Competitive Game Designs for Improving the Cost Effectiveness of Crowdsourcing

Markus Rokicki, Sergiu Chelaru, Sergej Zerr, Stefan Siersdorfer
L3S Research Center, Hannover, Germany
{rokicki,chelaru,siersdorfer,zerr}@L3S.de

ABSTRACT

Crowd based online work is leveraged in a variety of applications such as semantic annotation of images, translation of texts in foreign languages, and labeling of training data for machine learning models. However, annotating large amounts of data through crowdsourcing can be slow and costly. In order to improve both cost and time efficiency of crowdsourcing we examine alternative reward mechanisms compared to the "Pay-per-HIT" scheme commonly used in platforms such as Amazon Mechanical Turk. To this end, we explore a wide range of monetary reward schemes that are inspired by the success of competitions, lotteries, and games of luck. Our large-scale experimental evaluation with an overall budget of more than 1,000 USD and with 2,700 hours of work spent by crowd workers demonstrates that our alternative reward mechanisms are well accepted by online workers and lead to substantial performance boosts.

Categories and Subject Descriptors

K.4.4 [**Computers and Society**]: Electronic Commerce-Payment schemes

General Terms

Algorithms, Experimentation, Human Factors

Keywords

crowdsourcing, reward schemes, competitions, lotteries

1. INTRODUCTION

Gathering and exploiting collective knowledge online has become increasingly popular over recent years. Projects such as Wikipedia or Open Street Maps have demonstrated that a large number of non-experts can, under certain conditions, be as effective and precise as a small group of experts [24, 22]. Information providers make implicit use of collaborative knowledge for improving their services: Search engines leverage query logs for determining the popularity of web pages or for suggesting queries and advertisements to their customers. Online shops like Amazon and auctions like eBay correlate information from buyers to recommend new items. Games with a Purpose have been employed for gathering user input in large quantities: In the ESP game [26, 25], for instance, online players compete in image annotation tasks; in this way, the images indexed by Google in 2004 could be annotated in just 31 days. On the other hand, platforms like Amazon Mechanical Turk and CrowdFlower successfully make use of online workers and monetary incentives for accomplishing explicit tasks such as the annotation of multimedia content, translation of texts, or generation of training sets for machine learning.

Crowdsourcing platforms such as Amazon Mechanical Turk and CrowdFlower are based on a reward scheme where the payment of online workers is proportional to the number of accomplished tasks ("Pay-per-HIT"). Such platforms are used successfully in various contexts [29, 27], but can we get more value for money? In the "real world" gaming based motivation has been shown to be very successful in the past. A prominent example are lotteries where the prize money is rather small in comparison to the revenue generated by millions of participants. In the former Eastern Germany construction workers received special lottery tickets for a certain amount of hours spent building houses. Another example is the Speed Camera Lottery where a portion of fines levied against speeders would be pooled in a jackpot, with a random winner periodically drawn from the group of speed-limit adherents.[1] Apart from games of luck, also competition among workers can be a strong motivation. Inducement prizes awarded for instance by companies like Microsoft, Google, or Yahoo for ideas or code attract a huge numbers of participants and create an enormous overall value. The Netflix Prize[2] is an example for such an open competition, where 20,000 research teams were competing against each other, trying to implement the most effective collaborative filtering algorithm for a single prize of 1,000,000 USD.

To sum up, in the real world people are often attracted by scenarios where they can (a) compete, and/or, (b) can receive a relatively high reward with low probability. In this paper, we aim to carry over competitive and randomized reward mechanisms to (online) crowdsourcing based on monetary incentives. Reward schemes studied in this work range from "Winner-Takes-It-All" competitions, where only the top performer will earn a fixed prize money, over expo-

CIKM'14, November 3–7, 2014, Shanghai, China.

http://dx.doi.org/10.1145/2661829.2661946.

[1] http://www.thefuntheory.com/speed-camera-lottery-0
[2] http://en.wikipedia.org/wiki/Netflix_Prize

nentially decreasing rewards distributed among an extended set of top performers, to lottery-inspired mechanisms that introduce an additional random element. Furthermore, we study the influence of different information policies that vary the amount of knowledge and feedback given to a worker about the performance of their competitors. We use the classical "Pay-per-HIT" mechanism (as employed e.g. in Mechanical Turk) as a baseline and show that it can be clearly outperformed using elements borrowed from competitive games and games of luck. Although theoretical gamification scenarios for crowdsourcing were considered in recent literature (see Section 2 for details), we are the first to perform a large-scale empirical evaluation in this context.

In this work we shed light on how reward mechanisms can be employed in a cost effective manner for solving online tasks. We are aware that conventional games of luck are associated with increased life stress for players and can negatively affect their social environment [14, 18]. To this end, exploiting individuals depending on receiving living wages for the tasks they perform on crowdsourcing platforms would be inappropriate. As with almost every technology, gamification in crowdsourcing scenarios requires sensible handling and constructive usage to make it of great benefit for research communities operating under budget constraints on problems that are of interest for the public.

Outline. The remainder of this paper is organized as follows: In Section 2 we discuss related work on crowdsourcing and games with a purpose. In Section 3 we describe our reward strategies and information policies in competitive as well as lottery games. The evaluation of our strategies is presented in Section 4 where we first describe the experimental setup along with the core results, and then delve deeper into details about worker behavior and performance. Finally, in Section 5 we conclude and describe directions of our future work.

2. RELATED WORK

Literature in economics and computer science typically refers to crowdsourcing as a "principal-agent" problem where the agents are workers with various skill levels which solve tasks for another person - the principal. Crowdsourcing has a wide range of applications, including the annotation of data sets, conduction of user surveys, and collaborative gathering of data collections. Online platforms such as Amazon Mechanical Turk[3] and CrowdFlower[4] provide principals and agents with a framework for publishing and selecting tasks and transferring payments. Since 2011, TREC offers a crowdsourcing trek [2] addressing various issues related to gathering document-relevance labels for information retrieval systems [4, 3]. This includes HIT design, user filtering, and the fusion of worker judgments. In this paper, we embed crowdsourcing techniques within a competitive reward framework combined with random mechanisms in order to increase the effectiveness of annotations.

