Improving Reliability of Crowdsourced Results by Detecting Crowd Workers with Multiple Identities

Ujwal Gadiraju* and Ricardo Kawase†

* L3S Research Center, Leibniz Universität Hannover,
Appelstr. 9a, Germany
gadiraju@L3S.de
† mobile.de GmbH / eBay Inc.,
Berlin, Germany
rkawase@team.mobile.de

Abstract. Quality control in crowdsourcing marketplaces plays a vital role in ensuring useful outcomes. In this paper, we focus on tackling the issue of crowd workers participating in tasks multiple times using different `worker-ids` to maximize their earnings. Workers attempting to complete the same task repeatedly may not be harmful in cases where the aim of a requester is to gather data or annotations, wherein more contributions from a single worker are fruitful. However, in several cases where the outcomes are subjective, requesters prefer the participation of distinct crowd workers. We show that traditional means to identify unique crowd workers such as `worker-ids` and `ip-addresses` are not sufficient. To overcome this problem, we propose the use of *browser fingerprinting* in order to ascertain the unique identities of crowd workers in paid crowdsourcing microtasks. By using browser fingerprinting across 8 different crowdsourced tasks with varying task difficulty, we found that 6.18% of crowd workers participate in the same task more than once, using different `worker-ids` to avoid detection. Moreover, nearly 95% of such workers in our experiments pass gold-standard questions and are deemed to be *trustworthy*, significantly biasing the results thus produced.

Keywords: Crowdsourcing, Microtasks, Multiple Identities, Quality Control, Reliability

1 Introduction

With the ubiquity of the Internet these days, and the existing need for human intelligence, *crowdsourcing* has empowered millions of people around the globe by providing crowd workers an alternative source to earn their livelihood. A considerable number of real-world applications have showcased the value of this paradigm, ranging from mapping satellite imagery[1] to disaster relief and

[1] `http://www.digitalglobeblog.com/2014/03/10/missingmalayairjet/`

management initiatives[2]. While inumerable examples of profitable crowdsourcing initiatives exist at present, ensuring high quality of results and inhibiting malicious activity are pivotal challenges.

In this work, we aim to tackle a specific kind of potentially malicious activity in paid crowdsourcing marketplaces. Several crowdsourced tasks often require participation from unique crowd workers. This is clearly apparent in surveys and other tasks that require subjective judgments from individuals. For instance, a requester[3] would not want multiple judgments from the same crowd worker in a task that gathers an opinion census of a newly launched product. However, through our experiments presented in this paper, we note that a significant number of workers tend to complete the same task multiple times (by using distinct `worker-ids`) in order to maximize their monetary gains. We define these workers as '*repeaters*'. We reason that crowd workers who exhibit such behavior are primarily driven by monetary incentives. Recent work has shown that over the last 3 years *surveys* are one of the most prominent types of crowdsourced tasks, gaining wide popularity on Amazon's Mechanical Turk[4] (AMT) [1]. Hence, this is an important and timely problem to tackle. By posing as different workers (due to different worker-ids), the same individual can complete a given task any number of times within the task constraints. Workers benefit in the following two ways by doing so, with varying implications.

- *Completion Time.* Familiarity with the task due to repeated participation can result in workers requring much lesser time to complete a given task.
- *Monetary Rewards.* Workers consequently multiply the rewards attained on task completion.

Requesters on the other hand suffer from such repeated participation of workers in a given task in the following ways.

- In the best case of repeated participation, workers can complete tasks in a quick manner resulting in a reduced overall task completion time. If the repeated participation by workers is only motivated by an objective to maximize rewards by completing tasks with genuine effort, this can improve the results [5]. However in the contrasting case, workers with alternative intentions can sabotage a task repeatedly.
- Repeated participation by crowd workers in tasks where distinct workers are expected, implies that requesters bear costs without receiving qualitative returns on their investment. On detection of such activity in a post-processing manner, requesters may need to incur additional cost overheads to gather new judgments.

Existing methods on crowdsourcing platforms rely on user `ip-addresses` to prevent workers from participating in tasks multiple times if so specified by a

[2] `http://www.mission4636.org/`

[3] A *requester* is one who deploys a task on a crowdsourcing platform in order to gather responses from the crowd.

