Proceedings Paper:
Gadiraju, U., Kawase, R., Dietze, S. et al. (1 more author) (2015) Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In: Begole, B., Kim, J., Inkpen, K. and Woo, W., (eds.) CHI '15 Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 33rd Annual ACM Conference on Human Factors in Computing Systems, 18-23 Apr 2015, Seoul, Korea. ACM , pp. 1631-1640. ISBN 978-1-4503-3145-6

https://doi.org/10.1145/2702123.2702443

Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys

Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze
L3S Research Center
Leibniz Universität Hannover, Germany
{gadiraju, kawase, dietze}@L3S.de

Gianluca Demartini
Information School,
University of Sheffield, United Kingdom
g.demartini@sheffield.ac.uk

ABSTRACT

Crowdsourcing is increasingly being used as a means to tackle problems requiring human intelligence. With the ever-growing worker base that aims to complete microtasks on crowdsourcing platforms in exchange for financial gains, there is a need for stringent mechanisms to prevent exploitation of deployed tasks. Quality control mechanisms need to accommodate a diverse pool of workers, exhibiting a wide range of behavior. A pivotal step towards fraud-proof task design is understanding the behavioral patterns of microtask workers. In this paper, we analyze the prevalent malicious activity on crowdsourcing platforms and study the behavior exhibited by trustworthy and untrustworthy workers, particularly on crowdsourced surveys. Based on our analysis of the typical malicious activity, we define and identify different types of workers in the crowd, propose a method to measure malicious activity, and finally present guidelines for the efficient design of crowdsourced surveys.

Author Keywords

Crowdsourcing; Microtasks; Online Surveys; User Behavior; Malicious Intent

ACM Classification Keywords

H.1.2 User/ Machine Systems: Human factors

INTRODUCTION

Over the last decade, crowdsourcing has gained rapid popularity, because of the data-intensive nature of emerging tasks, requiring validation, evaluation and annotation of large volumes of data. While developing a sound definition of crowdsourcing, Estelles and Guevara [1] suggest that microtasks are of variable complexity and modularity, and entail mutual benefit to the *worker*[1] and the *requester*[2]. Accumulating small contributions through such microtasks facilitates the accomplishment of work that is not easily automatable, through rather minor contributions of each individual worker.

[1]A user that performs tasks in exchange of monetary rewards on a crowdsourcing platform.

[2]A user that deploys tasks to be completed on a crowdsourcing platform, also called a *task administrator*.

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea.
Copyright 2015 ACM 978-1-4503-3145-6/15/04$15.00
http://dx.doi.org/10.1145/2702123.2702443

With the ubiquity of the internet, it became possible to distribute tasks at global scales, leading to the recent success of crowdsourcing, being later defined as an 'online, distributed problem-solving and production model [2].'

In the recent past, there has been a considerable amount of work towards developing appropriate platforms and suggesting frameworks for efficient crowdsourcing. Amazon's Mechanical Turk[3], and CrowdFlower[4] are good examples of such platforms. An increasing number of research communities benefit from using crowdsourcing platforms in order to either gather distributed and unbiased data [3], to validate results, evaluate aspects, or to build ground truths [4].

While the demand for using crowdsourcing to solve several problems is on an upward climb, there are some obstacles that hinder requesters from attaining reliable, transparent, and non-skewed results. Herein, a primary nuisance is introduced through *malicious workers*, understood by [5, 6, 7] as workers with ulterior motives, who either simply sabotage a task or try to quickly attain task completion for monetary gains.

Gold-standards are the typically adopted solution to improve task performance [8]. In general practice, gold-standards are questions where answers are known apriori to the task administrators. Thus, if a worker fails to provide the correct answer for a particular question, he is automatically flagged as an *untrustworthy worker*[5]. However, with the flourishing crowdsourcing market, we believe that malicious activities and adversarial approaches will also become more advanced and popular, overcoming common gold standards. Quality control mechanisms should thereby account for a diverse pool of workers that exhibit a wide range of behavioral patterns. Methods have been designed and used in order to tackle poor worker performance in the past [9, 10]. However, there is a need to understand the behavior of these workers and the kinds of malicious activity they bring about in crowdsourcing platforms. In this paper, we present our work towards analyzing the behavior of malicious microtask workers, and reflect on guidelines to overcome such workers in the context of online surveys. An *online survey* is a questionnaire that can be completed over the Internet by a target audience.

We deployed a survey to 1000 workers in the crowd, and present evidence that a large number of workers are untrustworthy. This evidence shows that simple gold-standards

[3]https://www.mturk.com/mturk/

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

[5]Note that being an untrustworthy worker does not necessarily imply being a malicious worker.

might not be enough to provide reliable data or results. Then we conducted an analysis of both trustworthy and untrustworthy workers; we classified the behavior of the workers based on the different types of activity exhibited. To gain further insights into the prevalence of different kinds of malicious workers in the crowd, experts manually and exhaustively annotated the workers into established classes.

The main contributions of our work are listed below.

- Resulting from our analysis of workers, we present different types of malicious behavior exhibited in the crowd. This understanding of the prevalent kinds of malicious activity will be an aid in future task design.
- We suggest a novel method to measure the *maliciousness* of a worker based on the acceptability of her responses.
- We present a detailed analysis of the flow of malicious behavior of workers throughout the task, and define a *tipping point* which marks the starting point of a workers' malicious tendency.
- Finally, we propose a set of guidelines for the efficient, fraud-proof task design of *surveys*.

RELATED LITERATURE

Quality and Reliability of Workers

Behrend et al. showed the suitability of crowdsourcing as an alternative data source for organizational psychology research [11]. Kittur et al. promoted the suitability of crowdsourcing user studies, while cautioning that special attention should be given to the task formulation [12]. Although these works outline shortcomings of using crowdsourcing, they do not consider the impact of malicious activity that can emerge in differing ways. In our work, we show that varying types of malicious activity is prevalent in crowdsourced surveys, and propose measures to curtail such behavior.

