\right)}{\frac{\partial}{\partial I} \sum_{i=1}^{N} v_i \left(y_i, \varepsilon^0, I^0, \tilde{y^i}, p\right)} \\ \frac{bK(\delta, \beta, |\chi|, |\mathcal{Q}|, N)}{(\varepsilon^0)^{b+1}} &= \frac{\sum_{i=1}^{N} \gamma_i \left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right)}{\sum_{i=1}^{N} \eta_i \left(1 + \ln y_i - \mathrm{E}\left[\ln y_i\right]\right)} \\ &= \frac{\mathrm{E}\left[\gamma_i\right] + \mathrm{Cov}\left[\gamma_i, \ln y_i\right]}{\mathrm{E}\left[\eta_i\right] + \mathrm{Cov}\left[\eta_i, \ln y_i\right]} \end{aligned} \tag{28}$$

The full solution is

$$I^0\left(.\right) = 1 - \left\{ \frac{1}{b} K(\delta, \beta, |\chi|, |\mathcal{Q}|, N)^{1/b} \frac{\mathrm{E}\left[\gamma_i\right] + \mathrm{Cov}\left[\gamma_i, \ln y_i\right]}{\mathrm{E}\left[\eta_i\right] + \mathrm{Cov}\left[\eta_i, \ln y_i\right]} \right\}^{b/(b+1)} \tag{29}$$

and

$$\varepsilon^0\left(.\right) = \left\{ bK(\delta, \beta, |\chi|, |\mathcal{Q}|, N) \frac{\mathrm{E}\left[\eta_i\right] + \mathrm{Cov}\left[\eta_i, \ln y_i\right]}{\mathrm{E}\left[\gamma_i\right] + \mathrm{Cov}\left[\gamma_i, \ln y_i\right]} \right\}^{1/(b+1)}. \tag{30}$$

Figure 1 illustrates the solution to the social planner's problem when the statistical agency operates the PMW algorithm, as operationalized by Gupta et al. (2012). Substituting the expressions in Theorem 2 and simplifying equation (21) yields the implemented form of the PPF

$$I = 1 - \frac{8\sqrt{3}\left(\ln|\chi|\right)^{1/4} \sqrt{\ln\left(\frac{4}{\delta}\right) \ln\left(\frac{|\mathcal{Q}|}{\beta}\right)}}{\left(N\varepsilon\right)^{1/2}}. \tag{31}$$

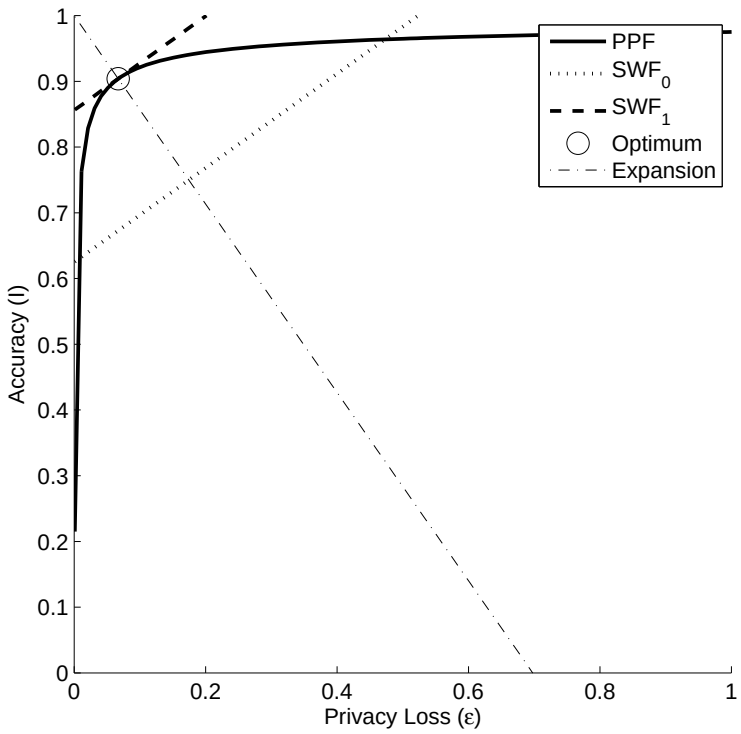

Figure 1: An Illustrative Solution to the Social Planner's Problem

The social welfare function is based on the indirect utility function in equation (27). The solid line represents the production possibilities frontier, equation (31). The dashed lines are contour plots of the social welfare function (24) at representative non-optimal (SWF_0) and optimal (SWF_1) attainable levels of social welfare. The expansion path is the straight line from minimal accuracy, $I = 0,$ to a point on the privacy loss axis between $\varepsilon = 0.6$ and $\varepsilon = 0.8$.

