Gadiraju, U.; Demartini, G.; Kawase, R.; Dietze, S.: Human Beyond the Machine: Challenges and Opportunities of Microtask Crowdsourcing. In: IEEE Intelligent Systems 30 (2015), Nr. 4, S. 81-85. DOI: https://doi.org/10.1109/MIS.2015.66

Human Beyond the Machine: Challenges and Opportunities of Microtask Crowdsourcing

Ujwal Gadiraju, *L3S Research Center*

Gianluca Demartini, *University of Sheffield*

Ricardo Kawase and Stefan Dietze, *L3S Research Center*

The notion of *crowdsourcing* existed long before the term itself was coined. The idea that small contributions from a group of individuals can be accumulated to accomplish some work or attain an objective has been observed in different realms of daily life for a few centuries. Over the years, crowdsourcing has emerged as a useful paradigm, showcasing the adage that "the whole is greater than the sum of its parts."

The Web's emergence as a sociotechnical system has dramatically changed the scale and scope of crowdsourcing, opening the possibility of reaching crowds at an unprecedented global scale. Crowdsourcing has proven to be an increasingly important source of knowledge and data, as documented by prominent examples such as Wiki-pedia, where many authors contribute discrete and diverse information to form an authoritative reference knowledge base. The recent Wiki-data project (www.wikidata.org) and the popular ReCAPTCHA work by Luis von Ahn and colleagues[1] are exemplary. Moreover, a recent report by eYeka *(http://tinyurl.com/mqsrets)* on the state of crowdsourcing in 2015 shows high adoption in business contexts, documenting that 85 percent of the top global brands (as defined at www.bestglobalbrands.com) use crowdsourcing for various purposes.

Particularly in research and science, crowd-sourcing has found noteworthy applications in solving real-world problems, ranging from intricate tasks such as protein folding and bio-molecule design[2] and mapping outer space[3] to aiding disaster relief initiatives[4] and assembling dictionaries.[5]

With the ubiquity of the Internet, and the concomitant accessibility of established microtask crowdsourcing platforms such as Amazon's Mechanical Turk (MTurk; www.mturk.com) and, more recently, CrowdFlower (www.crowdflower.com), researchers and practitioners are actively turning toward paid crowdsourcing to solve data-centric tasks that require human input. Popular applications include building ground truths, validating results, and curating data. These developments make it possible to build novel intelligent systems that combine the scalability of machines over large amounts of data with the power of human intelligence to solve complex tasks, such as audio transcription, language translation, and an-notation. For the rest of this article, *crowdsourcing* will refer to *paid microtask crowdsourcing* (wherein crowd workers receive monetary compensation for successfully completing a micro-task). For a thorough discussion of related terms and tasks in the context of human computation and collective intelligence, see the work of Alexander Quinn and Benjamin Bederson.[6]

Owing to the diversity in the crowd in terms of workers' motivations, demographics, and competencies, both microtask design and quality control mechanisms play an unparalleled role in determining the effectiveness of crowdsourcing systems.[7] These two realms, which specifically are concerned with the requesters' perspective on micro-tasks, have thereby attracted much interest and are our focus in this article.

We summarize and discuss findings from some previous work related to microtask performance and design.[8]–[9][10] Although crowdsourcing presents many open challenges, including ethical concerns we focus on performance-related aspects. In this article, we provide an overview of frequently crowdsourced microtasks, malicious activity observed in crowds, and open research challenges in the field.

