Request Summary

Requesters: Huichuan Xia, Syracuse University

Yang Wang, Yun Huang, Syracuse University

Anuj Shah, , Carnegie Mellon University

Title: Privacy Violations in Crowd Work

Abstract:

Crowd work is a major aspect of the rising gig economy. Crowd work platforms such as Amazon Mechanical Turk (MTurk) and CrowdFlower have millions of ordinary people (i.e., crowd workers) around the world performing tasks (e.g., answering a survey, testing a website) to get paid. These platforms are widely used by companies, academic researchers, and other individuals to provide tasks for the crowd workers. While the literature has raised ethical issues in crowd work, little is known about **people's *actual* experiences of privacy challenges and violations in crowd work**.

Using MTurk, the most popular crowd work platform, as a concrete example, we conducted a survey of crowd workers' privacy experiences with 435 MTurk workers from around the world. Our respondents reported their *actual* experiences with a wide range of **privacy violations, such as sensitive information collection, manipulative data aggregation and profiling, unauthorized secondary use and sharing, as well as deceptive practices such as phishing and scam.** To our knowledge, this is the first empirical study that reports actual privacy violations that people experienced in crowd work. We published these results in this year's *ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*.

Our follow-up work has analyzed actual tasks on MTurk and found that these tasks can **violate crowd workers' privacy but also other people's privacy**. For instance, some tasks asked crowd workers to dig out personal information of other individuals.

***Implications for policy*:** The privacy policies of crowd work platforms such as MTurk are vague from a crowd worker's perspective. We recommend that every crowd work task should be required to have its own privacy policy in which it clearly describes who the task requester is, what a crowd worker needs to do, what data will be collected, shared and used for what purpose, etc. The tasks descriptions on MTurk provide no or vague information about these important points, preventing crowd workers from making informed decisions about whether to perform certain tasks. The crowd work platforms should also enforce measures (e.g., suspend the requester) when the task privacy policy is violated.

***Implications for privacy design*:** First and foremost, there is little or inadequate privacy protection in crowd work platforms. These platforms should do a better job at screening out malicious tasks if they are already doing some screening. These platforms should also warn crowd workers about tasks that might violate their privacy. We believe the platforms can build tools to automatically mark problematic tasks based on user reports/complaints as well as natural language processing and machine learning techniques. Similar tools can be built to inform and remind benign task requesters when they unknowingly or unintentionally design tasks that might violate people's privacy.

Publication:

Xia, H., Wang, Y., Huang, Y., Shah, A. (2017), "Our Privacy Needs to be Protected at All Costs: Crowd Workers' Privacy Experiences on Mechanical Turk," *Proceedings of the ACM: Human-Computer Interaction (PACMHCI): Volume 1: Issue 2: Computer-Supported Cooperative Work and Social Computing (CSCW).*

"Our Privacy Needs to be Protected at All Costs": Crowd Workers' Privacy Experiences on Amazon Mechanical Turk

HUICHUAN XIA, Syracuse University
YANG WANG, Syracuse University
YUN HUANG, Syracuse University
ANUJ SHAH, Carnegie Mellon University

Crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) are widely used by organizations, researchers, and individuals to outsource a broad range of tasks to crowd workers. Prior research has shown that crowdsourcing can pose privacy risks (e.g., de-anonymization) to crowd workers. However, little is known about the specific privacy issues crowd workers have experienced and how they perceive the state of privacy in crowdsourcing. In this paper, we present results from an online survey of 435 MTurk crowd workers from the US, India, and other countries and areas. Our respondents reported different types of privacy concerns (e.g., data aggregation, profiling, scams), experiences of privacy losses (e.g., phishing, malware, stalking, targeted ads), and privacy expectations on MTurk (e.g., screening tasks). Respondents from multiple countries and areas reported experiences with the same privacy issues, suggesting that these problems may be endemic to the whole MTurk platform. We discuss challenges, high-level principles and concrete suggestions in protecting crowd workers' privacy on MTurk and in crowdsourcing more broadly.

CCS Concepts: • **Information systems** → **Crowdsourcing**; • **Security and privacy** → *Human and societal aspects of security and privacy*; • **Human-centered computing** → **Computer supported cooperative work**;

Additional Key Words and Phrases: Crowdsourcing; Privacy; Amazon Mechanical Turk (MTurk)

ACM Reference format:
Huichuan Xia, Yang Wang, Yun Huang, and Anuj Shah. 2017. "Our Privacy Needs to be Protected at All Costs": Crowd Workers' Privacy Experiences on Amazon Mechanical Turk. *Proc. ACM Hum.-Comput. Interact.* 1, 2, Article 113 (November 2017), 22 pages.
https://doi.org/10.1145/3134748

1 INTRODUCTION

Crowdsourcing typically refers to the practice of obtaining inputs or contributions to projects or tasks from an undefined pool of people (a.k.a., crowd workers) [21]. Crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) have enabled an increasing number of organizations, researchers or other individuals to crowdsource a wide variety of tasks, such as tagging photos, answering surveys, transcribing audio/video files, and creating or testing designs [11, 12, 20].

However, crowdsourcing may pose potential privacy risks, such as de-anonymization of crowd workers [27, 38]. For instance, the news media reported that MTurk was leveraged by politicians to mine data from crowd workers and their Facebook friends and to subsequently match their profiles

2573-0142/2017/11-ART113
https://doi.org/10.1145/3134748

to existing voter datasets [19]. Crowdsourcing-related phenomena such as participatory sensing [31], open collaboration [15], and citizen science [5, 6] have also raised privacy issues, such as surveillance and reputational damage, for crowd workers. Additionally, crowd workers may be used for activities that compromise the privacy and security of others. For example, spammers can use crowd workers to decipher CAPTCHAs [18].

Despite the growing body of literature that raises potential privacy risks in crowdsourcing, little is known about how crowd workers themselves experience and perceive privacy. This is an important question to answer because it not only can provide new insights into what kinds of privacy issues are actually occurring but also can inform the design of effective mechanisms to address these issues. To answer this question and help bridge the gap in the crowdsourcing literature, we conducted an online survey of MTurk crowd workers about their privacy concerns, experiences and expectations on MTurk.

As an exploratory study, we chose to focus on MTurk because it is arguably one of the most popular crowdsourcing platforms, having a large number of crowd workers from around the world. To the best of our knowledge, our study is one of the first empirical investigations of crowd workers' own privacy perceptions and experiences on MTurk. In total, we had 435 valid responses from MTurk workers from the US, India, and other countries or areas such as Mexico, France, South Africa, Venezuela, and Hong Kong.

This paper makes two main contributions. First, the study provides empirical evidence that privacy issues are real and affect MTurk crowd workers around the world. Specifically, our findings uncover a number of privacy concerns and experiences that MTurk workers have, many of which have not been reported or are under-reported in the crowdsourcing context, such as targeted ads, phishing, malware, scams, profiling, secondary use of collected data, and stalking. The findings also identify what crowd workers expect of MTurk (the crowdsourcing service provider) and task requesters on the platform to do to protect their privacy. Second, we discuss challenges, high-level privacy principles and concrete suggestions in protecting crowd workers' privacy on MTurk and in crowdsourcing more broadly.

2 BACKGROUND AND RELATED WORK

In this section, we present how MTurk works and review the prior literature on privacy issues in crowdsourcing in general, as well as on MTurk in particular. We then briefly explain why our study fills a significant gap in the literature.

2.1 How MTurk Works

MTurk is a commercial crowdsourcing platform provided by Amazon. MTurk allows requesters to publish tasks (a.k.a., Human Intelligence Tasks or HITs), set the total number of workers that are needed for a task and the amount of compensation for finishing the task, and specify optional criteria for what kinds of crowd workers can take the task (e.g., only workers who registered with MTurk from certain countries are eligible). Once a worker takes a task, he or she can choose to return it (in which case another worker can take it) or choose to finish the task by submitting the required results. Once the worker submits his or her results, the requester can check the quality of the worker's submission and decide whether to approve the submission and pay the worker the predetermined compensation. If the requester rejects the submission, the worker will not be paid. Additionally, the rejection will lower the worker's overall approval rate, which is one of the quality control criterion frequently used by requesters to specify eligible workers who can take their task.

2.2 Privacy in Crowdsourcing-Related Phenomena

Privacy has been recognized as a key challenge in the crowdsourcing literature. Durward et al. propose privacy as a central pillar of the ethical dimensions of crowdsourcing [13]. Halder presents the evolution of crowdsourcing and suggests that its use in public governance can heighten privacy risks and challenges [17] (e.g., using Ushahidi [56], a crowdsourcing platform primarily designed for social activism, during political campaigns or in response to natural disasters). The literature on crowdsourcing with smart phones (a.k.a., mobile crowdsourcing [10]) has also raised privacy issues such as location tracking [56].

