Estimating Preferences Over Data to Inform Statistical Disclosure Control Decisions

Elan Segarra*†

April 2024

Abstract

This project provides a framework and empirical example for the estimation of consumer demand for published statistics, and incorporation of these estimates into the process of statistical disclosure control (SDC). When implementing SDC methods, data providers are tasked with balancing the benefits of published statistics with the risks of re-identification of entities in the confidential micro-data. Typically, the benefit side of this calculation is reduced to maximizing the number of published statistics, as opposed to assessing which statistics might be more useful to downstream consumers. In the context of the cell suppression problem, data providers must choose complementary suppressions to protect against secondary disclosure attacks, and in the context of differential privacy data providers must choose how to allocate their privacy budgets across different sets of output. Incorporating valuations over potentially published statistics can help inform these decisions. Consumer demand for statistics is modeled using a discrete choice nested logit model where individual statistics can vary by characteristics such as their conditioning variables (e.g. labor data sliced by occupation versus industry). To illustrate its feasibility, the framework is applied to the Census of Fatal Occupational Injuries. Preferences are estimated using standard approaches on page-view data of public CFOI webpages, and the parameter estimates are used to compute valuations which are leveraged in both cell suppression and differential privacy approaches to protecting CFOI tables.

JEL Codes: C81, D83, H41

Key Words: statistical disclosure control; secondary disclosure; formal privacy; differential privacy; preference estimation; public goods

*Bureau of Labor Statistics, Office of Compensation and Working Conditions; Contact: Segarra.Elan@BLS.gov

†The views expressed herein are those of the author(s) and do not necessarily reflect those of the Federal Government, Department of Labor, or the Bureau of Labor Statistics. All results have been reviewed to ensure that no confidential information is disclosed.

1 Introduction

When data providers are tasked with protecting the confidentiality of entities whose information contributes to published statistics they are implicitly forced to find a balance between public benefits and private re-identification risk. While there has been ample research introducing new methods to protect confidentiality, there has been relatively little study of how to set the associated parameters that capture how much risk we are willing to accept or where the risk should be concentrated. Without more research on how to factor in the public benefits of accurate statistics there is a strong potential to inadvertently overprotect the statistics we value most. In this paper we provide a novel framework for the estimation of data consumer's valuations and how to incorporate these into statistical disclosure control (SDC) methods. We additionally apply the proposed framework to a real-world data set, the Census of Fatal Occupational Injuries, to demonstrate the entire pipeline from preference estimation, to statistics valuation, to parameter decisions on two common SDC methods.

The remainder of this paper is organized as follows. Section 2 lays out the model of consumer preferences over published statistics including strategies for estimation. Section 3 describes two common SDC methods, cell suppression and differential privacy, and how to generate valuations to inform these methods. Section 4 presents an application of the entire framework to the Census of Fatal Occupational Injuries, and section 5 concludes.

2 Model

In this section we describe a model that encapsulates the preferences of data consumers and serves as the foundation for both the consumption behaviors over published data products and for decisions surrounding statistical disclosure control methods.

2.1 Preferences Over Statistics

To establish the connection between core preferences over individual statistics and the consumption habits of data users, we start be defining three objects.

Definition 1. A *statistic*, indexed by $s \in \mathcal{S} = \{1, 2, \cdots n_S\}$, is a single scalar communicating information about the data set of interest.

Definition 2. A *publication*, indexed by $p \in \mathcal{P} = \{1, 2, \cdots n_P\}$, is a set of statistics, $\mathcal{S}_p \subseteq \mathcal{S}$, collected together and formatted as a single document for consumption by the public.

Definition 3. A *market*, indexed by t, is a set of P_t publications from which a data consumer can choose at time t.

The above definitions permit amply heterogeneity across what information could be communicated from the underlying data source. To solidify concepts Example 1 presents a situation involving unemployment rates. For most national statistical agencies (NSAs) and data providers

these statistics are likely to be simple descriptive measures such as counts or averages of various variables, however the framework also permits more complicated objects such as elasticity estimates from a large structural economic model.

Example 1. The estimated unemployment rate of 5.4%, for the construction industry in March of 2024 (Bureau of Labor Statistics 2024), is a *statistic* in our framework. If we were to collect all the unemployment rates across various sectors (e.g. by 2 digit NAICS codes) and present them in either a table or a chart, this object would constitute a *publication*. Finally, we could imagine multiple different tables or charts that each present unemployment rates sliced along difference dimensions, e.g. by industry, occupation, or geography. Collectively, these would make up a *market* of different options that the downstream data consumer could choose among in their quest to learn about the unemployment situation.

Statistics and publications each have associated characteristics, denoted by $x_s \in \mathcal{X} \subseteq \mathbb{R}^k$ and $z_p \in \mathcal{Z} \subseteq \mathbb{R}^r$ respectively. Characteristics of a statistic could include the value of the statistic, indicators of whether it conditions on a specific industry, or the type of statistic (e.g. count versus average versus rate). For a publication, characteristics can include the number of statistics in the publication, how many are suppressed, how the statistics are presented (e.g. table versus plot), or even whether they involve a comparison across time versus a single period. For additional examples refer to Table 1 for a list of the features leveraged in the application.