[4] http://www.mturk.com/mturk/

requester. However, crowd workers can change their IP addresses at will, thereby limiting the effectiveness of such methods. We present a novel method for quality assurance in paid microtask crowdsourcing. We propose the adoption of *browser fingerprinting* in order to identify crowd workers that participate in microtasks multiple times with distinct `worker-ids`. The concept of browser fingerprinting has evolved from device fingerprinting over the last decade, emulating the forensic essence of human fingerprints; the ability to uniquely identify different individuals.

The main contributions of this work are two-fold. First, by using browser fingerprints we expose the existence of the crowd workers who repeatedly complete a given task. We show that a substantial share of microtask participants are repeaters. Secondly, we show how current quality control mechanisms that rely on `worker-ids` and `ip-addresses` to restrict repeated participation are insufficient to determine the unique identiy of crowd workers.

2 Related Literature

We discuss related works in two different realms; (i) prior work related to browser fingerprinting and identifying multiple online identities, and (ii) those relevant to ensuring quality in crowdsourcing tasks.

2.1 Browser Fingerprinting & Identifying Multiple Online Identities

Over the last decade there have been advances in the reliable detection of unqiue web browsers. Eckersley investigated the version and configuration information that web browsers transmit upon request, in order to study the extent to which browsers are subject to device fingerprinting [2]. The variables considered in the hashing of browser fingerprints as prescribed by Eckersley included the following: the user agent string transmitted by HTTP, the HTTP accept headers, whether or not cookies are enabled, screen resolution, timezone, browser plugins, plugin versions and MIME types, system fonts and a partial supercookie test. The author showed that the distribution of browser fingerprints of users in their collection contained at least 18.1 bits of entropy, effectively meaning that in their experimental collection, only one in 286,777 other browsers shared its fingerprint.

Mowery and Shacham proposed the rendering of text and WebGL scenes to a `<canvas>` element, and thereby examining the pixels produced in order to tie a browser more closely to a user's operating system and hardware [14]. Mulazzani et al. proposed an efficient method to identify browsers by JavaScript engine fingerprinting [15]. In this paper, our proposal to use browser fingerprinting to detect the undesirable repeated participation of workers in crowdsourcing tasks, is inspired by these prior works. The contribution of our work in this context is the evidence we provide through rigorous experimentation, indicating the effectiveness of browser fingerprinting in improving the reliability of crowdsourcing results.

Prior works have also addressed the problem of identifying multiple identities in various online contexts. Gani et al. proposed a framework to detect multiple identities in social networks based on machine learning models and interaction between users [8]. Kafai et al. showed how online gamers use multiple accounts and identities in order to make more money or maximize rewards [10]. More recently, Yamak et al. proposed supervised machine learning algorithms to detect multiple identities of users in collaborative projects online [21]. In contrast to such previous works, in this paper we address the novel context of identifying *repeaters* in crowdsourcing microtasks.

2.2 Quality Assurance in Crowdsourcing

Several prior works have focused on methods from varying perspectives to improve the quality of crowdsourced work. Kittur et al. reflected on the measures required to ensure reliability in crowdsourced user measurements from the task design point of view [12]. Authors have also studied the effect of task pricing on the quality of results produced. Faradani et al. [4] focused on the duality between task completion time and pricing, and model quality as a tradeoff between these aspects. Wang et al. proposed a method to measure worker quality, based on which they determined the fair payment level for a worker [20]. Mason et al. showed that increasing monetary incentives of crowdsourced tasks attracts more workers but does not improve the quality of the results produced [13]. Oleson et al. proposed the usage of *gold-standard* questions to ensure reliability of responses and improve the quality of crowd work [16]. Eickhoff et al. proposed guidelines to inhibit spammers in crowdsourced tasks [3]. However, as shown by Gadiraju et al., the use of gold-standards alone are insufficient to curtail malicious activity in the crowd. The crowd consists of a significant number of *smart deceivers*, who take special precautions to avoid detection [7]. The authors studied the implications of task design as well as worker behavior on the quality of crowdsourced results. Other works have also investigated the motivation behind participation in crowdsourcing microtasks and the impact of motivation on performance of workers [18, 11]. In contrast to these prior works, we study the problem of workers repeatedly participating in tasks using distinct `worker-ids` to maximize their earning and avoid the scrutiny of quality control mechanisms. This problem has not been explicitly addressed and studied in prior works. We propose to detect such crowd workers who repeatedly participate in tasks (thereby called *repeaters*), by using browser fingerprinting of workers.

