Please cite the updated version of this work:
M. Hirth, T. Hossfeld, P. Tran-Gia.
"Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms."
Mathematical and Computer Modelling, 2012. DOI: 10.1016/j.mcm.2012.01.006

University of Würzburg
Institute of Computer Science
Research Report Series

Cheat-Detection Mechanisms for Crowdsourcing

Matthias Hirth, Tobias Hoßfeld, Phuoc Tran-Gia

Report No. 474 August 2012

University of Wuerzburg, Germany
Institute of Computer Science
Chair of Communication Networks
Am Hubland, D-97074 Würzburg, Germany
{matthias.hirth|hossfeld|trangia}@informatik.uni-wuerzburg.de

Please cite the updated version of this work:
M. Hirth, T. Hossfeld, P. Tran-Gia.
"*Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms.*"
Mathematical and Computer Modelling, 2012. DOI: 10.1016/j.mcm.2012.01.006

Cheat-Detection Mechanisms for Crowdsourcing

Matthias Hirth, Tobias Hoßfeld, Phuoc Tran-Gia
University of Wuerzburg, Germany
Institute of Computer Science
Chair of Communication Networks
Am Hubland, D-97074 Würzburg, Germany
{matthias.hirth|hossfeld|trangia}@informatik.uni-wuerzburg.de

Abstract

Crowdsourcing is becoming more and more important for commercial purposes. With the growth of crowdsourcing platforms like MTurk, Microworkers or Innocentive, a huge work force and a large knowledge base can be easily accessed and utilized. But due to the anonymity of the workers, they are encouraged to cheat the employers in order to maximize their income. Thus, this paper presents two crowd-based approaches to detect cheating workers. Both approaches are evaluated with regard to their detection quality, their costs and their applicability to different types of typical crowdsourcing tasks.

Keywords: Crowdsourcing, Cheat-Detection

Please cite the updated version of this work:
M. Hirth, T. Hossfeld, P. Tran-Gia.
"Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms."
Mathematical and Computer Modelling, 2012. DOI: 10.1016/j.mcm.2012.01.006

1. Introduction

With the rise of the Internet and the tremendous growth of its user base, a huge workforce with a large amount of knowledge developed. This is already exploited in projects like Wikipedia[13], where users created a high quality encyclopedia by sharing this knowledge, or OpenStreetMap[10] which offers detailed maps from all over the world based on information gathered by its users.

A new approach to use this workforce and the wisdom of the crowd is referred to as *crowdsourcing*. Crowdsourcing can be viewed as a further development of outsourcing. With outsourcing, entrepreneurs choose specialized subcontractors to accomplish certain parts of a development or production process. In contrast, "crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call" [7]. According to Jeff Howe's definition, the main differences to outsourcing are that the entrepreneur does not know who accomplishes his task and that the workers do not form an organized group like a firm but are members of a large anonymous crowd accessible via crowdsourcing platforms. In traditional work organization, the employer delegates work to the workers, but in the crowdsourcing approach, the worker chooses which tasks he wants to work for.

At the beginning crowdsourcing was often used for non-profit applications. But with the development of platforms like MTurk [4] or Microworkers[12], which offer an easy access to a huge amount of workers, crowdsourcing became also interesting for commercial usage. The main advantage of commercial crowdsourcing in comparison to the traditional approach is the possibility to get work done very quickly by accessing a large and relatively cheap workforce or to get innovations for "free". However, the work results are not reliable. Some workers try to submit incorrect results in order to maximize their income by completing as many jobs as possible in the least amount of time. Sometimes a small amount of incorrect results can be tolerated, but in general they have to be filtered. This has to be done by the employer or by a trustworthy employee and consumes a lot of time and money compared to the original crowdsourcing jobs. Therefore, techniques have to be developed to detect cheating workers and invalid work results.

In this paper, we present two approaches to detect cheating workers. As the approaches are also based on crowdsourcing, they are easily integrable in common crowdsourcing platforms. Besides a general description of the methodology, we evaluate the quality of our cheat-detection and discuss the costs of the proposed solutions. To depict the applicability of the approaches, we give some use cases for common crowdsourcing tasks and formulate general guidelines how to use our findings for improving the result quality of these tasks.