Marshall et al. profiled Turkers who take surveys, and examined the characteristics of surveys that may determine the data reliability [13]. Similar to their work, we adopt the approach of collecting data through crowdsourced surveys in order to draw meaningful insights. Our analysis quantitatively and qualitatively extends their work, and additionally provides a sustainable classification of malicious workers that sets precedents for an extension to different categories of microtasks.

Through their work, Ipeirotis et al. motivated the need for techniques that can accurately estimate the quality of workers, allowing for the rejection or blocking of low-performing workers and spammers [5]. The authors presented algorithms that improve the existing techniques to enable the separation of bias and error rate of the worker. Baba et al. reported on their study of methods to automatically detect improper tasks on crowdsourcing platforms [14]. The authors reflected on the importance of controlling the quality of tasks in crowdsourcing marketplaces. Complementing these existing works, our work propels the consideration of both aspects (task design as well as worker behavior), for effective crowdsourcing.

Dow et al. presented a feedback system for improving the quality of work in the crowd [15]. Oleson et al. present a method to achieve quality control for crowdsourcing, by providing training feedback to workers while relying on programmatic creation of gold data [8]. However, for gold-based quality assurance, task administrators need to understand the behavior of malicious workers and anticipate the likely types of worker errors with respect to different types of tasks. Understanding the behavior of workers, is therefore an important objective of this paper.

In the realm of studying the reliability and performance of crowd workers with respect to the incentives offered, Mason et al. investigated the relationship between financial incentives and the performance of the workers [16]. They found that higher monetary incentives increase the quantity of workers but not the quality of work. A large part of their results align with our findings presented in the following sections.

Worker Traits, Tasks Design and Metrics

Researchers in the field have acknowledged the importance and need for techniques to deal with inattentive workers, scammers, incompetent and malicious workers.

Ross et al. studied the demographics and usage behaviors characterizing workers on Amazon's Mechanical Turk [17]. Kazai et al. defined types of workers in the crowd by type-casting workers as either *sloppy*, *spammer*, *incompetent*, *competent*, or *diligent* [18]. By doing so, the authors expect their insights to help in designing tasks and attracting the best workers to a task. While the authors use worker-performance in order to define these types, we delve into the behavioral patterns of workers.

Wang et al. presented a detailed study of *crowdturfing systems*, which are dedicated to organizing workers to perform malicious tasks [19]. While the authors of this paper investigated systems solely dedicated to malicious activities, in our work, we explore and analyze the prevalence of malicious workers and activities on regular crowdsourcing platforms. In their work, Eickhoff et al. aimed to identify measures that one can take in order to make crowdsourced tasks resilient to fraudulent attempts [6]. The authors concluded that understanding worker behavior better is pivotal for reliability metrics. Understanding malicious workers, is in fact the main goal of this paper.

Difallah et al. reviewed existing techniques used to detect malicious workers and spammers and described the limitations of these techniques [9]. Buchholz and Latorre proposed metrics for the post-hoc exclusion of workers from results [20]. In another relevant work by Eickhoff et al., the authors proposed to design and formulate microtasks such that they are less attractive for cheaters [21]. In order to do so, the authors evaluated factors such as the type of microtask, the interface used, the composition of the crowd, and the size of the microtask. While our work presented in this paper complements the prior work done by Eickhoff et al. it is significantly different, in that we investigate the behavioral patterns of trustworthy and untrustworthy workers, and suggest remedies to detect and inhibit their prominence based on the specific type of behavior. Notably, we introduce novel metrics such as *maliciousness* of a worker, to quantify the behavioral patterns thus observed.

Yuen et al. present a literature survey on different aspects of crowdsourcing [22]. In addition to a taxonomy of crowdsourcing research, the authors present a humble example list of application scenarios. Their short list represents the first steps towards task modeling. However, without proper organization regarding types, goals and work-flows, it is hard to reuse such information to devise strategies for task design. As a step forward, in earlier work Gadiraju et al. proposed a comprehensive and exhaustive taxonomy for the different types of microtasks [23]. By studying the various kinds of behavior exhibited by trustworthy and untrustworthy workers in the crowd, in this work, we present a closer and detailed understanding of workers that will aid in developing anti-adversarial techniques.

BACKGROUND

We build on the previous work done by Gadiraju et al. [23], where the authors analyzed the nature of crowdsourced tasks. Firstly, the rationale behind the choice of workers to complete a job and the nature of the jobs themselves were studied. *Monetary reward* was found to be the most crucial factor that motivates workers across different task types, in their choice to complete a task. Additionally, *ease of completion* of a task is a driving force in the task selection process of a worker. An interesting topic, a high reward, a less time consuming task also play a role in the choice of task of a worker in the crowd, albeit to a less prominent extent.

Secondly, a generic umbrella classification of microtasks, which is conceptualized based on the final goal of the tasks was proposed. This goal-oriented taxonomy splits the tasks into six high-level categories:

- *Information Finding*: tasks that require workers to simply find pieces of information by following instructions.
- *Verification and Validation*: tasks that require workers to verify certain aspects as per the given instructions.
- *Interpretation and Analysis*: tasks that require workers to provide information that is subject to their individual interpretation.
- *Content Creation*: tasks that require workers to generate new content.
- *Surveys*: tasks that require workers to answer several questions based on their opinion and background.
- *Content Access*: tasks that require workers to simply access some online content.

This top classification encompasses different kinds of microtasks that vary according to specific goals, however, at this level the classification is considered to be exhaustive. From the analysis of Gadiraju et al. [23], we learnt that microtask workers are dictated by their top priorities; to maximize monetary gain and minimize effort. In particular, the indifference of workers towards their reputation leads to many microtask workers becoming malicious. It is clear that many workers attempt to exert a minimum amount of effort to receive their reward. Unfortunately, in many cases the minimum effort is not enough for a task administrator to accumulate good or even acceptable results. What is equally clear, is that a task administrator must therefore try to prevent alternatives that allow workers to receive their rewards without providing valid results, i.e. prevent *cheating*.