The optimum shown in Figure 1 is the solution to equation (28). It shows the Pareto optimal bundles of (ε, I) associated with model parameters $(\delta, \beta, |\chi|, |\mathcal{Q}|, N)$. For purposes of exposition, we have chosen parameter values that yield a quantitatively intuitive and useful solution.

The PPF in Figure 1 is displayed for a database that measures an arbitrary histogram over a publicly-labeled collection of non-overlapping bins that span the range of a discrete variable measured on a cardinal scale (income in one of our applications). We allow normalized linear queries, $f_t(x) = \frac{1}{N} m_t^T x$ of the form

$$m_t^T = \begin{pmatrix} 1 & \dots & 1 & 1 & 0 & \dots & 0 \end{pmatrix}$$

where m_t is $(|\chi| \times 1)$, and there can be an arbitrary number of initial ones followed by the complementary number of zeros. The query with m^T set to all ones is redundant, so we will remove it from $\mathcal{Q}$. The m_t are a set of basis queries for estimating the cumulative distribution of a discrete variable measured on a cardinal scale. They can be interpreted as allowing the user to ask a sequence of questions of the form "What proportion of the population has values less than or equal to the upper bound of a particular bin?" A finite sequence of such queries is sufficient to estimate the histogram or cumulative distribution function of the underlying discrete variable to arbitrary accuracy in the absence of the privacy-protecting query release mechanism.

The remaining assumptions are

- the population in the database is the 2010 population of the United States ages 18 to 64: $N =$ 194,000,000;
- the size of the sample space is the number of bins into which the cardinal discrete variable is tabulated: $|\chi| =$ 1,000;
- the size of the query set is the number of permissible queries of the allowable

form: $|\mathcal{Q}| = |\chi| - 1 = 999$;

- the probability that a query cannot be answered accurately is $\beta = 0.01$;
- the differential privacy failure bound is $\delta = 0.9/N$.[9]

Using estimates of $\mathrm{Cov}\left[\gamma_i, \ln y_i\right]$ and $\mathrm{Cov}\left[\eta_i, \ln y_i\right]$ based on our analysis of General Social Survey data in Section 6, we obtain optimal accuracy of $I^0 = 0.904$ and privacy loss of $\varepsilon^0 = 0.067$, which is the social optimum plotted in Figure 1.

6 Applications

To investigate the normative content of our model, we use data from the General Social Survey (GSS) and the Cornell National Social Survey (CNSS) to empirically quantify the distribution of the indirect utility function parameters.[10] These surveys ask a representative sample of American adults questions about their income, health status, and attitudes towards data accuracy and privacy protection. We equate the answers provided to the latent marginal utilities in our model. Under these assumptions, we can use the survey responses to measure the marginal rate of substitution in the social welfare function. Equating this rate to the marginal rate of transformation from our PPF allows us to characterize the socially optimal data accuracy and privacy protection implied by the survey responses.

[9] Dwork and Roth (2014, p. 18) argue that values of $\delta > 1/N$ are dangerous since they, in principle, allow publication of at least one exact record in the database.

[10] For the GSS (Smith and Kim 2011), see http://www3.norc.org/gss+website/. 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. See also Appendix A.4.

Our goal is to provide some empirical guidance about the optimal rate at which a public data provider should trade off privacy loss for statistical accuracy. We present results for two applications where privacy loss and accuracy of statistical information are both highly salient: (1) publication of income distribution statistics and (2) publication of relative health status statistics.

Our results provide guidance to data providers in choosing how to manage data privacy. We note, however, that the data from these 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 in measuring the optimal trade-off between privacy loss and data accuracy will require more and better information on individual preferences, 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).

