# Groupsourcing: Team Competition Designs for Crowdsourcing

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

## ABSTRACT

Many data processing tasks such as semantic annotation of images, translation of texts in foreign languages, and labeling of training data for machine learning models require human input, and, on a large scale, can only be accurately solved using crowd based online work. Recent work shows that frameworks where crowd workers compete against each other can drastically reduce crowdsourcing costs, and outperform conventional reward schemes where the payment of online workers is proportional to the number of accomplished tasks ("pay-per-task"). In this paper, we investigate how team mechanisms can be leveraged to further improve the cost efficiency of crowdsourcing competitions. To this end, we introduce strategies for team based crowdsourcing, ranging from team formation processes where workers are randomly assigned to competing teams, over strategies involving self-organization where workers actively participate in team building, to combinations of team and individual competitions. Our large-scale experimental evaluation with more than 1,100 participants and overall 5,400 hours of work spent by crowd workers demonstrates that our team based crowdsourcing mechanisms are well accepted by online workers and lead to substantial performance boosts.

## Categories and Subject Descriptors

H.5.m [**Information Interfaces & Presentation (e.g., HCI)**]: Miscellaneous; H.1.2 [**Models and Principles**]: User / Machine Systems—*Human information processing, Human Factors*

## General Terms

Algorithms, Experimentation, Human Factors

## Keywords

crowdsourcing; competitions; gamification; groupsourcing; reward schemes; teams


*WWW 2015,* May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741097.

## 1. INTRODUCTION

The rapid development of information technologies, algorithms, and computer hardware has enabled the solution of many real life problems through efficient processing of large amounts of data and application of various methods in the area of artificial intelligence such as machine learning or decision theory. However, for many tasks such as semantic image annotation, mood analysis, or relevance ranking, humans still significantly outperform automatic methods. Projects such as Wikipedia or Open Street Maps which leverage collaborative human power have demonstrated that a large crowd of non-experts can, under certain conditions, be as effective and precise as a small group of experts [19, 17]. Information providers make implicit use of collaborative knowledge for improving their services: Search engines exploit query logs for determining the popularity and improving the automatic ranking 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 can help *intrinsically* motivating users to produce input in large quantities: In the ESP game [21, 22], 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's Mechanical Turk and CrowdFlower successfully make use of online workers and *monetary* incentives for accomplishing tasks such as the annotation of multimedia content, translation of texts, or generation of training sets for machine learning and other contexts [23, 25]. Such crowdsourcing platforms are based on a reward scheme where the payment of online workers is proportional to the number of accomplished tasks ("pay-per-task") which can be slow and costly. In our recent work [18] we successfully employed competitive scenarios to motivate users and reduce crowdsourcing costs, but can we obtain even more value for money?

Humans often collaborate in order to achieve a competitive advantage, and being a part of a team can additionally boost the motivation. Well organized teamwork can shorten the time required for completing a particular task by distributing the workload across team members; in addition, some complex tasks require diverse skills and can only be successfully solved by a group of people with expertise in different fields. Inducement prizes awarded for instance by companies like Microsoft, Google, or Yahoo for ideas or code, attract huge numbers of participants and create an enormous overall value. Participants often organize themselves

in teams in order to increase their chances. In the DARPA Network Challenge [20] red weather balloons, spatially distributed in a large area in the US, had to be located. An overall prize of 40,000 USD was awarded to the winner and participants were allowed to form search teams. In this way each participant could increase his probability of a successful task completion, but on the other hand would have to share the prize with other team members. The Netflix Prize is another 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; interestingly, some of the participating groups merged into one team ("The Ensemble"), combined their algorithms, and achieved competitive results. In our work [18] on crowdsourcing competitions we also noticed a tendency of some workers to cooperate in order to achieve a competitive advantage (manifesting in attempts to share accounts), which we had to prevent at that time due to our experimental design.

In this work we investigate how group mechanisms can be leveraged to improve the cost efficiency of crowdsourcing competitions. To this end, we first consider a strategy for team formation where workers are randomly assigned to one of a fixed set of teams, and teams compete against each other for a number of monetary prizes. Secondly, we explore an agglomerative strategy involving self-organization where workers (initially starting as one-man teams) can decide by themselves on which team they want to join, or whom to include in their team. Finally, we introduce and evaluate different combinations of team and individual competitions in a large-scale study.

*Outline.* The remainder of this paper is organized as follows: In Section 2 we discuss related work on games with a purpose and crowdsourcing designs. In Section 3 we formalize our crowdsourcing setting and describe our strategies for team competitions. 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 team formation and worker dynamics. Finally, in Section 5 we conclude and describe directions of our future work.

## 2. RELATED WORK

Crowdsourcing has a wide range of applications, including the annotation of data sets, conduction of user surveys, and collaborative gathering of data collections. Since 2011, TREC offers a special trek "TrecCrowd" [1] addressing various issues related to gathering document-relevance labels for information retrieval systems [3, 4]. This includes human task design, user filtering, and the fusion of worker judgments. In this paper, we investigate how collaborative, game-based crowdsourcing designs, can increase the effectiveness of annotations.

Typically, crowd workers can have different qualifications and perform tasks of different types not equally well. In their seminal work [13] 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 [14] the same authors show the correlation between behavior and personality of workers and the accuracy of their work. In [16] the effect of the amount of micro-payments on quantity and quality of annotations is studied. In contrast, our work focuses on how competitive reward mechanisms in collaborative scenarios influence annotation behavior and the cost efficiency of crowdsourcing.

Gamification techniques can provide additional motivation for the workers. The authors in [2] have focused on principal aspects of game mechanics embedded in standard IR tasks including information seeking, crowdsourcing, and user engagement. He et al. [11] 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 leaderboards. Dinesh et al. [17] discuss additional incentive structures to encourage user activity within an online platform. A number of earlier works tackle the same problems. Eickhoff et al. [10] 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 comparison, in this work we study non-linear reward distribution combined with team based strategies in order to increase the effectiveness of crowdsourcing.