Based on this prior knowledge of the workers' preferences, and the taxonomy of microtasks, in this paper we analyze the malicious behavior of workers in a specific class of microtasks: *Surveys*. We specifically choose to study this category, since surveys present the most difficult challenges with respect to ensuring accurate responses from workers. This is due to the inherently subjective nature of most surveys. Thus, *gold standards* cannot be applied easily. For example, in an '*Information Finding*' task, the task administrator might typically be able to ensure the validity of workers' responses by employing questions for which the answers are priorly known (gold standard). Thus, verifying the character of the worker. On the other hand, in a simple demographic survey, which is subject to receive multiple valid responses for a single question from the target audience, such practice is infeasible.

In this work, we aim to address research questions (RQs) by using the following definitions.

Definition 1. *Malicious workers* are workers deemed to have ulterior motives that deviate from the instructions and expectations as defined a priori by the microtask administrator.
Definition 2. *Untrustworthy workers* are workers who provide wrong answers in response to one or more simple and straightforward attention-check or gold standard questions.

RQ #1: Do untrustworthy workers adopt different methods to complete tasks, and exhibit different kinds of behavior?
RQ #2: How can task administrators benefit from the prior knowledge of plausible worker behavior?
RQ #3: Can behavioral patterns of malicious workers in the crowd be identified and quantified?

DATA COLLECTION

We obtained response-based data from workers, that is representative of the usual occurrence in the crowd, by deploying a survey using the CrowdFlower platform. We gather information about typically crowdsourced jobs, and worker preferences through their responses.

Since we aim to study the behavior of malicious workers, we do not use all the existing quality control mechanisms provided by CrowdFlower[6]. By doing so, we do not inhibit the general behavioral patterns of malicious workers.

Survey Design

To begin with, the survey consisted of questions regarding the demographics, educational and general background of the workers. Next, questions related to previous tasks that were successfully completed by the workers, are introduced. In total the survey contained 34 questions, spanning a mixture of open-ended, multiple-choice, and Likert-type questions designed to capture the interest of the workers; we collected the responses from 1000 workers.

We asked the crowd workers open-ended questions, about two of their most recent successfully completed tasks, as shown in Figure 1.

[6] `http://www.crowdflower.com/overview`

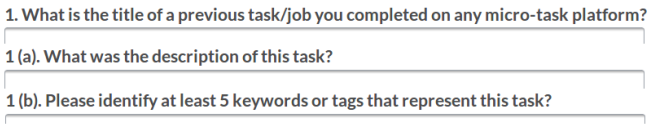

Figure 1. Open-ended questions to workers about their previous tasks.

We rely on open-ended questions in order to assess the malicious behavior depicted by untrustworthy workers. In addition, state-of-the-art qualitative research methods [24], have indicated that relying on recent incidents is highly effective, since respondents answer such questions with more details and instinctive candor. We pay all the contributors from the crowd 0.2 \$ per unit, irrespective of whether or not we discard their data for further analysis.

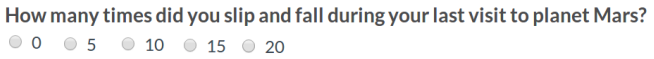

Figure 2. Engaging workers and checking their alertness by using *attention check* questions.

We used *attention check* questions, such as the one in Figure 2. The humor-evoking attention check questions interspersed with the regular questions, are known to keep the participants engaged, as indicated by Marshall et al. in previous work [13]. At the same time, these questions are also used to identify untrustworthy workers. We thereby follow the design recommendation given my Kittur et al. [12], where the authors suggest the importance of having explicitly verifiable questions. The authors also express the usefulness of having questions that require users to process the content, such as generating keywords or tags (see Figure1).

An important aspect to consider during the design of this survey, was to avoid influencing the workers into providing bad responses due to a poorly designed task. We took special care in order to ensure that the instructions and questions in the task were adequately clear. We gave workers unlimited time to complete the task. In addition, based on the optional feedback about the survey, received from 686 workers using the inbuilt facility on CrowdFlower, we observe that the workers were satisfied with the instructions $(4.1/5)$, ease of the job $(3.8/5)$, and the payments they received $(3.8/5)$ with an overall job satisfaction of $(4/5)$. Thus, we ensure that the workers behavior is not adversely effected by either a lack of time to complete the task, or the task design in general.

While in many crowdsourced tasks, the use of gold standards is widely applied to filter out malicious workers [8], this is not always applicable. For example, in crowdsourced tasks such as *Surveys* where none of the answers are known to the task administrators apriori, and in many cases, there is neither a correct nor a wrong answer. Note that we do not use other sophisticated means to curtail regular crowd worker behavior, in order to capture a realistic composition of workers (both trustworthy and otherwise).

DATA ANALYSIS

In this section, we first plot general results of the crowdsourced task. Later, we classify the behavior of trustworthy and untrustworthy workers in the crowd. We identify 432 untrustworthy workers by using test questions similar to the one depicted in Figure 2, who fail to pass at least one of two simple questions. These untrustworthy workers are then studied further in comparison to trustworthy workers, to determine plausible malicious traits.

Where are the workers from?

The Crowdflower platform forwards tasks to several different third-party crowdsourcing platforms, called '*channels*'[7]. In order to achieve coverage and results that are representative of the general crowdsourcing market, we do not impose any restrictions with respect to the channels. 50% of the workers who participated in the task used the 'Neodev' channel, while almost 25% of the workers used 'Clixsense'.

Since our survey was deployed in English, the first restriction we enforce via the platform is the language of the worker. However, it is difficult to accurately tell whether a worker is proficient in a given language. A simple workaround provided by the platforms exploits the location of the workers. Although imperfect, it is a reasonable assumption that a person located in an English speaking country, e.g. United States or Australia, is proficient in English (at least to an extent to understand and respond correctly to the questions in the task). Figure 3 shows the country distribution of the workers who participated in our task. We can observe that India leads by a large margin, followed by the USA, and Pakistan.

These numbers are anticipated since crowdsourcing is renown for widely employing workers from developing countries. In Figure 3, we divide the workers into two groups, trustworthy and untrustworthy, solely based on their responses to the attention check questions. In terms of percentage, we see that Pakistan, Sri Lanka, USA, and India lead in the number of workers who did not pass the attention checks.

