10-2015

Privacy in crowdsourced platforms

Thivya KANDAPPU
Singapore Management University, thivyak@smu.edu.sg

Arik Friedman

Vijay Sivaraman

Roksana Boreli

Chapter 4
Privacy in Crowdsourced Platforms

Thivya Kandappu, Arik Friedman, Vijay Sivaraman and Roksana Boreli

4.1 Introduction

Crowdsourcing has emerged, in recent years, as a means of outsourcing various tasks to groups of individuals that are recruited online. As defined in [23]:

> Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.

Crowdsourcing is increasingly used in a large number of application areas, from user opinion surveys and other information collection (including, e.g., testing of new designs) to contribution of content, for example, photos or other media, and even funding of new ventures via crowdfunding. A list of crowdsourcing companies and their classification includes over 20 task categories and around 170 companies that provide crowdsourcing activities.[1] These services have been used widely—both academic and market researchers have been increasingly relying on crowdsourcing platforms for conducting surveys, to gain new insights about customers and populations.

[1] http://www.resultsfromcrowds.com/features/crowdsourcing-services/.

T. Kandappu (✉) · V. Sivaraman
University of New South Wales, Sydney, Australia
e-mail: t.kandappu@unsw.edu.au

V. Sivaraman
e-mail: vijay@unsw.edu.au

A. Friedman · R. Boreli
National ICT Australia, Sydney, Australia
e-mail: arik.friedman@nicta.com.au

R. Boreli
e-mail: roksana.boreli@nicta.com.au

S. Zeadally and M. Badra (eds.), *Privacy in a Digital, Networked World*,
Computer Communications and Networks,
DOI 10.1007/978-3-319-08470-1_4

In this chapter, we focus on a narrower set of platforms that deal with collection and aggregation of information, like the Amazon Mechanical Turk[2] (AMT) platform that enables completion of human intelligence tasks or the Google Consumer Surveys platform[3] that enables large-scale market surveys.

In the vast majority of platforms, workers provide the information in a quasi-anonymous way, as there is no direct relationship between the requester (the company that requires completion of specific tasks) and the workers. Although the majority of such platforms use a payment (or micropayment) system, all direct interactions with the requesters are done thorough pseudonyms (worker IDs). Nevertheless, the release of personal information and opinions, albeit in small increments, can over time be accumulated (by the requesters or by the platform) to identify and profile individuals. This gradual loss in privacy may be undesirable for many workers, and even harmful (in social, financial, or legal ways) for some. Furthermore, the threat comes not only from requesters, but also from the platform itself, which can exploit the profiling for its own ends, or cede the information gained about the workers to another entity for monetary gain.

In this chapter we concentrate on the privacy issues of workers in crowdsourcing platforms. After a short overview of crowdsourcing platforms in Sect. 4.2, we start in Sect. 4.3 with a brief review of privacy risks in online systems. We then discuss how these risks apply to crowdsourcing platforms, focusing on the potential for personally identifying information (PII) exposure. These risks are illustrated through an example of a real world attack, conducted through a series of survey tasks in AMT. Following this, we present in Sect. 4.4 an overview of solutions that enhance privacy in online services in general, and which could also be applicable to crowdsourcing platforms.

In Sect. 4.5 we focus on a specific proposal for a privacy-preserving crowdsourcing platform [27] that relies on obfuscation, and describe the design choices surrounding obfuscation techniques, worker privacy levels, privacy loss quantification, worker privacy depletion, cost settings, and worker utility estimation. We also present the implementation details for a prototype of the system. We summarize in Sect. 4.6 the challenges that still need to be addressed to enhance worker privacy in crowdsourcing platforms and conclude the chapter in Sect. 4.7.

4.2 Crowdsourcing Platforms

Crowdsourcing platforms are leveraged to obtain feedback on goods and services, and to collect content or ideas, by soliciting contributions from an online community, rather than from more traditional sources like company employees or suppliers. A classification of crowdsourcing was presented in [47], distinguishing

[2] https://www.mturk.com/mturk/welcome.

[3] http://www.google.com/insights/consumersurveys/home.

between *integrative* crowdsourcing, where clients seeks to build databases or information bases (like data collection or translation of simple texts) and *selective* crowdsourcing, where a problem may be solved by relying on competencies of the crowd-based contributors.

Platforms like AMT, Crowdflower,[4] and oDesk[5] are used to crowdsource from online workers tasks like deciphering images, ranking websites, and answering surveys. AMT, launched in 2005, is extensively used by researchers in experimental psychology to conduct low-cost large-scale behavioral studies by obtaining opinion survey data from paid volunteer populations. Today AMT engages over 500,000 workers from 190 countries.[6] The Google Consumer Surveys platform utilizes a "surveywall" approach, where access to premium content is gated and enabled for users upon completion of survey questions. The Consumer Surveys platform customers include over 130 publishers (online newspapers and magazines) only in the US based market.[7] Crowdsourcing has also gained popularity in the research community. As of March 2014, Google Scholar counts more than 10,000 academic publications that involve crowdsourced experiments via AMT.

Given that mobile devices have become an integral part of people's daily lives, mobile crowdsourcing has also gained high popularity, especially in the area of environmental monitoring. For example, mCrowd [56] is an iPhone-based mobile crowdsourcing platform that enables mobile workers to contribute to geolocation-aware image collection, road traffic monitoring, and so on, which exploit the sensors available on iPhones. Txteagle [14] is a mobile crowdsourcing marketplace used in Kenya and Rwanda for translations, polls, and transcriptions. Waze[8] is a mapping app that relies on users' contributions to provide real-time traffic information. It has 15 million active users who upload their live driving data by default, so others can benefit by seeing the speed at which the contributors are moving. OpenSignal[9] allows its users to map cellular coverage, find Wi-Fi hotspots, test and improve their mobile reception, and obtain faster data rates. OpenSignal has been downloaded 3.7 million times and has about 700,000 active users (at the time of writing).

A crowdsourcing system typically comprises the following actors: *workers* (users, or contributors), who are the individuals forming the crowd that provides the data, or accomplishes selected tasks; *requesters*, who are the companies or individuals that need a set of tasks to be completed; and the *crowdsourcing platform*, which manages the crowdsourcing process, including matching workers to requesters and handling worker compensation.

---

[4] https://crowdflower.com/.

[5] https://www.odesk.com/.

[6] https://www.requester.mturk.com/tour, accessed March 27, 2014.

[7] http://www.forbes.com/sites/stevecooper/2013/03/29/qa-with-paul-mcdonald-co-creator-of-google-consumer-surveys/.

[8] https://www.waze.com/.

[9] http://opensignal.com/.

4.3 Privacy Issues in Online Services

In recent years, a number of real-world attacks have shown the importance of taking privacy into consideration when contributing data in online services. In this section, we discuss the main risks to worker privacy in online services and then focus on specific risks in crowdsourcing services, with a detailed description of an example attack.

We note that privacy issues may exist not only for the workers, but also for the requesters. For example, a company that requires a set of tasks to be accomplished by the crowd, may wish to keep such tasks, or the content of the data it shares, private. This problem can be encountered, for example, in services related to image classification and text translation [53]. However, as the vast majority of studies and real-world breach examples address worker privacy, in this chapter we consequently focus on worker privacy issues.

