Crowdsourcing a HIT: Measuring Workers' Pre-Task Interactions on Microtask Markets

Jason T. Jacques and **Per Ola Kristensson**
School of Computer Science
University of St Andrews
St Andrews, Fife, UK
{jtj2, pok}@st-andrews.ac.uk

Abstract

The ability to entice and engage crowd workers to participate in human intelligence tasks (HITs) is critical for many human computation systems and large-scale experiments. While various metrics have been devised to measure and improve the quality of worker output via task designs, effective recruitment of crowd workers is often overlooked. To help us gain a better understanding of crowd recruitment strategies we propose three new metrics for measuring crowd workers' willingness to participate in advertised HITs: conversion rate, conversion rate over time, and nominal conversion rate. We discuss how the conversion rate of workers—the number of potential workers aware of a task that choose to accept the task—can affect the quantity, quality, and validity of any data collected via crowdsourcing. We also contribute a tool—*turkmill*—that enables requesters on Amazon Mechanical Turk to easily measure the conversion rate of HITs. We then present the results of two experiments that demonstrate how conversion rate metrics can be used to evaluate the effect of different HIT designs. We investigate how four HIT design features (value proposition, branding, quality of presentation, and intrinsic motivation) affect conversion rates. Among other things, we find that including a clear value proposition has a strong significant, positive effect on the nominal conversion rate. We also find that crowd workers prefer commercial entities to non-profit or university requesters.

Introduction

Effective crowdsourcing of commercial tasks and data collection for research purposes depends on both the quality of work produced and the number of workers who can be recruited in a reasonable time at an acceptable price (Doan, Ramakrishnan, and Halevy 2011). One of the more popular crowdsourcing and microtask markets is Amazon Mechanical Turk (MTurk). MTurk provides a platform allowing requesters to list Human Intelligence Tasks (HITs) for completion by a large pool of crowd workers in exchange for a monetary reward. MTurk's equivalence to lab studies is a researched area and authors, including Kittur et al. (2008), Heymann and Garcia-Molina (2011), and Paritosh (2012), suggest including verifiable responses to allow an objective method of monitoring participants' in-task behavior, and using a filtering strategy to eliminate suspect results.

In this paper we suggest that a clearer understanding of the wider crowd worker population, including those who choose *not* to participate might provide us with two key benefits. First, it can help us develop more effective crowd worker recruitment strategies and thereby maximize the utilization of the crowd worker pool. Second, it can inform us of self-selection biases among crowd workers and thereby help us carry out more accurate replications of crowd experiments.

Related Work

To better understand how workers approach tasks on MTurk, Rzeszotarski and Kittur (2011) propose a method of comprehensive metadata collection of crowd workers' actions as they complete HITs. This information is then used to predict the quality of work submitted. The work of Rzeszotarski and Kittur (2011) on 'instrumenting the crowd' demonstrates the potential benefit of designing tasks that aid the worker in their completion. Huang et al. (2010) propose a solution for predicting and improving work quality that uses worker feedback to automate task design improvements from their initially seeded design. In order to model output variability as a function of the number of available workers at a given point in time, the authors propose an availability model based on worker time zone using IP-based geo-location of participant data (Huang et al. 2010).

Some work has also been done on understanding how workers search for HITs and how pricing influences workers' engagement (Chandler and Horton 2011; Chilton et al. 2010). Kaufmann et al. (2011) consider worker motivations and show that intrinsic motivation factors are a significant force in worker behavior. Understanding worker motivation is fundamental to creating a clear task which will attract high-quality workers in sufficient quantities. However, while Kaufmann et al. (2011) highlight factors that bring workers to MTurk, without appropriate measures of the scope of the marketplace and available worker populations, the application of these ideas to understand what brings workers to individual HITs can be problematic.

Chilton et al. (2010) track HIT lifetimes by scraping the MTurk HIT search pages. This allows them to create a measure of a HIT's exposure to workers as a function of the position in which it was displayed when ordered by newest first. However, Chilton et al. (2010) acknowledge that workers have the option to change the sort order and this impacts the accuracy of these estimates. Understanding how workers arrive at a HIT provides additional context to evaluations. However, the essentially limitless possibilities for search terms make this a difficult measure to use when evaluating worker interest in a given type of task.

