Privacy in Human Computation: User awareness study, Implications for existing platforms, Recommendations, and Research Directions

Mirela Riveni[1], Christiaan Hillen[2], Schahram Dustdar[1]

[1] Distributed Systems Group, TU Wien
1040 Vienna, Austria
{m.riveni,dustdar}@infosys.tuwien.ac.at
[2] Independent Researcher
christiaan@bitweasel.net

Abstract. Research and industry have made great advancements in human computation and today we can see multiple forms of it reflected in growing numbers and diversification of platforms; from crowdsourcing ones, social somputing platforms (in terms of collaborative task-execution) and online labor/expert markets, to collective adaptive systems with humans-in-the-loop. Despite the advancements in various mechanisms to support effective provisioning of human computation, there is still one topic that seems to be close to neglected both in research and the current design and development of human computation systems, namely *privacy*. In this work, we investigate this problem. Starting from the fact that user awareness is crucial for enforcing privacy-respecting mechanisms, we conducted an online survey-study to asses user privacy-awareness in human computation systems and in this paper provide the results of it. Lastly, we provide recommendations for developers for designing privacy-preserving human computation platforms as well as research directions.

Keywords: Privacy, Privacy Awareness, Social Computing, Human Computation, Crowdsourcing, Collective Intelligence

1 Introduction

Human computation is a concept that has already gained its momentum and its application is evolving in rapid pace with the development of different types of platforms and mechanisms for effective utilization of human intelligence online. Human computation as a term is first coined by Luis von Ahn in [36]. It has been defined as the utilization of human intelligence for tasks, activities and problems that cannot yet be executed and solved by artificial intelligence. Hence, some call human computation *artificial artificial-intelligence*. In this work, we put several concepts that include online task-execution by people under the umbrella of the Human Computation concept (varying a little from the taxonomy presented in [23] by Quinn et al.). Those concepts are: Crowdsourcing, which involves simple

task execution by a large number of (anonymous) people (see a survey in [38]); Social Computing, with which we imply online computations where multiple people are involved in complex task execution (see examples in [10],[25]); Online Labor Markets and Expert Networks; Human-based services in mixed systems in which people provide their services/skills within Service Oriented Architectures [29]; as well as human computation in Collective Adaptive Systems (CAS) [35], [39]. The last type, Collective Adaptive Systems, are distributed large-scale systems that are flexible in terms of number and type of resources, including human resources, and include complex task execution with a high number of interactions between resources. In these systems, privacy is even more relevant than in crowdsourcing, where tasks are simple and interactions are not common. Moreover, the utilization of cloud services that are inherent in the definition of CAS, ads a lot of weight to the importance or privacy; using public clouds requires a high amount of trust on the providers as user data, and generated content and artifacts are usually dependent on using cloud services but are also stored on the Cloud. Security mechanisms for clouds, along with regulations on data management, are of paramount importance if privacy is to be preserved.

While there is a solid amount of work on building human computation systems and mechanisms that support efficient management, such as task assignment and management (e.g., routing and delegations), worker management, incentive and payment models for workers online and quality assurance, little research has been conducted on the privacy implications in these systems. Privacy, however, is a human right and as such these users are also entitled to it. Article 12 of the Preamble of the Universal Declaration of Human Rights states for example that "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation." The design of human computations systems as well as everything else on the web, needs to be guided by privacy preserving principles.

The key contributions of this paper are: 1) a discussion of why and how user data is utilized in particular human computation systems, 2) an analysis of user privacy-awareness on human computation platforms through the results of a user study that we conducted with an online survey, and 3) recommendations for human computation stakeholders, along with some research directions.

The remainder of this paper is organized as follows: In Section 2 we present related work. Section 3, presents a discussion on privacy-implications in human computation, strictly speaking about data collection, data utilization and privacy risks. In Section 4 we present our user study and make an analysis of user privacy awareness on human computation platforms. We present a few recommendations for privacy-preserving mechanisms in Section 5 and provide some considerations for possible research direction in Section 6. We conclude the paper in Section 7.

2 Related Work

Smith et al. have studied individual privacy concerns in organizations and have identified multiple dimensions of these concerns that they have presented in [31].

In that work, the authors list four factors critical to consider when assessing user privacy concerns: concern over data *collection*, *errors* in user data, *unauthorized secondary use* of user data (e.g., when data is used for other purposes then stated), and *improper access* to user data. Since then, a number of other models, frameworks of user privacy concerns have been presented, (some building upon the work of Smith et al.) such as [19]. On the other hand, theories and models of how to control privacy and enforce privacy-aware mechanisms are also present in existing literature. Authors in [20] present such a privacy-control theory and discuss ways of its application in online environments.

Fischer-Hübner and Martucci in [12] have inspected privacy implications for Social Collective Intelligence Systems (SCIS) by presenting an overview of the European Data Protection Legal Framework and relating the privacy rules provided by this framework with SCIS systems and mechanisms supported by these systems, such as user profiling for reputation scores, incentive models, as well as data provenance. Reputation, incentive models and data provenance are all listed as risks for user privacy as these mechanisms are inherently designed to work with user profiling. However, the authors list and discuss some of the available tools and technologies that can enable SCIS platform providers to respect and preserve user privacy, mainly via pseudonyms and anonymity. One example is by allowing users to use different pseudonyms for different roles, i.e., context-based pseudonyms that can be used only once per role (e.g., skill type) and thus prevent misbehavior by malicious users. In addition, the authors describe anonymous credential protocols that can be used to create new credentials whenever a user wants, with less or different certificate attributes, which cannot be linked to an original certificate by the verifier and the issuer. Moreover, Fischer-Hübner and Martucci also present Privacy Policy Languages (such as PPL) with which platforms can make negotiations and come up to an agreement with platform-users on *how*, *by whom* (and *what*) data can be accessed, processed and logged.