6.1 Publication of Income Distribution Statistics

We began the normative analysis with our specification of the indirect utility function in Section 5.2. The solution to the planner's problem in equation (28) depends upon properties of individual preferences that are amenable to estimation from survey data. Any survey or experimental mechanism that elicits preferences to-

wards privacy loss, γ_i, relative income, η_i and income itself, $\ln y_i$ can in principle be used to quantify the optimal accuracy and differential privacy settings from the social planner's problem. To this end, we use three variables from the 2006 GSS:

- Family income, reported in quartiles.
- `DATABANK`, which records responses on a four-category Likert scale to the following question: "The federal government has a lot of different pieces of information about people which computers can bring together very quickly. Is this a very serious threat to individual privacy, a fairly serious threat, not a serious threat, or not a threat at all to individual privacy?"
- `SCIIMP3`, which records responses on a four-category Likert scale to the following question: "How important [is] the following in making something scientific? The conclusions are based on solid evidence."

We use `DATABANK` as a proxy measure of the latent preference for privacy γ_i and `SCIIMP3` as a proxy measure of the latent preference for privacy η_i. We compute the polychoric correlations between each preference measure and income:

- $\text{Corr}\left[\gamma_i, \ln y_i\right] = -0.144\,(\pm 0.031)$
- $\text{Corr}\left[\eta_i, \ln y_i\right] = 0.189\,(\pm 0.037)$

Interestingly, marginal preferences for privacy are negatively correlated with income, while marginal preferences for evidence-based science, interpreted here as a proxy for accuracy, are positively correlated with income. We do not consider

distributional effects in our analysis, but note in passing that these statistics suggest increasing accuracy at the expense of privacy might favor citizens with higher incomes. This results depends on the context, as our analysis of health statistics in the next section shows.

Taken together, we can equate the marginal rate of transformation between data accuracy and privacy loss to the marginal rate of substitution, as in equation (28).

$$MRT\left(\varepsilon^0, I^0\right) = \frac{\mathrm{E}\left[\gamma_i\right] + \mathrm{Cov}\left[\gamma_i, \ln y_i\right]}{\mathrm{E}\left[\eta_i\right] + \mathrm{Cov}\left[\eta_i, \ln y_i\right]} = 0.720. \tag{32}$$

At the social optimum, a one-unit increment in data accuracy requires a 0.720 unit incremental privacy loss. In making this calculation, we assume that the logarithms of (y, γ, η) are normally distributed about $\left(\mu_{\ln y}, 0, 0\right)$ with unit variances, so that $\mathrm{E}\left[\gamma_i\right] = \mathrm{E}\left[\eta_i\right] = 1$. The same assumption is made in computing the polychoric correlations, as the location of the Likert scale is arbitrary with respect to the underlying latent variables governing preferences.

Optimal accuracy depends on the shape and position of the PPF through the constants $K(\delta, \beta, |\chi|, |\mathcal{Q}|, N)$ and b according to equation (29). If data are provided through the PMW mechanism, and using the same parameters that generate Figure 1, the optimal accuracy and privacy are $I^0 = 0.904$ and $\varepsilon^0 = 0.067$. This is a reasonably tight optimal privacy parameter and a relatively tight accuracy value. Economically, this social optimum implies that individuals prefer to have their place in the income distribution well-protected in the published data and are prepared to accept income distribution estimates that are accurate to, approximately, the decile of the distribution.

We now use the estimated social optimum derived from the GSS data to evaluate the welfare cost of choosing suboptimally low privacy loss at the expense of data accuracy. Choosing a point along the PPF that represents a 3 percent decrease in I, that is the point $I = 0.880$, which is equivalent to a 25 percent increase in α, and the corresponding value of privacy loss on the PPF, $\varepsilon = 0.043$, results in an expected change in utility of -0.008 per person. Given the role of income in the indirect utility function, an income transfer of 0.008 log points (approximately 0.8 percent) to each person would offset the aggregate welfare loss. The estimated parameters imply that a decrease in the accuracy with which income data are published corresponding to approximately one-half decile of accuracy while maintaining the privacy protection level at the original optimum results in an aggregate utility loss of approximately one percent. There will be variation in the actual welfare loss for any person because of the heterogeneity in privacy and accuracy preferences.