Work by Kanich et al. (2011) makes use of preview statistics as part of their analysis of the potential risks of malware infections for crowd workers. However, they did not consider how preview data might be used to measure the available workforce for a given task. Related, using their comprehensive tool, *Turkalytics*, Heymann and Garcia-Molina (2011) gathered a wide variety of worker interaction data and, as part of a general overview, note an inverse correlation between the number of previews and the time since posting a HIT. In this paper we suggest that this preview data can be used to better understand the *available* worker population. While *Turkalytics* provides comprehensive data collection, the authors note that the solution suffers from a number of technical limitations and is not publically available at the time of writing[1].

Contributions

In this paper we propose *conversion rate*, *conversion rate over time*, and specifically the *nominal conversion rate* as new metrics for measuring crowd worker behavior and interest in specific tasks.

To collect this data we developed *turkmill* as an alternative, non-invasive tool for extracting pre-task interaction data retroactively from the standard log files generated by externally hosted HITs.

Second, we explain how the metrics we propose can help mitigate the problem of crowd latency and maximize the utilization of the crowd worker pool in human computation systems and crowdsourced user studies alike.

Finally, we use our tool and the metrics we propose in this paper to investigate how various aesthetic features and motivational factors affect conversion rates.

Measuring Worker Behavior

For tasks completed using MTurk, requesters are provided a downloadable comma-separated values (CSV) file of the results submitted by workers, allowing simplistic measures of HIT completion rate to be calculated. However, in the same way that a commodity item might sell faster in a city than a village, or a seasonal item might sell faster when in season, MTurk is subject to additional market factors. To understand *which factors* influence the completion rate in context, a more comprehensive understanding of the worker population is required. As of January 2011, Amazon indicated a workforce in excess of 500,000 workers[2], yet many of these workers may never see a particular task. Workers may not be online, available, eligible, or interested in any given HIT.

Webserver log data has been shown to be a useful method for recreating a user's "clickstream" and understanding user interactions with website and web-based applications (Büchner and Mulvenna 1998; Ting et al. 2007). To extract conversion rate data from HITs we developed *turkmill* – named as a portmanteau of MTurk and sawmill, a facility for processing physical wooden *logs*. *Turkmill* is a command line tool that processes standard Apache HTTPd log files and extracts each worker's IP address, worker ID, assignment ID, preview time, accept time, and, optionally, location using IP-based geo-location.

Turkmill works backwards through the log files and uses the *assignmentId* query-string parameter provided by MTurk to identify both the initial preview time, where it is set to *ASSIGNMENT_ID_NOT_AVAILABLE*, and the accept time of a HIT, where it is set to a unique identifying string. To associate the initial, anonymous, preview with the later accept time, *turkmill* uses the IP address to locate the closest pair. Previews that do not have an accept time are written out with the missing data, indicating a worker who chose to preview the HIT but decided not to accept.

Conversion Rate

Using the preview data we can measure the *conversion rate* (*cr*) for a HIT. We define it as the ratio of workers who complete a HIT to all those who were exposed to the HIT. Conversion rate provides us with a scaled metric that includes a measure of the potential worker population. Thereby it enables a direct comparison of features between HITs that may have been completed on different days, at different times, or with different eligibility requirements. While conversion rate is a useful metric for measuring the

[1] http://turkalytics.com

[2] https://forums.aws.amazon.com/thread.jspa?threadID=58891

popularity of a task, it does not consider variations that might occur during a HITs lifetime, nor does it consider the complex task management carried out by the MTurk platform, which we describe later in this paper.

Conversion Rate over Time

Conversion rate, as a scaled measure, allows both comparisons between different HITs, and comparisons within a HIT. By plotting the cumulative conversion rate over time we can visualize how the conversion rates develops. This provides us with an increased insight into external factors influencing the number of workers accepting a task and the probability that they will do so.

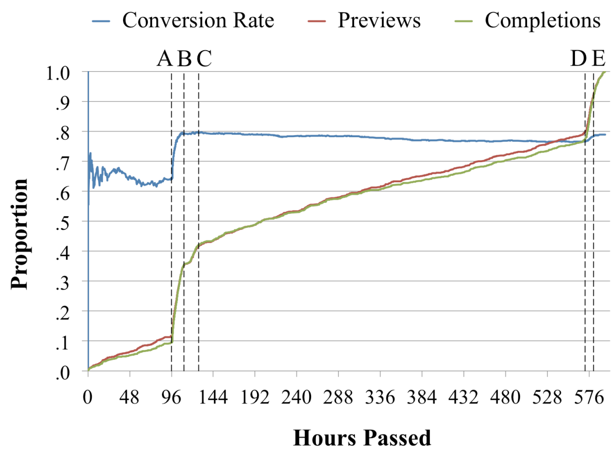

Figure 1. Conversion Rate, Previews, and Completions over time for a long-term demographic study.

