Loki: A Privacy-Conscious Platform for Crowdsourced Surveys

Thivya Kandappu†*, Vijay Sivaraman†, Arik Friedman*, Roksana Boreli*
† School of Electrical Engineering & Telecommunications, UNSW, Sydney, Australia.
* NICTA, Sydney, Australia.
{t.kandappu@student., vijay@}unsw.edu.au, {arik.friedman, roksana.boreli}@nicta.com.au

***Abstract*—Emerging platforms such as Amazon Mechanical Turk and Google Consumer Surveys are increasingly being used by researchers and market analysts to crowdsource large-scale survey data from on-line populations at extremely low-cost. However, by participating in successive surveys, users risk being profiled and targeted, both by surveyors and by the platform itself. In this paper we propose, develop, and evaluate the design of a crowdsourcing platform, called Loki, that is privacy conscious. Our contributions are three-fold: (a) We propose Loki, a system that allows users to obfuscate their (ratings-based or multiple-choice) responses at-source based on their chosen privacy level, and gives surveyors aggregated population averages with known statistical confidence. (b) We develop a novel selection mechanism, which the platform can use to give surveyors accurate population estimates within a cost budget, while ensuring fairness in privacy loss amongst users. (c) We evaluate our scheme both off-line using a large dataset of movie ratings, and on-line via experimentation with 131 real users using a prototype implementation on mobile devices. Our work represents a first step towards incorporating privacy protection in emerging platforms for crowdsourced survey data.**

I. Introduction

In recent years, both academic and market researchers have been increasingly relying on online crowdsourcing platforms for conducting surveys, to gain new insights about customers and populations. The Amazon Mechanical Turk (AMT) [1] platform is being used by researchers in experimental psychology to conduct low-cost large-scale behavioral studies by obtaining opinion survey data from paid volunteer populations, well beyond the walls of university campuses [2]. The recently launched Google Consumer Surveys platform [3] enables large-scale market surveys by gating access to premium content using a "surveywall" that requires users to answer "one question at a time". Though users of such systems receive compensation (in-cash or in-kind) for contributing data, they lose privacy with each personal fact or opinion revealed in the course of the surveys they participate in.

The release of personal facts and opinions, albeit in small increments, can over time be accumulated (by the surveyors or by the platform) to profile individuals. This gradual loss in privacy may be undesirable for many users, and even harmful (in social, financial, or legal ways) for some. Furthermore, the threat comes not only from surveyors, but also from the platform itself, which can exploit the profiling for its own ends, or cede it to another entity for gain. Prior solutions that have advocated the use of a trusted third party (or proxy) [4], [5], [6], [7] fall short on this account, since they lead to a gradual transfer of personal information to the external entity (e.g., the ISP, Amazon, Google, or a government body), which may even collude with surveyors. Solutions like anonymization also fall short as they can be circumvented via the use of auxiliary information [8], [9], [10], and also make it difficult to compensate users for contributing their data. What is therefore needed is a method that does not need to rely on users being anonymous, but rather empowers them to obfuscate their data at source without having to trust external parties, thereby allowing them to control the rate at which they leak their private information to the rest of the world.

In this paper we propose a system, called "Loki", in which users do not need to trust anyone other than themselves, and can obfuscate their responses at source prior to submission to the platform. At the outset, local obfuscation may seem trivially achievable; for example, continuous-valued (e.g., ratings-based) responses can be obfuscated by adding random noise (e.g., [11]), and discrete-valued (e.g., multiple-choice) responses can be obfuscated by randomly flipping the true response (e.g., "randomized response" [12]). However, obfuscation by users creates two new challenges for the platform: (a) the measured survey outcomes will differ from the true responses, and the platform therefore needs to quantify and manage the inaccuracy so that surveyors get acceptable confidence in the results, and (b) the platform needs to ensure fairness in privacy depletion across its users (i.e., some users should not lose much more privacy than others), so as to continue enjoying the long-term support of its entire user base. Managing this trade-off between service quality (to surveyors) and privacy fairness (to users) requires a non-trivial mathematical framework, and motivates the research undertaken in this paper.

We acknowledge that obfuscation does not prevent privacy leakage, it merely slows it down. If a user is asked the same question in several ways, the independent noise (of known distribution) added to each response can be filtered out with more certainty due to the larger number of samples available. It may be possible to counter this filtering by adding dependent noise to responses, but this is infeasible in practice, as the underlying correlations between questions are extremely difficult to quantify and are often unknown. In this work we therefore take the simpler approach of adding independent noise to user responses, and treat the quantified cumulative privacy loss as an upper bound rather than an exact value. Further, we restrict our focus to ratings-based questions and multiple-choice questions, noting that the underlying method of adding noise is general and applicable to other question types in which the response set is numeric or countable (i.e., almost anything other than free-text responses).

Our specific contributions in this paper are: **(a)** We propose the architecture of Loki, a privacy-preserving crowdsourcing platform, and describe our design choices surrounding obfuscation techniques, user privacy levels, privacy loss quantification, user privacy depletion, cost settings, and user utility estimation.

(b) We develop a novel algorithm by which the platform can select the best subset of users for a given survey to achieve the desired balance between cost, accuracy, and privacy. This algorithm ensures that the best accuracy is obtained within a stipulated cost budget, while maintaining fairness in privacy depletion across users. **(c)** We evaluate our user selection algorithm off-line on a large dataset of movie ratings, and show that it achieves higher accuracy for a given cost budget, and ensures consistency in accuracy over successive surveys. Additionally, we develop a prototype of our Loki system, including the surveyor web-interface, broker platform, and client apps for iOS and Android based mobile devices. We validate our solution by conducting an experiment with 131 volunteers, to demonstrate the validity of our approach in protecting user privacy, while obtaining reasonably accurate aggregated responses even in small-scale settings.

We believe our work is among the first to explicitly incorporate privacy into data crowdsourcing platforms such as AMT or Google Consumer Surveys, allowing utility, privacy and cost to be balanced in a controllable way. The rest of the paper is organized as follows: existing privacy preserving solutions are reviewed in §II. We present our system architecture in §III, and develop an algorithm for user selection in §IV. We evaluate our algorithm offline using the NetFlix dataset in §V, and describe our prototype implementation and associated insights in §VI. The paper is concluded in §VII.