Rzeszotarski and Kittur proposed *task fingerprinting*; collecting user activity logs through mousetracking in crowdsourcing tasks, in order to infer worker cognition and effectiveness [19]. While task fingerprints are extremely useful to model crowd workers and understand their cognitive processes, they cannot reliably be used to identify unique workers, since multiple workers can depict identical behavior in a given task. In a closely related application of browser fingerprinting, Rainer and Timmerer used browser fingerprinting in order to ensure unique participation in their experiments regarding QoE in multimedia streaming over HTTP [17]. However, the authors employ browser fingerprinting without

measuring the actual effectiveness of the method. In this paper, we investigate the applicability of browser fingerprinting as a quality control mechanism in crowdsourcing microtasks.

3 Preliminary Validation Study – Multiple Accounts Usage by Crowd Workers

To first determine the legitimacy of multiple accounts usage by crowd workers on crowdsourcing platforms, we surveyed workers on CrowdFlower[5], a premier crowdsourcing platform.

3.1 Survey Design

CrowdFlower allows task requesters to restrict the participation of workers based on their reputation (in terms of *levels*, where `level-3` workers have the best reputation, followed by `level-2` and `level-1`). Thus, we considered the three different levels of worker participation by deploying identical surveys corresponding to each restriction. We collected responses from 100 crowd workers in each case, and rewarded them with 5 USD cents for responding to 5 questions in the survey. The workers were urged to respond honestly in the instructions, and the objective of the survey (i.e., to understand the usage of multiple accounts by workers) was conveyed accurately. The workers were first asked whether they used multiple accounts with different `worker-id`s to access more work. Then the workers were asked to select the types of tasks in which they typically used multiple accounts (if at all) among the following; *content access, content creation, information finding, interpretation and analysis, surveys, verification and validation* [6]. Workers were also asked the frequency with which they used multiple accounts with different `worker-id`s on a 5 point Likert-scale ranging from *1: Never* to *5: Always*. Finally, we asked workers how many active multiple accounts they used among the following options; $(1, 2, 3, 4, 5, > 5)$.

To verify the reliability of responses from the workers, we embedded a test question within the 5 questions in the survey. Here the workers were asked explicitly to enter the word 'COOL' in a corresponding text box. Those workers who failed to do so were deemed to be unreliable and we do not consider their responses in our analysis. We found 2 workers in each of `level-1`, `level-2`, and 1 worker in `level-3` who failed the test question.

3.2 Results

We found that 12.25% of crowd workers in `level-1`, 7.14% of the workers in `level-2` and 11.11% of workers in `level-3` claimed to use multiple accounts with different `worker-id`s to access more work.

[5] `http://www.crowdflower.com/`

Figure 1 presents the frequency with which workers in each level use multiple accounts with different `worker-id`s. We note that although the vast majority of workers in each level claim to never use multiple `worker-id`s, few workers do indicate a moderate to high usage of such multiple accounts for participation.

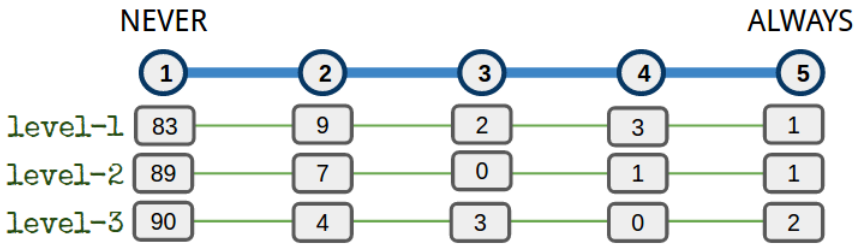


Fig. 1: Frequency (on a Likert-scale) with which workers corresponding to each of the 3 levels on CrowdFlower use different `worker-id`s to participate in tasks.

We found that some workers in each of the three levels actively used more than one `worker-id` to participate in tasks, as shown in Table 1a. Those workers who use multiple accounts did not depict a signficant affinity towards a particular type of task for such repeated participation, as shown in Table 1b. This suggests that task type is not necessarily an important feature that facilitates or drives the repeated participation of workers.

Table 1: (a) Number of `worker-id`s actively used by workers, and (b) frequency of the usage of multiple `worker-id`s across different types of tasks corresponding to each of three levels on CrowdFlower.