Privacy has also started to draw attention in the literature of other crowdsourcing-related phenomena such as open collaboration and citizen science. Forte et al. identify several privacy-related threats in open collaboration platforms such as Wikipedia. These threats include surveillance, loss of employment, harassment, safety threats, and reputational damage [15]. Bowser et al. suggest that citizen science projects (e.g., Zooniverse) where ordinary citizens voluntarily contribute to scientific research (e.g., identifying the species of animals from images) without any payment could pose privacy risks, such as revealing a volunteer's location, compromising her personally identifiable information (PII), and using her data for other purposes [5]. Unlike the above platforms, our platform of interest (MTurk) is a general-purpose crowdsourcing platform where a multitude of workers can perform a broad range of tasks for monetary compensation. Thus, understanding the privacy experiences and concerns of MTurk workers could inform the design of many crowdsourcing systems and their supporting policies.

2.3 Privacy Risks for Crowd Workers

It is typically believed that task requesters and crowd workers are anonymous to each other [27]. However, recent work has demonstrated that a crowd worker can be de-anonymized. For instance, Lease et al. report that MTurk can compromise its crowd workers' privacy through data triangulation: since a worker's MTurk ID is shared with all other Amazon services, the corresponding Amazon shopping profile can be found by searching the MTurk ID [38]. Kandappu et al. show that by using a sequence of surveys on a crowdsourcing platform (CrowdFlower), requesters can deliberately and gradually uncover crowd workers' identities [28].

The desire to provide and receive quality responses on MTurk motivates greater information disclosure, thus creating additional privacy concerns [30]. Social transparency is defined as "the availability of social meta-data surrounding information exchange" [54]. In the context of crowdsourcing, social transparency is relevant to how much information requesters and workers know about each other, and how much workers know about each other when completing collaborative tasks. In a study of social transparency on MTurk, Marlow and Dabbish found social transparency between requesters and workers may affect how much effort workers put into the task and thus the quality of their submissions [40]. For example, the more demographic information requesters reveal to workers, the more effort workers put into the tasks [40]. In another experiment on MTurk, Huang and Fu found that more social transparency among paired crowd workers can lead to better task performance [22]. However, they also note that more social transparency and thus more information sharing raises privacy concerns [22].

2.4 Crowdsourcing for Malicious Purposes

Previous work has also suggested methods by which crowdsourcing can be used for unethical purposes. For instance, Harris warns that unethical requesters could feasibly manipulate the crowd into performing questionable actions, divulging confidential information, and deciphering CAPTCHAs on the requester's behalf for little compensation [18]. In addition to these analyses of

conceptual risks, empirical studies have demonstrated that these risks may be realized. For instance, Lasecki et al. ask MTurk crowd workers to code behavioral videos, highlighting the possibility that crowd workers may learn and reveal other people's sensitive information [35]. Lasecki et al. also show that MTurk workers can be manipulated to perform malicious tasks, such as extracting or identifying someone's credit card number or handwriting characters [36, 37]. While these studies show that crowd workers can inadvertently impinge upon others' privacy via MTurk, our study suggests that MTurk workers themselves suffer from privacy threats and violations.

To summarize, these invaluable studies demonstrate that 1) today's crowdsourcing systems are susceptible to privacy violations against crowd workers, and 2) task requesters can use crowd workers for activities that violate others' privacy and security. However, little is known about how crowd workers themselves perceive and experience privacy in crowdsourcing and on MTurk in particular. Our study aims to bridge this gap by investigating crowd workers' privacy perceptions and experiences on MTurk.

3 METHODOLOGY

We conducted an online survey of MTurk crowd workers about their privacy concerns, experiences, and expectations on MTurk. The study was approved by the Syracuse University IRB office. Our survey had a diverse sample (N=435) of MTurk crowd workers from around the world.

3.1 Survey Flow and Questions

The survey was written in English and hosted on Qualtrics, an online survey platform. We first showed respondents the consent form for this study. Inspired by a prior study [47], before our respondents entered the survey we also emphasized that detailed responses would support our analysis and improve MTurk workers' future experience. While we did not specify how much detail we required, we stated that respondents who provide detailed experiences and insights could receive an extra 50-cent bonus in addition to the one-dollar compensation for finishing the survey.

The survey included a set of open-ended questions for our respondents to describe their experiences with MTurk. For instance, we asked them what they like about this crowdsourcing platform and what they think should be improved. We then asked them whether they have any privacy concerns about MTurk, and if so, what those concerns are. We asked similar questions about their privacy expectations on MTurk. We then asked our respondents if their privacy had ever been compromised on MTurk and to provide concrete examples. If they had no such experiences, we asked them to provide any experience of this kind that they had heard about. Finally, we asked about our respondents' demographics, such as gender and age, as well as their history of performing MTurk tasks.

3.2 Respondent Recruitment

We recruited respondents from MTurk that had a 95% or higher approval rate. Because the majority of MTurk workers are from the US and India [12, 48], we used a stratified sampling strategy to obtain a diverse sample. Specifically, we presented our survey as three tasks with the same questions. For the first survey task, we constructed the task so that only crowd workers from the US could answer it. Similarly, for the second survey task, we only allowed respondents from India to answer it. For the third survey, we only allowed respondents from a country/area other than the US or India to answer it. This stratified sample allowed us to explore whether certain privacy concerns, experiences or expectations are unique to crowd workers in a specific geographic area or applicable across different areas. The three survey tasks were conducted in parallel on MTurk from June to August in 2016.

To ensure the quality of responses included in our analysis, we adopted best practices [46] to screen out duplicate respondents. We also checked each response to make sure that it was meaningful. For example, we filtered out responses with consistent answers of "n/a" or "no." Through this quality check, we removed ten responses from the US, 19 from India, and two from the other countries and areas. In the end, we had 435 eligible respondents. The average completion time of our survey was 25 minutes. We gave the 50-cent bonus to 18 respondents for their very detailed responses.

3.3 Respondent Demographics

Out of the 435 total eligible respondents, there were 194 from the US, 181 from India, and 60 from other countries/areas such as Australia, Brazil, Hong Kong, and Venezuela. With regard to gender, 41% were female, and 59% were male. The average age of our sample was 34 years (Median=32, SD=9.7). The three subsamples within our larger participant pool had similar age distributions (median age 31 or 32), but the US sample had a larger percentage of female respondents (US: 53% female) than the other two samples (India: 29% female, and other countries: 35% female).

We also asked our respondents about their experiences with performing MTurk tasks. The majority of our respondents had substantial experience with MTurk both in terms of the history and the frequency of performing MTurk tasks. Regarding how long they have worked on MTurk, only 3.4% of respondents said less than one month, 33.6% said one month to one year, 31.5% said one year to three years, and 31.5% said more than three years. More specifically, 58.2% and 81.8% of respondents said more than one year from the US and India, respectively; 71.7% of the respondents in other areas had between one month and one year of experience with MTurk. In terms of how frequently they performed MTurk tasks, 54.3% of respondents completed one task per day, and 34.9% completed multiple tasks per day. The three specific subsamples exhibited similar frequencies of self-reported task completion.

3.4 Data Analysis

We conducted a thematic analysis of the answers to our open-ended questions. Thematic analysis is "a method for identifying, analyzing, and reporting patterns (themes) within data" and is commonly used to analyze qualitative data [7].

First, we immersed ourselves in the data by reviewing it multiple times with an eye toward the kinds of privacy issues being expressed. We then wrote simple descriptive notes on these privacy issues and engaged in several rounds of discussion.

Next, we open coded the data in an inductive fashion. Codes emerged and were selected through an iterative process and discussion between the coders. Some codes were dropped because they appeared to occur once, such as the code of "privacy concern about being targeted by terrorism." Some codes were collapsed because they appeared to represent two highly similar concepts, such as concerns about "spamming ads" and "targeted ads."

We then explored the connections between codes and identified about 20 recurring themes, such as de-anonymization, profiling, spam, scams, platform security, malware, secondary use of data, requester-worker confrontation, unlinkability, and task screening. These themes were not necessarily mutually exclusive. For example, some respondents' discussion of requesters collecting personal email addresses were related to their concerns about sensitive information inquiries, de-anonymization, and spamming. Finally, we reviewed the candidate themes by reading corresponding responses to check whether they coherently represented the underlying themes.

Once we converged upon a final set of themes encompassing the privacy issues present in our dataset, two researchers coded distinct subsets of the participants' responses according to these themes. The researchers also coded each response as expressing concern or experience (or both).

