Suboptimal Provision of Privacy and Statistical Accuracy When They are Public Goods

J. M. Abowd, I. M. Schmutte, W. Sexton, L. Vilhuber   Correspondence: william.n.sexton@census.gov

I. Goal and Contribution

Goal: To explain why population statistics are provided by public statistical agencies rather than private firms

To do so, we focus on inefficiencies in how private providers trade off data privacy and accuracy.

- Increasing the accuracy of published statistical summaries necessarily results in a loss of privacy for the data owners.
- Data publication is based on differential privacy.
- Privacy protection and accuracy are public goods.

We find that private provision results in suboptimally low data accuracy.

- The external benefit of data accuracy to all consumers is not captured by the willingness-to-pay of the consumer with the greatest private value.
- The provider buys just enough data-use rights (privacy loss) to sell the data accuracy to the consumer with the highest valuation.

II. Modeling Privacy and Accuracy

ε-differential privacy:

Definition 1. *Query release mechanism M satisfies ε-differential privacy if for $\varepsilon > 0$, for all pairs of neighboring databases D, D', all queries $Q \in \mathcal{Q}$, and all $B \in \mathcal{B}$*

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

where $\mathcal{B}$ are the measurable subsets of $\mathbb{R}$, and the randomness in M is due exclusively to the mechanism.

(α, β)-accuracy:

Definition 2. *Query release mechanism M satisfies (α, β)-accuracy if for $Q \in \mathcal{Q}$ and a output from $M(D,Q)$,*

$$Pr\Big(|a - Q(D)| \leq \alpha \;\Big|\; D, Q \Big) \geq 1 - \beta$$

where $a, Q(D) \in \mathbb{R}$.

III. Model (Consumer)

There are N private individuals:

- each possesses a single bit of information, b_i, and is endowed with income, y_i.
- each consume one unit of the published statistic, which has accuracy $I = (1-\alpha)$. Each is charged at the market price p_I, for her "share" of I, denoted I_i.
- preferences are given by the indirect utility function

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

The term p_ε is the common price per unit of privacy. And, $(\eta_i, \gamma_i) > 0$, are the individual's marginal preferences for data accuracy and privacy loss and are not known to the data provider, but their population distributions are public information.

IV. Model (Producer)

Ghosh and Roth (2015) prove that publishing

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

gives an (α, β)-accuracy estimate of the population mean, $\bar{b}$, requiring privacy loss $\varepsilon_i = \varepsilon(I) = \frac{1/2 + \ln(1/\beta)}{(1-I)N}$ from $H(I) = N - \frac{(1-I)N}{1/2 + \ln(1/\beta)}$ members of the population.

- Purchasing data-use rights from those with the smallest γ_i, is a minimum-cost, envy-free VCG mechanism.

Under said VCG mechanism, the total cost of producing I is

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

where Q is the quantile function with respect to the population distribution of privacy preferences, F_γ.

V. Competitive Market Equilibrium

A private profit-maximizing, price-taking, firm sells $\hat{s}$ with data accuracy I at price p_I. Then, profits $P(I)$ are

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

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

$$p_I = Q\left(\frac{H(I)}{N}\right) H(I)\varepsilon'(I) + \left[Q\left(\frac{H(I)}{N}\right) + Q'\left(\frac{H(I)}{N}\right)\left(\frac{H(I)}{N}\right)\right] H'(I)\varepsilon(I) \qquad (1)$$

where the solution is evaluated at I^{VCG}. As long as the cost function is strictly increasing and convex, the existence and uniqueness of a solution is guaranteed.
At market price p_I, consumer i's willingness to pay for data accuracy will be given by solving

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

Consumers are playing a classic free-rider game.

1. the only person willing to pay for the public good is one with the maximum value of η_i.
2. all others will purchase zero data accuracy but still consume the data accuracy purchased by this lone consumer.

Hence, equilibrium price and data accuracy will satisfy

$$p_I = \bar{\eta} = \frac{dC^{VCG}\left(I^{VCG}\right)}{dI}, \text{ where } \bar{\eta} = \max \eta_i.$$

However, the Pareto optimal consumption of data accuracy, I^0, solves

$$\sum_{i=1}^{N} \eta_i = \frac{dC^{VCG}\left(I^0\right)}{dI}. \qquad (2)$$

Marginal cost is positive, $\frac{dC^{VCG}(I^0)}{dI} > 0$, and $\sum_{i=1}^{N} \eta_i > \bar{\eta}$; therefore, data accuracy will be under-provided by a competitive supplier when data accuracy is a public good as long as marginal cost is increasing. More succinctly, $I^{VCG} < I^0$. Therefore, privacy protection must be over-provided, $\varepsilon^{VCG} < \varepsilon^0$.

VI. Suboptimality Theorem

Theorem 1. *If preferences are as in III, the query response mechanism and cost function for the VCG mechanism are as displayed in IV, the population distribution of γ is given by F_γ (bounded, absolutely continuous, everywhere differentiable, and with quantile function Q satisfying the conditions such that (1) has a solution), 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 $I^{VCG} < I^0$, where I^0 is the Pareto optimal level of I solving equation (2), and I^{VCG} is the privately-provided level when using the VCG procurement mechanism.*