In their seminal work [15] Kazai et al. study how payment, worker qualification, and required effort influence the output of tasks in Amazon Mechanical Turk. In the context of relevance labeling, the authors show that increasing the rewards, reducing the required effort, and filtering workers based on qualification requirements can increase the accuracy of the output. In [16] the same authors show the correlation between behavior and personality of workers and the accuracy of their work. In [19] the influence of the amount of micro-payments on quantity and quality of annotations is studied. In contrast, our work focuses on how competitive reward mechanisms influence annotation behavior and the cost efficiency of crowdsourcing.

A recent workshop [1] has focused on principal aspects of game mechanics embedded in standard IR tasks including information seeking, crowdsourcing, and user engagement. He et al. [12] introduce a user interface for studying search behavior within a gamified setting where users receive points for finding relevant documents, and where scores are announced in leader boards. Dinesh et al. [20] discuss additional incentive structures to encourage user activity within an online platform. A number of earlier works tackle the same problems. Eickhoff et al. [11] employ a game based approach for crowdsourcing of query result relevance labels and image cluster labels. The authors show that entertainment can be a powerful incentive in crowdsourcing and can partially replace financial rewards. In contrast, in our work we study monetary reward mechanisms combined with various information policies in order to increase the effectiveness of crowdsourcing. In [9, 21] the authors propose using lottery tickets for engaging crowd workers. In [9] a small scale survey was conducted and the results show that about one third of the users prefer a payment strategy involving lotteries over the standard payment mechanism. This confirms the claims in [23], where the author mentions that a non-negligible amount of workers engage themselves into crowdsourcing tasks for both money and fun. Our work provides a large-scale experimental evaluation on boosting the cost efficiency of crowdsourcing through payment strategies (lottery as well as non-lottery) in combination with different information policies.

There is a body of work on crowdsourcing theory in the area of business, economics, and e-commerce. For instance, [5] provides an analysis of crowdsourcing contests within the software development portal TopCoder[5]. In [10] the authors study the influence of the reward amount; they find that participation rates increase as a function of the offered reward. Other work shows that, while workers are generally attracted by high rewards, they also tend to choose tasks with low rewards that better suit their abilities and to maximize their outcome by balancing reward and workload [28]. In [13] this issue is further addressed by splitting the crowd into groups of workers with different abilities in a scenario where workers compete with each other in the context of bug detection. Although related to our research purposes, these works target crowd based software development contexts where just a *single* solution is selected at the end. In contrast, our work focuses on crowdsourcing of annotations in the context of information retrieval and data mining, where the workload is divided across multiple workers.

Finally, a number of works have proposed theoretical stochastic models in the context of crowdsourcing. In [8], the authors introduce models for the effectiveness of winner-take-all scenarios, and consider aspects such as the optimal choice of the prize money. In [7] the same authors

[3] https://www.mturk.com

[4] http://www.crowdflower.com/

[5] http://www.topcoder.com

concentrate on scenarios where every non-zero effort of workers is rewarded and, similar to the scenario studied in this paper, the final output consists of the cumulative effort of all workers. Archak et al. [6] present a model for designing crowdsourcing contests with optimized reward distributions to improve the quality of the best submission. However, in contrast to our paper, none of these works provide an experimental evaluation of their concepts.

To the best of our knowledge, we are the first to suggest different information policies for crowdsourcing competitions, and to conduct systematic real-world studies of the effects of policies and reward distributions on both quantity and quality of annotations.

3. CROWDSOURCING COMPETITION DESIGNS

In this section, we formalize our crowdsourcing scenario and describe different strategies for distributing rewards among users both in a competitive and random fashion. Furthermore, we describe the different information policies (relating to information revealed about fellow workers) we employed in our framework.

3.1 Problem Setting

We consider a scenario with a crowd consisting of n workers $W = \{w_1, \ldots, w_n\}$, and with a fixed (monetary) budget M for paying workers. This budget is distributed among the workers depending on the values $v(w_i)$ produced by them. These values $v(w_i)$ can, for instance, correspond to the number of correctly solved crowdsourcing tasks. A worker w_i receives a reward $r(w_i)$ with $\sum_{i=1}^{n} r(w_i) = M$. Our goal is to maximize the overall value $V = \sum_{i=1}^{n} v(w_i)$ produced by the workers in W.

3.2 Strategies for Reward Distribution

In terms of strategies we distinguish between the baseline approach of paying per task, competitive approaches where payment is received depending on the performance based on the position of workers in a leaderboard, and random approaches where a set of winners is determined in a worker lottery.

The Baseline: Linear Reward Assignment. The commonly used strategy in crowdsourcing is to distribute rewards proportional to the individual values produced by workers, i.e. each worker w_i receives a reward $r(w_i) = (v(w_i)/V) \cdot M$. In practice this is typically implemented by fixing a reward rate c (e.g. money per task solved) and an overall value $V = M/c$ to be produced (e.g. number of tasks to be solved), resulting in a reward $r(w_i) = v(w_i) \cdot c$ for worker w_i. This strategy corresponds to the usual payment scheme as, for instance, employed for Amazon Mechanical Turk HITs ("Pay-per-HIT").

Competitive Strategies. In this paper we explore competitive strategies where workers are ranked according to their produced values, and the reward of a worker will depend on his rank. Formally, let $rank(w_i) \in \{1, \ldots, n\}$ be the rank of worker w_i, with a rank of j corresponding to the jth highest value produced across all workers. A simple special case consists of paying the whole budget to the top-ranked worker ("Winner-Takes-It-All"), i.e. $r(w_i) = M$ if $rank(w_i) = 1$ and 0 otherwise. More generally, we compute the reward $r(w_i)$ as a monotonically decreasing function $\Gamma(rank(w_i))$ of the worker's rank. Similar to "real world" competitions we discount lower ranks, i.e. top performers receive more money per solved tasks, and Γ is a convex function. Examples for Γ could be (appropriately normalized) negative exponential or negative polynomial functions of the rank. In practice, instead of precise mathematical functions, one would rather provide workers with an easy to understand payment scheme containing "round" numbers for rewards and number of winners; in addition, one has to account for the fact that very small rewards can become meaningless. In our concrete experiment with an "exponential reward"-like strategy, for instance, we chose to pay out 25, 10, 5, 5, 1, 1, 1, 1, 0.5, and 0.5 USD to the top-10 users (overall budget: $M = 50$ USD).