4.3.1 Risks of Re-Identification

Privacy risks related to public release of anonymized data sets have been demonstrated by a number of real-world events. As a prominent example, in 2006, AOL[10] released a 2 GB file containing 21 million web search queries from 650,000 users, conducted over a period of three months [4]. The consequences of the data release were devastating from the privacy perspective. AOL took down the data within days of publication due to public outcry, but the data has already been downloaded, reposted,[11] and made searchable by a number of sites. In a matter of days, the identity of user 4417749 had been unmasked by *New York Times* reporters [1]. Besides harm to the users whose names and social security numbers were published,[12] the AOL search log release may have had other harmful consequences the extent of which is difficult to assess, such as: loss of user trust in AOL, as well as, possibly, in other search engines; increased anxiety regarding the privacy of online activities for users; and hesitation of other companies to share their data to enable broader innovation [21]. Following the release of this private data set, the CTO of AOL resigned, two employees were dismissed [24], and a class action lawsuit was filed.

Similarly, in 2006, DVD rental company Netflix announced a contest with a $1 million prize for the best movie recommendation algorithm, and made an anonymized dataset of user ratings available to all interested participants [5]. The Netflix prize data release included over 100 million ratings given by over 480,000 users to 17,700 movies. Despite the anonymization of the dataset, Narayanan and

[10] http://www.aol.com/.

[11] See, for example, http://www.gregsadetsky.com/aol-data/.

[12] http://superjiju.wordpress.com/2009/01/18/aol-search-query-database/.

Shmatikov [39] have shown how to de-anonymize several users in the published dataset by cross-correlating anonymized Netflix ratings with non-anonymous movie ratings on the Internet Movie Database (IMDb) website. While the ratings of movies users made on IMDb did not pose privacy risks, as they were made public deliberately by the users, the re-identification of these users in the Netflix dataset exposed also their private ratings on Netflix. The study demonstrated how little auxiliary information is needed for reliable cross-correlation: for example, with eight known movie ratings, 99 % of the records could be uniquely identified; two ratings and their dates are sufficient for re-identification of 68 % of the records.

Overall, prior research, including [39], has shown analytically that even a relatively small amount of background information about an individual can facilitate a fairly reliable de-identification of that individual, in a seemingly well-anonymized dataset.

4.3.2 Risks of Profiling and Data Misuse

The growing amount of information collected about individuals is increasingly utilized for profiling and subsequent targeting of users. For example, the popular loyalty and rewards cards enable retailers to collect details of users' consumption patterns, track their shopping habits, and mine the data to determine users' interests and needs. Australian retailer Woolworths recently stated in an industry publication that it has managed to "overlay" its insurance company's car crash database and its Everyday Rewards statistics, to reveal which consumers were best to target for insurance purchases [54]. Woolworths also shares its anonymous data with Quantium, a company that sells this data to its clients for direct marketing [54].

As a specific example of customer data use, it was shown in [22] how Target can successfully predict whether a female customer is expecting a child. Target assigns every customer a Guest ID number, tied to her credit card, name, or email address, and becomes a depository for her history of purchases and any demographic information collected from her, or bought from other sources. Using this data, Target assigns a score to every female customer to indicate the likelihood that she may be pregnant. More importantly, it can also estimate the due date, so that coupons can be timed to very specific stages of a customer's pregnancy.

In a second example, the US based political media firm, Engage,[13] is able to predict who users will vote for, how likely users are to go to the polls, and the potential for them to change their vote. They have reported, during the previous US elections, that if users use Spotify to listen to music, Tumblr to consume content, or Buzzfeed to keep up on the latest in social media, there is a high likelihood that they will vote for President Obama. On the other hand, if they buy things on eBay, play

[13] http://enga.ge/.

FarmVille, or search the Web with Bing, they are more likely to favor Mitt Romney.

Finally, a team of British researchers have developed an algorithm that uses tracking data from people's phones to predict where they will be in 24 h [40], with an average error of just 20 meters. The researchers combined tracking data from individual participants' phones with similar data from their friends, that is, other people in their contact list.

While these examples represent a small fraction of the ways in which companies are using data to predict user behavior, the proliferation of personal data is likely to drive a rapid increase in the business of prediction. As more of user movements, browsing patterns, purchase history, and social media interactions become recorded, more companies will find ways to use this data to profile users and exploit this knowledge for profit. These capabilities incentivize a multi-billion dollar industry of data brokers to collect and sell personal user data, with little or no transparency, and often without the knowledge and consent of the individuals to whom the data pertains [43].

4.3.3 Privacy Issues in Crowdsourcing Platforms

The examples outlined in the previous sections relate to online services in general. In this section, we consider specific risks in crowdsourcing platforms and provide an example of how a requester can attain knowledge about the personal details of workers in an anonymous crowdsourcing system.

4.3.3.1 The Lack of Worker Anonymity Guarantees

In a technical report, Lease et al. [32] have identified a direct loss of worker anonymity on AMT. In AMT, requesters and workers are identified with a 14-character alphanumeric string. However, Lease et al. have observed that the same string that identifies a worker in AMT is also the unique identifier of that account across all Amazon services. Therefore, any public information associated with an Amazon account, such as name and picture on the public Amazon profile, product reviews and ratings, or a wish list, will be easily accessible via that account's Web URL.[14] The use of the same account to access both AMT and other Amazon services allows workers to use the proceeds for their AMT work towards purchases on Amazon's website. Moreover, Lease et al. pointed out that the term "anonymous" has never been used on AMT's website and policies, and while these policies express Amazon's concern for workers' privacy, they do not state explicit guarantees of worker anonymity. However, it is unclear whether workers are aware

[14] www.amazon.com/gp/pdp/profile/<WorkerID>.

of the tight connection between their alphanumeric identifier on AMT and their public information on other Amazon services. In fact, a thread on Turker Nation (a forum dedicated to AMT discussions), predating Lease et al.'s work, reflects the surprise of workers who learned about this relation.[15]

4.3.3.2 De-Anonymization and Privacy Loss via Inference Attacks

Lease et al. [32] suggested that having worker IDs that are not linked to other (Amazon) services may mitigate the current direct anonymity loss issue for workers. However, such measures may not be sufficient to eliminate all threats to worker anonymity and privacy. Kandappu et al. [26] have shown how privacy risks like those explored in Sect. 4.3.1, could easily apply to existing crowdsourcing platforms to de-anonymize workers and obtain sensitive private information, in a short time period and at very low cost, by correlating responses across multiple surveys.

The inference attack in [26] comprised of launching a series of survey tasks in AMT (through the third-party aggregator CrowdFlower[16]). The first survey queried workers for their opinions on astrology services, and in the process obtained their star-sign and day/month of birth. The second survey purportedly conducted market research of online match-making services, and thereby obtained the workers' gender and year of birth. With the third survey, on mobile phone coverage, the researchers obtained workers' zip code information.

The surveys were designed with sufficient redundancy to help identify and filter out workers who gave random responses. Further, these surveys were posted independently over several days, and workers were unlikely to have known that they were conducted by the same entity. The researchers used the unique IDs (constant across all surveys) to link workers who took all the three surveys above and to obtain a combination of their personal details, that is, their date of birth, gender, and zip code. We note that previous studies [20, 51] have shown the effectiveness of using these attributes in re-identification of individuals.

A fourth survey was then launched, asking workers about their smoking habits and coughing frequency. Overall, of the 400 unique workers who took the surveys, 72 could be linked from the first three surveys, and the respiratory health (and likelihood of tuberculosis) for 18 of these individuals could be inferred from the fourth survey using their unique ID, resulting in a potentially serious breach of privacy. This experiment took only a few days and cost less than $30; one can only imagine what the scale of privacy loss could be, were this experiment to be conducted by entities with larger resources.

Finally, the above experiment was followed up with another survey, where workers were asked if they would participate in a survey, if they knew they could be

[15] http://turkernation.com/archive/index.php/t-6065.html.