II. Related Work

Early research works on data anonymization proposed sanitizing user data by masking or removing personally identifiable information (PII). However, mechanisms like k-anonymity [13] and variants like ℓ-diversity [14] were shown to be vulnerable to composition attacks [15], and do not provide adequate privacy protection in the presence of auxiliary information. It was shown that intuitive anonymization techniques are not effective in protecting user identity, as individual users can be re-identified [16], [9], [10].

For the distributed settings, a number of research works studied *input perturbation*, where privacy is obtained by adding noise to user data at the source. A generic approach proposed in [17], [11] is to add random distortion values drawn independently from a known distribution, e.g., the uniform distribution. A number of improvements in this technique were subsequently proposed [18], [19]. However, it was shown [20] that an adversary may analyze the data and filter out some of the noise, effectively reducing the bounds of uncertainty introduced by the noise and compromising the privacy guarantees.

More recently, differential privacy [21], [22] has emerged as a promising way of providing provable privacy guarantees under arbitrary adversary conditions, i.e., an adversary who may have any level of computational power and background knowledge. Differential privacy is typically achieved by adding a calibrated level of noise to the response on a query over a centralized database [21]. These mechanisms have been extended to the distributed setting [23], [24], where they leverage cryptographic techniques to generate differentially private noise in a distributed manner; however, they do not seem to scale well.

Several works have proposed architectures that rely on a trusted third party, or an honest-but-curious party, to assist in scalable data processing tasks while preserving user privacy. Guha et al. [5] presented an online advertising system that uses a proxy to hide user activities to the desired level, while enabling a number of standard online advertising features, including defense against click-fraud. Riederer et al. [7] focused on monetization of personal information and Toubiana et al. [25] proposed a combination of localized (browser-based) user profiling and a cryptographic billing system, also for online advertising. The system presented by Chen et al. [4] is the closest to our proposal, targeting personal query collection and aggregation from a subset of the overall user population; however their system relies on the broker being honest, which is not required in our approach.

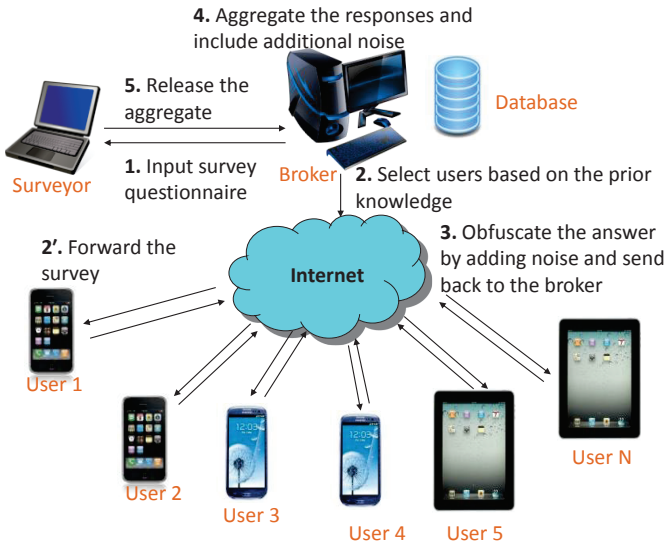


Fig. 1. System components and basic protocol

III. System Design

A. Architecture and Entities

The proposed system (Fig. 1) is closely modeled along the lines of AMT, and comprises three entities: surveyors, users, and the broker platform. **Surveyors** acquire data from users using a set of questions in a survey form (our work in this paper is restricted to ratings-based and multiple-choice questions). The surveyor pays the broker to run the survey, specifying an upper bound on total cost. The surveyor expects sufficient accuracy (utility) of the aggregated response, in that it should represent the averaged opinion over the entire user population. **Users** respond to questions in the surveys they choose to participate in, using a supplied application (app) installed on their personal device (smart phone/tablet). The app allows users to obfuscate their responses at source. The users' monetary compensation may in general depend on their choice of privacy level – higher privacy levels entail higher obfuscation and hence lower payment. We do not deal with intentional lying (or cheating) by users to get higher compensation. The **broker** provides a platform for launching surveys to users. It receives payment from surveyors, and passes it on to participants (less a commission). The broker has a dual objective: to provide accurate population estimates to surveyors (this requires it to keep track of how good a predictor of population behavior each user has been in the past), and to deplete privacy fairly across users so as to extend their lifetime in the system (this requires it to keep track of cumulative privacy depletion for a user across earlier surveys).

B. Design Choices

1) Obfuscation Technique: We use simple and natural techniques by which the user client locally obfuscates the answer before reporting it to the broker. For *ratings-based* questions, Gaussian noise $\mathcal{N}(0, \gamma^2)$ is locally added to the user response. We chose Gaussian over uniform as it has unbounded range, and hence does not compromise user privacy in boundary cases. We preferred it over Laplace since it is additive, i.e. the sum of Gaussian noise terms is still Gaussian. Further, note that the mean of the noise is chosen to be zero for convenience, so as not to introduce any bias one way or the other. The standard deviation γ will be adjusted based on the user's privacy chosen privacy level. For *multiple-choice* questions, we use the randomized response technique [12], whereby the user's true selection is preserved with probability $1 - p$, and with probability p ($p < 0.5$) the response is changed uniformly randomly to one of the other choices. Again, the value of p is dictated by the user's chosen privacy level, described next.

