2014
02

Working Paper
INSTITUTO DE POLÍTICAS Y BIENES PÚBLICOS (IPP)

Human computing via online labor markets. The perils and promises of crowdsourcing in data-rich ecosystems.

Matthijs den Besten,
Catalina Martínez,
Nicolas Besson,
Stéphane Maraut
&
Jean-Michel Dalle

CSIC
Consejo Superior de Investigaciones Científicas
CCHS
Instituto de Políticas y Bienes Públicos

INSTITUTO DE POLÍTICAS Y BIENES PÚBLICOS – CSIC

Instituto de Políticas y Bienes Públicos (CSIC-IPP)
Consejo Superior de Investigaciones Científicas
C/ Albasanz, 26-28
28037 Madrid (España)

Tel: +34 91 6022300
Fax: +34 91 3045710

http://www.ipp.csic.es

How to quote or cite this document:

den Besten, M., Martínez, C., N. Besson, N., Maraut, S. & Dalle, J.-M. (2014) Human computing via online labor markets. The perlis and promises of crowdsourcing in data-rich ecosystems. Instituto de Políticas y Bienes Públicos (IPP) CSIC, Working Paper. 2014-02.

Available at: digital.csic.es

Human computing via online labor markets
The perils and promises of crowdsourcing in data-rich ecosystems

Matthijs den Besten
Montpellier Business School/Montpellier Research in Management, Montpellier, France
m.den-besten@montpellier-bs.com

Catalina Martinez
Institute of Public Goods and Policies (CSIC-IPP), Madrid, Spain
catalina.martinez@csic.es

Nicolas Besson
Alcméon, Sophia-Antipolis, Nice, France
nicolas.besson@alcmeon.com

Stéphane Maraut
Independent researcher, Madrid, Spain
stephane.maraut@gmail.com

Jean-Michel Dalle
Université Pierre et Marie Curie, Paris, France
Centre de Recherche en Gestion, Ecole Polytechnique, France
jean-michel.dalle@upmc.fr

Abstract

In this article, we focus on the need for "human computing" in data-rich ecosystems, notably as a consequence of data variety and typically for name disambiguation, and explore ways to manage it via online platforms for paid crowdsourcing. Based on several studies of Amazon Mechanical Turk, a well-established platform for matching data treatment tasks to human beings willing to carry them out, we illustrate the difficulties involved as requesters compete for the attention of workers. We suggest that researchers should shift from a technical analysis and tentative resolution of human computing perils and pitfalls, towards a more economic and managerial analysis of human computing platforms understood as online labor markets, notably in their multi-sided nature and with respect to how they manage the attention of online workers.

Keywords: Online Labor Markets; Crowd-Sourcing; Big Data; Amazon Mechanical Turk; Human Computing
MSC: 62-04, 00-02
JEL: C81, C90, O30

1. Introduction

Big data, taken as shorthand for data-intensive activities in digital ecosystems, has been heralded as a great opportunity for management (McAfee and Brynjolfsson, 2012) and has already had a big impact in many economic sectors. However, what is new about big data is not only its volume, and the speed at which it is generated (its velocity), but also its *variety* (Chen et al., 2013), for which the different spellings of the same name is a good example. Quite paradoxically then, with regards to dealing with this variety of the data in data-rich ecosystems, humans, a.k.a. the general public, can be enlisted to help with data processing in big data environments (Brynjolfsson *et al*. 2014), because dealing with such variety and heterogeneity is still particularly hard for computers, but much less so for human beings. One would thus like to embed the "manual" approach in the "computational" method, leaving only the part where humans perform better to humans (Lewis *et al*., 2013). This mixing of computer and human efforts is sometimes referred to as "heteromation" (Ekbia & Nardi, 2014) and more general to "human computing" (Zittrain, 2008). Irrespective of naming, this division of labor between humans and computers ultimately turns into a work of co-creation between the data analyst and the crowd he or she employs: a difficult task (Striukova & Rayna 2015). It is therefore of utmost importance to understand how firms can engage with crowd workers (Raasch 2011; Lauritzen *et al*. 2013). Furthermore, human computing is generally established via platforms acting as intermediaries, and whose role in data-rich ecosystems is thus crucial (Rong *et al*. 2013).

In order to analyze the role of human computing platforms, we focus in this article on the most widely used platform for paid crowdsourcing, Amazon Mechanical Turk (AMT). We first describe the role that a platform like AMT can play with respect to name disambiguation, a process typically called the "name game" in scientometrics, and the many known pitfalls of using AMT in this respect. We further present a protocol used otherwise with this objective, and stress specifically the conclusions and difficulties that have been encountered in doing so, that we suggest stem from a partial misunderstanding and naïveté of the existing literature with respect to what online human computing platforms actually entail. Namely, understanding and finding ways to circumvent the many pitfalls associated with AMT should not blur the fact that, in essence, AMT is a platform via which "requesters" have access to an online labor market, which we suggest should be analyzed as such. In this respect, we stress two particularly important aspects of the "human computing" labor market associated with AMT: first, using empirical data from several months of activity on AMT, that AMT is a 2-

sided platform that has probably self-selected a category of workers interested mostly in repetitive simple and easy tasks, with a surprising low sensitivity to price signals; and second, using a tool developed otherwise to simulate AMT, that we ignore much of the basic determinants of workers' allocation of attention on AMT. We conclude that, much more than it is often the case, further investigations are needed from economists and management scientists to clarify the role and the functioning of human computing labor markets.

2. Amazon Mechanical Turk as a human computing platform

a. *Using AMT for name disambiguation*

We first report on our past experiments with AMT, used as a human computing platform to address the variety of name spellings. AMT is an online labor market in which people get paid a small amount for each task they accomplish (Vakharia and Lease 2013). Concretely, AMT is a platform connecting requesters (in our case, the data analyst) with workers. Workers look for jobs that fit their abilities and needs (Schultze *et al*. 2012). "Requesters" convey the difficulty of the tasks they publish, and the rewards connected to their completion through the description of these tasks: framing the tasks correctly is therefore of crucial importance. And what makes framing difficult beyond the variability of skills available in the workforce, is that, apart from money, workers are also motivated for intrinsic reasons (fun, developing skills, building up a track record, etc.). Moreover, requesters are in competition with other requesters for the attention of workers, while workers and requesters in general have access to very limited information about each other, which further complicates the task of establishing a person-job fit. Experienced users of the platform manage to have access to private information, however: requesters can maintain a list of workers that have worked for them and workers can maintain a list of requesters they have worked for. Consequently, based on their experience, users can build a list of trusted partners. In addition, external services have emerged to help requesters frame their tasks and to help workers vet requesters (Irani & Silberman 2013). The latter may also provide workers with an occupational community and a professional identity (Lehdonvirta & Mezier 2013).