The paper is structured as follows. Section 2 gives a quick overview of the concept of crowdsourcing and the research already done in this area. In Section 3 we present our two approaches for work validation, which are evaluated in Section 4. The costs of the approaches are analyzed in Section 5, which also contains relevant examples of use cases. The paper is concluded in Section 6.

2. Background and related work

In the following, we give a quick overview of the general ideas and common terms of crowdsourcing. We show typical examples of crowdsourcing tasks and introduce a rough categorization of these tasks, based on the required worker skills.

2.1. Crowdsourcing Scheme and Terminology

Applying the crowdsourcing approach, an employer does not selectively choose a worker for a certain task, but offers the task to a large crowd of workers who can freely choose to work on this task or not. In general, there has to be a mediator between the employers and the workers, the so called *crowdsourcing platform* which is schematically depicted in Figure 1. Well known examples of these platforms are the aforementioned MTurk and Microworkers.

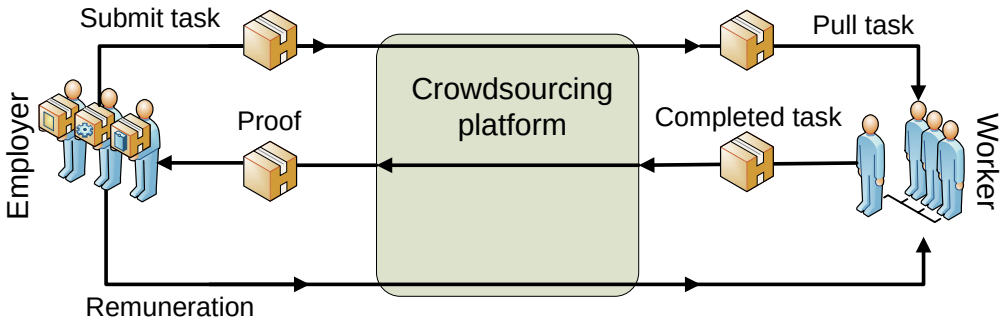

Figure 1: Crowdsourcing scheme

An employer submits a *task* to the crowdsourcing platform and defines how much the workers will be paid per task and how the workers have to proof a completed task. Usually the same task has to be done many times. Therefore, multiple copies of the same task are organized as a *campaign*. Random workers from the crowd choose to work on the task and after completion submit the required poof to the crowdsourcing platform. The work proof is forwarded to the employer, who pays the worker if the task was completed correctly.

2.2. Typical Crowdsourcing Tasks and Their Categorization

Crowdsourcing can be used for various purposes which can be roughly categorized into *routine tasks*, *complex tasks*, and *creative tasks*. Routine tasks are jobs which do not require any level of qualification, like bookmarking a web page using social bookmarking services as digg [1] or delicious [14], clicking on advertisements on web pages, relevance evaluation [3], or posting a given text into a forum. *Complex tasks*, like rewriting a given blog or text annotation [8], need some general skills, in contrast to *creative tasks* where highly specialized skills are required. *Creative tasks* include writing an article on a given topic or even research and development [2].

Usually, detecting cheating workers is more difficult for complex tasks than for routine tasks. Assume a routine task, where a worker has to digg a web page p and leave a comment about p on the digg web page. The worker has to submit his digg user name

in order to proof that the task is completed. It is easy to automatically check the digg page whether the given user name left a comment on p or not. This is exemplary for routine tasks, where verification is often simple and easy to automatize. For a complex or a creative task the verification of the workers results is more complicated. Assume a complex task, where a worker as to rewrite a given text and a creative task where a worker has to write a text on a given topic. In both cases the worker's texts have to be read and rated according to their content and their writing. This can not be automatized and especially for the complex task the reviewer also needs some background knowledge to judge the relevance of the worker's text.

2.3. Related Work

As pointed out, crowdsourcing applications suffer from untrustworthy workers, which try to submit an invalid poof in order to receive a payment without completing the required task. Kittur et al. [9] showed that the quality of some tasks can be increased by adding verification questions in order to detect cheating workers. Also Chen et al. [5] point out that cheating workers are a problem, but they stated that there is no systematic way to cheat their crowdsourcing platform for Quality of Experience tests. However, they do not describe general cheat-detection mechanisms. In [11], Ahn and Dabbish present a crowd-based image labeling game, which is now used in an adapted version by Google's Image Labeler[6]. In this game a picture is shown simultaneously to two random workers, who suggest labels for the image. If both suggest the same label, it is associated with the picture. Ahn and Dabbish argue that cheating is not possible in this application, because the game is played by a huge number of people and the partners are chosen randomly. Therefore, they are very unlikely to know each other and to collaborate.