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[1]. In [9] 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 prefer tasks with low rewards that better suit their abilities and tend to maximize their outcome by balancing reward and workload [24]. In [12] 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 and rewards are divided across multiple teams of workers.

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 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, none of these works provide an experimental evaluation of their concepts.

Finally, in our previous work [18] we introduced non-linear payment strategies for *individual* crowdsourcing competitions in combination with different information policies for boosting the cost efficiency.

[1] `http://www.topcoder.com`

To the best of our knowledge, we are the first to examine *team based* crowdsourcing competitions, and to conduct systematic real-world studies on the effects of collaboration and uneven reward distributions on both quantity and quality of annotations.

# 3. DESIGN OF TEAM-BASED CROWDSOURCING COMPETITIONS

In this section, we formalize our crowdsourcing scenario and describe different strategies for distributing rewards among workers in team competitions. Furthermore, we describe the information policies (relating to information revealed about the own and other teams as well as 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

We start by describing two baseline approaches: The first is just a formalization of the usual "pay-per-task" strategy within our framework, the second is from our recent work on individual crowdsourcing competitions. We then extend the latter approach towards team competitions.

*The Simple Baseline: Linear Reward Assignment.* The commonly used strategy in crowdsourcing is to distribute rewards proportionally 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-task").

*The Enhanced Baseline: Individual Competitions.* In our own recent work [18] we explored various competitive as well as randomized strategies where workers are ranked according to their produced values, and the reward of a worker will depend on his rank. In the following, we describe the best performing of these strategies. Formally, let $rank(w_i) \in \{1, \ldots, n\}$ be the rank of worker $w_i$, with a rank of $j$ corresponding to the $j$th highest value produced across all workers. The reward $r(w_i)$ is computed as a monotonically decreasing function $\Gamma(rank(w_i))$ of the worker's rank. Similar to "real world" competitions lower ranks are discounted, i.e. top performers receive more money per solved tasks, and $\Gamma$ is a convex function. In practice, instead of mathematical functions in closed form, one would rather provide workers with an easy to understand payment scheme containing "round" numbers for rewards and number of winners (e.g. 50, 25, 15, 8, and 2 USD for the top-5).

Part of the leaderboard content is revealed to a worker; this "medium" *information policy* turned out to be most effective in our previous crowdsourcing competitions [18]. Specifically, workers are shown their scores and rank along with a snapshot consisting of their $k$ neighbors above and below them in the leaderboard. The rationale is to trigger competitive behavior within part of the leaderboard while avoiding too much frustration for workers with lower scores. In [18], we showed that this form of competition clearly outperforms the conventional linear reward assignment described in the previous paragraph; therefore, we will compare our new methods in this paper to this enhanced baseline.

*Balanced Team Strategy.* In competitions among *individual* workers, relatively low performing workers (e.g. in terms of annotation speed) can quickly fall behind their competitors and become demotivated, resulting in a diminished overall value $V$ produced. In this paper we introduce *team based* competitions as a means of motivating a larger number of lower performers, rather than just incentivizing a relatively small number of high-performing individuals.

We first consider a balanced strategy for team formation, where we set up a fixed number $\tau$ of teams (each with initial team size 0) and assign each new worker (from the stream of registering workers) uniformly at random to one of the teams containing the lowest number of workers. This guarantees an even distribution of workers across the (disjoint) teams in $T = \{t_1, \ldots, t_\tau\}$. In order to avoid large numbers of inactive team members we introduce a low initial threshold for the value produced by worker $w_i$ before assigning him to a team. Instead of ranking individual workers, we now rank *teams* according to the sum of the values produced by their members, with $rank(t_i)$ defined analogous to the individual ranks described in the previous paragraph. A *team reward* $r(t_i)$ is assigned to each team, using, as for individual competitions, a monotonically decreasing, convex function. A worker $w_j$ receives a reward $r(w_j)$ that is proportional to the contribution to his team $t(w_j)$ - more specifically: $r(w_j) = r(t(w_j)) \cdot v(w_j) / \sum_{w_l \in t(w_j)} v(w_l)$, where $v(w_j)$ is the value produced by $w_j$.

We carry the previously described information policy for individual competitions over to the team scenario, i.e. during a competition members of a team see their positions along with scores of the $k$ neighbor teams above and below them in the leaderboard. In addition, workers are kept updated about their current shares of the team reward.

*Self-Organizing Team Strategy.* In order to avoid demotivation through team assignments imposed by the system, we also explore a strategy where workers can decide by themselves on which team they want to join, and whom to include in their team (or, alternatively, to stay single workers). Initially, each worker forms a one-man team and becomes its administrator. With the start of the competition begins the *team formation phase* where teams are built up in an agglomerative way as follows: The administrator of a team can invite the administrator of a second team; if the invitation is accepted both teams are merged and the inviting administrator becomes the (single) administrator of the whole team. This procedure can be repeated to form larger teams. The value produced by a team is equal to the sum of the values produced by its members during the whole competition, i.e.

values of merged teams are added up. Similar as for the balanced strategy, we require workers to produce a value above some low initial threshold before becoming involved in team formation activities. Note that workers (including administrators) are already actively involved in crowdsourcing tasks in the team formation phase. In order to support decision making, administrators gain access to statistics on the other teams - specifically, their sizes and produced values.

After the team formation phase and until the end of the competition no further merging of teams is possible, and administrators become "normal" workers without a dedicated view on team statistics. Except for administrators in the team formation phase, the information policy is the same as described for the balanced team strategy. We also rank teams and distribute rewards across workers in the same way as described in the previous paragraph.