(a)

# worker-ids	level-1	level-2	level-3
1	76	80	82
2	4	6	6
3	6	5	4
4	5	2	4
5	3	2	0
6 or more	4	3	3

(b)

Task Type	level-1	level-2	level-3
Content Access	4	4	1
Content Creation	3	3	3
Information Finding	8	7	3
Interpretation & Analysis	2	2	3
Surveys	9	8	5
Verification & Validation	5	7	6

Through surveying 300 crowd workers on CrowdFlower, we found evidence of the usage of multiple `worker-id`s by workers in order to access and complete more work, thereby maximizing their monetary rewards. As motivated earlier, this may however be an undesirable aspect depending on the task at hand.

4 Objectives and Methodology

By addressing the following research questions in this work, we propose the application of browser fingerprinting for improving quality assurance in crowdsourcing practice.

- **RQ1:** What proportion of crowd workers participating in tasks tend to be truly distinct workers?
- **RQ2:** How does *task difficulty* effect workers who feign their identity to complete tasks multiple times?
- **RQ3:** Do crowd workers who ineligibly repeat tasks have a significant impact on the results produced?

4.1 Experimental Setup and Task Design

Based on the responses of workers in the preliminary study, we did not find a correlation between the task type and the tendency of workers to use multiple `worker-id`s. With an aim to investigate the research questions stated earlier, we consider the task of logical reasoning. We first gather basic background information from the crowd workers through demographic questions. These are followed by 15 questions in the domain of *logical reasoning*. We used logical reasoning questions from $A + Click$[6]. The logical reasoning questions were based on the Common Core Standards[7], which is a set of academic standards in mathematics and English. These learning objectives indicate what a student should know and be able to do at the end of each grade. We chose this setup since the progressing grade-level is a clear indicator of increasing difficulty in the logical reasoning questions. Such a setup would enable us to explore the impact of *task difficulty* on the repeated participation of workers.

In order to assess the impact of *task difficulty* among crowd workers, we deployed 8 microtasks that are designed similarly except for the difficulty level of the logical reasoning questions. Herein, we used graded questions from $A+Click$ to procure logical reasoning questions from the level of Grade 5 to Grade 12. An example question is presented in Figure 2. We did not consider lower grades than the 5^{th}, since initial experiments revealed that workers tend to perform with a 100% accuracy in those grades. In order to separate *trustworthy* workers (TW)[8] from *untrustworthy* workers (UW)[9], we intersperse attention-check questions (example shown in Figure 3) as recommended by Gadiraju et al. [7].

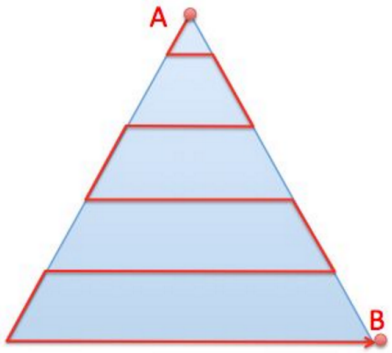
Prior research has shown that having verifiable questions such as tags is a recommended way to design tasks and assess crowdsourced results [12]. The last two questions in the task asked crowd workers to provide as many tags as possible for two different pictures. Note that the order in which different questions were

[6] `http://www.aplusclick.com/`

[7] `http://www.corestandards.org/`

[8] Workers who correctly answer all 3 attention check questions embedded in the task.

[9] Workers who incorrectly answer at least 1 of the 3 attention check questions embedded in the task.



A bug goes from point A to point B along the indicated path. How many times does the insect turn left?

- 9
- 4
- 10
- 5

Fig. 2: An example logical reasoning question from $A + Click$ that was administered to crowd workers in the task corresponding to Grade 5.

This is an attention check question. Please select the third option.

Apple Ball Cat Dog

Fig. 3: Attention-check questions to identify *untrustworthy* workers.

asked did not have an impact on the results reported in our work. Thus, we do not focus on this further. We paid the crowd workers according to a fixed hourly wage of 7.5 USD, for completing the tasks successfully. Corresponding to each of the 8 graded tasks that we deployed on CrowdFlower, we gathered 250 responses from independent crowd workers, resulting in a total of 2000 workers overall. We did not restrict the participation of workers based on the CrowdFlower *levels*.

Finally, we extracted the browser fingerprints of crowd workers through a Javascript implementation[10]. As shown by Peter Eckersley [2], browser fingerprinting can anonymously identify a web browser with an accuracy of over 95%.

5 Results and Discussion

5.1 Can We Trust the Trustworthy Workers?

Table 2 presents the number of trustworthy workers (TW) determined by using the attention check questions across the different grades. We can clearly see that the percentage of TW is quite high. However, as we have motivated earlier

[10] `http://valve.github.io/fingerprintjs/`

in this paper, a repeater represents a breach in trust. If a particular task is designed to collect a limited number of responses from an individual worker, then each worker is eligible and expected to provide only those limited number of responses. Not respecting such clearly prescribed limits would amount to a 'violation of trust'.