Randomized Reward Strategies. We also study randomized strategies which are inspired by scenarios such as gambling and lotteries where these types of rewards are successfully used. Carrying this over to crowdsourcing, we explore reward schemes where workers can earn "lottery" tickets, with each ticket corresponding to a produced value σ (e.g. a certain number of correctly solved tasks). In this way, worker w_i producing a value $v(w_i)$ will obtain $\lfloor v(w_i)/\sigma \rfloor$ lottery tickets. After the competition a number of lottery tickets are randomly drawn and winners obtain monetary awards $r(w_i)$ depending on the outcome. Similar to competitive strategies a simple special case consists of drawing a single lottery ticket and assigning the whole budget M to the winner, i.e. $r(w_i) = M$ if worker w_i is an owner of the (unique) winning ticket and 0 otherwise. For this "Winner-Takes-It-All" situation the probability p_i of worker w_i winning is proportional to the number of tickets earned by him: $p_i = \frac{\lfloor v(w_i)/\sigma \rfloor}{\sum_{j=1}^{n} \lfloor v(w_j)/\sigma \rfloor}$. This can be generalized by randomly drawing k tickets and paying out rewards for each of these tickets. In our concrete experiment, for instance, we randomly selected $k = 10$ of the earned tickets and randomly distributed the prize money of 50 USD as previously described for the "exponential reward"-like strategy. Note that, in contrast to "traditional" lotteries (with a randomly drawn set of numbers), in our payment scheme the whole budget M is finally paid out to workers because random selections are conducted over earned tickets only.

3.3 Information Policies

In contrast to the classical linear "Pay-per-HIT" scheme, for both competitive and randomized strategies the payment of workers largely depends on the performance of their fellow workers: In the competitive case rewards are determined by the rank relative to other workers; in the randomized case the probabilities of winning rewards depend on the overall number of tickets earned by users. How much information about the performance of his fellow workers should we provide to a worker during a competition? Our experiments show that such information can both motivate and demoralize workers. In this section we describe the different information policies we explored.

Information Policies for Competitive Strategies. For competitive strategies we distinguish between open, restricted, and medium policies. In the *open* information policy we provide the worker, in principle, with all of the relevant information about his fellow workers during the compe-

tition. This includes constantly showing the updated top-10 leaderboard of workers along with their positions and scores (values produced). Knowing the position of other workers can trigger competitive behavior and motivate workers to "catch up" or to "defend" their status. On the other hand, workers that are too far "behind" can be demotivated. In the *restricted* information policy no information about the fellow workers is provided. Workers can only see their own score and do not learn anything about scores of other workers and their relative position during the competition. While this can help avoiding the demoralization of players with low scores, it can also prevent desired competitive behavior. Situated in between the open and restricted policy is the *medium* policy where only part of the information about other workers is revealed. To this end, we tested scenarios where workers were shown their position along with a snapshot consisting of their k neighbors above and below them in the leaderboard. The rationale was to trigger competitive behavior within part of the leaderboard while avoiding too much frustration for workers with lower scores.

Information Policies for Randomized Strategies. For the randomized strategies described above the chances of winning do only depend on the number of tickets earned by the worker and the overall number of earned tickets. To this end, we tested an *open* policy where workers get continuous updates about their own tickets and the overall number of tickets in the system, and a *restricted* policy where a worker can just see his own tickets during the competition. In contrast to the competitive strategies and due to the simpler information related structure of the lottery competitions, we did not see the potential for some medium policy. One interesting point to note is that, although the chances of winning do not depend on the specific distribution of tickets across workers, we received many comments requesting ticket based leaderboards. This is an example for irrational elements and artifacts in crowdsourcing that we plan to further explore in future work.

4. EXPERIMENTS

In this section we evaluate the reward distribution strategies and information policies described in Section 3 using two crowdsourcing scenarios: captcha translation and face recognition. The objective of our evaluation was to study the cost efficiency of strategies, competitive behavior of workers, and acceptance of the different reward schemes.

4.1 Setup

Crowdsourcing Tasks. For the *captcha translation* scenario the tasks consisted of translating 10 captchas shown on a Web page, which were drawn from a pool of 100,000 captchas generated using the *Cage*[6] library. Our application checked the correctness of labeled captchas and provided immediate feedback to the workers. Workers obtained one point per correctly translated captcha, and the application provided continuous feedback on their scores. With the captcha task we deliberately chose a rather dull scenario because we wanted to focus on monetary incentives in this work rather than gamification and intrinsic motivation. In our setup we do not have to consider annotation quality for our captcha translation task, and measure only the correctly solved instances.

However, in order to additionally study more quality related aspects in our context we launched a *face recognition* task where workers were asked to identify a person on a given reference photo among a set of 10 test photos. The images were retrieved from the PubFig[7] database which was created for face verification [17]. Out of originally 58,797 images with faces of 200 celebrities, 37,004 images where available on the Web. We reviewed the dataset manually and removed 637 images showing placeholders as well as 133 images we deemed unsuitable because the correct person was not shown on the image. As a quality check mechanism we randomly introduced a "honeypot" task within each batch of 100 tasks that was manually selected beforehand. After workers finished a batch they were shown the honeypot and their own input for it. If workers correctly solved the honeypot, 100 points were added to their score, otherwise 20 points were subtracted (with a cut-off threshold of 0 for the score, i.e. we did not introduce negative scores).

Implementation and Settings. We announced the crowdsourcing tasks on the Amazon Mechanical Turk platform and on a mailing list consisting of participants from previous competitions about one day before the competition started. The workers were choosing the tasks autonomously, as common for crowdsourcing platforms such as Mechanical Turk. As most of our tested reward strategies are not supported by Mechanical Turk or CrowdFlower, we ran the actual competition using an external application on servers at our institute. Each worker was assigned a user code which he could enter in Mechanical Turk in order to claim his prize money after a competition was finished. The experiments were performed sequentially with a duration of four days per experiment (and one run per configuration) and a break of at least one day in between two competitions in order to avoid extensive user fatigue.