*Combining Teams and Individual Competitions.* In order to account for different worker preferences we can combine team and individual strategies. To this end, the overall price money $M$ is split into a part for a team competition and a separate part for an individual competition. Workers conduct crowdsourcing tasks as before and obtain scores corresponding to their values produced. From these scores both the individual ranks as well as team ranks and shares are computed. The overall reward of the worker is the sum of his individual reward and his team share.

# 4. EXPERIMENTS

In this section we evaluate the team strategies described in Section 3 using face recognition as crowdsourcing scenario. The objective of our evaluation was to study the cost efficiency of strategies, competitive behavior of workers, and team dynamics.

## 4.1 Setup

*Crowdsourcing Task.* 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[2] database which was created for face verification [15]. Out of originally 58,797 images with faces of 200 celebrities, 37,004 images were available on the Web. We reviewed the dataset manually and removed 663 images showing placeholders as well as 135 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 solved the honeypot correctly, 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 CrowdFlower platform and 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 CrowdFlower. As the 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 upon registration which had to be submitted to the CrowdFlower task in order for the account to be activated, thus ensuring that each participant could be paid.

We started with two preliminary experimental rounds consisting of an individual competition and a group based competition. We observed an initial strong interest by workers for the novel type of scenario resulting in high participation rate, which quickly decline in the second experiment. Furthermore we received feedback from workers concerning the overlong duration of the experiments (originally six days) and missing communication features. To remedy these issues we reduced the experiment running time to three days and added automatic notifications for the chat. The final experiments were performed sequentially with a duration of three 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.* As a baseline, we chose the challenging "Individual Competition" strategy **ind** which has been shown to substantially outperform simple "pay-per-task" in [18]. We will provide evidence that reaching comparable performance using conventional "pay-per-task" would be unrealistic. We tested the team competition strategies "Balanced Teams" (**balanceTS**) and "Self-Organizing Teams" (**selfTS**) described in Section 3.2. For each strategy the prize money of $M = 100$ USD was divided into shares of 50, 25, 15, 8, and 2 USD for the top-5 performing teams or individuals. In addition, we conducted two experiments (coined **ind-balanceTS** and **ind-selfTS**) where we extended the team strategies with rewards for individual workers; to this end, the 100 USD prize money was equally divided into an individual and a team part with prizes of 25, 12, 7, 4, and 2 USD for the top-5 individuals/teams. Workers had to solve at least one batch correctly in order to qualify for participation in a team (for *balanceTS*) or for team formation activities (for *selfTS*). For all strategies we employed a medium information policy showing the scores of the three neighbor teams above and below the workers' teams in the leaderboard.

We chose a schedule with the most simple strategy at the beginning and incrementally increased the complexity over the course of the experimental series. More specifically, we conducted the experiments in the order *ind*, *balanceTS*, *ind-balanceTS*, *selfTS*, and *ind-selfTS*. Although this sequence of the experiments can have an influence on the outcomes, in contrast to a random order, it allows users to incrementally adapt to our scenarios and prevents a bias due to user fatigue which might have been in favor of our new strategies otherwise.

## 4.2 Results

In our experiments, overall 1.6 million images were matched correctly by 1,164 participants from 91 different countries (amounting to over 5,400 hours of work). The workers were accessing our competitions from several parts of the world, with a majority from Europe (38.0%) and Asia (34.7%). The country represented most numerously was India (14.4%), followed by Venezuela (6%) and the United States (4.2%).

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

| Experiment | No. Images | Cent / Hour | Cent / 100 Images |
|---|---|---|---|
| **Baseline** | | | |
| ind | 298,332 | 9.895 | 3.352 |
| **Balanced Teams** | | | |
| balanceTS | 327,073 | 8.967 | 3.057 |
| ind-balanceTS | 391,620 | 8.059 | 2.553 |
| **Self-Organized Teams** | | | |
| selfTS | 282,942 | 9.346 | 3.534 |
| ind-selfTS | 295,158 | 9.929 | 3.388 |

Table 1: Aggregate outcomes of the experimental rounds.

*Aggregate Results.* Table 1 shows the number of correctly matched images for each of the reward strategies, along with the average amount of money spent per correctly annotated image (both measuring the produced value), and hours of work (measuring monetary recompense for the participants). The main results are the following:

- Using the enhanced baseline (*ind*), we obtained almost 300,000 correctly annotated images at an average cost of 33.5 cent for 1,000 correctly annotated images. Note that using the standard "pay-per-task" scheme, this would result in a task that requires a contributor to annotate over 30 images for a reward of only 0.01 USD.
- The winning strategy - balanced teams combined with individual rewards (*ind-balanceTS*) - clearly outperforms the other strategies, and beats the *ind* baseline by a margin of 30%. This reduces the amount paid for 100 correctly annotated images by almost 24%.
- The pure team strategy *balanceTS* still outperforms the baseline *ind* by around 10%, in terms of value produced as well as time invested by participants.
- The remaining self-organizing team strategies were performing similar to the baseline, without being able to beat it.

Overall, the balanced team strategy with automatic group formation consistently outperformed the self-organizing team strategy. One reason for this might be the additional overhead and complexity for workers due to activities such as team building, communication, and coordination for the *selfTS* and *selfTS-ind* strategies. For instance, our logs revealed that participants in the *selfTS* experiment spent more than 10% more time per value produced in comparison to the baseline. Another reason might be the formation of a very small number of large and dominant teams which we avoid in the balanced case. In section 4.3 we will analyze issues of team dynamics in more detail.