Overview Of Frequently Crowdsourced Microtasks

To further the understanding of crowdsourced tasks on popular crowd-sourcing platforms such as MTurk and CrowdFlower, we categorized typically crowdsourced tasks into a two-level taxonomy, [8] using responses from 567 crowd workers regarding their previously completed tasks. The top level comprises goal-oriented classes, and the

second level describes the workflow that a worker must adopt to complete a microtask. We describe the top-level classes below:

- *Information finding* (IF) tasks delegate the process of searching to satisfy an informational need to the workers in the crowd. (For ex-ample, “Find information about a company in the UK.”)
- *Verification and validation* (VV) tasks require workers in the crowd to either verify certain aspects as per the given instructions or confirm the validity of various kinds of content. (For example, “Is this a spam bot? Check whether the Twitter users are either real people or organizations or spam user profiles.”)
- *Interpretation and analysis* (IA) tasks rely on the crowd workers to use their interpretation skills during task completion. (For example, “Choose the most suitable category for each URL.”)
- *Content creation* (CC) tasks usually require workers to generate new content for a document or website. They include authoring product descriptions or producing question-answer pairs. (For ex-ample, “Suggest names for a new product.”)
- *Surveys* (S) about a multitude of aspects ranging from demographics to customer satisfaction are crowd-sourced. (For example, “Mother's Day and Father's Day survey [18-to 29-year-olds only!]”)
- *Content access* (CA) tasks require the crowd workers to simply access some content. (For example, “Click on the link and watch the video,” or “Read the information by following the website link.” In these tasks, the workers are asked to consume some content by accessing it, but to do nothing further.)

Note that, depending on the final goal of the task, tasks with the same workflow can belong to different classes at the top level of the micro-task taxonomy.

Microtask Evolution Over the Years

The most widely used microtask crowdsourcing platform for academic purposes is Mechanical Turk. There has been a constant evolution of its usage patterns since it was launched in 2005. We presented a large-scale analysis of log data from this micro-task

crowdsourcing platform, [9] showing how the use of microtask crowd-sourcing has evolved over the past *five years.*

Analyzing data from www.mturk-tracker.com about 130 million tasks posted on MTurk between 2009 and 2014, we observed the evolution of platform usage in terms of task type, task pricing, work volume, and required workers. Our main findings are as follows:

- Task reward has increased over time, reaching$0.05 as the most popular reward level in 2014.
- Tasks requesting audio transcriptions are now the most popular tasks in the platform.
- In terms of task type, CA tasks have become less popular over time, while surveys are becoming more popular, especially for workers based in the US.
- Most tasks do not specify a requirement on where the crowd worker is physically located.
- The number of active requesters (that is, those who publish tasks to be completed) has increased over time, with a rate of 1,000 new requesters per month over the past two years.
- The overall amount of tasks being published and completed on the platform is considerable: 10,000 new tasks are published and 7,500 tasks are completed per hour. Certain requesters can obtain a work throughput of up to thousands of tasks completed per minute.
- Newly published tasks are almost 10 times more attractive for workers compared to old tasks.

Our findings indicate that requesters should engage with workers and publish large volumes of HITs to more quickly obtain data back from the crowdsourcing platform.

Breaking Bad: Typology of Workers and Malicious Activity

Our findings clearly indicate rapid growth in microtask crowdsourcing. Given the crowd workers' inherent characteristics, with respect to their demographics, diversity, and motives, quality control mechanisms that can make crowdsourcing processes more effective play a crucial role. Previous work has asserted that crowdsourced microtasks

are often hindered by the presence of spammers and malicious workers who aim to complete micro-tasks as quickly as possible to gain monetary rewards.[11], [12] We delved into worker behaviors that determine per-formance, and therefore the quality, of crowdsourced results.[10] Crowd workers exhibit distinct behaviors that affect the overall quality of the work. We hypothesize that by understanding the behavioral patterns of trustworthy workers (workers who pass gold-standard test questions) and untrustworthy workers (workers who fail one or more gold-standard test questions), requesters can inhibit malicious workers from adversely affecting the task output.