2) User Privacy Levels: In theory, any number of privacy options can be available to the user. However, to keep it simple, we advocate a set of four privacy levels: *none*, *low*, *medium*, and *high*. The chosen privacy level determines the amplitude of the noise that is added to obfuscate the true user response. Needless to say, the higher the privacy level, the larger the obfuscation parameter (γ or p above). These are illustrated with examples below:

Example 1. *Consider a 5-point Likert scale commonly used in psychology studies, with the possible response values including: 1 (strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), and 5 (strongly agree). A reasonable selection of obfuscation parameter might be: $\gamma = 0$ for no privacy, $\gamma = 3$ for low privacy, $\gamma = 6$ for medium privacy, and $\gamma = 12$ for high privacy (note that the reported responses will consequently be real-valued rather than integers).*

Example 2. *Consider a multiple choice question with five options. A reasonable selection of obfuscation parameter might be: $p = 0$ for no privacy, $p = 0.1$ for low privacy, $p = 0.3$ for medium privacy, and $p = 0.4$ for high privacy.*

3) Privacy Loss Quantification: Given the obfuscation techniques above, we need a way to quantify the privacy loss for a user who answers a particular survey at a particular privacy level, and to accumulate the user's privacy loss across multiple surveys. For this purpose, we utilize the rich mathematical framework provided by differential privacy [21]. Differential privacy provides strict bounds on the sensitivity of the outcome of the computation to any particular record in the input. Consequently, the output of a differentially private computation does not allow inference of any specific input record, irrespective of the adversary's computational power or the available background knowledge. Formally, a function K provides (ϵ, δ)-differential privacy [23] if for any two datasets A and B differing in a single record, and for all outcomes S:

$$\Pr[K(A) \in S] \leq exp(\epsilon) \times \Pr[K(B) \in S] + \delta \ . \quad (1)$$

The degree of privacy is controlled by the parameter ϵ, while δ allows the condition in Equation 1 to be relaxed for unlikely events. Differential privacy maintains composability, meaning that if two computations maintain (ϵ_1, δ_1) and (ϵ_2, δ_2) differential privacy respectively, then executing both would amount to $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$ differential privacy.

In our case, the differential privacy constraint is applied to each survey answer, i.e., the difference between two datasets A and B amounts to a difference in a single answer. For *rating based* questions, the privacy guarantees of Gaussian noise $\mathcal{N}(0, \gamma^2)$ can be mapped to (ϵ, δ)-differential privacy measures through the relation [23]:

$$\frac{\epsilon\gamma^2}{2R^2} + \ln(\epsilon\gamma^2) \geq \ln \frac{1}{\delta} \ , \quad (2)$$

where R is the range of the user's possible answers. To illustrate by an example:

Example 3. *Following from Example 1, the 5-point Likert scale based ratings with privacy levels {no, low, medium, high} respectively used $\gamma = \{0, 3, 6, 12\}$. Since $R = 4$, and fixing $\delta = 0.01$, the privacy settings correspond to differential privacy guarantees of $\epsilon = \{\infty, 3.42, 0.85, 0.21\}$ respectively.*

For *multiple choice* questions obfuscated using the randomized response technique, the mapping from the probability measure p to (ϵ, δ) can be derived from (1) as:

$$\epsilon \geq \ln(1 - p - \delta) - \ln(p) + \ln(n - 1) \ . \quad (3)$$

Example 4. *Following from Example 2 of a multiple choice question with five options, privacy settings {no, low, medium, high} respectively used $p = \{0, 0.1, 0.3, 0.4\}$. Fixing $\delta = 0.01$, the privacy settings correspond to differential privacy guarantees of $\epsilon = \{\infty, 3.57, 2.22, 1.77\}$ respectively.*

4) User Privacy Depletion: We have shown above how a user's privacy loss in a survey can be translated to differential privacy metrics (ϵ, δ). Fortunately, these metrics are composable (i.e., additive), and the user's privacy loss over successive surveys can therefore easily be estimated by accumulating these metrics over the user's lifetime. We emphasize here that our objective in quantifying a user's privacy loss is so that the broker can try to be fair in privacy loss across users (as per the algorithm developed in the next section); we treat the differential privacy metrics as an upper bound that captures the relative privacy loss for each of the users. In the rest of this work, we will fix the value of δ at 0.01, and use ϵ for comparing privacy loss across users. Further, for cases where users choose privacy level "none", we set $\epsilon = 0$ (rather than the theoretically correct value of ∞), since the users are explicitly indicating that they do not value privacy for that survey, and the effect of this survey on their cumulative privacy loss should not be accounted for.

5) Cost Settings: A user i, who contributes data in response to a survey questionnaire, receives a compensation c_i. Users who choose a higher privacy level (and consequently add more noise to their responses) may receive lower compensation than those who choose a lower level of privacy.

Example 5. *Following from Example 1 that uses a 5-point Likert scale, the privacy levels none, low, medium, and high could correspond to user payments c_i of \$0.8, \$0.4, \$0.2 and \$0.1 respectively. The unit of cost is arbitrary and can be scaled appropriate to the complexity or value of the survey.*

6) User History and Utility: Despite noise addition by users to obfuscate individual answers, some characteristics of user behavior can be discerned by the broker over time. As an example, noise added by a user to n successive ratings-based questions, each with iid noise $\mathcal{N}(0, \gamma^2)$, can be averaged

by the broker to estimate the user's mean noise $\mathcal{N}(0, \gamma^2/n)$ that has lower variance. This fact can be leveraged by the broker to estimate metrics such as the "error" of the user's ratings, i.e., to determine on average how close the user's ratings in the past have been to the population averages. This in turn indicates how representative this user is of the general population, and helps the broker estimate the "value" of the user towards obtaining an accurate population estimate. In the following section, we will use this notion of user "value" to select users in a way that balances the accuracy, cost, and privacy needs for each survey.

IV. Privacy Preserving Selection Mechanism

We develop a practical method for the broker to select users to participate in each survey so as to balance cost, accuracy, and privacy. We outline the approach for ratings-based questions (continuous-valued); the analysis for multiple-choice questions (discrete-valued) is presented in the full version [26].

A. Quantifying Estimation Error

The broker is tasked with estimating the population average of a statistic (e.g., movie rating, product popularity, disease prevalence). Due to the cost constraint set by the surveyor, the broker can query only a subset of users S from the universal set of users U, and this selection is based on accuracy, cost, and privacy depletion. We begin by estimating the accuracy of the statistic due to sampling and user noise addition.