*Worker Contributions.* Figure 1 depicts the contributions of individual workers to the overall produced value (i.e. fraction of correctly annotated images by the top-10 ($1 \leq R \leq 10$) and the remaining ($R > 10$) participants). The team based strategies benefit from the "fat tail" of the contribution distribution, where lower ranked participants were motivated by the performance of their fellows and contributed more value compared to individual strategies. For self-organizing teams (*selfTS*) we observe that lack of individual

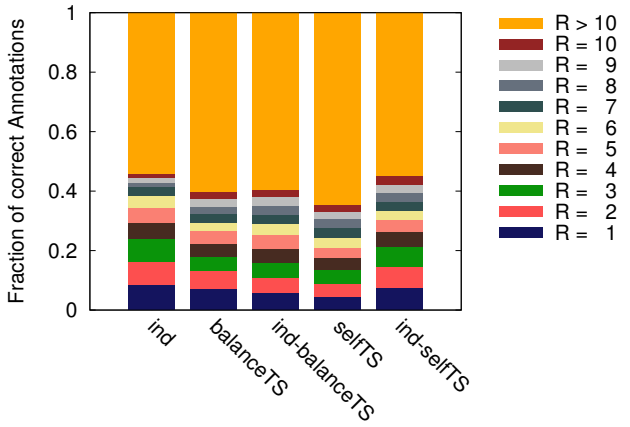



Figure 1: Individual worker contributions to the overall value in different experiments.

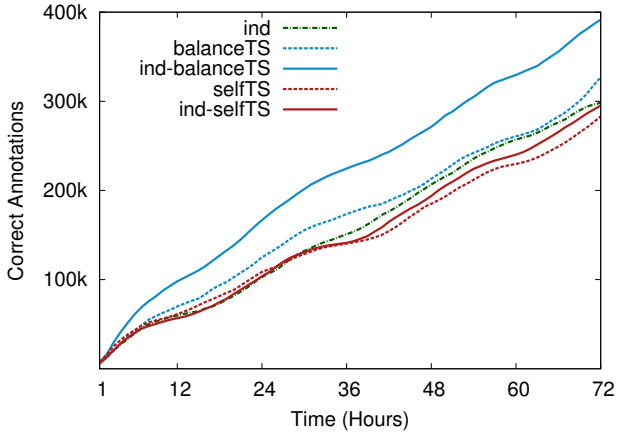
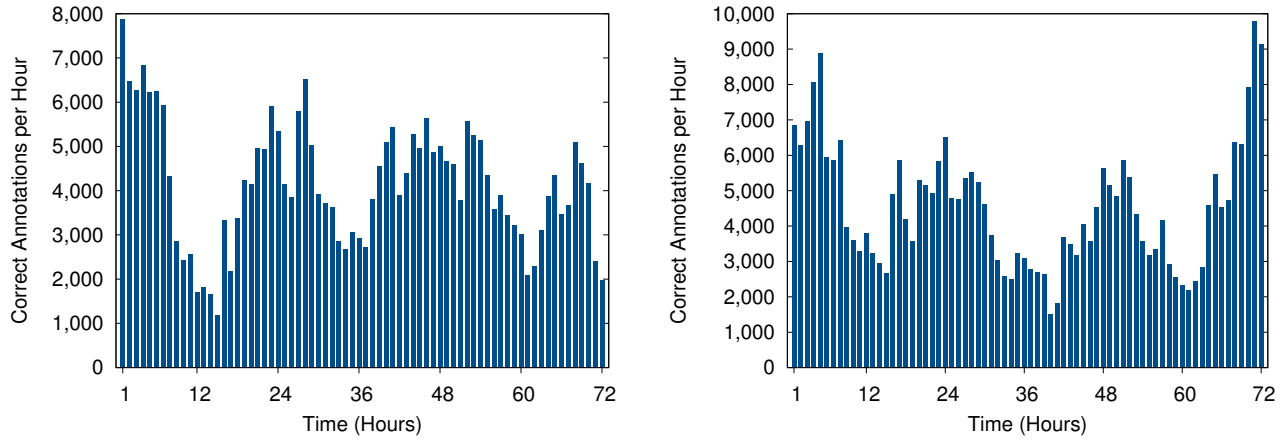



Figure 2: Cumulative number of correctly annotated images per experiment over time.

rewards and very large teams in the experiment can demotivate individual top performers. This effect can be remedied by combining group and individual competitions as shown for the *ind-selfTS* strategy.

