A direct empirical investigation of the determinants of online labour supply in Amazon Mechanical Turk[*]

Jean-Michel Dalle
UPMC & CRG-CNRS and Ecole Polytechnique

Thomas Lacroix
Supélec & Master IREN

Mathieu Lacage
Alcméon

Matthijs den Besten
Montpellier Business School, Labex Entreprendre & CRG-CNRS and Ecole Polytechnique

1. Introduction and overview

The part of the crowdsourcing phenomenon that relies on more structured labour markets, for which “Amazon Mechanical Turk” (here, AMT) is a leading example, raises many issues, notably political, ethical and economic (e.g. Kittur et al., 2013). Yet, besides its characterization by Ipeirotis and followers (Ipeirotis, 2010), and despite a burgeoning literature, a direct empirical investigation of the propensity of online workers to select among available projects, and thus of many of the determinants of “online work supply” in AMT, is still largely missing, at least as far as we know.

We try to address this issue in this paper with a dataset gathered by crawling AMT’s website every 3 minutes during a period of about 2 months, which allows us to investigate temporal data on several hundreds of real AMT projects (section 2), and hence to measure the mean speeds at which individual tasks disappear (are “executed”) from available projects. In this context, we investigate the “problem of problem choice”, as we had suggested to name it after C.S. Pierce in the context of scientific communities (Carayol & Dalle, 2007), i.e., we inquire about the *determinants of turkers’ choice among projects* – among the many problems that are offered to them on AMT. For, the global allocation of online efforts in online communities (Dalle & David, 2005; den Besten, Dalle et Galia, 2008) actually *results* from the aggregation of all of the workers’ individual choices among available problems. In AMT, compared to other communities such as open-source software or Wikipedia where coordination between workers is instrumental (den Besten & Dalle, 2014; Rossi et al., 2010), choices are affected by the pricing of tasks and and by other characteristics of tasks and projects, even if there is coordination among workers on dedicated forums.

In section 3, we present preliminary evidence according to which the pricing of individual tasks does not seem to influence workers choice, at least directly, contrary to volume i.e. to the number of individual tasks in a given AMT project or HIT Group. This observation is consistent with a population of workers who would maximize their wages by increasing their productivity over time through the selection of selecting groups of tasks on which they could focus and specialize for a sufficiently long amount of time, and coherent with Franklin et al

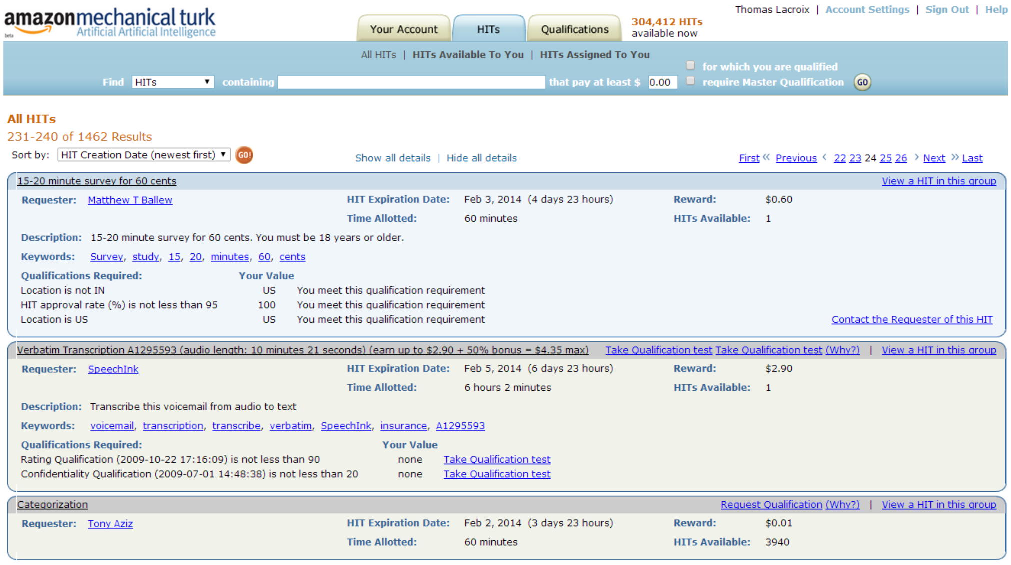
[*] The support of ANR grant OCKTOPUS is very gratefully acknowledged.