Motahari et al. in [21] have listed privacy threats in ubiquitous social computing by underlining the social inference threats in social computing, where a user can be identified for example through contextual information (e.g., location) or social links. Authors in [13] present a privacy model together with a framework for task-recommendation in mobile crowdsourcing. The model is based on enabling workers to share information (e.g., location) with a recommendation server by choosing how much and what type of information they wish to share. Task recommendations are based on the information shared by workers. However, authors conclude the obvious, namely that achieving a high efficiency in task-recommendations means low level of privacy. Another work that presents a privacy-aware framework is presented in [33] by To et al. The authors present a task-assignment algorithm that preserves location privacy for mobile spatial-crowdsourcing tasks, that is, for tasks that require workers to be at a specific location. Toch in [34] investigates privacy preferences of users in mobile context-aware applications through crowdsourcing and presents a method to calculate user privacy tendencies. He suggests building distributed systems to tackle privacy risks (with the computation of user privacy tendencies being executed on

the client side). Privacy preservation in decentralized systems is discussed in [5]. Privacy activists also advocate for decentralization and zero-knowledge systems, Balkan for example has written the Ethical Design Manifesto [4], which states: "Technology that respects human rights is decentralised, peer-to-peer, zero-knowledge, end-to-end encrypted, free and open source, interoperable, accessible, and sustainable.".

Langheinrich in [17] discusses some Privacy by Design principles, focusing on ubiquitous systems. He states that the Principle of Openness, or Notices, is an important principle during data collection. Users have to be informed when they are being monitored. In addition, Langheinrich discusses that consent should be required in a more flexible way than the "you can use our services only if you consent to our terms/policy". Users have to be able to use services while opting out of unwanted features.

3 Personal Data on Human Computation Systems

3.1 Collected data

We investigated the collected data from some of the existing crowdsourcing/expert-labor market platforms (by actually creating accounts), such as Amazon Mechanical Turk, Microworkers, Freelancer, Upwork, PeoplePerHour, TopCoder and uTest. The information required to *build up a profile* and/or *verify a profile* sees variations from platform to platform. The following list provides some data required to build and verify profiles generalized through platforms; not every platform requires everything on the list, but everything on the list is required in various platforms.[3]

- Full mailing address - Sometimes even documents are required to prove address, such as utility bills or bank statements.
- A government issued ID - Passport, ID card or Driving license.
- Photograph
- Code verification along with a users face on a photograph - In some platforms workers need to send a photograph of themselves where they hold a piece of paper with a code provided by a platform written on the paper.
- Educational experience - In some of the platforms filling out at least one educational experience is mandatory.
- Job title - In some platforms filling out a job title is mandatory.
- Bank account information
- Data from mobile-sensing - Depending on the application domain, other sensitive data may be collected at runtime while users are working on tasks

[3] We have investigated the information required by platforms in order to gain a better insight and to guide our discussion in this work along with providing improvement suggestions for researchers and industry alike, which could bring more privacy-aware platforms in the future. We do not intend to imply malicious use of the users' data by the aforementioned platforms as we have not conducted an investigation regarding the manner of usage of the collected data.

or just wearing a smart device. For example in mobile crowdsourcing applications, such as crowd-sensing, location information, health information (that users share through wearables to applications) and other data may be collected which can be used to identify and profile a user.
- Device and connection data - Basic system fingerprinting such as IP address, browser type and operating system.

The EU Data Protection Directive (DIRECTIVE 95/46/EC) [8], stipulates that personal data is "any information relating to an identified or identifiable natural person". Although this directive will be repealed by the General Data Protection Regulation [24] (with the enforcement on 25 May 2018), the principal definition for personal data remains the same, as can be found in Article 4 of the Regulation (under (1)). Moreover, identification is the singling out of an individual within a data-set [22], even if his or her name or other attributes that we typically associate with an identity remain unknown. Consequently, most of the aforementioned information can be used to identify a person and that means this data is personal and thus should be kept private.

3.2 Reasons for collecting personal data

The General Data Protection Regulation (REGULATION (EU) 2016/679) [24] defines profiling as "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyze or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behavior, location or movements". In Human Computation, the collection of user data and building user profiles is usually not conducted for big-profit purposes (such as selling user profiles to advertisers) but for designing and developing mechanisms for effective work-management on human computation provisioning systems. In the following we discuss some of these mechanisms.

Task-Assignment and Formation of collectives A number of existing works have presented (expert) discovery and ranking algorithms for service-oriented architectures with human-provided services ([28], [11]), as well as non-service oriented systems in which tasks assignment is based on qualifications. Research in team formation in expert networks is also being investigated ([2], [9]). These mechanisms are all based on logged and historical data about workers, for discovering appropriate ones for specific tasks in individual-crowdsourced work or specific collective-work.

Management Mechanisms Managing individual and team-based worker performance in human computation is also important for complex systems with automated processes even for the human-in-the-loop coordination. Adaptation mechanisms in human computation within SOA for example, are discussed in [30], and an adaptation mechanism for elastic collectives based on a trust model

is presented in [26]. Algorithms that calculate worker's performance can be used within delegation mechanisms, where a task is delegated to another worker that may or may not belong to the initial collective. Consequently, a new worker can be added to an existing collective at run-time. Research on both task assignment and adaptation algorithms that include measurement of worker performance is in big part based on ***trust*** and ***reputation*** models. Some of these trust and reputation models include not only metrics that can be measured automatically (such as task success-rate) but also *social trust*, which mostly is defined and calculated as a trust-score that is given to a worker by collaborators or acquaintances and/or by work-requesters/clients based on their satisfaction by the results. Social trust is often subjective and in most cases requires that the person rating a worker knows the worker personally, so in some trust and reputation models worker identities are needed to be known. Hence, in some cases reputation mechanisms are in conflict with privacy, and we need to find ways of bringing these two concepts together and provide privacy aware reputation strategies.