1) The Estimated Measure: Denote by $x_i \in \mathbb{R}$ the input of user $i \in U$. The desired population average θ is given by $\theta = \sum_U x_i/|U|$. The broker estimates this statistic by sampling a subset of users S. Further, each user i sends obfuscated input $\hat{x}_i = x_i + n_i$ to the broker, whereby the true input x_i is combined with noise n_i taken from $\mathcal{N}(0, \gamma_i^2)$, where γ_i depends on the user's chosen privacy level. The broker's estimate $\hat{\theta}$ of the population average is then given by $\hat{\theta} = \sum_S \hat{x}_i/|S| = \sum_S (x_i + n_i)/|S|$. The mean squared error in the estimator is given by:

$$RMSE^2 = (\hat{\theta} - \theta)^2 = \left[\frac{\sum_S n_i}{|S|} + \left(\frac{\sum_S x_i}{|S|} - \theta \right) \right]^2 \quad . \quad (4)$$

When selecting S, the broker therefore accounts for two influencing factors: the level of privacy required by each user, which determines the error due to privacy-related noise (first term above), and the expected sampling error (second term above). As the broker has the prior history of each of the users, it can evaluate how well each user's answers reflect the real aggregate θ, as discussed next.

2) User and Group Error History: The "value" of a user depends on how accurately the user's responses reflect those of the population at large. To quantify this, we consider the user error, i.e., the difference Δ_i between the user's response and the true population average, given by $\Delta_i = x_i - \theta$. Treating the user error Δ_i as a random variable, we can estimate its mean μ_i and variance σ_i^2 from the history of prior responses $H_i = \{\hat{x}_{ij}\}$ of the user using $\mu_i = \mathbb{E}[\Delta_i] = \sum_{j: x_{ij} \in H_i} (x_{ij} - \theta_j)/|H_i|$ and $\sigma_i^2 = \mathrm{Var}[\Delta_i] = \sum_{j: x_{ij} \in H_i} (x_{ij} - \theta_j - \mu_i)^2/|H_i|$, where θ_j denotes the true population average in a past survey question q_j. New users can be assigned a default value of user error.

Similarly, we can define the value of a group of users S to reflect how closely the average rating by this group matches the true population rating. The average rating by the group is defined as $x_S = \sum_S x_i/|S|$ (when not all users in the group answer a question, for convenience we take the average only over users who do answer). Denoting by Δ_S the group error, which quantifies the difference between this group's average rating and the population average, we have $\Delta_S = x_S - \theta$. The mean and variance of the group error can be deduced from the prior history $H_S = \{\hat{x}_{Sj}\}$ of this group using $\mu_S = \mathbb{E}[\Delta_S] = \sum_{j: x_{Sj} \in H_S} (x_{Sj} - \theta_j)/|H_S|$ and $\sigma_S^2 = \mathrm{Var}[\Delta_S] = \sum_{j: x_{Sj} \in H_S} (x_{Sj} - \theta_j - \mu_S)^2/|H_S|$.

The estimation of the user and group errors above assumes perfect knowledge of the true user responses x_i and the population averages θ_j. In reality the broker only has the noisy user/group responses ($\hat{x}_i$ or $\hat{x}_S$), as well as noisy population estimate $\hat{\theta}_j$ for prior survey questions. The mean (μ_S) and variance (σ_S^2) of the true group error can be approximated with the mean ($\hat{\mu}_S$) and variance ($\hat{\sigma}_S^2$) of the computed errors, using the fact that the noise is independent of user responses and has zero mean: $\hat{\mu}_S \approx \mu_S$ and $\hat{\sigma}_S^2 \approx \sigma_S^2 + \frac{\sum_S \gamma_i^2}{|S|^2} + \frac{\sum_U \gamma_i^2}{|U|^2}$. The expectation of the error in Eq. (4) is then derived as:

$$\begin{aligned} \mathbb{E}(RMSE^2) = & \quad \mathbb{E}\left[\left(\frac{\sum_S n_i}{|S|} \right)^2 \right] + \mathbb{E}\left[(x_S - \theta)^2 \right] = \\ = & \quad \frac{\sum_S \gamma_i^2}{|S|^2} + \sigma_S^2 + \mu_S^2 \approx \hat{\mu}_S^2 + \hat{\sigma}_S^2 - \frac{\sum_U \gamma_i^2}{|U|^2} \quad . \quad (5) \end{aligned}$$

In general, as the size of the set S increases, i.e., more users are surveyed, the error above decreases; in fact, it can be verified that when $S = U$, then $\mu_S = \sigma_S^2 = 0$ (since $x_U = \theta$ by definition), and the sampling bias is eliminated, with estimation error arising only from the noise added by users. We will now see how the estimation error balances with cost and fairness.

B. Balancing Cost, Accuracy, and Privacy Fairness

1) Optimizing a Single Survey: As described in §III-B5, each user chooses a privacy setting, which incurs a privacy cost (ϵ_i, δ_i). The privacy protection is obtained by adding noise with variance γ_i^2. The privacy setting is also associated with monetary compensation c_i. Given the user choices, the broker proceeds to select a group of users to be included in the survey, based on two constraints:

Monetary cost constraint: A surveyor sets an overall cost C for a survey. The broker selects n_j users who picked the j-th privacy setting associated with cost c_j. To stay within the overall cost bound, the broker ensures $\sum_j n_j c_j \leq C$.

Privacy constraint: For each user, the cumulative privacy loss throughout the system lifetime is capped at $(\epsilon_{max}, \delta_{max})$. Each user i in survey j incurs a known privacy cost $(\epsilon_{ij}, \delta_{ij})$ depending on the selected privacy level. The accumulated privacy loss for user i is therefore $(\sum_j \epsilon_{ij}, \sum_j \delta_{ij})$ where the summation is over all the past surveys taken by this user. The residual privacy budget for the user is consequently $(R_i^{(\epsilon)}, R_i^{(\delta)})$, where $R_i^{(\epsilon)} = \epsilon_{max} - \sum_j \epsilon_{ij}$ and $R_i^{(\delta)} = \delta_{max} - \sum_j \delta_{ij}$. To guarantee that the user's cumulative privacy

loss stays within the lifetime privacy budget, the broker must ensure that for the new survey, $\epsilon_i \leq R_i^{(\epsilon)}$ and $\delta_i \leq R_i^{(\delta)}$.