As an example of these insights Figure 1 shows the conversion rate, the proportion of previews to date, and the proportion of completions over the first 600 hours (25 days) of a simple demographic survey carried our using MTurk. The survey consisted of 1,985 previews with an overall conversion rate of 78.94%.

The difference seen in regions A-B and B-C of Figure 1 highlight the importance of conversion rate as an indicator of participant interest in the task. While a significant increase in workers *might* correlate with an increased interest, or preference, for a task, this is not always the case. For participation to increase, only the population of available workers available must increase. This can happen if, for example, more workers become available during weekends. However, for an increase in conversion rate the *intent* of those workers must change—they must be primed to accept tasks.

This anomalous behavior is again seen in regions D-E and E-end of Figure 1, however the impact is limited by the much larger dataset at this point, and the cumulative nature of conversion rate. In the case of B-C and D-E, we conjecture this priming might be due to online word-of-mouth discussions between participants on services such as TurkerNation[3] and MTurk Forum[4]. However, while workers are known to communicate “Great HITs” online, we have been unable to identify any worker communication with regard to this specific study at the times identified.

In the intervening period, region C-D, the conversion rate remains relatively stable showing only a slow decline, likely due to a natural aging-out of the task in the MTurk search results.

How MTurk Manages HITs

In addition to the variations that occur over the HIT lifetime, Amazon carries out task management. To ensure only the requested number of assignments are submitted for each task, MTurk counts the number of tasks available. This measure is exposed through their requester API as the element *NumberofAssignmentsAvailable*, part of the HIT’s data structure (Amazon Inc. 2012). This counter decreases by one for each submitted assignment, but also decreases by one for each active[5] preview. Once the available assignments reach zero, the HIT is no longer listed on the MTurk worker interface, and direct links instead indicate that there are no more assignments available. Should a user decide to return the task (drop out), this number then increases again and the task will again be available for other workers.

This simplistic control of assignments in progress, coupled with a reduction in parallel assignment completion, can lead to misleading conversion rates during the end phase of a HIT lifetime if it is calculated as a simple ratio of completions to previews.

Nominal Conversion Rate

To avoid including this noise in the calculation, and to comparably quantify the appeal of a HIT, the measured conversion rate must be representative of the nominal conversion rate (*ncr*) for that HIT. To compensate for the high variability present at the beginning of the task, and to counter for any end-of-HIT effects, we define *ncr* as the mean conversion rate for the interquartile time period of the lifespan of the HIT. For the study shown in Figure 1 the $ncr = 0.782$, lower than the final and aggregate conversion rate, $cr = 0.789$, which has been influenced by the sharp increase in the number of interested workers, but equally, had the study been smaller or not been ongoing, could have been affected by a downward end-of-HIT effect. In other words, *ncr* is a relatively robust estimator of the volatile raw conversion rate curve.

Formally, we define *ncr* as:

[3] http://turkernation.com
[4] http://mturkforum.com
[5] Our informal testing indicates that a preview is “active” for 60 seconds, after which it becomes inactive and the number of available assignments increase.

$$x_{NCR} = \frac{1}{k-j}\sum_{i=j}^{k} x_i\,, \quad t_j - t_0 > \frac{\Delta t}{4}, \quad t_k - t_0 < 3\frac{\Delta t}{4},$$

where x is an array of time-ordered conversion rates, t is a corresponding ordered array of the times for which those conversion rates were calculated, and Δt is the lifespan of the HIT.

Experiment 1: Presentation and Structure

To gain a general understanding of how the *preview* of a HIT affects worker decisions to accept a task we considered a comparison of common presentational features of HITs: layout and appearance, branding, and disclosure of the value proposition. These three variants were compared against a common baseline task.