Data scientists have embraced AMT in particular for a wide variety of activities ranging from data collection (e.g. Snow *et al*., 2008), and image analysis (Maisonneuve and Chopard, 2012), to interview transcription (Marge *et al*., 2010), and copy-editing (Bernstein *et al*., 2010). AMT has been heralded as a quick and easily accessible means for doing behavioral experiments (Mason and Suri 2012). Rand (2012) reviews a number of replication studies and draws the conclusion that AMT is reliable as a platform to run experiments on (for instance Sprouse (2011) reports no difference for linguistic judgments with respect to syntax between AMT and laboratory settings).

We used AMT in the context of a recent research project as a means to address the "Names-Game": in innovation studies, the use of disambiguation techniques to reclassify patent data at the inventor level is called the "Names-Game" (Raffo and Lhuillery 2009).[1] Matching is absolutely non-trivial as names are sometimes misspelt and the same person can be referred to in a variety of ways, and scientometric research hinges on the ability to link research outputs to the researchers responsible for them in data-intensive environments. The difficulty lies in deciding whether different works with similar author names belong to the same person or not. Hence, given limited resources, automated methods are typically preferred for matching and disambiguation (Smalheiser and Torvik 2009). See Cuxac *et al*., Gurney *et al*. (2012), and Wang *et al*. (2012a) for recent examples. Yet, manual matching is considered to yield higher levels of accuracy (Veve 2009). This activity was pioneered by Trajtenberg *et al*. (2006), and considerable efforts have indeed been devoted by different research groups in the past years to disambiguate inventors listed in patents and identify academic researchers amongst them. This has been mainly done in three different ways: i) matching inventors to research staff lists (Thursby *et al*. 2009; Lissoni *et al*. 2008; Lissoni *et al*. 2009); ii) searching for the "professor" title in the inventors' name fields (Schmoch 2007; Czarnitzki *et al*. 2007; Von Proff *et al*. 2012); and iii) matching inventors to authors of scientific publications (Noyons *et al*. 2003a, 2003b; Schmoch *et al*. 2012; Dornbusch *et al*. 2013; Maraut and Martinez 2014).[2] In the light of all this literature, we have explored the possibility to ask anonymous reviewers in crowdsourcing platforms to carry out authorship disambiguation manually, in a cost-efficient and reliable way.

b. The many known pitfalls of using AMT

In AMT, workers get to select the tasks they want to carry out among the ones that are available. Typically, a limited number of workers will end up doing the brunt of the work (Bernstein *et al*., 2010). It is possible for the requester to require that workers pass a qualification first. Alonso and Mizzaro (2012) find that workers who have passed a test are

[1] For information on most recent developments see the European Science Foundation Research Networking Programme – Academic Patenting in Europe (APE-INV) at http://www.esf-ape-inv.eu/ .
[2] See also NSF project to link MEDLINE papers with USPTO patents: http://www.nsf.gov/awardsearch/showAward?AWD_ID=0965341

more likely to complete the tasks. Furthermore, Wang *et al*. (2012b) find that the workers who have past the qualification tests deliver work of slightly higher quality. In addition to prescreening, Chandler *et al*. (2014) suggest gradually selecting workers on the basis of past engagement. There are several types of qualification tests. The most prevalent is to test for past performance in terms of proportion of work that has been approved. Ipeirotis (2010a) observes that this test is very easy to trick. Other tests concern (self-reported) skills and the location of workers derived from their IP-address. Demartini *et al*. (2012) find that while in general Indian workers performed worse than their American counterparts, for items related to local Indian news they performed better. Recently, AMT added a “master” qualification to the menu of tests that can be set. AMT attributes this qualification to workers it considers trustworthy. Despite (or because of) the opacity of the criteria on the basis of which the qualification is attributed, restricting the tasks to “master” workers will likely lead to higher quality results (Ipeirotis, 2012). Note however that Amazon charges a higher fee for work carried out this way. Furthermore, as for instance reported by Chandler *et al*. (2014) workers are slower to react.

In order to attract the attention of workers, it helps if the tasks are relatively easy to grasp: the quality of task formulation strongly influences the quality of results obtained in AMT (Kittur *et al*., 2008). It also helps if there are not too many other tasks competing for attention. Ipeirotis (2009) observed that most tasks are launched during weekdays and that most workers are active during weekend. If this still holds, it would be better to launch the task during the weekend. It also helps to offer higher pay than other requesters. According to Horton and Chilton (2010) a higher effort level can be expected in return for a higher pay. They also discovered that a number of workers clearly prefer earning total amounts that are evenly divisible by 5 and speculate that this might be because these workers pursue earning targets. The quality of the work does not seem to be affected by the level of payment, however (Mason and Watts, 2009; Mason and Suri, 2012). Nevertheless, Acemoglu *et al*. (2014) suggest one could implement a dynamic pricing mechanism in which tasks that have not been completed because they appear too cumbersome will be offered again at a higher price. In order to improve quality, Shaw *et al*. (2011) find that it helps to indicate that payment will be linked to the extent in which responses conform to responses given by peers. Redundancy in responses can also help counting the cheating, which, according to Eickhoff and de Vries (2013), has become more prevalent recently. The introduction of the “master qualification” mentioned before might serve to combat this practice. Hirth *et al*. (2013) blame the relative anonymity of workers in combination with an appeal limited to the profit motive. Kittur *et al*. (2008) already observed that the best way to prevent cheating is to make it more difficult than simply playing along. Among the other measures to improve quality, Ipeirotis (2010b) advices that one should announce the rules of the game clearly in the task description an announce sanctions if deficiencies are observed. Franklin *et al*. (2011) warn however that refusing to pay ex post may provoke a backlash from workers, who rate requesters on dedicated forums such as TurkOpticon and Turker Nation. Finally, Kittur *et al*. (2008) found a significant increase in the quality of the data obtained after the inclusion of additional questions with verifiable answers. If answers can be verified automatically this can be used for immediate feedback. Otherwise it can be used to identify misbehavior ex post (Shaw *et al*., 2011). According to one worker interviewed by Kittur *et al*. (2012) tasks are often monotonous. The resulting boredom may be a cause for abandoning the task (Sun *et al*., 2011). The inclusion of additional questions may also serve to alleviate this boredom.

c. *A prototype for crowdsourcing name disambiguation*

Based on these insights, we designed a prototype using the AMT platform for name disambiguation. AMT workers were asked replicate some of the manual checks done in the process of building the database of Spanish author-inventors described in Maraut and Martinez (2014). Building this database had entailed combining information from more than 15.000 patent applications and 150.000 scientific publications, with no limitation in terms of fields, regions or types of institutions. An added difficulty was the mix of specific features of Spanish names (e.g. multiple surnames) and the frequent existence of input errors due to poor understanding of the Spanish name patterns, apart from the lack of structure of person and institution name fields in large bibliographic databases. All this suggested that human intervention could be particularly helpful to build training and validation sets in semi-supervised machine learning techniques.

