Analyzing the Amazon Mechanical Turk Marketplace

Panagiotis G. Ipeirotis[1]

New York University

Introduction

Amazon Mechanical Turk (AMT) is a popular crowdsourcing marketplace, introduced by Amazon in 2005. The marketplace is named after, "Mechanical Turk" an 18th century "automatic" chess playing machine, which was handily beating humans in chess games. Of course, the robot was not using any artificial intelligence algorithms back then. The secret of the "Mechanical Turk" machine was a human operator, hidden inside the machine, who was the real intelligence behind the intelligent behavior exhibited by the machine.

The Amazon Mechanical Turk is also a marketplace for small tasks that cannot be easily automated today. For example, humans can easily tell if two different descriptions correspond to the same product, can easily tag an image with descriptions of its content, or can easily transcribe with high quality an audio snippet. However, such simple tasks for humans are often very hard for computers. Using AMT, it is possible for computers to use a programmable API to post tasks on the marketplace, which are then fulfilled by human users. This API-based interaction gives the impression that the task can be automatically fulfilled, hence the name "Mechanical Turk."

In the marketplace, employers are known as "*requesters*" post tasks, which are called "*HITs*," an acronym of "Human Intelligence Tasks." The HITs are then picked up by online users, referred to as "*workers*," who complete them in exchange for a small payment, typically a few cents per HIT.

Since the concept of crowdsourcing is relatively new, many potential participants have questions about the AMT marketplace. For example, a common set of questions that pop up in an "introduction to crowdsourcing and AMT" session are the following:

- Who are the workers that complete these tasks?
- What type of tasks can be completed in the marketplace?
- How much does it cost?
- How fast can I get results back?
- How big is the AMT marketplace?

For the first question, about the demographics of the workers, past research (Ipeirotis, 2010; Ross *et al.* 2010) indicated that the workers that participate on the marketplace are mainly coming from the United States, with an increasing proportion coming from India. In general, the workers are representative of the general Internet user population but are generally younger and, correspondingly, have lower income and smaller families.

---

[1] Panagiotis G. Ipeirotis is an Associate Professor at the Department of Information, Operations, and Management Sciences at Leonard N. Stern School of Business of New York University. His recent research interests focus on crowdsourcing. He received his Ph.D. degree in Computer Science from Columbia University in 2004, with distinction. He has received two Microsoft Live Labs Awards, two "Best Paper" awards (IEEE ICDE 2005, ACM SIGMOD 2006), two "Best Paper Runner Up" awards (JCDL 2002, ACM KDD 2008), and is also a recipient of a CAREER award from the National Science Foundation. This work was supported by the National Science Foundation under Grant No. IIS-0643846

At the same time, the answers for the other questions remain largely anecdotal and based on personal observations and experiences. To understand better what types of tasks are being completed today using crowdsourcing techniques, we started collecting data about the marketplace. Here, we present a preliminary analysis of the findings and provide directions for interesting future research.

The rest of the paper is structured as follows. First, we describe briefly the data collection process and the characteristics of the collected dataset. Then we describe the characteristics of the requesters in terms of activity and posted tasks, and we also provide a short analysis of the most common tasks that are being completed on Mechanical Turk today. Next, we analyze the price distributions of the posted HITs and analyze the HIT posting and completion dynamics of the marketplace. We conclude by presenting an analysis of the completion time distribution of the HITs on Mechanical Turk and present some direction for future research and some design improvements that can improve the efficiency and effectiveness of the marketplace.

Data Collection

We started gathering data about the marketplace of AMT in January 2009 and we keep collecting data until today. The process of collecting data is the following: Every hour we crawled the list of "HITs Available" on AMT and we kept the status of each available *HIT group* (*groupid, requester, title, description, keywords, rewards, number of HITs available within the HIT group, qualifications required, time of expiration*). We also stored the HTML content of each HIT. Following this approach, we could find the new HITs being posted over time, the completion rate of each HIT, and the time that they disappear from the market either because they have been completed or because they expired or because requester canceled and removed the remaining HITs from the market.[2] A shortcoming of this approach is that it cannot measure the redundancy of the posted HITs. So, if a single HIT needs to be completed by multiple workers, we can only observe it as a single HIT.

The data are also publicly available through the website http://www.mturk-tracker.com.

From the period of January 2009 till April 2010, we collected 165,368 HIT groups, with a total of 6,701,406 HITs, from 9,436 requesters. The total value of the posted HITs was $529,259. These numbers, of course, do not account for the redundancy of the posted HITs, or for HITs that were posted and disappeared between our crawls. Nevertheless, they should be good approximations (within an order of magnitude) of the activity of the marketplace.