We model preferences using a standard nested logit model of preferences over the statistics and publications. Let data consumer i have indirect utility from publication p in market t given by

$$U_{ipt} = \underbrace{\frac{1}{|\mathcal{S}_p|}\sum_{s\in\mathcal{S}_p} x_{st}\beta + z_{pt}\alpha + \xi_{pt}}_{\equiv\delta_{pt}} + \varsigma_{ig} + (1-\sigma)\epsilon_{ipt} \tag{1}$$

As introduced earlier, x_{st} and z_{pt} are vectors of observable attributes of the statistic and publication respectively, while β and α are parameters representing consumer preferences over these observable characteristics. The first error term, ξ_{pt}, includes any unobservable (to the econometrician) preferences, and together with the first two terms make up the mean utility across individuals, δ_{pt}.

The final two error terms represent unobservable consumer level idiosyncratic shocks. The nest or group shock, ς_{ig}, is common across all publications found in nest g and models the potential correlation between certain sets of publications (e.g. if a group of publications is all accessed via the same primary website). The idiosyncratic error, ϵ_{ipt}, includes all other unobserved consumer level preferences. These two shocks are assumed to follow a Cardell distribution and a Type 1 Extreme Value distribution respectively which ensures that the total idiosyncratic error, $\varsigma_{ig} + (1-\sigma)\epsilon_{ipt}$, follows a Type 1 Extreme Value distribution (Cardell 1997). This distributional assumption is desirable because we will see shortly that it results in a closed form solution for the share of each publication chosen.

Note that the preferences over the statistic's characteristics enter the utility function as an average over included statistics. This modeling choice helps to address the issues inherent in

comparing publications that condition on different dimensions with different numbers of cases. For example, a publication of unemployment rates by state would naturally have 50 statistics while a publication of unemployment rates by occupation codes might have more than 500 statistics which might artificially inflate the estimated valuation of conditioning on occupation. By averaging the contribution we can disentangle the effect of the conditioning dimension chosen (e.g. conditioning on state versus occupation) from the resulting number of statistics reported. For use cases where the valuation of the number of statistics is important, we suggest including this characteristic in z_{pt}. For parsimony sake, throughout the remainder of the paper we will collapse this average into $\tilde{x}_{pt}$ so that

$$U_{ipt} = \tilde{x}_{pt}\beta + z_{pt}\alpha + \xi_{pt} + \varsigma_{ig} + (1-\sigma)\epsilon_{ipt} \quad \text{where} \quad \tilde{x}_{pt} = \frac{1}{|\mathcal{S}_p|}\sum_{s\in\mathcal{S}_p} x_{st}. \tag{2}$$

Though the line between statistic characteristics and publication characteristics can sometimes be ambiguous, especially when a characteristic is constant across statistics within a publication, the linearity of the model implies that the distinction is often only superficial.

Given the model for preferences over statistics and publications we now discuss the consumption behavior of consumers. Following Berry (1994), we can derive the share of consumption of each of these options under the distributional assumption described earlier. Under the previous utility model, if consumers in market t are presented with P_t publication options, $M_t = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{P_t}\}$, and they choose exactly one option that maximizes their indirect utility then the share, y_{pt}, of consumers choosing publication p is given by

$$y_{pt} = \frac{\exp\left(\frac{\delta_{pt}}{1-\sigma}\right)}{D_g^{\sigma}\sum_h D_h^{1-\sigma}} \qquad \text{where} \qquad D_g = \sum_{k\in g}\exp\left(\frac{\delta_{kt}}{1-\sigma}\right). \tag{3}$$

Though it is rare that we observe the individual choices of all consumers, there are many scenarios in which we observe either the aggregate counts of certain choices or the percentage of consumers that make a certain choice. Equation (3) provides a link between the underlying consumer preference parameters and potentially observable information.

2.2 Identification and Estimation of Preferences

The primary objects of interest in the model are the preference parameters, β and α, since they can be used to derive valuations over potentially published statistics and publications (as we will illustrate in Section 3.3). The identification and estimation of these parameters can be achieved in several ways as illustrated by the breadth of the literature on discrete choice models (see **?**). For simplicity sake we take the route of Berry (1994) by transforming the nonlinear relationship of equation (3) into a linear one that relates the observable characteristics with an observable outcome variable. This "inversion step" is one of the most attractive features of the nested logit model and likely motivates its utilization across a number of disciplines and fields. Following a little bit of algebra, it can be shown that

$$\ln y_{pt} - \ln y_{0t} = \tilde{x}_{pt}\beta + z_{pt}\alpha + \sigma \ln y_{p|g} + \xi_{pt} \tag{4}$$

where $y_{p|g}$ is the share of publication p chosen among options in nest g.

There are two beneficial consequences of this transformation. The first is that establishing identification of the model parameters, (β, α, σ), is made easier given the familiarity of linear models. In particular, since the left hand side is observable then standard econometric theory dictates that the parameters are identified provided we can find suitable instruments for the observable variables, $(\tilde{x}_{pt}, z_{pt}, \ln y_{p|g})$, with respect to the unobservable error ξ_{pt}. The existence of such instruments will depend on the specific context and characteristics of the particular data set being used to estimate these preferences (see section 4.3 for an example in the application). In some situations the characteristics may be plausibly exogenous, however it is important to note that $y_{p|g}$ is endogenous by construction and will always require instrumentation.

The second advantage of this form is that it permits the use of more tractable estimation strategies, such as instrumental variables regression, rather than the more demanding methods typically required for nonlinear models. These methods have closed form solutions and therefore do not suffer from the various difficulties inherent to optimization-based estimators.

3 Statistical Disclosure Control

The methods for protecting the confidentiality of the microdata are myriad and diverse. Though the results of estimating user preferences can no doubt inform several different approaches to SDC, we focus on two of the most predominant methods found in practice. The first is cell suppression, where potentially disclosive cells are omitted from publication, and the second is differential privacy, which typically involves injecting noise into published statistics. In this section we describe both approaches and layout how estimates from the preference framework can be used to bolster and guide the tuning of these methods.