For a new survey, we can therefore pose the selection of a set S of users to survey as an optimization problem:

$$\arg\min_{S \subseteq U} RMSE \quad (6)$$
$$\text{s.t. } \sum_j n_j c_j \leq C \text{ and } \forall i \in S : \epsilon_i \leq R_i^{(\epsilon)} \wedge \delta_i \leq R_i^{(\delta)} \ ,$$

where the RMS error is obtained from Equation (5). For the special case when a user chooses a "no privacy" setting, which in theory translates to an unconstrained loss in privacy ($\epsilon \to \infty$), we make the practical choice of using $\epsilon = 0$, $\delta = 0$, reflecting that the user is not concerned about the privacy implications in this case.

2) Optimizing Across Multiple Surveys: When considering a series of surveys, additional factors may influence the broker's choices, beyond the cost and privacy constraints. In particular, *Quality of Service (QoS)* across surveys aims to keep an (ideally) constant RMS error over successive surveys that can be maintained and guaranteed to the surveyors, while *fairness* aims to balance the residual level of privacy across users, since privacy can be seen as a non-renewable resource, which should be equally depleted across users. QoS considerations may motivate the broker to select for a survey users with low error, but this may deplete such users' privacy budget rapidly. Consequently, those users may be excluded from participation in subsequent surveys, resulting in deterioration of QoS over time.

To express the importance of QoS and fairness, we introduce a "fairness parameter" $\alpha \in [0, 1]$ that is set by the broker. We then combine the monetary and privacy cost of user i into an overall cost F_i, given by:

$$F_i = (1 - \alpha)\frac{c_i}{C} + \alpha \cdot \max\left[\frac{\epsilon_i}{R_i^{(\epsilon)}}, \frac{\delta_i}{R_i^{(\delta)}}\right] \ . \quad (7)$$

The first term considers the monetary cost of the user for this survey, as a fraction of the budget available for the survey. The second term considers the privacy depleted by this user's participation in the survey, as a fraction of their residual privacy budget. The overall cost is therefore a dimensionless score that takes a linear combination of the two costs, weighted by the fairness parameter α. When $\alpha \to 0$, monetary cost is of primary concern and fairness in privacy depletion is ignored. Conversely, when $\alpha \to 1$, monetary cost is ignored and users with a low residual privacy budget are assigned high cost, disfavoring them for selection so as to maintain fairness in privacy depletion. The next section presents the selection algorithm that uses this combined cost metric.

C. Algorithm for User Selection

For a new survey, the proposed algorithm is executed to select the set of users who yield the best accuracy within the given cost constraint, while also maintaining fairness in privacy depletion amongst users. Our initial construction of this set assumes that (a) all selected users will actually take the survey, and (b) we can correctly predict the privacy level choice of each user according to their past history. In reality, these assumptions may not hold, but the algorithm can be easily modified to refine the set based on actual user feedback.

Evaluating all possible subsets $S \subseteq U$ of users to determine the optimum would be intractable. Instead, we propose a greedy heuristic approach, by which the broker constructs the set S incrementally, each time adding the user who would be most cost effective, while taking into account the QoS and fairness considerations. Given a set of users $S \subseteq U$, Equation (5) evaluates the expected error $RMSE^{(S)}$ of the set, based on past performance. Adding the user i to the set would result in the set $S \cup \{i\}$, for which the expected error $RMSE^{(S \cup \{i\})}$ can be evaluated as well. The difference $\Delta RMSE(S, i) = RMSE^{(S)} - RMSE^{(S \cup \{i\})}$ encapsulates the reduction in error by inclusion of the user i in the set. We can then compute β_i, the improvement in RMS error per unit of cost, for the user i:

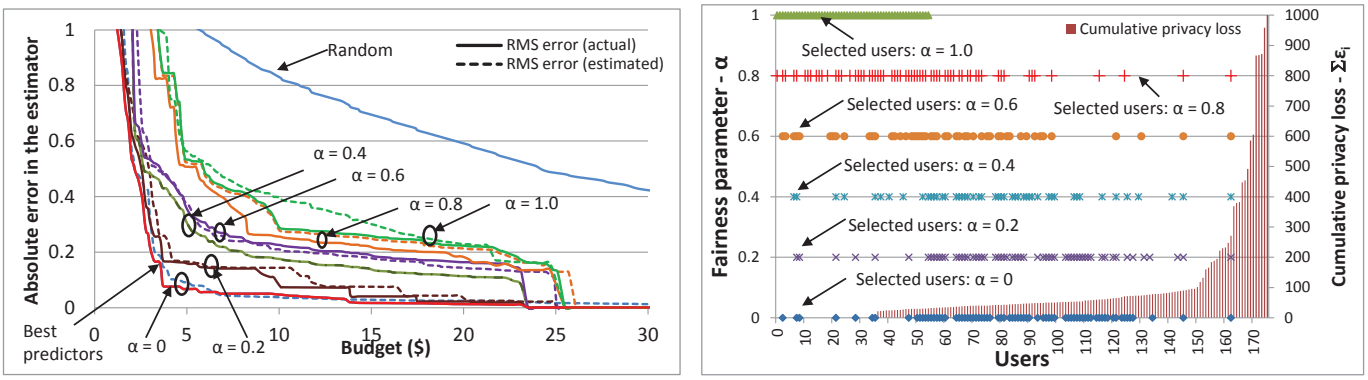
$$\beta_i(S) = \frac{\Delta RMSE(S, i)}{F_i} \ , \quad (8)$$

where the user cost F_i is given by Equation (7) and includes both monetary and privacy costs. In the greedy selection process, the broker picks the user with the highest β_i at each step (i.e., the user with the highest expected accuracy gain per unit of cost). By starting with an empty set of users, and iteratively adding users one by one, the broker can construct the target set S, until the monetary cost limit C is reached. Note that users who have depleted their lifetime privacy budget are not eligible for selection. Algorithm 1 describes this process.