Several hypotheses could be raised from these results. However further analysis of the influence of demographics, political, and economic factors are out of the scope of this paper. We are interested in analyzing and uncovering the universal user behavior that leads us to a coherent understanding of malicious activities. This can therefore provide us the required competence to restrict such malicious activity.

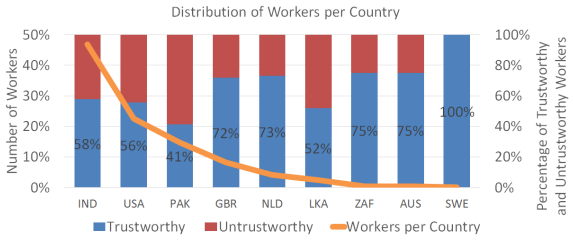

Figure 3. Distribution of the workers per country (left axis), and the distribution of trustworthy and untrustworthy workers (right axis).

Analyzing Malicious behavior in the Crowd

Prior research has shown that by devising typologies, we can provide a better structure to organize knowledge and study

[7] `http://www.crowdflower.com/labor-channels`

the relationships between disorderly concepts [25]. We analyze the implicit behavioral patterns of malicious workers by the means of their responses. Based on aspects such as (i) the eligibility of a worker to participate in a task, (ii) whether responses from a worker conform to the pre-set rules, or (iii) whether responses fully satisfy the requirements expected by the administrator, we determine the following types of behavioral patterns.

- **Ineligible Workers** (*IE*). Every microtask that is deployed on crowdsourcing platforms presents the workers in the crowd with a task description and a set of instructions that the workers must follow, for successful task completion. Those workers who do not conform to the priorly stated pre-requisites, belong to this category. Such workers may or may not provide valid responses, but their responses cannot be used by the task administrator since they do not satisfy the pre-requisites. For example, consider a pre-requisite in our survey, '*Please attempt this microtask ONLY IF you have successfully completed 5 microtasks previously*'. We observed that some workers responded to questions regarding their previous tasks with, '*this is my first task*', clearly violating the pre-requisite.

- **Fast Deceivers** (*FD*). Malicious workers tend to exhibit a behavior that is strongly indicative of the intention to earn easy and quick money, by exploiting microtasks. In their attempt to maximize their benefits in minimum time, such workers supply ill-fitting responses that may take advantage of a lack of response-validators. These workers belong to the class of *fast deceivers*. For example, workers who copy-paste the same response for different questions. In our survey, some workers copy-pasted the title of our survey, '*What's your task?*', in response to several unrelated open-ended questions. Some others simply entered gibberish such as '*adasd*', '*fygv fxc xdgj*', and so forth.

- **Rule Breakers** (*RB*). Another kind of behavior prevalent among malicious workers is their lack of conformation to clear instructions with respect to each response. Data collected as a result of such behavior has little value. For instance, consider the question from our survey, '*Please identify at least 5 keywords that represent this task*'. In response, some workers provided fewer keywords. In such cases, the resulting response may not be useful to the extent intended by the task administrator.

- **Smart Deceivers** (*SD*). Some eligible workers that are malicious, try to deceive the task administrators by carefully conforming to the given rules. Such workers mask their real objective by simply not violating or triggering implicit validators. For example, consider the instruction, '*Transcribe the words in the corresponding image and separate the words with commas*'. Here, workers that intentionally enter unrelated words, but conform to the instructions by separating the words with commas, may neutralize possible validators and achieve successful task completion. While this type of workers behave to an extent like *fast deceivers*, the striking difference lies in the additional attempt of *smart deceivers* to hide their real goal and bypass any automatic validating mechanisms in place. In our survey, some workers provided irrelevant keywords such as '*yes, no, please*', '*one, two, three*', and so forth to represent their preferred task-types. Some of these workers take special care to avoid triggering attention-check or gold standard type questions.

- **Gold Standard Preys** (*GSP*). Some workers who abide by the instructions and provide valid responses, surprisingly fail to surpass the gold standard questions. They exhibit non-malicious behavior, only to be tripped by one or more of the gold standard test questions. This may be attributed to the inattentiveness of such workers.

568 workers passed the gold standard questions (*trustworthy*) and 432 workers failed to pass at least one of the two test questions (*untrustworthy*). On analyzing each response from the workers, we found that only 335 of the trustworthy workers gave perfect responses (*elite workers*). A panel of 5 experts were presented the responses of each worker from the remaining 665 non-elite workers (233 trustworthy and 432 untrustworthy workers), and they manually classified the workers into the different classes, according to the class behavioral patterns described earlier. The inter-rater agreement between the experts during the classification of workers as per Krippendorffs Alpha is 0.94.

Based on majority voting and the agreement between the experts, we finalize the worker classification without discrepancies. 73 untrustworthy workers and 93 trustworthy workers were classified into 2 different classes, while the rest were classified into unique classes. Note that a worker can depict different kinds of behavior and thereby belong to multiple classes. Figure 4 presents the experts' classification of these workers as per the different types of behavioral patterns.

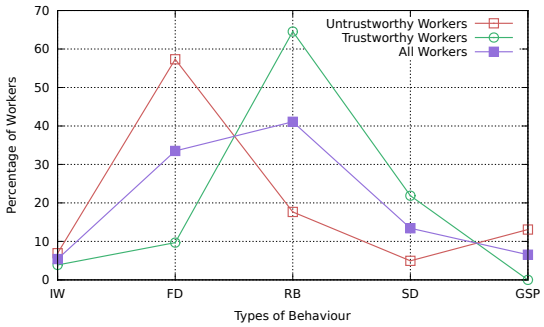

Figure 4. Distribution of non-elite workers as per their behavior.

More than 70% of all 665 workers classified are either *rule breakers* or *fast deceivers*. Nearly 60% of untrustworthy workers are *fast deceivers*, who intend to bypass response validators in order to earn monetary rewards easily. This is consistent with the findings of Kaufmann et al., wherein the authors establish that the number of workers who are mainly attracted by monetary rewards represent a significant share of the crowd [26]. About 65% of all non-elite trustworthy workers are *rule breakers*, who do not conform to the instructions laid out by the task administrators and thereby provide partially correct or limitedly useful responses.