Method

The baseline task was devised consisting of two demographic questions—age and gender—and eight simple mathematical expressions, two addition, subtraction, multiplication, and division operations, with single digit operands. The task was posted to MTurk with the title “Simple Mathematical Test”, a reward of 10¢ (USD), and a maximum duration of one hour. The baseline task included a value proposition—the amount of work required to receive the reward—and a progress bar (during the question-phase only). The task was also branded with a university logo, and used a multi-page template with a professional appearance. Figure 2 illustrates how the baseline task looked when workers previewed it on MTurk.

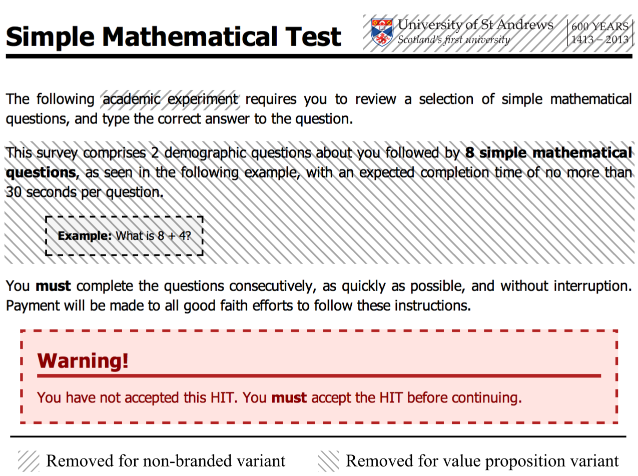

Figure 2. Baseline task presentation in Experiment 1. Striping shows elements removed in variants.

For each of our variants we removed or modified a feature of our idealized HIT. To verify the importance of the value proposition we created a new HIT, which consisted of a modified baseline HIT, excluding any indication of the amount of work to be done. In this variant we removed the example task (indicating difficulty), any reference to the number of questions in the task, and the progress bar (indicating duration).

In our second variant we considered branding. Branding a task provides some measure of endorsement and suggests a level of accountability in the event of negative outcomes for the worker (such as a block, or non-payment). For this variant we created a new HIT that consisted of a modified baseline HIT that excluded any university branding and any other acknowledgement that the task was part of a research project at all.

Our third variant considered the general aesthetics and structure of a task. The aesthetic features of a task may suggest a level of security, professionalism, and trustworthiness to workers previewing and undertaking a task. For this variant we created a new HIT that consisted of a modified baseline HIT that had all style information from the HTML markup removed, with the exception of red coloration on warning text. The multi-page design was collapsed onto a single page, as seen in Figure 3.

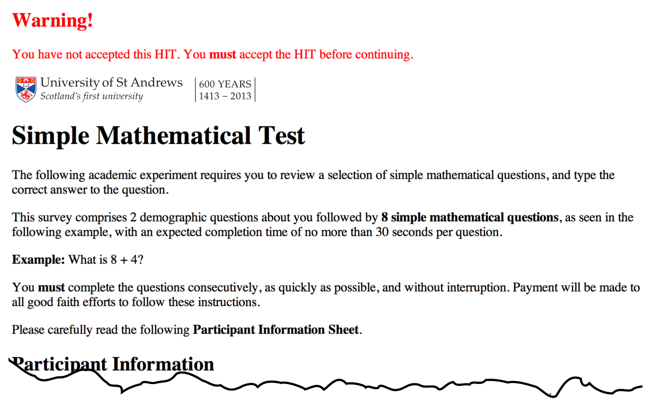

Figure 3. Presentation and structure variant in preview. As a one-page task, the questions followed.

To prevent workers from completing more than one variation of the task, and similar to Mason and Watts (2009), each experiment, comprising its variants, was tested using a single listing on the MTurk requester interface and a simple redirector was used to transparently redirect the content of the HIT to the current variant.

To further validate any differences, and to contrast a common eligibility requirement, we launched the HITs in two distinct microtask market regions: workers resident in the United States and workers resident elsewhere. In total we launched eight HIT designs: four HIT design variants limited to workers resident in the United States, and four HIT design variants that excluded American workers.

To achieve this, two mutually exclusive HITs, requesting 50 assignments, were launched on the MTurk market place, one with the eligibility requirement “Location is US”, and the other “Location is not US”. To ensure all previews would represent a potential worker that was able to accept and complete the HIT, the eligibility requirement was configured to prevent previews from those not eligible to complete the task. After 24 hours had passed, the redirector was set to the next task in the series and the number of assignments requested was topped-up to allow a maximum of 50 submissions for the current variant.