Quality of Service Keeping quality at a desired level also requires monitoring of workers. (see an example of SLA based QoS monitoring for crowdsourcing in [14]).

Misbehavior prevention Personal data is sometimes used in building mechanisms for preventing worker misbehavior, such as *Sybil attacks*- cases in which workers can easily create multiple profiles and make more gain by executing the same tasks multiple times.

Incentive Mechanisms Developing and applying appropriate incentive models for workers based on monetary and non-monetary gains is also based on data collection, and sometimes personal data, for example when reputation is used as an incentive. (See work on incentives and rewarding mechanisms in Social Computing in [27].)

Payments In most platforms that have implemented monetary incentives, user bank/credit account is required so that workers can be paid. However, some platforms allow users to delete this data after they withdraw the required amount from it (such as microworkers.com, in the case of requesters/clients). This should be a standard practice for all data types as well. A regulation example for this is the GDPR's *Right to erasure* described in Article 17. We are of the opinion that human computation platforms should consider the provisioning of mechanisms with which subjects would have the possibility to directly modify, update, or erase data without making a request to the controller (the owner of the crowdsourcing platforms), just like microworkers allows the deletion of bank account information without any explicit requests.

All the aforementioned mechanisms require monitoring of workers. However, in research frameworks and systems, no collection of personal data is mentioned explicitly, and the privacy aspect is not tackled, except in specific research presented in Section 2. On the other hand, almost all existing platforms in industry require users to share their personal data. Thus, we advocate for all the aforementioned mechanisms to be considered together with their privacy-related implications. In the next section we examine privacy implications by analyzing a few risk-factors that we identified as most relevant.

4 Privacy Risks

User Privacy Policy Awareness Users usually agree to Privacy Policies without actually reading them, and they do not read them because they are too long or too complex to understand ([37]). This contributes to their use of platforms without being aware of what they are entitled to, regardless of whether their privacy rights are violated or not.

Lack of Transparency in Privacy Policies Very often, users are not given complete information about what is considered under *personal information*, about how their personal data will be used, whether it will be shared with third parties and for how long will this information be stored on a Human Computation platform-provider's servers or on the servers of their service-providers. Often, the use of personal data is defined in policies in a vague way. Consider for example the statement "we may share certain data...", using words such as "certain data" without concretely defining what type of data the statement is referring to means getting a clear consent by users to use whatever personal data of users that the provider owns. Many privacy policies also contain phrases such as the following: "we may share information with third parties for industry analysis, research and other *similar purposes*"; the terms "similar purposes" give the providers the freedom to use personal data in any purpose fitting their needs, without the explicit consent of the platform users. In addition, consider this statement: "we may use your personal information from other services and connect to your account information when necessary", this is clearly a way to de-anonymize users even if the platform is designed to use pseudonyms, as combining various data-sets (e.g., by email addresses) de-anonymizes users. Selling user profiles to the ad industry is also a possibility, although many human computation platforms do not have an ad-based business model. However, companies called *data brokers* continuously collect data about people from multiple online (and even offline) sources and sell that data to clients in various business domains. Data that they collect can be also retrieved from websites with log-ins, *browser* and *device fingerprints*. Because privacy policies are not straightforward, people can not be sure whether data brokers are not leveraging some data from human computation platforms as well.

Profiling User information is collected through information that they share with platforms as well as by automatically collecting data by tracking. Among other uses such as computing reputation scores, this data may also be utilized to group people in different categories, by various contexts (e.g., country of residence, gender). This may be used to set up rules for task-assignment, and discriminate certain groups during task-assignment and rewarding. As we mention in our study results, there are cases where workers from certain countries are rewarded less than others for the same tasks. In addition, consider for example a real danger through crowdsourced tasks for online-monitoring of certain locations with the purpose of detecting and reporting criminal activities, or even a different setup, such as identifying wanted offenders from a set of pictures that are posted in a crowdsourced task online. If these platforms are profiling workers, criminals might find ways to identify workers, and workers' lives could be put at risk. Furthermore, if information from political crises-response crowdsourcing sites[4] about people reporting incidents in war-struck areas fall into the wrong hands, it might also pose risks for the reporters, who might be non-tech savy citizens. These may seem as extreme examples, but serve well to validate privacy concerns. On the other hand, many companies hide behind the term "anonymity", for example they do not require real names and allow people to register with pseudonyms while collecting other personal data, in this way wrongly convincing people that they actually work online without being identified. For example, authors of [6] cite a study revealing that the combination of zip code, gender and birth-date data had been unique for 216 million US citizens, and consequently citizens can be identified without any other additional data. In the same work, authors also cite another study showing that four data points, such as four sets of time and location data could be used to uniquely identify people. Thus, leaving out some data while collecting other type of data does not mean that anonymity is achieved, and in most cases people are not transparently informed of this fact.

Lack of Control In current systems, users do not control how their private information is used (whether it is shared, sold or misused), and have no control over who accesses that information. They have to be content with what they read on privacy policies (when they read them). They are not given control to their own data, to update or delete their data when they want. Moreover, stored information is sometimes not secured enough, rules and regulations are not always respected, and data stored on foreign servers belonging to a different jurisdiction than a person's residence country (over which the user may not be given a choice), can be misused (e.g., due to security breaches, unencrypted data, unethical employees or security agencies).