[2] Identifying expired HITs is easy, as we know the expiration time of a HIT. Identifying "cancelled" HITs is a little trickier: we need to monitor the usual completion rate of a HIT over time, and see if it is likely, at the time of disappearance, for the remaining HITs to have been completed within the time since the last crawl.

Top Requesters and Frequently Posted Tasks

One way to understand what types of tasks are being completed in the marketplace is to find the "top" requesters and analyze the HITs that they post. Table 1 shows the top requesters, based on the total rewards of the HITs posted, filtering out requesters that were active only for a short period of time.

We can see that there are very few active requesters that post a significant amount of tasks in the marketplace and account for a large fraction of the posted rewards. Following our measurements, the top requesters listed in Table 1 (which is 0.1% of the total requesters in our dataset), account for more than 30% of the overall activity of the market.

Requester ID	Requester Name	#HIT groups	Total HITs	Rewards	Type of tasks
A3MI6MIUNWCR7F	CastingWords	48,934	73,621	$59,099	Transcription
A2IR7ETVOIULZU	Dolores Labs	1,676	320,543	$26,919	Mediator for other requesters
A2XL3J4NH6JI12	ContentGalore	1,150	23,728	$19,375	Content generation
A1197OGL0WOQ3G	Smartsheet.com Clients	1,407	181,620	$17,086	Mediator for other requesters
AGW2H4I480ZX1	Paul Pullen	6,842	161,535	$11,186	Content rewriting
A1CTI3ZAWTR5AZ	Classify This	228	484,369	$9,685	Object classification
A1AQ7EJ5P7ME65	Dave	2,249	7,059	$6,448	Transcription
AD7C0BZNKYGYV	QuestionSwami	798	10,980	$2,867	Content generation and evaluation
AD14NALRDOSN9	retaildata	113	158,206	$2,118	Object classification
A2RFHBFTZHX7UN	ContentSpooling.net	555	622	$987	Content generation and evaluation
A1DEBE1WPE6JFO	Joel Harvey	707	707	$899	Transcription
A29XDCTJMAE5RU	Raphael Mudge	748	2,358	$548	Website feedback

Table 1: Top Requesters based on the total posted rewards available to a single worker (Jan 2009 - April 2010).

Given the high concentration of the market, the type of tasks posted by the requesters shows the type of tasks that are being completed in the marketplace: Castingwords is the major requester, posting transcription tasks frequently; there are also two other semi-anonymous requesters posting transcription tasks as well. Among the top requesters we also see two mediator services, Dolores Labs (aka Crowdflower) and Smartsheet.com, who post tasks on Mechanical Turk on behalf of their clients. Such services are essentially aggregators of tasks, and provide quality assurance services on top of Mechanical Turk. The fact that they account for approximately 10% of the market indicates that many users that are interested in crowdsourcing prefer to use an intermediary that address the concerns about worker quality, and also allow posting of complex tasks without the need for programming. We also see that four of the top requesters use Mechanical Turk in order to create a variety of original content, from product reviews, feature stories, blog posts, and so on.[3] Finally, we see that two requesters use Mechanical Turk in order to classify a variety of objects into categories. This was the original task for which Mechanical Turk was used by Amazon.

[3] One requester, "Paul Pullen", uses Mechanical Turk in order to paraphrase existing content, instead of asking the workers to create content from scratch.

The high concentration of the market is not unusual for any online community. There is always a long tail of participants that has significantly lower activity than the top contributors. Figure 1 shows how this activity is distributed, according to the value of the HITs posted by each requester. The x-axis shows the $\log_2$ of the value of the posted HITs and the y-axis shows what percentage of requesters has this level of activity. As we can see, the distribution is approximately log-normal. Interestingly enough, this is approximately the same level of activity demonstrated by workers (Ipeirotis, 2010).