[16] https://crowdflower.com/.

de-anonymized and profiled. Out of 100 workers who took this survey, 73 (including 15 of the 18 workers above, whose respiratory health could potentially be made public) responded that they were not aware that they could be profiled, and indicated that they would not have participated otherwise. These experiments illustrate that workers can be profiled easily and at low cost, despite their disapproval of such practices.

The release of personal facts and opinions, albeit in small increments, can over time be accumulated (by the requesters or by the platform) to profile individuals (e.g., the work carried out in [30] shows that a wide variety of people's attributes can be accurately inferred using their Facebook likes). This gradual loss of privacy may be undesirable for workers, and even harmful (in social, financial, or legal ways) for some. Furthermore, the threat comes not only from requesters, but also from the platform itself, which can exploit the profiling for its own ends, or cede it to another entity for gain.

We note that AMT's policies[17] explicitly forbid using the platform to collect personally identifiable information, or requiring workers to disclose their identity, directly or indirectly. However, partial information (e.g., gender or age), which is not sufficient for identifying an individual on its own, may be legitimately acquired for the purpose of a specific survey. Despite the policy restrictions, it may be difficult to track and enforce limitations on subsequent combining and (mis)use of this information.

4.4 Overview of Existing Solutions

We now present an overview of recent technological advances in defining and protecting individuals' privacy and data confidentiality (visibility of the data values used for aggregation) in data publishing and aggregation. We note that in crowdsourcing, as in most services that rely on users' data, there is a need to balance the privacy of individual participants with the greater good for which the aggregate data can be used.

4.4.1 Anonymization

Early research works on data anonymization proposed sanitizing user data by masking or removing PII such as name and address, and quasi-identifiers such as gender and zip code. *k*-anonymity [46, 52] takes a "blend into the crowd" approach to privacy, and requires that every combination of quasi-identifiers appears in at least *k* data instances. This is achieved by generalization of such identifiers, for

[17] https://www.mturk.com/mturk/help?helpPage=policies, accessed March 27, 2014.

example, by limiting the zip code to four or fewer, rather than five, recorded digits. Further refinements of *k*-anonymity include *l*-diversity [37], *t*-closeness [33], and other variants, which introduce additional restrictions on the released data values. It was demonstrated, however, that such intuitive anonymization techniques are not effective in protecting user privacy, as individual users can be re-identified via the use of background information [10, 39, 44, 51], as shown by the AOL and Netflix data release examples from Sect. 4.3.1. To date, safe release of anonymized data for analysis purposes is still an open research problem. In addition, using such techniques in crowdsourcing scenarios may not be practical, as users need to be identifiable so that they can be compensated for contributing their data.

4.4.2 Data Obfuscation

Data obfuscation techniques protect user privacy by perturbing the data contributed by individuals.

4.4.2.1 Randomized Response

A traditional method to obfuscate data is by randomization—this can be done by adding noise sampled from a selected distribution, by multiplying with noise or by projecting the data, to alter the individual values of the records. This method relies on the ability to recover the probability distribution of the aggregate (non-noisy) data, which can subsequently be used for data analysis. The earliest work on randomization was presented in [34, 55], where it was used to eliminate the untruthful answer bias. A generic approach proposed in [2, 3] is to add random distortion values drawn independently from a known distribution, for example, the uniform distribution. A number of improvements to this technique were subsequently proposed [15, 16].

We note that randomization methods apply noise to the records in a *data-independent way*, thereby this technique can be utilized at the *source* of data collection. Thus, perturbation of the records does not require a trusted server. However, it was shown 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 [29].

4.4.2.2 Differential Privacy

Differential privacy [13] is a privacy model based on the principle that the output of a computation should not allow inference about any record in the input, irrespective of an adversary's computational power or the available background knowledge. This guarantee is obtained by constraining the effect that any single record could

have on the outcome of the computation. Consequently, the promise of differential privacy is that the probability of a "bad" outcome resulting from a computation on the data will be almost unaffected by the specific value of any particular record in the dataset. Yet in aggregate, these records would still provide useful information. Formally, a mechanism K provides (ε, δ)-differential privacy [12] (or simply ε-differential privacy for $\delta = 0$) if for any two datasets A and B differing in a single record, and for all outcomes S:

$$\Pr[K(A) \in S] \leq exp(\varepsilon) \times \Pr[K(B) \in S] + \delta. \tag{4.1}$$

The parameter ε controls the level of privacy, where smaller values of ε provide stricter bounds on the influence of any particular input record on the outcome, and therefore provide better privacy—adding or removing any particular record would hardly change the probability of obtaining a given outcome, so the outcome would not reveal much about any underlying record. The parameter δ allows the condition in Eq. (4.1) to be relaxed for unlikely events, allowing ε-differential privacy to be breached in some rare cases. One of the prevalent methods to achieve differential privacy is by adding noise to the outcome of a computation. The noise is calibrated according to the influence that any record may have on this outcome such that Eq. 4.1 holds, as further described in Sect. 4.5.2.2. Differential privacy maintains composability, that is, if two computations maintain $(\varepsilon_1, \delta_1)$ and $(\varepsilon_2, \delta_2)$ differential privacy respectively, then executing both would amount to $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$ differential privacy.

The practical implications of differentially private analysis were studied in many application domains, including network trace analysis [38], intelligent transportation systems [28], collaborative security mechanisms [42], and distributed stream monitoring [17]. Most applicable to the *crowdsourcing* scenarios are the distributed differential privacy mechanisms [12, 41], that provide strong privacy guarantees in distributed settings. Rastogi and Nath in [41] designed a two-round protocol based on the threshold homomorphic cryptosystem, and Shi et al. in [48] applied cryptographic techniques to allow an untrusted aggregator to compute sums without learning anything about the user inputs. Both designs presented in [41, 48] achieve distributed differential privacy while reducing the computational load per user. These systems leverage cryptographic techniques to generate differentially private noise in a distributed manner, but unfortunately do not scale well. In Sect. 4.5 we describe in greater detail a system that relies on differential privacy to track privacy loss in a crowdsourcing scenario.

4.4.3 Cryptographic Mechanisms

Cryptographic mechanisms are commonly used in conjunction with obfuscation to achieve both data confidentiality and privacy [41, 48]. Chen et al. [8] proposed a system that performs statistical queries over private client data (distributed on local

databases, e.g., on client devices), where the analyst communicates with clients via an honest-but-curious proxy. On first connection, a proxy assigns a unique ID to a client. Answers are provided as binary values corresponding to a set of buckets, that is, the potential values that a query may result in. Each binary value is encrypted using the Goldwasser-Micali (GM) cryptosystem [19], a probabilistic public-key cryptosystem that ensures each encrypted value is represented by a different cyphertext. The analyst combines the decrypted client answers (that are also obfuscated using a differentially private mechanism) to produce the result. The authors extended these concepts in [7] to propose the SplitX system. This again includes an analyst, a set of clients who locally store their data and a set of intermediate entities: an aggregator and two mixes. However, SplitX uses a simple XOR-based crypto-mechanism and a series of split messages (each message is split into two, which are sent in parallel to any of the two intermediate nodes). This provides the additional properties of anonymity and unlinkability and enables considerably improved system performance.

A different approach to enabling confidentiality is to apply a secret sharing cryptographic mechanism to the distributed private data, and perform Secure Multi-Party Computation (MPC). However, this approach is only resilient to a specific proportion of honest-but-curious attackers who collude to learn the private data and/or the result of aggregation. MPC has been applied to crowdsourcing platforms using the Sharemind implementation, as described in [6].

