PLOS ONE

RESEARCH ARTICLE

A Statistical Approach to Provide Individualized Privacy for Surveys

Fernando Esponda[1]*, Kael Huerta[2], Victor M. Guerrero[2]

1 Computer Science Department, Instituto Tecnológico Autónomo de México (ITAM), Mexico City, Mexico, **2** Statistics Department, Instituto Tecnológico Autónomo de México (ITAM), Mexico City, Mexico

* fernando.esponda@itam.mx

Abstract

In this paper we propose an instrument for collecting sensitive data that allows for each participant to customize the amount of information that she is comfortable revealing. Current methods adopt a uniform approach where all subjects are afforded the same privacy guarantees; however, privacy is a highly subjective property with intermediate points between total disclosure and non-disclosure: each respondent has a different criterion regarding the sensitivity of a particular topic. The method we propose empowers respondents in this respect while still allowing for the discovery of interesting findings through the application of well-known inferential procedures.

Citation: Esponda F, Huerta K, Guerrero VM (2016) A Statistical Approach to Provide Individualized Privacy for Surveys. PLoS ONE 11(1): e0147314. doi:10.1371/journal.pone.0147314

Editor: Gang Han, Texas A&M University, UNITED STATES

Received: September 28, 2015

Accepted: December 31, 2015

Published: January 29, 2016

Data Availability Statement: All the data are available from the Graphic, Visualization, & Usability Center (GVU) database http://www.cc.gatech.edu/gvu/user_surveys/survey-1997-10/datasets.

Funding: This work was supported by Asociación Mexican de Cultura A.C.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Globalization and the internet era have brought with them a huge array of opportunities for data driven statistical applications and data driven science. An increasing volume of digital information is stored about individuals: their preferences, diseases, relationships, and even their current location; and an even greater amount about phenomena ranging from traffic to radioactivity that is picked up by sensors. Its uses provide great benefits to governments, scientists, health specialists and marketers alike, but by the same token, it has made the preservation of privacy a more urgent matter: data are long-lived, ubiquitously accessible, and with the advent of Big Data mining, exploitable in unimaginable ways. The benefits of the widespread efforts for data collection and its privacy related challenges are well captured by the President's Council of Advisors on Science and Technology's report on Big Data and Privacy [1].

Surveys are a useful recourse for collecting data in a directed fashion, be it from individuals or from machines. One important challenge experimenters face is when the data to be collected are sensitive in nature, as the subjects might refuse to participate or could participate with a strong response bias: imagine collecting data related to venereal diseases, or the radioactivity levels in a certain geographical area; or the speed at which your car is being driven. Additionally respondents should be reasonably protected from potentially harmful and unexpected uses of their information.

There are a few recourses to safeguard privacy in such scenarios, among which we have: anonymity, cryptography, and information reduction techniques. Anonymity is a method in

which the de-identification of the respondent is guaranteed from the onset; the survey contains all the relevant data except that which can be used to associate the answers to a specific interviewee. The drawbacks of this technique lie in the difficulty of providing such guarantees in an effective and believable manner, respondents might still be weary or even incapable of answering a very sensitive question, and researchers forgo the possibility of conducting longitudinal studies. Anonymity can also be provided after the data collection has taken place (see [2, 3]). The objective of these techniques is to anonymize (de-identify) the sensitive data and allow them to be disclosed; its disadvantage from the vantage point of surveying—during data collection—is that it doesn't provide many guarantees to entice truthful participation.

Cryptography based methods are generally applied after the survey has been conducted (with the exception of multiparty computation techniques, which we classify in the anonymity group [4]) and their aim is to ensure that the survey data—including the respondent's information—can't be examined but by authorized parties. However, authorizations change over time and cryptographic keys can be stolen, misplaced or misused; this, together with the increasing lifetime expectancy of data, makes their long-lived privacy unlikely. From the surveying standpoint cryptographic techniques are hard to explain and therefore to trust by average individuals and, as is the case for anonymity, respondents still have to answer the sensitive question directly (see [5, 6] for examples).

Finally, information reduction techniques are used during the application of the survey and work by requiring less information from the interviewed, enough to compute population statistics but not enough to impute specific sensitive answers to specific respondents. In this way surveys are de-sensitized and respondents can provide their identification data for longitudinal studies without the fear that their answers will come back to haunt them (for examples see [7–14]). Their main disadvantage is that they are not applicable to every kind of survey; that by collecting less information they require bigger samples to maintain accuracy; and that they use a one-size fits all privacy scheme, which squanders information that some respondents may be willing to surrender and forces the more hesitant ones to bias their participation or response. Additionally, techniques such as randomized response techniques and negative surveys have suffered from successfully explaining to respondents how the survey should be answered. However, we believe that this shortcoming is quickly being surpassed by the widespread use of electronic devices that collect data in such a way that the complexity and awkwardness of randomizing devices is hidden from respondents. Furthermore an increasing amount of sensitive data are being collected from sensors to which these techniques can be applied transparently [15–18].

In this paper we focus on an information reduction technique that addresses the fact that the sensitivity of a question or topic is a subjective matter and allows different respondents to disclose a different amount of information for the same question. Our method is a generalization of the Negative Survey technique [10]; we present the surveying technique as well as some of its key statistics and leave the specifics of a survey design outside the scope of this work. We consider that our instrument is appropriate for collecting data from people and from devices, and that it can be applied straightforwardly to the latter but that much work is needed to make its guarantees clear and its administration transparent to the former. Furthermore, we believe that this technique can be successfully employed for answering database queries in a private fashion (where the respondents are the individual fields of each database entry) and thus used for reducing the privacy concerns of already collected data while preserving some of their value. We provide a simulation study using a publicly available database in order to show the accuracy of the technique and how it could be used to collect data—simulating database entries as respondents—or to disclose data sources that contain sensitive attributes.

In Section we briefly explain the Negative Survey technique and follow, in Sections and, with a generalization that allows the experimenter to set the level of privacy to the survey and with a scheme that enables the respondent to decide on its own on the appropriate level. In Section we provide a simplified method for computing the relevant statistics of our instrument for a special design case, which we believe will be widely applicable, and in Section we introduce our instrument that empowers each participant to elect the amount of information to disclose. Section presents the results of our simulation study using real data and we finish with a discussion of the current work and some of its possible directions.