6.2 Publication of Health Status Statistics

6.2.1 Restating and Solving the Planner's Problem

We adopt a similar approach to the publication of relative health status statistics, using an alternative specification for the indirect utility function assumes that relative health status $\ln h_i$ interacts with income, data accuracy and privacy loss in a manner that makes the individual better off with more accurate knowledge of her relative health status but worse off for the privacy loss that entails. An equation

similar to equation (27) has the desired properties:

$$
\begin{aligned}
&v\left(y_i, \varepsilon, I, h_i, p\right) = \\
&-\textstyle\sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln y_i - \mathrm{E}(\gamma)\varepsilon - (\gamma_i - \mathrm{E}(\gamma))(1 - \ln h_i)\varepsilon \\
&+ \mathrm{E}(\eta) I + (\eta_i - \mathrm{E}(\eta)) \ln y_i I,
\end{aligned}
\tag{33}
$$

where the notation is as in equation (27), and h_i is an ordinal measure of the relative health status of individual i. We again assume that the logarithms of (y, h, γ, η) follow a joint normal distribution with zero means (except for $\ln y$) and unit variances in the population. Equation (33) reflects the assumption that the marginal disutility of decreased privacy in health statistics is higher when relative health status is poor. That is, sick individuals value privacy of health information statistics more than healthy individuals. We also assume that individuals with higher incomes place a higher value on the accuracy of health status statistics, reflecting the intuition that their stock of health capital has higher long-run value.

Substitution of equation (33) into equation (26) yields

$$
MRT\left(\varepsilon^0, I^0\right) = \frac{\mathrm{E}\left[\gamma_i\right] - \mathrm{Cov}\left[\gamma_i, \ln h_i\right]}{\mathrm{E}\left[\eta_i\right] + \mathrm{Cov}\left[\eta_i, \ln y_i\right]}. \tag{34}
$$

6.2.2 Statistical Results

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 also includes

questions that elicit an assessment of subjective health status along with attitudes toward the privacy of personal health information and the value of accurate health statistics.

We use the following questions from the CNSS:

- Income, measured in nine bins;
- `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?" 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 of the ordinal variables to estimate:

- $\text{Corr}\left[\gamma_i, \ln h_i\right] = 0.015\,(\pm 0.021)$
- $\text{Corr}\left[\gamma_i, \ln y_i\right] = 0.009\,(\pm 0.020)$
- $\text{Corr}\left[\eta_i, \ln y_i\right] = 0.176\,(\pm 0.020)$

It is of some interest that the correlation between accuracy preferences (η) and income $(\ln y)$ is almost identical in the CNSS and GSS samples. On the other hand, the correlation between the privacy preferences (γ) and income has the opposite sign in the two samples implying that the correlation of privacy preferences with income is sensitive to the context in which the data quality question is posed. Concern about the privacy of health status is negligibly correlated with health status and with income. Concern for the quality of medical information, is positively correlated with income. Making the relevant substitutions into equation (34), and, as before, imposing the model restriction $\mathrm{E}(\gamma) = \mathrm{E}(\eta) = 1$ yields

$$MRT\left(\varepsilon^0, I^0\right) = 0.837,$$

which implies that a one-unit increment in the accuracy of health data requires a 0.837 increase in privacy loss. The estimated shadow price of increased health status accuracy, therefore, exceeds the estimated shadow price of increased income distribution accuracy at the social optimum.

To characterize the social optimum, we continue to assume that the data are provided through the PMW mechanism, and use the following parameters:

- the population in the database is unchanged: $N = 194{,}000{,}000$;
- the size of the sample space is the number of bins into which the discrete variable on health status crossed with a single binary characteristic is tabulated: $|\chi| = 10$;
- the size of the query set is the number of permissible queries of the allowable

form: $|\mathcal{Q}| = |\chi| - 1 = 9$;

- the probability that a query cannot be answered accurately is unchanged: $\beta = 0.01$;

- the differential privacy failure bound is unchanged: $\delta = 0.9/N$.

The data publication assumptions are based on allowing any arbitrary linear query about the contents of up to nine cells in the full cross-tabulation. Once again, these assumptions are sufficient to allow fully accurate publication of the underlying 5×2 cross-tabulation in the absence of privacy constraints.