Lack of Ownership Users do not own their own data that they have shared with the platforms, rather platform providers do. Similar to the risks in not

[4] See for example: https://syriatracker.crowdmap.com/

being able to control data, not owning data means intentional or unintentional sharing with third parties, access to user data by unintended parties as well as transferring/selling user information to other parties and monetizing people's data, which sometimes can be done without users' knowledge and approval.

Lack of Security Last but not the least, in addition to the aforementioned factors, security is of paramount importance for protecting privacy. Data control and ownership do not have any effect on privacy protection if user-information is unencrypted. Needless to say, security protocols for securing internet connections and data encryption protocols should be standard features that every human computation platform should support.

5 Study

5.1 Method: Survey Design and Distribution

We conducted a study to asses user privacy-awareness in human computation with an online questionnaire, hosted on a server at Technische Universität Wien/TU Wien. We asked participants a series of questions that we designed specifically to get their opinion on their private data collected and utilized on the platforms, to get their knowledge on privacy implications on these platforms as well as their concerns.

We disseminated our survey in two ways: 1) by sharing it with fellow researchers and colleagues by email, and with acquaintances and friends on social networks by asking them to fill it in (if registered as users on these systems) or send the survey to people that they know are using these systems as requesters or workers; and 2) by creating a task/campaign at Mircoworkers (https://microworkers.com/) and a HIT batch on Amazon MechanicalTurk, asking workers to fill in our survey (at a given link). We did two rounds of the survey, as we came up with some more questions that we saw relevant during our study. Thus, the first round of the survey had 20 questions, 16 of which were designed to asses user privacy-awareness, and 4 were statistical questions to get demographic data. We submitted this survey on Microworkers and to researchers and freelancers through private communication, while the second round of the survey had 5 additional questions, and was conducted only Amazon MechanicalTurk. Where we have less participants for the newly added questions we mention it when discussing the particular questions.

Microworkers has a design that allows requesters to select what user-base to chose depending on worker country-information and makes payment recommendations according to country-dependent ratings. We created four tasks/campaigns and asked users from four different country-groups to fill in our survey. For the first three groups of workers, we paid workers $0.42 per task, as by investigating other studies on these platforms (that have used payments between $0.10 and $1) we concluded that this amount was enough for people that are interested on the topic to accept the task and low enough to discourage misbehavior by those who may want to fill in the survey without interest and spam the

Table 1. Demographics of participants

Education	**Percentage**
Primary School	1.97%
High-School	20.6%
Undergraduate studies/BSc/BA	40.6%
MSc/MA/Specialty training	15.6%
Dr/PhD	6.86%
Postdoctoral researcher	1.4%
Other	12.7%
IT Knowledge	**Percentage**
Expert/Professional	2-%
Medium level (good IT skills but not expert/professional)	59%
Knowledge to get around online	21%

Table 2. Most common collected data

Collected data	**% of users who want to hide the data**
Name and Surname	24%
E-mail address	22%
Phone number	61%
Birthdate	26%
Photograph	54%
Location information/Mailing address	35%
Utility bills (sometimes used to verify address)	53%
A government issued ID	68%
Bank account information	54%
None of the above	1%

results. In spite of recommendations for lower payment for a group of workers coming from countries rated lower, we decided to pay workers of a lower rated group of countries the same amount of $0.42 and not lower. However, our survey-task for a forth group of workers residents of high rated countries was rejected for that amount as the minimum payment was $1.00 per task for surveys such as ours, so we paid that amount to get our survey completed by a fourth group, as we needed different demographics. In this regard, we strongly encourage ethical payment methods by crowdsourcing platforms and ethical payment behavior by work requesters/clients. Some of these mechanisms could be, *equal pay per task-type* for all workers (regardless of country ratings), or individual worker payments based on *quality of results*. Nevertheless, this side-experiment allowed us to qualify the answers that were submitted in reply to our survey, for example filling in optional questions that required more elaboration. In this context, we noticed no difference in the answers of higher rated countries compared to lower rated groups. In fact, some users that were paid less gave elaborated answers while none of the higher paid participants did this. Workers on Amazon MTurk were paid $1.

To filter submissions as well as to avoid spammers and malicious users who fill in the survey without reading the questions and answers, we included a few questions that helped us asses (to some level) the honesty of participant answers. One such "testing" question was added to check a related "yes or no question", the testing question included radio buttons with more elaborating statements to be chosen if the user answered with "Yes" on the related question, and included a radio button with the statement "I answered with "No" on the previous question" to be selected if a participant has answered with "No" on the related question. In addition, we added a "Yes or No question" asking survey participants if they have read our consent form for the survey and excluded submissions of participants who answered with "No". A few participants had filled our survey multiple times. We counted only one submission from these participants and excluded all duplicates.

5.2 Results and Analysis

Demographics We had a total of 204 participants, the answers of three of which we excluded as a consequence of their negative answers to the question with which we requested participant consent for the survey (through reading our consent form). 105 participants were workers from Microworkers, 78 from Amazon MTurk (engaged in the second round of our study), whereas 21 were participants that we contacted by mail/social networks. Most of our participants were residents of US, EU countries and India.

Table 1 shows the level of education and IT knowledge of our participants. Participants with a PhD and Postdoc level of education were users of human computation platforms as requesters of work (for research purposes).

We asked participants to fill in the names of up to three platforms that they use and we got the following variety of platforms as answers: Microworkers, Amazon Mechanical Turk, Upwork, InnoCentive, Elance, Guru, 99designs, CrowdFlower, clickworker, RapidWorkers, ShortTask, Testbirds, cashcrate, fiverr, scribie, TranscribeMe, foulefactory, ideaCONNECTION, and OneSpace. Most of our participants worked on two or three platforms.