3.1 Cell Suppression

Cell suppression for disclosure control is the simple process of omitting any statistics, such as cells in a larger table, that are deemed to sufficiently threaten the confidentiality of the underlying microdata. Suppression methods were considered by NSAs long ago (Fellegi 1972) and research into their subtleties and implementation also has a long history (Fischetti and Salazar 2001).

Implementing cell suppression typically consists of a two step procedure. First, cells or statistics whose publication are deemed sensitive are marked for primary suppression according to some rule defined by the data provider. Potential measures of sensitivity are numerous and the thresholds for primary suppression vary both by the context of the microdata and the NSA's level of aversion to disclosure risk. In the second step, the remaining cells that are being considered for publication must be reviewed to ensure that their publication does not implicitly reveal any cells that were suppressed in step one. These implicit revelations come about from the natural relationships among the presented statistics, for example row or column totals in the same or a related table. A selection of cells are then identified for complementary (also known as secondary) suppression to

prevent disclosure risk of all suppressed cells. However, there is often room for choice about which cells can be marked for complementary suppression and this is where estimated valuations over the statistics can provide guidance.

The complete cell suppression problem (CSP) is as follows. Let $x \in \mathcal{X} = \mathbb{R}^n$ represent a vector of n statistics being considered for publication. Suppose they satisfy the known bounds, $x_{lb} \leq x_s \leq x_{ub}$, and the known linear constraint $Ax = b$. The linear constraint represents the relationships among the statistics such as published totals of other published statistics. Let $q^{(ps)} \in \mathcal{Q} = \{0,1\}^n$ be a primary suppression vector such that $q_s^{(ps)} = 1$ indicates that statistic s is marked for primary suppression. Similarly, let $q^{(cs)} \in \mathcal{Q} = \{0,1\}^n$ indicate the complementary suppressions, and finally let $q \in \mathcal{Q} = \{0,1\}^n$ indicate either suppression (i.e. $q_s = \mathbb{1}\{q_s^{(ps)} = 1 \text{ or } q_s^{(cs)} = 1\}$). Let the value associated with a given set of published statistics be given by $f : \mathcal{X} \times \mathcal{Q} \to \mathbb{R}$. A common assumption of this valuation function is the independence of valuations across statistics within a publication so that $f(x,q) = \sum_s v_s(1 - q_s)$ where v_s is the valuation (or weight) of publishing statistic s. Finally, most CSP statements also include a set of bounds defining the desired level of protection guaranteed on suppressed cells which may or may not be cell specific. Given cell specific bounds, lb_s and ub_s, we want to guarantee that the implied value of a suppressed cell s contains the interval $(x_s - lb_s, x_s + ub_s)$.

The CSP consists of finding an optimal set of suppressions conditional on an intruder doing their best to identify suppressed cells. The intruder's problem is to take the available information, i.e. the published statistics ($x^* = \{x_s : q_s = 0\}$) and known constraint ($Ax = b$), and determine the set of implied potential values of the suppressed statistics. Due to the linearity of the constraint this takes the form of determining the interval, $(\underline{x}_s, \overline{x}_s)$, of potential values for each suppressed statistic:

$$\underline{x}_s(x^*, q) = \min_{r \in \mathbb{R}^n} r_s \quad s.t. \qquad \overline{x}_s(x^*, q) = \max_{r \in \mathbb{R}^n} r_s \quad s.t.$$
$$Ar = b, \quad x_{lb} \leq r \leq x_{ub} \qquad\qquad Ar = b, \quad x_{lb} \leq r \leq x_{ub}$$
$$r_j = x_j^* \text{ for } q_j = 0 \qquad\qquad r_j = x_j^* \text{ for } q_j = 0$$

The data provider's objective is to choose complementary suppressions that maximize the value of the published data product while ensuring the solution to the intruder's problem satisfies the level of protection defined earlier:

$$\begin{aligned} \max_{q \in \mathcal{Q}^*} f(x,q) \qquad \text{where} \qquad \mathcal{Q}^* = \{\, q \in \mathcal{Q} : & \\ & q_s = 1 \quad \forall\, s \text{ with } q_s^{(ps)} = 1, \\ & \underline{x}_s(x^*, q) \leq x_s - lb_s \quad \forall\, s \text{ with } q_s = 0 \\ & \overline{x}_s(x^*, q) \geq x_s + ub_s \quad \forall\, s \text{ with } q_s = 0 \,\} \end{aligned} \tag{5}$$

The domain over which this optimization occurs, $\mathcal{Q}^*$, consists of the set of suppressions that respect the designated primary suppressions and the protection bounds.

Though the objective statement is straightforward, the CSP is NP-hard (Kelly et al. 1992) which means it can quickly become intractable when involving large sets of statistics. Substantial research

has been conducted to provide and compare methods for solving this problem, including heuristic methods to accommodate scenarios where data size makes complete algorithms unfeasible. For reviews of methods and work in the area of implementation see Castro (2012) and Castro (2023).

3.2 Differential Privacy

Since the publication of Dwork et al. (2006) there has been significant research effort put toward expanding, assessing, and implementing the space of SDC methods with formal privacy guarantees. The majority of the approaches involve a stochastic element, such as noise injection, to help mask the likelihood of identifying individual entities in the microdata from published statistics. For the purposes of this paper we focus on the original property of pure ε differential privacy (ε-DP), however, many of the extensions and relaxations of ε-DP can also be used in the contexts we describe with some alteration.