4.4.4 Compensating Users for Privacy Loss

Rather than limiting privacy loss when collecting and using personal information, an alternative approach is to accept this loss and compensate the users accordingly, so that they are incentivized to share information. Laudon [31] proposed a market for personal data, which relies on individual ownership of this information. In fact, several start-ups, such as Reputation.com [50], Handshake [36], and Datacoup [49], are endeavoring to make such markets a reality.

Further to this, Ghosh and Roth [18] initiated a study of markets for private data, where the privacy of users, as measured by differential privacy, is the sold commodity. Specifically, they consider a setting where the data is binary, and the aggregator wishes to estimate the sum of bits. They proposed the *FairQuery* mechanism, which achieves the optimal accuracy given a budget B, among the set of all truthful, individually rational envy-free fixed purchase mechanisms. Dandekar et al. [11] generalized these results to linear predictors (with inputs in $\mathbb{R}^n$), and observed that while these settings are similar to the knapsack auction mechanism, they also pose the challenge that privacy costs exhibit externalities. That is, the privacy cost of an individual depends also on which other individuals are being compensated.

One of the challenges highlighted by Gosh and Roth [18] is that the data collector will only get the information of individuals who value their privacy at a lower

cost than that offered by the buyer. This introduces a selection bias, which could lead to inaccurate results. Another challenge is that an individual's cost for privacy may be correlated to private information, and therefore might be itself private information. In fact, they showed that, in general, it is not possible for any individually rational direct revelation mechanism to compensate individuals for their privacy loss due to unknown correlations between their cost functions and their private data. Ligett and Roth [35] proposed to circumvent this impossibility result by considering a "take-it-or-leave-it" framework, where a surveyor randomly samples members of the underlying population, and offers them the same price in return for participating in the survey. These offers are repeated with fresh population samples and with increasing prices, until a sufficient rate of participation is obtained. This model captures that individuals may also experience a cost when information about their cost function (i.e., the cost they associate with privacy loss) is revealed. The model also captures that individuals can suffer negative utility even when they choose not to participate in surveys, as this choice may be correlated with their private information.

All the aforementioned works [11, 18, 35] assumed that users cannot lie about their private information, but can lie about their costs. The truthful mechanisms ensure that individuals do not mis-report their cost functions in an attempt to maximize their payment. In contrast, Chen et al. [9] did not assume that players provide truthful answers. Instead, they considered settings where users may choose to lie, but also have a direct interest in the outcome of the mechanisms. They explicitly modeled privacy in the participants' utility functions, and designed truthful mechanisms with respect to it. The mechanisms leverage the users' interest in the outcome, such that the payoff overcomes the users' value for privacy. Essentially, the privacy parameter should only be set to be small enough such that the privacy costs are outweighted by the participants' preferences for outcomes.

Finally, Riederer et al. [45] introduced a mechanism of transactional privacy, which enables end-users to sell or lease portions of their personal information (on a strictly opt-in basis) in exchange for monetary compensation. This compensation is determined in an auction, where data aggregators place bids based on their valuation of the user's information. The users can then decide what and how much information will be disclosed to the aggregators, and the data can be sold multiple times.

4.5 Loki: Privacy Preserving Crowdsourcing Platform

In Sect. 4.4 we surveyed general techniques for protecting user privacy in data analysis, which are also applicable to crowdsourcing platforms. In this section we consider in depth "Loki," a system proposed by Kandappu et al. [26, 27], which focuses on facilitating the crowdsourcing applications in a privacy preserving way.

4.5.1 Architecture and Entities

The proposed system (Fig. 4.1) comprises three entities: requesters, workers, and the broker platform.

Requesters acquire data from workers using a set of questions in a survey form (the work largely focuses on ratings-based questions and multiple-choice questions). The requester pays the broker to run the survey, specifying an upper bound on total cost. The requester aims for high accuracy (utility) in the aggregated response for any survey, so that it closely represents the feedback of the entire population.

Workers respond to questions in the surveys, using a supplied application (app) installed on their personal device (smart phone/tablet). The app allows workers to obfuscate their responses at source. The workers' monetary compensation may in general depend on their choice of privacy level—higher privacy levels entail higher obfuscation and hence lower payment. Loki does not deal with intentional lying (or cheating) by workers to get higher compensation; however, lying may make the worker a worse predictor of the population average, reducing the chances that the algorithm (described in Sect. 4.5.3.4) will select this worker for subsequent surveys, thereby offsetting the monetary gains from cheating.

The **broker** provides a platform for launching surveys. It receives payment from requesters, and passes it on to workers (less a commission). The broker has a dual

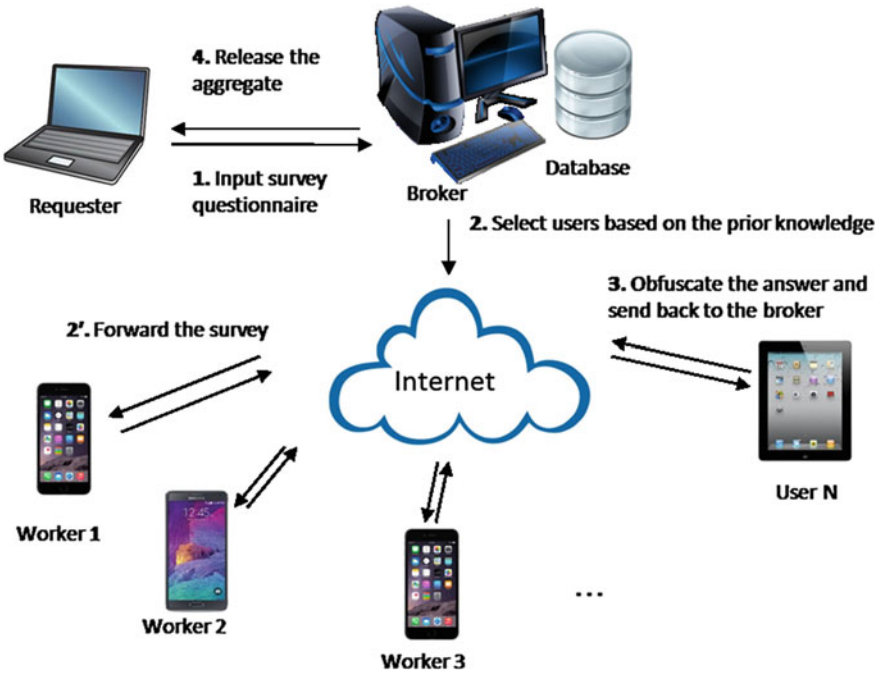

Fig. 4.1 Loki: System components and the basic protocol

objective: to provide accurate population estimates to requesters, and to extend the lifetime of workers in the system. The broker keeps track of workers' performance (i.e., how good a predictor of population behavior each worker has been in the past) and privacy (cumulative depletion of a privacy "budget" due to participation in surveys), so it can balance the trade-off within the cost budget.

4.5.2 Design Choices

4.5.2.1 Obfuscation and Worker Privacy Levels

In Loki, the worker 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 worker response. Gaussian distribution was chosen over uniform as it has unbounded range, and hence does not compromise worker privacy in boundary cases. It was preferred over Laplace noise since it is additive, that is, 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 is adjusted based on the worker's privacy chosen privacy level. For *multiple-choice* questions, Loki relies on the randomized response technique [55], whereby the worker'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 worker's chosen privacy level, described next.

For the sake of simplicity, Loki uses 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 worker response. The higher the privacy level, the larger the obfuscation parameter (γ or p above).

