A Taxonomy of Microtasks on the Web

Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze
{gadiraju,kawase,dietze}@L3S.de
L3S Research Center, Leibniz Universität Hannover, Germany

ABSTRACT

Nowadays, a substantial number of people are turning to crowdsourcing, in order to resolve tasks that require human intervention. Despite a considerable amount of research done in the field of crowdsourcing, existing works fall short when it comes to classifying typically crowdsourced tasks. Understanding the dynamics of the tasks that are crowdsourced and the behaviour of workers, plays a vital role in efficient task-design. In this paper, we propose a two-level categorization scheme for tasks, based on an extensive study of 1000 workers on CrowdFlower. In addition, we present insights into certain aspects of crowd behaviour; the task affinity of workers, effort exerted by workers to complete tasks of various types, and their satisfaction with the monetary incentives.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Crowdsourcing; Microtasks; Taxonomy; Incentive; Effort; Affinity

1. INTRODUCTION

Crowdsourcing is evolving rapidly as a means for solving problems that require human intelligence or human intervention. Over the last decade, there has been a considerable amount of work towards establishing suitable platforms and proposing frameworks for fruitful crowdsourcing. Amazon's Mechanical Turk[1] and CrowdFlower[2] are exemplary reflections of such platforms.

A large number of researchers have used these platforms in order to gather distributed and unbiased data, to validate results or to build ground truths. However, the literature inspecting the actors involved in the crowsourcing process is rather scarce. Only a few noteworthy works have investigated best practices, the reliability of data [4], or have proposed comprehensive strategies and guidelines [7] (see Section 2).

As a consequence, without adequate knowledge of how one can effectively and efficiently exploit the wisdom of the crowd through crowdsourcing platforms, several research works are hindered by chaotic results leading to doubtful conclusions. Thus, it is essential that task administrators[3] are, to some extent, educated in crowdsource task modeling. To pave a way towards solving this problem, we present our work comprising the categorization of crowdsourced tasks, and reflect on guidelines and principles of crowsourcing tasks.

First, based on a crowdsourced survey, we present strong evidence that a large number of *workers*[4] are untrustworthy. This evidence shows that simple gold-standards (See Section 4) might not be enough to provide reliable data.

Second, after a manual exhaustive selection of reliable responses, we provide a data analysis showing that, in fact, *workers* are in a modern *gold-rush*, prioritizing monetary reward over their affinity to tasks.

Finally, we venture into determining a fine-grained goal-oriented categorization of crowdsourced tasks. This facilitates a greater understanding of the dynamics between the administrators of such tasks and the workers in the crowd, thereby increasing reliability of the results and data quality.

2. RELATED WORK

Marshall et al. profile Turkers who take surveys, and examine the characteristics of surveys that may determine the data reliability [4]. Similar to their work, we adopt the approach of collecting data through crowdsourced surveys in order to draw meaningful insights.

Kazai et al. [3], use behavioural observations to define the types of workers in the crowd. They type-cast workers as either *sloppy*, *spammer*, *incompetent*, *competent*, or *diligent*. By doing so, the authors expect their insights to help in designing tasks and attracting the best workers to a task. Along the same lines, Ross et al. [6] study the demographics and usage behaviors, characterizing workers on Amazon's Mechanical Turk. Complementing such existing works, as well as in contrast, our work focuses on task modeling rather

[1]https://www.mturk.com/mturk/
[2]http://www.crowdflower.com/

[3]A user responsible for deploying tasks on a crowdsourcing platform.
[4]A user that performs tasks for monetary rewards on a crowdsourcing platform.

HT'14, September 1–4, 2014, Santiago, Chile.

http://dx.doi.org/10.1145/2631775.2631819 .

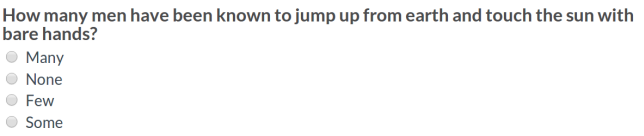

Figure 1: Engaging workers and checking their alertness by using questions.

than user modeling. Nevertheless, we hypothesize that the consideration of both aspects is essential for effective crowdsourcing.