Privacy Awareness We posed three type of questions assessing user privacy awareness and concerns: the first was related to a) *data collection, control and ownership of data;* the second was related to b)*anonymity online* and the third group of questions was related to c) *regulations and policies.* In the following we discuss the results.

a)Data collection, Usage Concerns and Security Regarding data collection, we asked participants to state their level of concern regarding the fact they they have to share sensitive data to register and verify their accounts. The *level of concern* question was set up as a 1-5 Likert scale. 19% of participants stated that they are somewhat concerned with the information that they are obliged to share when they register on the platforms, while 25% were not concerned at all. Table 3 gives a more detailed overview of the answers regarding participant concerns over the collection of their personal data. In addition, to get more detailed information, we listed some of the data types (mentioned in Section 2) collected by platforms and asked participants to select which of the given data type they would want to hide. They could chose multiple types. Most of the participants answered that they would prefer hiding: **a government issued ID** (68%), **bank account information** (54%) and **phone numbers** (61%). More details regarding the results for this question are given in Table 2.

Next, we asked participants if they ever provide false information when registering and creating their profiles and 17.7% reported that they do. Regarding the reasons for providing false information, 19% reported that they do not feel comfortable revealing some specific information about them, and 3% reported that they provide some false information in order to create secondary accounts.

Users' knowledge on where their data is stored is important for assessing their privacy awareness; hence, we asked participants whether they know and whether

Table 3. Data concerns

Survey statements	Very	Somewhat	Neutral	Not concerned	Not at all	Likert score
How much are you concerned with the personal information that you are obliged to share so that you can register on these platforms?	19%	8%	39%	9%	25%	2.78
How much are you concerned with the personal information that you need to share so as to build your profile and verify your identity on these platforms?	19%	22%	25%	12%	22%	2.83
Are you concerned that your information will be misused (by the /platform that you are registered with)?	6%	11%	23%	28%	32%	2.16

Table 4. Security related statements

Survey statements	I strongly agree	I agree	Neutral	I disagree	I strongly disagree	Likert score
Having in mind that not only my personal information but also content that i produce or expect as a result from engaging on a microwork platform is sensitive, I expect the platform I perform micro-work on, the prove on a regular basis (every three months?) that it is secure, by having an independent pen-test performed and have the results published.	15.06%	24.68%	36.98%	9.59%	13.69%	3.18
"I would like to have the option to receive payouts in a privacy-friendly cryptocurrency." Please select a choice from 1 to 5, 1 indicating that you are not concerned with secure and private payouts, 5 indicating that you strongly agree with the statement.	10.96%	20.55%	32.88%	10.95%	24.66%	2.82

they are concerned if their data is stored on platform providers' own servers, or if platform-providers utilize Cloud services (in which case an agreement should exist between platform providers and Cloud providers for protecting user information not sharing user data with other parties), and if data is stored in a location with a different jurisdiction (in which case different data protection regulations exist). Participants were given three answer choices and they reported as follows, 32.29% said "I admit I have never thought about these things and frankly I am not concerned.", 33.86% chose "I admit I have never thought about these things but I became concerned now.", and 33.85% answered with "I have thought about these things and am concerned." Table 4 shows security-related statements that we added for the second round of the survey and the replies from 78 participants recruited from Amazon MTurk.

b)Anonymity Anonymity is of course fundamentally different from privacy. Privacy means that people *may* be identified online, but it should be their choice regarding how much and in what way their data is shared and utilized. Nevertheless, the two concepts are invariably related. Hence, we examined opinions on anonymity as well, and asked participants what would be the reasons in the case

they prefer to work anonymously. Some workers that are working on more complex tasks (e.g., projects such as those posted on 99designs, Freelancer) and not on micro-tasks stated that they would not prefer to work anonymously online, using statements such as "working anonymously is not effective". Consequently, we assume that these workers do care about reputation as reputation mechanisms bring more clients and work. However, some replied that they would want to work anonymously in cases if they are working on some projects on which they would not want to put their name on but they are well paid, and communication and collaboration with clients is satisfactory. In addition, one participant answered that it would be nice if users are provided with the option to work anonymously online whenever they chose to (opt-in/opt-out).

On the other hand, most of the workers who work on micro-tasks answered that they would want to work anonymously for several reasons: they do not want their name to be associated to the type of work they do, to protect their banking information, they do not want the companies with which they work full-time to know that they are doing a side-job. We quote some answers stating other contexts of concern for anonymity: "I would want to work anonymously so there was no bias towards me based on my demographics and/ or social class. I also would prefer to remain anonymous in case scammers entered the platform pretending to collect data, but instead, they were going to participants homes etc.", "I like minimizing my digital footprint as much as possible", "When doing microwork online, you do work for various people, potentially over dozens of people a day. I'd rather not have my sensitive information potentially available to all of them, when I'm forced to provide demographic information for much of the work anyway", "I'd not want to have that information available for marketers or to be available to be sold. I'd not want other organizations to be able to access such information and use it to send me ads or other materials" and others. In addition some answered they would want to work anonymous and because they do not want their earnings to be reflected on their taxes. Related to the latter, one participant stated that he uses foreign money transfer services, such as Payoneer, to avoid taxes for online work.

Furthermore, some participants stated their concern of their information being leaked to other parties. One particular participant stated that the reason he would want to work anonymously is that he can not be certain by whom and how his private information will be used, he added the statement "I want to control my "web" identity as i want.".