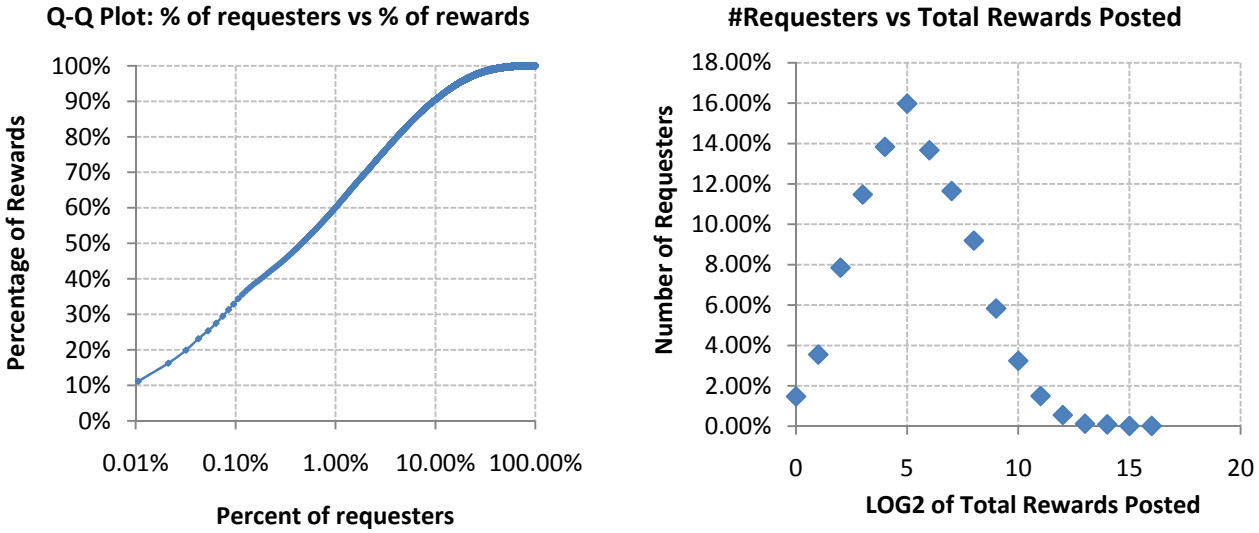

Figure 1: Number of requesters vs. total rewards posted.

For our analysis, we wanted to also examine the marketplace as a whole, to see if the HITs submitted by other requesters were significantly different than the ones posted by the top requesters. For this, we measured the popularity of the keywords in the different HITgroups, measuring the number of HITgroups with a given keywords, the number of HITs, and the total amount of rewards associated with this keyword. Table 2 shows the results.

Our keyword analysis of all HITs in our dataset indicates that transcription is indeed a very common task on the AMT marketplace. Notice that it is one of the most "rewarding" keywords and appears in many HITgroups, but not in many HITs. This means that most of the transcription HITs are posted as single HITs and not as groups of many similar HITs. By doing a comparison of the prices for the transcription HITs, we also noticed that it is a task for which the payment per HIT is *comparatively* high. It is unclear at this point if this is due to the high expectation for quality or whether the higher price simply reflects the higher effort required to complete these transcription HITs.

Beyond transcription, Table 2 indicates that classification and categorization are indeed tasks that appear in many (inexpensive) HITs. Table 2 also indicates that many tasks are about data collection, image tagging and classification, and also ask workers for feedback and advice for a variety of tasks (e.g., usability testing of websites).

Keyword	Rewards	Keyword	#HITGroups	Keyword	#HITs
data	$192,513	castingwords	48,982	product	4,665,449
collection	$154,680	cw	48,981	data	3,559,495
easy	$93,293	podcast	47,251	categorization	3,203,470
writing	$91,930	transcribe	40,697	shopping	3,086,966
transcribe	$81,416	english	34,532	merchandise	2,825,926
english	$78,344	mp	33,649	collection	2,599,915
quick	$75,755	writing	29,229	easy	2,255,757
product	$66,726	question	21,274	categorize	2,047,071
cw	$66,486	answer	20,315	quick	1,852,027
castingwords	$66,111	opinion	15,407	website	1,762,722
podcast	$64,418	short	15,283	category	1,683,644
mp	$64,162	advice	14,198	image	1,588,586
website	$60,527	easy	11,420	search	1,456,029
search	$57,578	article	10,909	fast	1,372,469
image	$55,013	edit	9,451	shopzilla	1,281,459
builder	$53,443	research	9,225	tagging	1,028,802
mobmerge	$53,431	quick	8,282	cloudsort	1,018,455
write	$52,188	survey	8,265	classify	1,007,173
listings	$48,853	editing	7,854	listings	962,009
article	$48,377	data	7,548	tag	956,622
research	$48,301	rewriting	7,200	photo	872,983
shopping	$48,086	write	7,145	pageview	862,567
categorization	$44,439	paul	6,845	this	845,485
simple	$43,460	pullen	6,843	simple	800,573
fast	$40,330	snippet	6,831	builder	796,305
categorize	$38,705	confirm	6,543	mobmerge	796,262
email	$32,989	grade	6,515	picture	743,214
merchandise	$32,237	sentence	6,275	url	739,049
url	$31,819	fast	5,620	am	613,744
tagging	$30,110	collection	5,136	retail	601,714
web	$29,309	review	4,883	web	584,152
photo	$28,771	nanonano	4,358	writing	548,111
review	$28,707	dinkle	4,358	research	511,194
content	$28,319	multiconfirmsnippet	4,218	email	487,560
articles	$27,841	website	4,140	v	427,138
category	$26,656	money	4,085	different	425,333
flower	$26,131	transcription	3,852	entry	410,703
labs	$26,117	articles	3,540	relevance	400,347
crowd	$26,117	search	3,488	flower	339,216
doloreslabs	$26,117	blog	3,406	labs	339,185
crowdflower	$26,117	and	3,360	crowd	339,184
delores	$26,117	simple	3,164	crowdflower	339,184
dolores	$26,117	answers	2,637	doloreslabs	339,184
deloreslabs	$26,117	improve	2,632	delores	339,184
entry	$25,644	retranscribe	2,620	dolores	339,184
tag	$25,228	writer	2,355	deloreslabs	339,184
video	$25,100	image	2,322	find	338,728
editing	$24,791	confirmsnippet	2,291	contact	324,510
classify	$24,054	confirmtranscription	2,288	address	323,918
answer	$23,856	voicemail	2,202	editing	321,059