In their work, Yuen et al. present a literature survey on different aspects of crowdsourcing [8]. In addition to a taxonomy of crowdsourcing research, the authors present a humble example list of application scenarios. Their short list represents the first steps towards task modeling. However, without proper organization regarding types, goals and work-flows, it is hard to reuse such information to devise strategies for task design. We solve this issue by providing an articulated categorization in terms of goals and workflows.

In the realm of studying the reliability of crowd workers, and gauging their performance with respect to the incentives offered, Mason et al. investigate the relationship between financial incentives and the performance of the workers [5]. They find that higher monetary incentives increase the quantity but not the quality of work of the crowd workers. A large part of their results align with our findings presented in Section 4.

In their work, Geiger et al. propose a taxonomic framework for crowdsourcing processes [2]. Based on 46 crowdsourcing examples they conceive a 19-class crowdsourcing process classification. As stated by the authors, they focus exclusively on an organizational perspective, thus providing valuable insights mainly for stakeholders running crowdsourcing platforms. With a different focus, our proposed categorization intends to primarily assist microtask administrators in effectively using such platforms.

3. APPROACH

We aim to analyze tasks that are typically crowdsourced by exploiting response-based data from the workers.

We deployed a survey using the CrowdFlower platform in order to gather information about typically crowdsourced jobs. To begin with, the survey consisted of questions regarding the demographics, educational and general background of the workers. Next, questions related to previous tasks that were successfully completed by the workers, are introduced. The survey consisted of a mixture of open-ended, direct, and Likert-type questions designed to capture the interest of the workers. We restrict the participation to 1000 workers. We ask the crowd workers open-ended questions, about two of their most recent successfully completed tasks. State-of-the-art qualitative research methods [1], have indicated that relying on recent incidents is highly effective, since respondents answer such questions with more details and instinctive candor. We pay all the contributors from the crowd, irrespective of whether or not we discard their data for further analysis.

In addition, to keep the participants engaged we intersperse the regular questions with humour-evoking and amusing bits as shown in Figure 1. At the same time, these questions are also used to filter out spammers or workers with malicious intentions. We do not use other sophisticated means to curtail regular crowd worker behaviour, in order to capture a realistic composition of workers (both trustworthy and otherwise), although malicious workers are discarded from our data analysis.

4. DATA ANALYSIS

In this section we present our findings from the analysis of the data, collected through the crowd sourcing process described in Section 3.

From the 1000 workers that participated in the survey, we consider the responses from 490 in our analysis. The responses from 433 workers are pruned out of consideration based on their failure to pass at least one of the so called '*gold standard*' tests. A gold standard question, is one that is designed in order to prevent malicious workers from either progressing in a task, or to identify and discard such workers during analysis. For example, consider the question in Figure 1. Malicious workers often pick and choose from the available options at random. By using two such hidden tests, we prune out workers with ulterior intentions.

We manually curated the responses from the remaining workers, and found that 77 workers tried to cheat their way through to task completion by copy-pasting the same bits in response to all open-ended questions.

Of the 490 trusted workers, 76% were found to be *Male* participants while 24% were *Female*. The average age of the male and female crowd workers was similar, at 33.7 and 33.1 years respectively. We found that 88% workers cared about their reputation as crowd workers, while 12% workers claimed that they did not care about their reputation. These workers contributed to the description of 980 tasks that they successfully completed in the past. Workers claimed that in hindsight, they could have performed better in 534 of these instances, while they responded that they could not have performed better in 282 of those tasks. Workers were unsure about a possible improvement in their performance corresponding to 164 tasks.

4.1 How do workers choose their tasks?

An interesting research question, which we set out to answer through this work, was to find out what factors influence a worker's choice in the tasks she picks to complete. Based on our survey, we gather the three most commonly stated factors that determine a worker's choice in task. An indicator could be the *monetary incentives* offered on task completion. The *interestingness* of the task itself and the *time required* to complete a task are the other factors that surfaced. However, the distribution of significance of these factors is not known. In order to determine this, we capture the responses from the workers for this question on a 5-point Likert-scale from *No Influence: 1* to *Strong Influence: 5*.

The aggregated results of the Likert-scale show that *monetary reward* (4.02) is significantly ($p < 0{,}01$) the most crucial factor for a worker while determining which task to complete. The factors *Time required* (3.76) for the completion of a task and *topic* (3.69) come in next with marginal difference between them.

