Crowdsourcing, Collaboration and Creativity

While many organizations turn to human computation labor markets for jobs with black-or-white solutions, there is vast potential in asking these workers for original thought and innovation.

By Aniket Kittur

DOI: 10.1145/1869086.1869096

Crowdsourcing has become a powerful mechanism for accomplishing work online. Hundreds of thousands of volunteers have completed tasks including classifying craters on planetary surfaces (http://clickworkers.arc.nasa.gov), deciphering scanned text (ReCaptcha http://recaptcha.net), and discovering new galaxies (http://galaxyzoo.org). Crowdsourcing has succeeded as a commercial strategy as well, with examples ranging from crowdsourcing t-shirt designs (Threadless http://www.threadless.com) to research and development (Innocentive http://www2.innocentive.com).

One of the most interesting developments is the creation of general-purpose markets for crowdsourcing diverse tasks. For example, in Amazon's Mechanical Turk, tasks range from labeling images with keywords to judging the relevance of search results to transcribing podcasts. Such micro-task markets typically involve small tasks (on the order of minutes or seconds) that users self-select and complete for monetary gain. These markets represent the potential for accomplishing work for a fraction of the time and money required by more traditional methods.

Crowdsourcing has worked especially well for certain kinds of tasks, typically, ones that are fast to complete, incur low cognitive load, have low barriers to entry, are objective and verifiable, require little expertise, and can be broken up into independent subtasks. Examples in Mechanical Turk include tagging images, evaluating search results, finding contact information, and labeling data for use in machine learning.

However, these simple tasks are only tappint the lowest levels of human intelligence. Are there opportunities to extend general purpose crowdsourcing markets to complex tasks that require significant intelligence or creativity? Innocentive, for example, has already shown that some companies may be willing to pay tens of thousands of dollars for one-off creative research and development—however, this represents a relatively small pool of highly expert workers and highly motivated companies.

The potential opportunity is enormous: tapping into hundreds of thousands or millions of workers across the globe to accomplish creative work on an unprecedented scale. In this article, I discuss ways to accomplish tasks that go beyond those that have traditionally been successful in micro-task markets, using subjective user studies and collaborative translation as examples. I also highlight some challenges and directions for future work that may be helpful for guiding other researchers interested in the field. I focus primarily on research-relevant tasks, given that those will be what most readers will find of most interest; however, many of the techniques presented here may apply to accomplishing complex work tasks as well.

CROWDSOURCING SUBJECTIVE TASKS

There are many potential tasks for which there is no one right answer. Researchers collect user ratings, elicit opinions, give surveys, and run experiments in which each individual's opinion may be different but equally valid. The possibilities of accomplishing tasks that don't have a predefined answer or solution through crowdsourcing in a matter of hours instead of weeks and for pennies instead of dollars is, with good reason, quite enticing. However, getting good results is no trivial matter. Subjective tasks