Table 2: The top-50 most frequent HIT keywords in the dataset, ranked by total reward amount, # of HITgroups, and # of HITs.

Price Distributions

To understand better the typical prices paid for crowdsourcing tasks on AMT, we examined the distribution of the HIT prices and the size of the posted HITs. Figure 2 illustrates the results. When examining *HIT groups*, then we can see that only 10% of the HITgroups have a price tag of 2 cents or less, 50% of the HITs have price above 10 cent, and that 15% of the HITs come with a price tag of $1 or more.

However, this analysis can be misleading. In general, HITgroups with high price only contain a single HIT, while the HITgroups with large number of HITs have a low price. Therefore, if we compute the distribution of HITs (not HITgroups) according to the price, we can see that 25% of the HITs create on Mechanical Turk have a price tag of just 1 cent, 70% of the HITs have a reward of 5 cents or less, and 90% of the HITs come with a reward of less than 10 cents. This analysis confirms the common feeling that most of the tasks on Mechanical Turk have tiny rewards.

Of course, this analysis simply scratches the surface of the bigger problem: How can we automatically price tasks, taking into consideration the nature of the task, the existing competition, the expected activity level of the workers, the desired completion time, the tenure and prior activity of the requester, and many other factors? For example, how much should we pay for an image tagging task, for 100,000 images, in order to get it done within 24 hours? Building such models will allow the execution of crowdsourcing tasks to become easier for people that simply want to "get things done" and do not want to tune and micro-optimize their crowdsourcing process.