Maraut and Martinez (2014) used a semi-supervised technique, combining automated matching techniques with human validation of dubious matches, and ended up identifying more than 4.000 author inventors. They first built clusters automatically by assigning a similarity score to author-inventor pairs on the basis of a weighted combination of a variety of matching and disambiguation indicators, where name matching indicators rely on approximate string matching using complex edit distance measures and entity resolution techniques and disambiguation indicators rely on contextual information, such as institutional affiliation, discipline and geographical location.[3] And then, by way of the clustering step, author and inventor identifiers likely to belong to the same person were grouped together and linked back to their corresponding patent applications and publications. In the original methodology, experts then intervened to manually check the dubious matches identified in the clusters, so that false pairs could be excluded and validated matches could be later used to improve the disambiguation recursively and improve precision of the final dataset. Dubious matches would be document pairs from whose validity is more difficult to assess due to common names, spelling mistakes, mobility (different affiliations) or multi-disciplinarity (different areas of specialization) of their corresponding authors and inventors. Since the number of dubious matches to be checked increases with the size of the initial sample, expert validation raises substantially the cost, time and effort needed to develop a large database of this kind.

With respect to AMT, Wang *et al*. (2012b), who had previously explored the design of AMT tasks to tackle entity resolution. They tested two ways to present the task: as a pair of records for which similarity has to be judged and as a list of records for which the distinct entities need to be enumerated. In both cases they provided workers with the opportunity to indicate the reasons for their choice. Overall, the pair based presentation appeared to be more popular with workers despite the fact that workers who had opted for the cluster-based presentation managed to complete these tasks more quickly than their colleagues who had opted for the pair-based task. Given the observation by Georgescu *et al*. (2012) that "most problems arise from workers being too quick and not paying enough attention to the task" the pair-based presentation would seem the better choice. The overall results for both variants in terms of precision and recall were very similar however. We opted for a cluster-based rather than a pair based presentation because the clues provided by the coherence of a publication record seemed important in this context. Moreover as each cluster may cover a large number of pair-based comparisons, the cluster representation requires far less separate tasks, and so it becomes feasible to offer a higher reward per task. The task presentation was similar to the one proposed by Wang *et al*. (2012). We also ask workers to identify the gender of the inventors in the list, as the analysis of the responses could give further indications with regards to the trustworthiness of the workers. It also might make the task more interesting to some people.

[3] For a detailed description of the methodology see Maraut and Martinez (2014), which makes use of a density-based technique known as DBSCAN (Ester et al. 1996) that relies on the notion of density reachability and connectivity.

We then asked AMT workers to review a number of randomly selected clusters from the database produced by Maraut and Martinez (2014) prior to expert validation, and compared their responses of to those of the experts engaged by Maraut and Martinez, where the latter were used as a sort of 'gold standard'. More precisely, we randomly selected 99 not-yet validated clusters of patents and publications likely to correspond to the same person, according to name similarity, affiliation, discipline, etc. which cover a total of 2106 distinct patent-publication pair comparisons. Clusters including pairs with uncommon names of authors and inventors that were exactly matched (e.g. no spelling mistakes) were deliberately excluded from this sample in order to avoid offering too simple tasks to AMT workers. The information on publications and patents provided to AMT workers to carry out each task includes: i) document type (whether the particular document presented is a patent application or journal article); ii) first and last names of authors and inventors (as they appeared in the original documents); iii) non-name information available in the documents (address, institution of affiliation in the case of journal articles, name of patent applicant in the case of patents); and iv) original document title (of the patent or the journal article). The title of the document is linked to a version of the document available online, so that the worker can get additional information if needed (e.g. abstract, coauthors). In AMT, 'checking whether all the patents and publications included in a given cluster belong to the same person' was the name of the AMT granular 'task' (HIT: human intelligence task), while requests to AMT workers were submitted in 'batches', each batch comprising several individual tasks.

d. Experiments, results, and issues

In order to test our protocol with respect notably to the price offered, we first submitted several batches of 10 tasks each, each of which could be resolved by a maximum of five workers with a prospective reward per task of 0.05 USD and 0.10 USD respectively, limiting their visibility to workers with a good record. These batches failed to elicit sufficient response, as in each only three tasks were completed over the next days by one worker in the first batch and three distinct workers respectively in the second batch. We then submitted two more batches offering a much higher reward of 0.50 USD per task completed. We made the first batch visible to all workers on the platform; the second only to those for whom more than 60% of past work had been approved by requesters. This time, all tasks were completed within a day after publication by 17 workers.

We next launched nine different batches of either 10 or 50 distinct tasks, with a reward per task fixed at 0.20 USD where each task was made available to a maximum of five different workers. The nine batches were published sequentially in the same week, from Thursday 3 April 2014 to Thursday 10 April, at randomly selected times of the day, with at least 20 hours difference between them. Each batch remained posted for 24 hours at the AMT website. With a reward of 0.20 USD per task the maximum reward a worker could thus earn per batch was five times higher for the batches with 50 clusters (10 USD) than for the batches with 10 clusters (2 USD). The batches were only visible for workers showing good quality working records, more precisely, the qualification required was greater than 98% approval rate for more than 100 approved tasks. The batches were publicized with the term 'Spanish names' as a keyword on the AMT platform in order to be able to attract workers with some knowledge of Spanish. In addition, "disambiguation" and "record linkage" were given as keywords.