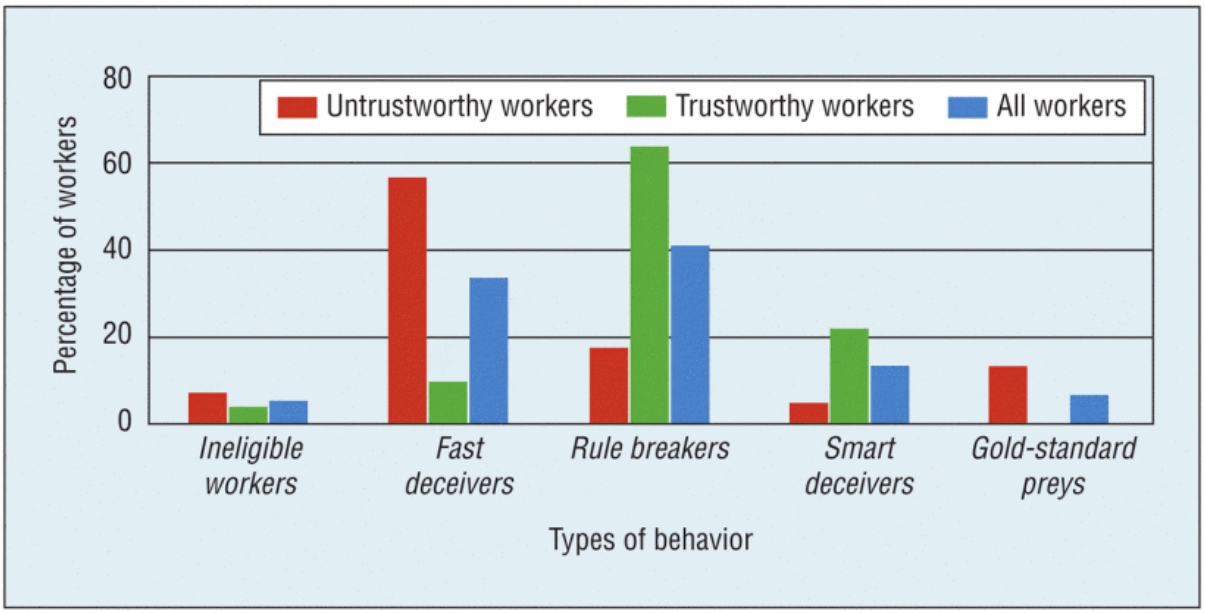

Figure 1.

Distribution of trustworthy and untrustworthy workers as per the behavioral patterns exhibited.[10]

On the basis of the responses we received from 1,000 workers in a crowdsourced survey task, we determined the following behavioral typology of trustworthy and untrustworthy crowd workers.[10]

- *Ineligible workers* (IE). Microtask requesters present instructions to the workers that they should follow to complete a given task suc-cessfully. The workers who do not qualify as per previously stated requisites belong to this category.

- *Fast deceivers* (FD). Malicious workers are characterized by behavior that suggests a zeal to earn quick money by exploiting microtasks. This is apparent from some workers who adopt the "fastest-response-first" approach, such as copy-pasting the same response in multiple fields.
- *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.
- *Rule breakers* (RB). A behavior prevalent among malicious workers is the lack of conformation to clear instructions with respect to each response. Data collected as a result of such behavior has little value, because the resulting responses might 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 nonmalicious behavior but stumble at one or more of the gold-standard test questions because of inattentiveness, fatigue, or boredom.

Figure 1 depicts the distribution of the workers whose responses were studied and consequently classified, as per the behavioral pattern they exhibited. It is interesting to note that fast deceivers are the most prevalent type of untrustworthy workers in crowdsourced surveys.

In addition, we found a strong correlation between a worker's task completion time and the malicious intent exhibited. We also introduced a worker's tipping point as an early indicator of possible malicious activity, which can help us detect and reduce malicious activity and hence improve worker performance. For more information see our previous work.[10]

To restrict the participation of ineligible workers, task administrators should employ prescreening methods. Stringent validators should be used to ensure that fast deceivers 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 cannot pass off inaccurate or nearly fair responses. Lexical validators can enforce workers to meet the task's exact requirements and prevent ill-fitting responses. Smart deceivers can be restricted by using psychometric approaches (for

instance, repeating or rephrasing the same question periodically and cross-checking whether the respondent provides the same response).

Crowdsourcing: the Right Hammer for Every Nail?