Negative Surveys

Negative surveys, introduced in [10], is a method for applying a multiple choice questionnaire with t exhaustive and mutually exclusive categories—see [19, 20] for refinements on the technique and [21–25] for some applications. The technique is useful when the query in question is sensitive in nature and might cause response and non-response bias. In essence, the approach consists of negating the original question and having respondents choose one among the $t - 1$ options that now apply to him/her with a known probability distribution (see the example in Fig 1). Negative surveys provide a scheme that is expected to help reduce response and non-response biases and that will safeguard sensitive information in the most secure way: by not collecting it in the first place.

Note that not all positive questions can be translated straightforwardly to negative questions as some categories might be sensitive in themselves; for example, when asked "How many sexual partners have you **not** had?" the category "Between 0 and 2" reveals more than some respondents might like. In such scenarios the sensitive category should be replaced with a *dummy* category, a sink so to speak, and the design matrix (see Section) adjusted accordingly.

What is interesting about this method is that even though the negative version provides less information about each respondent, meaningful population statistics can still be estimated. It enables an experimenter to learn something about the population without being able to impute a sensitive answer to a particular individual. However, sensitivity to questions is a relative matter as not everybody places the same burden on the same topic. In the following section we generalize the negative survey scheme in such a way that respondents can decide how much to reveal allowing experimenters to take advantage of the information that is willingly provided.

The amount of money you owe is:

[ ] Between 0 and 2,000

[ ] Between 2,001 and 5,000

[ ] Between 5,001 and 7,000

[ ] 7,001 or more

(a) Positive Question

The amount of money you owe is **not**:

[ ] Between 0 and 2,000

[ ] Between 2,001 and 5,000

[ ] Between 5,001 and 7,000

[ ] 7,001 or more

(b) Negative Question

Fig 1. A positive question and its corresponding negatively framed question. Respondents are asked to choose only one option in both cases.

doi:10.1371/journal.pone.0147314.g001

Interviewer Defined Privacy: Multiple-answer Questionnaire

In this section we generalize the one-answer negative survey model—where one and only one of the categories must be chosen—to the case in which a respondent chooses k_0 of the available options. We discuss the case in which the corresponding positive setup has t exhaustive and mutually exclusive alternatives and where k_0 is previously fixed by the experimenter taking any value between 1, corresponding to the one-answer model, and $t - 1$, corresponding to a positive survey. By letting k_0 vary we have a variety of models for the same question, each affording a different amount of privacy. This scheme may be suitable for disclosing data that were previously collected but for which we wish to provide a certain, uniform level of obfuscation. Later, in Section, we extend this design to allow for a record (or respondent) level of privacy.

As with the one-answer scheme, let X be a random variable denoting the category to which the respondent truly belongs—and does not wish to fully disclose. Let $\pi_j = P(X = j)$ be the probability that X takes on the value j with $j \in \{1, 2, \ldots, t\}$, $\sum_{j=1}^{t} \pi_j = 1$ and $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_t)^T$.

Let Y be a random variable denoting the k_0 categories that the respondent has revealed *not* to belong to. This variable takes its values from the set of all combinations of t values taken k_0 at a time. We refer to this space as $\boldsymbol{\Omega}_{k_0}$ with cardinality $\alpha = \binom{t}{k_0}$ and denote each element of $\boldsymbol{\Omega}_{k_0}$ as $\bar{\omega}_i = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{ik_0})$. Each ω_{ir}, with $r = 1, 2, \ldots, k_0$, refers to a category that has been discarded by the respondent and $\bar{\omega}_i$ to the set of all simultaneously discarded categories—the respondent's answer to the negative survey. Finally exactly one of the events $\{Y = \bar{\omega}_i\}$ occurs for each application of the negative survey with probability λ_i such that $\sum_{i=1}^{\alpha} \lambda_i = 1$.

Consider n independent repetitions of the experiment and let N_i be the random variable denoting the number of occurrences of $\{Y = \bar{\omega}_i\}$, then $\sum_{i=1}^{\alpha} N_i = n$. Together they constitute the random vector $\boldsymbol{N}$ which follows a multinomial distribution with parameters n and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_\alpha)^T$, i.e., $\boldsymbol{N} \sim$ Multinomial$(n, \boldsymbol{\lambda})$. We then have

$$\begin{aligned} \mathrm{E}(N_i) &= n\lambda_i \\ \mathrm{Var}(N_i) &= n\lambda_i(1 - \lambda_i) \\ \mathrm{Cov}(N_i, N_j) &= -n\lambda_i\lambda_j \quad \text{if } i \neq j. \end{aligned} \tag{1}$$

The Maximum Likelihood (ML) estimator for λ_i is given by $\hat{\lambda}_i = N_i/n$ and

$$\begin{aligned} \mathrm{E}(\hat{\lambda}_i) &= \mathrm{E}\left(\frac{N_i}{n}\right) = \lambda_i \\ \mathrm{Var}(\hat{\lambda}_i) &= \mathrm{Var}\left(\frac{N_i}{n}\right) = \frac{1}{n}\lambda_i(1 - \lambda_i) \\ \mathrm{Cov}(\hat{\lambda}_i, \hat{\lambda}_j) &= \mathrm{Cov}\left(\frac{N_i}{n}, \frac{N_j}{n}\right) = -\frac{1}{n}\lambda_i\lambda_j \quad \text{if } i \neq j. \end{aligned} \tag{2}$$

Assuming each individual answers truthfully we can write the conditional probabilities as

$$p_{ij} = \begin{cases} 0 & \text{if } i = j \\ P(Y = \bar{\omega}_i | X = j) & \text{otherwise} \end{cases}$$

and by the Law of Total Probability we can see that the probability of obtaining a specific combination $\bar{\omega}_i$, of i categories for all $i = 1, 2, \ldots, t$ is

$$\lambda_i = \sum_{j=1}^{t} P\{Y = \bar{\omega}_i | X = j\}\pi_j = \sum_{j=1}^{t} p_{ij}\pi_j \quad \text{for} \quad i = 1, 2, \ldots, \alpha$$

which we can write in matrix notation as

$$\lambda = \mathbb{P}\pi, \tag{3}$$

where $\mathbb{P}$ is the design matrix with dimension $t \times t$ whose element (i, j) is given by p_{ij}.