Forty five (45) different workers participated in this experiment, 14 worked in more than one batch (31%), and 727 tasks were completed in total, presenting 99 different clusters in an equal number of distinct tasks. Table 1 recalls the main features of the experiment together with information on number of responses and worker participation. As generally reported in similar contexts, the distribution of effort was highly uneven.[4]

Table 1. Main features of the experiment

Batch ID	**Time created**	**Different tasks offered**	**Redundancy of tasks offered**	**Different tasks completed by at least one worker**	**Total number of tasks completed (including redundant)**	**Number of different workers completing at least one task**	**Number of author-inventor pairs with worker judgment** (i)	**Amount paid to workers (USD)**(ii)
1480448	Thu Apr 03 22:40:16 GMT 2014	50	5	49	149	8	1726	29.80
1481211	Fri Apr 04 15:59:12 GMT 2014	50	5	47	64	3	836	12.80
1482480	Sat Apr 05 19:08:58 GMT 2014	10	5	10	20	3	841	4.00
1482891	Sun Apr 06 07:01:04 GMT 2014	50	5	45	45	1	498	9.00
1484530	Mon Apr 07 19:04:14 GMT 2014	50	5	49	109	9	1661	21.80
1485453	Tue Apr 08 10:31:01 GMT 2014	10	5	10	32	6	460	6.40
1487123	Wed Apr 09 08:12:19 GMT 2014	10	5	10	26	4	716	5.20
1488800	Thu Apr 10 06:23:30 GMT 2014	10	5	10	40	12	342	8.00
1489909	Thu Apr 10 22:32:05 GMT 2014	50	5	50	243	28	4837	48.60
				99	727	45	11917	145.40

Note: (i) The number of author-inventor pairs set out in the table for each batch corresponds to the number of all different combinations of article-author and patent application-inventor pairs for which a judgment from AMT workers can be inferred based on their responses to the cluster-based task format presented in the Annex; (ii) The cost indicated does not include fees paid to Amazon.

[4] The Gini coefficient for the number of tasks submitted per worker is 0.75.

Among the 99 distinct tasks presented, 13 were only offered in one batch, the other 86 appeared in more than one batch (23 in two; 28 in three; other 28 in four; and 7 in five different batches). Tasks differed in terms of number of documents (articles and patent applications) presented to workers as being potentially authored by the same person, ranging from small tasks with only 2 documents to the largest tasks with 61 documents, with an average number of documents per batch per task between 8 and 18 and high standard deviations (Table 2).

Table 2. Number of documents (articles + patent applications) per task

		Number of documents (articles + patent applications)			
Batch	Tasks	**Mean**	**Std. Dev.**	**Min**	**Max**
1480448	49	7.98	4.70	3	20
1481211	47	7.79	6.12	2	32
1482480	10	18.40	16.79	5	57
1482891	45	8.04	5.87	2	30
1484530	49	10.90	13.32	2	61
1485453	10	13.70	15.60	4	54
1487123	10	14.20	12.20	3	41
1488800	10	11.10	15.91	3	54
1489909	50	11.30	12.26	2	61

Since more than one worker could complete the same task in a given batch (up to the maximum of 5 workers allowed), there was more than one response even for the 13 tasks offered in only one batch. No task was systematically ignored by workers, but redundancy of responses was achieved only for 93 of them because 6 were completed only once. These six tasks are slightly larger than the rest, with 14 documents on average compared to a mean of 10 for the others, but the difference is not statistically significant, suggesting that other factors counted more than size for their lower uptake. As regards other features of workers' behavior and effort intensity, we find that responsiveness of workers improves over time in terms of the proportion of a batch that is completed after 24 hours[5] and also, in contrast to the findings reported above, that weekends seem to be worse than weekdays in terms of workers' activity.[6]

To assess the quality of responses, we calculated a rate of agreement with experts[7] at the pair level and tested whether its distribution differed significantly across different batches, tasks or workers, by using the Kruskal-Wallis rank test. This test compares ranks of observations from the lowest to the highest score across groups and tests if the rank sum for each group is the same or not (if the groups were equal, their rank sum would be equal too), where the null hypothesis is that the distribution of the outcome variable is identical across groups. As shown in Table 3, agreement rates differ significantly across batches, tasks and workers.

Table 3. Kruskal-Wallis tests of agreement with experts at the pair level across batches, tasks and workers

	Chi-squared	p	Chi-squared with ties	p
Batches	144.327 with 8 d.f.	<0.0001	409.151 with 8 d.f.	<0.0001
Tasks	990.357 with 98 d.f.	<0.0001	2807.549 with 98 d.f.	<0.0001
Workers	413.156 with 44 d.f.	<0.0001	1171.250 with 44 d.f.	<0.0001

Note: When some scores receive tied ranks, a correction factor is used and a slightly different value of chi-squared is obtained.

Differences across tasks might be due to differences in task complexity. The differences across batches and across workers, however, are likely to be at least in part due to the fact that each batch is a different collection of tasks and workers, depending on when they were active, had a choice among a different offer of tasks. Table 4 checks whether agreement rates in batches is different from agreement rates for the author-inventor pairs among them that also occur in other batches. It turns out that the proportion of workers in the batches who agree with the expert assessment of author-inventor similarity is not significantly different from the proportion of workers in the other batches agreeing with the expert in 7 out of 9 batches. Furthermore, the agreement between the expert and the judgment arrived at by the majority of workers in a batch is not significantly different from the rate of agreement elsewhere for any

[5] A Pearson product-moment correlation between launch dates of batches and their completion rate is equal to 0.7, significant at 5%.
[6] The two batches launched in the weekend have of 0.4 and 0.18 compare to the median and mean of 0.6 for batches launched during weekdays.
[7] The expert validated pairs included in the final dataset of Maraut and Martinez (2014), which can be considered as a gold standard for the AMT responses.

of the batches. So, even though the speed with which one obtains results might be dependent on factors beyond the control of the requester, the quality of results seems not to be affected.

Table 4. Expert agreement at the pair level, differences across batches

Batch	Available author-inventor pairs	Matched pairs in other batches	KS≠ (i)	p-value	Chi-squared (ii)	p-value
1480448	644	588	0.22	0.000	296.52	<0.001
1481211	584	570	0.38	0.000	196.47	<0.001
1482480	425	425	0.15	0.000	58.99	<0.001
1482891	505	505	0.19	0.000	327.12	<0.001
1484530	1258	1233	0.17	0.000	689.13	<0.001
1485453	166	166	0.07	0.779	108.71	<0.001
1487123	325	115	0.36	0.000	49.97	<0.001
1488800	100	97	0.10	0.681	73.77	<0.001
1489909	1353	1280	0.13	0.000	577.66	<0.001

Notes: (i) Kolmogorov-Smirnov test of dissimilarity among proportion of workers who agree with the expert. (ii) Test for independence between batches and their controls based on the frequency the majority of workers agree with experts on pairs.