The third most prevalent kind of untrustworthy workers are *smart deceivers*. Around 13% of all the classified workers take cautious steps in order to deceive task administrators and achieve task completion. These are malicious workers that tend to slip through most of the existing automated standards to prevent malicious activity, since they take special care to deceive the task administrators and receive the rewards at stake. This is made evident by the fact that over 20% of the non-elite trustworthy workers are *smart deceivers*, who give poor responses despite passing the gold standard questions.

Over 6% of all workers, seem to have failed the gold standard due to a lack of alertness (*gold standard preys*). This implies that a portion of workers' responses can be useful although the workers are deemed to be untrustworthy. Therefore, methods to identify and detect gold standard preys can benefit in maximizing the value of responses. This can be achieved either in a post-processing manner, or on the fly, at a relatively small additional cost.

Around 2.5% of workers attempt and complete tasks despite being clearly ineligible to take part (*ineligible workers*), as dictated by the pre-requisites. In our survey, such workers responded in languages other than in English, or in some cases claimed to have not completed any tasks before, thereby violating clearly stated pre-requisites.

Measuring Maliciousness of Workers

Next, we aim to measure the *maliciousness* of workers, as indicated by the acceptability of their individual responses.

Definition 3. The *acceptability* of a response can be assessed based on the extent to which a response meets the priorly stated expectations.

For example, consider the question, '*Enter the names of any 5 colors (separated by commas).*' A fully acceptable response to this question would be one which contains the names of 5 colors separated by commas (awarded a score of '1'). An *unacceptable* response on the other hand, is one which does not meet the requirements at all (awarded a score of '0'). So, in case of the same example, a response which does not contain names of colors would be completely unacceptable.

An important aspect to consider when measuring the maliciousness of a worker is interpreting the responses of the worker accurately. This means that we cannot reliably include subjective responses from the workers in such an analysis. For instance, consider a question with multiple checkbox options; any combination of responses to such a question may be acceptable. This means that in order to perform a reliable analysis, we have to consider only those responses with unambiguous corresponding acceptability. Therefore, we measure the maliciousness of workers by exploiting the acceptability of their responses to open-ended questions.

Experts manually annotated the responses from each worker for every open-ended question as either *acceptable* or *unacceptable*. The agreement between the experts was found to be 0.89 as per Krippendorf's Alpha. Figure 5 presents the average acceptability of workers' responses with respect to each open-ended question. Note that the questions *Q1*, and *Q4* ask workers to share the titles of their previously completed tasks on crowdsourcing platforms. *Q2* and *Q5* correspond to the description of these tasks, while *Q3* and *Q6* correspond to keywords representing these tasks. In the last open-ended question (*Q7*), the workers are asked to provide keywords pertaining to tasks that they prefer.

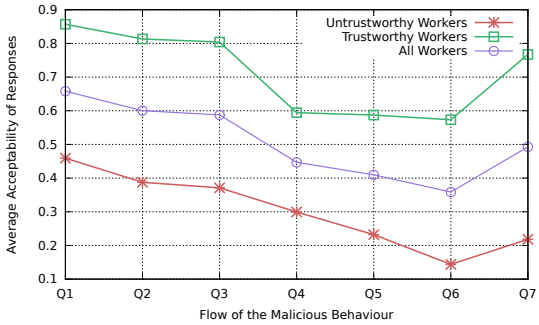

Figure 5. Average acceptability of responses from workers for each open-ended question *Q*.

Since we do not randomize the order of the questions for the different workers, we do not draw insights about the trend in acceptability through the course of the survey. However, we clearly observe that the acceptability of responses of the malicious workers reduces with the increase in required input from the workers (studied in literature as *task effort* [23]). It is easier for the workers to pass off a *title* as acceptable, than doing the same with either the *description* or *keywords* describing the task. On the whole, our findings indicate that the acceptability of individual responses of malicious workers decreases with an increase in the effort required to provide suitable responses.

Based on the acceptability of each response from a worker, we can compute the *average acceptability* (**A**) of a given worker pertaining to a task. In order to do so, we score each *acceptable response* with 1, and each *unacceptable response* with 0. Finally, we compute the *maliciousness* (**M**) of a worker using the following formula.

$$M_{worker} = 1 - (1/n \sum_{i=1}^{n} A_{r_i})$$

where,
n is the number of responses from the worker which are assessed, and A_{r_i} is the acceptability of response r_i.

$M_{worker} = 0$ indicates a completely non-malicious worker, while a worker is said to exhibit complete maliciousness if $M_{worker} = 1$. Figure 6 presents our findings regarding the distribution of workers with respect to the degree of their maliciousness, segmented by trustworthiness. In addition, the figure also depicts the corresponding average task completion time of the workers.

We can see that 50% of the untrustworthy workers exhibit a very strong maliciousness degree (greater than 0.8) while most of trustworthy workers (56%) have very low maliciousness. Nearly 20% of the untrustworthy workers exhibit a maliciousness degree between 0.4 and 0.6, while almost 15% indicate a high degree of maliciousness between 0.6 and 0.8.

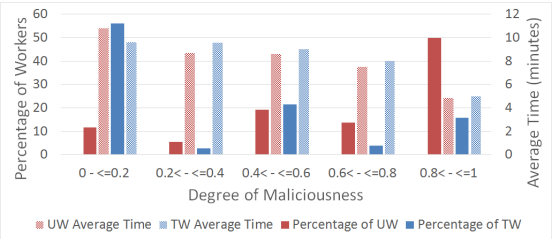

Figure 6. Degree of maliciousness of trustworthy (TW) and untrustworthy workers (UW) and their average task completion time.

In addition, we observe that the average task completion time of untrustworthy workers decreases with the increasing maliciousness. The same is observed for the trustworthy workers, where the group with highest maliciousness has the lowest average times. We find that for untrustworthy workers, the maliciousness and average task completion time show a high correlation of 0.51, as measured using Pearson Correlation. For trustworthy workers this correlation is moderate at 0.37.