Algorithm 1 Greedy User Selection Mechanism

Input: set of users U, each with cost c_u; overall cost bound C.
Output: set $S \subseteq U$ of survey participants.
$S = \emptyset$.
$P = \{i \in U : c_i \leq C \wedge \epsilon_i \leq R_i^{(\epsilon)} \wedge \delta_i \leq R_i^{(\delta)}\}$. ▷ candidate users within budget
while $P \neq \emptyset$ **do**
    $u \leftarrow \arg\max_{i \in P} \beta_i(S)$.
    $S \leftarrow S \cup \{u\}$, $P \leftarrow P \setminus \{u\}$.
    $C \leftarrow C - c_u$. ▷ remaining budget
    $P \leftarrow \{i \in P : c_i \leq C\}$.
end while
return S.

Algorithm 1 has complexity $O(KN^2)$, where K is the number of items that constitute prior history and N is the number of users. An alternative algorithm may simply sort users by value per unit cost, e.g., prior error for each user divided by (monetary and privacy) cost, and select the top few that can be afforded by the budget C. However, such an algorithm does not account for the correlations amongst user responses. For example, user A may always rate items below the population average, while user B is always above this value. Individually, each would have a high error, and a naive strategy would end up selecting neither. In contrast, Algorithm 1 works in a more subtle way, accounting for the performance of the group S as a whole rather than as the sum of the individuals in it. For the above example, if Algorithm 1 picks user A in one step, it would then favor picking user B in the next step since the two users together in S cancel out each other's error and have low joint error.



(a) Estimation error versus cost budget

(b) Privacy depletion across users

Fig. 2. Impact of selection policy on (a) accuracy versus cost, and (b) fairness in privacy depletion.

V. Offline Evaluation

We evaluate our algorithm on a large dataset of movie ratings to study the trade-offs between cost, utility, and fairness, and the long-term system performance.

A. Dataset and Methodology

We use as a survey answer set the Netflix dataset (http://www.netflixprize.com/), which contains over 100 million movie ratings (on a 5-point scale) from $480,000$ anonymized Netflix customers over $17,000$ movie titles, collected between Oct'98 and Dec'05. The movies released in 2004 (1436 in number) are used as historical information, and our objective is to estimate the population-wide average rating of movies released in 2005 within a specified cost budget C.

We consider a simplification of the privacy choice, with each user being permanently assigned into one of four privacy bins {none, low, medium, high} at random, with probabilities 13.8%, 24.4%, 38.9%, and 22.9% respectively (the probabilities were derived from our experimental study with real users, as described in §6). The bins are associated with zero-mean Gaussian noise with standard deviations $\gamma = 0, 3, 6, 12$ respectively (corresponding to $\epsilon = 0$, $\epsilon = 3.42$, $\epsilon = 0.85$ and $\epsilon = 0.21$), and respective payments of \$0.8, \$0.4, \$0.2, and \$0.1 for each user. Noise sampled from $\mathcal{N}(0, \gamma^2)$ is added to each of the users' movie ratings. To allow evaluation of the trade-offs over long periods of time, we only consider the users who have rated at least 50 movies.

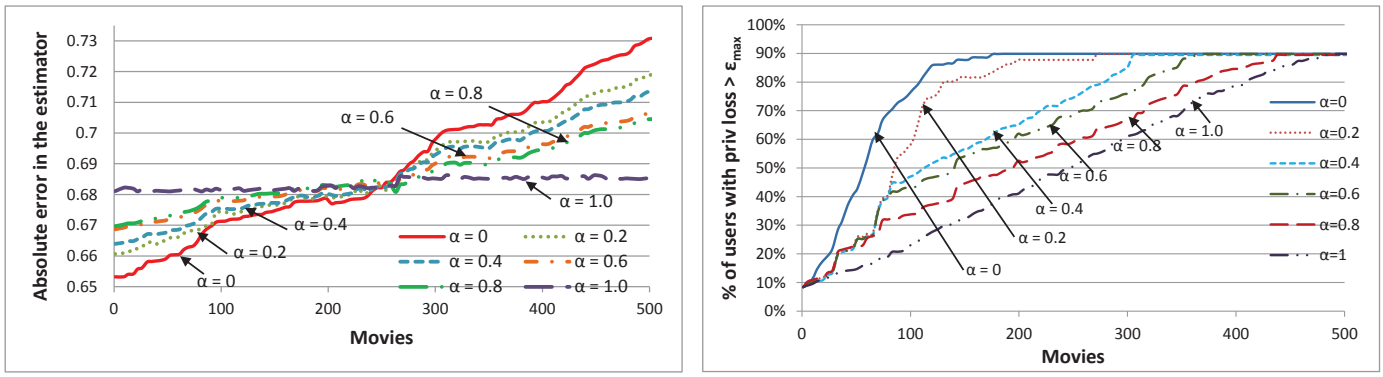
We derive the following measures, as described in §IV: (a) for user i, the "error" in rating is defined as the difference between the user's rating of a movie and the average rating of this movie by the population; we compute the mean $\hat{\mu}_i$ and variance $\hat{\sigma}_i^2$ of this error over all the movies rated by user i; (b) similarly, we define "error" for an arbitrary group S (a subset of the user population U) as the difference between the group's average rating of a movie and the average rating of the movie by the population; we compute the mean $\hat{\mu}_S$ and variance $\hat{\sigma}_S^2$ of this error over all the movies rated by (any non-empty subset of) the group S – note that this computation is done on demand as the set S evolves; and (c) lastly, for each user, we track the privacy loss over all past surveys taken by this user; the cumulative privacy loss $\sum \epsilon_i$ of user i is the sum of the ϵ's corresponding to the user's privacy choice in past surveys (recall that we assign $\epsilon = 0$ and $\delta = 0$ for users who choose the privacy setting "none").

B. Cost, Accuracy, and Fairness Trade-Offs

To evaluate the trade-offs for a single survey, we considered several movies from 2005. The results shown here correspond to the movie "Sleepover Nightmare" that was rated by 176 users (similar results were observed for other movies). Fig. 2(a) shows the estimation error $\mathbb{E}(RMSE)$ for varying values of the available budget C, for various selection policies. Our proposed selection mechanism is evaluated for different values of the fairness control parameter α. We also introduce two baseline selection strategies: random selection, in which a random set of users is selected subject to the cost constraint, and the "best predictors" selection, in which we choose the subset of the population which has the highest historical accuracy (i.e., is most representative of the population) subject to the cost constraint. For different α values, both the true error (i.e., difference between the estimate and the ground truth available to us in the dataset), depicted with solid lines, as well as the corresponding estimated error (computed using Equation 5), depicted with dashed lines, are shown in Fig. 2(a).