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). For 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.

In general, the worker can set the desired level of privacy for each conducted survey. For simplicity, we assume that the worker's choice is consistent across surveys (i.e., the worker tends to choose the same level of privacy for each survey), but our selection algorithm (described in Sect. 4.5.3) can be easily modified to adapt based on different user choices.

4.5.2.2 Quantification and Tracking of Privacy Loss

Loki quantifies the privacy loss for a worker who answers a particular survey at a particular privacy level, so it can be accumulated and tracked across multiple surveys. For this purpose, Loki relies on differential privacy, where the differential privacy constraint is applied to each survey answer. That is, the parameters ε and δ capture how easy or difficult it is to infer the original user response given the noisy survey response. 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 [12]:

$$\frac{\varepsilon\gamma^2}{2R^2} + \ln(\varepsilon\gamma^2) \geq \ln\frac{1}{\delta}, \tag{4.2}$$

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

Example 2 Following from the previous example, 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 $\varepsilon = \{\infty, 3.42, 0.85, 0.21\}$ respectively.

For *multiple choice* questions (with n options) obfuscated using the randomized response technique, the mapping from the probability measure p to (ε, δ) can be derived from (4.1) as:

$$\varepsilon \geq \ln(1 - p - \delta) - \ln(p) + \ln(n - 1). \tag{4.3}$$

Example 3 Following from the previous example of a multiple choice question with five options, the 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 $\varepsilon = \{\infty, 3.57, 2.22, 1.77\}$ respectively.

The differential privacy metrics are composable (i.e., additive), and the worker's privacy loss over successive surveys can therefore easily be upper bounded by accumulating these metrics over the worker's lifetime. These upper bounds capture the relative privacy loss for each of the workers, which the broker can rely on to ensure a fair distribution of the privacy loss across workers. In the rest of this chapter, we will fix the value of δ at 0.01, and use ε for comparing privacy loss across workers. Further, for cases where workers choose privacy level "none," ε is set to 0 (rather than the theoretically correct value of ∞), since the workers 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.

4.5.2.3 Cost Settings

A worker i, who contributes data in response to a survey questionnaire, receives a compensation c_i. Workers 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 4 Following from the previous example that uses a 5-point Likert scale, the privacy levels none, low, medium, and high could correspond to worker 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.

4.5.2.4 Worker History and Utility

Despite noise addition by workers to obfuscate individual answers, some characteristics of worker behavior can be discerned by the broker over time. As an example, noise added by a worker to n successive ratings-based questions, each with iid noise $\mathcal{N}(0, \gamma^2)$, can be averaged by the broker to estimate the worker'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 worker's ratings, that is, to determine on average how close the worker's ratings in the past have been to the population averages. This in turn indicates how representative this worker is of the general population, and helps the broker estimate the "value" of the worker towards obtaining an accurate population estimate. In Sect. 4.5.3, this notion of worker "value" is leveraged to select workers for each survey, in a balanced way.

4.5.3 Privacy-Preserving User Selection Mechanism

This section describes a practical method for the broker to select workers 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 authors' full version [25].

4.5.3.1 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 requester, the broker can query only a subset of workers S from the universal set of workers U, and this selection is based on accuracy, cost, and privacy depletion.

Denote by $x_i \in \mathbb{R}$ the input of worker $i \in U$. The desired population average θ is given by $\theta = \sum_U x_i/|U|$. The broker estimates this statistic by sampling a subset of

workers S. Further, each worker 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 worker's chosen privacy level. The broker's estimate $\hat{\theta}$ of the population average is then given by

$$\hat{\theta} = \sum_S \hat{x}_i/|S| = \sum_S (x_i + n_i)/|S|. \tag{4.4}$$

The mean squared error in the estimator is given by:

$$RMSE^2 = (\hat{\theta} - \theta)^2 = \left[\frac{\sum_S n_i}{|S|} + \left(\frac{\sum_S x_i}{|S|} - \theta\right)\right]^2. \tag{4.5}$$

When selecting S, the broker therefore accounts for two influencing factors: the level of privacy required by each worker, which determines the error due to privacy-related noise (first term above), and the expected sampling error (second term above).

The "value" of a worker depends on how accurately the worker's responses reflect those of the population at large. To quantify this, consider the worker error, i.e., the difference Δ_i between the worker's response and the true population average, given by $\Delta_i = x_i - \theta$. Treating the worker 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 worker using:

$$\mu_i = \mathbb{E}[\Delta_i] = \sum_{j:x_{ij} \in H_i} (x_{ij} - \theta_j)/|H_i|, \tag{4.6}$$

$$\sigma_i^2 = \mathrm{V}ar[\Delta_i] = \sum_{j:x_{ij} \in H_i} (x_{ij} - \theta_j - \mu_i)^2/|H_i|, \tag{4.7}$$

where θ_j denotes the true population average in a past survey question q_j. New workers can be assigned a default value of worker error.

Similarly, we can define the value of a group of workers S. The average rating by the group is defined as $x_S = \sum_S x_i/|S|$. 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:

$$\mu_S = \mathbb{E}[\Delta_S] = \sum_{j:x_{Sj} \in H_S} (x_{Sj} - \theta_j)/|H_S|, \tag{4.8}$$

$$\sigma_S^2 = \mathrm{V}ar[\Delta_S] = \sum_{j:x_{Sj} \in H_S} (x_{Sj} - \theta_j - \mu_S)^2/|H_S|. \tag{4.9}$$

The estimation of the worker and group errors above assumes perfect knowledge of the true worker responses x_i and the population averages θ_j. In reality the broker only has the noisy worker/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 $\left(\sigma_S^2\right)$ of the true group error can be approximated with the mean $(\hat{\mu}_S)$ and variance $\left(\hat{\sigma}_S^2\right)$ of the computed errors, using the fact that the noise is independent of worker responses and has zero mean:

$$\hat{\mu}_S \approx \mu_S, \tag{4.10}$$

$$\hat{\sigma}_S^2 \approx \sigma_S^2 + \frac{\sum\limits_S \gamma_i^2}{|S|^2} + \frac{\sum\limits_U \gamma_i^2}{|U|^2}. \tag{4.11}$$

The expectation of the error in Eq. (4.5) is then derived as:

$$\begin{aligned} \mathbb{E}(RMSE^2) =& \mathbb{E}\left[\left(\frac{\sum\limits_S n_i}{|S|}\right)^2\right] + \mathbb{E}\left[(x_S - \theta)^2\right] = \\ =& \frac{\sum\limits_S \gamma_i^2}{|S|^2} + \sigma_S^2 + \mu_S^2 \approx \hat{\mu}_S^2 + \hat{\sigma}_S^2 - \frac{\sum\limits_U \gamma_i^2}{|U|^2}. \end{aligned} \tag{4.12}$$

4.5.3.2 Balancing Cost and Accuracy in a Single Survey

As described in Sect. 4.5.2.3, each worker chooses a privacy setting, which incurs a privacy cost $(\varepsilon_i, \delta_i)$. 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 worker choices, the broker proceeds to select a group of workers to be included in the survey, based on two constraints:

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

Privacy constraint. For each worker, the cumulative privacy loss throughout the system lifetime is capped at $(\varepsilon_{max}, \delta_{max})$. Each worker i in survey j incurs a known privacy cost $(\varepsilon_{ij}, \delta_{ij})$. The accumulated privacy loss for worker i is therefore $(\sum_j \varepsilon_{ij}, \sum_j \delta_{ij})$ where the summation is over all the past surveys taken by this worker. The residual privacy budget for the worker is consequently $(R_i^{(\varepsilon)}, R_i^{(\delta)})$, where $R_i^{(\varepsilon)} = \varepsilon_{max} - \sum_j \varepsilon_{ij}$ and $R_i^{(\delta)} = \delta_{max} - \sum_j \delta_{ij}$. To guarantee that the worker's cumulative privacy loss stays within the lifetime privacy budget, the broker must ensure that for the new survey, $\varepsilon_i \leq R_i^{(\varepsilon)}$ and $\delta_i \leq R_i^{(\delta)}$.

