Make Hay While the Crowd Shines: Towards Efficient Crowdsourcing on the Web

Ujwal Gadiraju
«Supervised by Prof. Wolfgang Nejdl and Dr. Stefan Dietze»
L3S Research Center
Appelstr. 9a
30167 Hanover, Germany
{gadiraju}@L3S.de

ABSTRACT

Within the scope of this PhD proposal, we set out to investigate two pivotal aspects that influence the effectiveness of crowdsourcing: (i) microtask design, and (ii) workers behavior. Leveraging the dynamics of tasks that are crowdsourced on the one hand, and accounting for the behavior of workers on the other hand, can help in designing tasks efficiently. To help understand the intricacies of microtasks, we identify the need for a taxonomy of typically crowdsourced tasks. Based on an extensive study of 1000 workers on CrowdFlower, we propose a two-level categorization scheme for tasks. We present insights into the task affinity of workers, effort exerted by workers to complete tasks of various types, and their satisfaction with the monetary incentives. We also analyze the prevalent behavior of trustworthy and untrustworthy workers. Next, we propose behavioral metrics that can be used to measure and counter malicious activity in crowdsourced tasks. Finally, we present guidelines for the effective design of crowdsourced surveys and set important precedents for future work.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Crowdsourcing; Microtasks; Workers; Behavioral Patterns

1. PROBLEM AND MOTIVATION

In recent times, crowdsourcing has emerged as an important means to acquire human input in order to solve a wide range of problems. Researchers and practitioners alike have identified the value of crowdsourcing as a cost-effective paradigm. In his seminal book 'The Wisdom of Crowds' [13], James Surowiecki pointed out the importance of diversity and independence in the crowd to achieve useful results through crowdsourcing. The advent of crowdsourcing platforms such as Amazon's Mechanical Turk[1] and CrowdFlower[2] provide such important ingredients for effective crowdsourcing and elevate the ease with which one can reach out to millions of workers with minimum effort.

With the growing popularity of crowdsourcing, several users are confounded by the problem of designing microtasks. The inadequacy of existing guidelines for crowdsourcing corresponding to the different types of tasks mean that less-experienced users have to go through an arduous cycle of experimenting with reward schemes, task lengths and other such parameters in order to acquire suitable results. While there have been numerous works that have studied crowdsourcing in order to improve the quality and the reliability of the results produced [7], there are several challenges that remain unconquered. One such shortcoming is a granular classification of crowdsourced microtasks. A critical step towards improving the effectiveness of crowdsourcing is to understand its typical uses. Therefore, we aim to propose a taxonomy of microtasks that can prove to be useful for the user modeling of crowd workers, and the recommendation of tasks. Researchers can leverage this taxonomy in order to study both, popular microtasks as well as the types of worker behavior associated with these tasks.

A primary obstacle that hinders the optimal output from microtasks is the malicious activity that is prevalent among workers. Previous works have acknowledged the existence of spammers and malicious workers who aim to acquire quick rewards, and thereby provide responses which are either misleading or fall well short of the requirements. *Gold-standards* are a popular means to detecting and countering malicious activity [11]. However, with the ubiquity of the Internet and the blooming crowdsourcing market, malicious workers are also evolving and we believe adversarial approaches will become more sophisticated and popular, thereby overcoming commonly used gold-standards. Previous work by Difallah et al. portray the inadequacy of existing techniques in detecting malicious workers and spammers [1]. Mechanisms that counter such attempts of workers need to consider a wide range of workers with varying behavioral patterns.

It is important to note that malicious activity of crowd workers and the design of the corresponding microtasks go hand in hand; stringent task design can curtail possible malicious activity, while less strict tasks allow easy infiltration of malicious workers. Having said that, this is also largely subjective with respect to the type of microtasks, i.e., dif-

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741748 .

[1] http://www.mturk.com/
[2] http://crowdflower.com/

ferent types of microtasks may attract significantly different types of workers. In addition to this, malicious workers can sabotage a task in various ways. It is therefore of prime importance to understand the behavior of workers (especially malicious workers) in different types of microtasks, in order to improve the quality of results, reliability of workers and cost-effectiveness of the crowdsourced task in its entirety.