Crowdsourcing is becoming a ubiquitous approach to dealing with ma-chine-based computation's limitations by leveraging human intelligence at scale. However, as we have described, crowdsourcing-based solutions must deal with higher uncertainty with respect to worker performance. To make crowdsourcing approaches more efficient and effective, one must consider various aspects. In this article, we list the open challenges when aiming toward wider adoption of crowdsourcing.

In the absence of adequate measures to control crowd workers' performance, crowdsourced data has been shown to deviate far from the desirable. However, it may not always be possible to enforce constraints through task design. Depending on the task at hand, it is not always straightforward to control the performance in real time. Furthermore, not all tasks can be modeled as microtasks that are fit for crowd workers. Task decomposition to facilitate atomic microtask crowd-sourcing, especially in the case of complex tasks, is an interesting open challenge.

In the following, we introduce some of the open research challenges toward making crowdsourcing-based solutions more dependable.

Crowdsourcing Efficiency

To make crowdsourcing solutions scale to large amounts of data, it is key to design solutions that will retain crowd workers longer in the crowdsourcing platform [13] and prioritize work execution over the crowd.

Incentives, gamification, and sat-isfaction. To retain workers on the platform, one can leverage custom task-pricing schemes.[14] gamification techniques, [15] or competitive task designs[16] to recognize worker contributions, direct workers toward specific task types, and balance task difficulty to keep workers productive.

Scheduling tasks. To make sure most important tasks are completed as quickly as possible, scheduling techniques must be considered to prioritize tasks. Such techniques also need to consider worker properties such as training effects, context switch cost, and personal preferences.

Performance Monitoring

To obtain higher-quality crowdsourced data, we identify the need to create solutions that better profile crowd work-ers. This will let us understand which worker will perform with high quality on which tasks. This can be done by automatically detecting malicious workers and effectively routing tasks to the right workers in the crowd.

Detecting poorly performing workers. It is important to effectively detect low-quality workers in the crowd to remove their answers from the generated annotations. This can be done using advanced result aggregation techniques or supervised machine learning approaches.

Task routing. By profiling crowd workers over time, we can understand each worker's strengths and weak-nesses. We can then leverage such information to assign microtasks to the right workers in the crowd rather than randomly assigning them.

Crowdsourcing Ethics

Crowdsourcing is used in production for many commercial products. For a sustainable crowdsourcing environment, a crowdsourcing ethics culture is required that includes schemes to identify fair pricing of work and build long-term worker perspectives and careers.

Fair pricing. In terms of fair pricing, it's important to design methods to correctly price microtasks rather than relying exclusively on market dynamics (that is, balance of demand and offer). Estimating the effort that a microtask takes to complete is key, [17] but it also requires personalized estimations for different types of workers.

Long-term work perspective. From a long-term work perspective, crowd workers are satisfied of their rewards because of the short-term benefits they perceive. However, in the long term, there is no personal development program or social security scheme in place.[18] A better regulation of the crowdsourcing market is necessary-one that considers crowd worker careers and strives for improved transparency (such as in Turkopticon [19].)

Given the controversial discussion of such ethical concerns, a wider discourse of such topics is required to improve the performance of both the workers and the requesters in crowd-sourcing settings. With efforts underway to address such challenges, we anticipate a growing importance of crowdsourcing-based efforts in data-centric tasks, specifically to complement and validate machine-processed results with human intelligence.