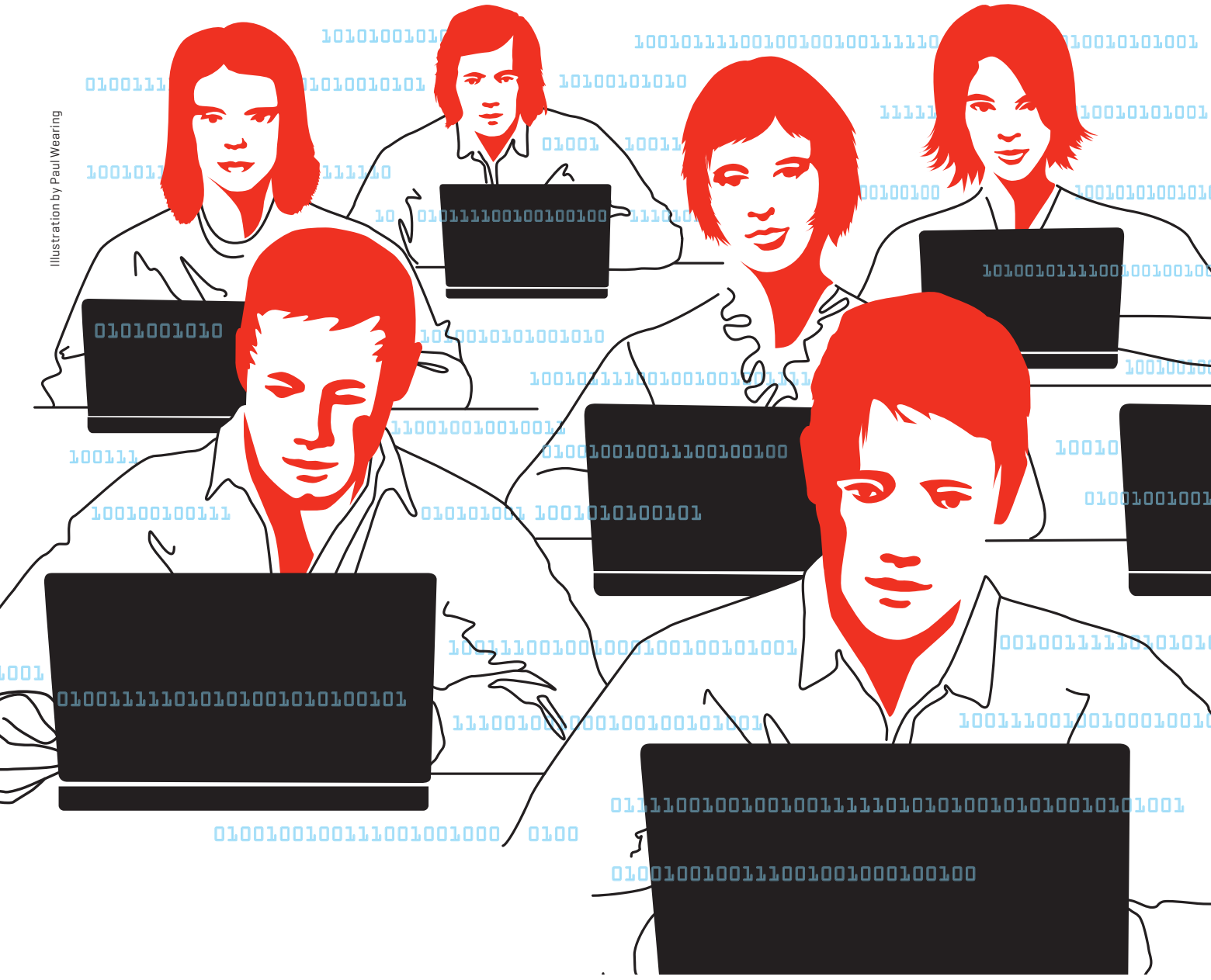


Illustration by Paul Wearing

are especially prone to users "gaming" the system, providing low-effort or random answers in order to reduce the time and work needed to get a reward. How would you know whether a given answer is a user's honest opinion or whether she is just clicking randomly?

My research partners and I first encountered this in 2007 when we tried to get turkers to rate the quality of Wikipedia articles, which has been a difficult thing for researchers to do with automated tools. With high hopes, we posted a number of articles to Mechanical Turk, paying turkers $0.05 to make judgments about their quality and to write what improvements they thought the article needed. We got results remarkably quickly: 200 ratings within two days.

But when we looked at the results, it was obvious that many people were not even doing the task. Almost half the responses to the writing question were blank, uninformative, or just copied and pasted, and about a third spent less than a minute on the task—less time than is needed to read the article, let alone rate it. This was not a promising beginning.

Is there something we could do to improve the quality of responses and reduce gaming? In traditional settings, there are a number of mechanisms that can mitigate gaming. For example, experimenters have long known that being in the same room as a participant can lead to better performance by the participant, as she knows the experimenter is monitoring her. Social norms, sanctions, and a desire to avoid looking bad can promote higher quality contributions in groups that know each other. The potential for additional future work can motivate high quality work in the present, as can formal or informal reputation systems that enable future employers to observe past performance. Job loss can create hardships, with time needed to find another job. Explicit contracts can be enforced through legal systems, causing high costs to those who do not honor them.

However, these mechanisms are less effective or absent in online microtask markets such as Mechanical Turk. Social norms and sanctions are largely absent as workers are interchangeable and usually identified only by their worker ID—a random string of numbers and letters.

Interestingly, there are social sanctions of employers by the workers in Mechanical Turk, who use forums such as turkernation.com to alert other workers about bad employers (for more, see "Ethics and Tactics of Professional Crowdwork," page 39). This is partially in response to the asymmetrical power balance in Mechanical Turk: employers can reject work for any reason with no recourse for the worker whose has already put in the effort but does not get paid. Unification sites like Turkernation let workers sanction employers who consistently do not approve legitimate work.

New identities are relatively easily created, though they do require some external validity, such as credit card verification. Monitoring is essentially non-existent since the worker may be in any physical location in the world. Furthermore, employers don't know whether a worker who has accepted the task is actually engaged in it or is multitasking or watching TV.