A similar set of tests of atypical expert agreement among workers does find significant differences for many of them. For 18 out of 45 workers the null-hypothesis of independence of worker agreement given majority agreement on the same author-inventor pairs cannot be rejected. Typically these workers complete only a few tasks, however, which suggests that experience counts.

As should be clear from these results, and although we believe that our experiments represent a significant step toward partly automating name disambiguation in a data-rich environment with crowdsourcing and human computing via AMT, they also point towards the difficulties faced by "requesters" with respect to the nature and the quality of the work that can be supplied through the platform. A preliminary, mostly business-oriented, answer to these difficulties, has been associated with the emergence of intermediary companies that post human intelligence tasks for the sake of others (Ipeirotis, 2010c). However, we would however like to argue that there is much more to understand here. The previous academic literature that has tried to analyze the pitfalls of AMT "technically" might have missed an important point: even if AMT is a set of algorithms, and even if the human computing steps can also be dealt with incentives mechanisms, AMT still is an online labor market, composed on online workers who are not naïve (Chandler et al., 2014), contrary maybe to the naïveté with which previous research has sometimes dealt with the economic and managerial aspects associated with any known labor market, even more so in the digital world where further issues are known to be prevalent. To name but 2 about which the next section will insist, the fact that AMT is a platform between requesters and workers implies a multi-sided nature that can result in self-selecting outcomes, and the fact that workers choose online on which problem they work implies that the management of their online attention is also a key issue.

3. "Human computing" actually rests on online labor markets

a. AMT as a two-sided platform

The experiments reported on in the previous section have shown a general dependence of work supply in AMT with respect to price signals, which is not at all surprising: what is more surprising is that, when we looked at the literature to find elements about the price elasticity of online work supply, we found only very few and largely inconclusive studies (Franklin et al., 2011; Mason & Watts 2010; Yan et al., 2010). We therefore gathered a dataset by crawling AMT's website every 3 minutes during a period of about 2 months in order to investigate temporal data on several hundreds of real AMT projects and to measure the mean speeds at which individual tasks disappear (are "executed") from available projects, that is to say, to 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), or else the *determinants of worker choice* among the many problems that are offered on AMT (Dalle et al., 2014). The global allocation of online efforts in online communities (Dalle & David, 2005; den Besten, Dalle et Galia, 2008) indeed 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 direct coordination between workers is instrumental (den Besten & Dalle, 2014; Rossi et al., 2010), choices are specially affected by the pricing of tasks and by other characteristics of tasks and projects, even if there is coordination among workers on dedicated forums. In this context, we found preliminary evidence according to which the pricing of individual tasks ("Price") does not seem to influence workers choice, at least directly, contrary to "Size" i.e. to the number of individual tasks in a given AMT project. This observation is consistent with workers simply maximizing their wages by increasing their productivity over time through the selection of groups of tasks on which they could focus and specialize for a sufficiently long amount of time, and coherent with Franklin et al (2011). Furthermore, and contrary to the length of the title given to AMT projects ("Length(Title)"), the length of *task descriptions* on AMT's website ("Length(Desc)") increases the speed at which they are executed, which could correspond to a preference from workers for more detailed descriptions when choosing among available tasks, and/or to the fact that tasks that are "better thought through", both in their description and in the process leading to their execution, seem able to attract workers.

Table 5. 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*

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, further limiting our dataset to successive pairs with positive price variation or to relatively small positive variations. Even then, the variation in price did not appear as significant. We further found, by experimenting directly on AMT, that workers appeared to be sensitive to price signals with respect to their problem of problem choice but 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 is compatible with the former, 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 who suggest that low-priced tasks tend to be addressed more rapidly.

In a sense, if they suggest that rational online workers simply tend to maximize their rewards by selecting easy and repeatable tasks, these results are not surprising at all, even if they point towards a somewhat different reality than the one initially envisioned by Mason & Watts (2009). However, they might also suggest a broader conclusion: that AMT, as a two-sided platform that actually allows for the existence of a finely granular labor market, might have self-selected a category of online workers that would specialize on large series of easy tasks, maybe because they have another occupation or because they need to be able to stop and switch easily between their online and their offline tasks, which could be typically the case for mothers of young children, be they in the US or in India. This interpretation clearly warrants further studies, all the more so as our results here are associated with regular workers, since they were selected according to the "Master" qualification that is granted by Amazon itself to its regular workers, and that implies a higher price billed by Amazon to Requesters (independently of the price paid to workers). However, it points toward the fact that online platform are online platforms, and that they cannot avoid all the economic and managerial consequences associated with online platforms in the digital world, be they dedicated to "human computing". Not only does price matter: furthermore, there are elements of self-selection between both sides of platforms, in relation with the difficulty of 'spot' online labor markets to convey information on the quality of workers and to deal with information asymmetries in the simplest "peaches or lemons", à la Akerlof, way; namely here, a very reasonable outcome might be that workers and requesters self-select on both sides, according to a specific kind of tasks – simple and repetitive ones in the case of AMT. A corollary, whose importance for the use of human computing in data-rich environment cannot be minimize, is then that AMT might not be an appropriate tool for other kinds of tasks, which is a straightforward conclusion but a conclusion that is not clear at all in the existing literature, though intensely burgeoning. Needless to say, there might be other platforms dedicated to other types of online work, and the market might simply have segmented itself: but we now present simulation results that point towards not only a methodology in order to analyze online labor markets, yet also to a higher level of generality of the interpretation we are suggesting.

b. Simulations show the relevance of online workers' attention

Simulation generally allows for the exploration of a larger variety of scenarios. Practically speaking, we developed a simulation tool to replicate AMT because it was thought to be useful when engaging with workers who expect to be paid notably as it allows for the testing of scenarios that might have a negative impact on the reputation of the requester if carried out in real;[8] and it allows for extending the scope of the scenarios to include information on the environment that a requester at AMT cannot control directly. In this respect, the primary aim of our simulator was to investigate what quality to expect from workers in different settings. However, results from our simulation model confirmed the importance of understanding the

[8] Requester reputation matters since workers will prefer to work for those with a good reputation (Silberman *et al*. 2010, Martin *et al*. 2014), just as worker reputation matters to requesters (Peer *et al*. 2013).

determinant of online work supply: not only with respect to its quantity and elasticity to price, but also with respect to, again, the problem of problem choice faced by online workers. How do they choose between only tasks? Indeed, our simulation exercises indirectly stressed the importance of workers online selection behaviors and of the management of their attention within the context of AMT.