Currently cheat-detection techniques are either highly specialized, are based on control questions which are evaluated automatically or rely on manual checking by the employer. In the next section, we present two generic crowd-based approaches for tasks where control questions are not applicable and manual re-checking is ineffective.

Please cite the updated version of this work:
M. Hirth, T. Hossfeld, P. Tran-Gia.
"Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms."
Mathematical and Computer Modelling, 2012. DOI: 10.1016/j.mcm.2012.01.006

3. Crowd Based Cheat Detection Mechanisms

We propose two different approaches of crowd-based cheat-detection techniques: A majority decision (MD) and an approach using a control group (CG) to re-checking the main task. In order to analyze these approaches we use the model described in the following. The most important variables used for this model, as well as the ones introduced for the evaluation in Section 4 and the cost model in Section 5 are summarized in Table 1 in the Appendix.

3.1. General Notion and Variables

In our model, the crowd consists of N individual workers and each worker can only work on one task in a campaign. As mentioned above, not every worker is good-willing and performs the task correctly. However, we can assume that these workers do not intend to falsify the task result but only give a random result in order to complete the task as fast as possible. To account for this, we assume that there is a probability p_c that a randomly chosen worker is a cheater, which means the worker will submit a random result. This result is wrong with a probability of $p_{w|c}$. If a wrong result is detected in a real campaign, it is not possible to decide whether the worker submitted it willingly or accidentally and the worker is treated as a cheater in both cases. Thus, in our model we can assume that only cheaters submit incorrect results, i.e. $p_{w|\overline{c}} = 0$. Good-willing users, who accidentally submit an invalid result can be modeled by adjusting p_c. Therefore, the probability of a wrong task result is $p_w = p_c \cdot p_{w|c}$. To clarify this imagine a multiple choice test with one correct answer out of four possibilities and a crowd of 100 workers including 10 cheaters. The probability of choosing a cheater is $p_c = 10\%$, the probability for picking a wrong answer when choosing randomly $p_{w|c} = 75\%$. This results in a probability for a wrong answer $p_w = 7.5\%$.

3.2. The Majority Decision Approach

A simple approach to eliminate incorrect results is to use a majority decision (MD). This means the same task is given to i different workers and the results are compared. The result which most of the workers submitted is assumed to be correct.

Figure 2 depicts the work flow of a campaign using the MD approach. Similar to a normal campaign, the employer submits his task to the crowdsourcing platform (1). However, the task is not only done by one worker, but the crowdsourcing platform duplicates the task several times and offers the same task to a large number of workers (2). Each of these workers submits his individual task result (3), which might be correct or incorrect. The crowdsourcing platform performs a majority decision to validate the correct result (4) and returns it to the employer (5). In this approach each worker submitting a result is paid.

As an example application of the MD approach, think of an relevance evaluation of an article, where 100 workers have to judge whether the text is related to a certain topic and 95 workers rate it off-topic. Even if there are some cheaters among the workers, the article is with a high probability off-topic.