The estimated error (dashed lines) closely reflects the true error (solid lines), and is hence of sufficient accuracy to be useful in the selection decision. As can be expected, random selection of users results in the lowest accuracy, and the selection of "best predictors" consistently yields near-perfect estimates, even by surveying as low as 37% of the population. Setting $\alpha = 0$ yields accuracy identical to the "best predictors" selection algorithm, but as α progressively increases (in steps of 0.2 in the figure), the error increases.

However, the loss in accuracy is compensated for in privacy fairness. Random selection results in best privacy as the whole population is equally utilized. However, the large estimation error makes this approach unusable. Fig. 2(b) shows in sorted order (on the right hand side y-axis) the cumulative privacy loss of 176 users who rated the movie, resulting from prior surveys they have participated in. About 20% of the users have no recorded privacy loss (their privacy setting is "none"), while about 5% of users have a cumulative privacy loss in excess of 400. For the current survey, our selection algorithm uses a value of $\epsilon_{\max} = 400$, so that 95% of users are eligible for selection, and a cost budget of \$40. Fig. 2(b) then shows, for selected values of α (on the left hand side y-axis), the users who are selected by the algorithm, marked on the corresponding horizontal lines. When $\alpha = 1$, selection is strongly biased



(a) Estimation error over time

(b) Privacy loss over time

Fig. 3. Time evolution of (a) estimation error and (b) fraction of users reaching privacy threshold.

towards users who currently have low privacy loss, thereby making the selection fair in balancing privacy depletion. As α decreases, selection progressively gives less regard to prior privacy depletion. Selection is more concentrated in the center, since users on the left have high cost whereas users on the right have lower residual privacy budget.

C. Long-Term Performance

To evaluate the performance in a series of surveys, we apply our selection algorithm, sequentially, to a set of 500 movies released in 2005, again using the movies from 2004 as prior history. The privacy threshold at which a user is deemed to have depleted too much privacy and is ineligible to participate in further surveys is set to $\epsilon_{\max} = 400$. We show the evolution of accuracy in estimating the true movie rating in Fig. 3(a), and the privacy depletion in Fig. 3(b), for various α settings. It can be observed in Fig. 3(a) that when α is low, the error is initially low, but rises rapidly with successive movies. This happens because the best performing users are selected for the initial movies (yielding low error), but deplete their privacy rapidly. This is reflected in Fig. 3(b), which shows that for low α the fraction of users who exhaust their privacy budget grows rapidly with the number of rated movies. Conversely, a choice of high α results in fairer depletion of privacy, prolonging the lifetime of users in the system and giving more consistent quality of estimates over time. In the long run, the broker therefore has an incentive to choose a larger α setting to ensure fairness and consistency in the quality of the results. We note that our algorithm allows this parameter to be chosen by the surveyor on a per-survey basis.

VI. Prototype Implementation

A. System Components

Our prototype consists of two parts – a front-end application (for both iPad/iPhone and Android platforms) for users to participate in surveys, and a back-end database/server that stores user data. The workflow of the app closely follows the description in §III-A, and screenshots for the iPhone app are shown in the full version [26]. The opening screen allows the user to login (and register to create an account if needed). A list of the available surveys are shown on the next screen, along with available privacy levels. The survey screen lists the questions – in our trial, users (students) are asked to rate lecturers on a Likert scale, ranging from 1 (very poor) to 5 (excellent). The user can select the lecturers he/she wants to rate and use the slide bar to enter the rating. Pressing "save" in the navigation bar transitions to the final screen, which shows the (obfuscated) responses. Noise is generated locally by the app (Gaussian noise of zero mean and standard deviation based on the chosen privacy-level), and cannot be changed by the user once it has been generated, to prevent the user from biasing the noise generation via repeated selection. Only obfuscated responses are uploaded by the app to the server.

The server was built using the Django web Framework (written in Python) and uses a MySQL database to store registered user details and surveys. Surveys (and associated filters) are entered into the database via a web-interface. The web-interface can be found on http://loki.eng.unsw.edu.au/. New surveys are pulled into the app each time it is launched by the user. The app interacts with the server via HTTP GET messages (securing the messages via encryption is left for future work). Due to the complexity of setting up a payment scheme, the implementation of a rewards system is left for future prototype versions.

B. Trial and Results

131 student volunteers from our university used our app to rate lecturers in the department. Our choice of survey was motivated by several reasons: (a) all users would know many of the lecturers, so we could get enough data points to enable statistical analysis; (b) we did not want to ask students overly private (e.g., personal health or relationships) or insufficiently private (e.g., movie ratings) questions, as that could bias their privacy choice towards too high or too low, whereas lecturer ratings gave students the right balance between incentive (to give useful feedback on lecturer quality) and risk (revealing their personal likes or dislikes could influence their future grades); and (c) we could corroborate the survey results with some form of ground truth (official ratings of lecturers) and correlate them with known facts (e.g., student grades).

User Perception of Privacy: We evaluated user perceptions of privacy by talking to each participant after the survey, and by analyzing their choices. Most participants said that they liked the way we presented privacy (4 levels), could understand easily how it operated (by addition of Gaussian noise with parameters as described earlier), and felt comfortable that their privacy was protected when they saw their noisy responses on the final screen of the app. Of the 131 students who took the survey, 18 (13.7%) chose no privacy, 32 (24.4%) chose low privacy, 51 (38.9%) chose medium privacy, and 30

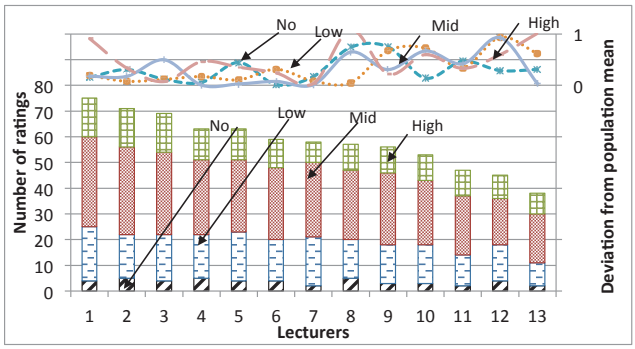


Fig. 4. Variation in mean across the bins for various lecturers

(22.9%) chose high privacy. We believe "medium" was the most popular choice because users perceive it as a "safer" option than any of the extreme values.