Table 2: Percentage of trustworthy workers (TW) out of 250 participating workers, across the different graded microtasks.

Grade	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	**G11**	**G12**
#TW	91.2%	86.4%	90.4%	82.8%	82.8%	86%	85.6%	87.6%

Figure 4 presents the distribution between the number of distinct `worker-ids` corresponding to the distinct fingerprints in all the tasks. We observe a power-law distribution with one fingerprint corresponding to 11 different `worker-ids`, and this gradually decreases to the majority of fingerprints corresponding to distinct `worker-ids`. It is clear that repeaters used distinct `worker-ids` in order to avoid detection by potential quality control mechanisms.

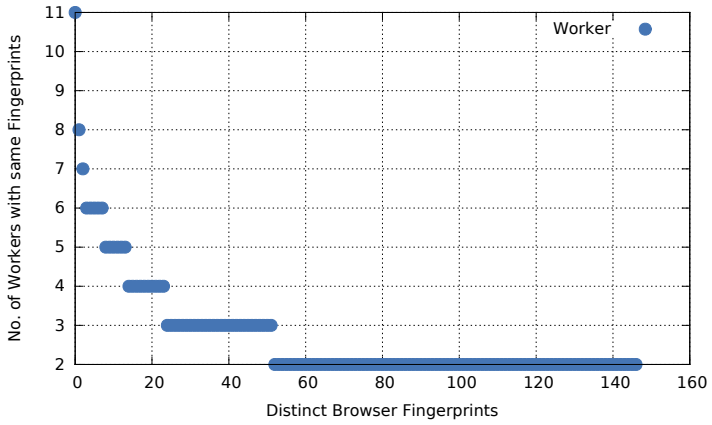


Fig. 4: Distribution of number of distinct `worker-ids` corresponding to each fingerprint across all grades (G5 through G12).

Having said that, Table 3a depicts the percentage of repeaters in each grade that are trustworthy (TW). We note that in each grade there are several repeaters, i.e., workers having different `worker-ids` with the same browser fingerprints. At first, this behavior might seem irrelevant; the average number of repeaters across all tasks is 6.18%. However, in some cases, these repeaters provide a substantial number of contributions (as also observed in Figure 4). For

example, nearly 18% of the total responses (contributions) collected in G6 correspond to answers from repeaters. We also observe that the fraction of repeaters in each grade are similar to our findings in the preliminary study.

Table 3: (a) Percentage of trustworthy (TW) repeaters, and share of their contributions in each task, and (b) performance of all workers, repeaters alone and non-repeaters alone in each task.

(a)

Task	% TW Repeaters	% Contributions
G5	4.13	6.82
G6	8.78	17.48
G7	7.38	10.91
G8	5.35	13.43
G9	6.72	13.64
G10	5.86	10.29
G11	5.46	12.87
G12	3.71	7.58
All	6.18	11.63

(b)

Task	All Workers	Repeaters	Non-Repeaters
G5	77.61	75.71	77.88
G6	60.45	57.40	61.68
G7	57.36	60.30	56.61
G8	40.00	43.17	39.16
G9	39.93	35.60	41.28
G10	45.59	46.83	45.28
G11	32.97	33.18	32.91
G12	28.72	24.00	29.50

Previous work by Gadiraju et al. proposed a classification of the most common type of untrustworthy workers [7]. The authors did not study *repeaters* as a particular case of untrustworthy workers. Under their proposed classification, repeaters would belong to the category of *Ineligible Workers* due to violating the pre-requisite that a worker is not eligible to perform a task more than a certain number of times. However, repeaters circumvent their ineligibility by using multiple `worker-ids` and do not demonstrate further untrustworthy characteristics. In fact, in terms of performance, we see little (non-significant) variations (see Table 3b). These results show that, despite the fact that *repeaters* are ineligible workers, they perform tasks with the diligence of an average trustworthy worker. This makes repeaters undetectable unless techniques such as fingerprints are employed.

Although no significant differences were found in terms of performance of the workers, the impact of repeaters becomes clear when one considers the demographics questions. Our demographics questions which included multiple choices for the age group (5 options), education (9 options), ethnicity (7 options) and gender (2 options), depict a significant change in all cases in the presence and absence of repeaters (with $p < 0.05$) in the distributions of at least one of the options provided. Based on these results, and considering that surveys are one of the most common types of crowdsourced tasks [1], we reflect on the susceptibility of surveys to the participation of repeaters, resulting in the generation of skewed and biased outcomes that can go unnoticed.