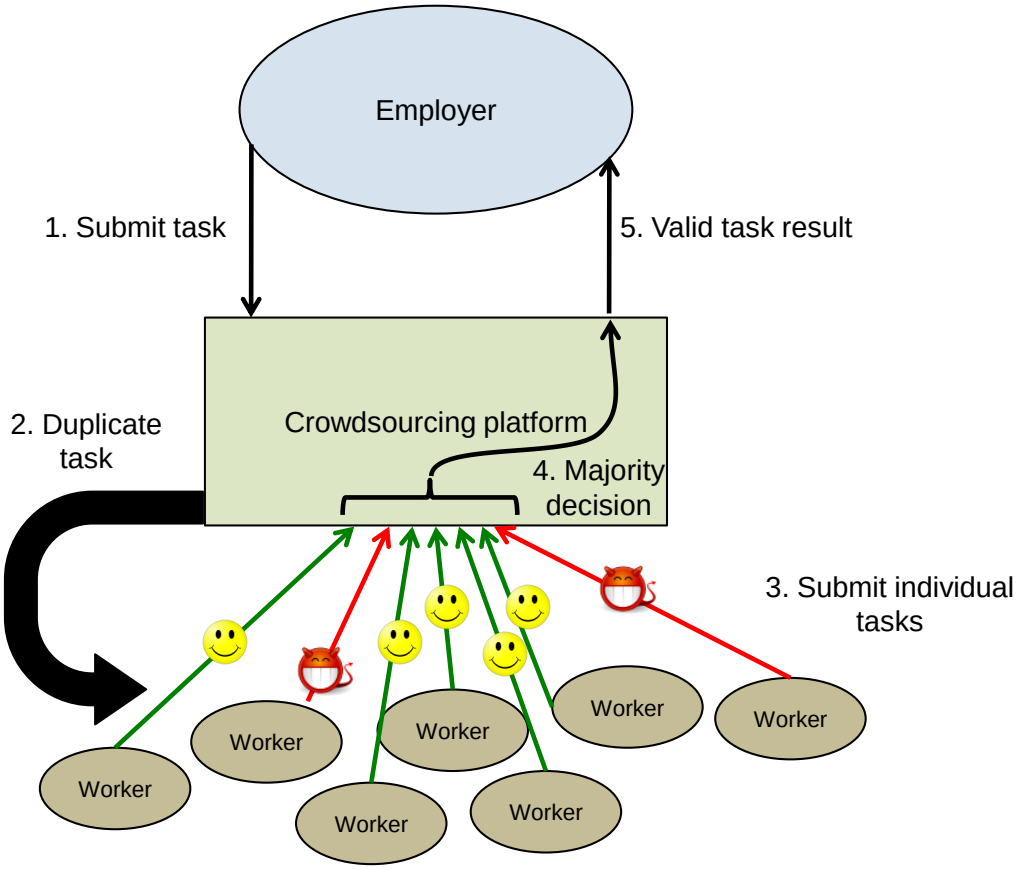

Figure 2: MD approach scheme

3.3. The Control Group Approach

Our second approach is based on a control group (CG). A single worker works on a main task and a control group consisting of j other workers re-checks the result, whether it is valid or not. A task is considered to be valid, if the majority of the control group decides the task is correctly done. An important point of this approach is that the main task and the "re-check" task are assumed to have different costs. Usually the main task is expensive, like writing an article to given keywords, while the control task is cheaper.

The work flow of a CG campaign is shown in Figure 3. The employer submits the main task to the crowdsourcing platform (1) and the task is chosen by a worker (2), who submits the required task result (3). This result is not directly returned to the employer, but the crowdsourcing platform generates a new campaign for validating this result. Therefore, the result of the main task is given to a group of workers, who rate it according to given criteria (4). The ratings of the different workers are returned to the crowdsourcing platform (5), which calculates the final rating of the main task using a majority decision (6). This is necessary, because the workers in the control group may be cheating and submit wrong ratings. If the main task is rated valid, the main worker is paid (7) and the result is returned to the employer (8).

A possible application of this approach is a task where a worker has to write a long article including some given keywords. This is a quite expensive task, as it is time intensive and the worker has to be creative. As the worker submits the completed article, a new