While our survey explicitly asked for concerns and experiences in separate questions, thematic analysis dictates that questions do not shape the interpretation of responses [7, 8]. A third researcher then gathered a random sample of these responses and coded them independently to ensure quality and agreement in the coding process. Upon comparing the former researchers' coding to that of the third independent researcher, we achieved a Krippendorff's alpha value of about 0.83, suggesting very good inter-rater reliability [33].

After identifying the set of privacy issues, we chose to organize them primarily based on a well-known taxonomy of privacy proposed by Daniel Solove [52]. The taxonomy was originally developed to assist courts and policymakers in identifying and clarifying privacy harms, as a concrete understanding is necessary to defend privacy rights against other interests (such as security) [52]. It has subsequently proven useful to structure privacy engineering guidelines. For example, Spiekermann and Cranor derive support from Solove's taxonomy in defining engineers' responsibility to build access control and risk management into privacy-enhancing systems, as well as explaining end users' conception of privacy breaches [53]. Marsh et al. suggest using the taxonomy to inform risk-benefit analyses that organizations may conduct when adopting new technologies. Specifically, they argue that its focus on privacy harms ensures privacy impact assessments account for how new technology affects the organization, individual, and our society [41]. With its emphasis on privacy violations, this framework allows us to understand the similarities and differences among privacy issues, to focus on the underlying activities (e.g., collection vs. processing), and to determine future improvements that target these activities.

The taxonomy includes four categories of "socially recognized privacy violations": information collection, information processing, information dissemination, and invasions [52]. Accordingly, we used the same four categories, plus a fifth category, "deceptive practices," to classify the privacy issues reported by our participants. This fifth category was added to accommodate issues such as phishing, malware, and scams that did not neatly fit within the original taxonomy yet related to each other in their deceptive nature. To ensure rigor in our classification of these issues within the taxonomy, two independent researchers coded each issue using all five categories. We achieved a Krippendorff's alpha value of 0.77, suggesting good inter-rater reliability [33].

4 FINDINGS

In this section, we present the main findings of our study, focusing on the various privacy issues that our respondents reported including their actual experiences and/or conceptual concerns with these issues. In addition, we present respondents' privacy expectations on MTurk as well as a comparison between respondents from the US and India in terms of their concerns, experiences, and expectations.

4.1 Privacy Issues Related to Information Collection

Information collection in our study context refers to the collection of information about MTurk workers via MTurk tasks. Our respondents highlighted three specific privacy issues related to information collection: inquiry of sensitive information, providing information for free, and targeted ads.

Inquiry of sensitive information. This privacy issue involves tasks that request sensitive information about MTurk workers, such as their identities, religious views, and financial information. This was a common privacy issue our respondents experienced. For example, one respondent recalled a task that requested a photo of his health insurance card:

> *"To participate in the task, I had to upload a picture of my health insurance card. The purpose of the study was to find out what people hated about having or getting a health insurance. But I feel like I had to give up very sensitive information."* (US, R16)

This respondent considered that his health insurance card contains very sensitive information about him. However, in the end, he confessed that he did provide such information because the task paid $25. In a way, the respondent decided to trade his sensitive information for money. Scholars have suggested that people often disclose their information because of immediate gratifications (e.g., monetary gain) even if they say they value their privacy [1]. Prior research has also shown that many crowd workers do work on MTurk to make ends meet [25]. Therefore, these workers might be more influenced by the immediate monetary gain to share their sensitive information. This example also raises another important question - is a picture of the worker's insurance card really necessary for the purpose of the study? If yes, the task should clearly explain why. If no, the task should not request it.

Some respondents were asked about their caste, which is considered a piece of highly sensitive information in India:

> *"A few tasks do ask about caste of the individual. Some of them are really done with cruel intention to hurt the individual psychologically."* (India, R88)

This respondent felt uncomfortable when being asked about her caste and suspected a malicious intention. In her view, asking about caste, such a piece of sensitive information, could even cause psychological damage to individuals.

Respondents from other countries also experienced sensitive information inquiries. For instance, a respondent talked about a task that requested his browsing history:

> *"There was this one HIT that required me to upload my browser history. I deleted most of the history and proceeded to upload it, only to find that the HIT couldn't even [sic] submitted."* (Venezuela, R39)

This respondent felt his browsing history was confidential information, motivating him to sanitize the history before submitting it. This case demonstrates that if workers have privacy concerns about a task, they may apply strategies to protect their privacy, such as manipulating the data. However, from the requester's perspective, such data might be less detailed or useful. More broadly, this implies that crowdsourcing systems need to protect workers' privacy if they want to maintain the quality of crowd contributions.

Providing information for free. Our respondents also disliked providing their information without getting paid. For instance, one respondent heard about other workers encountering this issue, but he avoided it himself:

> *"There's a particular requester that pops up from time to time that always promises an amount of money...I've never fallen for it but I've read other accounts by people who have. They have pages and pages asking for personal info and then they reject the HIT anyway and never even pay."* (US, R28)

This is a telling example that a requester could game the MTurk system to exploit workers and collect their information for free. It also highlights the power imbalance on MTurk where requesters arguably have the upper hand. Whether a MTurk worker or his or her submission is qualified is judged solely by the requester. Therefore, a worker may provide a large amount of personal information via a task but could still be deemed unqualified for the task.

Targeted ads. Targeted advertising refers to the practice of tracking people's online activities to build profiles of individual users and to provide ads based on these user profiles. Tracking people's online activities can be considered a form of (extensive) information collection. Some respondents reported encountering targeted ads after conducting certain MTurk tasks. For instance, one respondent said:

> *"I have done tasks before where you had to click an outside website. Afterwards I noticed some of the same ads I had previously seen on that site when browsing on my own. Seemed to me I had been targeted for ads specifically from doing that task."* (US, R164)

These tasks were likely designed to make the respondent form some kind of profile based on what the tasks asked them to do (e.g., visiting a particular site). Another respondent reported a similar experience but also felt his reputation was endangered:

> *"It was an advertisement work. We need to view the ads with 18+ content in a site given. The purpose was to increase the viewers of the ads so that they can earn money. However, after that work done whenever I browse the Internet, the ads comes up. It was actually very irritating because my brother used to use my laptop."* (India, R137)

This task made the respondent click and view adult ads, increasing the click-through rate of the ads and in turn providing advertisers with more revenue. By performing this task, the respondent also appeared to be interested in adult content, which led to similar ads following him across the Internet. This may make others think he actually has such an interest. In other words, the task tricked the respondent into forging a false image of himself, which may benefit the advertisers but could hurt his Internet experience and even his reputation.

4.2 Privacy Issues with Information Processing

Information processing refers to the practices of using, storing or manipulating data that has already been collected [52]. In our study context, it refers to how MTurk or requesters handle the data collected from workers. Our respondents shared their concerns about and/or experiences with de-anonymization, data aggregation, profiling, unauthorized secondary use, and insecurity in data processing.

De-anonymization. Workers' MTurk IDs can be linked to their Amazon accounts, which can then potentially de-anonymize their real identities, as one respondent noted:

> *"The anonymity of the user isn't quite so anonymous because Amazon uses the same ID for shopping accounts so it is quite easy to pull up name, location and other info that can pinpoint an individual. It's unnerving."* (US, R48)

Another respondent was aggravated that his real email address was revealed to the requester by Amazon:

> *"I had my personal identity revealed to a requester because I used the contact form through AMT. I figured Amazon would use their system to send the e-mail with a generated e-mail from the Amazon domain, but they showed my real e-mail directly to the requester! I still can't comprehend how Amazon allows this and they're the ones directly responsible for giving the information!"* (US, R46)

Amazon should consider changing its policy and corresponding technical implementations to make MTurk IDs unlinkable to other Amazon accounts and hide workers' personal information (e.g., email addresses) from requesters.

Data aggregation. Data aggregation means combining different information that a MTurk worker has provided. For example, one respondent spoke about the risk of combining different information about an individual together:

> *"People ask question here and there...like sexual preferences, racial views, gender views, income, all [sic] many other private things. One at a time is not much, however add them up and someone knows you better than your wife."* (US, R53)

As this respondent pointed out, each individual piece of his information might not seem significant, but the aggregate of these pieces of information can reveal sensitive details about him. In

addition, these sensitive details can then be used to build profiles of individual workers for various purposes.

Profiling. Profiling refers to building a dossier of an individual for certain purposes (e.g., providing targeted ads). Many respondents were concerned that gradually revealing their personal information across multiple tasks over time could allow for the combination of these pieces of information with their MTurk IDs, in turn enabling requesters to profile individual workers. For instance, a respondent explained:

> *"It concerns me that it's very possible that the aggregate of all the surveys I've completed, if linked together through my MTurk ID, could be used to build a profile on me, and used for whatever purposes."* (Venezuela, R39)

While this respondent did not explicitly specify who might have the capability to combine all the information a worker has provided to different tasks with his or her MTurk ID, at least in theory both requesters (who had multiple tasks that a worker had completed) and Amazon could do this.