Apart from the factors captured on the Likert-scale, we posed additional questions to the workers regarding why they chose to complete the particular tasks that they de-

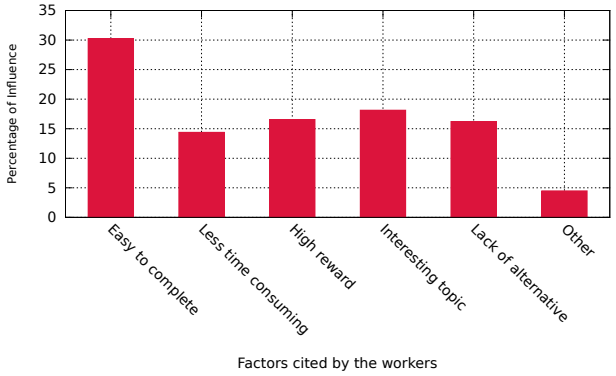

Figure 2: Factors that determine workers' choice of task based on their most recently completed tasks.

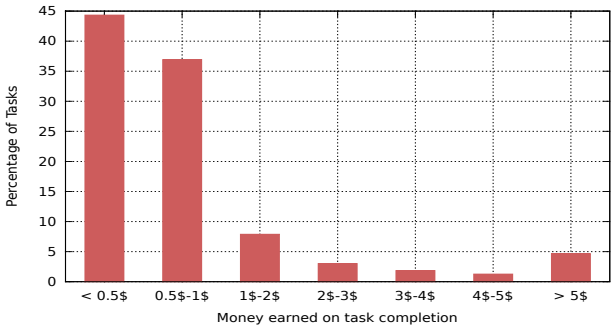

Figure 3: Money earned by workers on completing various tasks.

scribed in the survey (the workers' two most recently completed tasks). Figure 2 quantifies our findings. The *ease of completion* of a task is a driving force in the task selection process of a worker. An *interesting topic*, a *high reward*, a *less time consuming* task also play a role in the choice of task of a crowd worker, albeit to a less prominent extent. It is interesting to note that a significant number of crowd workers end up completing tasks due to the *lack of other alternatives*. Additionally, we facilitated for an open-ended response when workers chose the *Other* option. Through this we found a few other minor reasons that workers cited for choosing to complete their previous tasks. For example, a few workers said they wanted to increase their overall profile accuracy, i.e. their reputation on the crowdsourcing platform.

Considering that the monetary reward is highly influential in the workers' choice of completing tasks, it is interesting to investigate the precise amount of rewards offered by the tasks. Figure 3 presents the distribution of the money earned by the workers on completion of the 980 tasks, considered in the analysis. We note that most tasks that are deployed for crowdsourcing, offer either meagre (<0.5$) or small monetary rewards (between 0.5$ and 2$). This is reasonable from the point of view of the task deployers or administrators, since most of these tasks do not require a lot of effort from the crowd workers, as confirmed from our analysis (see Figure 4).

We clearly see that the tasks that offer bigger monetary awards (>3$) are also coincidentally the tasks in which the

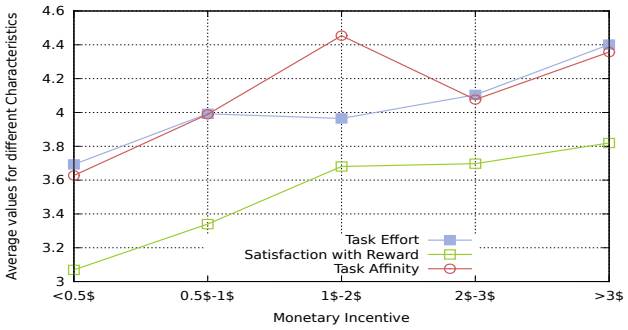

Figure 4: Distribution of effort required, task affinity, and satisfaction with reward of the workers with respect to varying task incentives.

crowd workers are required to exert the most amount of effort. From this, we infer that the monetary rewards offered for the tasks that are typically crowdsourced, are proportional to the amount of effort that is expected for task completion by the crowd workers.