The Tipping Point

In our study of the trustworthy and untrustworthy workers, we find that several workers provide acceptable responses to begin with, before depicting malicious behavior. We thereby investigate this tendency of workers to trail off into malicious behavior, and present our findings here.

Definition 4. We define the first point at which a worker begins to exhibit malicious behavior after having provided an acceptable response, as the *tipping point*.

The tipping point can be determined in terms of the number of responses at which the worker exhibits the first sign of malicious activity. In our analysis, we consider the open-ended questions. Note that we do not consider the workers who begin with providing unacceptable responses (we find 233 such untrustworthy workers, and 81 such trustworthy workers).

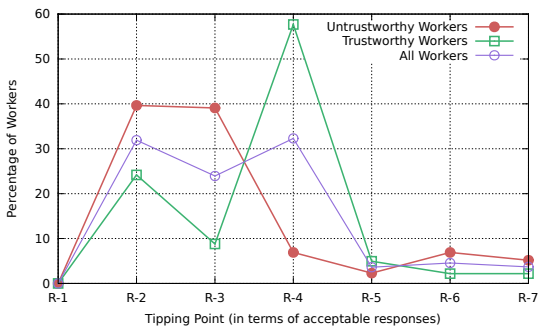

Figure 7. Distribution of Tipping Point of trustworthy and untrustworthy workers.

Figure 7 presents our findings. We can see that over 30% of all workers have a tipping point at their second response (R-2). This is largely due to the finding that almost 40% of all untrustworthy workers have a tipping point at R-2. Another 30% of all workers have a tipping point at their fourth response (R-4). Trustworthy workers largely contribute to this case. Nearly 60% of the non-elite trustworthy workers have a tipping point at R-4. On further analysis, we observe that these workers are mostly *rule breakers* who provide poor responses after the first set of questions about the previous tasks. Just below 25% workers depict tipping points at R-3, while under 5% of workers have a tipping point at R-5, R-6, and R-7. This shows that a significant number of malicious workers (especially untrustworthy workers) exhibit early signs of malicious activity, while a smaller percentage depict signs of malicious activity at a later stage.

Worker Maliciousness vs Tipping Point

We investigate the relationship between the maliciousness *(M)* of untrustworthy workers *(UW)*, trustworthy workers *(TW)* and their corresponding tipping points. We hypothesize that a worker with a greater maliciousness would have an earlier tipping point. Based on the analysis, we present our findings in Table 1.

Table 1. Relationship between the Maliciousness and Tipping Point of untrustworthy and trustworthy workers (percentage of workers having tipping point @R).

Maliciousness	UW	TW
$0 < M \leq 0.2$	40.9% @ R-7 31.8% @ R-6	28.5% @ R-7 28.5% @ R-5
$0.2 < M \leq 0.4$	43.47% @ R-3 21.73% @ R-6	30% @ R-5 30% @ R-3
$0.4 < M \leq 0.6$	66.19% @ R-3 25.35% @ R-2	88% @ R-4 5.1% @ R-3
$0.6 < M \leq 0.8$	71.05% @ R-2 28.95% @ R-3	60% @ R-3 40% @ R-2
$0.8 < M \leq 1$	100% @ R-2	100% @ R-2

We find that a majority of untrustworthy workers (40.9%) and trustworthy workers (28.5%) having a $M \leq 0.2$, have a tipping point at R-7. In case of the untrustworthy workers having $0.2 > M \leq 0.4$, 43.47% of workers have a tipping point at R-3, while 21.73% have a tipping point at R-6. In all the cases where $M > 0.4$, a great majority of workers have a tipping point at either R-3 or R-2. We observe a clear trend, which implies that the greater the maliciousness of a worker, the earlier is the 'tip' towards unacceptability. From this we learn that, a worker who provides poor responses in the beginning should be dealt with stricter measures, since there is a greater probability that the worker is malicious.

Worker behavior Beyond the Tipping Point

We analyzed the behavior of workers beyond their tipping points in order to verify whether the tipping point is a true indicator of further malicious activity from workers. Table 2 presents the amount of workers who depict malicious activity after their corresponding tipping points. We observe that over 95% of trustworthy and untrustworthy workers that have a tipping point at R-2, go on to provide at least one more unacceptable response.

Table 2. Percentage of workers that depict malicious activity after their corresponding Tipping Points.

Workers (in %)	R-2	R-3	R-4	R-5	R-6
Trustworthy	95.45	93.75	100	55.56	25
Untrustworthy	98.55	100	69.23	75	41.67

All trustworthy workers having a tipping point at R-4 and all untrustworthy workers having a tipping point at R-3, go on to provide at least one more unacceptable response. From Table 2, we learn that the tipping point is a good indicator of forthcoming malicious activity within the task.

Task Completion Time vs Worker Maliciousness

We also investigate the time that workers take in order to complete the task. In order to draw a comparison across the different types of behavior exhibited by workers, with respect to the time that they take for task completion, we use the average task completion time for each type of worker behavior. Apart from this, we also compare the maliciousness exhibited by each group of workers constituting the different types of behavior. We find that the *average task completion time* and the *average maliciousness* of untrustworthy workers show a high Pearson Correlation of 0.514.

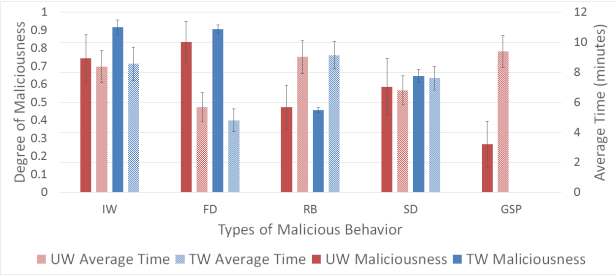

Figure 8. Comparison of Worker Maliciousness and Average Response Time of the different classes of malicious workers.