Tested Strategies. We tested the two competitive reward strategies ("Winner-Takes-It-All" (**wta**) and exponential reward strategy (**exp**)) and the two random strategies ("Winner-Takes-It-All" lottery (**wtaLott**) and exponential reward lottery (**expLott**)) described in Section 3.2 with an overall prize money of $M = 50$ USD, and combined these strategies with the information policies described in Section 3.3 (open (**open**), restricted (**res**), and medium (**med**) policy for competitive strategies; open and restricted policy of random strategies). A competition is thus defined by its reward strategy and information policy; in the following, we abbreviate, for instance, a "Winner-takes-It-All" strategy conducted under an open information policy as "*wta-open*". This results in a total of 10 (reward strategy - information policy) combinations. In addition, we conducted baseline experiments using a linear reward assignment ("Pay-Per-HIT") with a fixed amount of money paid per solved task (**baseline**). In order to study the influence of the absolute value of monetary rewards, we also conducted experiments with increasing prize money amounts of 10, 25, 50, and 100 USD per competition.

[6] http://akiraly.github.io/cage/

[7] http://www.cs.columbia.edu/CAVE/databases/pubfig/

4.2 Results for Captcha Task

In our captcha experiments overall 988,054 captchas (amounting to 2,165 hours of work) were translated by 441 participants from 17 different countries. The majority of the participants where from India (51.25%) and the US (44.44%).

Aggregate Results. Table 1 shows the number of translated captchas for each of the strategies along with the average amount of money spent per captcha and per hour of work. The main observations are the following:

- The exponential reward strategy with medium information policy (*exp-med*) clearly outperforms the other strategies, with almost three times as many captchas translated as for the *baseline* within the same time frame and with the price per captcha reduced by a factor of almost 2.5.[8]
- In general, the medium policy (*med*) shows clear advantages over the other information policies for all reward strategies. This confirms our intuition that the open policy (*open*) might demoralize lower-ranked workers, while the restrictive policy (*res*) does not provide sufficient (competition related) feedback to the workers.
- Exponential reward strategies (*exp*) in combination with open and medium policies outperform the winner-takes-it-all (*wta*) approaches. Further analysis of individual worker performances, described later in this section, shows that spreading the rewards over a wider range of ranks (instead of allowing just one single winner as in *wta*) seems to stimulate competitive behavior between lower ranked workers.

The lottery based strategies are outperformed by both the competitive strategies and the baseline. This might be explained by the relatively low rewards (with relative high probability of winning) in comparison to "real world" lotteries where typically risk affine persons are attracted by very high rewards paid out with low probability. Another explanation might be that the competitive character of our lotteries is just *implicitly* reflected by the odds of winning given the overall number of tickets in our system.

Worker Contributions. Figure 1 shows the fraction of captchas solved by the top-10 ($1 \leq R \leq 10$) and the remaining workers ($R > 10$) for each run. For most of the runs (including the baseline) the top-10 workers translated over 70 percent of the captchas. Another interesting observation is the performance of the top-1 users, who translated between 10 and more than 40 percent of the captchas. The contribution of the top-1 workers is especially high for the winner-takes-it-all scenarios; a more detailed temporal analysis of worker contributions described below reveals that these scenarios become quickly less attractive for

[8]For the baseline we chose a reward of 1 cent for 13 captchas; in contrast, for *exp-med* workers would have to annotate about 31 captchas for the same reward. Also note that within the given 4-day time frame workers where just willing to translate captchas for an overall amount of about 45 USD (i.e. the whole budget of 50 USD could not be utilized).

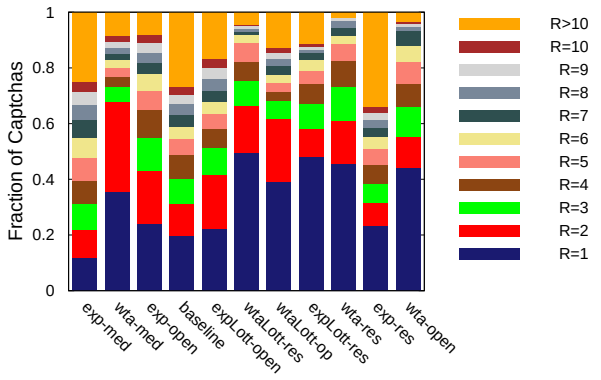


Figure 1: Contributions of individual workers in different rounds.

lower ranked workers. In contrast, the exponential reward strategies *exp-res* and our best performing strategy *exp-med*, where lower ranked workers also have a chance of earning rewards, result in a more balanced distribution of the workload across the top-10 workers. These strategies, along with the baseline, exhibit also the "fattest" tail for the distribution of the number of translated captchas contributed by workers not appearing in the top-10.

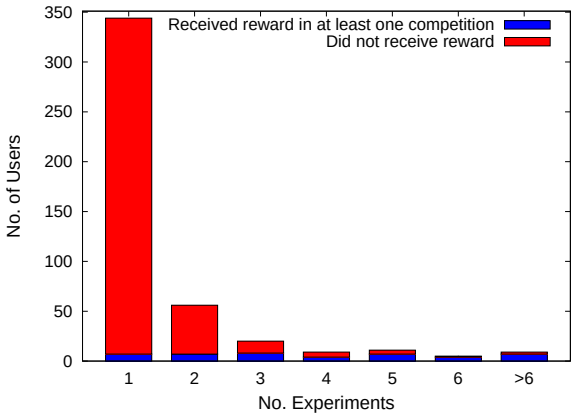


Figure 2: Participation statistics.

Figure 2 shows the number of workers with respect to the number of rounds they participated in; workers are further divided into the ones that won a monetary prize in at least one of the rounds and the ones who didn't. The limited number of prizes paid out in our lottery and competitive scenarios naturally resulted in a large number of workers that did not receive any rewards. As expected, for workers participating in multiple rounds the fraction that won in one of the rounds is higher, consistent with our observation that winning workers were more motivated to participate in further rounds.