Unauthorized secondary use. Unauthorized secondary use is the practice of using information collected about an individual for a purpose other than its original purpose without the individual's authorization [51]. In our study context, it means requesters use workers' data collected via MTurk tasks for a different purpose. The secondary use is done by the original requesters rather than other people. Several respondents shared their experiences with unauthorized secondary use. For instance, one respondent talked about her experience with a task that surprised her afterwards:

> *"The requester asked you to take a series of photos of yourself doing different tasks, like smoking a cigarette, sweeping etc. I did not think too much about it and completed the task. Six months later, I heard that the requester was using these photos in some type of art exhibit. I felt violated and lied to. I hated having my picture out there."* (US, R154)

This respondent did not know her photos would become public or be used for an art exhibit until six months after finishing the task. This is clearly an unauthorized secondary use of her data and violated her privacy expectations. While MTurk's policy prohibits fraudulent tasks, it does not require tasks to specify how the collected data will be used. This lack of requirement creates a hotbed for potential privacy violations.

Insecurity. Insecurity means the lack of security in information processing - for example, insecure storage of collected information [52]. Some respondents felt anxious because they had provided too much personal information to various tasks on MTurk; hence, if the platform gets hacked, their sensitive information may be compromised. For instance, one respondent explained his concern about the security of MTurk:

> *"I worry that if someone hacks some research surveys on MTurk, personal information may be compromised. There are some surveys which are regarding personal habits and about things like adult work, if this get leaked, it would definitely affect my reputation in the society."* (India, R87)

He was not only concerned about the leak of his sensitive information but also how such leakage could place him in a bad light.

4.3 Privacy Issues with Information Dissemination

Information dissemination refers to propagating or sharing information that has been collected [52]. Our data set primarily reveals concerns about unauthorized sharing.

Unauthorized information sharing. Unauthorized information sharing, in our study context, means that workers' information is shared by requesters to other parties without the workers' awareness and consent. For example, one respondent suspected that requesters may share or sell workers' email addresses collected from MTurk tasks:

> *"It's very usual they ask you too for your e-mail. Maybe they will never use it, but I don't know...they are a company, and we know company can earn money sharing private information, as the emails are."* (Spain, R23)

Targeted ads (discussed above) can also be considered a type of unauthorized information sharing when workers do not know that their actions during a task (e.g., visiting a site) may be shared with third parties interested in providing targeted ads. This issue of unauthorized information sharing speaks to the information asymmetry on MTurk [14] where workers may not be aware of requesters' intent or practices.

4.4 Privacy Issues with Invasion

Invasion means intrusion into people's private lives, "disturbing their tranquility or solitude" [52]. It can be informational (e.g., spamming) or physical (e.g., stalking). Our respondents shared their experiences with spamming, stalking, and requester-worker confrontations.

Spamming. Many respondents reported instances in which tasks asked for their email addresses and they were subsequently bombarded with spam messages. This represents a violation of the right to be let alone, a prominent conception of privacy [59]. For instance, one respondent shared his experience with spamming as a result of doing a MTurk task:

> *"Also one time I did a HIT for an app. The app was actually interesting but it needed my email to get the app. I put a burner email in there. It won't stop sending emails to that burner email and it's really annoying."* (US, R126)

While some other respondents suspected that doing certain tasks triggered spam emails, this respondent had clear evidence that the task led to spam emails.

Stalking. Stalking is another type of invasion, and it involves "repeated, persistent, unsolicited communications or physical approaches to the victim" [43]. Many respondents suspected that they might be stalked because they did some MTurk tasks. For instance, one respondent explained:

> *"I regretted when I started to realize it was probably some creep who wanted pictures of women. Not knowing who got my picture and my address is very scary. They could be stalking me or trying to find me."* (US, R5)

According to her recollection, she shared both her picture and her address, which could be used to identify and even stalk her. This kind of story suggests that MTurk tasks that are online may backfire in workers' offline lives.

Requester-worker confrontation. Confrontations between requesters and workers could occur online and/or offline. While no respondent reported their own experience of this issue, many shared such stories that they heard. For instance, one respondent told us about a dramatic fight between a requester and a worker that he had heard:

> *"It was a Turker who got rejected and he contacted the requester and I suppose things got a little heated. The requester then used his email address to find his Facebook account and leave harassing comments there, as well as on his Twitter account."* (US, R86)

This kind of confrontation is often triggered by a requester rejecting a worker's task submission. On MTurk, requesters often specify a minimum approval rate of workers who can do the task. Rejection of a worker's submission would not only disable the worker from receiving the payment for the current task but also worsen the worker's overall approval rate, deteriorating the worker's prospect of performing other MTurk tasks. This story also highlights how what happens on MTurk can spill over into other realms of a worker's online experiences (e.g., Facebook and Twitter).

4.5 Deceptive Practices: Phishing, Malware and Scams

Deceptive practices refer to tasks designed by malicious requesters that could harm workers, for instance, stealing valuable information from their computer. These practices include phishing, malware, and scams. Unlike the issue of insecurity which is about the security of a crowdsourcing platform, these deceptive practices focus on the malicious behavior of the requesters.

Phishing. Phishing is a type of fraudulent practice that tricks people into visiting a fake site that looks similar to the genuine site so that the users will divulge their sensitive information (e.g., passwords) to the malicious actors. Phishing is often done by crafting emails that seem to come from reputable or reliable sources. Several respondents described their experiences with phishing on MTurk. For example, one respondent explained tasks that lead to fake log-in screens:

> *"Requesters will put up jobs that are actually phishing. So when workers click, they get to a fake log-in screen and end up having their account and linked payment account compromised."* (US, R86)

Similar to common phishing attacks, both log-in credentials and sensitive financial information can be compromised, and victims may suffer financial losses.

Malware. Malware is software designed to harm computers. Our respondents shared their experiences of doing MTurk tasks that asked them to download or install software that turned out to be malware. For instance, one respondent talked about falling for such a trap on MTurk:

> *"I downloaded a .ASP file when I first started MTurk. For a measly $.50 I had to spend 3 days trying to get rid of a virus. In the end I wiped out my hardrive and bought a new copy of Windows and started from scratch because I couldn't get rid of it. That's $100 for a hard drive $125 for software and 3 days without a computer so I could make $.50!!!!!!"* (US, R104)

While this respondent later said that he reported the incident to MTurk, which had then solved the problem, it is still concerning that this kind of malicious tasks occurred on MTurk in the first place. These malicious tasks are clear violations of the MTurk policy, which says, "HITs that require Workers to download software that contains any malware, spyware, viruses, or other harmful code - is an example that violates Amazon Mechanical Turk policies" [3]. However, the policy does not seem to forbid tasks from requiring workers to download (legitimate) software. Therefore, the policy puts the burden on workers to identify problematic software, which is challenging to do. For instance, another respondent shared an experience that highlights this challenge of judging the legitimacy of software prior to downloading/using it:

> *"I encountered tasks that provide a link to another page and the tasks require to use an experimental app to install on the cellphone not knowing if it is a bad program that is going to steal personal info."* (Mexico, R2)

These real incidents suggest that more proactive measures by MTurk to filter out this kind of malicious task are imperative.

Scams. Scams are instances of fraud where requesters conceal their real purposes or break their promises. For example, one respondent explained how his PayPal account got hacked as a result of doing a MTurk task masqueraded as a legitimate task of network testing:

> *"A requester declared his purpose: 'Visit a research website and leave it open [No interaction required] - a network survey.' The task requested that we do a HIT about network testing and to leave out the screen...After the HIT, I started losing my money on PayPal...I think our privacy need[s] to be protected at all costs."* (South Africa, R18)

This example highlights another aspect of the power imbalance of MTurk where it is challenging for workers to know the real intent of a task (or its requester) and can fall for various scams masqueraded as legitimate tasks.

	Concern (%)	Experience (%)	Both (%)	Concern | Experience (%)
Information Collection	23.7	28.5	6.7	23
Sensitive info inquiry	21.6	25.5	6.7	26
Providing info for free	0.5	0.2	0.0	0
Targeted ads	2.5	5.7	1.1	20
Information Processing	30.1	8.0	4.1	51
De-anonymization	12.9	4.4	2.3	53
Data aggregation	4.1	1.6	1.1	71
Profiling	1.8	0.0	0.0	N/A
Unauthorized secondary use	14.5	2.3	0.5	33
Insecurity	12.2	2.1	0.2	11
Information Dissemination	14.5	2.3	0.5	20
Unauthorized sharing	14.5	2.3	0.5	20
Invasion	8.3	11.0	0.9	8
Stalking	4.8	3.2	0.5	14
Confrontation	0.2	0.0	0.0	N/A
Spamming	4.1	8.5	0.9	11
Deceptive Practices	9.7	6.9	1.6	23
Phishing	4.1	4.1	1.4	33
Malware	3.9	3.9	0.7	18
Scams	4.1	2.1	0.5	22

Table 1. The frequencies of different privacy issues. The first column lists the specific privacy issues grouped into five categories: information collection, processing, and dissemination as well as invasion and deceptive practices. The second column *concern* shows the percentage of respondents who expressed their concern about a particular issue. The third column *experience* shows the percentage of respondents who reported having experienced a particular issue. The fourth column *both* shows the percentage of respondents who reported having a concern about and having a personal experience with a particular issue. The fifth column *concern* | *experience* shows for those respondents who reported having experienced a particular issue, what is the percentage of them who also reported having a concern about that issue.