Figure 8 presents our findings with respect to the analysis described here. We observe that *fast deceivers* exhibit the most amount of maliciousness on average. Interestingly, they also take the least amount of time to complete the task. This is coherent with the type of behavior they exhibit, which is providing bad responses and achieve quick task completion (usually by copy-pasting same responses for multiple questions, or entering gibberish responses). On the other hand, we observe that *smart deceivers* also exhibit high malicious content, but they take more time to complete the task. We reason that this is due to the fact that *smart deceivers* take more precautions in order to bypass possible validators. *Gold standard preys* depict the least amount of maliciousness amongst all the types of untrustworthy workers. They also depict the highest average task completion time, indicative of a lower maliciousness. The *rule breakers* depict a high average task completion time, and a moderate maliciousness. This is attributed to their behavioral pattern; wherein the workers do not provide responses that meet the priorly stated requirements. *Ineligible workers* who complete the task, also depict a high maliciousness.

Caveats and Validity Threats

It is important to note that in this work, when we refer to 'maliciousness', we infer this based on the responses provided by a worker. There is no way to learn about the real intentions of a worker behind each response, based merely on the response itself.

While studying the major challenges that stand in the way of efficient crowdsourcing paradigms, Kittur et al. say that workers who are new and have relatively low expertise, as well as task administrators who do not provide clear instructions contribute to poor responses [27]. In order to ensure that we did not introduce unwanted bias due to the inexperience of workers (that could result in spiking the number of malicious workers), we ensured that all the questions in the crowdsourced survey were straightforward and easy to answer, even if a worker has little experience. Moreover, clear and thorough instructions were provided in the survey to aid the workers in completing the task.

We acknowledge that trustworthy workers may provide poor responses due to fatigue or boredom, as discussed in previous works. However, by varying the format of questions (open-ended, multiple choice check-boxes, Likert-type), limiting questions of the same type (two sets of 3 open-ended questions about previous tasks), and engaging the crowd with humor evoking attention-check questions, we attempt to curtail such bias.

The degree of *acceptability* corresponding to a worker's response is a metric that can be used at the discretion of the task administrator. In our case, we have computed the *acceptability* of a worker by awarding scores of '1' or '0' to each response, depending on whether a response is acceptable or unacceptable respectively. However, if a clear distinction with respect to the extent of usefulness of a response can be made, then a task administrator can use a continuous value between the closed interval of [0,1] in order to represent the acceptability of a response.

Since we do not randomize the order in which questions are answered by workers, we do not venture into analyzing the relationship between the type of question and the *tipping point*. We aim to extend this work with such an analysis in future. By doing so, we can empirically propose ideal lengths of tasks featuring different types of questions. Finally, more trials on different platforms, using varying design types would be ideal to further reinforce our findings.

DISCUSSION

Our experimental setup and findings are based on the task type, '*survey*'. A survey-type task inherently begets a general population of the crowd, without restricting participation due to the open design. Thus, the various kinds of trustworthy and untrustworthy workers presented in our work are representative of the general crowd. Having said that, the distribution of different kinds of untrustworthy workers depends on the type of task. This is due to the fact that a particular type of task may or may not be breached by some kinds of malicious workers, depending on the nature of the task and the gold standards being used.

Our experimental results showed that there was no significant correlation between the channels that the workers used for task completion and the behavioral patterns observed.

In our study of workers, we detect different types of worker behavior as described earlier. A key observation is that gold

standard test questions alone remain insufficient to curtail malicious activity. We find that trustworthy workers who pass test questions can still provide ill-fitting responses (as in case of *rule breakers*), or deceive the task administrators (as in case of *smart deceivers*). By understanding the various kinds of behavior prevalent in the crowd, administrators can design tasks much more effectively. Being aware of the different ways in which malicious workers attempt to cheat their way towards task completion, can help in developing mechanisms to counter such activity. For instance, we find that tasks of the 'survey' type are most prone to activity of the kind exhibited by *fast deceivers* and *rule breakers*. This urges the need for stringent response validation especially for open-ended questions, to curtail possible attempts to cheat by *fast deceivers*, and *rule breakers* who provide sub-optimal responses.

The responses from *Gold Standard Preys* are valid and acceptable, though they are tripped by the gold standard questions, owing to a possible lack of attentiveness. By detecting such workers and consuming their responses instead of discarding them, task administrators can enhance the value of responses received from the pool of workers in a task. This essentially means that, by detecting gold standard preys the value of responses can be maximized without increasing the costs for task completion.

The measurement of *maliciousness (M)* of the workers, as presented earlier can be extended to different types of tasks, since the method relies on determining the acceptability of the individual responses from a worker in the context of the task. Depending on his needs, a task administrator can choose to discard responses from workers based on their maliciousness, thus using *M* as a sliding window for filtering responses.

Task Design Guidelines

We propose the following guidelines in order to design tasks of the '*survey*' type efficiently. By adhering to these key guidelines, we claim that the malicious activity prevalent in tasks of this type can be curtailed to a significant extent.

- The *tipping point* can be used to identify workers who 'tip early in the job. By excluding such workers, the quality of the produced results can be improved.
- In order to restrict the participation of *ineligible workers*, task administrators could employ a commonly used pre-screening method.
- Stringent validators should be used in order to ensure that workers cannot bypass open-ended questions by copy-pasting identical or irrelevant material as responses. This is an important guideline to enforce for survey-type tasks, since open-ended questions are popular in surveys and the majority of malicious workers are *fast deceivers*.
- *Rule breakers* can be curtailed by ensuring that basic response-validators are employed, so that workers cannot pass off inaccurate responses, or nearly fair responses. Even trustworthy workers tend to tip through the course of a task, providing poor or partially accurate responses. This demands for methods to monitor the progress of workers. Such validators can enforce workers to meet the exact requirements of the task and prevent ill-fitting responses.
- Additional methods and careful steps are required to prevent malicious activity by *smart deceivers*. Since such workers take care to avoid being flagged, they present the most difficulties in detecting and containing. Only a small number of workers make the additional effort to deceive task administrators in surveys. Yet, these workers can be restricted by using psychometric approaches such as repeating or rephrasing the same question(s) periodically and cross-checking whether the respondent provides the same response.
- Surveys garner a fair number of *gold standard preys*. Therefore, a post-processing step should be accommodated in order to identify such workers and consider their acceptable responses if needed.