Finally, we found a moderately strong negative correlation between the difficulty-level of a task (an inherent function of progressive grades from G5 through G12) and the number of trustworthy `worker-ids` corresponding to re-

peaters (Pearson's $\mathbf{r} = -0.3$). This suggests that the more difficult that a task is, the less often trustworthy workers tend to repeat it using different `worker-ids`. Thus, easier tasks that require less effort, or those which provide a better cost-benefit ratio, are more prone to attract repeaters.

5.2 The Case of Account Sharing Among Crowd Workers

Across the different tasks we found 21 cases where multiple (2 or more) browser fingerprints were associated with the same `worker-id` within the same task. Out of these 21 anomalies, in 7 cases the different fingerprints were associated with the same `ip-address`, suggesting that the workers switched or altered some browser configuration. In the other 14 cases, the different `ip-addresses` and the corresponding different fingerprints suggest that multiple workers have access to the same user account, and thereby correspond to the same `worker-id`. This can be attributed to scenarios where the users have different sessions with the same login through virtual machines (which is less likely), or it is a shared account where different persons work together using different devices. Although this is a breach in the quality control mechanisms for crowdsourced tasks, in this paper we focus on the more frequent case of workers participating repeatedly by using different `worker-ids` rather than multiple workers using the same `worker-id`.

5.3 Pruning Workers Using IP Addresses

We investigate the number of *repeaters* that can be detected relying solely on the worker `ip-address`. We detect one repeater in the task corresponding to G7 and another in G10. Thus, we note that using a worker's `ip-address` alone as a means to identify unique crowd workers is not sufficient.

5.4 The Privacy Perspective

Although the experiments in this work have been carried out after establishing user-consent, covertly tracking users as a means of their browser fingerprints can be considered to be an unsolicited intrusion of their privacy.

Having said that, we argue in favor of using *browser fingerprinting* to detect *repeaters* who attempt to maximize their monetary benefits by completing tasks mutliple times. Through our 8 crowdsourced tasks involving 2000 workers, we observed a significant participation of repeaters (6.18%). These repeaters account for over 13% of the total contributions by virtue of their repeated participation. Repeaters skew the purpose of requesters, especially in subjective types of tasks such as surveys. Due to the fact that these fingerprints are merely required to uniquely identify workers within a task, the user data used to generate browser fingerprints can be consequently discarded on task completion. Moreover, by using hashing functions to generate browser fingerprints, one does not need to store the underlying data representing browser characteristics such as agent strings, headers, plugin details, system fonts, cookie settings, and so forth.

Due to these reasons, *browser fingerprinting* is a viable and effective method to prevent workers from violating task requirements in crowdsourced microtasks, thereby improving the quality of the results produced. Moreover, by relying solely on the hashed fingerprint, we can alleviate privacy concerns.

5.5 Caveats and Limitations

We acknowledge that the existing browser fingerprinting techniques are around 95% accurate. This means that there is room for a small percentage of errors. However, the elaborate hashing of various attributes that are considered for browser fingerprinting means that it is highly unlikely that two fingerprints will accidentally collide to be identical. Yet, a conservative approach can be the use of browser fingerprinting as a means to flag crowd workers for further scrutiny, rather than blocking potential *repeaters* immediately. Such an approach would also resonate with prior work that has called for less-aggresive means of dealing with sub-optimal or potentially malicious work.

Another limitation of this work stems from our inability to account for genuine explanations of repeated participation, as detected using browser fingerprinting. For example, a false positive could result from crowd workers working in Internet cafes [9], or family members sharing a computer.

6 Repeaters in Real-World Crowdsourcing Microtasks

We conducted an additional study to evaluate the occurrence of repeaters in real-world microtasks. We manually created a batch of 120 microtasks comprising of an equal distribution of all the different types. Table 4 presents some example tasks corresponding to each task type that were deployed. The different types were prescribed by a taxonomy proposed in previous work [6]. We deployed these tasks on CrowdFlower and collected 100 responses from distinct workers for each task, resulting in 12,000 human intelligence tasks (HITs). Once again we did not restrict participation of workers based on CrowdFlower *levels*. Table 4 also presents sample tasks that we created corresponding to each type; these tasks are noticeably designed to reflect real-world microtasks that have previously been deployed on crowdsourcing platforms such as Amazon's Mechanical Turk.