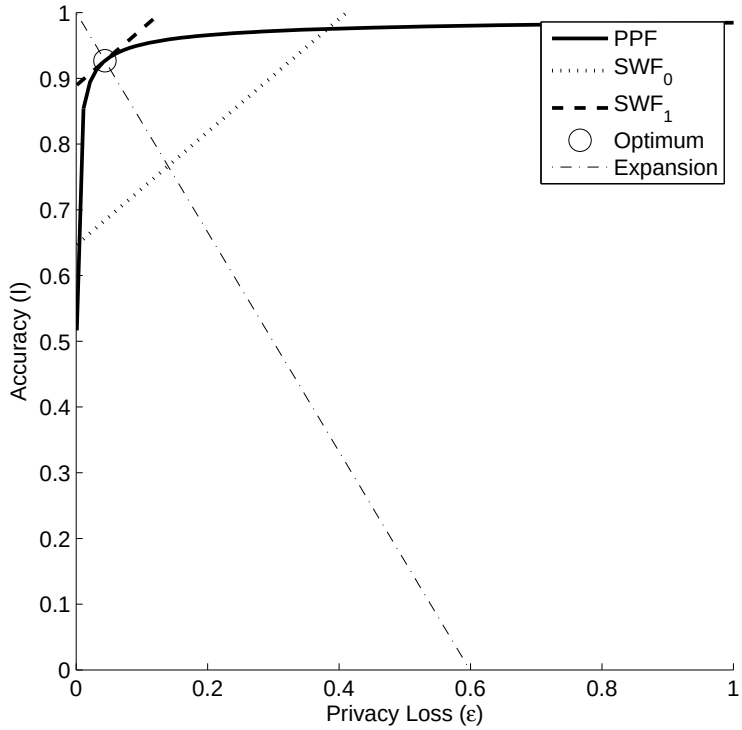

Figure 2: The Social Planner's Problem: Health Statistics

The optimal accuracy is $I^0 = 0.927$ and optimal privacy is $\varepsilon^0 = 0.044$. First, note that because health status is interacted with only a single binary variable in the publication tables allowed by our query space, this problem exhibits a PPF that allows both less privacy loss and more data accuracy for the same other resources, $K(\delta, \beta, |\chi|, |\mathcal{Q}|, N)$ and b. Since the slope of the PPF at the optimum is greater for the health status publication problem, 0.837, than for the income statistics publication problem, 0.720, the relative price of privacy loss, measured in accuracy per unit of foregone privacy, is greater for health status than for income. This equilibrium is displayed in Figure 2.

Once again, we consider the effect on social welfare of increasing privacy at the expense of accuracy. Increasing α by 25 percent to $\alpha = 0.091$, or, equivalently, a 2 percent decrease in I to 0.909, is accompanied by an increase along the PPF in ε to 0.028. The change in average welfare associated with this movement is -0.006 log points. Therefore, the social cost of choosing this suboptimally high level of privacy protection is equivalent to a 0.006 log point (approximately 0.6 percent) decrease in average income.

The simple tabulation of health status in this example complicates direct comparisons with the income distribution publication problem. However, there is no reason to limit the cross-tabulations of health status as we have done. We now consider a more complex publication strategy that exactly replicates the PPF assumptions used for the income distribution problem. This is simple mathematically but requires some subtlety in interpreting those parameters. Setting the sample space to $|\chi| =$ 1,000 implies that we are allowing cross-tabulations of the

five-category health status by 200 dimensions of other characteristics so that the data histogram represents a 5×200 contingency table. Setting the query space to $|\mathcal{Q}| = |\chi| - 1 = 999$ implies that any single linear query takes the form of asking about the contents of exactly one of the cells. In this case, the optimal accuracy is $I^0 = 0.895$ and optimal privacy is $\varepsilon^0 = 0.063$. The magnitudes of these estimates are now directly comparable to the values we found for the income distribution publication problem ($I^0 = 0.904$ and $\varepsilon^0 = 0.067$). When the resource commitments, $K(\delta, \beta, |\chi|, |\mathcal{Q}|, N)$ and b, are identical, the difference in the estimated social preferences between the specification in equation (27) and the one in equation (33) implies that published tables of health statistics should be five percent more private and one percent more accurate than published tables of income statistics.

For the more complex health status publication problem, we also compute the loss in total social welfare from decreasing the accuracy of the publication by 3 percent to $I = 0.869$, forcing the privacy parameter to $\varepsilon = 0.040$ along the PPF. The resulting loss of social welfare is -0.008 log points, or approximately 0.8 percent loss in total welfare.