Notice that for this set up we have $\alpha = \binom{t}{k_0}$ equations and only t unknowns and thus the system will be overdetermined for $1 < k_0 < t - 1$. We therefore use the Moore-Penrose pseudo-inverse to construct our estimator. Let $\mathbb{P}$ be the design matrix with known conditional probabilities, $\mathbb{P} = U\Sigma V^T$ be its singular value decomposition, and let $\mathbb{P}^\dagger = V\Sigma^\dagger U^T$ be the generalized inverse of $\mathbb{P}$ with its respective singular value decomposition, so that U and V are orthonormal matrices, while Σ is a diagonal matrix whose elements are the singular (nonnegative) values of $\mathbb{P}$. Then if $\hat{\lambda} = (\hat{\lambda_1}, \hat{\lambda_2}, \ldots, \hat{\lambda_\alpha})^T$ with the $\hat{\lambda}_i s$ estimated by ML, we obtain the following result:

Proposition 1. Given the system $\lambda = \mathbb{P}\pi$ where $\mathbb{P} = U\Sigma V^T$, then

$$\hat{\pi} = V\Sigma^\dagger U^T \hat{\lambda}$$

is the unbiased ML estimator for $\boldsymbol{\pi}$ whose variance is given by

$$\mathrm{Var}(\hat{\pi}) = \frac{1}{n} V\Sigma^\dagger U^T \left[\mathrm{Diag}(\lambda) - \lambda\lambda^T\right] \left(V\Sigma^\dagger U^T\right)^T.$$

One disadvantage of this method is that the singular value decomposition could be computationally costly when faced with big design matrices. An alternative method for computing the desired estimator can be used when the design matrix $\mathbb{P}$ has full rank, then $(\mathbb{P}^T\mathbb{P})$ is symmetric positive-semidefinite and we can estimate the population proportions $\boldsymbol{\pi}$ by ML as

$$\hat{\pi} = (\mathbb{P}^T\mathbb{P})^{-1}\mathbb{P}^T\hat{\lambda}$$

with variance

$$\mathrm{Var}(\hat{\pi}) = \frac{1}{n}\left(\mathbb{P}^T\mathbb{P}\right)^{-1}\mathbb{P}^T\left[\mathrm{Diag}\left(\lambda\right) - \lambda\lambda^T\right]\mathbb{P}\left(\mathbb{P}^T\mathbb{P}\right)^{-T}$$

using a computationally more efficient method, such as Cholesky decomposition.

Special Case: Equiprobable Design Matrix

In this section we examine the special case in which each of the $t - 1$ categories from which the respondent can pick is chosen with the same probability to form his/her answer set—for example with the assistance of a randomization device. In this case, assuming individuals answer truthfully and according to instructions, the probability of a respondent choosing a set containing its true category is zero and the probability of selecting a set of size k_0 that does not contain it, is inversely proportional to the number of such subsets

$$p_{ij} = P(Y = \bar{\omega}_i | X = j) = \begin{cases} \dfrac{1}{\binom{t-1}{k_0}} & \text{if } j \notin \bar{\omega}_i \\ 0 & \text{if } j \in \bar{\omega}_i \end{cases} \tag{4}$$

where $\bar{\omega}_i = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{ik_0})$ denotes the possible subsets from which the individual can choose. We again write the probability for each λ_i as

$$\lambda = \mathbb{P}\pi.$$

Let $\bar{\omega}_i' = (\omega_{i1}', \omega_{i2}', \ldots, \omega_{it}')$ be an indicator row vector of dimension t, indicating which categories have been discarded, such that

$$\omega_{ij}' = \begin{cases} 1 & \text{if } j \in \bar{\omega}_i \\ 0 & \text{if } j \notin \bar{\omega}_i \end{cases} \quad \text{for all } j = 1, 2, \ldots, t.$$

We can rewrite the design matrix as

$$\mathbb{P} = \frac{1}{\binom{t-1}{k_0}} \begin{bmatrix} \bar{\omega}_1' \\ \bar{\omega}_2' \\ \vdots \\ \bar{\omega}_\alpha' \end{bmatrix}$$

which yields a system of α equations, with the i^{th} equation described by

$$\lambda_i = \sum_{j=1}^{t} p_{ij}\pi_j = \frac{1}{\binom{t-1}{k_0}} \bar{\omega}_i'\pi, \tag{5}$$

with this in mind we can now find a more direct way to estimate $\boldsymbol{\pi}$.

Let λ_j' with $j = 1, 2, \ldots, t$ be the set of variables defined by:

$$\lambda_j' = \sum_{\{i|j\in\bar{\omega}_i\}} \lambda_i \qquad \text{for all } \bar{\omega}_i \in \Omega_{k_0} \tag{6}$$

that is, λ_j' is formed by the sum of the proportion of all sets that include category j as a member. Substituting in Eq (5) we get a system of t equations in t unknowns

$$\lambda_j' = \frac{1}{\binom{t-1}{k_0}} \sum_{\{i|j\in\bar{\omega}_i\}} \bar{\omega}_i'\pi \quad \text{for} \quad j = 1, 2, \ldots, t.$$

Note that for each equation we are adding each π_i, with $i \neq j$, a total of $\binom{t-2}{t-k_0-1}$ times and we can thus rewrite the above expression as

$$\lambda_j' = \frac{\binom{t-2}{t-k_0-1}}{\binom{t-1}{k_0}} \sum_{h\neq j} \pi_h = \frac{k_0}{t-1}(1-\pi_j)$$

Now let M_j with $j = 1, 2, \ldots, t$ be the number of times some respondents eliminated a set containing category j, then $\lambda_j' = M_j/n$ is proportional to the probability of selecting the j^{th} category. This result is expressed as follows.