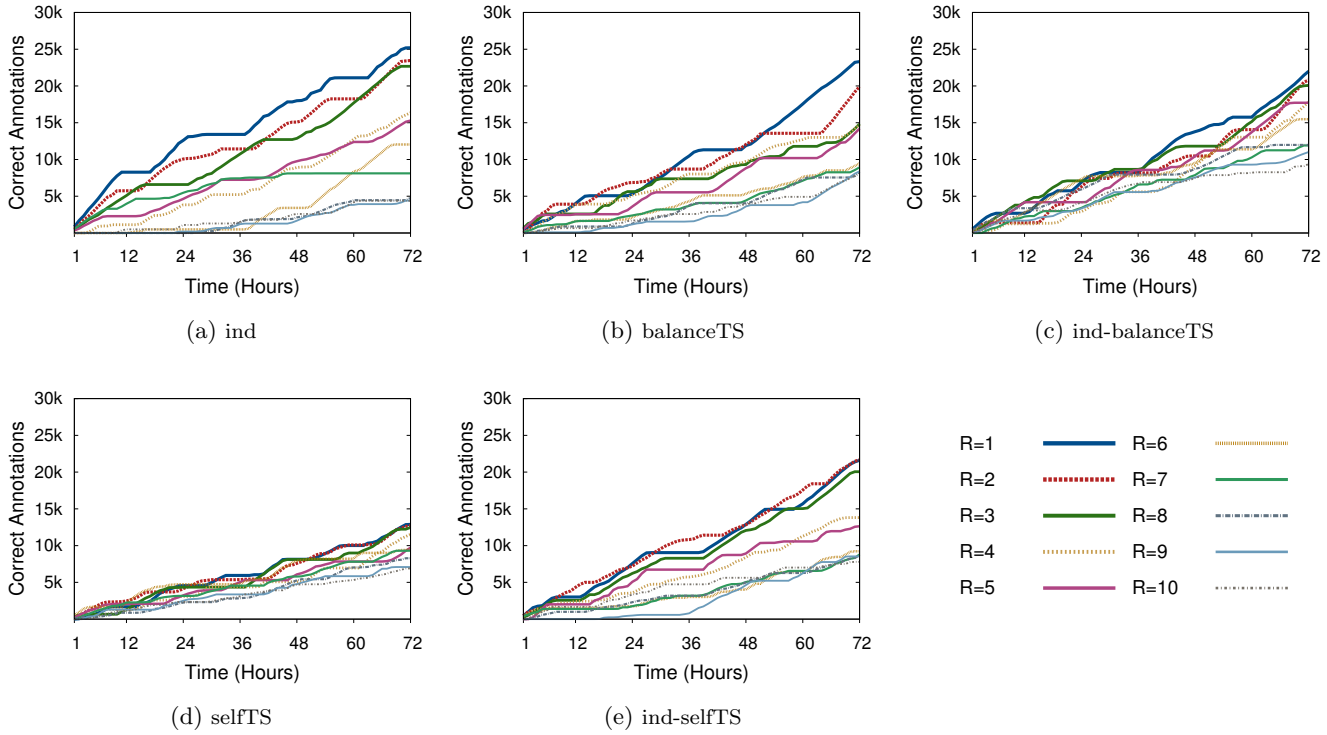
*Temporal Dynamics.* Figure 2 illustrates the temporal evolution of the cumulative counts of correctly solved images aggregated over the whole set of workers. The numbers grow mostly linearly over time with higher increments at the beginning of the competitions and periodicities which are probably due to day cycles, similar to our findings in [18]. However, due to the larger and more diverse user base of CrowdFlower compared to Mechanical Turk in our previous work, the periodicity is less pronounced (for further illustration see the non-cumulative view for the individual competition *ind* and balanced team competition *balanceTS* in Figure 3). In addition, Figure 3b reveals fierce competitions between workers for the *balanceTS* strategy (e.g. for team ranks 1 and 3 towards the end of the contest), leading to a boost in the amount of annotations.

We also studied the contribution of individual workers over time. Figure 4 breaks the cumulative annotation counts down for the top-10 individual workers. Compared to indi-




(a) ind (b) balanceTS

Figure 3: Temporal characteristics of annotations for individual competitions (*ind*) and balanced team competitions (*balanceTS*).




(a) ind (b) balanceTS (c) ind-balanceTS

(d) selfTS (e) ind-selfTS

Figure 4: Cumulative number of correctly annotated images for the top-10 participants over time.

vidual competitions we observe considerably higher scores for the lower ranked participants in team-based competitions. For our winning strategy (*ind-balanceTS*) in particular, the spread of the curves is much less pronounced, highlighting that in contrast to the individual competition *ind*, lower performers remain motivated even when confronted with stronger competitors. Except for the *selfTS* strategy, the top performers annotated more than 20,000 images each, indicating adequate motivation in the team scenarios.

*Annotation Quality.* In terms of correct annotations there is no noticeable difference between the baseline and any of our team-based strategies (see Table 2). For the large majority of batches with correctly solved honeypots the percentage of accurately recognized faces is over 95% for all of the strategies. The quality slightly increased in the course of our experimental runs, compensating to some degree the handicap of worker fatigue. Similarly, the number of batches with incorrectly solved honeypots declined over time as the

| Exp. | Correct Honeypots: No. of batches | Correct Honeypots: Correctly matched faces | Incorrect Honeypots: No. of batches | Incorrect Honeypots: Correctly matched faces |
|---|---|---|---|---|
| ind | 2,996 | 95.8% | 85 | 81.0% |
| balanceTS | 3,293 | 96.3% | 81 | 74.6% |
| ind-balanceTS | 3,964 | 96.6% | 66 | 85.2% |
| selfTS | 2,845 | 96.7% | 65 | 82.8% |
| ind-selfTS | 2,966 | 97.1% | 41 | 83.7% |

Table 2: Quality results.

users became more proficient at the task, whilst the fraction of correct annotations for these batches remained stable.[3]

## 4.3 Team Formation and Dynamics

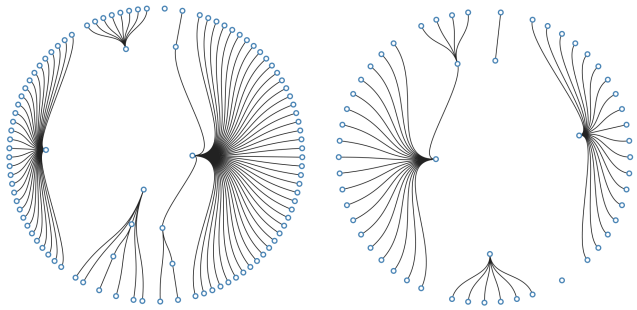
In the following we examine team structures and temporal dynamics in order to shed further light on the performance differences between team strategies.

*Team Composition.* Figure 5 shows the contributions of workers to their team for the top-10 teams. We observe a strong competition between the top teams towards the end of the *selfTS* and *balanceTS* experiments, resulting in very similar performance characteristics for the competitors. For the combined strategy *ind-balanceTS*, especially participants in lower ranked teams were additionally engaged through individual rewards and contributed more value. Although some of the top workers seemed to be demotivated by the relatively weak performance of their teams, generally more participants were highly motivated in the group based strategies in comparison to the *ind* baseline.

For both of the self-organized team strategies the emerging team structure turned out to be suboptimal: the competitions were dominated by few very large teams, offering only small shares to strong performers. This issue was alleviated with the introduction of individual rewards in *ind-selfTS*, where the contributions of the top performers are about twice as high as for *selfTS*. Designing and testing team formation conditions (such as stronger restrictions on teams sizes) that lead to more competitive team scenarios is a challenging task for future work.