The similarity of these health status results, based on the CNSS, to those of the income distribution problem, based on GSS data, is somewhat coincidental. The estimated correlations of income, health, and preferences for privacy and accuracy lead to a very similar measure of the average willingness to pay for privacy in foregone accuracy. We use the same calibration for the data production technology, so we find a very similar solution to the social planner's problem.

6.3 The Provision of Both Income Distribution and Health Status Statistics

It seems reasonable to 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 will not develop that model here.

Instead, to illustrate a problematic consequence of this technology, we use results on the composability of (ε, δ)-differential privacy to reason about expanding the set of published queries. Intuitively, differential privacy loss is additively composable across independent data releases. The details of composability are covered thoroughly in Theorems 3.14 and 3.20 of Dwork and Roth. These details are relatively complicated, and we defer a full discussion to future work.[11] We can however 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 solu-

[11] 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.

tions 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 Private Multiplicative Weights query release mechanism.

Consumers demand the statistics supplied by the government agency because of their interdependent preferences. They want to know where they fit in the in-

come 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 either the VCG or Lindahl mechanisms for private-provision of data. The reason is that the VCG mechanism 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 Lindahl mechanism for procuring data-use rights has a lower marginal cost of acquisition than VCG but also ignores the public-good aspect of the data accuracy. Neither mechanism accounts 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) relative 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 overprovision of privacy protection is harmful to the citizens when the demand for the statistical products of the agencies is derived from interdependent preferences.

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.

Bibliography

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

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, *SIGMOD '00 Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* p. 439450.

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): pp. 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.

Blum, A., Ligett, K. and Roth, A. (2008). A learning theory approach to non-interactive database privacy, *Proceedings of the 40th annual ACM symposium on Theory of computing*, STOC '08, ACM, New York, NY, USA, pp. 609–618.

Clark, A. E., Frijters, P. and Shields, M. A. (2008). Relative income, happiness, and utility: An explanation for the easterlin paradox and otherpuzzles, *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. **URL:** *http://www.ciser.cornell.edu/beta/CNSS/*

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 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): pp. 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.

Gupta, A., Roth, A. and Ullman, J. (2011). Iterative constructions and private data release, *CoRR* **abs/1107.3731**.

Gupta, A., Roth, A. and Ullman, J. (2012). Iterative constructions and private data release, *Proceedings of the 9th International Conference on Theory of Cryptography*, TCC'12, Springer-Verlag, Berlin, Heidelberg, pp. 339–356.

Hardt, M., Ligett, K. and McSherry, F. (2010). A simple and practical algorithm for differentially private data release, *CoRR* **abs/1012.4763**.

Hardt, M. and Rothblum, G. N. (2010). A multiplicative weights mechanism for privacy-preserving data analysis, *IEEE 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 Trans. Knowl. Discov. Data* **1**(1). **URL:** *http://doi.acm.org/10.1145/1217299.1217302*

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem, *Review of Economic Studies* **60**(3): 531–542.

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): pp. 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.

Smith, Tom W, P. M. M. H. and Kim, J. (2011). *General Social Surveys, 1972-2010: cumulative codebook*, National Data Program for the Social Sciences Series, no. 21, Chicago: National Opinion Research Center.

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.

APPENDIX

A.1 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

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

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

GR-Fact 1 (stated without proof, but see Dwork and Roth (2014, p. 42-43) for

a proof) says that ε_i-minimal differential privacy composes. That is, if algorithm $A(D)$ is ε_i-minimally differentially private, $T \subset \{1, \ldots, n\}$, and $D, D^{(T)} \in \{0,1\}^n$ with $H\left(D, D^{(T)}\right) = |T|$, then

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

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

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

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

vate 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 application of GR-Definition 2.3 ($\mathrm{Lap}\,(\sigma)$ noise addition) to construct an ε-differentially private response to the counting query based on GR-Theorem 3.3 with $|H| = (1-\alpha)\, n$ and the indices ordered such that $H = \{1, \ldots, |H|\}$. 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 $\mathrm{ceil}\,((1-\alpha)\, n)$. The resulting answer from the query response mechanism is