Thus, we can conclude that workers doing complex tasks are more inclined to identification than workers executing micro-tasks that are easy to execute. Lastly, an interesting answer we encountered was: "I would want to protect my privacy", even though the specific question was related to anonymity online. Consequently, participants associated anonymity with privacy.

c)Regulations and Policies To asses participant engagement in privacy issues we went a step further and asked them if they read privacy laws, directives and policies. Figure 1 provides their reports. Interestingly enough, more than 50% of participants reported that they do read privacy policies when they register on

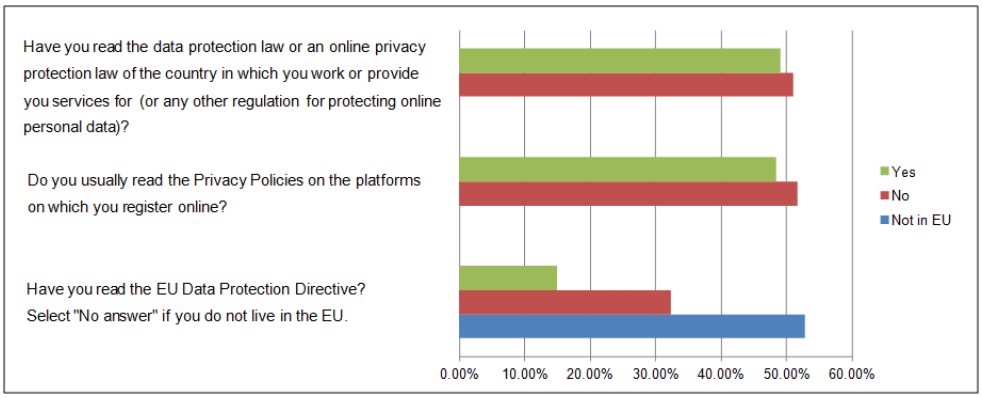

Fig. 1. User reports on regulations and policies

platforms. However, around 40% of participants reported that they do not read them. We take this result to be truthful as the number of participants is small. If we had a bigger number of participants we assume that this ratio would be significantly in favor of participants that do not read privacy policies because of complexity, as existing research suggests (see [15], [37]).

In order to get participants' opinion on platforms, research and regulations on privacy we included a question with a five Likert-scale agreement levels to chose from, for a few statements that we compiled. Most participants agreed that existing platform providers need to be more transparent about how they use personal information and they also agreed that research and industry should increase their efforts in providing mechanisms that will enable people to control and even own their data. For every statement we also asked participants if they are knowledgeable on the topics that the statements refer to or not. In total 57% answered that they need more information on the topics, 38% answered that they have knowledge on the topics and the rest did not answer. Detailed results are given in Table 5.

6 Suggestions

6.1 Recommendations

For platform providers, an important privacy-respecting guide in storing personal data is to only store that which is essential to the needs of the platform. With this, we mean the concept of data minimization; if n data points are enough to perform the task for which these points were collected, do not collect $> n$ data points. In the most general terms, according to Colesky et al., in [7], there are two directions of strategies to protect the privacy of clients: data oriented, and policy oriented. These two directions lead to eight high level strategies that can be applied to the collection of data in ways that respect the privacy of data subjects.

Table 5. Opinions on regulations, and approaches in research and industry

Survey statements	I strongly agree	I agree	Neutral	I somewhat disagree	I disagree	Likert score
Human computation Companies/Platforms should clearly state under which country/state law they operate.	28.13%	41.67%	25%	2.6%	2.6%	3.9
Companies/Platforms than enable and provide human computation should be more transparent about how they use my personal information	38.54%	29.67%	21.36%	7.31%	3.120	3.84
I am concerned about the privacy regulations/laws of the country in which i reside and work.	15.1%	30.73%	31.25%	18.23%	4.69%	3.3
Research and industry should increase their efforts in giving users more control over the use of their data.	30.73%	38.02%	19.79%	6.77%	4.69%	3.83
Research and industry should increase their efforts in enabling tools and mechanisms that will enable users to own their own data, in contrast to current standards where companies own users' data.	29.68%	33.85%	28.13%	4.17%	4.17%	3.81

One method of identity protection that lies in between tools for users, and something that developers will have to implement is attribute based credentials (ABC) [16]. In short, ABC provides the client with a set of credentials such as "over 18" or "holds an MSc in computer science". The classic example of the use of ABC's is in buying alcohol: A person must be of legal age in order to buy alcohol, but the only classic way to prove this is by showing an identity card [1]. This card contains more information than is needed, such as exact date of birth, social security number, and full name. The only requirement to know is "over 18", which in ABC's can be presented as an attribute. The beauty of this system, is that the attribute itself is not trackable, the next time the same client needs to prove the same attribute, this instance is unlinkable to the previous instance. Thus, attributes can be used to hold all information that is needed for the service, placing them under the control of the client, rather then storing them on the server in a user-profile. Well-known anonymization methods are the k-anonymity model, presented in [32], and t-closeness described in [18], which prevents attribute disclosure going beyond the limitations of k-anonymity .

However, these strategies do not solve all the practical problems. In the end, a client still has to provide at least some personal data in order to receive a reward for his work, such as a payment method. In this context, one could argue that cryptographic currencies such as BitCoin could be used to pay rewards, but these too have been shown not to be entirely anonymous [3]. In addition, there are other alternatives of online services that offer means to transfer funds, such as CashU, and Perfect Money. These services typically let one transfer funds from a legitimate source (such as a bank account) to them, and then allow transfer between accounts within the service itself. As such, these services can be seen as a 'Trusted Third Party' for money exchange. Although tracking is thus made

more difficult, it is still quite possible, as there is a single party that still needs to know enough to be able to transfer funds.

Ideally, a completely anonymous client would perform work, and be rewarded in an untraceable way. The technologies exist to make this possible, but as far as we know, no human-computation platform has implemented multiple privacy-preserving technologies yet, still relying on cheaper, faster, and easier management methods that (might) erode the privacy of clients.

The following section mentions a few possible research directions.