(2011). We further show that the length of task *descriptions* on AMT's website increases the speed at which they are executed, which could correspond either to a preference from workers for more detailed descriptions when choosing among available tasks, or to the fact that tasks that are "better thought through", both in their description and in the process leading to their execution, are able to attract and to retain workers more. We next present, in section 4, an experiment on AMT that suggests that AMT workers could be sensitive to price signals with respect to their problem of problem choice through the assessment of the *difficulty* of the tasks that they could execute via the price that has been set for those tasks. This finding appears compatible with our former observations, since assessing the difficulty of tasks through price is coherent with the strategy of workers who would seek to maximize their productivity by focusing on relatively easy and well-defined tasks, and with Yan et al. (2010)'s finding that low-priced tasks would be addressed more rapidly. Section 5 concludes and suggests limits for our work and the need for further investigations.

2. Dataset

Data collection

An elementary task on AMT is called a HIT, for "Human Intelligence Task". Requesters post HITs on AMT: similar HITs are grouped within "HIT groups" below the same title, same description, same reward, etc. Workers then are able to carry out HITs in exchange for the associated monetary reward. They can choose on which HIT groups to work from the list available, along with characteristics of each HIT group, on AMT website (see example on Figure 1).


amazonmechanical turk
Artificial Artificial Intelligence
beta
Thomas Lacroix | Account Settings | Sign Out | Help
Your Account
HITs
Qualifications
304,412 HITs
available now
All HITs | HITs Available To You | HITs Assigned To You
Find
HITs
containing
that pay at least $
0.00
for which you are qualified
require Master Qualification
GO
All HITs
231-240 of 1462 Results
Sort by:
HIT Creation Date (newest first)
GO!
Show all details | Hide all details
First « Previous ‹ 22 23 24 25 26 › Next » Last

15-20 minute survey for 60 cents | View a HIT in this group

Requester: Matthew T Ballew | HIT Expiration Date: Feb 3, 2014 (4 days 23 hours) | Reward: $0.60
Time Allotted: 60 minutes | HITs Available: 1

Description: 15-20 minute survey for 60 cents. You must be 18 years or older.

Keywords: Survey, study, 15, 20, minutes, 60, cents

Qualifications Required:	Your Value	
Location is not IN	US	You meet this qualification requirement
HIT approval rate (%) is not less than 95	100	You meet this qualification requirement
Location is US	US	You meet this qualification requirement

Contact the Requester of this HIT

Verbatim Transcription A1295593 (audio length: 10 minutes 21 seconds) (earn up to $2.90 + 50% bonus = $4.35 max) | Take Qualification test Take Qualification test (Why?) | View a HIT in this group

Requester: SpeechInk | HIT Expiration Date: Feb 5, 2014 (6 days 23 hours) | Reward: $2.90
Time Allotted: 6 hours 2 minutes | HITs Available: 1

Description: Transcribe this voicemail from audio to text

Keywords: voicemail, transcription, transcribe, verbatim, SpeechInk, insurance, A1295593

Qualifications Required:	Your Value	
Rating Qualification (2009-10-22 17:16:09) is not less than 90	none	Take Qualification test
Confidentiality Qualification (2009-07-01 14:48:38) is not less than 20	none	Take Qualification test

Categorization | Request Qualification (Why?) | View a HIT in this group

Requester: Tony Aziz | HIT Expiration Date: Feb 2, 2014 (3 days 23 hours) | Reward: $0.01
Time Allotted: 60 minutes | HITs Available: 3940

Figure 1 - AMT HIT list

We parsed the HIT Group list on AMT web site every 3 minutes from November 20th, 2013 until February 10th, 2014, in order to gather the information on available HIT groups and their current level of completion. This list was parsed in descending order of the number of HITs still available within each HIT group. Each time, we collected the following information for the 200 largest HIT groups:

- HIT id
- Title: short description of HITs within the HIT group
- Requester id

- Requester name
- HIT Expiration Date: the date when HITs will be removed
- Reward: the reward in dollars for successfully completing a single HIT
- Time allotted: how long a Worker can hold on to a HIT
- Number of HITs Available: the number of HITs currently available in the HIT Group
- Description: a more precise description of the HIT
- Keywords: classification of HITs to facilitate search
- Qualifications Required: the qualifications the worker need to satisfy to be allowed to execute the HITs