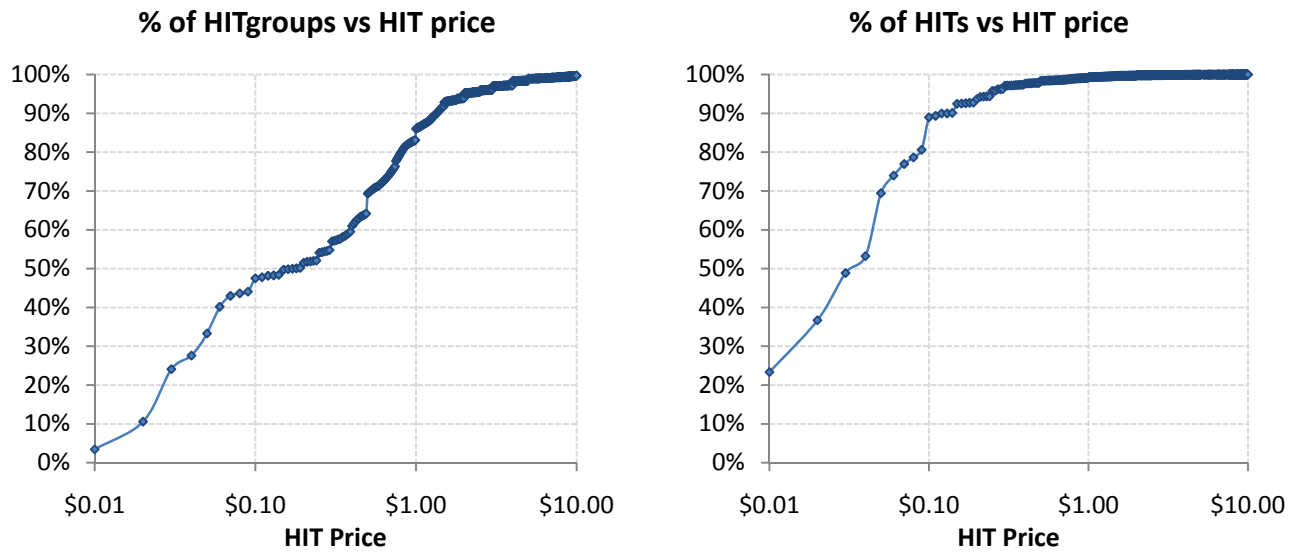

Figure 2: Distribution of HITgroups and HITs according to HIT Price.

Activity Dynamics on the AMT Marketplace: Posting and Serving Processes

What is the typical activity in the AMT marketplace? What is the volume of the transactions? These are very common questions from many people that are interested in understanding the size of the market and its demonstrated capacity[4] for handling big tasks.

One way to approach such questions is to examine the task posting and task completion activity on AMT. By studying the posting activity we can understand the demand for crowdsourcing, and the completion rate shows how fast the market can handle the demand. To study these processes, we computed, for each day, the value of tasks being posted by AMT requesters and the value of the tasks that got completed in each day.

We present first an analysis of the two processes (posting and completion), ignoring any dependencies on task-specific and time-specific factors. Figure 3 illustrates the distributions of the posting and completion processes. The two distributions are similar but we see that, in general, the rate of completion is slightly higher than the rate of arrival. This is not surprising, and is a required stability condition: if the completion rate was lower than the arrival rate, then the number of incomplete tasks in the marketplace would go to infinity. We observed that the median arrival rate is $1,040 per day and the median completion rate is $1,155/day. If we assume that the AMT marketplace behaves like an M/M/1 queuing system, and using basic queuing theory, we can see that a task worth $1 has an average completion time of 12.5 minutes, ***resulting in an effective hourly wage of $4.8***.

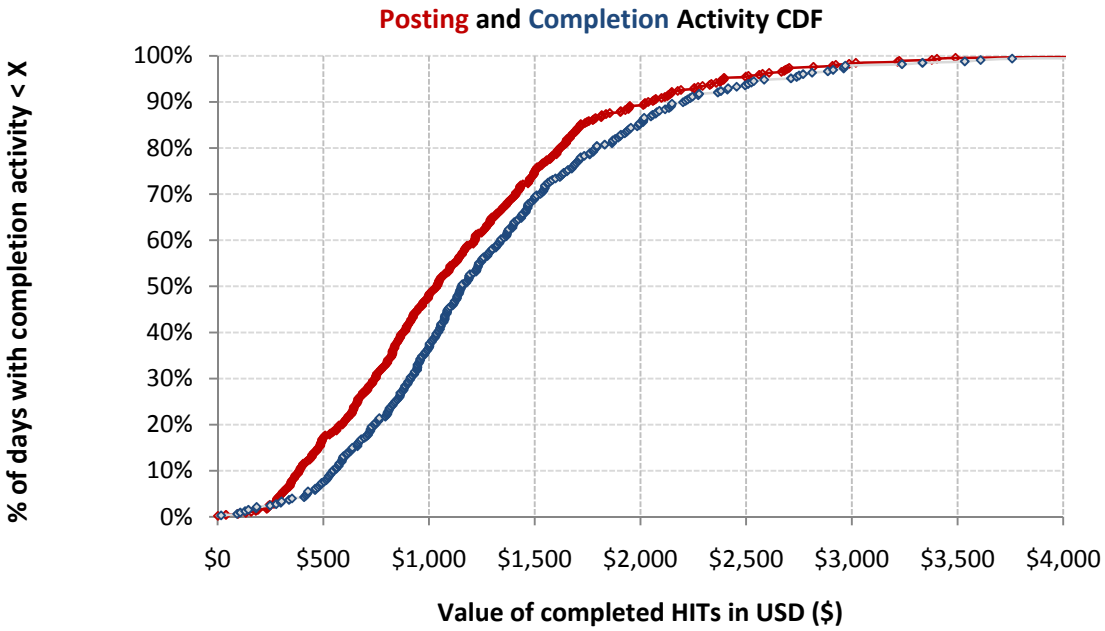

Figure 3: The distribution of the arrival and completion rate on the AMT marketplace, as a function of the USD ($) value of the posted/completed HITs.