For each microtask region, we set out three *a priori* hypotheses for our experiment:

H_1: The nominal conversion rate will decrease if we remove the value proposition for the HIT.

H_2: The nominal conversion rate will decrease if we remove university branding from the HIT.

H_3: The nominal conversion rate will decrease if we worsen the overall presentation and appearance of the HIT.

Results

In total, 428 previews were recorded of which 376 participants completed the task. To remove the occurrence of workers with multiple accounts participating more than once and potentially skewing the results (Heymann and Garcia-Molina 2011), the data was also filtered by IP addresses, preserving the latest action group (preview, accept, complete) only. The data was sorted by preview time, and any previews occurring after the preview time of the last completed assignment were removed. Previews occurring before the first completion, if any, were preserved.

Over the 24-hour lifetime of each task the "Location is US" HIT recruited 176 participants, 26 for the branding task, 50 for all other tasks. Of the 176 participants only one was removed due to a non-unique IP address. The nominal conversion rate was calculated for each condition: baseline, 87.6%; no value proposition, 56.3%; no branding, 94.5%; poor layout, 86.0%. Statistical significance testing was carried out using the individual conversion rates, from the inter-quartile time period, as used to calculate the *ncr*.

An omnibus ANOVA revealed a statistically significant difference between the conditions ($F_{3,43}$ = 88.748, $\eta_p^2 = .861$, $p < 0.05$). Tukey HSD *post-hoc* pair-wise comparisons showed that the baseline variant resulted in a statistically significant difference compared to the variant with no value proposition and the variant without university branding. The difference between the baseline variant and the layout variant was not statistically significant.

Over the 24-hour lifetime of each task the "Location is not US" HIT recruited 200 participants, 50 for each task variant. Of the 200 participants seven were removed due to non-unique IP addresses. The nominal conversion rate was calculated for each condition: baseline, 86.5%; no value proposition, 76.5%; no branding, 89.6%; poor layout, 93.3%. Statistical significance testing was carried out using the individual conversion rates, from the inter-quartile time period, as used to calculate the *ncr*.

An omnibus ANOVA revealed a statistically significant difference between the conditions ($F_{3,51}$ = 105.596, $\eta_p^2 = .861$, $p < 0.05$). Tukey HSD *post-hoc* pair-wise comparisons showed that the baseline variant resulted in a statistically significant difference compared to all variants: the variant with no value proposition, the variant without university branding, and the poor layout variant.

In summary, the first experiment revealed that including a value proposition significantly increased the nominal conversion rate. The impact of the overall presentation of the HIT was mixed as *ncr* was influenced differently for workers resident inside and outside the US. Finally, unexpectedly, including university branding decreased *ncr*.

Experiment 2: Intrinsic Motivation

To understand the unexpected negative effect of university branding on conversion rate, we carried out another experiment with the goal to better understand workers' intrinsic motivation to carry out a HIT.

Method

The effect of intrinsic motivators on worker performance has previously been investigated by Rogstadius et al. (2011), showing that work quality was improved when intrinsic motivation was higher. Workers were primed with two different introductory messages describing commercial and non-profit sponsors for the task and asked to count blood cells and malaria infections in a computer generated image. A minimal replication of this experiment was conducted to determine the effect of this priming on worker conversion rate when exposed during a HIT preview.

The experiment consisted of three variants of this task, with each worker seeing only one variant managed using the same redirection technique used in Experiment 1. The first variant indicated sponsorship by a fictional commercial organization, the second by a fictional non-profit, and the third without any sponsorship information. Workers were shown a simple introductory message explaining the task and introducing the fictional sponsor, if any. The participants were asked to count the number of blood cells in the image, and the number of malaria parasites. After submitting the form, the true nature of the experiment was disclosed and users were asked to indicate their age and gender. The layout presented to workers can be seen in Figure 4.

Replicating the experiment by Rogstadius et al. (2011), workers were not segregated by region into US and non-US workers. Due to the lower anticipated interest in the task, each variant was available for 48 hours. After 48 hours had passed, the redirector was set to the next task in the series and the number of assignments requested was topped-up to allow a maximum of 50 submissions for the current variant.

For these microtasks, we set out our *a priori* hypothesis for our experiment:

H_4: The nominal conversion rate will increase when intrinsic motivation is increased.