6.1 Results – Distribution of Unique Fingerprints and `worker-ids`

As in case of the previous study, we used a JavaScript implementation to generate browser fingerprints corresponding to each worker participating in the tasks. We analyzed the browser fingerprints and found a power law distribution between the number of distinct `worker-ids` corresponding to each fingerprint across all 120 tasks, as shown in Figure 5.

Once again we found that a significant portion of distinct browser fingerprints corresponded to more than one `worker-id` associated to the participating workers; nearly 18.5% of fingerprints corresponded to 2 or more `worker-id`s.

Table 4: Examples of different real-world microtasks that were deployed.

Task Type	Sample Tasks Deployed
Content Access	Watch the following video.
Content Creation	Transcribe the audio excerpt presented above.
Information Finding	Find the middle-names of the following famous persons by searching on the Web.
Survey	What is your age?
Interpretation & Analysis	Which of the following tweets has a neutral sentiment? Check all that apply.
Verification & Validation	Choose the words which are synonyms of 'HAPPY' in the following list.

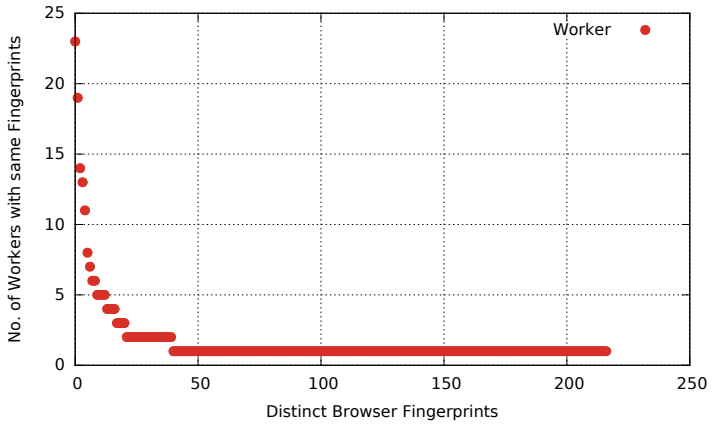


Fig. 5: Distribution of number of distinct `worker-ids` corresponding to each fingerprint across all 120 real-world microtasks of different types.

This shows that the usage of multiple accounts with different `worker-id`s can be observed in real-world microtasks.

6.2 Evaluation of Repeaters

Due to the lack of a given groundtruth with respect to whether or not the multiple `worker-ids` associated with a unique fingerprint are a result of workers using multiple accounts, we interviewed a random selection of such workers for the purpose of evaluating the accuracy in identification of repeaters. We randomly selected 10 workers from the pool of 193 workers who corresponded to sharing the same fingerprint with at least one more worker in the pool. We contacted these workers via e-mail and recruited them to a follow-up 15 minute Skype interview in return for 3 USD each. We promised to maintain the anonymity of workers and clarified the purpose of the interview beforehand. We carried out these interviews over 2 weeks following the completion of the tasks.

Workers were first asked about whether or not they participated in the tasks that they completed within this study. All 10 workers confirmed that they com-

pleted those tasks successfully. We then asked workers regarding the usage of multiple accounts to complete more tasks. 9/10 workers admitted to using multiple accounts to complete more work, and maximize their monetary rewards. Nearly all workers defended their actions since they claimed to have completed the tasks diligently each time they repeated it using a different *worker-id*. Through the interviews, it was apparent that workers were not aware of the unintentional consequences in skewing the reliability of results through their repeated participation. However, our findings suggest the high reliability of using browser fingerprinting to identify repeaters.

7 Conclusions and Future Work

In this paper we have showed that there are a significant number of repeaters that participate in crowdsourced tasks using distinct `worker-ids`. In the light of repeaters in crowdsourced microtasks, we present the following contributions and draw conclusions.

- Across 8 crowdsourced logical reasoning tasks with varying task difficulty and spanning 2000 workers, we have observed that over 13% of the workers are not distinct, but are a result of repeated participation from 6.18% of workers using different `worker-ids` (**RQ#1**). We found consistent results in further experiments using real-world microtasks.
- We found that there is a moderately high negative correlation between the task difficulty and the number of trustworthy (TW) repeaters. This means that with an increasing task difficulty the number of TW repeaters decreases (**RQ#2**). Thus, task requesters should be more prudent while deploying tasks that are relatively easy to complete; simple tasks have a greater propensity for repeated participation.
- Existing quality control mechanisms that rely on worker `ip-addresses` or `worker-ids` fail to detect repeaters. We have shown that browser fingerprinting can be used in order to identify repeaters in crowdsourced microtasks. Through our experimental tasks, we have found that repeaters significantly skew the demographic attributes within a given task, and thereby adversely affect the reliability of the results produced (**RQ#3**).