Temporal Dynamics. Figure 3 shows the temporal evolution of the cumulative counts for solved captchas aggregated over the whole set of workers. The numbers grow mostly linearly over time. However, several of the strategies such as

Exponential Reward				Winner-takes-it-all			
Experiment	No. Captchas	USD/Hour	Cent/Captcha	Experiment	No. Captchas	USD/Hour	Cent/Captcha
Performance-based payment							
exp-open	67,951	0.300	0.074	wta-open	25,424	0.880	0.197
exp-med	154,188	0.138	0.032	wta-med	81,742	0.322	0.061
exp-res	25,853	0.605	0.193	wta-res	33,085	0.6688	0.151
Lottery-based payment							
expLott-open	48,241	0.428	0.104	wtaLott-open	39,249	0.608	0.127
expLott-res	38,194	0.621	0.131	wtaLott-res	39,490	0.541	0.127

Baseline		
No. Captchas	USD/Hour	Cent/Captcha
58,635	0.352	0.077

Table 1: Aggregate outcomes of the main captcha translation rounds.

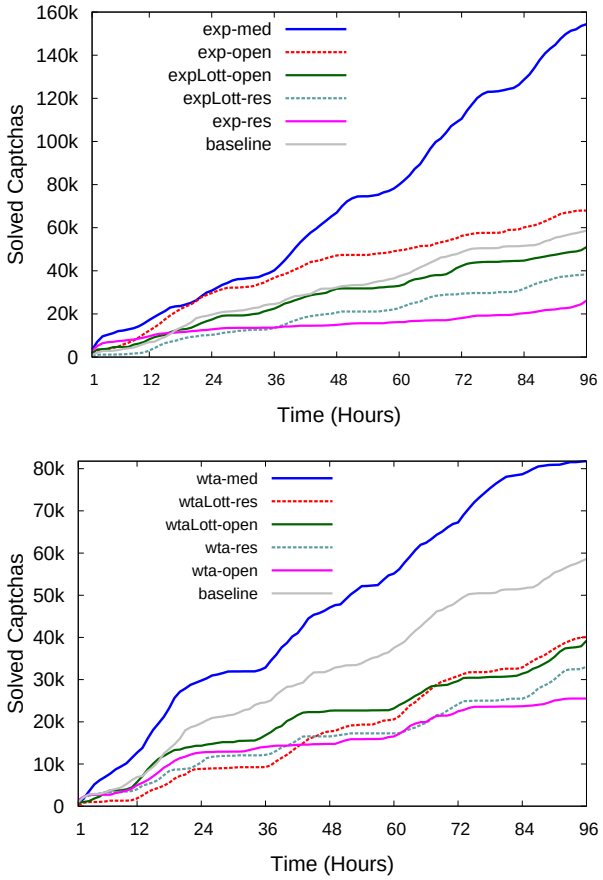
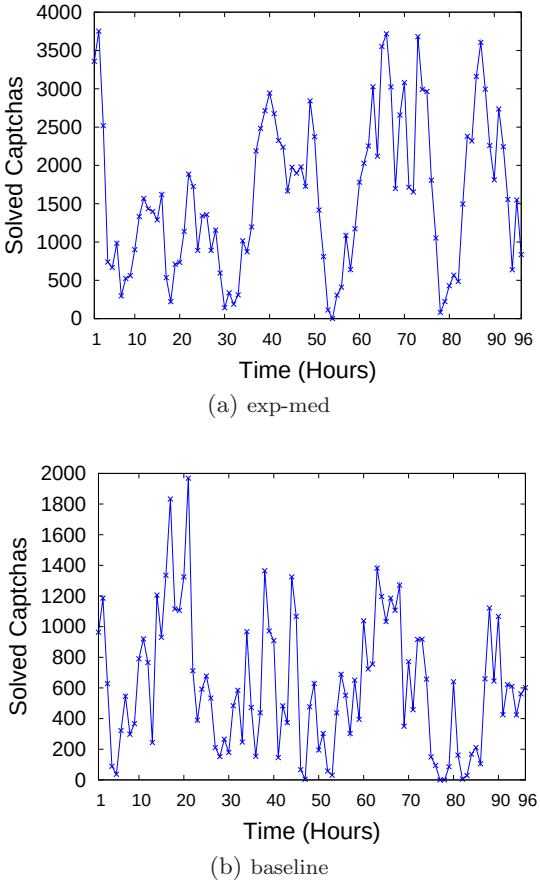