Proposition 2. Suppose an equiprobable design for a negative survey with k_0 answers, then an unbiased estimator of π_j is given by

$$\hat{\pi}_j = 1 - \frac{t-1}{k_0}\hat{\lambda}_j' \quad \text{with } \hat{\lambda}_j' = \frac{M_j}{n} \qquad \text{for all } j = 1, 2, \ldots, t$$

with corresponding variance and covariances given by

$$\mathrm{Var}(\hat{\pi}_j) = \frac{1}{n}\left(\frac{t-1}{k_0}\right)^2 \lambda'_j(1-\lambda'_j)$$

$$\mathrm{Cov}(\hat{\pi}_i, \hat{\pi}_j) = -\frac{1}{n}\left(\frac{t-1}{k_0}\right)^2 \sum_{\{r|i\in\varpi_r\}} \sum_{\{s|j\in\varpi_s\}} \lambda'_i\lambda'_j \quad \text{for } i \neq j.$$

We are particularly interested in obtaining a confidence interval for the population proportion of the sensitive category, say the j^{th} category. Since M_j is a random variable that behaves as Binomial(n, λ_j') for each $j = 1, \ldots, t$, we follow the Agresti and Coull's recommendation (see [26]) of using an adjusted Wald interval whose coverage probabilities are closer to the nominal levels than those of the unadjusted interval. The adjustment amounts to adding two successes and two failures when constructing a 95% confidence interval, but in general consists of replacing $\tilde{\lambda}'_j$ by $\tilde{\lambda}'_j = \tilde{M}_j/\tilde{n}$ with $\tilde{M}_j = M_j + z^2_{\alpha/2}/2$ and $\tilde{n} = n + z^2_{\alpha/2}$ where $z_{\alpha/2}$ denotes the upper $\alpha/2$ percentage point of a unit Normal distribution. Thus we deduce that a $100(1-\alpha)\%$ confidence interval for π_j, when n, t and k_0 are fixed, is given by

$$1 - \frac{t-1}{k_0}\left[\tilde{\lambda}'_j \pm z^2_{\alpha/2}\sqrt{\tilde{\lambda}'_j(1-\tilde{\lambda}'_j)/\tilde{n}}\right].$$

We conclude this section by analyzing the variance of our estimator in terms of the variance of the corresponding positive survey and the variance added by using a multiple answer negative survey

$$\begin{aligned} Var(\hat{\pi}_j) &= \frac{1}{n}\left(\frac{t-1}{k_0}\right)^2 \lambda'_j(1-\lambda'_j) \\ &= \frac{1}{n}\left(\frac{t-1}{k_0}\right)^2 \left[\frac{k_0(1-\pi_j)}{t-1}\right]\left(1 - \left[\frac{k_0(1-\pi_j)}{t-1}\right]\right) \\ &= \frac{\pi_j(1-\pi_j)}{n}\left(1 + \frac{t-k_0-1}{\pi_j k_0}\right). \end{aligned}$$

In contrast to the one-answer setup, where a good strategy to control the variance is by keeping the number of categories low, for the multiple-answer version we can improve its accuracy by reducing the difference between the number of possible options and the participation parameter k_0, which is to say, the privacy afforded to respondents. Furthermore, this parameter gives greater control to the experimenter as it allows a tradeoff between sample size, number of options, and privacy.

With the multiple-answer model we have a way for the experimenter to control the privacy of a survey. However, privacy is an inherently subjective measure and having k_0 fixed might oblige some to disclose more than what they are comfortable with (risking bias) and others to disclose less and waste potentially useful information. In the next section we build on the present model to address this issue.

Respondent Defined Privacy: Variable-answer Questionnaire

In the previous sections we examined a setup in which the interviewer established a survey wide level of privacy by specifying the number of options each respondent should choose. In what follows we build on these results to create a design that allows each participant to disclose as much information as he/she is confortable with while still providing useful information to the surveyor.

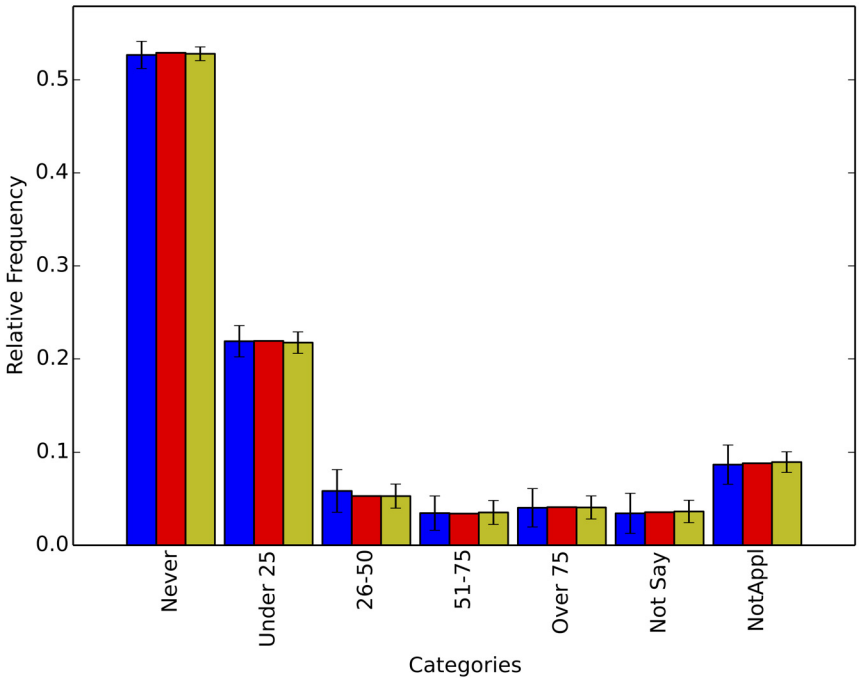
Let $\mathbb{P}_k$ be the design matrix for a multiple-answer negative survey scheme where k categories, the participation parameter, are eliminated by respondents. Let $\boldsymbol{\pi}_k = (\pi_{1k}, \pi_{2k}, \ldots, \pi_{tk})^T$ be the preference vector for the same scheme, where π_{jk} refers to the proportion of individuals that prefer category j, and let $\boldsymbol{\lambda}_k = (\lambda_{1k}, \lambda_{2k}, \ldots, \lambda_{\alpha k})^T$ be the probability vector of observing each combination of eliminated categories. We now obtain the following result.