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

$$\begin{aligned} &\arg\min_{S \subseteq U} RMSE \\ &\text{s.t.} \sum_j n_j c_j \leq C \text{ and } \forall i \in S : \varepsilon_i \leq R_i^{(\varepsilon)} \wedge \delta_i \leq R_i^{(\delta)}, \end{aligned} \tag{4.13}$$

where the RMS error is obtained from Eq. (4.12). For the special case when a worker chooses a "no privacy" setting, which in theory translates to an unconstrained loss in privacy ($\varepsilon \to \infty$), we make the practical choice of using $\varepsilon = 0$, $\delta = 0$, reflecting that the worker is not concerned about the privacy implications in this case.

Note that the upper bounds ε_{max} and δ_{max} are used to capture the relative privacy loss for each of the workers, which the broker relies on to ensure a fair distribution of the privacy loss across workers. Workers whose privacy budget is exhausted can be given a new identity which is unlinked to the previous one, and a new privacy budget, allowing them participation in future surveys. Another possible option is to increase all the workers' privacy budgets once a significant portion of the workers deplete their budget. Regardless of the broker's policy, workers can always choose to quit the system when they deem their cumulative privacy loss too high.

4.5.3.3 Balancing Cost, Accuracy, and Privacy Fairness 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 requesters, while *fairness* aims to balance the residual level of privacy across workers, since privacy can be seen as a non-renewable resource, which should be equally depleted across workers. QoS considerations may motivate the broker to select for a survey workers with low error, but this may deplete such workers' privacy budget rapidly. Consequently, those workers may be excluded from participation in subsequent surveys, resulting in deterioration of QoS over time.

To control the influence of QoS and fairness considerations, a "fairness parameter" $\alpha \in [0, 1]$ is set by the broker. The monetary and privacy cost of worker i are then combined into an overall cost F_i, given by:

$$F_i = (1 - \alpha)\frac{c_i}{C} + \alpha \cdot \max\left[\frac{\varepsilon_i}{R_i^{(\varepsilon)}}, \frac{\delta_i}{R_i^{(\delta)}}\right]. \tag{4.14}$$

The first term considers the monetary cost of the worker for this survey, as a fraction of the budget available for the survey. The second term considers the

privacy depleted by this worker's participation in the survey, as a fraction of the worker's residual privacy budget. 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 workers 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.

4.5.3.4 Algorithm for Worker Selection

For a new survey, Algorithm 4.1 is executed to select the set of workers who yield the best accuracy within the given cost constraint, while also maintaining fairness in privacy depletion among workers. The initial construction of this set assumes that (a) all selected workers will actually take the survey, and (b) Loki can correctly predict the privacy level choice of each worker 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 worker feedback.

Evaluating all possible subsets $S \subseteq U$ of workers to determine the optimum would be intractable. Instead, Loki uses a greedy heuristic approach, by which the broker constructs the set S incrementally, each time adding the worker who would be most cost effective, while taking into account the QoS and fairness considerations. Given a set of workers $S \subseteq U$, Eq. (4.12) evaluates the expected error $RMSE^{(S)}$ of the set, based on past performance. Adding the worker 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 worker i in the set. We can then compute β_i, the improvement in RMS error per unit of cost, for the worker i:

$$\beta_i(S) = \frac{\Delta\, RMSE(S, i)}{F_i}, \tag{4.15}$$

where the worker cost F_i is given by Eq. (4.14) and includes both monetary and privacy costs. The broker bootstraps the algorithm by choosing the worker with the highest accuracy gain per unit of cost. Then in the greedy selection process, the broker picks the worker with the highest β_i at each step. By starting with an empty set of workers, and iteratively adding workers one by one, the broker can construct the target set S, until the monetary cost limit C is reached. Note that workers who have depleted their lifetime privacy budget are not eligible for selection. Algorithm 5.3.4 has complexity $O(KN^2)$, where K is the number of items that constitute prior history and N is the number of workers.

Algorithm 4.1 Greedy Worker Selection Mechanism

Input: U: a set of workers, each with cost c_u; C: overall cost bound.
Output: $S \subseteq U$: a set of survey participants.
$S \leftarrow \emptyset$.
$P \leftarrow \{i \in U : c_i \leq C \wedge \varepsilon_i \leq R_i^{(\varepsilon)} \wedge \delta_i \leq R_i^{(\delta)}\}$. ▷ candidate workers 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.

4.5.4 Evaluation

To study the trade-offs between cost, utility, and fairness, and the long-term system performance, the algorithm was evaluated using the Netflix dataset,[18] a large dataset of movie ratings, as a survey answer set [27]. The dataset contains over 100 million movie ratings (on a 5-point scale) from 480,000 anonymized Netflix customers over 17,000 movie titles, collected between October 1998 and December 2005. The movies released in 2004 (1,436 in number) were used as historical information, and the objective was to estimate the population-wide average rating of movies released in 2005 within a specified cost budget C.

The experiments assumed a simple model of privacy choice, in which each worker was 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 the experimental study with real users, described in Sect. 4.5.4.1. While this experimental setup is different from the one discussed in Sect. 4.5.4.1, this allows us to evaluate performance on the basis of a privacy preference breakdown observed in real-world settings. In general, different settings can induce different preference distributions. The bins were associated with zero-mean Gaussian noise with standard deviations $\gamma = 0, 3, 6, 12$ respectively (corresponding to $\varepsilon = 0$, $\varepsilon = 3.42$, $\varepsilon = 0.85$, and $\varepsilon = 0.21$), and respective payments of \$0.8, \$0.4, \$0.2, and \$0.1 for each worker. Noise sampled from $\mathcal{N}(0, \gamma^2)$ was added to each of the workers' movie ratings.

Within these settings, Loki's selection mechanism was evaluated for different values of the fairness control parameter α. Figure 4.2(a) shows the estimation error $\mathbb{E}(RMSE)$ for varying values of the available budget C and for various selection

[18] http://www.netflixprize.com/.

policies. For different α values, both the true error (i.e., difference between the estimate and the ground truth available in the dataset), depicted with solid lines, as well as the corresponding estimated error (computed using Eq. 4.12), depicted with dashed lines, are shown in Fig. 4.2(a). The estimated error closely reflects the true error, and is hence of sufficient accuracy to be useful in the selection decision. The figure shows also two baseline selection strategies: random selection, in which a random set of workers is selected subject to the cost constraint, and a "best predictors" selection, in which the subset of the population that has the highest historical accuracy (i.e., is most representative of the population) is selected subject to the cost constraint. As can be expected, random selection of workers resulted in the lowest accuracy (users who are "bad predictors" or who provide very noisy answers to protect their privacy are just as likely to be chosen as more "valuable" users), and the selection of "best predictors" consistently yielded 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, the error increases.