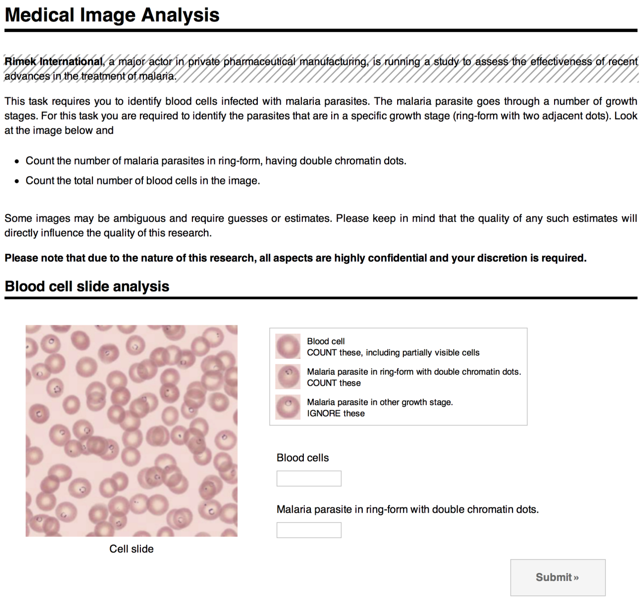

Figure 4. Intrinsic motivation task for Experiment 2. Striping shows text changed or removed in variants.

Results

Over the 48-hour lifetime of each variant, we collected 42 participants for the non-profit variant, 25 for the commercial variant, and 18 for the unsponsored variant. In addition, the variants collected an additional 107, 29, and 50 previews respectively. The nominal conversion rate was calculated for each condition: the non-commercial task, 18.7%; the unsponsored task, 31.7%; the commercial task, 48.9%. Statistical significance testing was carried out using the individual conversion rates, from the inter-quartile time period, as used to calculate the *ncr*.

An omnibus ANOVA revealed a statistically significant difference between the conditions ($F_{3,134}$ = 2673.055, η_p^2 = .976, $p < 0.05$). Tukey HSD *post-hoc* pair-wise comparisons showed that the non-profit variant resulted in a statistically significant difference compared to both the variant with no sponsorship information and the variant indicating a commercial sponsor.

Discussion

As expected, both US and non-US workers showed a significant reduction in conversion rate for tasks which did not indicate their value proposition. Without accurate knowledge of the effort required, workers appear reluctant to begin a task. The inclusion of some measure of the effort required to receive the reward is clearly essential, however attempting to use the maximum duration to indicate the value proposition would seem difficult as providing sufficient time for slower workers is essential to maximize worker up-take. Providing a clear statement of the task can be used to encourage worker interaction, especially so with American workers.

The importance of presentation is unclear. While US workers showed no significant preference between the baseline task and the one-page layout, a clear preference for a one-page layout was noted for non-US workers. While initially this might appear to be due to a further exposure of the value proposition, non-US workers placed less importance on this feature than their American peers. For this reason, we conjecture that this may be a due to different familiarity with a particular presentation style.

Contrary to our expectations, workers showed a preference for our unbranded task. With our initial expectation that a recognizable brand or institution would lend the benefits of their prestige and surety, an inverse relationship between branding and conversion rate was unexpected. To our knowledge, the authors are presently the only researchers using MTurk for crowdsourcing at their institution, and at the time of writing, maintain a 100% rating in all categories on the requester-ranking site, Turkopticon[6].

Our second experiment further investigated this anomaly, and demonstrated a clear preference for workers to distance themselves from non-commercial entities. When choosing a task, workers were more likely to complete an unsponsored variant over the non-profit variant, and preferred a commercially sponsored variant over both. This preference for a commercial institution suggests that intrinsic motivators are less important than the implications of our priming text for receiving the remuneration, namely that a commercial organization may be more reliable than a non-profit with regard to their obligation to pay.

The lower overall worker turnout for the motivation experiment (85 in 144 hours, compared to 376 in 96 hours) may be attributed to the more complex nature of the intrinsic motivation task. The HIT required an extended effort to carefully identify and count the individual units in the image, and this, combined with the medical nature of the task, may have lead to self-selection not present in our first experiment with its simpler mathematical task.