The main basic features of AMT are represented in our simulation model. As on AMT, tasks – for now, only "yes or no" questions, e.g., does this pair of names refer to the same person? – are organized in batches that have in common the requester that proposed them and the remuneration for each task executed in the batch. Furthermore a requester can set a number of assignments per task. With this system, each task can be submitted more than once, and to different workers, leaving it to the requester to aggregate the results obtained. We opted for a setting in which the marketplace of requesters and workers follows a turn-based discrete-time approach. At each step, we calculate the tasks proposed by each requester. Then we simulate the activity of each worker. First, we determine the tasks the worker sees, by applying a particular filter (for example: "decreasing number of tasks in batches" or "increasing remuneration"). The worker then chooses one or several batches, depending on various criteria (including remuneration or date of creation of the batch), and answers the questions. We included in our simulator the notion of expiration for a task (called "allotted time" on AMT). After a certain time (which can be chosen by the requester), if any worker has not executed a task, it will automatically be removed from the market. In order to determine the appropriate parameter settings, model candidates were benchmarked against results from the literature as suggested by Meyer (2011). We were notably able to reproduce an important stylized fact according to which finding that the number of tasks offered at the same time affects how many of them are completed by Franklin *et al*. (2011).

We notably used this simulation model to investigate the trade-off between quality assurance through the inclusion of test questions for which the correct answer is known, and quality assurance through the assignment of the same task to multiple workers. Workers were assigned a chance of successfully completing a task of between 50 and 100% and redundancy consisted in 5 assignments per task. To aggregate answers, the answer of each worker that worked on a given task was weighted by the estimated success rate of this worker. Two alternative methods were used to update estimated success rates: (1) including test tasks into our other tasks with a set ratio, a worker estimated success rate being the percentage of test questions he answered correctly; and (2) not placing test tasks into other tasks, a worker estimated success rate being the percentage of questions for which he answered the same as the majority. In this framework, it turned out that similar results can obtain with both methods and that, consequently, requesters can arbitrate between the cost of introducing tests tasks for verification of work and redundancy.

However, the amount of redundancy that yields similar quality to a given proportion of tests tasks critically depends on parameters and notably on the "scarcity index", i.e. only the amount of only work supply, understood both globally and with respect to each requester, simply because work supply has to be big enough to ensure that all assignments are completed. Clearly, our simulation model did not include at this stage a proper supply function, yet, that would determine scarcity endogenously i.e. based on factors such as price, requester identity, size, etc. In future work, we definitely plan to incorporate the empirical results presented in the previous section in a later version of the simulator. But the existing simulation expetiments still stressed how critical work supply was, not only with respect to the mere completion of the tasks, yet also with respect to more qualitative features such as the performance of quality assurance methods, since they are themselves dependent upon the actual completion rates.

Furthermore, other critical issues showed up when we imagined an environment with pre-defined size for every new batch created each period (10, 25, 50, 100, 200, 400 tasks per batch). The total number of tasks for every type of batch was the same (800, so we had for example 16 new batches of 50 tasks by period). This non-homogeneous distribution actually led to some incoherence when workers were simulated as using different sorting functions about batch size. We understood that we actually lacked information about the sorting functions that workers use, "biggest first" (in terms of numbers of HITs) being the default on AMT's landing page, at least as of 2014. Nor do we have clues of the number of pages a worker will see on AMT before choosing a task in a batch. As a consequence, we cannot control for the relative attractiveness of the batches that appear first according to the dominant sorting function in the simulations.

This points towards is a much more general consequence: we do not know how workers on AMT actually *choose* their tasks, not only with respect to price or to the number of HITs available (Size) but also, more generally, to the way their *attention* is managed in this online environment. Empirical findings reported in the previous section with respect to the influence of Size of batches can result not only from dynamic wage maximization from workers, but also from some behavioral consequences of the way attention is managed online, and of the way it affects online workers' problem of problem choice. Obviously, both interpretations can coexist and reinforce one another: it might be because AMT is populated mainly by workers seeking large batches of simple tasks that the default presentation favors big batches, and reciprocally. Here again, positive reinforcing feedbacks might have been, and might be, at work within AMT, with critical consequences with respect to how AMT can be used as a human computing platform in data-rich ecosystems, since AMT is, first and foremost, a two-sided platform that has to manage, even implicitly, the attention of online workers.

4. Conclusion

Human computing could considerably help implement semi-supervised machine learning methods in data-rich ecosystems, a key element of big data issues. However, it implies not only to find appropriate ways to organize this problem as the resolution of micro-tasks distributed among multiple workers on line through online platforms like AMT: it also implies for "requesters" to find suitable workers via online labor markets, the existence and the characteristics of which strongly depend on the platforms themselves – a conclusion that is not at all a surprise to economists and management scientists. However, whatever the many pitfalls already addressed in the intensely burgeoning literature on these matters, probably as a consequence of the importance of the promises of crowdsourcing and human computing in data-rich environments, and the many others on their way to be addressed, our own practical experience and experiments via AMT, about name disambiguation techniques, about pricing and through simulation, when analyzed through the lenses of the difficulties faced, provides evidence for a somewhat different reality.

Whatever the second-order optimization techniques that have been, are and can be developed with respect to using human computing through online platforms, further difficulties might still always arise until the first-order conditions associated with the economic and managerial nature of such environments acting as online labor markets, are neglected. Simply said, AMT is a multi-sided platforms and it is highly probable that workers have been self-selected along with tasks offered in a dynamic manner, maybe in comparison with other platforms, like oDesk, whose history might have been different; and even under these inherited constraints, the management of the attention of workers does play a crucial role. In concluding so, we do not feel to have unearthed a blandly new phenomenon, but we hope that we could contribute to shift the direction in which research is currently conducted with respect to human computing in a direction closer to economic and managerial sciences. A straightforward consequence is then that more studies are now needed in this respect, to which we hope we can contribute: a consequence which, though straightforward, is in this case, we believe, of acute relevance.

ACKNOWLEDGEMENTS

We would like to thank Nicolas Maisonneuve, Francesco Lissoni and Pluvia Zuniga. We acknowledge funding from the European Science Foundation (ESF) Research Networking Program Academic Patenting in Europe (APE-INV); the 2012 short-term visiting fellows program of CSIC Institute of Public Goods and Policies (CSIC-IPP); the Spanish Ministry of Economy and Competitiveness (MINECO) for the Spanish National Plan projects CSO2009-10845 and CSO2012-32844; and the Agence Nationale de Recherche (ANR) for project Ocktopus. In addition, this research received financial support from the same agency through the program "Investments for the Future" under reference number ANR-10-LabX-11-01. An earlier version of this work was presented at the 14th International Society of Scientometrics and Informetrics Conference (ISSI) in Vienna, Austria, 2013.