Let the underlying microdata be represented by $D \in \mathcal{D}$ and let $x \in \mathcal{X} = \mathbb{R}^n$ represent a vector of n statistics generated from the microdata by function g, so that $g(D) = x$. Rather than publish these exact statistics, the data provider instead considers using a *mechanism* which is a stochastic function mapping each statistic to a new perturbed statistic, $\mathcal{M} : \mathcal{X} \to \mathcal{X}^*$. This mechanism is said to have the property of ε-Differential privacy if it satisfies the following definition.

Definition 4. For $\varepsilon > 0$, the mechanism $\mathcal{M}$ is said to satisfy ε-*differential privacy* if for any two adjacent data sets, D and D', that differ in exactly one element and for all outcomes $A \in \mathcal{X}^*$ we have

$$\frac{Pr\left[\mathcal{M}(g(D)) \in A\right]}{Pr\left[\mathcal{M}(g(D')) \in A\right]} \leq e^{\varepsilon}.$$

This property essentially ensures that there is sufficient noise in the published statistics, x^*, such that it cannot be determined, up to a certain level, whether the microdata that generated the statistics does or does not contain a specific individual. In this definition ε, often called the privacy budget, can be thought of as the amount of privacy a data provider can spend in order to safely publish the information produced by the underlying microdata and mechanism. Setting smaller values of ε suggests that there is little room between publishing the statistics and re-identification of the underlying microdata, so a mechanism must typically inject a large magnitude of noise in order to achieve ε-DP. Conversely, larger values of ε suggest there is ample room for publishing the statistics with less worry of re-identification, which means mechanisms do not need to inject as much noise.

Beyond establishing a formal mathematical privacy guarantee, this property also has several other advantageous characteristics including composability and invariance to post-processing.

Lemma 1 (Composability)**.** *Let $\mathcal{M}_1, \ldots, \mathcal{M}_k$ be k independent mechanisms that each satisfy ε-DP with respective guarantees of $\varepsilon_1, \ldots, \varepsilon_k$. If we consider the joint mechanism, $\mathcal{M} = (\mathcal{M}_1, \ldots, \mathcal{M}_k)$, then $\mathcal{M}$ also satisfies ε-DP with guarantee $\varepsilon = \sum_{i=1}^{k} \varepsilon_i$.*

Lemma 2 (Invariance to post-processing)**.** *Let the mechanism $\mathcal{M}$ satisfy ε-DP and let g be any function of the output of $\mathcal{M}$. Then the mechanism $g(\mathcal{M})$ also satisfies ε-DP.*

Composability will be especially useful in later discussion regarding the allocation of a privacy budget. Refer to Dwork (2008) for a discussion of multiple facets of differential privacy.

3.3 Leveraging Preferences in SDC

Once preferences have been estimated according to the model and methods outlined in Section 2.1 they can be incorporated into the the SDC methods being utilized. In this section we describe how to leverage these estimates in each of the SDC approaches introduced, and discuss a few aspects of this process that should be considered. The approach outlined below consists first of establishing a valuation for a potential publication and then illustrating how that valuation can be used in each of the SDC methods previously described.

Suppose there are P potential publications under consideration consisting of $\mathcal{S}_1, \ldots, \mathcal{S}_P$ sets of statistics respectively. Note that each publication may consist of exactly one statistic ($|\mathcal{S}_p| = 1$). As described in Section 2.1, each of these publications have characteristics given by $\tilde{x}_p$ and z_p. Given consistent estimates for the primary preference parameters, $\widehat{\beta}$ and $\widehat{\alpha}$, one approach to generating a valuation for each of these publications would be to compute the expected indirect utility using equation (2)

$$\mathbb{E}\left[U_{ip} | \tilde{x}_p, z_p\right] = \tilde{x}_p \widehat{\beta} + z_p \widehat{\alpha},$$

however these valuations would be problematic for a number of reasons. The ordinal nature of utility coupled with the fact that they are normalized so the utility of outside option is 0 make the exact value of $\mathbb{E}\left[U_{ip}\right]$ difficult to interpret. Moreover, since there are no quantitative constraints, i.e. estimated utilities could be negative, the estimated utilities are impossible to compare outside of simple preference orderings.

To create valuations that are more amenable to comparison and usable in SDC methods we will transform them according to the discrete choice model described earlier.

Definition 5. Given P publications with respective characteristics $\tilde{x}_p$ and z_p and estimated preference parameters (β, α), let the *valuation* of publication p be denoted by v_p and given by

$$v_p = \frac{\exp\left(\mathbb{E}\left[\delta_p | \tilde{x}_p, z_p\right]\right)}{\sum_k \exp\left(\mathbb{E}\left[\delta_k | \tilde{x}_k, z_k\right]\right)} = \frac{\exp\left(\tilde{x}_p \beta + z_p \alpha\right)}{\sum_k \exp\left(\tilde{x}_k \beta + z_k \alpha\right)}. \tag{6}$$

This approach essentially supposes a hypothetical market that consists of only the P potential publications, and the valuation is then defined as the share of consumers that would choose that publication. This hypothetical market is simplified along two dimensions: all choices are assumed to share the same nest and there is no outside option. This ensures that valuations are dictated by the estimated substitution patterns between choices rather than with the outside option.

This definition of valuation not only maintains the preference ordering of the estimated mean utilities, but also has additional desirable properties. Under this definition all valuations are

bounded, $v_p \in [0,1]$, and are easily interpretable through the lens of the hypothetical market. For example, the relative valuations can be utilized more easily since one publication having a valuation that is twice another does in fact indicate twice the desirability within the context of the hypothetical market.