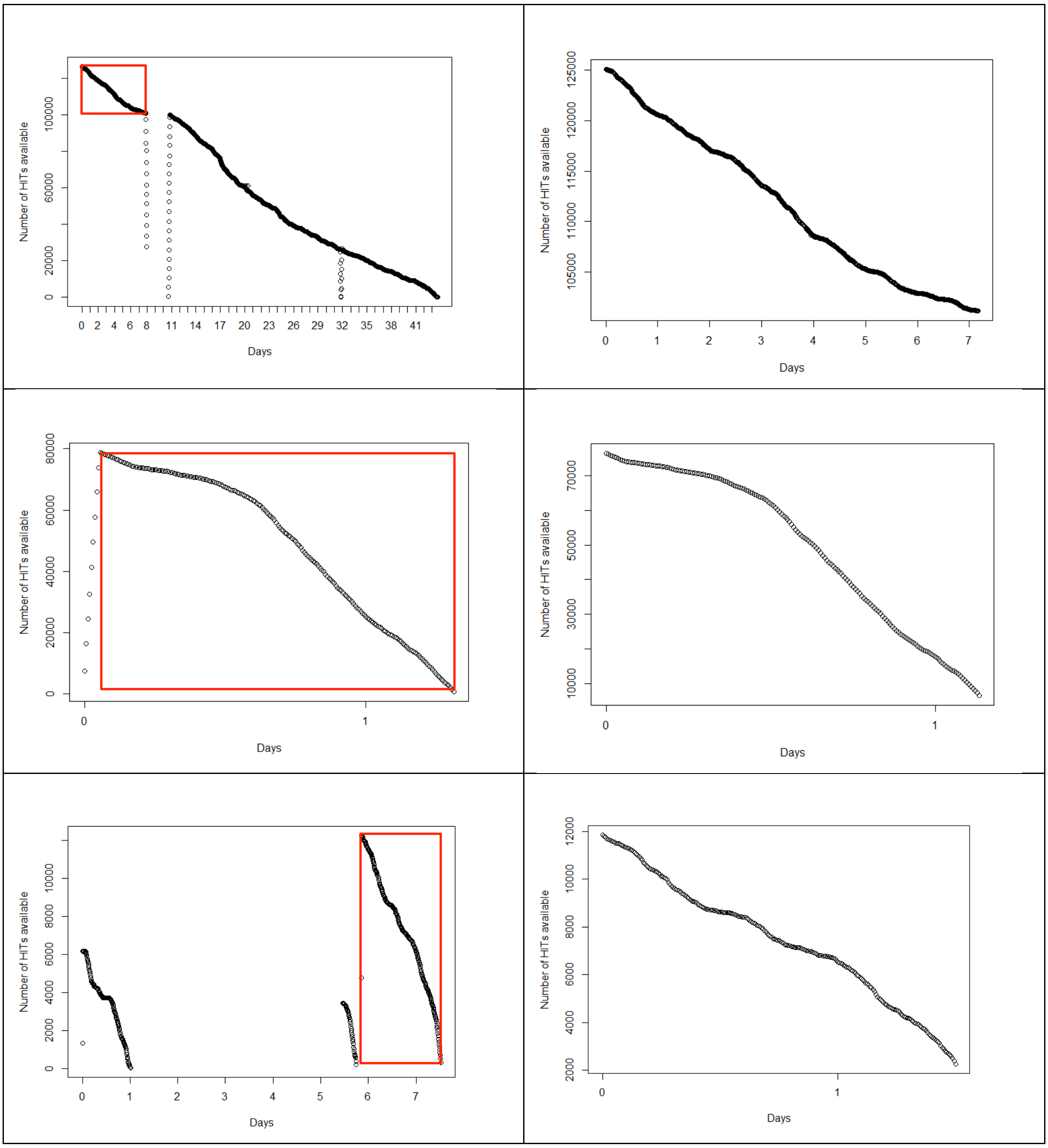


Figure 2 – "Best part" algorithm: 3 examples

In order to study a coherent population of workers and to exclude newcomers, we decided to focus our study on HIT groups that requested the *'Masters have been granted'* qualification from workers. This qualification is granted directly by AMT to recurrent AMT workers[1], and thus available to all Requesters and relatively widely used even if it implies that Requesters pay a significantly higher overhead (30% instead of 10%) to Amazon. Although its precise specifications have not been made publicly available by Amazon, there seems to be a agreement among AMT workers that it should correspond to workers who would have completed many, very probably thousands of jobs, and whose approval rate (approval of their work by who has requested it) would be very high and certainly above 95% if not 99%. In total, we gathered information about 3228 Hit Groups using this procedure.