*Team Formation Processes.* Figure 6 visualizes team formation events (specifically the merging into larger teams) along with the development of the overall value produced by the top-5 teams for the group strategies *selfTS* and *balanceTS*. Over the course of the team formation phase, a large number of participants decided to join one of the leading teams - instead of collaborating with smaller teams in order to compete for the lower prize ranks ("winning team joining" pattern). However, we also observed occasional mergings of smaller teams if a higher rank in the leaderboard could be achieved in that way ("competitive merging" pattern). Finally, for both strategies fierce competitions occurred towards the end of the experiment (e.g. between rank 1 and 2 as well as rank 3 and 4). A dendrogram view of the team formation for the top-5 teams in both self-organized teams

[3]The results for *balanceTS* are an exception: two of the workers tried to game the system by annotating randomly.



(a) selfTS (b) ind-selfTS

Figure 7: Team merging structure of top-5 teams for self-organizing team strategies (*selfTS* and *ind-selfTS*)

experiments is shown in Figure 7, where merges evolve from the outer to the inner part of the circles. The "winning team joining" pattern appeared in over 80% of all mergings.

In total, 550 and 284 invitations were posted for *selfTS* and *ind-selfTS*, where 75.4% and 57.3% were accepted after an average of 2.1 and 4.9 hours, respectively. These numbers exclude the invitations ignored or automatically canceled due to engagement in concurrent merging activities. The decline in the number of invitations in conjunction with the lower acceptance rate for subsequent experiments indicate that workers often preferred smaller teams after having participated in one of the very large teams before.

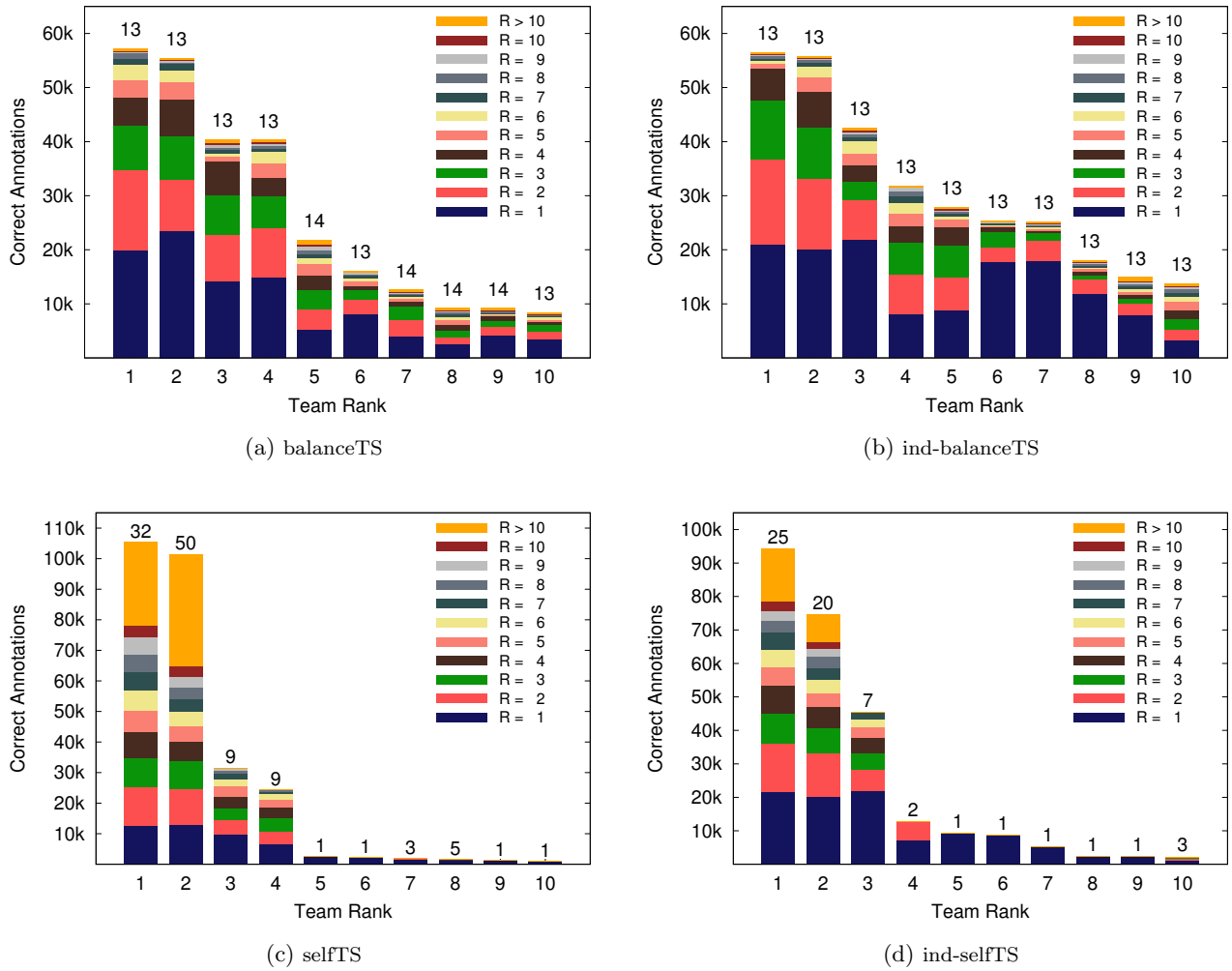
## 4.4 Worker Activity and Interaction

*Worker Activity.* Table 3 shows the number of participating contributors who solved at least one batch of 100 instances. The discrepancy between registered users and users who actually provided annotations is mostly due to particularities of CrowdFlower where each submission of a user code (i.e. registration) had to be rewarded with 1 cent[4]. Over the course of the experiments, we noticed user fatigue manifesting in a continuous, slight decrease in the number of participants, putting the baseline strategy at an advantage.