Given estimates of the preference parameters, $\widehat{\beta}$ and $\widehat{\alpha}$, we can easily construct estimated valuations using the plug-in version of equation (6):

$$\widehat{v_p} = \frac{\exp\left(\tilde{x}_p\widehat{\beta} + z_p\widehat{\alpha}\right)}{\sum_{k=1}^{P}\exp\left(\tilde{x}_k\widehat{\beta} + z_k\widehat{\alpha}\right)}. \tag{7}$$

Provided there is at least one publication with a finite mean utility, $\mathbb{E}[\delta_p]$, then this plug-in valuation estimator is continuous over the relevant parameter space. This fact coupled with the continuous mapping theorem implies that our valuation estimator is consistent for the true valuation, $\widehat{v_p} \to_p v_p$.

Leveraging these newly estimated valuations in the context of the cell suppression problem is straightforward. Since we need to assess the relative valuations of each cell within the table we define our publications so that each consists of exactly one cell in the relevant tables, $\mathcal{S}_p = \{s_p\}$, and therefore $\widehat{v}_p = \widehat{v}_s$. With the objective already defined in equation (5) the data provider can simply substitute these estimated valuations in for the weights:

$$\widehat{v}_s = \widehat{v}_p = \frac{\exp\left(\tilde{x}_p\widehat{\beta} + z_p\widehat{\alpha}\right)}{\sum_{k=1}^{P}\exp\left(\tilde{x}_k\widehat{\beta} + z_k\widehat{\alpha}\right)} \qquad \Rightarrow \qquad \max_{q^{\in}\mathcal{Q}^*}\sum_s \widehat{v}_s(1-q_s).$$

Solving this CSP can then be done with whichever method is appropriate given the context and computational resources available to the data provider.

In the context of differential privacy a simple heuristic approach can be used to incorporate the valuations. Given a previously established privacy budget, ε, the choice at hand is how to allocate this budget across the P potential publications. Assuming the data provider wishes to exhaust the entire budget on these publications, and not leave any privacy budget left for future publications, a natural approach is to simply allot the budget in proportion to the estimated valuations. Specifically, publication p would be given $\widehat{v_p}$ of the budget, i.e. for that publication use the established mechanism that provides $\widehat{v_p}\varepsilon$ differential privacy. The construction of the estimated valuations, which ensure $\sum_p \widehat{v_p} = 1$, coupled with the composability of independent mechanisms means that the joint mechanism that produces all publications together will satisfy ε differential privacy. This allocation will also imply that publications with higher valuations will have more of the privacy budget, less noise injected, and therefore more accuracy with respect to the underlying truth. In other words, those statistics deemed more valuable will be reported more accurately.

4 Application

In this section we take a real world data product and apply the entire process, from the estimation of preferences over statistics to the incorporation of those estimates into SDC methods. Though

this framework is not currently in use for the data set of focus, this application illustrates the feasibility and benefits of the framework described earlier.

4.1 Census of Fatal Occupational Injuries

For this application we focus on the Bureau of Labor Statistics's Census of Fatal Occupation Injuries (CFOI). The CFOI is an annual census of workplace fatalities that is compiled via a joint partnership between states and the federal government. Since 1992 this effort has included the collection of documents from numerous sources (e.g., death certificates, coroner reports, OSHA reports, and public news articles), verification of information, assessment for inclusion, and standardization of the accumulated data. With this microdata BLS publishes annual statistics related to the incidence of workplace fatalities across a number of dimensions including geography, injury characterization, industry, occupation, and demographics. The BLS also offers a comprehensive public data query tool[1] where individuals can query the microdata along custom dimensions that are not included in the typical annually published tables. For more details about the CFOI's scope and collection see the Handbook of Methods (Bureau of Labor Statistics 2020)

Protecting the confidentiality of the underlying CFOI microdata poses unique challenges that make it a prime candidate for studying disclosure risk. Beyond just being mandated by CIPSEA, mitigating disclosure risk is especially important since the context surrounding a workplace fatality can sometimes include sensitive health information. The commonplace reporting of fatalities in public new sources adds additional risk even when those news articles do not reveal specifics about the decedent. The risk is further elevated by the relative rarity of these events. For example, tabulations across seemingly broad dimensions can still result in table cells with small counts which can be potentially disclosive. Finally, while many statistics produced from other data products at the BLS involve statistical transformations which can help mitigate disclosure worries (e.g., occupational wage estimates are computed using weighted averages across sampled individuals), the statistics produced using CFOI are typically counts or simple unweighted incidence rates.

4.2 Google Analytics Data

To estimate preferences over statistics we will be leveraging Google Analytics (GA) data across numerous BLS websites which contain published statistics from the CFOI. GA is a resource provided by Google which tracks website traffic over several dimensions and provides various methods of querying this information so businesses and institutions can better understand when, where, and how their sites are being visited.

Figure 1 illustrates an example of data that can be extracted via GA. This figure plots the weekly count of unique views of a webpage that presents a table of fatality counts across industry and the event that resulted in the fatality.[2] Each times series represents the table for a different

[1] `https://data.bls.gov/cgi-bin/dsrv?fw`

[2] An example of this table for the most recent reference year, 2022, can be seen here: `https://www.bls.gov/iif/fatal-injuries-tables/fatal-occupational-injuries-table-a-1-2022.htm`

reference year. For example, we can see that for fatalities which occurred in reference year 2020 the table was first published at the end of 2021 where the time series begins collecting information. This figure alone depicts how pageviews might convey data preferences since we can see a sharp decline in visits for any reference year as soon as any new year's statistics are first published, suggesting a strong preference for the most recent year's statistics. This stylized fact coupled with the variation in presented information across CFOI webpages is what motivates the use of pageviews to estimate preferences over the data published.