Of course, this analysis is an oversimplification of the actual process. The tasks are not completed in a first-in-first-out manner, and the completion rate is not independent of the arrival rate. In reality, workers pick tasks following personal preferences or by being restricted by the web user interface of AMT. For example (Chilton et al. 2010) indicate that most workers use two of the main task sorting mechanisms provided by AMT to find and complete tasks ("recently posted" and "largest number of HITs" orders). Furthermore, the completion rate is not

[4] Detecting the true capacity of the market is a more involved task than simply measuring its current serving rate. Many workers may show up only when there is a significant amount of work for them, and be dormant under normal loads. Examining fully this question is beyond the scope of this paper.

independent of the arrival rate. When there are many tasks available, more workers come to complete tasks, as there are more opportunities to find and work for bigger tasks, as opposed to working for one-time HITs. As a simple example, consider the dependency of posting and completion rates on the day of the week. (Figure 4 illustrates the results.) The posting activity from the requesters is significantly lower over the weekends and is typically maximized on Tuesdays. This can be rather easily explained: since most requesters are corporations and organizations, most of the tasks are being posted during normal working days. However, the same does not hold for workers. The completion activity is rather unaffected by the weekends. The only day on which the completion rate drops is on Monday, and this is most probably a side-effect of the lower posting rate over the weekends. (There are fewer tasks available for completion on Monday, due to the lower posting rate over the weekend.)

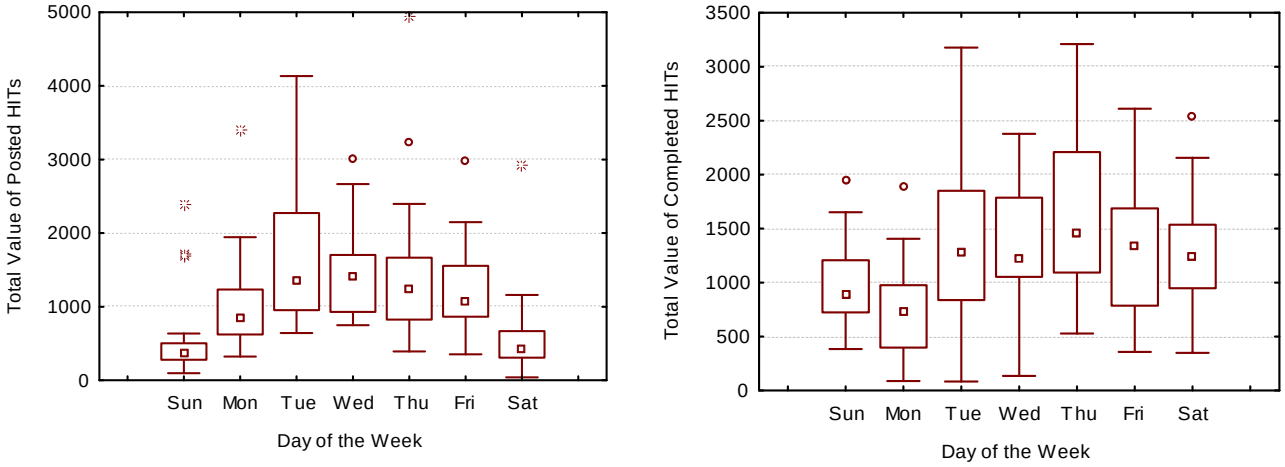

Figure 4: The posting and completion rate on AMT as a function of the day of the week

An interesting open question is to understand better how to model the marketplace. Work on queuing theory for modeling call centers is related, and can help us understand better the dynamics of the market and the way that workers handle the posted tasks. Next, we present some evidence that modeling can help us understand better the shortcomings of the market and point to potential design improvements.

Activity Dynamics on the AMT Marketplace: Completion Time Distribution

Given that the system does not satisfy the usual queuing assumptions of M/M/1 for the analysis of completion times, we analyzed empirically the completion time for the posted tasks. The goal of this analysis was to understand what approaches may be appropriate for modeling the behavior of the AMT marketplace.

Our analysis indicated that the completion time follows (approximately) a power law, as illustrated in Figure 5. We observe some irregularities, with some outliers at approximately 12 hours and at the 7-day completion times. These are common "expiration times" set for many HITs, hence the sudden disappearance of many HITs at that point. Similarly, we see a different behavior of HITs that are available for longer than one week: these HITs are typically "renewed" by their requesters by the continuous posting of new HITs within the same HITgroup.[5] Although it is still unclear what dynamics causes this behavior, the analysis by Barabási (2005) indicates that priority-based completion of tasks can lead to such power-law distributions.