4.2 Task Affinity vs Incentive

We define *task affinity* as the tendency of a crowd worker to like the task she chooses to complete. Next, we investigate how the incentive for a given task influences the task affinity of a crowd worker. From our analysis, presented in the Figure 4 we observe that there are two subtle kinds of behaviour exhibited by the workers in the crowd. Crowd workers tend to exhibit a greater affinity to those tasks which offer higher incentives (>3$). This is understandable since the workers can earn more money by completing such tasks. On the other hand, crowd workers also depict a significant amount of affinity for tasks that offer a reasonable amount of incentive (between 1$ and 2$). This can be explained by the fact that, these tasks require significantly lesser effort from the workers.

An interesting point to note, is that although the most number of tasks deployed on crowd sourcing platforms fall under the bracket of relatively low monetary incentives, thus resulting in such tasks being completed by most workers, the workers' affinity towards these given tasks is considerably low.

We consider that a workers' approval of the monetary reward corresponding to a given task, may be subject to change on task completion. This may be attributed to the difference in the amount of effort needed for task completion when compared to the anticipated effort by a crowd worker. We capture the average satisfaction of the crowd workers with the reward they receive on task completion. Our aggregated results are presented in Figure 4. We observe that the satisfaction is proportional to the incentive of the reward that is offered.

5. CATEGORIZATION OF TASKS

From the responses collected through the crowdsourced survey, we have manually established the following classes that describe typically crowdsourced tasks. The example task descriptions presented alongside each type of task, are extracted from the responses received from workers regarding their previous micro-tasks. We categorize the tasks into

6 high-level goal-oriented classes as presented next, with each class containing sub-classes of other types of tasks. The high-level categorization is drawn based on the 'goals' of a task, while the sub-classes are based on the 'work-flow' of tasks.

5.1 Categorization Scheme

- **Information Finding**(IF)- Such tasks delegate the process of searching to satisfy one's information need, to the workers in the crowd. For example, '*Find information about a company in the UK*', or '*Find the cheapest air fare for the selected dates and destinations*'.
- **Verification and Validation**(VV)- These are tasks that require workers in the crowd to either verify certain aspects as per the given instructions, or confirm the validity of various kinds of content. For example, '*Is this a Spam Bot? : Check whether the twitter users are either real people or organisations, or merely spam twitter user profiles*', or '*Match the names of personal computers and verify corresponding information*'.
- **Interpretation and Analysis**(IA)- Such tasks rely on the wisdom of the crowd to use their interpretation skills during task completion. For example, '*Choose the most suitable category for each URL*', or '*Categorize reviews as either positive or negative*'.
- **Content Creation**(CC)- Such tasks usually require the workers to generate new content for a document or website. They include authoring product descriptions or producing question-answer pairs. For example, '*Suggest names for a new product*', or '*Translate the following content into German*'.
- **Surveys**(S)- Surveys about a multitude of aspects ranging from demographics to customer satisfaction are crowdsourced. For example, '*Mother's Day and Father's Day Survey (18-29 year olds only!)*'
- **Content Access**(CA)- These tasks require the crowd workers to simply access some content. For example, '*Click on the link and watch the video*', or '*Read the information by following the website link*'. In these tasks the workers are merely asked to consume some content by accessing it, but do nothing further.

It is important to note that, in certain cases it may be possible for a particular job to belong to more than one of the aforementioned classes. For example, a survey about the perception of a product like the new iPhone from Apple could belong to the classes of *Surveys* as well as *Sentiment Analysis*.

Apart from these high-level categorization based on the goals of the tasks, Table 1 presents some sub-classes of the high-level classes, which are based on the work-flow of the tasks. Some sub-classes are explained below.