In this PhD proposal, we aim to study the behavior of malicious workers with respect to varying microtask types. In addition, we propose a method to measure the malicious activity in crowdsourced microtasks. Finally, we aim to delve into inherent traits of workers such as their competence, and study how these factors influence their behavioral patterns.

2. RELATED LITERATURE

We describe the relevant literature by dividing it into the two realms of (i) task design and (ii) workers behavior. Due to the space limitation we illustrate the most influential previous works.

2.1 Task Design and Quality of Results

Marshall et al. anlayzed workers who took surveys on Amazon's Mechanical Turk and examined how the characteristics of the surveys influenced the reliability of the data produced [9]. Inspired by this work, we adopt a similar approach to collect data through crowdsourced surveys in order conduct a meaningful analysis and arrive at sound insights.

Yuen et al. present a literature survey on different aspects of crowdsourcing [14]. The authors present a taxonomy of crowdsourcing research alongside a sample set of application scenarios. This short list represents the first step towards task modeling. However, without a well-defined structure of task types, goals or work-flows, it is challenging to reuse such information to devise optimal strategies for task design. We provide a solution to this problem by providing a clear taxonomy of microtasks in terms of goals and work-flows.

Geiger et al. proposed a framework for crowdsourcing processes based on 46 crowdsourcing examples [6]. The authors provided a 19-class crowdsourcing process classification that focuses exclusively on an organizational perspective. This provides useful insights for stakeholders responsible for running crowdsourcing platforms. On the contrary, we propose a categorization scheme that primarily supports task administrators in effectively utilizing such platforms.

Mason et al. studied the effect of varying financial incentives on the performance of workers [10]. The authors conclude that increasing monetary incentives of microtasks attracts more workers to the tasks but does not improve the quality of the results produced.

2.2 Workers Behavioral Patterns

Eickhoff et al. acknowledged the importance of understanding worker behavior in order to develop reliability metrics and design fraud-proof tasks [2]. Kazai et al. used behavioral observations to define the types of workers in the crowd [8]. By type-casting workers as either *sloppy*, *spammer*, *incompetent*, *competent*, or *diligent*, the authors expect their insights to help in designing tasks and attracting the best workers to a task. While the authors correlate these types to the personality traits of workers, we aim to unravel the behavioral patterns of workers through their responses.

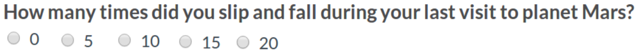
Ross et al. extended previous works on the usage behavior of workers on Mechanical Turk along with their changing demographics [12]. The authors reflect that the global scale at which tasks are crowdsourced can affect the quality of the responses. In our work, we aim to identify the main behavioral patterns that workers exhibit independent of their demographical identity.

Eickhoff et al. additionally evaluated factors such as the size and type of microtask, interface used and composition of the crowd [3]. Based on this the authors suggest to design tasks in a manner that discourages malicious workers. In our work, we aim to inhibit such malicious activity depending on the behavioral patterns of crowd workers.

3. APPROACH AND METHODOLOGY

Two main research questions (RQs) that we aimed to analyze through real crowdsourced data are as follows;
(**RQ1**) What kinds of tasks are typically crowdsourced?
(**RQ2**)What kinds of behavior do crowd workers exhibit?

In order to gather real crowdsourced data we deployed a survey using the CrowdFlower platform. The survey consisted of questions ranging from the demographics and background of the workers, to specific questions regarding the previous tasks that were successfully completed by them. These questions were modeled as a mixture of open-ended, direct, and Likert-type questions, with an aim to engage the workers. We restrict the participation to 1000 workers, and ask about two of their most recent successfully completed tasks in the form of open-ended questions. We pay all the contributors from the crowd, irrespective of whether or not we discard their data for further analysis. We use gold standard questions, as shown in Figure 1 in order to identify untrustworthy workers.



How many times did you slip and fall during your last visit to planet Mars?

○ 0 ○ 5 ○ 10 ○ 15 ○ 20

Figure 1: We use humor-evoking attention-check questions as gold standards in order to engage workers and also identify untrustworthy workers.

In order to gather realistic data from the trustworthy and untrustworthy workers alike, we do not use sophisticated measures to restrict the general crowd behavior. We take several precautions in order to avoid introducing any bias due to poor task design. We refer the reader to our work for further details [4, 5].