"Best part" algorithm

For the present study, we limited our dataset to HIT Groups for which we had more than 200 data points (i.e. more than 3 hours of activity). Furthermore, we developed a "best part" algorithm in order to clean the dataset, most notably to get rid of gaps in time series. For each given HIT group, we determine when it had the maximum number of HITs available, which becomes the first point of our cleaned-up subsequence; and we subsequently define the last point of this subsequence as the last point after the initial one until which there was no gap in the data and until which the number of HITs was never increasing (in order to exclude situations where HIT groups were "reloaded" with extra HITs while they were still active). Figure 2 presents 3 distinct examples of this data improvement procedure, where the "best part" kept for future analysis is marked by a red rectangle on the left-hand side pictures, of which the red-hand side pictures present zooms.

Circadian cycles

In addition, we conducted a spectrum analysis on our dataset. Figure 3 presents an example of the first order difference and the periodogram for the "best part" of the largest (measured in max number of HITs) in our dataset. It shows a clear circadian (24 hours) cycle (frequency of the highest peak on the right-hand side figure of Figure 3). A circadian cycle was also observable for several HIT groups in our dataset whereas for most others, either the limited time frame or other factors prevented its observation.

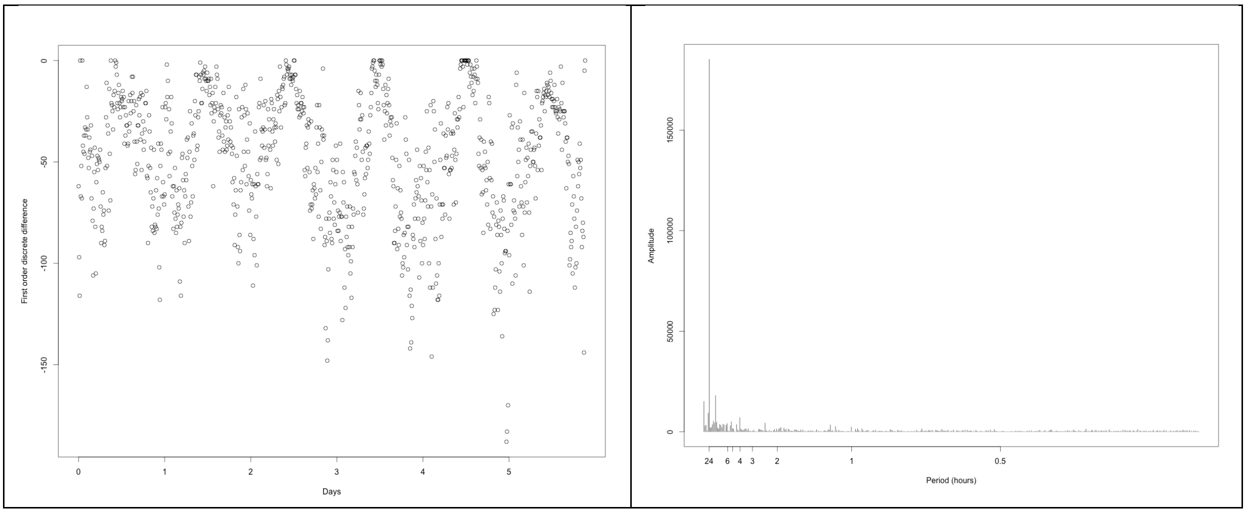


Figure 3 - First order discrete difference (left) and periodogram (right) of largest HIT group in our dataset

[1] http://mechanicalturk.typepad.com/blog/2011/06/get-better-results-with-less-effort-with-mechanical-turk-masters-.html

Mean speed estimation

In order to observe the intensity of work supply on each HIT Group, while also reducing the impact of circadian cycles, we focus here on measuring of the mean speed at which individual HITs are executed: that is to say, we compute the difference between the maximum number of HITs available and the minimum number of HITs available in each *cleaned* subsequence using our “Best Part” algorithm, and divide it by its duration. This approach differs and is complementary to the one of Ipeirotis (2010) and Wang et al. (2011), who make use of survival analysis, in that it does not focus on completion time and therefore is not affected by transient condition that affect completion near completion, or else the addition of new HITs in an existing HIT group, but rather on a factor that could be a better proxy for the attractivity of online tasks vis-à-vis workers. According to this investigation strategy, we removed 18 outliers from our dataset for which mean speed was not observed properly or whose value why dubious or idiosyncratic.

3. Factors affecting the supply of online labour

Figures 4a, b & c present the mean speed for each HIT Group plotted against HIT price, maximum number of HITs in the HIT group, and the logarithm of the latter. No clear pattern is visible on Figure 4a whereas Figures 4b and 4c suggest some level of correlation.