- Class *IF*/**Metadata Finding**- Such tasks require the users to find specific relevant information from a given data source. For example, '*Find e-mail addresses of corresponding employees from the company's websites*'.
- Class *VV*/**Content Verification**- In these tasks the crowd workers are required to verify, validate, qualify, or disqualify different aspects as dictated by the task administrators. For instance, '*Check if the following company websites describe the correct business*'.
- Class *IA*/**Categorization and Classification**- Such tasks involve the organization of entities into groups with the same features, or assigning entities to classes according to a predetermined set of principles. For example, '*Choose the most suitable category for each URL*'.
- Class *IA* or *CC*/**Media Transcription**- These tasks require the crowd workers to transcribe (put into written form) different media like images, music, video, and so forth. For example, '*See the images and find the year on which the wine bottle was manufactured*'. The tasks also include transcribing captchas. For instance, '*Type what you see in the following Captchas*'.
- Class *IA*/**Ranking**- Here the crowd workers are required to determine the most relevant entities with respect to the search query. For example, '*Search for the given terms and click on the best three results*'.
- Class *IA*/**Content Moderation**- Here the workers are required to moderate content for guideline violations, inappropriate content, spam, or others. Independent of the kind of media (text, photos, or videos), the crowd is asked to evaluate the content against a set of rules. For example, '*Moderate images for inappropriate content (sexually explicit content)*'.
- Class *IA*/**Sentiment Analysis**- Tasks that pertain to the assessment of the sentiment towards an entity or notion, fall under this category. For example, '*What do you think of the new Samsung tablet?*', or '*Identify if the tweets are positive, negative, or neutral*'.
- Class *CC*/**Data Collection and Enhancement**- Crowdsourcing is used to generate and enhance data. For instance, the crowd has been used in the past to create a dataset of colours by asking workers to annotate different hues and shades with labels[5].
- Class *S*/**Content Feedback**- In such tasks workers are asked to assess and provide feedback about products, entities, websites, and so forth. For example, '*Help us improve our website*'.
- Class *CA*/**Promoting**- In such tasks workers are asked to access and consume content. For example, '*Visit the webpage by clicking on the provided link*'.

5.2 Tasks as per Categorization Scheme

Based solely on the reliable data collected during the crowdsourcing process, we manually annotated each of the workers' previously completed tasks according to the categorization scheme. Figure 5 presents the distribution of these tasks, as per their categorization into the different proposed classes.

Note that certain tasks can rightly be classified into more than one class. For example, consider the task, '*Search for spam-like comments in the following content*'. This task can be classified into the classes *Verification and Validation*, as well as *Information Finding*. This is because, the goal of such a task could be either to ensure the content is spam-free, or merely find the spam comments. Consider the task, '*Identify biographies in the following*'. This task can be classified into the class *Interpretation and Analysis* since the identification of a biography relies on the workers interpretation of the classification. At the same time, the task can be classified into *Verification and Validation*, since the goal of the task could be to validate biographies.

[5] `http://www.crowdflower.com/blog/2008/03/our-color-names-data-set-is-online`

Table 1: Sub-classes of the proposed categorization for typically crowdsourced tasks.

Information Finding	Verification & Validation	Interpretation & Analysis	Content Creation	Surveys	Content Access
Metadata finding	Content Verification	Classification	Media Transcription	Feedback/Opinions	Testing
	Content Validation	Categorization	Data Enhancement	Demographics	Promoting
	Spam Detection	Media Transcription	Translation		
	Data matching	Ranking	Tagging		
		Data Selection			
		Sentiment Analysis			
		Content Moderation			
		Quality Assesment			

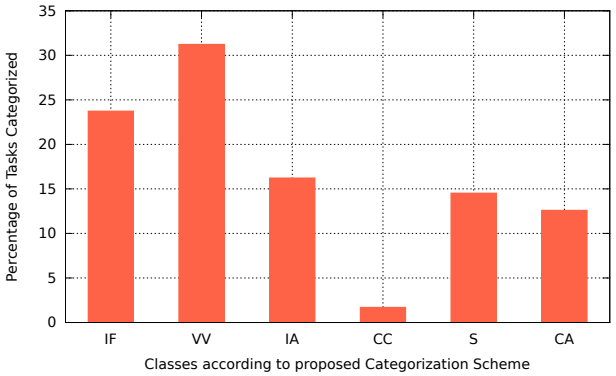

Figure 5: Distribution of tasks in the classes of the proposed Categorization Scheme.

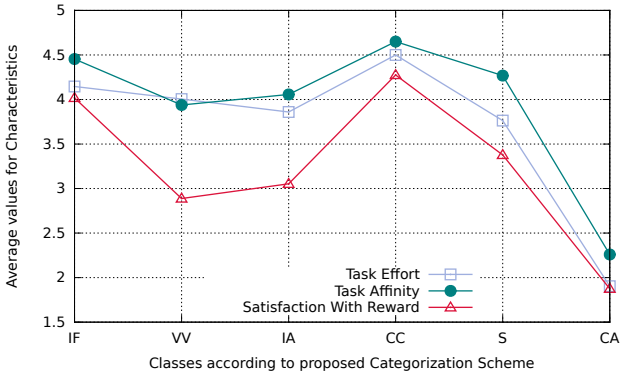

Figure 6: Distribution of task-related characteristics according to the proposed Categorization Scheme.