We analyze the responses from the trustworthy workers in order to propose a taxonomy of microtasks [4]. By studying the responses from the untrustworthy workers in contrast to trustworthy workers, we identify the behavioral patterns that workers exhibit in crowdsourced surveys [5].

In the future, we will deploy microtasks of different types as per the proposed taxonomy, in order to further study aspects of worker behavior such as the influence of a worker's competence in the behavioral patterns exhibited. An interesting aspect that has not been dealt with thoroughly thus far, is crowdsourcing highly complex tasks. We believe a study of worker competencies will help in designing methods to successfully crowdsource complex tasks. In addition, we aim to gather crowdsourced data from different types of tasks as per the proposed taxonomy, in order to further study the influence of task type on the behavior of workers.

Table 1: A two-level taxonomy for typically crowdsourced tasks.

Information Finding (IF)	Verification & Validation (VV)	Interpretation & Analysis (IA)	Content Creation (CC)	Surveys (S)	Content Access (CA)
Metadata finding	Content Verification	Classification	Media Transcription	Feedback/Opinions	Testing
	Content Validation	Categorization	Data Enhancement	Demographics	Promoting
	Spam Detection	Media Transcription	Translation		
	Data matching	Ranking	Tagging		
		Data Selection			
		Sentiment Analysis			
		Content Moderation			
		Quality Assesment			

4. RESULTS

In this section, we present the key results that we have arrived at during the course of this work. For more interesting results emerging from our work so far, we refer the readers to our work [4, 5].

4.1 Taxonomy of Microtasks

Based on the responses provided by the 567 trustworthy workers, we manually categorize crowdsourced tasks into a two-level taxonomy as presented in the Table 1. The top level consists of *goal-oriented* classes while the second level contains sub-classes of these that are based on the *work-flow* of the microtasks. The top level describes the overall objective of a given microtask, while the second level describes the process that a worker has to go through in order to complete the task successfully and help the task administrator achieve his goal. Therefore note that by definition, in this taxonomy work-flow oriented sub-classes can belong to multiple goal-oriented top level classes. This categorization scheme serves as first step towards establishing task-specific guidelines for the efficient design of microtasks.

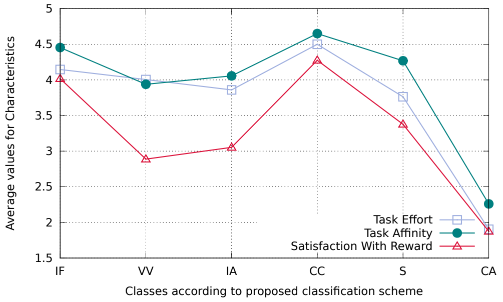


Figure 2: Distribution of task-related characteristics according to the proposed microtask taxonomy.

By leveraging the data collected from trustworthy workers, we also analyze task dependent characteristics of workers such as their *task affinity*, *task effort* and their *satisfaction* with the monetary reward. We find that workers tend to show greater task affinity (i.e., they like the tasks) which offer relatively higher monetary rewards. Figure 2 presents our findings regarding how these task dependent aspects vary across the different microtask types. We observe that workers tend to put more effort into the tasks that they have a greater affinity towards.

Additionally, we investigate the landscape of financial incentives that are offered to the workers based on our data. We study how such monetary incentives effect the task affinity of workers. Figure 3 presents a comparison between the task dependent characteristics with respect to the incentives.

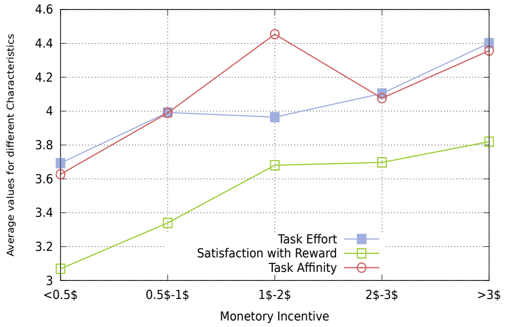


Figure 3: Distribution of effort required, task affinity, and satisfaction with reward of the workers with respect to varying task incentives.