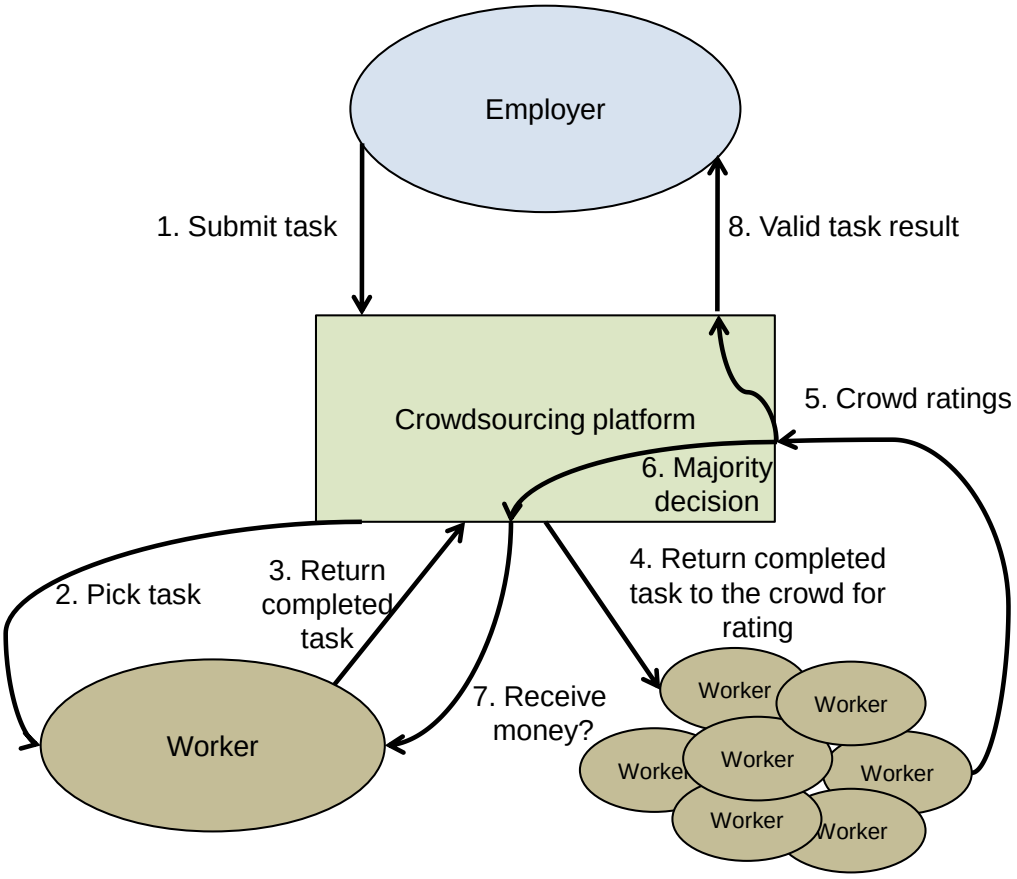

Figure 3: CG approach scheme

campaign is created during which 100 workers have to judge whether the article matches the initial keywords or not. This task is simple and therefore, less paid. If enough workers submit that the article matches the given keywords, it can be assumed to be valid.

4. Evaluation of the MD and the CG approach

Both approaches use a majority decision of m workers in order to verify the task result. Therefore, we have a closer look on how to choose an optimal m for the majority decision, in order to minimize the costs and maximize the reliability of the decision result. Afterwards, we have a closer look at the quality of the cheat-detection of the MD and the CG approach.

4.1. Group Size for Majority Decisions

For a majority decision we use m random workers from the total crowd of N workers. Each of these m workers is with a probability of p_c a cheater and submits an invalid result with the probability $p_{w|c}$. Thus, the number of invalid results X follows a binomial distribution $X \sim \text{BINOM}(m, p_w)$. In order to derive a correct majority decision, the number of incorrect results has to be smaller than $\frac{m}{2}$, i.e., the probability of a correct majority decision p_m is given by

$$\begin{aligned} p_m &= P(X < \frac{m}{2}) = P(X \leq \left\lfloor \frac{m-1}{2} \right\rfloor) \\ &= \sum_{k=0}^{\lfloor \frac{m-1}{2} \rfloor} \binom{m}{k} p_w^k (1-p_w)^{m-k}. \end{aligned} \tag{1}$$

The total cost of a majority decision increases with the number m of workers involved. Figure 4 depicts the dependency between p_m and m for a high probability of an invalid result $p_w = 0.4$. The calculated values of p_m for odd group sizes are marked with circles, the values for even group sizes with squares.

The probability of a correct majority decision p_m is not constantly increasing with the group size, but is dependent on the parity of the group. Smaller groups of an odd number of workers always achieve better results than slightly larger groups of an even number, a mathematical poof is given in the Appendix. Alternatively this effect can be explained by the stalemate when using even groups. While there is always a valid majority decision in odd groups, a stalemate can occur in a majority decision using even groups. In this situation, the majority decision is also assumed to be not correct as the decision process has to be repeated. The exploitation of this effect is helpful in any application of majority decisions, as it helps both to improve the reliability of the results and to lower the costs.