6.2 Research directions

Transparency with rules or SLAs System designers, developers and business actors need to come up with more transparent and direct ways of getting user consent (other than the current standard of publishing privacy policies). An interesting open challenge that we will tackle in our future work is our idea of enforcing user consent through Service Level Agreements (SLAs). Depending on the type of (human) tasks and whether their execution can be monitored and measured (see some metrics in [26], introducing SLAs may come as appropriate as a mechanism to monitor, manage and adapt human computation collectives.

In relation to the aforementioned SLA application, a possible research direction is investigating the inclusion of privacy clauses (e.g., from privacy policies)in SLAs so that users will be obligated to read them and give consent when negotiating SLAs. This could be a two-way negotiation, employers could regulate personal data and content/artifact privacy in relation to the workers as well as the system, and workers could regulate their personal data in relation to employers, other workers and the system.

Privacy preserving workflows Human-based computing in general and crowdsourcing in particular, in addition to issues with personal information, have issues with sensitive artifacts and data submitted for tasks. People who submit tasks may want to reveal only a part of the data. Thus, the design of workflows that provide enough knowledge for workers to be able to execute tasks but don't disclose the full context of requesters' work/interest, is an open (domain-dependent) research question (example work presented in [**?**].

Payment methods When people work in socio-technical systems individually and don't belong to an organization, even with the most efficient anonymization methods, e.g., on the assumption that all worker data is private, the payment methods are still an open question as they can be still used to identify a person.

Location Let us assume a person consents to his/her location data being collected, e.g., in a crowdsourced traffic management of a city. In this case, developers need to pay attention for example to set some location checkpoints,which would not be used to infer sensitive information, such as for example religion (checkpoints near religious buildings), hospitals, houses, and other institutions.

Evaluation methods An interesting research challenge are also evaluation methods for software, evaluating the included privacy-preserving mechanisms.

Raising people awareness about privacy Methods and techniques for raising awareness on privacy should not be a question tackled by experts working on social and legal areas only; it is crucial that computer science researchers approach these challenges as they develop software and disseminate their research.

This section provides only a discussion of possible mechanisms for preserving privacy, and possible research challenges, and it is not an attempt to provide an extensive list of tools, strategies and problems; the goal is to provoke and motivate researchers of human computation systems to tackle privacy challenges.

7 Conclusions

The goal of this research was to get an insight into user privacy-awareness for human computation platforms. As the reported results show, we may conclude that users are moderately concerned for their privacy on these platforms. This is partly because they are willing to show their reputation publicly, and partly because they are not informed enough about how their personal data is collected and processed. Most of our participants stated that they became concerned after they read the statements in our survey concerning privacy-implications in these systems.

The lack of privacy awareness is a key factor why corporations today can leverage the power of personal data collection and analysis, which in the majority of cases is done without peoples' knowledge or consent. Thus, academia, industry, and the civil society needs to focus more on improving awareness. In addition, privacy protection laws and regulations need to be enforced for privacy policies to make sense. Lastly, we recommend system developers and businesses to be guided by principles which respect users' privacy and include privacy-preserving settings by default.

Exploring and building mechanisms that will encourage users to read privacy policies and to prove that they have read them is an important research direction, because awareness and consent are the most important elements in privacy-protection and preservation.

Generalizing the importance of the topic, we advocate that research from every area in computer science should progress with having the context of privacy in mind, as the way we build applications and systems affect the direction in which our societies will further develop. Technology and society are in an interminable process of mutual effect on their change and transformation, in which process the construction of reality takes place; it is up to us to chose and build the type of reality we want to live in, and what better than when progress is accompanied by transparency, trust and consequently, autonomy.

Acknowledgment

We would like to thank all participants of our survey, for providing us with their opinions and enabling us to gain an overview of their privacy concerns.

References

1. Alpár, G., Jacobs, B.: Credential design in attribute-based identity management. In: Bridging distances in technology and regulation, 3rd TILTing Perspectives Conference. (2013) 189–204
2. Anagnostopoulos, A., Becchetti, L., Castillo, C., Gionis, A., Leonardi, S.: Power in unity: forming teams in large-scale community systems. In: CIKM. (2010) 599–608
3. Androulaki, E., Karame, G.O., Roeschlin, M., Scherer, T., Capkun, S.: Evaluating user privacy in bitcoin. In: International Conference on Financial Cryptography and Data Security, Springer (2013) 34–51
4. Balkan, A.: Ethical design manifesto *Ind.ie.* Online available: https://ind.ie/ethical-design/. Last access: 22.07.2017.
5. Buchegger, S., Crowcroft, J., Krishnamurthy, B., Strufe, T.: Decentralized Systems for Privacy Preservation (Dagstuhl Seminar 13062). Dagstuhl Reports **3**(2) (2013) 22–44
6. Christl, W., Spiekermann, S.: Networks of Control. 1 edn. Facultas (2016)
7. Colesky, M., Hoepman, J.h., Hillen, C.: A Critical Analysis of Privacy Design Strategies. In: IWPE, San Jose, CA, IEEE (2016)
8. Directive, E.: 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal of the EC **23**(6) (1995)
9. Dorn, C., Dustdar, S.: Composing near-optimal expert teams: a trade-off between skills and connectivity. On the Move to Meaningful Internet Systems: OTM 2010 (2010) 472–489
10. Dustdar, S., Bhattacharya, K.: The social compute unit. IEEE Internet Computing **15** (May 2011) 64–69
11. Dustdar, S., Truong, H.L.: Virtualizing software and humans for elastic processes in multiple clouds- a service management perspective. IJNGC **3**(2) (2012)
12. Fischer-Hübner, S., Martucci, L.A.: Privacy in Social Collective Intelligence Systems. In: Social Collective Intelligence: Combining the Powers of Humans and Machines to Build a Smarter Society. Springer International Publishing, Cham (2014) 105–124
13. Gong, Y., Guo, Y., Fang, Y.: A privacy-preserving task recommendation framework for mobile crowdsourcing. In: IEEE Global Communications Conference, GLOBECOM 2014, Austin, TX, USA, December 8-12, 2014. (2014) 588–593
14. Khazankin, R., Psaier, H., Schall, D., Dustdar, S.: Qos-based task scheduling in crowdsourcing environments. In: Proceedings of the 9th international conference on Service-Oriented Computing. ICSOC'11, Berlin, Heidelberg, Springer-Verlag (2011) 297–311
15. Kluver, L.: ICT and Privacy in Europe. Experiences from technology assessment of ICT and Privacy in seven different European countries. Final report October 16, 2006, European Parliamentary Technology Assessment network (EPTA), Wien (2006) Online available: http://epub.oeaw.ac.at/?arp=0x0013038d - Last access:26.6.2016.