4.6 Frequencies of Different Privacy Issues

To understand how common these particular privacy issues were among our respondents, we calculated the frequency of each issue. More specifically, Table 1 shows the percentages of respondents who reported having a concern about an issue, having a personal experience with an issue or reporting both. The table also includes a conditional percentage: the percentage of respondents reporting having experienced a specific issue, who also reported concerns about that issue. We also counted the frequencies of the five categories: information collection, processing, and dissemination, as well as invasion and deceptive practices. For instance, if one respondent expressed her concern about at least one of the constituent issues of information collection, we counted her as having the concern about information collection.

As we can see from Table 1, the two categories encompassing the most frequently reported concerns were information processing (30.1%) and collection (23.7%), whereas information collection

(28.5%) was the most frequently experienced category. At the level of specific issues, the issues for which the greatest number of respondents reported concerns were sensitive information inquiry (21.6%), unauthorized secondary use (14.5%), unauthorized sharing (14.5%), de-anonymization (12.9%), and insecurity (12.2%); the most frequently experienced issues were sensitive information inquiry (25.5%) and spamming (8.5%). These issues should be MTurk's priorities to address.

We also explored the association between having a concern about a privacy issue and having a personal experience with the issue (i.e., by calculating *concern* | *experience*). A higher value of *concern* | *experience* means that having experienced an issue is associated with a higher likelihood of having a concern about the issue. For example, 71% of the respondents who reported having a personal experience with data aggregation also raised privacy concerns about this issue. In contrast, a lower value of this percentage (e.g., insecurity) implies that the issue may not have registered as a concern in workers' mind even if they have experienced it. Given the various privacy issues our respondents reported, how did they think their privacy should be protected on MTurk?

4.7 Crowd Workers' Privacy Expectations on MTurk

When asked whether they have any privacy expectations on MTurk, 82% of respondents expressed some expectation(s).

Privacy expectations for MTurk. Our respondents mentioned many expectations for MTurk, including not sharing or misusing their information (25.5%), maintaining a secure platform (17%), helping them maintain their anonymity/pseudonymity (12.9%), ensuring their MTurk activities are unlinkable (4.4%), and screening for requesters and tasks (2.8%). For instance, a respondent said that he had provided very sensitive information to MTurk, and he therefore expects Amazon to keep the data private and safe:

> *"I have an expectation to privacy with MTurk because we gave MTurk our social security numbers and other super sensitive material such as this. Honestly? I really expect Amazon not to flash it around like mad."* (US, R29)

Privacy expectations for requesters. Our respondents also had expectations for the requesters, including not sharing workers' data to third parties (14.3%), not collecting sensitive information from workers (9.2%), and not using workers' data for a different purpose (6%).

For instance, one respondent felt requesters should not ask about his birth date, home address, phone numbers and email:

> *"I expect the requesters not asking for date of birth or the place we reside. These are very sensitive information as far as me concerned. Seeking mobile numbers and email ID also should be stopped."* (India, R21)

Consent and control for crowd workers. Some respondents felt that they need better empowerment in protecting their privacy. 2% of all respondents explicitly expressed their expectation that they need to provide consent before their data can be used or sold. For instance, one respondent explained:

> *"I expect for my information to not be sold or used by third parties that I have not given consent to."* (US, R152)

Similarly, 3% of respondents explicitly stated their expectation to have control over their data. For example, one respondent expected to control when requesters can use his data:

> *"I expect that my private information remains such, except in situations where I willingly give it to requesters for use."* (US, R13)

The key for consent and control is that workers themselves are the ones who make data sharing and usage decisions.

No privacy expectation. While most of our respondents had some privacy expectations on MTurk, 78 (18%) respondents had no such expectation, and their explanations were either because they trust MTurk/Amazon's brand and security (15.5%), or because they distrust MTurk or any MTurk requesters or anybody to protect their privacy (2.5%).

For the respondents who trust MTurk/Amazon, they saw no privacy issues or had no privacy expectations because they believed that MTurk/Amazon would do the right thing to protect them. For instance, one respondent explained:

> *"Amazon is a name of trust. They are going well in a right way. I feel so secured on using MTurk. So I didn't have any privacy expectation on MTurk."* (India, R163)

However, some respondents either did not trust Amazon or they did not trust any parties will protect their privacy. For instance, one respondent was vocal about his lack of trust:

> *"None. I don't trust anybody with my privacy. I simply don't reveal what I don't want to."* (Italy, R66)

4.8 Crowd Workers in US vs. India

Since most MTurk workers reside in the US and India [12, 24], we explored the differences between the responses of our respondents in these two countries. This US-India comparison was in part inspired by the prior literature that suggests that privacy perceptions and behaviors may vary across different countries [26, 29, 39]. However, we chose not to include our respondents from other countries in this country-based comparison because we had very few respondents from each country/area other than US and India.

Overall, more than two-thirds of US respondents shared some privacy concerns (67%), experiences of privacy losses (67%), and privacy expectations (81%) on MTurk, as compared to lower percentages of Indian respondents reporting privacy concerns (56%), experiences (51%), and expectations (60%). The two samples also shared the same list of most frequently mentioned privacy concerns (e.g., sensitive information inquiry and insecurity) and experiences (e.g., sensitive information inquiry and spamming). While the US sample almost always had a higher percentage of respondents reporting concerns or experiences about an issue, the differences between the two samples were not statistically significant according to Chi-Square tests.

With regard to privacy expectations, "no unauthorized sharing" was the most cited privacy expectation for both US (46.9%) and Indian (39.5%) respondents. However, 24.9% of Indian respondents said they did not have any privacy expectations on MTurk because they trust MTurk or Amazon, as compared to a significantly lower percentage for their US counterparts (6.2%) according to a Chi-Square test (p-value $< .0005$).

5 DISCUSSION

Our study uncovers real incidents in which crowd workers' privacy was violated on one of the largest crowdsourcing platforms, MTurk. These incidents point to various privacy issues. De-anonymization on MTurk [38] and worker-requester confrontations on Wikipedia [15] have been shown in the prior literature. However, other issues we identified, such as profiling, phishing, malware, and stalking, have seldom or never been discussed in the crowdsourcing context.

Our findings also show that these privacy issues are not just conceptual concerns that people may have or potential threats that could materialize, but rather something that our respondents from around the world had already experienced themselves on MTurk. These privacy issues are real and have resulted in significant consequences for workers on MTurk such as losses of sensitive/personal information, reputation, and financial assets. Some of these issues such as targeted ads may only affect the workers who were profiled and targeted by ads, while other issues such as malware may

affect people beyond the workers who were originally tricked (e.g., the compromised computers might infect other computers). In this case, workers are a means or "commodity" [2] to achieve a vicious goal. Furthermore, these issues seem to be endemic to MTurk/workers in general rather than to workers only in certain countries or areas. While our study focuses on MTurk, these issues may present in other crowdsourcing systems.

The top two privacy issues that our respondents experienced were sensitive information inquiry and spamming. Clearly, MTurk should prioritize these issues. The most frequent concerns were sensitive information inquiry, unauthorized secondary use, unauthorized sharing, de-anonymization, and insecurity. Our respondents' privacy expectations also mostly centered on these issues. Prior literature on information disclosure has shown that people's privacy concerns can affect their intention to disclose information (e.g., in online shopping sites) [50]. As we saw in our study, if a worker has privacy concerns about a task, he or she may make a more privacy-preserving but less detailed and potentially less useful submission (e.g., a worker cleaned most of his browser history before submitting the history). Quality of crowd work is crucial in crowdsourcing. This result suggests that addressing workers' privacy concerns is also important because that may encourage workers to provide more detailed submissions, potentially enhancing the quality of their work.