Accuracy of the Responses: To validate the accuracy of the responses, we obtained the university ratings (conducted by a trusted third party) for a small handful of lecturers, and found these to corroborate well with those obtained from our system. For example, one author of this paper obtained an average score of 4.72 based on the noisy responses from the 131 users of this system, which is only slightly higher than the average rating of 4.61 (out of 5) he has received from the university system over the past 3 years.

To illustrate how choice of privacy level affects the accuracy of the results, we sorted lecturers in decreasing order of the number of ratings they received, and considered the top 13 lecturers who were rated by at least 50 students. For each such lecturer, we plot in the top of Figure 4 the difference between the mean rating obtained from a given privacy bin and the overall mean rating (considering all bins). The figure also shows a histogram of the number of students rating each lecturer per privacy bin. We found that in general, when the number of students rating a lecturer is high, the mean score from each bin is fairly close to the overall mean. However, as the number of users in each bin falls, the ratings across the bins become more disparate, particularly for higher privacy bins. As pointed out earlier, the standard deviation of the mean inversely correlated with the square root of the number of samples constituting the mean. This trade-off between accuracy and privacy is inevitable, but our study shows that even with a relatively small sample size of 131 participants, the error in estimates is still reasonably small.

VII. Conclusion

In this paper we have proposed, evaluated, and prototyped a platform for crowdsourcing data in a privacy conscious way. We developed an architecture that does not require a trusted broker, allowing users to obfuscate their input based on comfort level, while giving surveyors accurate population estimates within their cost budget. We devised an algorithm that allows the broker to leverage prior user history to select the most suitable set of users for each survey, such that accurate population estimates are obtained within the specified cost budget, while being able to control the fairness in privacy depletion across users. We evaluated our selection mechanism off-line using a large dataset, and showed how the broker can use the fairness parameter in our algorithm to achieve consistent accuracy across successive surveys. We prototyped our system on mobile devices and conducted trials with 130 volunteers, demonstrating that our approach aligns well with user perception of privacy. We hope that our work will motivate crowdsourcing platform providers to integrate privacy protection mechanisms for the benefit of the users who contribute their data to support such platforms.

References

[1] Amazon Mechanical Turk. [Online]. Available: https://www.mturk.com/mturk/

[2] Experimental Psychology: The Roar of the Crowd, The Economist, 26 May 2012.

[3] Google Consumer Surveys. [Online]. Available: http://www.google.com/insights/consumersurveys/home

[4] R. Chen, A. Reznichenko, P. Francis, and J. Gehrke, "Towards Statistical Queries over Distributed Private User Data," in *NSDI*, 2012.

[5] S. Guha, B. Cheng, and P. Francis, "Privad : Practical Privacy in Online Advertising," in *NSDI*, 2011.

[6] K. Ligett and A. Roth, "Take it or Leave it: Running a Survey when Privacy Comes at a Cost," in *WINE*, 2012.

[7] C. Riederer, V. Erramilli, A. Chaintreau, B. Krishnamurthy, and P. Rodriguez, "For sale : Your Data, By : You," in *ACM Workshop on Hotnets*, 2011.

[8] A Face is Exposed for AOL Searcher No. 4417749. [Online]. Available: http://www.nytimes.com/2006/08/09/technology/09aol.html

[9] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in *IEEE Symosium on Security and privacy*, 2008.

[10] B. F. Ribeiro, W. Chen, G. Miklau, and D. F. Towsley, "Analyzing Privacy in Enterprise Packet Trace Anonymization," in *NDSS*, 2008.

[11] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," in *SIGMOD*, 2000.

[12] S. L. Warner, "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," in *Journal of the American Statistical Association*, 1965.

[13] P. Samarati and L. Sweeney, "Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression," CS Laboratory, SRI International, Tech. Rep. SRI-CSL-98-04, 1998.

[14] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramaniam, "l-Diversity: Privacy Beyond k-Anonymity," in *ICDE*, 2006.

[15] S. R. Ganta, S. Kasiviswanathan, and A. Smith, "Composition Attacks and Auxiliary Information in Data Privacy," in *KDD*, 2008.

[16] L. Sweeney, "Simple Demographics Often Identify People Uniquely," *Carnegie Mellon University, Laboratory for Internation Data Privacy, Pittsburgh, PA*, 2000.

[17] D. Agrawal and C. C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," in *PODS*, 2001.

[18] A. V. Evfimievski, R. Srikant, R. Agarwal, and J. Gehrke, "Privacy Preserving Data Mining of Association Rules," in *KDD*, 2002.

[19] A. V. Evfimievski, J. Gehrke, and R. Srikant, "Limiting Privacy Breaches in Privacy Preserving Data Mining," in *PODS*, 2003.

[20] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," in *ICDM*, 2003.

[21] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," in *TCC*, 2006.

[22] C. Dwork, "Differential Privacy: A Survey of Results," in *TAMC*, 2008.

[23] C. Dwork, K. Kenthapadi, F. Mcsherry, I. Mironov, and M. Naor, "Our Data, Ourselves: Privacy via Distributed Noise Generation," in *EUROCRYPT*, 2006.

[24] V. Rastogi and S. Nath, "Differentially Private Aggregation of Distributed Time-Series with Transformation and Encryption," in *SIGMOD*, 2010.

[25] V. Toubiana, A. Narayanan, D. Boneh, H. Nissenbaum, and S. Barocas, "Adnostic : Privacy Preserving Targeted Advertising," in *NDSS*, 2010.

[26] T. Kandappu, V. Sivaraman, A. Friedman, and R. Boreli, "Controlling Privacy Loss in Crowdsourcing Platforms, One Question at a Time," National ICT Australia, Tech. Rep. 1833-9646-7103, 2013.