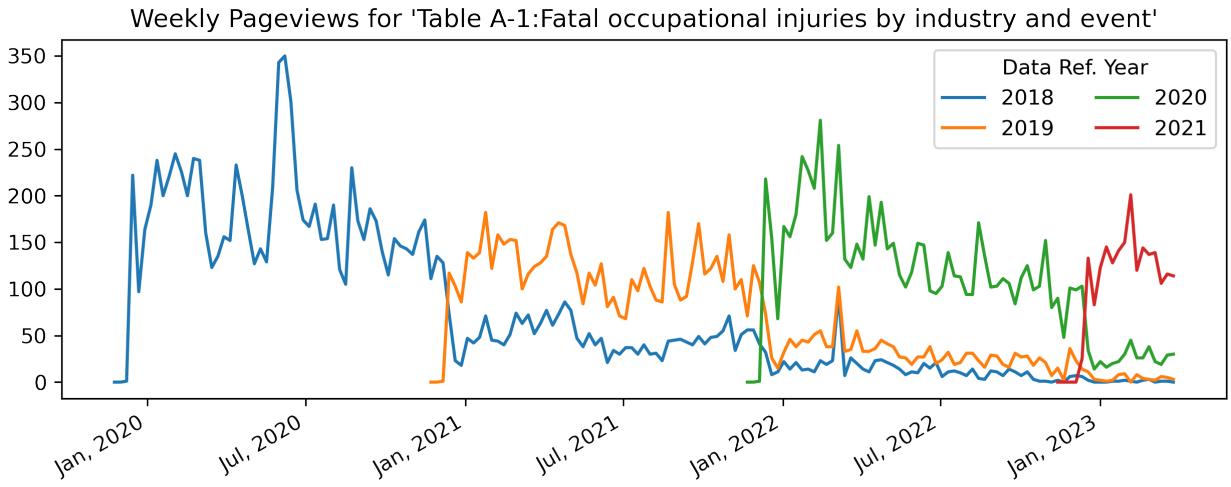


Figure 1: Google Analytics Data Example

For the purposes of this project we collected weekly counts of unique visitors to 28 different types of CFOI publications across more than 50 URLs from December 2017 to April 2023. Given that URLs have changed over time during website redesigns, ample effort was put into stitching together the site visitation histories for the same presented statistics across different URLs. For every URL the various characteristics of both the individual statistics and the entire presentation, i.e. publication, were manually coded. See Table 1 for a complete list of the characteristics that were constructed for this analysis.

4.3 Preference Estimation

Each website can be thought of as a collection of CFOI statistics that can be chosen and consumed once the page is visited. Therefore, each specific website is considered to be a publication as described in the model of section 2. The set of potential pages that can be chosen to visit at any given time will then constitute the market for our purposes. More specifically, a market is defined to be all identified BLS URLs within a given week that present CFOI statistics, which gives us a total of 277 markets across the time horizon of this estimation exercise. Nests are defined such that each represents a set of URLs that are all accessible from a common parent website. This results in the following 3 nests:

Variable (Characteristic) Descriptions

Variable	Level	Description
Emp. Status	statistic	Indicates conditioning on the employment status
Industry	statistic	Indicates conditioning on broad industry categories
Occupation	statistic	Indicates conditioning on the decedent's occupation
Gender	statistic	Indicates conditioning on the decedent's gender
Injury Event	statistic	Indicates conditioning on the broad injury event (event OIICS code)
Age	statistic	Indicates conditioning on the decedent's age
Geo. State	statistic	Indicates conditioning on the decedent's state of residence
Race/Ethnicity	statistic	Indicates conditioning on the decedent's race/ethnicity
Doc. Source	statistic	Indicates conditioning on the document sources used in collection
Stat. Type	statistic	Indicates if it is a count, rate, or percent (count is the base case)
Ref. YYYY	statistic	Indicates if conditioning on CFOI reference year YYYY
Current Year	statistic	Indicates conditioning on the most current reference year
Form	publication	Indicates if presented as a table, bar chart, or line chart (table is the base case)
Multiple Years	publication	Indicates if it includes data from multiple CFOI ref. years
Num. Cells	publication	The number of potential statistics included (i.e. includes suppressed cells)
Num. Stats.	publication	The number of published statistics included (i.e. omits suppressed cells)

Table 1: List of variable descriptions included in the estimated models the object (statistic vs publication) that each variable is associated with.

- Nest 1: Primary tables of CFOI statistics.[3]
- Nest 2: Interactive charts of CFOI statistics.[4]
- Nest 3: Other CFOI tables presented through BLS economic news releases.[5]

By using common parent pages to define nests, this allows the preference model to capture association across publication choices that is driven by common navigation routes throughout the BLS website.

Estimation of the preference parameters is done using instrumental variables regression after leveraging the inversion property described in equation (4). Given that the structure of the chosen

[3]See examples of included publications linked here: `https://www.bls.gov/iif/fatal-injuries-tables.htm`

[4]See examples of included publications linked here: `https://www.bls.gov/charts/census-of-fatal-occupational-injuries/`

[5]See examples of included publications linked here: `https://www.bls.gov/news.release/cfoi.nr0.htm`

tables and charts have been relatively static over time, we can safely assume that all characteristics of the publications are exogenous and do not require instrumentation. However, the nest share variable, $\ln s_{p|g}$, is endogenous by construction. As is common practice for the estimation of nested logit models, we use the number of choices in a nest as the instrument for the nest share to identify σ.