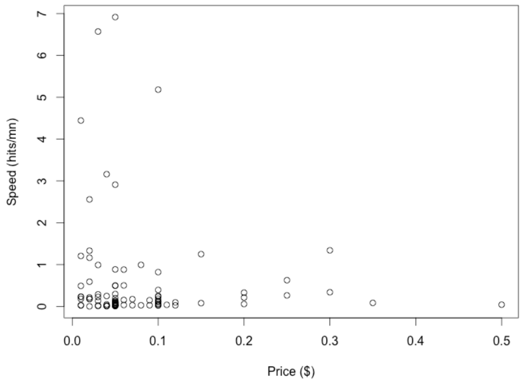


Figure 4a – Mean speed against HIT price

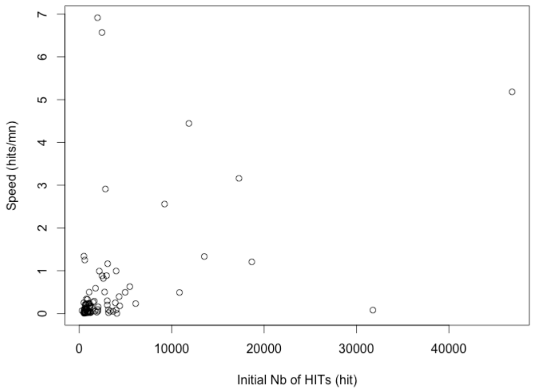


Figure 4b – Mean speed against maximum number of HITs

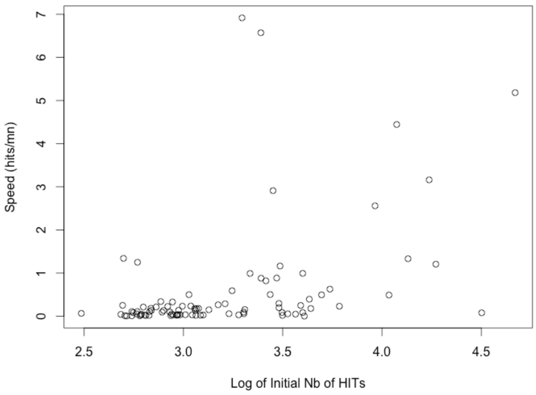


Figure 4c – Mean speed against log of maximum number of HITs

Table 1 presents several linear models with mean speed as the dependent variable. Independent variables include and estimation of the hourly wage, computed by dividing for each HIT group the individual reward of a HIT by the suggested "allotted time" for the execution of tasks in a given HIT group, evaluated in hours. These results suggest that the choices between alternative projects do not seem to be affected directly by individual task price. That said, results with regard to hourly wages unfortunately do not seem relevant, since the corresponding field in the description of tasks seems to have been filled by Requesters according to a "maximum time" rule and does not seem to be a good proxy of how long each individual tasks could take: typically, a significant proportion of requesters select 1 hour probably as a "default" strategy in this respect.

On the very contrary, workers' choices, and thus the supply of online work, appear elastic with respect to the size of HIT groups. We interpret this finding as suggesting that more workers select large HIT groups, and possibly also that workers could select HIT groups on which they can focus for several hours or days, potentially improving their productivity and thus their hourly wages.

Model	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Price	-1.5				-1.2	-0.3		
Hourly wage		-0.13					-0.14	-0.06
Size			8.5e-05***		8.5e-05***		8.6e-05***	
Log(Size)				1.25***		1.24***		1.24***
R^2	*0.009008*	*0.006114*	*0.1898*	*0.1897*	*0.1958*	*0.1901*	*0.1963*	*0.191*

Table 1 – Factors affecting online labour supply – OLS (*: < 0.001 significance level)**

Model	*1*	*2*	*3*	*4*
Price	-1.0	-0.1	-1.7	-0.8
Size	8.7e-05***		8.2e-05***	
Log(Size)		1.25***		1.19***
Length(Title)	-7.7e-03	-6.1e-03		
Length(Desc)			5.3e-03***	5.3e-03***
R^2	*0.2073*	*0.1974*	*0.3053*	*0.3*

Table 2 – Further factors affecting online labour supply – OLS (*: < 0.001 significance level)**

Table 2 presents several further linear models, still with mean speed as the dependent variable, that add the length (number of characters used) of the "title" and the "description" fields that appear in the description of HIT Groups (and are inserted by Requesters) to the list of dependent variables, in addition to price and volume. These models show that the length of the description, and not the length of the title, does affect positively the speed at which tasks are completed. This finding could be explained either by the fact that workers would select their projects based on the precision of their description, or by the fact that Requesters who are able to describe their tasks in more details would also present tasks to workers that would be more appealing because they would typically be better thought.