4.2 Behavior Typology of Workers

Next, we venture into unravelling the behavioral patterns of workers (especially malicious workers) based on their responses. We rely on the following factors to determine the behavior topology proposed in our work; (i) eligibility of a worker to participate in a task, (ii) adherence of responses to pre-stated rules, and the extent to which responses satisfy the expectations of the administrator.

Ineligible Workers (IE). Task administrators present instructions to the workers that they should follow to complete a given task successfully. The workers who do not qualify as per such priorly stated requisites belong to this category.

Fast Deceivers (FD). Malicious workers are characterized by their behavior that is highly suggestive of a zeal to earn quick money by exploiting microtasks. This is apparent from some some workers who adopt the 'fast-response-first' approach such as copy-pasting the same response for instance. Such workers belong to the class of fast deceivers.

Smart Deceivers (SD). Some eligible workers who are malicious, attempt to deceive task administrators by cleverly adhering to the rules. Such workers mask their real objective by simply not violating or triggering implicit validators, and belong to this category.

Rule Breakers (RB). A 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, since the resulting responses may not be useful to the extent intended by the task administrator.

Gold Standard Preys (GSP). Some workers who abide by the instructions and provide valid responses, surprisingly fall short at the gold standard questions. They exhibit non-malicious behavior, they stumble at one or more of the gold

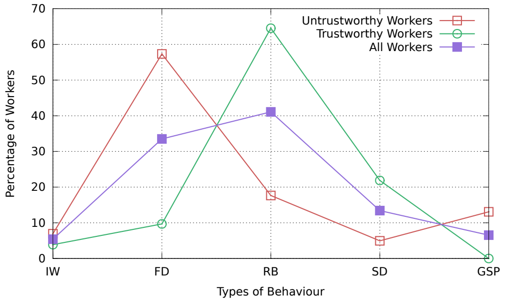


Figure 4: Distribution of workers as per the behavioral patterns exhibited.

standard test questions due to to their inattentiveness, fatigue or boredom.

We manually classify the 568 trustworthy workers (those who passed the gold standard) and 432 untrustworthy workers (those who failed at least one of the two gold standard questions) into the behavior topology described earlier. Figure 4 presents the distribution of trustworthy and untrustworthy workers across the different kinds of behavioral classes.

4.3 How Malicious is a Malicious Worker?

We measure the maliciousness of a worker based on the *acceptability*[3] of their individual responses. For this purpose, it is important to consider only those responses from a worker where the acceptability is not contentious or subjective. Thus, we only consider the responses from the workers to the open-ended questions (*Q1* to *Q7*). Experts manually annotated each response from the workers as either *acceptable:'1'* or *unacceptable:'0'*. The agreement between the experts was found to be 0.89 as per Krippendorf's Alpha.

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. We compute the *maliciousness* (**M**) of a worker using the following metric.

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

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 5 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 clearly see that a large percentage of untrustworthy workers exhibit a high degree of maliciousness (> 0.8), while the majority of trustworthy workers exhibit a very low degree of maliciousness (< 0.2).

4.4 The Tipping Point in Workers Behavior

We find that several workers provide acceptable responses to begin with, before showing signs of maliciousness. We

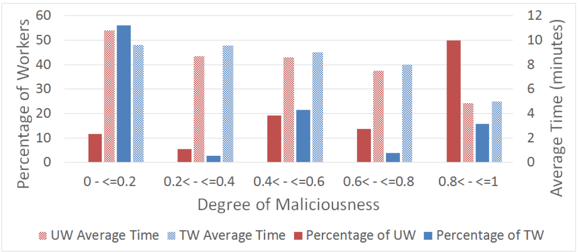


Figure 5: Degree of maliciousness of trustworthy (TW) and untrustworthy workers (UW) and their average task completion time.

thereby investigate this tendency of workers to trail off into malicious behavior. We define the first point at which a worker begins to exhibit malicious behavior after having provided an acceptable response, as the *tipping point.* We present the tipping point of workers based on our analysis of their responses (*R-1* to *R-7*) in the Figure 6. 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.