Prior literature has shown that cultural differences could affect people's intention and behavior of information disclosure [26, 39]. However, we did not observe a significant difference between the responses of our respondents in the US and those in India. The only exception was that more than 20% of our respondents in India did not report any privacy expectations on MTurk whereas only about 6% of the respondents in the US held the same view. We suspect that it may be related to the fact that more Indian respondents expressed their trust in and gratitude to MTurk than their US counterparts. We also notice that this finding is in line with a prior study which found that US MTurk workers are more concerned about online privacy in general than their Indian counterparts [29]. Next, we discuss the challenges for privacy protection in crowdsourcing.

5.1 Challenges for Privacy Protection in Crowdsourcing

Many challenges stem from the fact that it is difficult to monitor and filter out problematic requesters/tasks. This is due to the difficulty of automatically detecting malicious or privacy-invasive tasks, as well as a large number of requesters and tasks in a crowdsourcing platform such as MTurk. What is worse, the responsibility or burden of detecting problematic tasks seems to rest mainly on the shoulders of crowd workers. It can be challenging and costly for workers to uncover the real intent of a task. Part of the problem is due to the information asymmetry between requesters and workers, because the former may not (truthfully) reveal the purpose of a task. While MTurk's policies prohibit tasks with malicious intent and allow workers to report problematic tasks, crowd workers often need to actually perform the task in order to determine its real intent. Sometimes, the consequences may not occur or become known immediately after the task - for instance, using workers' uploaded pictures for a different purpose. Thus, crowd workers almost have to accept these tasks before realizing they are problematic. The reactive nature of current policies is obviously not ideal.

It is also concerning and even ironic that crowd workers are the ones who are supposed to detect and report the problematic tasks because they are arguably in a weaker position on MTurk or in a crowdsourcing ecosystem more generally. Crowd workers are often treated as an invisible API call on a crowd work platform such as MTurk, and they have to face the risks of working without payment, short task expiration time, arbitrary rejection, and unresponsiveness [49]. Crowd workers are also seen as a computational service, being bound by duty and payment at the will of the requesters who are prioritized by a crowd work platform such as MTurk [25]. Furthermore,

they are commoditized in a sense that no social protection or any employment benefit would cover them [4].

Crowd workers and requesters are often in an imbalanced power relationship in which the latter have the arbitrary dominion to qualify or reject the former's work [25, 49]. In addition, there is an information asymmetry between crowd workers and requesters where workers can have limited information about requesters and their tasks, whereas requesters are empowered to collect and know more information about workers [14]. Scholars have suggested that balancing the power dynamics and information asymmetry in crowd work can benefit both crowd workers and requesters [45]. For example, Turkopticon has become a popular tool for MTurk workers to rate requesters and counter the power imbalance [25, 49]. Online forums such as TurkNation have become a popular site for MTurk workers to communicate and unionize [42, 60]. Crowdsourcing platforms should consider incorporating these features into the platforms.

A sizable portion (about 40%) of our respondents did not express any privacy concerns on MTurk. While we did not have data to directly explain this result, we suspect two reasons which could be investigated in future research. First, some MTurk workers might be unaware of the potential privacy risks or issues that the majority of our respondents experienced. This lack of awareness can also make them vulnerable to problematic tasks and requesters. The challenge then becomes how to inform or educate them to raise their privacy awareness when privacy protection is not their immediate task or goal. Second, this reported lack of concern might also reflect crowd workers' different valuations of privacy and working strategies on MTurk. Given that MTurk is a competitive labor market where crowd workers try to get work, some crowd workers might adopt a strategy that seeks to earn more work at the expense of knowingly exposing themselves to privacy risks.

While people are often uncomfortable with the monetization of their personal information [9], Acquisti suggests that people may still disclose their information for the immediate gratification of monetary gain [1]. In a pay-to-work crowdsourcing platform such as MTurk, crowd workers might overemphasize the immediate gratification of monetary payment by doing a task while underestimating the potential future consequences. Given these challenges, how can MTurk better protect crowd workers' privacy?

5.2 Protecting Worker Privacy in Crowdsourcing

The privacy issues we identified along the lines of information collection, processing, and dissemination as well as invasion and deceptive practices are arguably not unique to MTurk or crowdsourcing but could occur in computer systems more generally. How can designers of crowdsourcing systems anticipate and address these privacy issues? High-level privacy principles such as *Fair Information Practices* (e.g., notice and consent) [16] have been used to guide privacy protection in computer systems in general and specific application domains such as ubiquitous computing [34]. We believe that these principles can benefit MTurk and crowdsourcing in general. Below we discuss both high-level principles and concrete ideas to mitigate the specific privacy issues we found on MTurk. It is worth noting that the principles are not platform dependent and the ideas can potentially be implemented by other crowdsourcing platforms.

Issues related to information collection. To address issues such as inquiry of sensitive information, providing information for free, and targeted ads, crowdsourcing systems should apply privacy principles such as collection limitation, proportionality, notice and purpose specification. *Collection limitation* means that "There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means" [44]. Following this principle, crowdsourcing platforms and requesters should consider what kinds of worker data are really necessary to collect and how the data can be collected properly. Similarly, *proportionality* states that "any application, system, tool or process should balance its utility with the rights to privacy (personal, informational,

etc.) of the involved individuals" [23]. To achieve proportionality, crowdsourcing requesters should be mindful about what worker data to collect in order to balance the utility of the collected data with the workers' privacy. A concrete idea of implementing these principles can be to provide requesters with an assessment tool that can allow requesters to list the goal of the task and the kinds of data to be collected, and automatically mark any sensitive data (e.g., address) so requesters can evaluate whether that data is really needed.

The principle of *notice* states that "privacy policy statements [should be] clear, concise, and conspicuous to those responsible for deciding whether and how to provide the data" [55]. More specifically, the principle of *purpose specification* states that the purpose of certain data practices (e.g., collection) should be clearly specified prior to data collection [44]. Following these principles, the description of a crowdsourcing task should clearly specify the purpose of the task. As we saw in our data, workers' privacy concerns or experiences may prevent them from contributing detailed data (e.g., cleaning one's browsing history before submitting), affecting the quality of crowdsourced work. Clear descriptions of why certain data is needed for the purpose of a task can mitigate these concerns and encourage workers to provide detailed data. For instance, prior literature has shown that clear descriptions of why a personalized system needs certain user data makes users more willing to share their data [32, 58].

A concrete idea of implementing notice and purpose specification is to design a privacy policy template for crowdsourcing tasks that clarify their data practices, such as what data will be collected and how the data will be used or shared and for what purposes. For example, to address the issue of providing information for free (especially for paid crowdsourcing), the task description should clearly describe any screening component used to select qualified workers, what data will be requested during the screening, and how that data will be handled (e.g., will the data be deleted immediately) especially when a worker is not selected to perform the tasks and receive payment after the screening. Crowdsourcing platforms can incorporate the privacy policy template directly into the task template for requesters. For crowd workers, tools can be designed to identify what (sensitive) information may be collected based on the task description so workers can make informed decisions about whether to work on the task or not.

Issues related to information processing. To address issues such as data aggregation, profiling and unauthorized secondary use, privacy principles such as purpose specification, data minimization and use limitation would be useful. *Data minimization* states that "Before deployment of new activities and technologies that might impact personal privacy, carefully evaluate them for their necessity, effectiveness, and proportionality: the least privacy-invasive alternatives should always be sought" [55]. Applying this principle in crowdsourcing means that data aggregation and profiling, which involve collecting and combining multiple pieces of information about an individual worker, should be cautiously examined for their necessity with regards to the purpose of the crowdsourcing tasks. An example of a concrete idea is a tool that could provide workers with an overview of what information they have disclosed to a requester over multiple tasks, or to all MTurk tasks over time. This would not only provide more transparency to the tasks but also provide useful information that could prompt workers to think more about their disclosure decisions and whether to submit answers to a task. *Use limitation* means "Personal data should not be disclosed, made available or otherwise used for purposes other than those specified" [44]. This principle directly addresses the issue of unauthorized secondary use. The privacy policy template presented as a concrete idea for notice and purpose specification is applicable here as well.

To address the issue of de-anonymization, the principles of anonymity and unlinkability are helpful. *Anonymity* means people's real identities cannot be identified [57]. *Unlinkability* means two activities or interaction steps of the same user cannot be linked together [57]. For crowdsourcing in general, these principles mean that a crowdsourcing platform should hide workers' personally

identifiable information (PII) from requesters and prohibit linking a worker's pseudonymous ID to his/her PII. For instance, MTurk should (1) hide workers' real names and email addresses when they message requesters (e.g., using an anonymous email address generated by MTurk) to avoid potential spam; and (2) make MTurk IDs unlinkable to accounts of other Amazon services such as shopping and Amazon Web Services (AWS).