Proposition 3. Given a personalized negative survey applied to a sample of n respondents and a weight vector $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_{t-1})^T$, the unbiased ML estimator for the preference population proportion $\boldsymbol{\pi}$ is given by

$$\hat{\pi} = \sum_{k=1}^{t-1} \zeta_k \hat{\pi}_k$$

where $\sum_{k=1}^{t-1} \zeta_k = 1$. Its covariance matrix is given by

$$\mathrm{Var}(\hat{\pi}) = \sum_{k=1}^{t-1} \frac{\zeta_k^2}{n} \mathbb{P}_k \left[\mathrm{Diag}(\lambda_k) - \lambda_k \lambda_k^T\right] \mathbb{P}_k^T.$$



Standard deviation results per category

	Never	Under 25	26-50	51-75	Over 75	Not Say	NotAppl	Average
Custom Privacy	0.0074	0.0115	0.013	0.0129	0.0124	0.0121	0.0111	0.0115
$k_0 = 1$	0.0146	0.0168	0.0229	0.0184	0.0206	0.0216	0.0211	0.0194
Reduction	49.45%	31.36%	43.52%	30.23%	39.83%	44.09%	47.31%	40.83%

Fig 2. Responses to the question: Some Web sites ask for you to register with the site by providing personal information. When asked for such information, what percent of the time do you falsify the information? Each category shows, from left to right, the relative frequency for the interviewer defined privacy survey with $k_0 = 1$; the true relative frequency; and the relative frequency for the respondent defined privacy survey with random preference. Each bar represents the average of 100 repetitions and the error bar the standard deviation.

doi:10.1371/journal.pone.0147314.g002

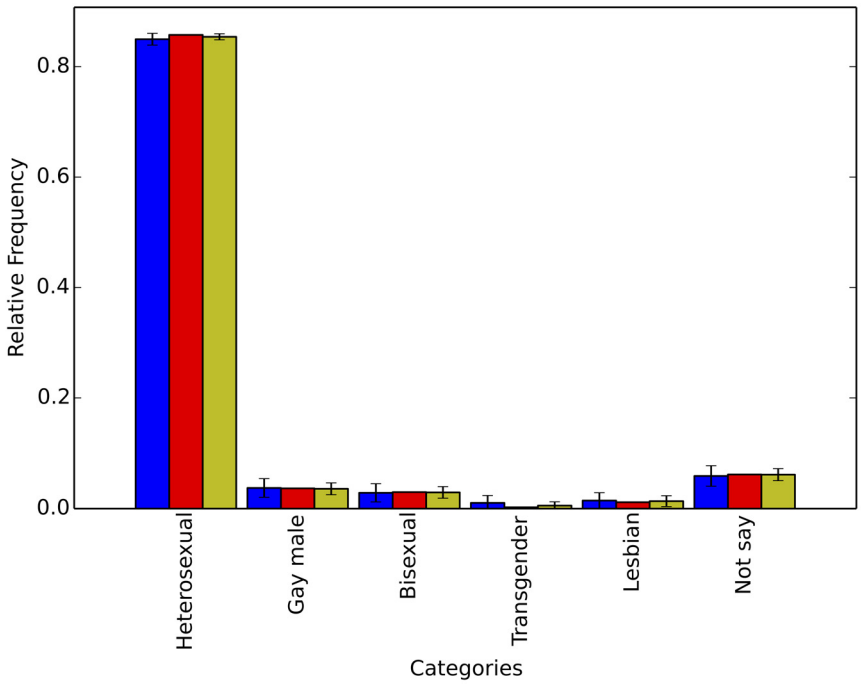
Recall that $\hat{\pi}_k$ is an unbiased estimator for $k = 1, \ldots, t-1$, therefore

$$\text{bias}(\hat{\pi}) = \text{E}\left(\sum_{k=1}^{t-1} \zeta_k \hat{\pi}_k - \pi\right) = \pi \sum_{k=1}^{t-1} \zeta_k - \pi = \mathbf{0}.$$

This models affords great flexibility by allowing the experimenter to weigh each multiple-answer estimator according to different scenarios; for instance:

- Each multi-answer estimator is weighed equally
- Estimators with higher variability, e.g. lower participation parameter, receive lower weight
- The weight is proportional to the number of samples in each multi-answer estimator

and choosing the one with the least variability. The only restriction being that the sum of all weights sum to 1. Once the survey has been conducted the desired proportions $\hat{\pi}$ can be computed as per the methods described in Section.



Standard deviation results per category

	Heterosexual	Gay male	Bisexual	Transgender	Lesbian	Not say	Average
Custom Privacy	0.0055	0.0107	0.0105	0.0068	0.0098	0.0109	0.009
$k_0 = 1$	0.0107	0.0169	0.0165	0.0133	0.0142	0.0186	0.015
Reduction	48.78%	36.51%	36.56%	48.65%	30.81%	41.3%	40.44%

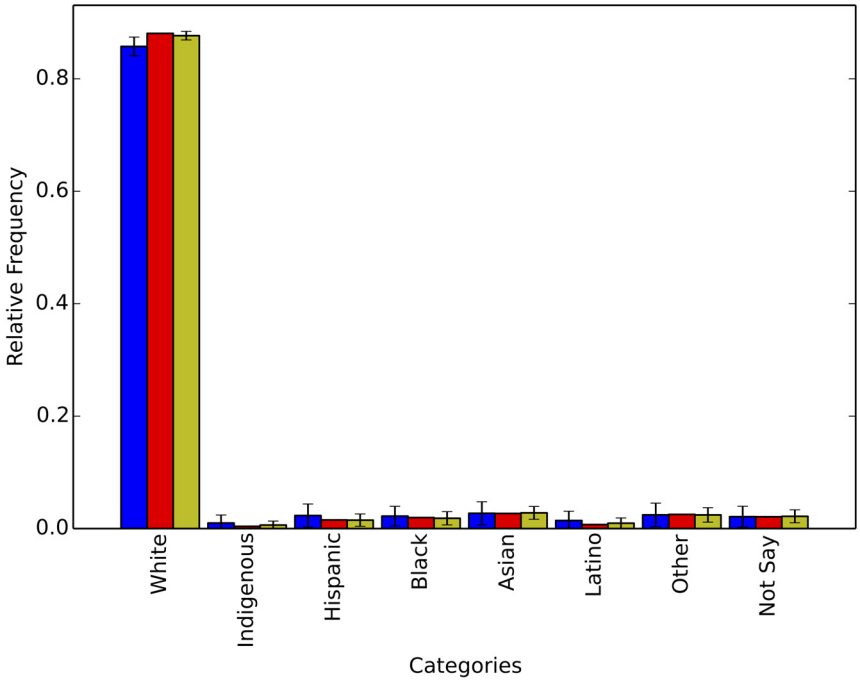
Fig 3. Responses to the question: How would you classify yourself? Each category shows, from left to right, the relative frequency for the interviewer defined privacy survey with $k_0 = 1$; the true relative frequency; and the relative frequency for the respondent defined privacy survey with random preference. Each bar represents the average of 100 repetitions and the error bar the standard deviation.