References

1. L. von Ahn et al. "Recaptcha: Human-based Character Recognition via Web Security Measures", Science, vol. 321 no. 5895 pp. 1465-1468 2008.
2. B.L. Ranard et al. "Crowdsourcing—Harnessing the Masses to Advance Health and Medicine a Systematic Review", J. General Internal Medicine, vol. 29 no. 1 pp. 187-203 2014.
3. E.C. McLaughlin "Image Overload: Help Us Sort It All Out NASA Requests" Aug. 2014 [online] Available: http://edition.cnn.com/2014/08/17/tech/nasa-earth-images-help-needed.
4. M. Zook et al. "Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake", World Medical & Health Policy, vol. 2 no. 2 pp. 7-33 2010.
5. N. Lanxon "How the Oxford English Dictionary Started Out Like Wikipedia", Wired, Jan. 2011 [online] Available: www.wired.co.uk/news/archive/2011-01/13/the-oxford-english-wiktionary.
6. A.J. Quinn B.B. Bederson "Human Computation: A Survey and Taxonomy of a Growing Field", Proc. SIGCHI Conf. Human Factors in Computing Systems, pp. 1403-1412 2011.
7. P.G. Ipeirotis F. Provost J. Wang "Quality Management on Amazon Mechanical Turk", Proc. ACM SIGKDD Workshop on Human Computation, pp. 64-67 2010.
8. U. Gadiraju R. Kawase S. Dietze "A Taxonomy of Microtasks on the Web", Proc. 25th ACM Conf. Hypertext and Social Media, pp. 218-223 2014.
9. D.E. Difallah et al. "The Dynamics of Micro-Task Crowdsourcing—The Case of Amazon MTurk", Proc. 24th Int'l Conf. World Wide Web, pp. 238-247 2015.
10. C. Eickhoff A.P. de "Increasing Cheat Robustness of Crowdsourcing Tasks", Information Retrieval, vol. 16 no. 2 pp. 121-137 2013.
11. G. Kazai J. Kamps N. Milic-Frayling "Worker Types and Personality Traits in Crowdsourcing Relevance Labels", Proc. 20th ACM Int'l Conf. Information and Knowledge Management, pp. 1941-1944 2011.
12. U. Gadiraju et al. "Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys", Proc. 33rd Ann. ACM Conf. Human Factors in Computing Systems, pp. 1631-1640 2015.
13. M. Rokicki et al. "Competitive Game Designs for Improving the Cost Effectiveness of Crowdsourcing", Proc. 23rd ACM Int'l Conf. Information and Knowledge Management, pp. 1469-1478 2014.

14. D.E. Difallah et al. "Scaling-Up the Crowd: Micro-Task Pricing Schemes for Worker Retention and Latency Improvement", Proc. 2nd AAAI Conf. Human Computation and Crowdsourcing, pp. 50-58 2014.
15. K. Siorpaes E. Simperl "Human Intelligence in the Process of Semantic Content Creation", World Wide Web, vol. 13 pp. 33-59 2010.
16. M. Rokicki S. Zerr S. Siersdorfer "Groupsourcing: Team Competition Designs for Crowdsourcing" , Proc. 24th Int'l World Wide Web Conf., 2015.
17. J. Cheng J. Teevan M.S. Bernstein "Measuring Crowdsourcing Effort with Error-Time Curves", Proc. 33rd Ann. ACM Conf. Human Factors in Computing Systems, pp. 1365-1374 2015.
18. A. Kittur et al. "The Future of Crowd Work", Proc. Conf. Computer Supported Cooperative Work, pp. 1301-1318 2013.
19. L.C. Irani M.S. Silberman "Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk" , Proc. SIGCHI Conf. Human Factors in Computing Systems, pp. 611-620 2013.

Ujwal Gadiraju

L3S Research Center

Ujwal Gadiraju is a PhD candidate in the L3S Research Center at the Leibniz Universitat Hannover. Contact him at gadiraju@13s.de

Gianluca Demartini

University of Sheffield

Gianluca Demartini is a lecturer in data science in the Information School at the University of Sheffield. Contact him at g.demartini@sheffield.ac.uk

Ricardo Kawase

L3S Research Center

Ricardo Kawase is a data analyst for the trust and safety team at mobile.de. He performed the work for this article as a researcher in the L3S Research Center at the Leibniz Universitat Hannover. Contact him at rkawase@team.mobile.de

Stefan Dietze

L3S Research Center

Stefan Dietze is a research group leader in the L3S Research Center at the Leibniz Universitat Hannover. Contact him at dietze@13s.de