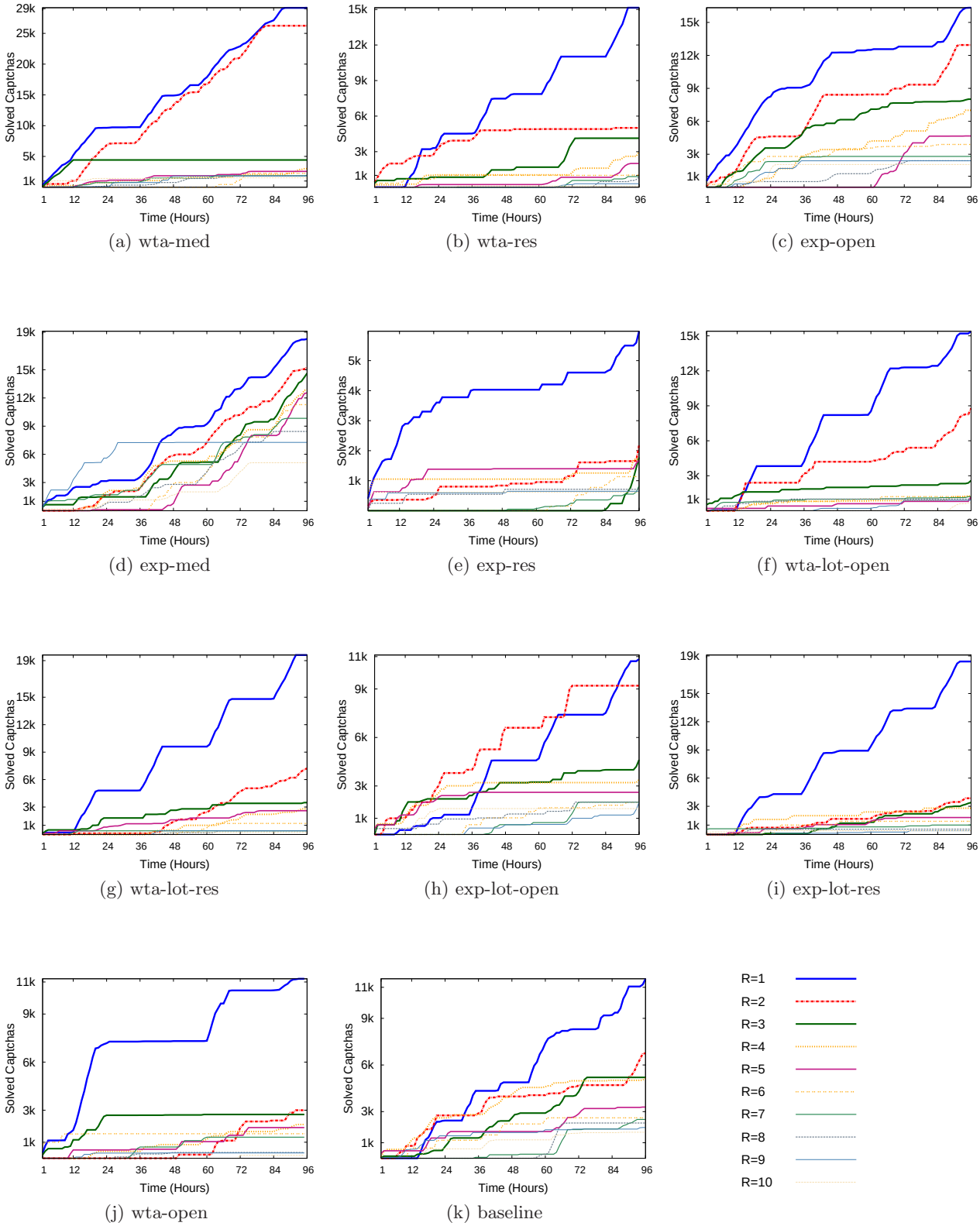
Figure 3: Cumulative number of annotated captchas per experiment over time for exponential reward based (top) and winner-takes-it-all based competitions (bottom).



(a) exp-med

(b) baseline

Figure 4: Temporal characteristics of annotations for the exponential reward strategy with medium information policy (*exp-med*) and the *baseline*.



(a) wta-med (b) wta-res (c) exp-open

(d) exp-med (e) exp-res (f) wta-lot-open

(g) wta-lot-res (h) exp-lot-open (i) exp-lot-res

(j) wta-open (k) baseline

Figure 5: Number of annotated captchas for top-10 workers over time.

our winning method exhibit a "quick start" phase in the first hours, showing that people are highly motivated at the beginning if a fixed starting point is announced beforehand. This can be seen even more clearly in the non-cumulative plots for the annotated captcha counts per hour in Figure 4 (provided for the examples of our exponential reward strategy *exp-med* and the *baseline*). In addition, we observe periodic phases where curves become almost "flat", i.e. less work is done in these phases (see also Figure 4 that makes that periodicity more explicit). This can be explained by the day cycle of the workers as there is a strong bias towards participants from India and the US.

We also studied the contribution of individual workers over time. Figure 5 breaks the cumulative captcha counts down for the top-10 workers. For our winning exponential reward strategy *exp-med* (Figure 5d) we observe a large number of "crossing curves" during all phases of the experiment, indicating strong competitive behavior compared to other strategies. This holds both for the top-2 and the remaining workers in the top-10. In contrast, activity is much lower towards the end for the lottery and winner-takes-it-all experiments. For *wta-med* (Figure 5a) we can observe a fierce competition between the top-2 workers; the remaining workers seem less motivated - presumably because they are already too far "behind" and there is just a single reward issued. It is also interesting to note that some workers seem to exhibit irrational behavior: In the *wta-open* experiment (Figure 5j), for instance, the top worker keeps annotating until the end, although he is already far "ahead" and can monitor the activity level of his competitors due to the open information policy.

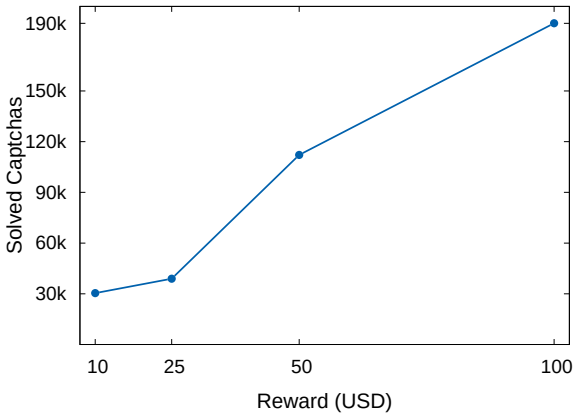


Figure 6: Work output with respect to budget.

Influence of the Overall Budget. For our best exponential reward strategy, *exp-med*, we also studied the influence of the total prize money by varying the budget M. Figure 6 indicates a roughly linear increase of overall amount of work conducted with respect to prize money spent. However studying the influence of the overall reward more in-depth requires additional experiments which we have to leave as future work, including both multiple repetitions of the conducted experiments and the introduction of an extended range of budgets.

	Correct Honeypots		Incorrect Honeypots	
	No. of batches	**Correctly matched faces**	**No. of batches**	**Correctly matched faces**
exp-med	474	92.1%	19	69.9%
baseline	149	93.5%	7	56.6%

Table 2: Quality results for face recognition.

4.3 Results for Face Recognition Task

In order to examine a more realistic scenario where quality-related issues have to be taken into account and where no immediate feedback about the correctness of annotations can be provided to workers, we evaluated our approach for the face recognition task described in Section 4.1. To this end, we compared the best performing configuration from the previous subsection (exponential reward strategy with medium information policy - *exp-med*) with the "pay-per-task" *baseline*. As before, for *exp-med* we distributed a prize money of 50 USD among the top-10 workers. For the baseline we paid 15 cents per batch of 100 face annotations with correctly solved honeypot, and a penalty of 3 cents for each incorrect honeypot (corresponding to the 20% of the points subtracted in the competitive *exp-med* scenario). In total 72.29 USD were spent and 61,310 images were annotated correctly (amounting to 289.6 hours of work) by 155 participants from 9 countries (48.4% of them coming from the US and 47.1% from India). We considered only results from participants that completed at least one batch.