doi:10.1371/journal.pone.0147314.g003

A Simulated Survey

In this section we simulate a survey to demonstrate the use and results of our method. As discussed in the previous sections, negative surveys may be used to collect sensitive information from a group of people or a group of sensors; in this example we simulate such a collection process by taking a real database, where each record contains the response of an individual to a survey and apply to it a negative questionnaire, in such a way that the actual value for each record is substituted by a set of "negative" categories— categories that do *not* contain the record's actual value. We then use these to estimate the frequency of each category for the database and compare it to the actual distribution.

The following simulations use data from the Graphic, Visualization & Usability Center's (GVU) 8th WWW User Survey, available on-line from [27] which archives the survey responses of 10,108 web users with regards to their general demographic information. Some of the categories are intrinsically sensitive in nature, such as sexual preference and race, and we use them to showcase our technique. We use the variable response setup described in Section and assume a uniform distribution between 1 and $t - 1$ for the privacy preferences —1 being the most private eliminating only 1 category from the possible answers and $t - 1$ the least



Standard deviation results per category

	White	Indigenous	Hispanic	Black
Custom Privacy	0.0077	0.0071	0.011	0.0118
$k_0 = 1$	0.0167	0.0142	0.0206	0.0175
Reduction	54.02%	49.85%	46.52%	32.66%

	Asian	Latino	Other	Not Say	Average
Custom Privacy	0.0114	0.0094	0.0129	0.0115	0.0103
$k_0 = 1$	0.0206	0.0165	0.0209	0.0186	0.0182
Reduction	44.52%	43.35%	38.42%	37.92%	43.41%

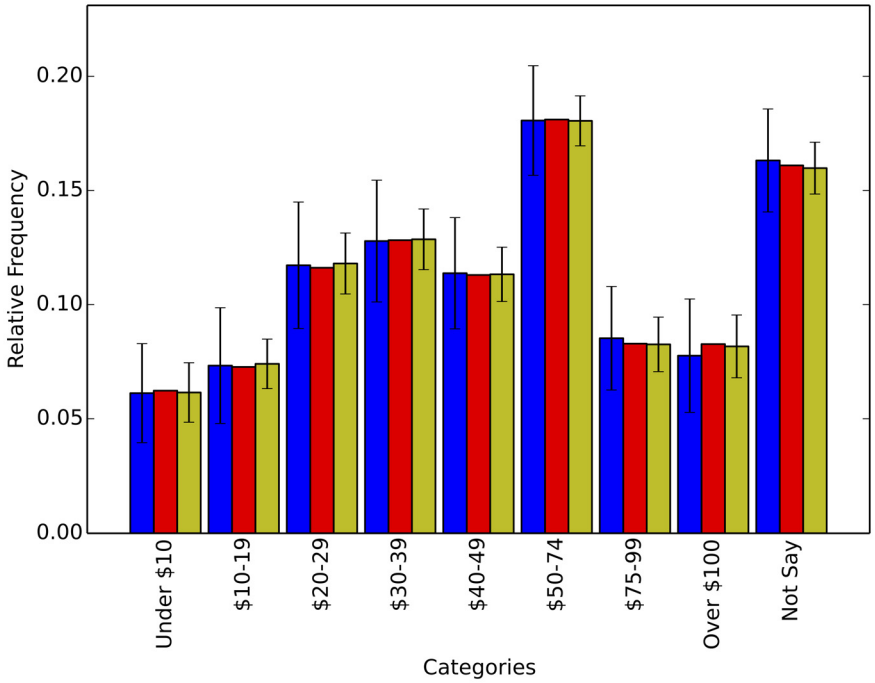
Fig 4. Responses to the question: How would you classify yourself? Each category shows, from left to right, the relative frequency for the interviewer defined privacy survey with $k_0 = 1$; the true relative frequency; and the relative frequency for the respondent defined privacy survey with random preference. Each bar represents the average of 100 repetitions and the error bar the standard deviation.

doi:10.1371/journal.pone.0147314.g004

private which exposes the true, positive category. Each multiple answer estimator is weighed in proportion to the number of responses with a given privacy preference. The code used to generate the simulation is available from the authors upon request. Each simulation was run 100 times with different random numbers and its summary statistics are reported by means of graphs.

Figs 2 through 6 show the results of the simulations. We selected five different fields from the database corresponding to sensitive questions and whose answers show different distributions, thus allowing us to test our approximation under different scenarios. We also include the results of the interviewer defined privacy survey with $k_0 = 1$, corresponding to the standard application of a negative survey, to demonstrate how the instrument presented here leverages the extra information disclosed by participants to achieve higher precision (lower variance).

Each experiment was run 100 times and the figures report the average relative frequency estimation as well as the standard deviation of the estimation. As expected, the average estimation is very close to the real relative frequency, but there is marked decrease in the standard deviation (above 40% decrease on average per experiment) from the interviewer defined