In order to investigate these issues further, we searched our dataset for occurrences where the same HIT Group (same title, description and Requester, same qualification for workers) had been posted several times with different prices for individual tasks. We found several hundred such situations: focusing further, in order to improve the consistency of our data, on HIT Groups that had been continuously observed at least 10 successive periods of time (1/2 hour), for which we could compute a non-null mean speed, we were left with 182 sequential pairs of observations. Removing further 33 outliers with either a very high difference in mean speed or a very high added or removed volume of individual hits, figure 5a and 5b present mean speed variation depending on price variation (5a) and volume variation (5b). Table 3 then confirms visual impressions using several linear models based on this data, with here the difference of mean speeds as the dependent variable and, as dependent variables, the difference of price, in absolute numbers and as a ratio and the difference in volume. Unsurprisingly, variables measuring the lengths of the "title" and "description" fields were never significant when added in these models. Nor was the variation in price significant when we limited our dataset to successive pairs with positive price variation or to relatively small positive variations.

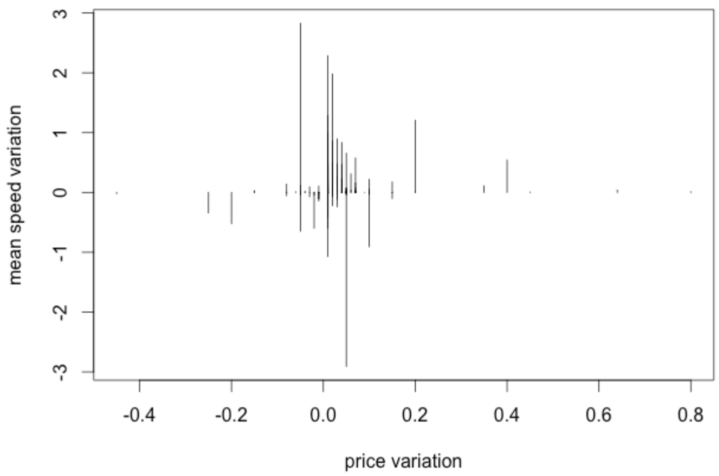


Figure 5a – Mean speed variation against HIT price variation

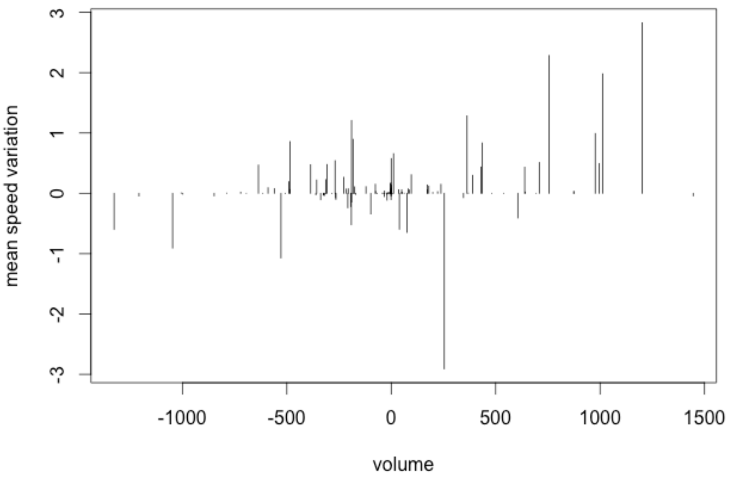


Figure 5b – Mean speed variation against variation of volume

Model	*1*	*2*	*3*
DeltaPrice	0.21		0.13
PriceRatio		1.32e-02	
DeltaSize	4.1e-04***	4.1e-04***	
SizeRatio	-7.7e-03		7.5e-02*
R^2	*0.1142*	*0.1125*	*0.0285*

Table 3 – Further factors affecting mean speed variation – OLS (*: <0.05 & *: < 0.001 significance levels)**

Altogether, these results point towards a general attractivity of HIT Groups mostly relying on their size and quality, be it the quality of their description or their intrinsic (with respect to how workers can process HITs) quality. It wouldn't be correct to say that price as such does not matter, even if the fact that changes in the pricing of individual tasks do not statistically impact the mean speed at which HITs get worked out is puzzling, since our results are coherent with the fact that workers would try to maximize their wages through selecting easier to address tasks on which they could focus for a longer period of time. That said, they definitely call for more investigations with respect to how the pricing of individual tasks might affect the choices and decisions of workers, all the more so since our results are in this respect not aligned with previous others and notably with Mason & Watts (2011), whose experiment could therefore have suffered from a selection bias, i.e. self-selecting workers with a strong sensitivity to price.