To address the issue of insecurity (in crowdsourcing platforms), the security principle is invaluable. The principle of *security* states that "Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data" [44]. To apply this principle, crowdsourcing platforms should adopt or implement state-of-the-art security measures such as security protocols (e.g., HTTPS), authentication, access control, and encryption of worker data. This is particularly crucial for paid crowdsourcing platforms such as MTurk where monetary compensations are involved and workers' financial information is stored and used.

Issues related to information dissemination. To address issues such as unauthorized sharing, the principles of purpose specification and onward transfer are relevant. As discussed before, purpose specification ensures that the purpose of a crowdsourcing task is plainly stated so that workers can make a more informed judgment about whether certain data practices (e.g., data sharing) are appropriate. *Onward transfer* means that "Personal data should not be transferred to a third party if it does not ensure an adequate level of protection" [16]. An example concrete design idea is to build a tool for workers to identify and visualize potential information flows based on task descriptions.

Issues related to invasions. To mitigate issues such as spamming, stalking and confrontation, the principles of purpose specification, use limitation, and enforcement/redress are useful. Purpose specification and use limitation would expect requesters to clarify why workers' contact information (e.g., email address) is collected and how it will be used. In the crowdsourcing context, the principle of *enforcement/redress* means that effective privacy protection must include recourse mechanisms for workers to make complaints, which would then be investigated and resolved; verification mechanisms for crowdsourcing platforms and requesters to demonstrate that their privacy protections of worker data are implemented as they claim; and remedy mechanisms that crowdsourcing platforms are obliged to apply when problems arise (e.g., when requesters stalk or confront workers) [16]. Similar to MTurk, crowdsourcing platforms in general should have a policy for requesters in terms of what they can and cannot do, should allow workers to report problematic tasks/requesters and have a procedure to investigate and resolve reported issues. The platform should enforce its policy and hold violating requesters accountable.

Issues related to deceptive practices. To mitigate issues such as phishing, malware and scams, the principles of security and enforcement/redress are helpful. It would be beneficial for crowdsourcing platforms to proactively monitor and filter out potentially problematic tasks and requesters. For example, a concrete idea is to design tools that use Natural Language Processing (NLP) techniques to parse and analyze the description of a task or the actual task (e.g., survey questions) to detect potential privacy risks (e.g., asking for sensitive personal information, use of collected data in a public or inappropriate fashion). Based on workers' reports of problematic tasks as ground truth training data, the tool could gradually build models to predict the privacy invasiveness of a task and inform workers. If a task involves visiting/using an external site or downloading a piece of software, the tool could leverage security analysis techniques (e.g., static/dynamic analysis of the software in a sandboxed environment) to proactively analyze the site or software. Similar to app markets (e.g., Google Play, Apple Store), crowdsourcing platforms that host various tasks should take the responsibility of examining the validity or riskiness of tasks before workers work on them.

5.3 Implications for the CSCW Community

Crowdsourcing has become an active research area in CSCW. Kittur et al. suggest that how to protect crowd workers' privacy should be one of the key research questions in crowdsourcing research [30]. Crowd workers are often treated as an API call [49], as a computational service [25], or as a commodity [2]. These conceptualizations highlight the imbalanced power dynamics [25] and information asymmetry [14] between crowd workers and requesters. Our study provides novel and rich empirical evidence that many MTurk workers' privacy has actually been violated and details how their privacy was violated. These results can inform the design of future privacy-friendly crowdsourcing systems.

In addition, crowdsourcing platforms such as MTurk have increasingly become a popular venue for academic researchers including those in the CSCW/HCI field to easily and quickly collect research data. While academic researchers are likely to think about the ethics of their research and have obtained ethics (e.g., IRB) approval, it is still important to remind researchers like ourselves to be mindful about the privacy implications of the research we conduct on MTurk, such as the data we collect and the proportionality of data collection and usage. In short, data practices should be proportional and justifiable to the scientific value or purpose of the research. Many respondents mentioned that some research tasks (e.g., surveys) asked them very sensitive information. While these research tasks may need to collect sensitive data for the very purpose of the research, these data practices should be made readily clear to the workers so they will not be surprised. We advocate that it is academic researchers' responsibility to set a good example (e.g., adopting a privacy policy or a privacy-enhancing tool if MTurk does not enforce them) for how crowd workers' data should be collected and managed. If academic researchers adopt these privacy-preserving practices, that may help create privacy-friendly social norms on MTurk so requesters in general will consider adopting them.

5.4 Limitations and Future Work

Our study is one of the first to investigate how crowd workers themselves think about and experience privacy in crowdsourcing and on MTurk in particular. As such, it is exploratory in nature and has many limitations. First, our findings are based on self-reported responses to a series of open-ended survey questions. While we have collected a large and diverse sample of crowd workers' privacy concerns, experiences, and expectations in the survey, interviewing crowd workers can be valuable in future research to complement our findings by providing richer data regarding workers' experiences. Second, our study focuses on MTurk, arguably one of the most popular crowdsourcing services. However, we do not claim our findings can be generalized to all crowdsourcing systems because there are other types of crowdsourcing, for instance, those that do not involve monetary incentives. Studying various types of crowdsourcing services would be another promising direction for future research. Finally, we did not explicitly ask our respondents whether the problematic tasks they experienced were from organizations, academic researchers, or other individuals. Future research can study this factor as it can shed light on the kinds of requesters that need more supervision and guidance.

6 CONCLUSION

Crowdsourcing platforms such as MTurk are empowering millions of Internet users around the world to perform tasks or contribute to research for payment. However, crowd workers' privacy can be at risk. Our study provides a novel perspective from the workers themselves in terms of how they think about and experience privacy on MTurk. Our findings uncover many privacy-related issues (e.g., phishing, malware, scams) that have actually occurred on MTurk. These findings not

only highlight the insufficiency of current privacy protection on MTurk but also inform future crowdsourcing policies and designs to better protect crowd workers' privacy. We will end by quoting a MTurk worker from our study: "*our privacy needs to be protected at all costs*."

7 ACKNOWLEDGEMENT

We thank our respondents for sharing their experiences and insights. We are also very grateful to Linlin Yu, Qiuyan Liu, Qunfang Wu, and Yaxing Yao for their assistance as well as the anonymous reviewers for their thoughtful feedback. Finally, we thank "Spamgirl," the forum administrator of Turker Nation, as well as Kevin Crowston and Jason Dedrick, for their advice on our research. This work was supported in part by NSF Grant CNS-1464347.

REFERENCES

[1] Alessandro Acquisti. 2004. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM conference on Electronic commerce*. ACM, 21–29.
[2] A Aloisi. 2015. Commoditized Workers. The Rising of On-Demand Work, a Case Study Research on a Set of Online Platforms and Apps. In *4th Conference of the Regulating for Decent Work Network*.
[3] Amazon. 2017. Amazon Mechanical Turk's General Policies. (2017). https://www.mturk.com/mturk/help?helpPage=policies
[4] Birgitta Bergvall-Kåreborn and Debra Howcroft. 2014. Amazon Mechanical Turk and the commodification of labour. *New Technology, Work and Employment* 29, 3 (2014), 213–223.
[5] Anne Bowser, Katie Shilton, E Warrick, and J Preece. 2017. Accounting for privacy in citizen science: Ethical research in a context of openness. *Proceedings of Computer-Supported Cooperative Work and Social Computing (CSCW) 2017* (2017).
[6] Anne Bowser, Andrea Wiggins, Lea Shanley, Jennifer Preece, and Sandra Henderson. 2014. Sharing data while protecting privacy in citizen science. *Interactions* 21, 1 (2014), 70–73.
[7] Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development*. Sage.
[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[9] Juan Pablo Carrascal, Christopher Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira. 2013. Your browsing behavior for a big mac: Economics of personal information online. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 189–200.
[10] Georgios Chatzimilioudis, Andreas Konstantinidis, Christos Laoudias, and Demetrios Zeinalipour-Yazti. 2012. Crowdsourcing with smartphones. *IEEE Internet Computing* 16, 5 (2012), 36–44.
[11] Kevin Crowston. 2012. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches*. Springer, 210–221.
[12] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G Ipeirotis, and Philippe Cudré-Mauroux. 2015. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 238–247.
[13] David Durward, Ivo Blohm, and Jan Marco Leimeister. 2016. Is There PAPA in Crowd Work?-A Literature Review on Ethical Dimensions in Crowdsourcing. (2016).
[14] Alek Felstiner. 2011. Working the crowd: employment and labor law in the crowdsourcing industry. *Berkeley Journal of Employment & Labor Law* 32, 1 (2011).
[15] Andrea Forte, Nazanin Andalibi, and Rachel Greenstadt. 2017. Privacy, anonymity, and perceived risk in open collaboration: a study of Tor users and Wikipedians. *Proceedings of Computer-Supported Cooperative Work and Social Computing (CSCW). Portland, OR. CSCW* (2017), 12.
[16] FTC. 2000. *Privacy Online: Fair Information Practices in the Electronic Marketplace. A Report to Congress*. Federal Trade Commission.
[17] Buddhadeb Halder. 2014. Evolution of crowdsourcing: potential data protection, privacy and security concerns under the new media age. *Revista Democracia Digital e Governo Eletrônico* 1, 10 (2014), 377–393.
[18] Christopher G Harris. 2011. Dirty deeds done dirt cheap: a darker side to crowdsourcing. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*. IEEE, 1314–1317.
[19] Davies Harry. 2015. Ted Cruz using firm that harvested data on millions of unwitting Facebook users. *Guardian* (2015).
[20] Paul Hitlin. 2016. Research in the crowdsourcing age, a case study. *Pew Research Center. http://www. pewinternet. org/2016/07/11/research-in-the-crowdsourcing-age-a-casestudy* (2016).
[21] Jeff Howe. 2006. The rise of crowdsourcing. *Wired magazine* 14, 6 (2006), 1–4.