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

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

A.2 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 $v_i(y_i, I, \phi, y, p)$ is homogeneous of degree zero in (p, y), note that for all $\lambda > 0$

$$\begin{aligned}
v_i\left(\lambda y_i, \varepsilon, I, \lambda y^{\tilde{}i}, \lambda p\right) &= -\sum_{\ell=1}^{L} \xi_\ell \ln(\lambda p_\ell) + \ln(\lambda y_i) - \gamma_i (1 + \ln(\lambda y_i) - \mathrm{E}[\ln \lambda y_i])\,\varepsilon \\
&\quad + \eta_i (1 + \ln(\lambda y_i) - \mathrm{E}[\ln \lambda y_i])\, I \\
&= -\sum_{\ell=1}^{L} \xi_\ell \ln \lambda - \sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln \lambda + \ln y_i \\
&\quad - \gamma_i (1 + \ln \lambda + \ln y_i - \mathrm{E}[\ln \lambda] - \mathrm{E}[\ln y_i])\,\varepsilon \\
&\quad + \eta_i (1 + \ln \lambda + \ln y_i - \mathrm{E}[\ln \lambda] - \mathrm{E}[\ln y_i])\, I \\
&= -\sum_{\ell=1}^{L} \xi_\ell \ln p_\ell + \ln y_i - \gamma_i (1 + \ln y_i - \mathrm{E}[\ln y_i])\,\varepsilon \\
&\quad + \eta_i (1 + \ln y_i - \mathrm{E}[\ln y_i])\, I \\
&= v_i\left(y_i, \varepsilon, I, y^{\tilde{}i}, p\right) \qquad (35)
\end{aligned}$$

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

For all $\lambda > 1$

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

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.3 The Private Multiplicative Weights Algorithm

We provide a complete description of the PMW mechanism based on the presentation in Gupta et al. (2012). The algorithm was introduced in Hardt and Rothblum (2010). Gupta et al. compose their Algorithm 1 (Online Query Release Mechanism) and their Algorithm 4 (Multiplicative Weights Iterative Database Construction) to get the complete algorithm that Hardt and Rothblum call Private Multiplicative Weights.

To maintain consistency with the presentation in Sections 3 and 5, we present the PMW algorithm using an unnormalized histogram to represent both the con-

fidential and synthetic databases, and normalized linear queries operating on both the confidential and synthetic databases. Hardt and Rothblum (2010) and Gupta et al. (2012) present the algorithm using a normalized histogram to represent the synthetic database (which is then rescaled to the population size) and unnormalized queries operating on the original unnormalized histogram. With linear queries, the choice of where to normalize is arbitrary. All symbols in the Algorithm *Private Multiplicative Weights* have the same meaning as in the text.

Algorithm *Private Multiplicative Weights*

Input: An unnormalized histogram, x, from a database whose elements have cardinality $|\chi|$; number of record in the original database, $||x||_1 = N$; differential privacy parameters $\varepsilon > 0$ and $0 < \delta < 1$; accuracy parameter, $0 < \alpha < 1$; accuracy failure probability $0 < \beta < 1$; a number, k, of normalized linear queries to answer adaptively from $\mathcal{Q} \subseteq \mathcal{F}$ with cardinality $|\mathcal{Q}|$. Each normalized linear query, $f_t(x) \equiv \frac{1}{N} m_t^T x$ where $m_t \in [-1, 1]^N$, which may be specified interactively.

1. Set the Laplace scale parameter $\sigma = \phi_\sigma(B(\alpha; \chi), \varepsilon, \delta)$.
2. Set query threshold $T = \phi_T(k, \beta, \sigma)$.
3. Set weight parameter $\mu = \phi_\mu(\alpha)$.
4. Set stopping threshold $B = B(\alpha, |\chi|)$.
5. Initialize the synthetic database: $\tilde{x}_0 = \frac{N}{|\chi|} u_{|\chi|}$, where $u_{|\chi|}$ is the unit vector of length $|\chi|$.
6. **for** $t \leftarrow 1$ **to** k
7. Get query f_t.
8. Sample A_t from $\mathrm{Lap}(\sigma)$.
9. Compute the noisy answer to f_t using the original database, $\hat{a}_t \leftarrow f_t(x) + A_t$.
10. Compute the answer to f_t using the synthetic database, $\tilde{a}_t \leftarrow f_t(\tilde{x}_{[t-1]})$.
11. Compute the difference between the noisy and synthetic answers: $d_t \leftarrow \hat{a}_t - \tilde{a}_t$.
12. **if** $|d_t| \leq T$ **then** set $w_t \leftarrow 0$ and output $\tilde{a}_t$ (no privacy budget expenditure because the synthetic data answer was close enough).
13. **if** $|d_t| > T$ **then do**
14. $w_t \leftarrow 1$ (update mechanism: expend some of the privacy budget to update the synthetic data).
15. output $\hat{a}_t$
16. **for** $i \leftarrow 1$ **to** $|\chi|$
17. **if** $d_t > 0$ define $r_t[i] \leftarrow m_t[i]$
18. **else** $d_t \leq 0$ define $r_t[i] \leftarrow (1 - m_t[i])$.