The main results are the following:

- For the baseline we obtained 14,724 correctly recognized faces for a price of 22.29 USD. The baseline was clearly outperformed by the *exp-med* strategy with 46,586 correctly recognized faces for a price of 50 USD. This corresponds to correct annotation per USD rates of about 660 for the baseline and 930 for *exp-med*, and, in addition, shows that in the baseline case users were just willing to annotate about one third of the captchas as for *exp-med* in the 4 day time frame - even for a clearly higher reward per captcha.
- In terms of correct annotations there is no noticeable difference between the baseline and *exp-med* (see Table 2): The large majority of batches for those scenarios had correctly solved honeypots and the percentage of correctly recognized faces in these batches is 93.2% and 92.1% for the baseline and *exp-med*, respectively. On the other hand *exp-med* has a slightly lower error rate for batches where the honeypot was incorrectly solved. Thus, the quality did not decline in our competitive scenario.

The *exp-med* setup also attracted more participants: 61 workers completed at least one batch, in comparison to 45 for the baseline. The mentioned differences between the methods are further illustrated in Figure 8 showing the cumulative number of correctly matched faces for both experiments. Figure 7 provides additional insights on the temporal dynamics of the top-10 performers. Similar to the captcha scenario we observe a large number of "crossings" as well as simultaneous bursts of activities in the first 36 hours of the *exp-med* experiment, indicating a vivid competition in contrast to the baseline scenario.

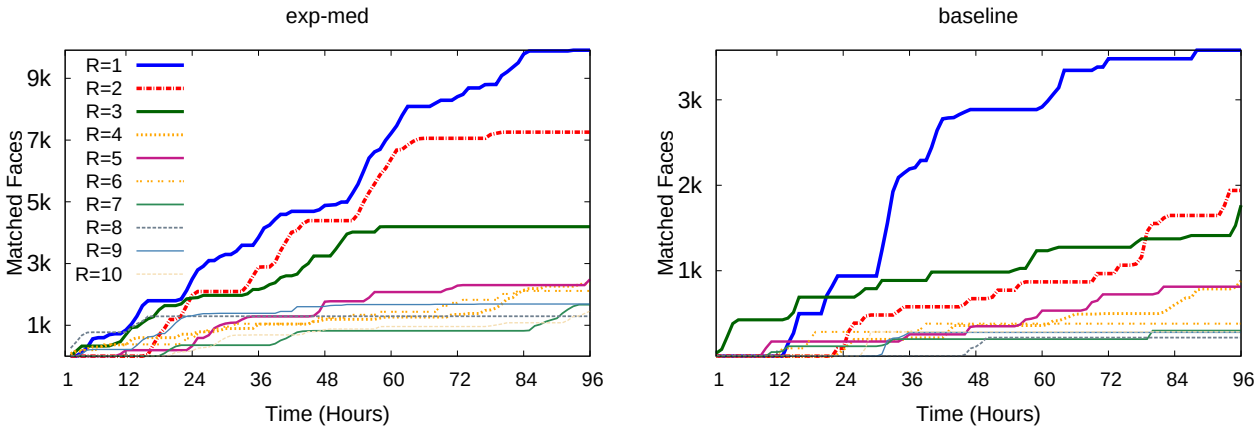


Figure 7: Cumulative number of recognized faces for the top-10 workers (*exp-med* strategy vs. baseline).

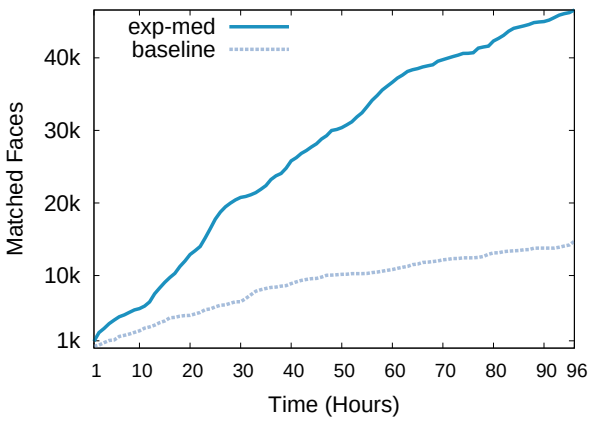


Figure 8: Cumulative number of recognized faces over time.

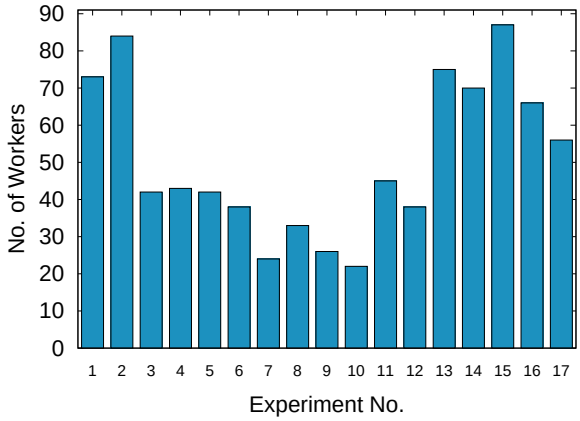


Figure 9: Participation over the course of the experiments.

4.4 Worker Activity and Direct User Feedback

Worker Activity. Over the course of the experiments we observed variations in the number of participants per experiment. These numbers are shown in Figure 9 in chronological order. At the very beginning we observe a high amount of participants, which might be explained by curiosity of workers regarding the new reward models which are not common in Mechanical Turk. Most of the experiments (3-12) show rather moderate variation in the number of participants. We also observe that some of the experiments were exceptionally attractive for users (2 and 13 are both *exp-med* for captchas with a 50 USD budget, 14 is *exp-med* for captchas with a 100 USD budget, and 15 is *exp-med* for face recognition with a 50 USD budget). Finally No. 16 and 17 were the baseline experiments (for captchas and face recognition, respectively) which attracted a comparable number of workers.

Direct User Feedback. In general, the user feedback was quite positive and the communication was very friendly. The tasks were described with terms like "interesting" or "mind exercise" but also as challenging in the case of the captcha task. Multiple users commented favorably on competitions (`"i love this game thank you very much [...]"`, `"it was good experience"`).