Figure 8 shows the number of workers with respect to the number of experiments they participated in; workers are further divided into the ones that won a monetary prize in at least one of the experiment and the ones who didn't. Similar to our observations in [18], workers who had won prizes in earlier rounds, were more motivated to participate in further rounds, with 11.5% of all workers engaged in at least 3 competitions.

*User Feedback.* During the experiments we continuously received feedback from the participants. Besides administrative issues (such as missing or inappropriate images), participants also suggested new features such as team chat functionalities which we implemented right after the preliminary rounds. User comments also provided us with valuable insights into details of the competition design. In a preliminary balanced team experiment, for instance, workers were assigned to a team immediately upon login, which was often

[4]During the *selfTS* and *ind-selfTS* experiments, we additionally observed periods without any registrations after the start of the competition due to technical problems of the platform.



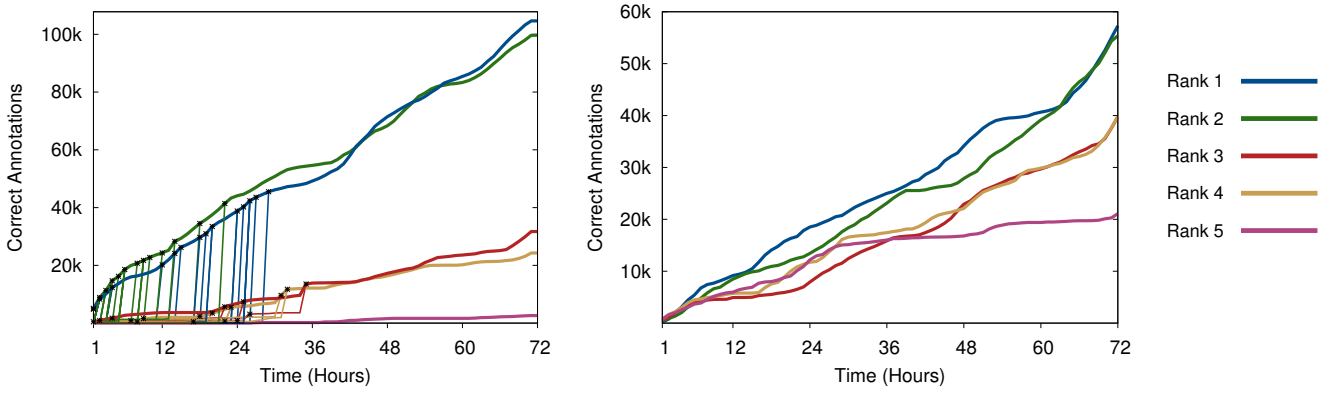

(a) balanceTS

(b) ind-balanceTS

(c) selfTS

(d) ind-selfTS

Figure 5: Contributions to the top-10 teams for the team based strategies. We show the number of correct annotations and the contribution of each individual team member (ordered by rank within the team). The size of each team is shown above the respective column.




(a) selfTS

(b) balanceTS

Figure 6: Temporal development of correct annotations of the top-5 teams for *selfTS* (left) and *balanceTS* (right). For *selfTS* we also show merging events during the team formation phase (i.e. in the first 36 hours).

| Exp. | registered | annotated | solved batch |
|---|---|---|---|
| ind | 839 | 448 | 300 |
| balanceTS | 785 | 382 | 282 |
| ind-balanceTS | 732 | 342 | 273 |
| selfTS | 586 | 219 | 181 |
| ind-selfTS | 763 | 316 | 222 |

Table 3: Participation statistics

not considered fair ("`we have a lot of members who has 0 points.(...) So we are at disadvantage!`"). Consequently, in our main series of experiments, users were only assigned to a team after first solving one batch of tasks. To ensure setup consistency throughout the experiments we had to discard numerous requests, such as the suggestion to adopt the interface for mobile devices.

*Team Communication.* Over the course of the experiments, about 2,500 messages were posted by over 200 of the participants. Conflicts, such as discussions about reward shares, were scant and most of the communication remained positive and friendly "`Thank you all for your hard work. May we meet again in another competition either as a team or a competitor. Good luck to you all.`". Lots of messages were very constructive, including motivating statements like "`Let's go team !!! we are 5, team A are 3. We can reach them !!!`" and "`chat later, focus on work fast.`".

Team members supported each other with working advice and clarification of the rules:

**user 1:** `What if I answer wrong?`
**user 2:** `we will lose 20 points :) go easy press F11 and than u will see better`

as well as discussing their own strategy with others:

**user 3:** `I'm trying to get to number 5 spot because he/she stopped clicking.`
**user 2:** `Yeah but u need 2000 thousand more buddy, and you know that he/she will be careful now :/ she will check again to see if you will attack and then he/she will start doing more [...]`
**user 3:** `good point`

Furthermore, many of the participants collaborated on team administration and coordination issues in a rather democratic way. For instance, team administrators sought out the opinions of their team members regarding new invitations and teams sizes ("`Ok. help me with the decisions`", "`Don't invite anyone else so we don't have to split the reward with so many people! [...]`"). Team leaders received input from members about experiences made in previous competitions, such as suggestions for inviting former co-workers ("`<XXX> has worked with me.i still can watch the old competitions.he scored 8400 points,so you decide :)`"). Strategic decisions were discussed about catching up with competitors, defending the current position, or pausing to keep the share proportions stable.

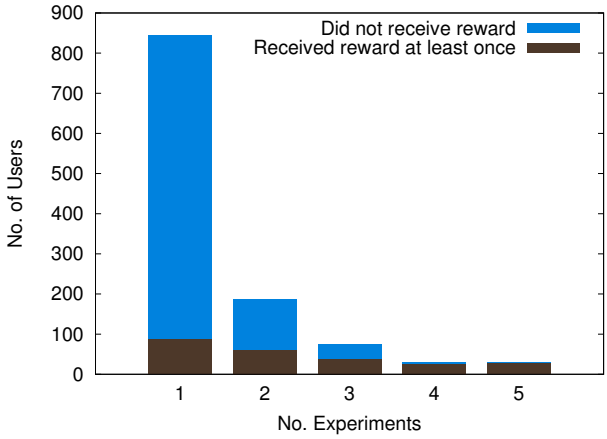



Figure 8: Worker participation.