19. Update: $y_t[i] \leftarrow \frac{1}{N}\tilde{x}_{t-1}[i] \times \exp(-\mu r_t[i])$.
20. Normalize: $\tilde{x}_t[i] \leftarrow N \times \frac{y_t[i]}{\sum_i y_t[i]}$.
21. **end for**
22. **end if**
23. Update the count of the number of update loops $z \leftarrow \sum_{\tau=1}^{t} w_\tau$.
24. **if** $z < B$ then continue.
25. **else** terminate.
26. **end for**

In our notation

$$\phi_\sigma(B(\alpha;\chi),\varepsilon,\delta) \equiv \frac{1000\sqrt{B(\alpha,|\chi|)}\ln(4/\delta)}{\varepsilon}$$

and

$$\phi_T(k,\beta,\sigma) = 4\phi_\sigma(B(\alpha;\chi),\varepsilon,\delta)\ln(2k/\beta).$$

The reader is referred to Gupta et al. (2012) for the functional forms of the original definition of the algorithm. Here we want to highlight the key ideas as they relate directly to the notation we use in our analysis. Gupta et al. (2012) derive a closed-form solution for the stopping threshold $B = B(\alpha^*; N, |\chi|) = 4N^2 \ln|\chi| / (\alpha^*)^2$, using their definition of the accuracy parameter, which for clarity we call α^* here– the accuracy of an unnormalized query. Converting to normalized queries gives $B(\alpha,|\chi|) = 4\ln|\chi|/\alpha^2$ since our accuracy parameter, α, equals α^*/N. The limit B puts an upper bound on the number of updates to the synthetic database that can be answered with the differentially private query answer $\hat{a}_t$ in order to achieve the accuracy α for all queries in the set $\mathcal{Q}$ within the (ε,δ)-differential privacy budget (their Theorem 8). Via their Definition 4 and the proof of their Theorem 8 (see Gupta et al. (2011) for the proofs), they establish that after $B(\alpha,|\chi|)$ invocations of the update mechanism PMW provides (α,β)-accurate answers to all k queries with (ε,δ)-differential privacy.

Following Gupta et al., we set $k = |\mathcal{Q}|$. These answers can either be released as $\{\hat{a}_{t^*}\}$ or as $f_{t^*}(\tilde{x}_B)$, where t^* are the indices of t at which the algorithm set $w_t = 1$, f_{t^*} is the associated query from $\mathcal{Q}$, and $\tilde{x}_B$ is the terminal value of the synthetic database. Note that $|\{\hat{a}_{t^*}\}| = B(\alpha,|\chi|)$. Their proof uses a potential function approach to show that the synthetic database will answer queries with sufficient accuracy such that after B update steps it will always be the case that $|d_t| \leq T$, so that the privacy budget is exactly exhausted. With this threshold in hand, the application to our Theorem 2 of their general privacy and accuracy results follows directly. We omit the parallel argument that the accuracy bound for the Median

Mechanism (MM) is

$$\alpha \propto \frac{(\ln|\chi|)^{1/4}(\ln|\mathcal{Q}|)^{3/4}}{(N\varepsilon)^{1/2}} \tag{36}$$

since we do not use this mechanism in our applications.

A.4 Data Sources

The raw data are from the General Social Survey (GSS) and the Cornell National Social Survey(CNSS). The input data files sources are:

- General Social Survey: obtained from the NORC GSS download site: http://publicdata.norc.org/GSS/DOCUMENTS/OTHR/GSS_stata.zip. We used the 2012 R2 release. Our analysis is restricted to variables collected only in 2006.
- 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/.