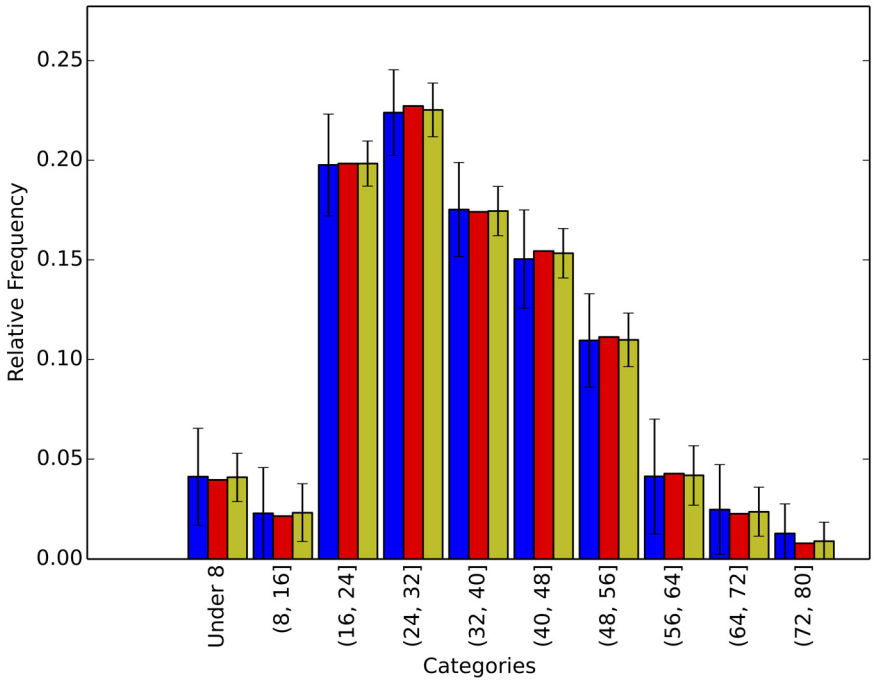
Standard deviation results per category

	Under $10	$10-19	$20-29	$30-39	$40-49
Custom Privacy	0.013	0.0108	0.0134	0.0133	0.0119
$k_0 = 1$	0.0217	0.0253	0.0277	0.0267	0.0244
Reduction	39.93%	57.22%	51.78%	50.19%	51.19%

	$50-74	$75-99	Over $100	Not Say	Average
Custom Privacy	0.0109	0.0119	0.0138	0.0113	0.0123
$k_0 = 1$	0.024	0.0227	0.0248	0.0225	0.0244
Reduction	54.44%	47.33%	44.6%	49.66%	49.59%

Fig 5. Responses to the question: Please indicate your current household income in U.S. dollars. Each category shows, from left to right, the relative frequency for the interviewer defined privacy survey with $k_0 = 1$; the true relative frequency; and the relative frequency for the respondent defined privacy survey with random preference. Each bar represents the average of 100 repetitions and the error bar the standard deviation.

doi:10.1371/journal.pone.0147314.g005



Standard deviation results per category

	Under 8	(8, 16]	(16, 24]	(24, 32]	(32, 40]
Custom Privacy	0.012	0.0144	0.0113	0.0135	0.0124
$k_0 = 1$	0.0243	0.0228	0.0255	0.0215	0.0236
Reduction	50.45%	36.97%	55.61%	37.36%	47.67%

	(40, 48]	(48, 56]	(56, 64]	(64, 72]	(72, 80]	Average
Custom Privacy	0.0124	0.0134	0.0148	0.0122	0.0095	0.0126
$k_0 = 1$	0.0247	0.0234	0.0287	0.0224	0.0148	0.0232
Reduction	49.66%	42.58%	48.39%	45.55%	35.68%	44.99%

Fig 6. Responses to the question: What is your age? We discretized the answers into 10 equally sized bins and replaced the "Not Say" category for zeros. Each category shows, from left to right, the relative frequency for the interviewer defined privacy survey with $k_0 = 1$; the true relative frequency; and the relative frequency for the respondent defined privacy survey with random preference. Each bar represents the average of 100 repetitions and the error bar the standard deviation.

doi:10.1371/journal.pone.0147314.g006

privacy set up to the respondent defined privacy scheme. This decrease shows how our method is able to harness the extra information provided by some of the participants to reduce its variance. Finally Fig 7 shows the result of one, randomly chosen, simulated survey to illustrate how a particular application of our technique on a particular question might look.

Discussion

The need for privacy-preserving surveying techniques stems from the desire to eliminate biases related to asking sensitive questions as well as from the duty to protect respondents from unintended consequences. The amount of potentially sensitive data that are now produced daily and stored indefinitely by both humans and sensors has fueled the need for a richer toolkit of data collection techniques. In this paper we introduced a surveying technique that is mindful of the subjectivity inherent in assessing the sensitivity of a question and that empowers a respondent, be it a human or a sensor, to select the amount of information to disclose; in essence, our method allows a question to be answered partially in accordance to its perceived

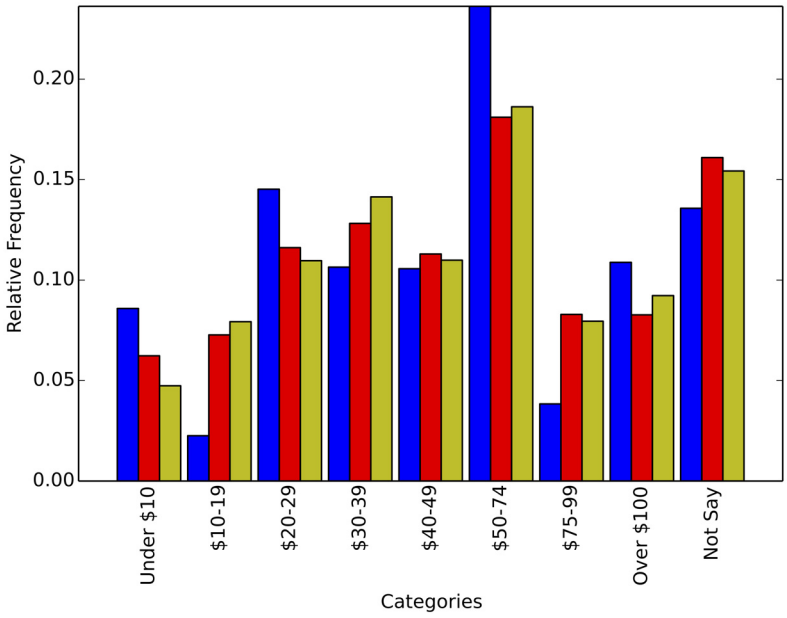


Fig 7. This figure shows the result of only one randomly chosen experiment for the household income question in order to illustrate how the approximation of our method might look for a particular survey. Each category shows, from left to right, the relative frequency for the interviewer defined privacy survey with $k_0 = 1$; the true relative frequency; and the relative frequency for the respondent defined privacy survey with random preference.

doi:10.1371/journal.pone.0147314.g007

intrusiveness. This technique will enable experimenters to leverage sensitive data in a more efficient way by maintaining sensitivity related biases low without sacrificing the information of those willing to disclose it. Data so gathered will also have a long lasting privacy assurance since, by itself, it is not enough to impute the sensitive characteristic to a particular respondent.