To better characterize this power-law distribution of completion times, we used the maximum likelihood estimator for power-laws. To avoid biases, we also marked as "censored" the HITs that we detected to be "aborted before completion" and the HITs that were still running at the last crawling date of our dataset. (For brevity, we omit the details.) The MLE estimator indicated that the most likely exponent for the power-law distribution of the completion times of Mechanical Turk is α=-1.48. This exponent is very close to the value predicted theoretically for the queuing model of (Cobham, 1954), in which each task upon arrival is assigned to a queue with different priority. Barabási (2005) indicates that the Cobham model can be a good explanation of the power-law distribution of completion times only when the arrival rate is equal to the completion rate of tasks. Our earlier results indicate that for the AMT marketplace this is not far from reality. Hence the Cobham model of priority-based execution of tasks can explain the power-law distribution of completion times.

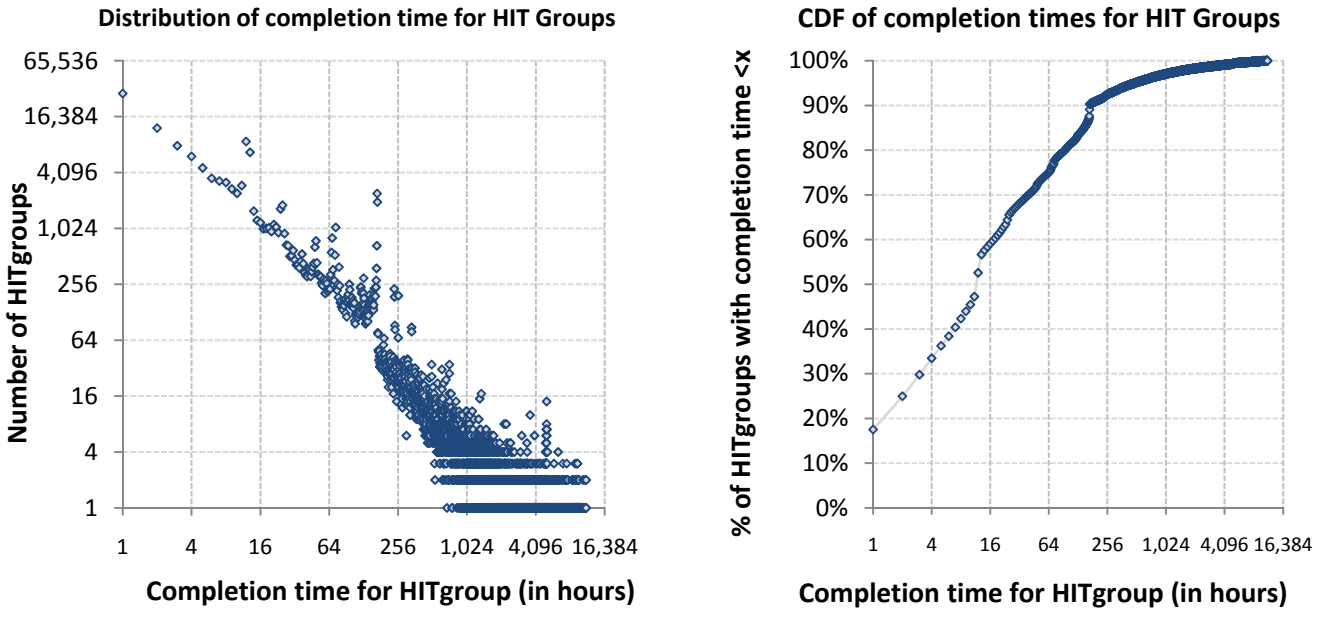

Figure 5: The distribution of completion times for HITgroups posted on AMT. The distribution does not change significantly if we use the completion time per HIT (and not per HITgroup), as 80% of the HIT groups contain just one HIT.

[5] A common reason for this behavior is for the HIT to appear in the first page of the "Most recently posted" list of HITgroups, as many workers pick the tasks to work on from this list (Chilton, 2010).

Unfortunately, a system with a power-law distribution of completion times is rather undesirable. Given the infinite variance of power-law distributions, it is inherently difficult to predict the necessary time required to complete a task. Although we can predict that for many tasks the completion time will be short, there is a high probability that the posted task will need a significant amount of time to finish. This can happen when a small task is not executed quickly, and therefore is not available in any of the two preferred queues from which workers pick tasks to work on. The probability of a "forgotten" task increases if the task is not discoverable through any of the other sorting methods as well.