[22] Shih-Wen Huang and Wai-Tat Fu. 2013. Don't hide in the crowd!: increasing social transparency between peer workers improves crowdsourcing outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 621–630.

[23] Giovanni Iachello and Gregory D. Abowd. 2005. Privacy and Proportionality: Adapting Legal Evaluation Techniques to Inform Design in Ubiquitous Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM, New York, NY, USA, 91–100. https://doi.org/10.1145/1054972.1054986

[24] Panagiotis G Ipeirotis. 2010. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students* 17, 2 (2010), 16–21.

[25] Lilly C Irani and M Silberman. 2013. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 611–620.

[26] Allen C Johnston, Merrill Warkentin, and Xin Luo. 2009. National culture and information privacy: the influential effects of individualism and collectivism on privacy concerns and organizational commitment. In *Proceedings of the International Federation of Information Processing (IFIP), International Workshop on Information Systems Security Research*. 88–104.

[27] Thivya Kandappu, Arik Friedman, Vijay Sivaraman, and Roksana Boreli. 2015. Privacy in Crowdsourced Platforms. In *Privacy in a Digital, Networked World*. Springer, 57–84.

[28] Thivya Kandappu, Vijay Sivaraman, Arik Friedman, and Roksana Boreli. 2014. Loki: a privacy-conscious platform for crowdsourced surveys. In *2014 Sixth International Conference on Communication Systems and Networks (COMSNETS)*. IEEE, 1–8.

[29] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara B Kiesler. 2014. Privacy Attitudes of Mechanical Turk Workers and the US Public. In *SOUPS*. 37–49.

[30] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of Computer-Supported Cooperative Work and Social Computing (CSCW)*. ACM, 1301–1318.

[31] Predrag Klasnja, Sunny Consolvo, Tanzeem Choudhury, Richard Beckwith, and Jeffrey Hightower. 2009. Exploring privacy concerns about personal sensing. In *International Conference on Pervasive Computing*. Springer, 176–183.

[32] Alfred Kobsa and Max Teltzrow. 2005. Contextualized Communication of Privacy Practices and Personalization Benefits: Impacts on Users' Data Sharing Behavior. In *Privacy Enhancing Technologies: Fourth International Workshop, PET 2004, Toronto, Canada*, David Martin and Andrei Serjantov (Eds.). Vol. LNCS 3424. Springer Verlag, Heidelberg, Germany, 329–343.

[33] K Kripendorff. 2004. Reliability in content analysis: Some common misconceptions. *Human Communications Research* 30 (2004), 411–433.

[34] Marc Langheinrich. 2001. Privacy by Design - Principles of Privacy-Aware Ubiquitous Systems. In *Proceedings of the 3rd International Conference on Ubiquitous Computing (UbiComp '01)*. Springer-Verlag, London, UK, UK, 273–291. http://dl.acm.org/citation.cfm?id=647987.741336

[35] Walter S Lasecki, Mitchell Gordon, Winnie Leung, Ellen Lim, Jeffrey P Bigham, and Steven P Dow. 2015. Exploring Privacy and Accuracy Trade-Offs in Crowdsourced Behavioral Video Coding. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1945–1954.

[36] Walter S Lasecki, Jaime Teevan, and Ece Kamar. 2014. Information extraction and manipulation threats in crowd-powered systems. In *Proceedings of Computer-Supported Cooperative Work and Social Computing (CSCW)*. ACM, 248–256.

[37] Walter S Lasecki, Jaime Teevan, and Ece Kamar. 2015. The cost of asking crowd workers to behave maliciously. In *Proc. the AAMAS Workshop on Human-Agent Interaction Design and Models*.

[38] Matthew Lease, Jessica Hullman, Jeffrey P Bigham, Michael S Bernstein, Juho Kim, Walter Lasecki, Saeideh Bakhshi, Tanushree Mitra, and Robert C Miller. 2013. Mechanical turk is not anonymous. *Available at SSRN 2228728* (2013).

[39] Paul Benjamin Lowry, Jinwei Cao, and Andrea Everard. 2011. Privacy concerns versus desire for interpersonal awareness in driving the use of self-disclosure technologies: The case of instant messaging in two cultures. *Journal of Management Information Systems* 27, 4 (2011), 163–200.

[40] Jennifer Marlow and Laura A Dabbish. 2014. Who's the boss?: requester transparency and motivation in a microtask marketplace. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2533–2538.

[41] Steven Marsh, Ian Brown, and Fayaz Khaki. 2009. Privacy Engineering Whitepaper: A Report from a Special Interest Group of the Cyber Security KTN. (2009). https://ssrn.com/abstract=1763248

[42] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 224–235.

[43] Bran Nicol. 2006. *Stalking*. Reaktion Books.

[44] OECD. 1980. *Recommendation of the Council Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*. OECD.

[45] Jacki O'Neill and David Martin. 2013. Relationship-based Business Process Crowdsourcing?. In *14th International Conference on Human-Computer Interaction (INTERACT)*. Springer, 429–446.
[46] Eyal Peer, Gabriele Paolacci, Jesse Chandler, and Pam Mueller. 2012. Screening participants from previous studies on Amazon Mechanical Turk and Qualtrics. *Unpublished Manuscript* (2012).
[47] Emilee Rader and Rebecca Gray. 2015. Understanding user beliefs about algorithmic curation in the Facebook news feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 173–182.
[48] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*. ACM, 2863–2872.
[49] M Silberman, Lilly Irani, and Joel Ross. 2010. Ethics and tactics of professional crowdwork. *XRDS: Crossroads, The ACM Magazine for Students* 17, 2 (2010), 39–43.
[50] H. Jeff Smith, Tamara Dinev, and Heng Xu. 2011. Information Privacy Research: An Interdisciplinary Review. *MIS Quarterly* 35, 4 (Dec. 2011), 989–1016. http://dl.acm.org/citation.cfm?id=2208940.2208950
[51] H Jeff Smith, Sandra J Milberg, and Sandra J Burke. 1996. Information privacy: measuring individuals' concerns about organizational practices. *MIS quarterly* (1996), 167–196.
[52] Daniel J Solove. 2005. A taxonomy of privacy. *University of Pennsylvania Law Review* 154 (2005), 477.
[53] Sarah Spiekermann and Lorrie Faith Cranor. 2009. Engineering Privacy. *IEEE Transactions on Software Engineering* 35, 1 (2009), 67–82.
[54] H Colleen Stuart, Laura Dabbish, Sara Kiesler, Peter Kinnaird, and Ruogu Kang. 2012. Social transparency in networked information exchange: a theoretical framework. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 451–460.
[55] USACM. 2006. *USACM Policy Recommendations on Privacy*. Technical Report. U.S. Public Policy Committee of the Association for Computing Machinery.
[56] Yang Wang, Yun Huang, and Claudia Louis. 2013. Respecting user privacy in mobile crowdsourcing. *ASEScience* 2, 2 (2013), 39–50.
[57] Yang Wang and Alfred Kobsa. 2009. Privacy-Enhancing Technologies. In *Social and Organizational Liabilities in Information Security*, Manish Gupta and Raj Sharman (Eds.). IGI Global, 203–227.
[58] Yang Wang and Alfred Kobsa. 2013. A PLA-based privacy-enhancing user modeling framework and its evaluation. *User Modeling and User-Adapted Interaction* 23, 1 (March 2013), 41–82. https://doi.org/10.1007/s11257-011-9114-8
[59] Samuel D Warren and Louis D Brandeis. 1890. The right to privacy. *Harvard law review* (1890), 193–220.
[60] Ming Yin, Mary L Gray, Siddharth Suri, and Jennifer Wortman Vaughan. 2016. The communication network within the crowd. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1293–1303.

Received April 2017; revised July 2017; accepted November 2017