Workers can choose between employers and jobs easily with no switching costs. Although some rudimentary reputation systems do exist (such as tracking the proportion of work rejected), workers with even low reputations can often find jobs to complete. There are no explicit contracts. Even after a worker has accepted a job, she can return it any time for any reason without consequence.

In the absence of external mechanisms for enforcing quality responses in subjective tasks, we turned to the design of the task itself. Specifically, we had two key criteria for task design. First, we wanted it to take the same amount of effort for a worker to enter an invalid but believable response as a valid one written in good faith. Second, we wanted to signal to the workers that their output would be monitored and evaluated.

To meet these criteria, we altered the rating task. Instead of subjective ratings followed by subjective feedback about what could be improved, we required turkers to complete three simple questions that had verifiable, quantitative answers, such as how many references/images/sections the article had. We also asked turkers to provide between four and six keywords summarizing the article. Importantly, we selected these questions to align with what Wikipedia experts claimed they used when rating articles (such as examining the references or the article structure), with the goal that by answering these questions, they would have a reasonable judgment of the quality of the article. We placed the verifiable questions before the subjective questions so workers would have the opportunity to develop this judgment before even having to think about subjective questions. Finally, since these questions have concretely verifiable answers, they signal that workers' responses can and will be evaluated—preventing gaming in the first place and potentially increasing effort (criteria 2).

"Instead of subjective ratings followed by subjective feedback about what could be improved, we required turkers to complete three simple questions that had verifiable, quantitative answers."

Re-running our experiment with the new task design led to dramatically better results. The percent of invalid comments dropped from 49 percent to 3 percent, improving by more than a factor of 10. Time spent on the task also more than doubled, suggesting increased effort. This was borne out by a positive and statistically significant correlation between turker ratings and those of expert Wikipedians. Finally, we found that we tapped a more diverse group, with more users contributing and a more even spread of contributions across users. (Details of the study can be found in "Crowdsourcing User Studies With Mechanical Turk," in *Proceedings of the ACM Conference on Human-factors in Computing Systems*, 2008.)

Although our initial results were rocky, in the end we found that crowdsourcing markets like Mechanical Turk can be quite useful for subjective tasks, as long as the task is designed appropriately to prevent gaming and ensure quality responses.

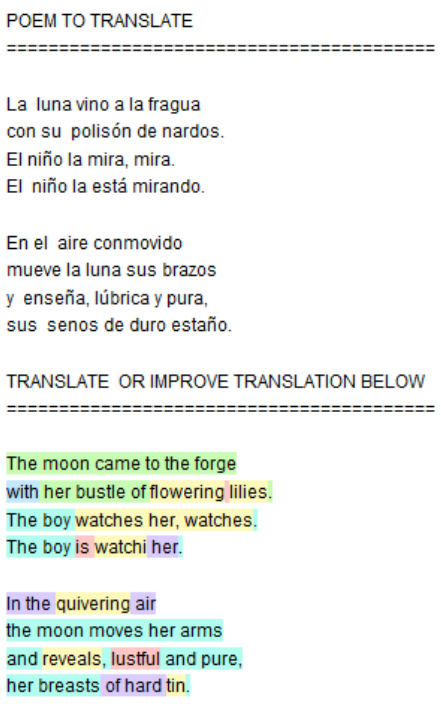
This opens up possibilities for entire new domains of tasks to be accomplished an order of magnitude faster and cheaper through crowdsourcing than traditional means. Imagine being able to do user testing of design prototypes and getting feedback within days from hundreds of users for dozens of prototypes, or running dozens of experiments a month, and accessing participants from across the globe rather than the low-diversity undergraduate participant pool available at most universities.

COLLABORATIVE CROWDSOURCING

One common assumption about Mechanical Turk is that turkers must work independently of each other. Most tasks involve turkers each making an independent judgment about an object (such as providing a label for an image) with their judgments aggregated afterward.

However, even interdependent tasks do not involve turkers interacting with each other. For example, the company CastingWords accomplished podcast transcriptions in a serial fashion: one turker may do the initial transcription; the transcription is automatically split into segments; other turker workers verify or improve the segments. Throughout, turkers never have to interact with each other despite using the results of each others' work. This is a reasonable approach when a requester does not know who will accept a task, when they will complete it, what the quality of the work will be, and when there are few dependencies such that work can be easily split up and done in parallel.

Figure 1: Workers collaborated to translate a poem using Etherpad, and the results are shown. Each color indicates a different worker's contribution.



Many collaborative tasks in the real world, however, involve people interacting with each other. Examples range from scientists collaborating on a discovery to students collaborating on a report to volunteers writing articles together on Wikipedia.