This result indicates that it is necessary for the marketplace of AMT to be equipped with better ways for workers to pick tasks. If workers can pick tasks to work on in a slightly more "randomized" fashion, it will be possible to change the behavior of the system and eliminate the "heavy tailed" distribution of completion times. This can lead to a higher predictability of completion times, which is a desirable characteristic for requesters. Especially new requesters, without the necessary experience for making their tasks visible, would find such a characteristic desirable, as it will lower the barrier to successfully complete tasks as a new requester on the AMT market.

We should note, of course, that these results do not take into consideration the effect of various factors. For example, an established requester is expected to have its tasks completed faster than a new requester that has not established connections with the worker community. A task with a higher price will be picked up faster than an identical task with lower price. An image recognition task is typically easier than a content generation task, hence more workers will be available to work on it and finish it faster. These are interesting directions for future research, as they can show the effect of various factors when designing and posting tasks. This can lead to a better understanding of the crowdsourcing process and a better prediction of completion times when crowdsourcing various tasks.

Higher predictability means lower risk for new participants. Lower risk means higher participation and higher satisfaction both for requesters and for workers.

Conclusions

Our analysis indicates that the AMT is a heavy-tailed market, in terms of requester activity, with the activity of the requesters following a log-normal distribution; the top 0.1% of the requesters amount for 30% of the dollar activity and with 1% of the requesters posting more than 50% of the dollar-weighted tasks. A similar activity pattern also appears from the side of workers (Ipeirotis, 2010). This can be interpreted both positively and negatively. The negative aspect is that the adoption of crowdsourcing solutions is still minimal, as only a small number of participants actively use crowdsourcing for large-scale tasks. On the other hand, the long tail of requesters indicates a significant interest for such solutions. By observing the practices of the successful requesters, we can learn more about what makes crowdsourcing successful, and increase the demand from the smaller requesters.

We also observe that the activity is still concentrated around small tasks, with 90% of the posted HITs giving a reward of 10 cents or less. A next step in this analysis is to separate the price distributions by type of task and identify the "usual" pricing points for different types of tasks. This can provide guidance to new requesters that do not know whether they are pricing their tasks correctly.

Finally, we presented a first analysis of the dynamics of the AMT marketplace. By analyzing the speed of posting and completion of the posted HITs, we can see that Mechanical Turk is a price-effective task completion marketplace, as the estimated hourly wage is approximately $5. Further analysis will allow us to get a better insight of "how things get done" on the AMT market, identifying elements that can be improved and lead to a better design for the marketplace. For example, by analyzing the waiting time for the posted tasks, we get significant evidence that workers are limited by the current user interface and complete tasks by picking the HITs available through one of the existing sorting criteria. This limitation leads to a high degree of unpredictability in completion times, a significant shortcoming for requesters that want high degree of reliability. A better search and discovery interface (or perhaps a better task recommendation service, a specialty of Amazon.com, can lead to improvements in the efficiency and predictability of the marketplace.

Further research is also necessary in better predicting how changes in the design and parameters of a task can affect quality and completion speed. Ideally, we should have a framework that automatically optimizes all the aspects of task design. Database systems hide all the underlying complexity of data management, using query optimizers to pick the appropriate execution plans. Google Predict hides the complexity of predictive modeling by offering an auto-optimizing framework for classification. Crowdsourcing can benefit significantly by the development of similar framework that provide similar abstractions and automatic task optimizations.

References

- Mechanical Turk Monitor, http://www.mturk-tracker.com.
- Barabási, A.-L. 2005. The origin of bursts and heavy tails in human dynamics. Nature, 435:207-211.
- Cobham, A. 1954. Priority assignment in waiting line problems. J. Oper. Res. Sec. Am. 2, 70–76.
- Chilton, L. B., Horton, J. J., Miller, R. C., and Azenkot, S. 2010. Task search in a human computation market. In Proceedings of the ACM SIGKDD Workshop on Human Computation (Washington DC, July 25 - 25, 2010). HCOMP '10. ACM, New York, NY, 1-9.
- Ipeirotis, P. 2010. Demographics of Mechanical Turk. Working Paper CeDER-10-01, New York University, Stern School of Business. Available at http://hdl.handle.net/2451/29585
- Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In Proceedings of the 28th of the international Conference Extended Abstracts on Human Factors in Computing Systems (Atlanta, Georgia, USA, April 10 - 15, 2010). CHI EA '10. ACM, New York, NY, 2863-2872.