As a next step, we analyze the average effort that a worker needs to exert to complete a task, the task affinity, as well as the workers' satisfaction with the reward, for each of the high-level classes. The findings are presented in Figure 6. Understandably, tasks of the class *Content Creation* require the most amount of effort from the crowd workers, while those of the class *Content Access* require the least amount of effort. It is interesting to note that crowd workers like to work on tasks of the class type *Information Finding* and *Surveys* in addition to *Content Creation*. The most disparity between the effort exerted for task completion by the workers and their satisfaction with the reward, corresponds to the classes of *Verification and Validation*, and *Interpretation and Analysis*.

5.3 Tasks with Ulterior Motives

During our manual analysis of the data collected we identified several tasks with deceitful hidden motives. While such tasks may indicate legitimacy to some extent due to the work-flow they suggest to workers, there is a clear ulterior goal of deliberately manipulating third party results. For example, improving the popularity or general sentiment of particular content. In most cases, these tasks fall in the classes *Content Access* and *Content Creation*, further being masked with an additional goal such as a *Survey*. For example, '*Search for some particular terms in Google, and click on the link of our Website*', '*Watch this video on Youtube and click like*', or '*Give a five start rating to this product*'.

We also verified that in many circumstances, these tasks are followed by a survey which contains a failure guaranteed gold standard. For example, '*What's your age?*', whereas the only correct answer is an unrealistic number. At this point, the malicious task administrator has already collected his desired data and the system prevents workers from getting their reward. As a principle, crowdsourcing platforms discourage the deployment of such tasks. We hypothesize that our modeling categorization can improve the identification of deceitful tasks by the system and also by potential workers. This is a critical issue to be addressed in order to improving crowdsourcing practice.

6. CONCLUSIONS & FUTURE WORK

In this paper, we presented a meta-crowdsourcing profiling study. First, we identified high levels of malicious workers. Almost 44% of the workers did not manage to correctly answer simple attention check questions. It is important to highlight that we deliberately model our crowdsourcing task to allow such malicious behavior. These results highlight the importance of gold standards in crowdsourcing tasks and need for intelligent task modeling strategies.

At the same time, based on the manually verified reliable responses, we collected enough data to characterize workers' behaviour and preferences. Further, we thoroughly study the types of tasks that are typically crowdsourced, and as a result, we propose a goal-oriented *Categorization Scheme* for crowdsourced tasks. A fine-grained categorization of crowdsourced tasks has important implications for the user modeling of crowd workers, and recommendation of tasks. The proposed categorization of tasks, including the findings from our extensive analysis, aid in future task design and deployment. By drawing from our findings related to the task dependent characteristics, like task affinity, task effort, incentive required, and so forth, one can design tasks with higher success rates (i.e., maximizing the quality of the results with respect to the given reward).

7. REFERENCES

[1] Corbin, J., and Strauss, A. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage, 2008.

[2] Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., and Schader, M. Managing the crowd: Towards a taxonomy of crowdsourcing processes. In *AMCIS* (2011).

[3] Kazai, G., Kamps, J., and Milic-Frayling, N. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (2011), ACM, pp. 1941–1944.

[4] Marshall, C. C., and Shipman, F. M. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference* (New York, NY, USA, 2013), WebSci '13, ACM, pp. 234–243.

[5] Mason, W., and Watts, D. J. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter 11*, 2 (2010), 100–108.

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

[7] Willett, W., Heer, J., and Agrawala, M. Strategies for crowdsourcing social data analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2012), CHI '12, ACM, pp. 227–236.

[8] Yuen, M.-C., King, I., and Leung, K.-S. A survey of crowdsourcing systems. In *PASSAT/SocialCom 2011, Privacy, Security, Risk and Trust (PASSAT), 2011 IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, MA, USA, 9-11 Oct., 2011* (2011), pp. 766–773.