Our findings have important implications in crowdsourced tasks, especially when the tasks are subjective. It is vital to detect and prevent repeated participation of workers in a task in order to ensure reliable and unbiased results in crowdsourced microtasks. In the imminent future, we plan to investigate the usage of browser fingerprinting in tandem with other quality control mechanisms.

Acknowledgments

This research has been supported in part by the European Commission within the H2020-ICT-2015 Programme (AFEL project, Grant Agreement No. 687916).

References

1. D. E. Difallah, M. Catasta, G. Demartini, P. G. Ipeirotis, and P. Cudré-Mauroux. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *Proceedings of the 24th International Conference on World Wide Web*, pages 238–247. International World Wide Web Conferences Steering Committee, 2015.
2. P. Eckersley. How unique is your web browser? In *Privacy Enhancing Technologies, 10th International Symposium, PETS 2010, Berlin, Germany, July 21-23, 2010. Proceedings*, pages 1–18, 2010.
3. C. Eickhoff and A. P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.
4. S. Faradani, B. Hartmann, and P. G. Ipeirotis. What's the right price? pricing tasks for finishing on time. In *Human Computation, Papers from the 2011 AAAI Workshop, San Francisco, California, USA, August 8*, 2011.
5. U. Gadiraju and S. Dietze. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, March 13-17, 2017*, pages 105–114, 2017.
6. U. Gadiraju, R. Kawase, and S. Dietze. A taxonomy of microtasks on the web. In *25th ACM Conference on Hypertext and Social Media, HT '14, Santiago, Chile, September 1-4, 2014*, pages 218–223, 2014.
7. U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 1631–1640, 2015.
8. K. Gani, H. Hacid, and R. Skraba. Towards multiple identity detection in social networks. In *Proceedings of the 21st International Conference on World Wide Web*, pages 503–504. ACM, 2012.
9. M. Gawade, R. Vaish, M. N. Waihumbu, and J. Davis. Exploring employment opportunities through microtasks via cybercafes. In *2012 IEEE Global Humanitarian Technology Conference, GHTC 2012, Seattle, WA, USA, October 21-24, 2012*, pages 77–82, 2012.
10. Y. B. Kafai, D. A. Fields, and M. Cook. Your second selves: avatar designs and identity play in a teen virtual world. In *Proceedings of DIGRA*, volume 2007, 2007.
11. N. Kaufmann, T. Schulze, and D. Veit. More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk. In *AMCIS*, volume 11, pages 1–11, 2011.
12. A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.
13. W. A. Mason and D. J. Watts. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation, Paris, France, June 28, 2009*, pages 77–85, 2009.
14. K. Mowery and H. Shacham. Pixel perfect: Fingerprinting canvas in html5. *Proceedings of W2SP*, 2012.
15. M. Mulazzani, P. Reschl, M. Huber, M. Leithner, S. Schrittwieser, E. Weippl, and F. Wien. Fast and reliable browser identification with javascript engine fingerprinting. In *Web 2.0 Workshop on Security and Privacy (W2SP)*, volume 5, 2013.
16. D. Oleson, A. Sorokin, G. P. Laughlin, V. Hester, J. Le, and L. Biewald. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Human*

Computation, Papers from the 2011 AAAI Workshop, San Francisco, California, USA, August 8,, 2011.

17. B. Rainer and C. Timmerer. Quality of experience of web-based adaptive http streaming clients in real-world environments using crowdsourcing. In *Proceedings of the 2014 Workshop on Design, Quality and Deployment of Adaptive Video Streaming*, pages 19–24. ACM, 2014.
18. J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.
19. J. M. Rzeszotarski and A. Kittur. Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16-19, 2011*, pages 13–22, 2011.
20. J. Wang, P. G. Ipeirotis, and F. Provost. Quality-based pricing for crowdsourced workers. 2013.
21. Z. Yamak, J. Saunier, and L. Vercouter. Detection of multiple identity manipulation in collaborative projects. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 955–960. International World Wide Web Conferences Steering Committee, 2016.