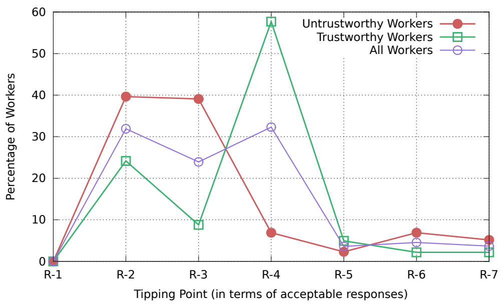


Figure 6: Distribution of Tipping Point of trustworthy and untrustworthy workers.

On further analysis of the activity of workers beyond their tipping point, we find that the tipping point is a very good indicator of forthcoming malicious activity [5].

4.5 Task Design Guidelines for Surveys

Based on each kind of behavioral pattern that we observe within the topology, we recommend ways to make a crowdsourced survey effective.

- The *tipping point* can be used to identify workers who 'tip' early in the job. We can improve the quality of the produced results greatly by identifying such workers and discarding them.
- To restrict the participation of *ineligible workers*, task administrators should employ pre-screening methods.
- An important guideline to enforce for survey-type tasks due to the popularity of open-ended questions is to curtail malicious activity from *fast deceivers.* 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.
- *Rule breakers* can be curtailed by ensuring that basic response-validators are employed, so that workers

[3] The acceptability of a response is determined by the extent to which the response meets the expectation as stated.

cannot pass off inaccurate responses, or nearly fair responses. Lexical validators can enforce workers to meet the exact requirements of the task and prevent ill-fitting responses.

- Since *smart deceivers* take special precautions to avoid being detected, they present the biggest hindrance in overcoming. Although only a meagre portion of workers make the additional effort to deceive task administrators in surveys, these workers can be restricted by using psychometric approaches (for instance, repeating or rephrasing the same question(s) periodically and cross-checking whether the respondent provides the same response).
- Finally, we note that surveys garner a fair number of *gold standard preys*. We recommend post-processing step that can be accommodated in order to identify such workers and consider their acceptable responses to boost the reliability and quality of results.

5. CONCLUSIONS AND FUTURE WORK

5.1 Key Contributions

In our work thus far, we have proposed a two-level taxonomy of microtasks. This fine-grained categorization of crowdsourced tasks has important implications in the field of crowdsourcing. It serves as an essential starting point to analyze several aspects of both task design and worker behavior across well-defined types of tasks.

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 kinds of malicious behavior which go beyond existing works and are better-justified through our data. A thorough understanding of these aspects helps us to design tasks that can counter malicious activity effectively, thereby benefiting task administrators as well as ensuring adequate utilization of the crowdsourcing platforms.

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. Our contributions also include a set of guidelines for requesters to efficiently design crowdsourced surveys by limiting malicious activity.

5.2 Upcoming Work

As part of our future work, we will develop automatic methods to identify workers as per 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. We will therefore study malicious behavior for each type of task in the proposed taxonomy.

There are no existing principles based on which a task administrator can adjust important parameters such as *length* of a task, or the *incentive* to be offered in order to obtain optimal results in the presence of any limiting constraints. We aim to delve into this problem and develop learned models that can recommend ideal schemes to support task administrators in adjusting task related parameters for specific tasks. Finally, we aim to apply our findings in data-driven crowdsourcing applications at scale, in order to help solve real world problems such as knowledge-base curation, event validation, entity search and so forth more efficiently (in terms of costs, accuracy and reliability of the results).

6. REFERENCES

[1] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*, pages 26–30, 2012.

[2] 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)*, pages 11–14, 2011.

[3] C. Eickhoff and A. P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137, 2013.

[4] 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.

[5] U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of CHI'15, CHI Conference on Human Factors in Computing Systems*, 2015.

[6] D. Geiger, S. Seedorf, T. Schulze, R. C. Nickerson, and M. Schader. Managing the crowd: Towards a taxonomy of crowdsourcing processes. In *AMCIS*, 2011.

[7] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM, 2010.

[8] 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*, pages 1941–1944. ACM, 2011.

[9] 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*, WebSci '13, pages 234–243, New York, NY, USA, 2013. ACM.

[10] W. Mason and D. J. Watts. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108, 2010.

[11] 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*, 11:11, 2011.

[12] 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*, pages 2863–2872. ACM, 2010.

[13] J. Surowiecki. *The wisdom of crowds*. Random House LLC, 2005.

[14] 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 social computing (socialcom)*, pages 766–773. IEEE, 2011.