Application of Conversion Rate Metrics

As a measure of worker up-take, conversion rate indicates how many of the workers aware of a HIT choose to accept and complete a task. For research studies, improving the conversion rate has an effect on both sample quality and,

[6] http://turkopticon.differenceengines.com

potentially, quantity. The ability to monitor worker intent, and potentially influence it, can allow for larger, more representative samples to be collected. Reducing the self-selection of workers can lead to higher quality data and studies (Paritosh 2012). Improved understanding of these factors can result in a more representative sample of workers, allowing more confidence in generalizations and fewer unknown and uncontrolled confounds. Including conversion rate metrics in studies may also improve replicability.

For commercial tasks, the time taken to complete all assignments in a HIT can be an important factor in the value of work produced. Many common tasks, such as audio transcription, may only have value to the requester if completed before the information being transcribed is no longer current. Maximizing utilization of the available worker pool during a task minimizes the time taken to complete all component assignments of the HIT, regardless of the current availability of workers.

Complete human computation systems, which rely on workers to interpret and solve computationally difficult problems, are an intense area of research for practical applications of crowdsourcing. However, due to the unreliable nature of crowdsourcing, *crowd latency* is often a problem (Bernstein et al. 2010; Bernstein et al. 2011; Franklin et al. 2011; Gingold, Shamir, and Cohen-Or 2012).

For example, *CrowdDB* is a crowdsourced extension to the SQL data access language that allows human computation resources to be accessed through a familiar interface (Franklin et al. 2011). However, such a system has a huge reliance on rapid turnaround times for queries if they are to be of practical real-world use. Franklin et al. (2011) highlight their difficulty in recruiting workers to these tasks, citing an insufficient worker pool and workers wary of new requesters. Without actual conversion rate metrics to provide an understanding of the worker pool, an accurate measure of precisely *what* is limiting the availability of workers, and why workers appear to be reluctant to accept tasks, cannot be meaningfully quantified.

While conversion rates can allow *post-hoc* analysis of user interface elements, measuring conversion rates and worker population estimates in real time can also allow crowdsourcing systems to *dynamically price tasks* to meet requester requirements, such as speed of task execution, or executing the task in the most cost efficient manner. This type of optimization would be transferable to other crowdsourced processing systems which experience significant issues with task turn-around time, and other algorithms with crowd latency issues (Bernstein et al. 2010; Gingold, Shamir, and Cohen-Or 2012).

Limitations and Future Work

Turkmill extracts pre-task interaction data from standard Apache log files, however this approach requires that HITs are hosted on an external webserver on which the log files can be accessed. This type of task requires that the HIT is launched using the MTurk API or the MTurk command line tools. While this allows for non-invasive and even retroactive data analysis, this approach requires some level of technical expertise in HIT implementation. This limitation is not endemic to the data itself, and Heymann and Garcia-Molina (2011) demonstrate that preview information can be captured using a more technically complex, but reusable, client-side framework allowing data to be captured from a wider variety of HIT implementations, and for these HITs to be created using the MTurk web-based requester interface.

Turkmill also suffers from its own technical limitations and is limited by the IP addresses listed in the standard Apache log format. For interaction that occurs before MTurk reveals the *workerId*, *turkmill* must use the IP address as a unique identifier to collate the separate actions into a collective click-stream. To distinguish concurrent workers *turkmill* uses the IP address combined with a nearest-neighbor approach to collate these individual events, and this may lead to events in overlapping click-streams from the same IP address being misattributed to other users from the same location. *Turkmill* supports IPv6 addressing for webserver logs including these addresses. With the expected gradual growth of IPv6 connectivity and the decline of shared IP addresses for originating Internet traffic, this type of issue is expected to become less frequent.

Finally, as a new metric for crowdsourced studies, this paper has made only the broadest comparisons between very narrowly defined HIT features. The unexpected, inconsistent importance of presentation style for US and non-US workers would certainly benefit from further exploration. A more complete investigation of the different expectations between worker populations may provide a fruitful avenue for future research. Similarly, the unexpected preference of workers to avoid non-commercial HITs, and the somewhat concerning implications, is certainly an area where increased study and further confirmation is necessary before we can draw definitive conclusions.