The loss in accuracy is compensated for in privacy fairness. To evaluate the performance in a series of surveys, the selection algorithm was applied, sequentially, to a set of 500 movies released in 2005, again using the movies from 2004 as prior history. Figure 4.2 shows the privacy depletion, for various α settings. When α

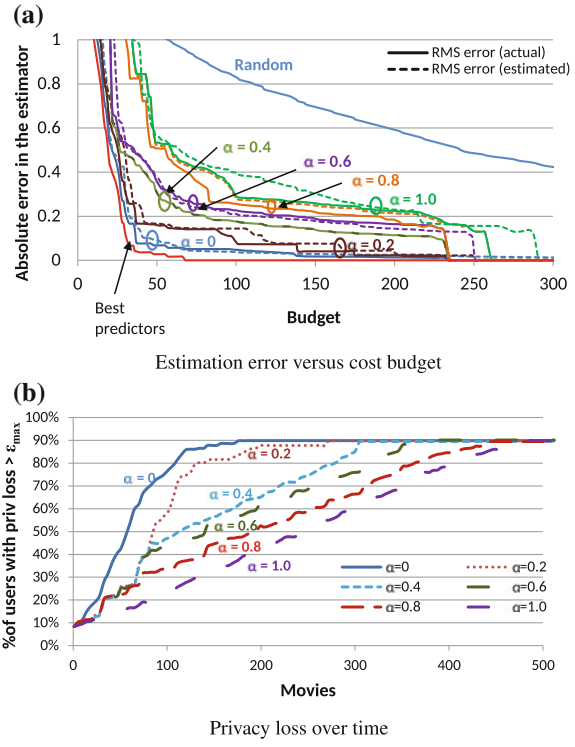

Fig. 4.2 Impact of the selection policy on (**a**) accuracy versus cost in a single survey, and (**b**) the estimation error over multiple surveys

is low, the error is initially low, but rises rapidly with successive movies. This happens because the best performing workers are selected for the initial movies (yielding low error), but deplete their privacy rapidly. Conversely, a choice of high α results in fairer depletion of privacy, prolonging the lifetime of workers 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. Note that the algorithm allows this parameter to be chosen by the surveyor on a per-survey basis.

4.5.4.1 Prototype Implementation

Loki was also implemented as a prototype to evaluate the system with real users, in an experiment involving 131 volunteers [27].

The prototype[19] consists of a front-end application for workers to participate in surveys (screenshots of the iPhone app are shown in Fig. 4.3), and a back-end database/server that stores worker data and communicates with the app. Gaussian noise is generated locally by the app using the Box-Muller method, and obfuscated responses are uploaded to the server. The *no*, *low*, *medium*, and *high* privacy settings correspond to $\gamma = 0, 3, 6,$ and 12 and $\varepsilon = 0, 3.42, 0.85$ and 0.21 respectively. The server was built using the Django (Python) Web Framework and uses a MySQL database to store worker details and surveys.

The system was trialed with 131 volunteers, all 3rd and 4th year undergraduate students studying Electrical Engineering at UNSW. Of the 131 students who took a lecturer assessment survey, 18 (13.7 %) chose no privacy, 32 (24.4 %) chose low privacy, 51 (38.9 %) chose medium privacy, and 30 (22.9 %) chose high privacy. Medium was chosen by most since users perceived it as a "safer" option than any of the extreme values.

The accuracy of the responses was validated by comparing them to the university-run rating mechanism, and by comparing the ratings across the privacy bins in the system. In general, the standard deviation of the mean lecturer rating falls with the square root of the number of samples constituting the mean. The evaluation showed that even with a relatively small sample size of 131 participants, the error in estimates was still reasonably small.

4.6 Challenges and Opportunities

Crowdsourcing is an emerging and promising model for information gathering and problem solving that is already transforming industry and scientific practices, allowing researchers access to human resources in a scope that was not possible

[19]Available at https://itunes.apple.com/au/app/loki/id767077965?mt=8.

Fig. 4.3 Screenshots of iPhone app showing (**a**) list of surveys and privacy levels available to the workers, (**b**) the questions and ratings entered by the worker, and (**c**) the worker responses uploaded after noise addition

before. Unfortunately, while the opportunities of crowdsourcing are still being explored, little attention has been given to the privacy implications that crowdsourcing platforms may impose for their workers. The topic of privacy in online systems has earned much research attention in recent years, but the study of privacy in the context of crowdsourcing systems is still in its infancy. We outline below some of the challenges that are yet to be addressed in this area.

Understanding privacy risks in crowdsourcing platforms. The research community has come a long way in the last few years in understanding the implications of the "Big Data" revolution on users' privacy in online services, and the power of data mining tools to uncover information in ways that may break users' expectations of privacy. Crowdsourcing platforms, which provide the ability to solicit information from a large community of workers, are not devoid of these risks, yet only little work has been conducted to understand how such risks apply to crowdsourcing. Exploration of existing crowdsourcing platforms and the privacy risks involved in using them would be vital for educating both workers and requesters on these risks and for designing proper privacy-enhancing mechanisms in crowdsourcing platforms.

Understanding the role of anonymity in crowdsourcing platforms. Among the privacy risks involved in the use of crowdsourcing platforms, anonymity plays a unique role. On one hand, some instances of crowdsourcing, like academic studies that are vetted by Institutional Review Boards, include an integral requirement to minimize the privacy risks that the human subjects are exposed to, including safeguarding their anonymity. On the other hand, establishing a link between virtual

worker accounts and real people could play a vital role in establishing the reliability of the gathered information and in mitigating worker fraud. Finding the right balance between these conflicting goals is one of the challenges that existing crowdsourcing platforms will need to face as this area keeps evolving. This conflict also presents an opportunity for specialized crowdsourcing services that emphasize one aspect over another, depending on the specific subject area, for example, crowdsourcing services that impose harsh restrictions to guarantee worker anonymity in (highly sensitive) human studies, versus services that forgo worker privacy to provide better matching between requesters and skilled workers.

Designing privacy-preserving crowdsourcing mechanisms. Participation in surveys and disclosure of information in crowdsourcing platforms exposes workers to risks of privacy loss, and these risks increase as workers participate in an increasing number of surveys and give away more information over time. While any particular piece of information may seem insignificant, the aggregated data, linked to the same worker, may be collected over time and may reveal significant amounts of information about the worker. While regulations and legally-binding terms of use may be sufficient to prevent privacy-invading data misuse by honest parties, they may not be effective in preventing privacy loss in the face of data theft or human error. Therefore, proposing and evaluating mechanisms for enhancing privacy in crowdsourcing applications is vital for protecting worker privacy in such platforms, or at least for educating workers and giving them more control over the rate of privacy loss. Such research could draw from existing works on privacy-preserving computations, and adapt them to the distributed nature of crowdsourcing applications.

Worker compensation for privacy loss. Existing crowdsourcing platforms tend to ignore the impact of privacy concerns on worker participation, and set a fixed price per task. This policy may consequently drive away workers who value their privacy above the suggested compensation, and introduce a bias towards workers who place a lower value on their privacy. Compensating workers for their privacy loss may somewhat mitigate this problem, but it introduces many other challenges: the workers' privacy choices may become a source of privacy leak even before participation in the survey; workers may not be truthful about their privacy costs and may provide false answers in surveys, to protect their privacy while maximizing compensation; and participation in multiple surveys over time may call for different mechanisms than those studied so far in single-query settings. While several works have started investigating such problems, many of these questions are still open.

4.7 Conclusion

We provided in this chapter an overview of the state-of-the-art of privacy in crowdsourcing platforms, including existing frameworks that can be leveraged to enhance user privacy in these platforms, and the challenges that are yet to be addressed. The research community has made great strides in recent years

developing new semantic definitions of privacy, given various realistic characterizations of adversarial knowledge and reasoning. However, while research and technology play a critical role in privacy protection for personal data, they do not solve the problem in its entirety. In the future, technological advances must dovetail with public policy, government regulations, and developing social norms. Many challenges still remain, and we believe that this will be an active and important research area for many years to come.