In real-world collaborations, interaction is the norm rather than the exception. There are many advantages to interacting groups, such as the ability to communicate and coordinate on the fly rather than having to follow prespecified plans or rules, motivational gains from identifying with a group, and the bonds formed from interacting with other group members and helping behavior between them. There may be interesting benefits from breaking the assumption of independence and enabling workers to collaborate interactively in crowdsourcing.

However, such benefits are by no means certain. For example, would workers participating in a financial market really help each other without any financial incentives?

We examined this question in the context of a problem that is both difficult overall and especially difficult to do in a crowdsourcing context: collaborative translation. Unlike the short, simple, objective, and verifiable tasks that are typical of Mechanical Turk, translation can be complex, challenging, time-consuming, highly subjective, and impossible to verify automatically. It is also highly interdependent, requiring a consistent voice and approach throughout. However, if we were able to harness the power of the crowd for translation, there could be many potential benefits ranging from supporting disaster relief efforts (as already demonstrated by CrowdFlower and Samasource in crowdsourcing translation in the Haitian relief efforts, see page 10) to providing essential training data to help machine translation research.

In order to support turkers working interactively on translations we used Etherpad, an open-source platform for real-time collaborative editing that gave us the ability to track the participation patterns of each worker, support keystroke-by-keystroke real-time editing, show the contributions of each worker in a different color, and support chat between workers. Using this platform we had turkers come and add to or improve a translation of the famous Spanish poet Garcia Lorcas' poem "Romance de la luna," paying them $0.15 for their contributions. We started turkers with the original untranslated Spanish, as well as the first two sentences from a published English translation (see **Figure 1**).

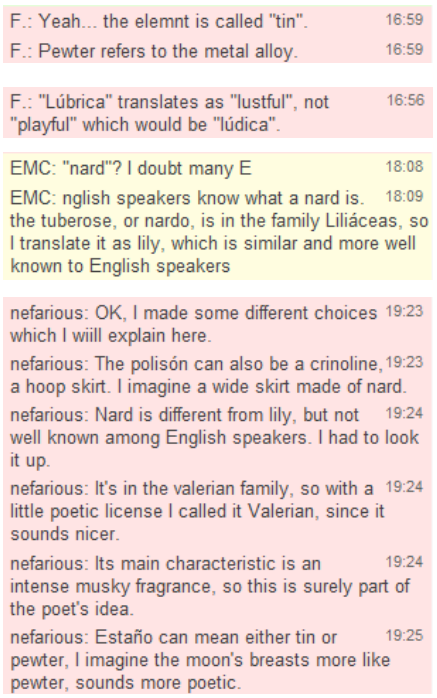
As an aside, we found a marked improvement in acceptance and completion of the task when it was already "seeded" with a couple sentences that workers thought were contributed by others. This may be because it didn't seem quite as overwhelming, and new workers had a place to start; or it may have signaled that it was legitimate work; or that it was a reasonable amount of work since others had done it. Such seeding may be a useful technique for other tasks as well.

Within hours, more than a dozen turkers were working on the translation interactively, seeing each others' edits reflected in real time on the pad. After 48 hours, we took the results and had a different set of turkers compare it to a published translation, rating which they thought was a better translation of the original poem. Interestingly, 14 out of 16 bilingual raters pre-

Figure 2: Comparing the original poem (left), a published translation by Havard, 1990 (center), and final crowdsourced version from our experiment (right), the crowdsourced version was preferred by 14 out of 16 bilingual raters.

Romance de la luna, luna	Translation 1	Translation 2
La luna vino a la fragua con su polisón de nardos. El niño la mira, mira. El niño la está mirando.	The moon came to the forge in her bustle of spikenard. The boy stares at her. The boy is staring hard.	The moon came into the forge with her bustle of flowering lilies. The boy watches her, watches. The boy gazes upon her.
En el aire conmovido mueve la luna sus brazos y enseña, lúbrica y pura, sus senos de duro estaño.	In the feverish air the moon sways her arms, showing, lewd and spotless, her cruel, tin breasts.	In the quivering air the moon rotates her arms and, lustful and pure, reveals her breasts of hard tin.
Huye luna, luna, luna. Si vinieran los gitanos, harian con tu corazón collares y anillos blancos.	Run away, moon, moon, moon. If the gypsies find us, they would cut out your heart to make necklaces, silvery rings,	Fly moon, moon, moon, If the gypsies came, of your heart they would make white rings and necklaces.
Niño, déjame que baile. Cuando vengan los gitanos, te encontrarán sobre el yunque con los ojillos cerrados.	Child, let me dance. When the gypsies come, they will find you on the anvil with your tiny eyes shut tight.	Child, let me dance. When come the gypsies, on the anvil they'll find you with your lovely eyes closed.
Huye luna, luna, luna, que ya siento sus caballos.	Run away, moon, moon, moon. I can hear their horses.	Fly moon, moon, moon For I already feel their horses.