Conclusion

To maximize the available workforce in online labor markets, such as MTurk, it is essential to understand what features make a task attractive to workers. This paper has proposed three new standard metrics for crowdsourced studies: conversion rate, conversion rate over time, and nominal conversion rate. Conversion rate metrics help nuance and contextualize crowdsourcing studies and provide us with a basic understanding of the population under study. Using these metrics we carried out two experiments to better understand worker engagement. We found that including a clear value proposition has a strong significant, positive effect on the nominal conversion rate. We also found that crowd workers prefer commercial entities to non-profit or university requesters. For experimental research, conversion rate metrics help nuance our understanding of the sample recruited for study. For human computation systems, conversion rates can help identify design features that affect crowd latency, performance, and cost.

To help other researchers, we will be releasing our tool *turkmill* under the MIT open source license and it will be available from: http://www.cs.st-and.ac.uk/~jtj2/turkmill

Acknowledgments

This work was supported by a studentship from the Engineering and Physical Sciences Research Council. We also thank the Scottish Informatics and Computer Science Alliance for additional support.

References

Amazon Inc. 2012. *Amazon Mechanical Turk API Reference*.

Bernstein, Michael S., Joel Brandt, Robert C. Miller, and David R. Karger. 2011. Crowds in Two Seconds: Enabling Realtime Crowd-powered Interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, 33–42. UIST '11. New York, NY, USA: ACM.

Bernstein, Michael S., Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a Word Processor with a Crowd Inside. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, 313–322. UIST '10. New York, NY, USA: ACM.

Büchner, A. G., and M. D. Mulvenna. 1998. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining. *ACM Sigmod Record* 27 (4): 54–61.

Chandler, D., and J. Horton. 2011. Labor Allocation in Paid Crowdsourcing: Experimental Evidence on Positioning, Nudges and Prices. In *Human Computation, Volume WS-11-11 of Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Chilton, Lydia B., John J. Horton, Robert C. Miller, and Shiri Azenkot. 2010. Task Search in a Human Computation Market. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 1–9. HCOMP '10. New York, NY, USA: ACM.

Doan, Anhai, Raghu Ramakrishnan, and Alon Y. Halevy. 2011. Crowdsourcing Systems on the World-Wide Web. *Communications of the ACM* 54 (4) (April 1): 86.

Franklin, M. J., D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. 2011. CrowdDB: Answering Queries with Crowdsourcing. In *SIGMOD Conference*, 61–72.

Gingold, Yotam, Ariel Shamir, and Daniel Cohen-Or. 2012. Micro Perceptual Human Computation for Visual Tasks. *ACM Trans. Graph.* 31 (5) (September): 119:1–119:12.

Heymann, Paul, and Hector Garcia-Molina. 2011. Turkalytics: Analytics for Human Computation. In *Proceedings of the 20th International World Wide Web Conference*, 477–486. WWW '11. New York, NY, USA: ACM.

Huang, Eric, Haoqi Zhang, David C. Parkes, Krzysztof Z. Gajos, and Yiling Chen. 2010. Toward Automatic Task Design: a Progress Report. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 77–85. HCOMP '10. New York, NY, USA: ACM.

Kanich, C., S. Checkoway, and K. Mowery. 2011. Putting Out a Hit: Crowdsourcing Malware Installs. In *Proceedings of the 5th USENIX Workshop on Offensive Technologies*.

Kaufmann, N., T. Schulze, and D. Veit. 2011. More Than Fun and Money. Worker Motivation in Crowdsourcing–a Study on Mechanical Turk. In *Proceedings of the Seventeenth Americas Conference on Information Systems, Detroit, MI*.

Kittur, Aniket, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 453–456. CHI '08. New York, NY, USA: ACM.

Mason, Winter, and Duncan J. Watts. 2009. Financial Incentives and the "Performance of Crowds." In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 77–85. HCOMP '09. New York, NY, USA: ACM.

Paritosh, P. 2012. Human Computation Must Be Reproducible. In *CrowdSearch 2012 Workshop at WWW 2012*.

Rogstadius, J., V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic. 2011. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. In *Proc. of AAAI Conference on Weblogs and Social Media*, 11:321–328.

Rzeszotarski, Jeffrey M., and Aniket Kittur. 2011. Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, 13–22. UIST '11. New York, NY, USA: ACM.

Ting, I, Lillian Clark, Chris Kimble, Daniel Kudenko, and Peter Wright. 2007. APD-A Tool for Identifying Behavioural Patterns Automatically from Clickstream Data. In *Knowledge-Based Intelligent Information and Engineering Systems*, edited by Bruno Apolloni, Robert Howlett, and Lakhmi Jain, 4693:66–73. KES '07. Springer.