There were also critical responses, with some users having concerns about the reward mechanisms, reward amount, or duration of the competitions (`"is this trick for gettin free of cost captcha solvers [...]"`, `"you are giving very little amount"`, `"i want duration 1 or 2 hour only. one week is not possible"`). In such cases the users did not participate further in the competitions. There were also a number of technical suggestions, ranging from requests for a mobile application to more subtle changes in the reward distributions.

In our experiments we noticed some workers dominating the competitions for multiple consecutive rounds (comment from another worker: `"So...you are saying that the same person won again??"`). In the preliminary rounds, concerns were raised regarding workers sharing the same account, leading to an unfair advantage. We reacted to this by prohibiting simultaneous logins altogether, although it was still possible to take *turns* annotating using the same account. However, collaborations between workers are not necessarily negative and explicitly leveraging them opens promising avenues for future research.

5. CONCLUSIONS AND FUTURE WORK

We have studied various alternative reward distribution strategies for crowdsourcing compared to the commonly used "pay-per-task" approach. These strategies are based on competitive as well as randomized aspects, and build on ideas borrowed from inducement prizes, lotteries, and games of luck. Our evaluation shows substantial performance boosts in both examined crowdsourcing scenarios compared to the baseline: Our best approach results in three times as many annotations than for the *baseline* within the same time frame and with a price per captcha reduced by a factor of almost 2.5 for the captcha translation task, and about 40% more annotations per dollar for the face recognition task. We also found that the amount of information revealed about other workers during the competitions plays a crucial role: Our results indicate that the sweet spot (a "medium" information policy) lies between a restrictive policy (no information about other workers revealed) and an open policy (all of the relevant information revealed). An in-depth analysis of the characteristics of individual workers sheds additional light on motivational and competitive aspects.

In our future work we aim to explore the influence of parameters such as exact amount of prizes, duration of competitions, and amount information about fellow workers revealed. Furthermore, we plan to study quality insurance mechanisms in the context of competitive crowdsourcing. To this end, we plan to explore and compare different strategies based on honeypots, redundant inputs from different workers, and trust models for workers. Finally, we aim to study strategies for coping with worker fatigue - a phenomenon that might be further amplified in scenarios where competition among workers creates additional pressure and stress. In order to tackle these problems we want to explore different mechanisms based on the introduction of strategic breaks, small guaranteed rewards, and task diversification.

6. ACKNOWLEDGMENTS

This work is partly funded by the European Research Council under ALEXANDRIA (ERC 339233) and by the European Commission FP7 under CUBRIK (grant agreement No. 287704).

7. REFERENCES

[1] *GamifIR '14: 1st International Workshop on Gamification for Information Retrieval*, NY, USA.

[2] TREC Crowdsourcing Task 2013. `https://sites.google.com/site/treccrowd/home`.

[3] O. Alonso and R. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In ECIR'11.

[4] O. Alonso and S. Mizzaro. Using crowdsourcing for trec relevance assessment. *Information Processing and Management, 2012*.

[5] N. Archak. Money, glory and cheap talk: Analyzing strategic behavior of contestants in simultaneous crowdsourcing contests on topcoder.com. In *WWW'10*.

[6] N. Archak and A. Sundararajan. Optimal design of crowdsourcing contests. In *ICIS'09*.

[7] R. Cavallo and S. Jain. Efficient crowdsourcing contests. In *AAMAS'12*.

[8] R. Cavallo and S. Jain. Winner-take-all crowdsourcing contests with stochastic production. In *HCOMP'13*.

[9] L. E. Celis, S. Roy, and V. Mishra. Lottery-based payment mechanism for microtasks. In *HCOMP'13*.

[10] D. DiPalantino and M. Vojnovic. Crowdsourcing and all-pay auctions. In *EC'09*.

[11] C. Eickhoff, C. G. Harris, A. P. de Vries, and P. Srinivasan. Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In *SIGIR'12*.

[12] J. He, M. Bron, L. Azzopardi, and A. de Vries. Studying user browsing behavior through gamified search tasks. In *GamifIR '14*.

[13] H. Jiang and S. Matsubara. Improving crowdsourcing efficiency based on division strategy. *WI'12*.

[14] J. Joukhador, A. Blaszczynski, and F. Maccallum. Superstitious beliefs in gambling among problem and non-problem gamblers: Preliminary data. *Journal of Gambling Studies*, 2004.

[15] G. Kazai. In search of quality in crowdsourcing for search engine evaluation. In *ECIR'11*.

[16] G. Kazai, J. Kamps, and N. Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *CIKM'11*.

[17] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *ICCV'09*.

[18] D. Lam. An exploratory study of gambling motivations and their impact on the purchase frequencies of various gambling products. *Psychology and Marketing*, 2007.

[19] W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". *SIGKDD'09*.

[20] D. Pothineni, P. Mishra, A. Rasheed, and D. Sundararajan. Incentive design to mould online behavior: A game mechanics perspective. In *GamifIR'14*.

[21] J. P. Rula, V. Navda, F. E. Bustamante, R. Bhagwan, and S. Guha. No "one-size fits all": Towards a principled approach for incentives in mobile crowdsourcing. In *HotMobile '14*.

[22] N. Savage. Gaining wisdom from crowds. *Magazine Communications of the ACM, 2012*.

[23] K. Stoddart. Behind the scenes of crowdsourcing: Who are crowd workers?, 2012. Available at `http://www.crowdsource.com/blog/2012/11/behind-the-scenes-of-crowdsourcing-who-are-crowd-workers/`.

[24] J. Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[25] L. von Ahn and L. Dabbish. Designing games with a purpose. *Magazine Communications of the ACM, 2008*.

[26] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI'04*.

[27] P. Welinder and P. Perona. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *CVPRW'10*.

[28] J. Yang, L. A. Adamic, and M. S. Ackerman. Crowdsourcing and knowledge sharing: Strategic user behavior on taskcn. In *EC'08*.

[29] M.-C. Yuen, I. King, and K.-S. Leung. A survey of crowdsourcing systems. In *Passat Workshop at socialcom'11*.