Although some of the workers appreciated having control in the self-organizing team strategy ("`[...] this system.. it's stable and perfect.. all in our hands(public) but not of system automatically selecting arranging them in teams..`"), workers worried about inactive team mates ("`those people who haven't done much kinda piss me off`"). In our future work, we consider also offering a mechanism for democratically ruling out members.

## 5. CONCLUSIONS AND FUTURE WORK

We have studied various team competition strategies for crowdsourcing, including balanced teams, self-organizing teams, and combination of team and individual strategies. Our evaluation shows substantial performance boosts for team-based scenarios: Our best approach results in 30% more annotations than the recent state-of-the-art baseline, with no decrease in the quality of annotations. Our results indicate that balancing team sizes plays a crucial role in fostering engagement in competitions. In addition, the combination of team and individual competition helps to overcome declines in performance through weak teams. Our analysis of group formation processes, temporal dynamics in competitions, and user interactions sheds further light on worker motivation and behavior.

In our future work we aim to explore additional team formation mechanisms, including democratic decision processes, options for hiring and firing, and hierarchical worker structures. Furthermore, we plan to conduct more in-depth studies on the influence of parameters such as maximum team size, reward distribution, and duration of competitions. Finally, we want to study crowdsourcing scenarios that require closer collaboration and stronger expertise in different areas. To this end, we aim to introduce more sophisticated communication mechanisms that allow for effective coordination between team members and teams.

## 6. ACKNOWLEDGMENTS

This work is partly funded by the European Research Council under ALEXANDRIA (ERC 339233) and by the European Commission FP7 under QualiMaster (grant agreement No. 619525).

## 7. REFERENCES


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

[2] *GamifIR '14: Proceedings of the First International Workshop on Gamification for Information Retrieval*, New York, NY, USA, 2014. ACM.

[3] O. Alonso and R. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 153–164, Berlin, Heidelberg, 2011. Springer-Verlag.

[4] O. Alonso and S. Mizzaro. Using crowdsourcing for trec relevance assessment. *Information Processing & Management*, 48(6):1053–1066, Nov. 2012.

[5] N. Archak. Money, glory and cheap talk: Analyzing strategic behavior of contestants in simultaneous crowdsourcing contests on topcoder.com. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 21–30, New York, NY, SA, 2010. ACM.

[6] N. Archak and A. Sundararajan. Optimal design of crowdsourcing contests. In *Proceedings of the International Conference on Information Systems*, ICIS 2009, Phoenix, Arizona, USA, 2009. Association for Information Systems.

[7] R. Cavallo and S. Jain. Efficient crowdsourcing contests. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '12, pages 677–686, Richland, SC, 2012. International Foundation for Autonomous Agents and Multiagent Systems.

[8] R. Cavallo and S. Jain. Winner-take-all crowdsourcing contests with stochastic production. In *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*, Palm Springs, CA, USA, 2013. AAAI.

[9] D. DiPalantino and M. Vojnovic. Crowdsourcing and all-pay auctions. In *Proceedings of the 10th ACM Conference on Electronic Commerce*, EC '09, pages 119–128, New York, NY, USA, 2009. ACM.

[10] C. Eickhoff, C. G. Harris, A. P. de Vries, and P. Srinivasan. Quality through flow and immersion: Gamifying crowdsourced relevance assessments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 871–880, New York, NY, USA, 2012. ACM.

[11] J. He, M. Bron, L. Azzopardi, and A. de Vries. Studying user browsing behavior through gamified search tasks. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, GamifIR '14, pages 49–52, NY, USA, 2014. ACM.

[12] H. Jiang and S. Matsubara. Improving crowdsourcing efficiency based on division strategy. In *Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 425–429, Los Alamitos, CA, USA, 2012. IEEE Computer Society.

[13] G. Kazai. In search of quality in crowdsourcing for search engine evaluation. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 165–176, Berlin, Heidelberg, 2011. Springer-Verlag.

[14] G. Kazai, J. Kamps, and N. Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1941–1944, New York, NY, USA, 2011. ACM.

[15] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *Proceedings of the 12th IEEE International Conference on Computer Vision*, ICCV 2009, pages 365–372, Piscataway, NJ, USA, 2009. IEEE Computer Society.

[16] W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". *SIGKDD Explorations Newsletter*, 11(2):100–108, May 2010.

[17] D. Pothineni, P. Mishra, A. Rasheed, and D. Sundararajan. Incentive design to mould online behavior: A game mechanics perspective. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, GamifIR '14, pages 27–32, New York, NY, USA, 2014. ACM.

[18] M. Rokicki, S. Chelaru, S. Zerr, and S. Siersdorfer. Competitive game designs for improving the cost effectiveness of crowdsourcing. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, CIKM '14, New York, NY, USA, 2014. ACM.

[19] N. Savage. Gaining wisdom from crowds. *Communications of the ACM*, 55(3):13–15, Mar. 2012.

[20] J. C. Tang, M. Cebrian, N. A. Giacobe, H.-W. Kim, T. Kim, and D. B. Wickert. Reflecting on the darpa red balloon challenge. *Commununications of the ACM*, 54(4):78–85, Apr. 2011.

[21] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 319–326, New York, NY, USA, 2004. ACM.

[22] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commununications of the ACM*, 51(8):58–67, Aug. 2008.

[23] P. Welinder and P. Perona. Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 25–32, June 2010.

[24] J. Yang, L. A. Adamic, and M. S. Ackerman. Crowdsourcing and knowledge sharing: Strategic user behavior on taskcn. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, EC '08, pages 246–255, New York, NY, USA, 2008. ACM.

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