REFERENCES

Acemoglu, D., Mostagir, M., & Ozdaglar, A. (2014). *Managing Innovation in a Crowd* (No. w19852). National Bureau of Economic Research.

Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D., Crowell, D. & Panovich K. (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, 2010. doi:10.1145/1866029.1866078.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679.

Brynjolfsson, E., Geva, T., & Reichman, S. (2014). Crowd-Squared: Amplifying the Predictive Power of Large-Scale Crowd-Based Data. *Available at SSRN.*

Chandler, J., Mueller, P., Paolacci, G. (2014), Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers., *Behavior Research Methods*. DOI 10.3758/s13428-013-0365-7

Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2013). Big data challenge: a data management perspective. *Frontiers of Computer Science*, *7*(2), 157-164.

Chen, J. J., Menezes, N. J., Bradley, A. D., & North, T. A. (2011). Opportunities for crowdsourcing research on Amazon Mechanical Turk. *Interfaces*, *5*(3).

Cuxac, P., Lamirel, J. C., & Bonvallot, V. (2012). Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics*, 1-12.

Czarnitzki, D., Hussinger, K, & Schneider, C. (2007), "Commercialising academic research: the quality of faculty patenting", *Industrial and Corporate Change*, 20, 1403-1437.

Day, George S. "Closing the marketing capabilities gap." *Journal of Marketing* 75.4 (2011): 183-195.

Den Besten, M., Martinez, C., Maisonneuve, N. & Maraut, S. (2013), Crowdsourcing the Names-Game: a prototype for name disambiguation of author-inventors. ISSI 2013 Proceedings.

Dornbusch, F., Schmoch, U., Schulze, N. & Bethke, N. (2013), Identification of university-based patents: A new large-scale approach, *Research Evaluation*, 22(1), 52-63.

Demartini, G., Difallah, D.E., Cudré-Mauroux, P., 2012. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking, in: Proceedings of the 21st International Conference on World Wide Web, WWW '12. ACM, New York, NY, USA, pp. 469–478.

Eickhoff, C., & Vries, A. P. de. (2013). Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, *16*(2), 121–137. doi:10.1007/s10791-011-9181-9

Ekbia, H., & Nardi, B. (2014). Heteromation and its (dis) contents: The invisible division of labor between humans and machines. *First Monday*, *19*(6).

Ester, M., Kriegel, H.P., Sander, J. & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, U. M. Fayyad (Ed.), Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.

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, 61–72. SIGMOD '11. New York, NY, USA: ACM, 2011. doi:10.1145/1989323.1989331.

Franzoni, C., & Sauermann, H. (2014). Crowd Science: The Organization of Scientific Research in Open Collaborative Projects, *Research Policy, 43*(1), 1-20. http://dx.doi.org/10.1016/j.respol.2013.07.005

Geiger, D., & Schader, M. (2014). Personalized task recommendation in crowdsourcing information systems—Current state of the art. *Decision Support Systems*.

Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., & Schader, M. (2011). Managing the crowd: towards a taxonomy of crowdsourcing processes. *AMCIS 2011 Proceedings - All Submissions*. Paper 430. Available at http://www.crowdsourcing.org/document/managing-the-crowd-towards-a-taxonomy-of-crowdsourcing-processes/5534

Georgescu, M., Pham, D. D., Firan, C. S., Nejdl, W., & Gaugaz, J. (2012). Map to humans and reduce error: crowdsourcing for deduplication applied to digital libraries. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1970–1974). New York, NY, USA: ACM. doi:10.1145/2396761.2398554

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213-224.

Gurney, T., Horlings, E., & Van Den Besselaar, P. (2012). Author disambiguation using multi-aspect similarity indicators. *Scientometrics*, *91*(2), 435-449.

Horton, J. & Chilton, L. (2010) "The Labor Economics of Paid Crowdsourcing." 1001.0627 (January 5, 2010). http://arxiv.org/abs/1001.0627.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399-425.

Hirth, M., Hoßfeld, T., & Tran-Gia, P. (2013). Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Mathematical and Computer Modelling*, *57*(11), 2918-2932.

Ipeirotis, P. (2009). When to Post Tasks on Mechanical Turk? | A Computer Scientist in a Business School [WWW Document], URL http://www.behind-the-enemy-lines.com/2009/08/when-to-post-tasks-on-mechanical-turk.html (accessed 1.28.13).

Ipeirotis, P. (2010a). Getting High Quality Results on MTurk | A Computer Scientist in a Business School [WWW Document]. URL http://www.behind-the-enemy-lines.com/2010/03/getting-high-quality-results-on-mturk.html (accessed 1.28.13).

Ipeirotis, P. (2010b). Be a Top Mechanical Turk Worker: You Need $5 and 5 Minutes | A Computer Scientist in a Business School [WWW Document]. URL http://www.behind-the-enemy-lines.com/2010/10/be-top-mechanical-turk-worker-you-need.html (accessed 1.28.13).

Ipeirotis, P. G. (2010c). Analyzing the Amazon Mechanical Turk marketplace. ACM XRDS, 17, 16–21.

Ipeirotis, P. (2012). Mechanical Turk changing the defaults: The game has changed | A Computer Scientist in a Business School [WWW Document]. URL http://www.behind-the-enemy-lines.com/2012/12/mechanical-turk-changing-defaults-game.html (accessed 1.28.13a).

Irani, L. C., & Silberman, M. (2013, April). Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 611-620). ACM

Kittur, A., Chi, E.H. & Suh, B. (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, 2008. doi:10.1145/1357054.1357127.

Kittur, A., Nickerson, J., Bernstein, M.,Gerber, E., Shaw, A., Zimmerman, J., Lease, M. & Horton, J. (2012), The Future of Crowd Work SSRN Working Paper, December 19, 2012. http://papers.ssrn.com/abstract=2190946.

Lauritzen, G. D., Salomo, S., & La Cour, A. (2013). Dynamic boundaries of user communities: exploiting synergies rather than managing dilemmas. *International Journal of Technology Management*, *63*(3), 148-168.

Lehdonvirta, V., & Mezier, P. (2013). Identity and Self-Organization in Unstructured Work. The Dynamics of Virtual Work Working Paper 1, University of Herftfordshire, UK.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, *57*(1), 34-52.