A selection of parameter estimates is presented in Table 2. Two sets of estimates are provided, with the only difference being that model (2) estimates heteroskedasticity-robust standard errors while model (1) does not. All included characteristics are statistically significant, and the estimate for σ falls within the interval $(0, 1)$ without enforcing a constraint on the estimation. Direct interpretation of the magnitudes of these estimated preference parameters is difficult given the nature of utility. However, their ordinality does communicate several interesting aspects of the preferences of the average consumer of CFOI statistics. For example, regarding which dimensions are of highest interest to slice CFOI statistics along these estimates tell us consumers prefer (from highest to lowest) employment status, gender, industry, geography, event, occupation, race, and then finally age. They also prefer count based statistics over percentages or rates, and tables over

Table 2: Preference Estimation Results

	(1)	(2)
Employment Status	0.518**	0.518**
	(0.037)	(0.038)
Industry	-0.112**	-0.112**
	(0.030)	(0.033)
Occupation	-0.459**	-0.459**
	(0.034)	(0.036)
Gender	0.069	0.069
	(0.072)	(0.058)
Injury Event	-0.453**	-0.453**
	(0.025)	(0.029)
Age	-0.525**	-0.525**
	(0.042)	(0.038)
Geo. State	-0.350**	-0.350**
	(0.088)	(0.107)
Race/Ethnicity	-0.509**	-0.509**
	(0.034)	(0.034)
Document Source	-0.958**	-0.958**
	(0.054)	(0.061)

** indicates significance at the 0.01 level.

	(1)	(2)
Type: Percent	-0.482**	-0.482**
	(0.075)	(0.062)
Type: Rate	-0.179**	-0.179**
	(0.036)	(0.034)
Form: Bar Chart	0.337**	0.337**
	(0.051)	(0.060)
Form: Line Chart	0.660**	0.660**
	(0.058)	(0.051)
Multiple Years	-0.274**	-0.274**
	(0.021)	(0.025)
Current Year	1.170**	1.170**
	(0.059)	(0.080)
Nest Shares	0.358**	0.358**
	(0.029)	(0.037)
Constant	-3.308**	-3.308**
	(0.182)	(0.118)
Mkt FE	Yes	Yes
Robust SE	No	Yes
N	5870	5870

** indicates significance at the 0.01 level.

bar charts or line charts. Finally we see a strong preference for data about the most current year's fatalities which mirrors the stylized fact depicted in Figure 1.

4.4 Statistical Disclosure Control

Turning to statistical disclosure control, we now demonstrate how the preference parameter estimates of the previous section are employed. For both cell suppression and differential privacy we will illustrate this using a pair of potentially publishable tables displayed in Table 3. Though the data contained in these tables is entirely fictitious, they represent a simplified version of the types of publications about CFOI that BLS currently produces. Table 3a collects counts of fatalities by age bin and year while Table 3b zooms in on reference year 2021 and collects fatality counts by age and industry.

If published in their current state, there is potential disclosure risk since some of the small cells may result in the identification of the underlying decedents. For example, knowing there was exactly one occupational fatality in 2021 with a decedent under the age of 20 may, when coupled with other public information, be sufficient to identify this individual. To mitigate this risk we will consider the two SDC methods, suppression and differential privacy, as discussed in Section 3.

Table 3: Fictitious Potential Publications

(a) Fatalities by Age and Year

	Year			
Age	2019	2020	2021	Total
< 20	2	3	8	13
20-34	29	27	34	90
35-54	51	46	55	152
≥ 55	49	43	57	149
Total	131	119	154	404

(b) Fatalities by Age and Industry

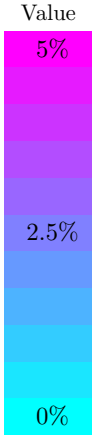
	Industry (for 2021)			
Age	Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

Starting with the suppression approach to SDC the first step is to identify cells for primary suppression. Suppose the disclosure methods dictated that primary suppressions must be applied to any cell with fewer than 4 fatalities, then 4 of the 40 cells would need to be suppressed. As discussed in section 3.1, additional complementary suppressions are required to protect the suppressed cells, however there is flexibility about which cells could be used. Using the estimated preference parameters we can estimate the valuation of each cell according to (7). As a reminder, these estimated valuations are derived from pretending a consumer were given the option to choose 1 out of these 40 statistics (i.e. each is considered to be a separate publication) and defining the estimated valuation as the expected share of consumers to choose that cell. Table 4 presents a visual representation of the estimated valuations of each cell that are implied by the previously estimated preference parameters. Various stylized facts are revealed, such as the higher value placed

Table 4: Potential Publications with Cell Valuations

(a) Fatalities by Age and Year

	Year			
Age	2019	2020	2021	Total
< 20	2	3	8	13
20-34	29	27	34	90
35-54	51	46	55	152
≥ 55	49	43	57	149
Total	131	119	154	404



(b) Fatalities by Age and Industry

	Industry (for 2021)			
Age	Cons.	Mfg.	Trade	Total
< 20	4	3	1	8
20-34	15	9	10	34
35-54	29	15	11	55
≥ 55	31	12	14	57
Total	79	39	36	154

on marginal cells versus inner joint cells in either table, and also higher value being placed on the finer industry breakdown in Table 4b as compared to the aggregated breakdown in Table 4a.