16. Koning, M., Korenhof, P., Alpár, G.: The abc of abc- an analysis of attribute-based credentials in the light of data protection, privacy and identity (2014)
17. Langheinrich, M. In: Privacy by Design — Principles of Privacy-Aware Ubiquitous Systems. Springer Berlin Heidelberg, Berlin, Heidelberg (2001) 273–291
18. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In Chirkova, R., Dogac, A., Özsu, M.T., Sellis, T.K., eds.: ICDE, IEEE Computer Society (2007) 106–115
19. Malhotra, N.K., Kim, S.S., Agarwal, J.: Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. Info. Sys. Research **15**(4) (December 2004) 336–355
20. Moloney, M., Bannister, F.: A privacy control theory for online environments. In: 42st Hawaii International International Conference on Systems Science (HICSS-42 2009), Proceedings (CD-ROM and online), 5-8 January 2009, Waikoloa, Big Island, HI, USA. (2009) 1–10
21. Motahari, S., Manikopoulos, C., Hiltz, R., Jones, Q.: Seven privacy worries in ubiquitous social computing. In: Proceedings of the 3rd Symposium on Usable Privacy and Security. SOUPS '07, New York, NY, USA, ACM (2007) 171–172
22. Party, D.P.W.: Opinion 4/2007 on the concept of personal data. Brussels, Belgium: European Commission (2007)
23. Quinn, A.J., Bederson, B.B.: Human computation: A survey and taxonomy of a growing field. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '11, New York, NY, USA, ACM (2011) 1403–1412
24. Regulation, E.: Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Official Journal of the EC **119** (2016)
25. Riveni, M., Truong, H.L., Dustdar, S.: On the elasticity of social compute units. In Jarke, M., Mylopoulos, J., Quix, C., Rolland, C., Manolopoulos, Y., Mouratidis, H., Horkoff, J., eds.: Advanced Information Systems Engineering. Volume 8484 of Lecture Notes in Computer Science. Springer International Publishing (2014) 364–378
26. Riveni, M., Truong, H.L., Dustdar, S.: Trust-aware elastic social compute units. In: Trustcom/BigDataSE/ISPA, 2015 IEEE. Volume 1., IEEE (2015) 135–142
27. Scekic, O., Truong, H.L., Dustdar, S.: Incentives and rewarding in social computing. Commun. ACM **56**(6) (June 2013) 72–82
28. Schall, D., Dustdar, S.: Dynamic Context-Sensitive PageRank for Expertise Mining. In: Social Informatics: Second International Conference, SocInfo 2010, Laxenburg, Austria, October 27-29, 2010. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg (2010) 160–175
29. Schall, D., Truong, H.L., Dustdar, S.: Unifying human and software services in web-scale collaborations. IEEE Internet Computing **12**(3) (2008) 62–68
30. Skopik, F., Schall, D., Psaier, H., Dustdar, S.: Adaptive provisioning of human expertise in service-oriented systems. In: Proceedings of the 2011 ACM Symposium on Applied Computing. SAC '11, New York, NY, USA, ACM (2011) 1568–1575
31. Smith, H. Jeff, M.S.J., Burke, S.J.: Information privacy: Measuring individuals' concerns about organizational practices. MIS Q. **20**(2) (June 1996) 167–196
32. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(5) (2002) 557–570
33. To, H., Ghinita, G., Shahabi, C.: A framework for protecting worker location privacy in spatial crowdsourcing. Proc. VLDB Endow. **7**(10) (June 2014) 919–930

34. Toch, E.: Crowdsourcing privacy preferences in context-aware applications. Personal and Ubiquitous Computing **18**(1) (2014) 129–141
35. Truong, H.L., Dustdar, S.: Context-Aware Programming for Hybrid and Diversity-Aware Collective Adaptive Systems. In: Business Process Management Workshops: BPM 2014 International Workshops, Eindhoven, The Netherlands, September 7-8, 2014, Revised Papers. Springer International Publishing, Cham (2015) 145–157
36. Von Ahn, L.: Human Computation. PhD thesis, Pittsburgh, PA, USA (2005) AAI3205378.
37. Williams, T.L., Agarwal, N., Wigand, R.T.: Protecting private information: Current attitudes concerning privacy policies. (2015)
38. Yuen, M.C., King, I., Leung, K.S.: A survey of crowdsourcing systems. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, IEEE (2011) 766–773
39. Zeppezauer, P., Scekic, O., Truong, H.L., Dustdar, S.: Virtualizing Communication for Hybrid and Diversity-Aware Collective Adaptive Systems. In: Service-Oriented Computing - ICSOC 2014 Workshops: WESOA; SeMaPS, RMSOC, KASA, ISC, FOR-MOVES, CCSA and Satellite Events, Paris, France, November 3-6, 2014, Revised Selected Papers. Springer International Publishing, Cham (2015) 56–67