Figure 3: Amazon Mechanical Turk workers show an ability to communicate and coordinate surprisingly well on a translation project.



ferred the crowdsourced work over the published translation (see **Figure 2**). The total cost for this translation was less than $5.

There were many additional interesting things we found about the process. Certain passages were more difficult to translate than others, and we could see which these were as they were continually changed. For example, "En el aire conmovido" went through ten different revisions starting with "amidst the shaken air" and ending with "in the quivering air" by ten different contributors. In contrast, the first sentence ("La luna vino a la fragua") was translated as, "The moon came to the forge," and was not subsequently changed, suggesting it was easier to translate in the context of the poem. Such meta-information could provide valuable feedback for translation research or judging the ease or quality of translations.

The process of turkers working together interactively was at least as interesting. Many of them coordinated their edits with others, explaining and asking advice using chat. Some examples are shown in **Figure 3**. They also found the process enjoyable and rewarding, making statements such as "this was fun" or "sweet" in the chat, and emailing us further changes to the poem after the experiment was over, despite already having been paid.

Another unexpected outcome happened when we left the pad open after an experiment completed. Some turkers removed the original poem and replaced it with a new poem which others then came and translated for free! Together, these results show the potential for gains in effort, motivation, coordination, and quality that can be achieved by letting people work together collaboratively rather than treating them only as simple processors or algorithms. It's especially remarkable given the context of the environment, which was a market-based crowdsourcing platform in which the primary driver is money. Is there more value than we at first recognized in supporting the social nature of crowdsourcing, especially for creative tasks?

CROWDSOURCING RESEARCH TO COME

Crowdsourcing is already a powerful approach to solving a variety of problems faster and cheaper than traditional methods, but how far we can push it? For example, instead of treating crowd workers as simple sensors or processors, can we harness their human intelligence and creativity? Instead of focusing on simple, independent tasks, can we generalize to more complex tasks where people need to work together? In the limit, one can imagine a future in which the *de facto* method of doing work is through crowdsourcing markets, with experts and non-experts collaborating in *ad-hoc* groups to accomplish tasks.

We need to learn much more about the possibilities and limits of crowdsourcing. One important area of research may be understanding the motivational and reward structure of crowd workers and how they generalize across different kinds of markets. For example, workers on Mechanical Turk working for pennies are very different from those in Innocentive, the R&D crowdsourcing market in which solutions may pay tens of thousands of dollars.

Even for less extreme cases, such as people working for the kinds of virtual cash provided by Facebook games, such as Farmville, there are different demographics and motivations than those on Mechanical Turk. Furthermore, motivation may not be solely based on external rewards. Intrinsic motivations such as fun, camaraderie, and meaning may be just as powerful mechanisms for motivation, and may have especially beneficial effects on quality. However, more research is needed to understand the factors involved in task acceptance and completion.

Another important area of research is how to structure tasks to meet the needs of different markets. In Mechanical Turk, for instance, work that takes a long time, incurs high cognitive demands, or has an uncertain payment structure or description tends to be less attractive than short, simple tasks with high certainty of reward, even if the reward is lower overall. Parceling work into short, simple subtasks may be optimal in such markets, while chunking work to avoid the switching costs of people choosing another task to do may be better suited to other markets. Matching the structure of the work to the characteristics of the market could lead to faster throughput and better quality.

Crowdsourcing markets like Mechanical Turk already show enormous potential, but they could go much farther. There are plenty of opportunities for new researchers to help realize these possibilities in fields such as economics, sociology, psychology, computer science, human-computer interaction, and policy. And there has never been a better time to step out of the crowd and get involved.

Biography

Aniket Kittur is an assistant professor in the Human-Computer Interaction Institute at Carnegie Mellon University. He received his PhD in cognitive psychology from University of California-Los Angeles, studying the cognitive processes underlying sensemaking activities such as learning, memory, and insight. His research focuses on harnessing the efforts of many individuals to make sense of information together in ways that exceed individual cognitive capacities in domains including Wikipedia, crowdsourcing markets, and scientific collaboration. His research employs multiple complementary techniques, including experiments, statistical and computational modeling, visualization, data mining, and machine learning.