We focused on a specific kind of survey—multiple choice questionnaire with exhaustive and mutually exclusive categories— and based the technique on Negative Surveys. We provided the necessary tools to estimate the population proportion and variance of each category, but left out how the questions on the survey should be worded when the potential respondents are humans rather than electronic devices. Additionally we conducted a simulation study that shows the accuracy of our instrument for real data distributions and points to its possible application for de-sensitizing previously collected sensitive data.

Acknowledgments

The authors gratefully acknowledge the support provided by Asociación Mexicana de Cultura, A. C.; the GTRC and the GVU Center for providing the data for our experimental exposition; and the article's reviewers for their helpful comments.

Author Contributions

Conceived and designed the experiments: FE KH VMG. Performed the experiments: FE. Analyzed the data: FE VMG. Wrote the paper: FE KH VMG.

References

1. PCAST. Presidents Council of Advisors on Science and Technology; 2014. https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.

2. Sweeney L. K-anonymity: A Model for Protecting Privacy. Int J Uncertain Fuzziness Knowl-Based Syst. 2002 Oct; 10(5):557–570.
3. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. In: ICDE'06. Proceedings of the 22nd International Conference on Data Engineering; 2006. p. 24–24.
4. Feigenbaum, J, Pinkas, B, Ryger, RS, Jean, FS. Secure Computation of Surveys. In: EU Workshop on Secure Multiparty Computation (SMP). Amsterdam, The Netherlands; 2004.
5. Schneier B. Applied Cryptography: Protocols, Algorithms, and Source Code in C. New York, NY, USA: John Wiley and Sons, Inc.; 1994.
6. Basharat I, Azam F, Muzaffar AW. Database Security and Encryption: A Survey Study. International Journal of Computer Applications. 2012; 47(12):28–34. doi: 10.5120/7242-0218
7. Warner SL. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association. 1965; 60(309):63–69. doi: 10.1080/01621459.1965.10480775 PMID: 12261830
8. Chaudhuri A, Mukerjee R. Randomized Response: Theory and Techniques. Marcel Dekker Inc.; 1988.
9. Droitcour J, Larson E, Scheuren F. The Three Card Method: Indirect Estimation of Sensitive Survey Items. In: Proceedings of the Annual Meeting of the American Statistical Association; 2001. p. 1–5.
10. Esponda F, Guerrero VM. Surveys with Negative Questions for Sensitive Items. Statistics and Probability Letters. 2009; 79:2456–2461. doi: 10.1016/j.spl.2009.08.019
11. Raghavarao D, Federer WT. Block Total Response as an Alternative to the Randomized Response Method in Surveys. Journal of the Royal Statistical Society, Series B. 1979; 41(1):40–45.
12. Chaudhuri A, Christofides T. Indirect Questioning in Sample Surveys. Springer; 2013. doi: 10.1007/978-3-642-36276-7
13. Trappmann M, Krumpal I, Kirchner A, Jann B. Item Sum: A New Technique for Asking Quantitative Sensitive Questions. Journal of Survey Statistics and Methodology, Advance Access. 2013;.
14. Glynn AN. What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment. Public Opinion Quarterly. 2014; 77:159–172. doi: 10.1093/poq/nfs070
15. Krumm J. A survey of computational location privacy. Personal and Ubiquitous Computing. 2009; 13 (6):391–399. doi: 10.1007/s00779-008-0212-5
16. Mun M, Hao S, Mishra N, Shilton K, Burke J, Estrin D, et al. Personal data vaults: a locus of control for personal data streams. In: Proceedings of the 6th International COnference. ACM; 2010. p. 17.
17. Di Pietro R, Viejo A. Location privacy and resilience in wireless sensor networks querying. Computer Communications. 2011; 34(3):515–523. doi: 10.1016/j.comcom.2010.05.014
18. Othman SB, Trad A, Alzaid H, Youssef H. Secure and energy-efficient data aggregation for wireless sensor networks. International Journal of Mobile Network Design and Innovation. 2013; 5(1):28–42. doi: 10.1504/IJMNDI.2013.057146
19. Bao Y, Luo W, Zhang X. Estimating positive surveys from negative surveys. Statistics & Probability Letters. 2013; 83(2):551–558. doi: 10.1016/j.spl.2012.10.032
20. Bao Y, Luo W, Lu Y. On the dependable level of the negative survey. Statistics & Probability Letters. 2014; 89:31–40. doi: 10.1016/j.spl.2014.02.011
21. Kulik L. Privacy for real-time location-based services. SIGSPATIAL Special. 2009; 1(2):9–14. doi: 10.1145/1567253.1567256
22. Groat MM, Edwards B, Horey J, He W, Forrest S. Enhancing privacy in participatory sensing applications with multidimensional data. In: 2012 IEEE International Conference on Pervasive Computing and Communications (PerCom). IEEE; 2012. p. 144–152.
23. Aoki S, Iwai M, Sezaki K. Limited Negative Surveys: Privacy-Preserving Participatory Sensing. In: 2012 IEEE 1st International Conference on Cloud Networking (CLOUDNET). IEEE; 2012. p. 158–160.
24. Shunsuke A, Sezaki K. Negative Surveys with Randomized Response Techniques for Privacy-Aware Participatory Sensing. IEICE Transactions on Communications. 2014; 97(4):721–729.
25. Liu R, Tang S. Negative Survey-Based Privacy Protection of Cloud Data. In: Advances in Swarm and Computational Intelligence. Springer; 2015. p. 151–159.
26. Agresti A, Coull BA. Approximate is better than exact for interval estimation of Binomial proportions. The American Statistician. 1998; 52(2):119–126. doi: 10.1080/00031305.1998.10480550
27. Graphic UCG Visualization. 8th WWW User Survey; 1997. Accessed: 2015-09-10. http://www.cc.gatech.edu/gvu/user_surveys/survey-1997-10/datasets/.