4. Pricing influences workers' assessment of task difficulty

In this respect, we conducted an experiment on AMT by creating a HIT group named "HIT difficulty sentiment", where individual HITs asked workers (with *'Masters had been granted'* qualifications) to evaluate the "Estimated difficulty of the HIT, willingness to do it", choices being between Very Difficult / Difficult / Normal / Easy / Very Easy. Workers were given the following information on HITs, the list of which was composed of 109 HIT groups in our dataset:

- Time: Time allotted to the HIT (in minutes)
- Title: Title of the HIT
- Desc: Description of the HIT
- Price: The reward for the completion of the HIT (see below)

The reward was \$0.02 for rating 5 items. We executed this experiment in 2 batches on April 6th, 2014. In one of the two batches, we hid the price. 5 different workers evaluated each HIT group. Figure 6a and 6b present the average of these 5 estimations for each HIT group and for the two batches (without and with price).

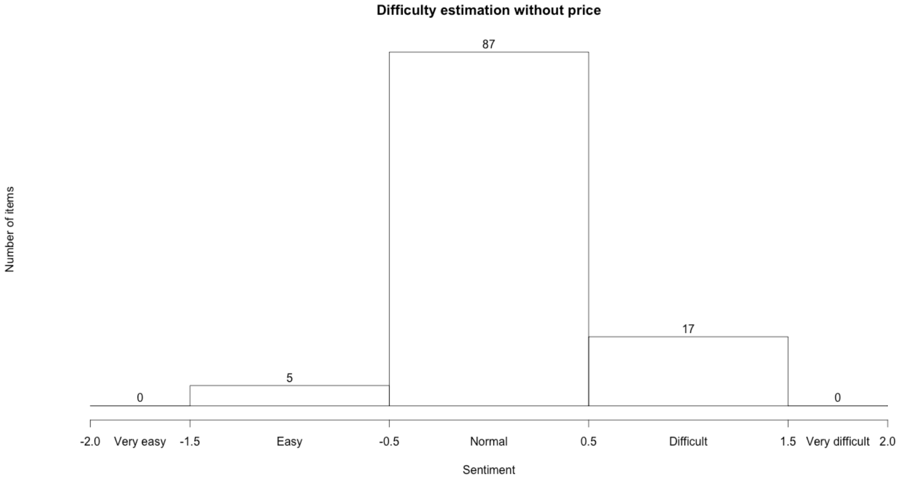


Figure 6a: workers evaluate difficulty without any information on pricing

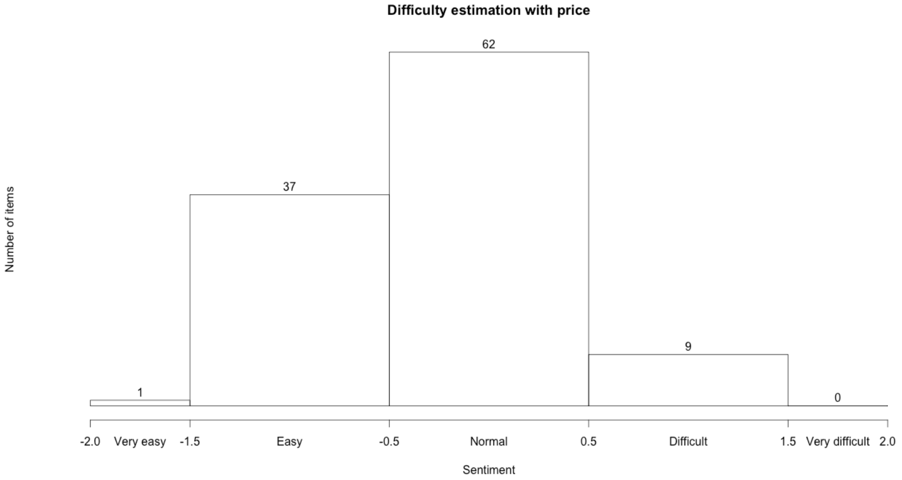


Figure 6b: workers are given price information when evaluating difficulty

When workers do not have access to information about the price of individual tasks, many of the HITs are estimated to have a normal difficulty, a few are deemed difficult (17) and only 5 are considered easy. On the contrary, when they have access to pricing information, many more tasks (37) are classified as "easy" and fewer (9) as difficult. The correlation between estimated difficulty and price (Figure 6a) is almost null (0.006) and non significant in the first batch, and estimated at 0.19 and significant at the 5% level in the second (Figure 6b).