CONCLUSIONS AND FUTURE WORK

The ubiquity of the internet, allows to distribute *crowdsourcing* tasks that require human intelligence at an increasingly large scale. This field has been gaining rapid popularity, not least because of the data-intensive nature of emerging tasks, requiring validation, evaluation and annotation of large volumes of data. While certain tasks require human intelligence, humans can exhibit maliciousness that can disrupt accurate and efficient utilization of crowdsourcing platforms. In our work, we aim to understand this phenomenon.

We have studied the behavior of malicious workers in the crowd by showcasing the task type of *Surveys*. Based on our analysis, we have identified different types of malicious behavior (**RQ #1**), which go beyond existing works and are better-justified through our data. An understanding of these aspects helps us to efficiently design tasks that can counter malicious activity, thereby benefiting task administrators as well as ensuring adequate utilization of the crowdsourcing platforms (**RQ #2**). By conducting an extensive analysis, we introduce the novel concepts of measuring 'maliciousness' of workers in order to quantify their behavioral traits, and 'tipping point' to further understand worker behavior (**RQ #3**). Our contributions also include a set of guidelines for requesters to efficiently design crowdsourced surveys by limiting malicious activity.

As part of our future work, we will develop machine learning methods to identify workers according to their behavior and classify them into the different types established in this work. Next, we intend to present an extensive set of methodologies and guidelines for effective task design and deployment on crowdsourcing platforms. In order to do so, we will delve into understanding malicious behavior for each type of task in the taxonomy introduced by Gadiraju et al. [23], namely, *information finding*, *verification and validation*, *interpretation and analysis*, and *content creation*.

REFERENCES

1. E. Estellés-Arolas and F. González-Ladrón-de Guevara, "Towards an integrated crowdsourcing definition," *Journal of Information science*, vol. 38, no. 2, pp. 189–200, 2012.
2. D. C. Brabham, "Crowdsourcing as a Model for Problem Solving," *Convergence: The International*

Journal of Research into New Media Technologies, vol. 14, no. 1, pp. 75–90, Feb. 2008.

3. K. Ntalianis, N. Tsapatsoulis, A. Doulamis, and N. Matsatsinis, "Automatic annotation of image databases based on implicit crowdsourcing, visual concept modeling and evolution," *Multimedia Tools and Applications*, vol. 69, no. 2, pp. 397–421, 2014.
4. G. Kazai, J. Kamps, and N. Milic-Frayling, "An analysis of human factors and label accuracy in crowdsourcing relevance judgments," *Inf. Retr.*, vol. 16, no. 2, pp. 138–178, 2013.
5. P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 2010, pp. 64–67.
6. C. Eickhoff and A. de Vries, "How crowdsourcable is your task," in *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the fourth ACM international conference on web search and data mining (WSDM)*, 2011, pp. 11–14.
7. R. Gennaro, C. Gentry, and B. Parno, "Non-interactive verifiable computing: Outsourcing computation to untrusted workers," in *Advances in Cryptology–CRYPTO 2010*. Springer, 2010, pp. 465–482.
8. D. Oleson, A. Sorokin, G. P. Laughlin, V. Hester, J. Le, and L. Biewald, "Programmatic gold: Targeted and scalable quality assurance in crowdsourcing." *Human computation*, vol. 11, p. 11, 2011.
9. D. E. Difallah, G. Demartini, and P. Cudré-Mauroux, "Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms." in *CrowdSearch*, 2012, pp. 26–30.
10. S. Dow, A. Kulkarni, S. Klemmer, and B. Hartmann, "Shepherding the crowd yields better work," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012, pp. 1013–1022.
11. T. S. Behrend, D. J. Sharek, A. W. Meade, and E. N. Wiebe, "The viability of crowdsourcing for survey research," *Behavior research methods*, vol. 43, no. 3, pp. 800–813, 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*. ACM, 2008, pp. 453–456.
13. C. C. Marshall and F. M. Shipman, "Experiences surveying the crowd: Reflections on methods, participation, and reliability," in *Proceedings of the 5th Annual ACM Web Science Conference*, ser. WebSci '13. New York, NY, USA: ACM, 2013, pp. 234–243.
14. Y. Baba, H. Kashima, K. Kinoshita, G. Yamaguchi, and Y. Akiyoshi, "Leveraging crowdsourcing to detect improper tasks in crowdsourcing marketplaces," in *Twenty-Fifth IAAI Conference*, 2013.
15. S. Dow, A. Kulkarni, B. Bunge, T. Nguyen, S. Klemmer, and B. Hartmann, "Shepherding the crowd: managing and providing feedback to crowd workers," in *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2011, pp. 1669–1674.
16. W. Mason and D. J. Watts, "Financial incentives and the performance of crowds," *ACM SigKDD Explorations Newsletter*, vol. 11, no. 2, pp. 100–108, 2010.
17. J. Ross, L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?: shifting demographics in mechanical turk," in *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2010, pp. 2863–2872.
18. G. Kazai, J. Kamps, and N. Milic-Frayling, "Worker types and personality traits in crowdsourcing relevance labels," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1941–1944.
19. G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Serf and turf: crowdturfing for fun and profit," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 679–688.
20. S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating." in *INTERSPEECH*, 2011, pp. 3053–3056.
21. C. Eickhoff and A. P. de Vries, "Increasing cheat robustness of crowdsourcing tasks," *Information retrieval*, vol. 16, no. 2, pp. 121–137, 2013.
22. M.-C. Yuen, I. King, and K.-S. Leung, "A survey of crowdsourcing systems," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. IEEE, 2011, pp. 766–773.
23. U. Gadiraju, R. Kawase, and S. Dietze, "A taxonomy of microtasks on the web," in *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 2014, pp. 218–223.
24. J. Corbin and A. Strauss, *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage, 2008.
25. R. L. Glass and I. Vessey, "Contemporary application-domain taxonomies," *IEEE Software*, vol. 12, no. 4, pp. 63–76, 1995.
26. N. Kaufmann, T. Schulze, and D. Veit, "More than fun and money. worker motivation in crowdsourcing - a study on mechanical turk," in *AMCIS*, 2011.
27. A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton, "The future of crowd work," in *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2013, pp. 1301–1318.