Inserting these valuations into the CSP optimization defined in (5) is fairly straightforward. For a problem of this size, nearly any optimization method can be employed to arrive at an optimal set of complementary suppressions given the cell valuations. While a solution can be easily found, careful observation will reveal that there are in fact multiple optima in this case. This stems from the fact that cell valuations are constant across sets of cells which leads to several options for complementary suppressions that are each optimal. For example in Table 4b, to protect the (Trade Industry, Age < 20) cell we can pick either (Trade Ind., Age 20-34), (Trade Ind., Age 35-54), or (Trade Ind., Age ≥ 55) for complementary suppression since they all have identical estimated valuations. Rather than a flaw in the design, this highlights that the granularity in which you are able to estimate valuations will depend on the variation across publications that exists in the data set used to estimate the preference parameters. While the CFOI publications covered in the GA data do contain tables that breakdown fatalities by age bin, there are not separate publications that look at each age bin so we are unable to identify and estimate a separate valuation by age group. Future work that utilizes requests for data from the CFOI public query tool, which is highly specific, would be able to estimate more granular preference parameters and provide more variation across the estimated valuations.

Next we turn to differential privacy as an alternative SDC method for protecting Tables 3a and 3b. Starting with the assumption that the data provider has already determined the value of ε that encodes their acceptable level of aggregate risk, the choice to be made concerns how to allocate this privacy budget across these potential applications. For pedagogical simplicity suppose the data provider plans to treat each table as a separate publication and is considering how much noise to inject into the cells of each table.

As with the cell suppression approach, we can construct the valuations of each table via our hypothetical market, and we find that Table 3a and Table 3b have respective valuations of 0.361 and 0.639 respectively. Note that the valuation of a table is not the sum of the valuations of the

cells due to the nonlinearity of equation (7). Following the allocation approach described in Section 3.2, we then protect Table 3a using a mechanism with privacy budget 0.361ε and protect Table 3b using a privacy budget of 0.639ε.

The exact mechanism leveraged can be chosen according to the constraints and context. A rudimentary mechanism would add noise using a Laplace distribution, $Lap(\lambda)$, parameterized by λ with density $f_{Lap}(x) = \frac{1}{2\lambda} \exp\{\frac{-|x|}{\lambda}\}$. Specifically, we could add independent draws from $Lap(\frac{1}{0.361\varepsilon})$ to the cells of the first table and draws from $Lap(\frac{1}{0.639\varepsilon})$ to the cells of the second table. This specific mechanism would yield a few oddities such as potentially negative cell values and inconsistencies across totals and tables. However, other more complex ε-DP mechanisms for adding noise to contingency tables exist (see **?** and **?**) which could be applied using the privacy budget allocations specified here.

5 Conclusion

In this paper we have tackled the question of how to improve the implementation of statistical disclosure control methods by incorporating consumer valuations for the statistics being produced. Our focus is on informing decisions surrounding the intensive margin of SDC methods, e.g. which complementary cells to suppress or how to allocate a privacy budget, rather than on the extensive margin, e.g. which cells to mark for primary suppression or how to choose the overall privacy budget ε. We introduced a simple model of consumer choice over publications, consisting of sets of statistics, that can be estimated and used to produce expected valuations over new potential statistics or publications. The incorporation of these valuations into SDC methods was presented for two common SDC approaches: suppression and differential privacy. Finally we supplied an application using a real-world data set, the Census of Fatal Occupational Injuries, where pageview consumption patterns were leveraged to demonstrate a potential data source for preference estimation, and how the estimated preferences could be used to find optimal solutions to the Cell Suppression Problem or allocate a given privacy budget across potential publications.

Given the relatively nascent field of research studying how to better account for the consumer's value of published information in SDC decisions, there are many more facets of this discussion to investigate and ways to further extend the relatively simple framework introduced here. Though we took cell suppression and differential privacy as prime examples of SDC methods, there are many other existing methods for which the presented framework might be applied or adapted. It would also be worthwhile to study other sources of information from which we can elicit estimated valuations of published data, such as the distribution of published empirical work using data sets or the analysis of public social media posts that mention statistics published by a data provider.

References

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 25(2):242–262.

Bureau of Labor Statistics (2020). Handbook of Methods: Census of Fatal Occupational Injuries (CFOI). `https://www.bls.gov/opub/hom/cfoi/`.

Bureau of Labor Statistics (2024). Table A-31: Unemployed persons by industry, class of worker, and sex. `https://www.bls.gov/web/empsit/cpseea31.htm`. Accessed 2024-04-28.

Cardell, N. S. (1997). Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity. *Econometric Theory*, 13(2):185–213.

Castro, J. (2012). Recent advances in optimization techniques for statistical tabular data protection. *European Journal of Operational Research*, 216(2):257–269.

Castro, J. (2023). Thirty years of optimization-based sdc methods for tabular data. *Transactions on Data Privacy*, 16(1):3–13.

Census Bureau (2021). Census bureau sets key parameters to protect privacy in 2020 census results. `https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html`. Press Release: 2021-06-09.

Dwork, C. (2008). *Differential Privacy: A Survey of Results*, pages 1–19. Springer Berlin Heidelberg.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). *Calibrating Noise to Sensitivity in Private Data Analysis*, pages 265–284. Springer Berlin Heidelberg.

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

Fischetti, M. and Salazar, J. J. (2001). Solving the cell suppression problem on tabular data with linear constraints. *Management Science*, 47(7):1008–1027.

Kelly, J. P., Golden, B. L., and Assad, A. A. (1992). Cell suppression: Disclosure protection for sensitive tabular data. *Networks*, 22(4):397–417.

US Congress (2002). Confidential Information Protection and Statistical Efficiency Act. `https://www.bls.gov/bls/cipsea.pdf`.