These results suggest that AMT workers use the pricing of individual tasks as a signal in order to assess the difficulty of tasks when choosing among HIT groups, particularly low prices in order to assess easiness. As a consequence, workers might be collectively driven towards high volume but low price tasks since they would select HIT Groups in terms of their expected wages, and since low price would signal easiness, in a sense "like" a more detailed description. This interpretation could explain why we do not find any elasticity of work supply in our dataset with respect to the pricing of individual tasks, even when this pricing changes for the same tasks.

5. Conclusion and future work

The preliminary results presented here warrant further investigations on a larger dataset and on a longer period of time, with a larger exploration of the impact of qualifications requested from workers since the use of "Masters" in our study might have introduced some bias. It would also be helpful to design experiments allowing for a direct observation of how much time workers spend on individual tasks, and of whether this time tends to decrease over time as they specialize on a given HIT Group. Furthermore, investigation of the problem of problem choice in online labour markets such as AMT might also benefit from experiments that would directly present choices to workers. These experiments should be designed in order to avoid selection biases, as it is now clear that the population of AMT workers is heterogeneous and that the design of experiments might have a significant effect on workers that self-select to perform given tasks.

Provided the results presented here would not be infirmed by such further studies, they would suggest that AMT as an online labour markets is largely driven towards highly repetitive and simple tasks, perhaps because AMT itself, compared maybe to other platforms, would have retained a biased sample of workers based on the nature of HITs proposed i.e. in a "path-dependent" and "two-sided" manner. It would then be left to researchers to inquire whether this phenomenon is specific to AMT or whether a "crowding" of interesting work would be generally associated with crowdsourcing platforms functioning as online labour markets with price systems, compared typically to non-monetary crowdsourcing, since the "nature of the crowd", so-to-say, would depend on the nature of the tasks it is offered.

6. Bibliography

Carayol, N., & Dalle, J. M. (2007). Sequential problem choice and the reward system in Open Science. Structural Change and Economic Dynamics, 18(2), 167-191.

Dalle, J. M., & David, P. A. (2005). The allocation of software development resources in 'open source' production mode, In Joe Feller, Brian Fitzgerald, Scott Hissam, Karim Lakhani, eds., Perspectives on Free and Open-Source Software, MIT Press, pp. 297-307.

den Besten, M., & Dalle, J. M., (2014). Coordination by reassignment in the Firefox community, Proceedings of the European Conference on Information Systems (ECIS) 2014, Tel Aviv, Israel.

den Besten, M., Dalle, J. M., & Galia, F. (2008). The allocation of collaborative efforts in open-source software. Information Economics and Policy, 20(4), 316-322.

Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S., & Xin, R. (2011). CrowdDB: answering queries with crowdsourcing. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (pp. 61-72). ACM.

Hossain, M. (2012, May). Users' motivation to participate in online crowdsourcing platforms. In Innovation Management and Technology Research (ICIMTR), 2012 International Conference on (pp. 310-315). IEEE.

Ipeirotis, P. G. (2010). Analyzing the amazon mechanical turk marketplace. XRDS: Crossroads, The ACM Magazine for Students, 17(2), 16-21.

Kaufmann, N., Schulze, T., & Veit, D. (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, Michigan August 4th-7th.

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., ... & Horton, J. (2013). The future of crowd work. In Proceedings of the 2013 conference on Computer supported cooperative work (pp. 1301-1318). ACM.

Mason, W., & Watts, D. J. (2010). Financial incentives and the performance of crowds. ACM SigKDD Explorations Newsletter, 11(2), 100-108.

Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In CHI'10 Extended Abstracts on Human Factors in Computing Systems (pp. 2863-2872). ACM.

Rossi, A., Gaio, L., den Besten, M., & Dalle, J. M. (2010). Coordination and division of labor in open content communities: the role of template messages in Wikipedia. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on (pp. 1-10). IEEE.

Wang, J., Faridani, S., & Ipeirotis, P. (2011). Estimating the completion time of crowdsourced tasks using survival analysis models. Crowdsourcing for search and data mining (CSDM).

Yan, T., Kumar, V., & Ganesan, D. (2010). Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones. In Proceedings of the 8th international conference on Mobile systems, applications, and services (pp. 77-90). ACM.