Lissoni, F., Llerena, P., McKelvey, M. & Sanditov, B. (2008), “Academic patenting in Europe: new evidence from the KEINS database”, *Research Evaluation*, 17, 2, 87-102.

Lissoni, F., Lotz, P., Schovsbo, J. & Treccani, A. (2009), "Academic patenting and the professor's privilege: evidence on Denmark from the KEINS database," *Science and Public Policy*, 36, 8, 595-607.

Litman, L., Robinson, J., & Rosenzweig, C. (2014). The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior Research Methods*, 1-10.

Maisonneuve, N., & Chopard, B. (2012). Crowdsourcing satellite imagery analysis: study of parallel and iterative models. In *Geographic Information Science* (pp. 116-131). Springer Berlin Heidelberg.

Marge, M., Banerjee, S. & Rudnicky, A.I. (2010). “Using the Amazon Mechanical Turk to Transcribe and Annotate Meeting Speech for Extractive Summarization.” In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, 99–107. CSLDAMT ’10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. http://dl.acm.org/citation.cfm?id=1866696.1866712.

Maraut, S. & Martinez, C. (2014), “Identifying author – inventors from Spain: methods and a first insight into results”, *Scientometrics*, 101, 445-476doi: 10.1007/s11192-014-1409-1..

Martin, D., Hanrahan, B. V., O'Neill, J., & Gupta, N. (2014, February). Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 224-235). ACM.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. doi:10.3758/s13428-011-0124-6

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

Mason, W., & Watts, D.J. (2011). "Collective Problem Solving in Networks." SSRN eLibrary (January 19, 2011). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1795224.

McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60.

Meyer, M. (2011). Bibliometrics, Stylized Facts and the Way Ahead: How to Build Good Social Simulation Models of Science? *Journal of Artificial Societies and Social Simulation*, *14*(4), 4.

Noyons, E., Buter, R., van Raan, A., Schmoch, U., Heinze, T., Hinze, S. & Rangnow, R. (2003a), "Mapping excellence in science and technology across Europe Nanoscience and Nanotechnology", Report of project EC-PPN CT-2002-0001 to the European Commission, October 2003.

Noyons, E., Buter, R., van Raan, A., Schmoch, U., Heinze, T., Hinze, S. &. Rangnow, R. (2003b), "Mapping excellence in science and technology across Europe Life Sciences", Report of project EC-PPLS CT-2002-0001 to the European Commission, October 2003.

Peer, E., Vosgerau, J., & Acquisti, A. (2013). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods*, 1-9.

Raasch, C. (2011). The sticks and carrots of integrating users into product development. *International Journal of Technology Management*, *56*(1), 21-39.

Raffo, J. & Lhuillery, S. (2009), "How to play the "Names Game": Patent retrieval comparing different heuristics" *Research Policy*, 38 (2009) 1617–1627

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179. doi:10.1016/j.jtbi.2011.03.004

Rong, K., Lin, Y., Shi, Y., & Yu, J. (2013). Linking business ecosystem lifecycle with platform strategy: a triple view of technology, application and organisation. *International Journal of Technology Management*, *62*(1), 75-94.

Royle, J., & Laing, A. (2014). The digital marketing skills gap: Developing a Digital Marketer Model for the communication industries. *International Journal of Information Management*, *34*(2), 65-73.

Schmoch, U. (2007), Patentanmeldungen aus deutschen Hochschulen, Studien zum deutschen Innovationsystem 10-2007, Karlsruhe.

Schmoch, U., Dornbusch, F., Mallig, N., Michels, C., Schulze, N. & Bethke, N. (2012), Vollstandige Erfassung von Patentanmeldungen aus Universitaten. Bericht an das Bundesministerium fur Bildung und Forschung (BMBF). Revidierte Fassung, Karlsruhe: Fraunhofer ISI.

Schulze, T., Krug, S., & Schader, M. (2012). Workers' task choice in crowdsourcing and human computation markets.*Thirty Third International Conference on Information Systems*, Orlando 2012

Shaw, A.D., Horton, J.J., Chen, D.L., 2011. Designing incentives for inexpert human raters, in: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, CSCW '11. ACM, New York, NY, USA, pp. 275–284.

Silberman, M., Ross, J., Irani, L., & Tomlinson, B. (2010, July). Sellers' problems in human computation markets. In *Proceedings of the acm sigkdd workshop on human computation* (pp. 18-21). ACM.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual review of information science and technology*, *43*(1), 1-43.

Snow, R., O'Connor, B., Jurafsky, D. & Ng, A.Y. (2008). "Cheap and Fast—but Is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks." In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 254–263. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. http://dl.acm.org/citation.cfm?id=1613715.1613751.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167. doi:10.3758/s13428-010-0039-7

Striukova, L., & Rayna, T. (2015). Open innovation 2.0: Is co-creation the ultimate challenge?. *International Journal of Technology Management*.

Thursby, J., Fuller, A. W. & Thursby, M. (2009) "US faculty patenting: Inside and outside the university," *Research Policy,* 38,1, 14-25.

Trajtenberg M., Shiff G. &. Melamed, R. (2006), "The 'names game': Harnessing inventors' patent data for economic research", NBER working paper 12479.

Vakharia, D., & Lease, M. (2013). *Beyond AMT: An Analysis of Crowd Work Platforms* (arXiv e-print No. 1310.1672). Retrieved from http://arxiv.org/abs/1310.1672

Veve, M. (2009). Supporting Name Authority Control in XML Metadata: A Practical Approach at the University of Tennessee, *Library Resources & Technical Services*, 53, 1, 41-52.

Von Proff, S., Buenstorf, G. & Hummel, M. (2012), "University Patenting in Germany before and after 2002: What Role Did the Professors' Privilege Play?", *Industry and Innovation*, 19: 23-44.

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F., & Pinheiro, D. (2012a). A boosted-trees method for name disambiguation. *Scientometrics*, 93(2), 391-411.

Wang, J., Kraska, T., Franklin, M. J., & Feng, J. (2012b). CrowdER: crowdsourcing entity resolution. *Proc. VLDB Endow.*, *5*(11), 1483–1494.

Winkler, W. E. (2006). Overview of record linkage and current research directions, Statistical Research Division U.S. https://www.census.gov/srd/papers/pdf/rrs2006-02.pdf. Retrieved Jan 2014.

Zittrain, J. (2008). Ubiquitous human computing. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 366(1881), 3813-3821.