References

1. A face is exposed for AOL searcher No. 4417749. http://www.nytimes.com/2006/08/09/technology/09aol.html
2. Agrawal D, Aggarwal CC (2001) On the design and quantification of privacy preserving data mining algorithms. In: PODS
3. Agrawal R, Srikant R (2000) Privacy-preserving data mining. In: SIGMOD
4. Arrington M (2006) AOL proudly releases massive amounts of private data. In: TechCrunch
5. Bennett J, Lanning S (2007) The Netflix prize. In: KDD Cup and workshop
6. Bogdanov D (2013) Sharemind: programmable secure computations with practical applications. PhD thesis, University of Tartu
7. Chen R, Akkus I, Francis P (2013) SplitX: high-performance private analytics. In: ACM SIGCOMM
8. Chen R, Reznichenko A, Francis P, Gehrke J (2012) Towards statistical queries over distributed private user data. In: NSDI
9. Chen Y, Chong S, Kash IA, Moran T, Vadhan SP (2013) Truthful mechanisms for agents that value privacy. In: EC, pp 215–232
10. Coull SE, Right CV, Monrose F, Collins MP, Reiter MK (2007) Playing Devil's advocate: inferring sensitive information from anonymized network traces. In: NDSS
11. Dandekar P, Fawaz N, Ioannidis S (2012) Privacy auctions for recommender systems. In: WINE
12. Dwork C, Kenthapadi K, Mcsherry F, Mironov I, Naor M (2006) Our data, ourselves: privacy via distributed noise generation. In: *EUROCRYPT*
13. Dwork C, McSherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: TCC, pp 265–284
14. Eagle N (2009) txteagle: Mobile crowdsourcing. In: Internationalization, design and global development. Springer
15. Evfimievski AV, Gehrke J, Srikant R (2003) Limiting privacy breaches in privacy preserving data mining. In: PODS
16. Evfimievski AV, Srikant R, Agarwal R, Gehrke J (2002) Privacy preserving data mining of association rules. Knowl Discov Data Mining
17. Friedman A, Sharfman I, Keren D, Schuster A (2014) Privacy-preserving distributed stream monitoring. In NDSS, San Diego, USA
18. Ghosh A, Roth A (2011) Selling privacy at auction. In: 12th ACM Conference on electronic commerce
19. Goldreich O, Micali S, Wigderson A (1987) How to play any mental game. In: 19th Annual ACM symposium on theory of computing, STOC '87, pp 218–229. ACM
20. Golle P (2006) Revisiting the uniqueness of simple demographics in the US population. In: ACM Workshop on privacy in the electronic society
21. Hafner K (2006) Researchers yearn to use AOL logs, but they hesitate. The New York Times
22. Hill K (2012) How target figured out a teen girl was pregnant before her father did. Forbes

23. Howe J (2008) Crowdsourcing: why the power of the crowd is driving the future of business, 1st edn. Crown Publishing Group, New York
24. TZ Jr. AOL technology chief quits after data release. The New York Times
25. Kandappu T, Sivaraman V, Friedman A, Boreli R (2013) Controlling privacy loss in crowdsourced platforms. Technical Report, NICTA
26. Kandappu T, Sivaraman V, Friedman A, Boreli R (2013) Exposing and mitigating privacy loss in crowdsourced survey platforms. In: ACM CoNEXT student workshop
27. Kandappu T, Sivaraman V, Friedman A, Boreli R (2014) Loki: a privacy-conscious platform for crowdsourced surveys. In: COMSNETS
28. Kargl F, Friedman A, Boreli R (2013) Differential privacy in intelligent transportation systems. In: WiSec, pp 107–112, New York, NY, USA, 2013. ACM
29. Kargupta H, Datta S, Wang Q, Sivakumar K (2003) On the privacy preserving properties of random data perturbation techniques. In: International conference on data mining
30. Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. In: National Academy of Sciences
31. Laudon KC (1993) Markets and privacy. In: ICIS, pp 65–75
32. Lease M, Hullman J, Bigham JP, Bernstein MS, Kim J, Lasecki W, Bakhshi S, Mitra T, Miller RC (2013) Mechanical turk is not anonymous. SSRN
33. Li N, Li T, Venkatasubramanian S (2007) t-closeness: privacy beyond k-anonymity and l-diversity. In: International conference on data engineering
34. Liew CK, Choi UJ, Liew CJ (1985) A data distortion by probability distribution. In: ACM TODS
35. Ligett K, Roth A (2012) Take it or leave it: running a survey when privacy comes at a cost. In: 8th International workshop of internet and network economics
36. Lomas N (2013) Handshake is a personal data marketplace where users get paid to sell their own data. Techcrunch, 2 September 2013. Accessed 27 March 2014
37. Machanavajjhala A, Gehrke J, Kifer D, Venkatasubramaniam M (2006) l-Diversity: privacy beyond k-anonymity. In: International conference on data engineering
38. McSherry F, Mahajan R (2010) Differentially-private network trace analysis. SIGCOMM Comput Commun Rev 40(4):123–134
39. Narayanan M, Shmatikov V (2008) Robust De-anonymization of large sparse datasets. In: IEEE symposium on security and privacy
40. Oremus W (2012) What happens when our cellphones can predict our every move? Slate
41. Rastogi V, Nath S (2010) Differentially private aggregation of distributed time-series with transformation and encryption. In: SIGMOD
42. Reed J, Aviv AJ, Wagner D, Haeberlen A, Pierce BC, Smith JM (2010) Differential privacy for collaborative security. In: EUROSEC, ACM, pp 1–7
43. S. report for chairman Rockefeller. A review of the data broker industry: Collection, use, and sale of consumer data for marketing purposes. Committee on Commerce, Science, and Transportation, United States Senate, 18 December 2013. Accessed 27 March 2014
44. Ribeiro BF, Chen W, Miklau G, Towsley DF (2008) Analyzing privacy in enterprise packet trace anonymization. In: Network and distributed system security symposium
45. Riederer C, Erramilli V, Chaintreau A, Krishnamurthy B, Rodriguez P (2011) For sale: your data, by: you. In: ACM Workshop on Hotnets
46. Samarati P, Sweeney L (1998) Generalizing data to provide anonymity when disclosing information (abstract). In: PODS, p 188
47. Schenk E, Guittard C (2011) Towards a characterization of crowdsourcing practices. J Innov Econ (1):93–107
48. Shi E, Chan THH, Riefel EG, Chow R, Song D (2011) Privacy-preserving aggregation of time series data. In: Network and distributed system security symposium
49. Simonite T (2014) Sell your personal data for 8$ a month. MIT Technol Rev. Accessed 27 March 2014
50. Simonite T (2013) If facebook can profit from your data, why can't you? MIT Technol Rev. Accessed 27 March 2014

51. Sweeney L (2000) Simple demographics often identify people uniquely, laboratory for internation data privacy. Carnegie Mellon University, Pittsburgh
52. Sweeney L (2002) k-Anonynity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl-Based Syst 2002
53. Varshney LR (2012) Privacy and reliability in crowdsourcing service delivery. In: SRII global conference, pp 55–60
54. Wallace N, Whyte S (2013) Supermarket spies: big retail has you in its sights. In: The Sydney Morning Herald
55. Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc
56. Yan T, Marzilli M, Holmes R, Ganesan D, Corner M (2009) mCrowd: a platform for mobile crowdsourcing. In: ACM SenSys