We analyze the possible cost savings in more detail. To this end, we calculate how many workers are needed for the 99% quantile of a correct majority decision (1) in a group with even parity and (2) in a group with odd parity. For both groups the number of workers is dependent on p_w. The results are depicted in Figure 5; note that the y-axis is in logarithmic scale. The required size of the even group is shown by the dark continuous line, the size of the odd group by the dashed line and the difference of both group sizes by the light continuous line.

From Figure 5 we can conclude two findings. First, we can at least save one worker when using an odd instead of an even group size without reducing the quality of our majority decision. Second, the higher p_w is the more workers can be saved using an odd group size. In the remainder of this paper we will only use odd group sizes in majority decisions.

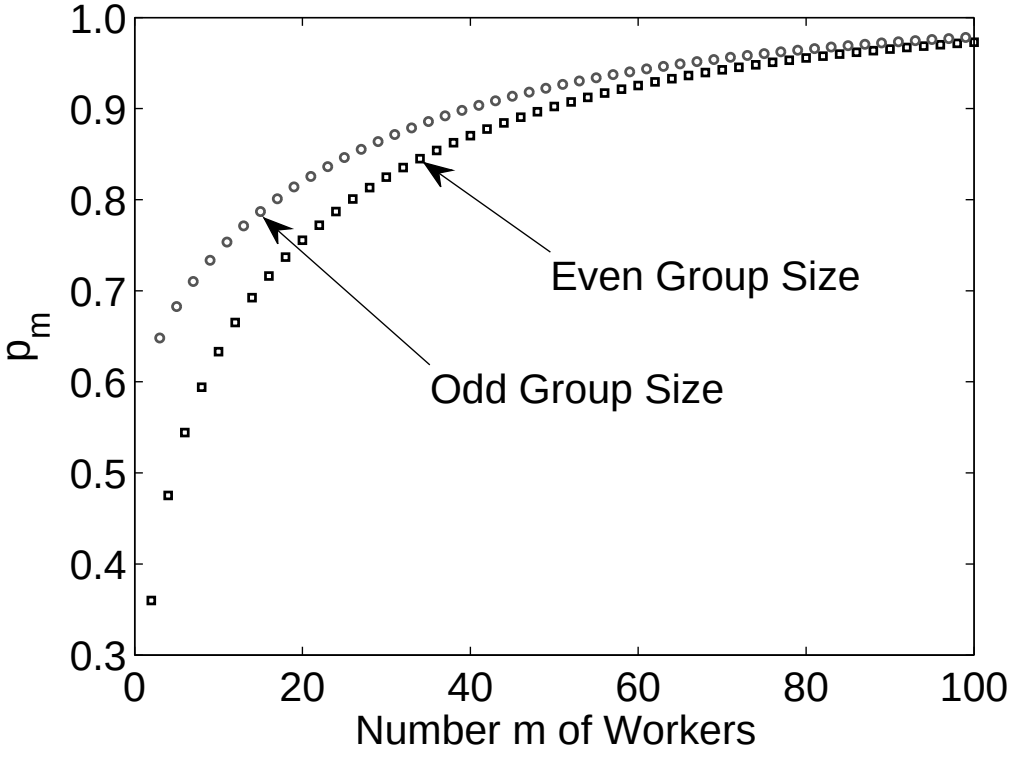

Figure 4: Probability of a correct majority decision p_m dependent on the number m of workers involved ($p_w = 0.4$)

4.2. Quality Comparison of MD and CG Approach

One of the main questions of this work is, whether the MD or the CG approach gives better results in terms of cheat-detection quality. In order to compare both approaches, we use the same number of workers m for the MD approach and for the control group of the CG approach, i.e. $i = j = m$.

4.2.1. MD Approach

Having a look at the MD approach only two results have to be evaluated, a correct and an incorrect decision. The probability for a correct result using the MD approach p_{MD} is the same as the one given in Equation 1. Therefore,

$$p_{\text{MD}} = p_m. \tag{2}$$

Consequently, the probability for an incorrect MD decision $\overline{p_{\text{MD}}}$ is given by

$$\overline{p_{\text{MD}}} = 1 - p_m. \tag{3}$$

4.2.2. CG Approach

The CG approach is more complicated. Here we have to differentiate as both the main worker and the control crowd may try to cheat. This results in four possible cases. In order to describe these cases, we introduce the following notation:

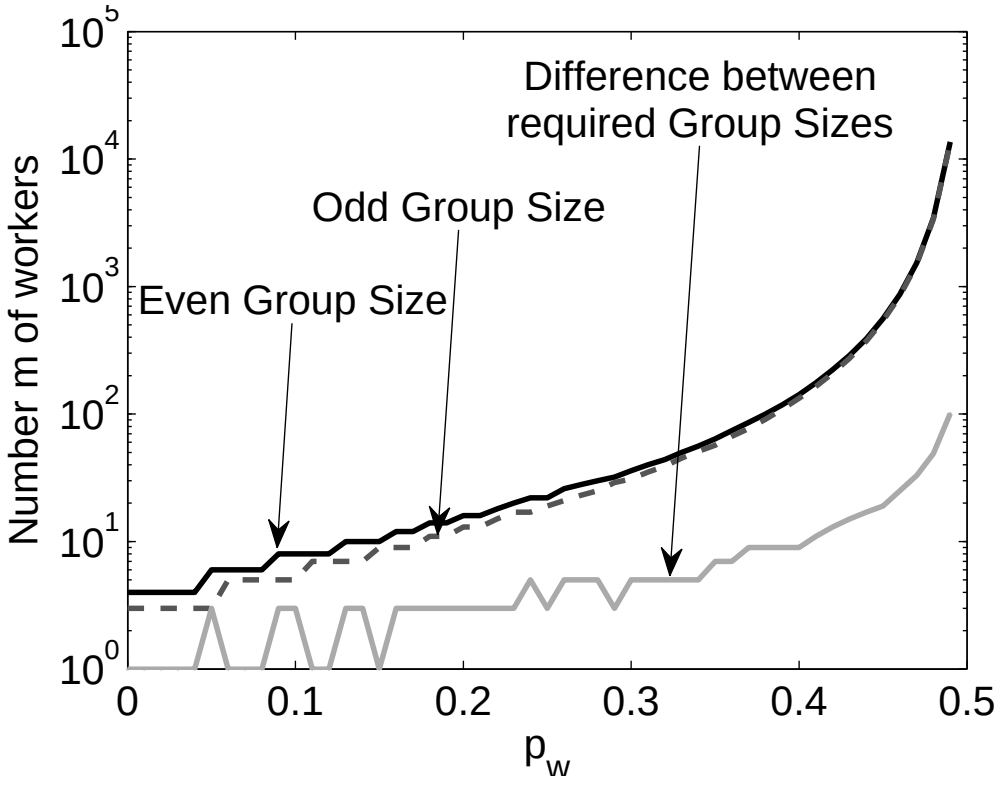

Figure 5: Required number m of workers for correct majority decision (99% Quantile)

p_{ab} with:

$$a = \begin{cases} r, & \text{worker submits right result} \\ w, & \text{worker submits a wrong result} \end{cases}$$

$$b = \begin{cases} d, & \text{control group makes a right decision} \\ \bar{d}, & \text{control group makes a wrong decision} \end{cases}$$

Using the notation, the possible results of the CG approach are:

(i) Worker submits a wrong result and the control group decides the result is invalid:

$$p_{wd} = p_w \cdot p_m$$

(ii) Worker submits a wrong result and the control group decides the result is valid:

$$p_{w\bar{d}} = p_w \cdot (1 - p_m)$$

(iii) Worker submits a right result and the control group decides the result is valid:

$$p_{rd} = (1 - p_w) \cdot p_m$$

(iv) Worker submits a right result and the control group decides the result is invalid:

$$p_{r\bar{d}} = (1 - p_w) \cdot (1 - p_m)$$

The probability for a correct decision using the control crowd approach p_{CG} is hence given by

$$p_{\text{CG}} = p_{wd} + p_{rd} = p_m, \tag{4}$$

and the probability for a wrong decision using the control crowd approach $\overline{p_{\text{CG}}}$ by

$$\overline{p_{\text{CG}}} = p_{w\bar{d}} + p_{r\bar{d}} = 1 - p_m. \tag{5}$$

Comparing Equation 2 and Equation 4, as well as Equation 3 and Equation 5, we can see that the both the MD and the CG approach offer the same quality of cheat-detection quality:

$$p_{\text{MD}} = p_{\text{CG}} = p_m \tag{6}$$

However, they differer among there applicability for different crowdsourcing tasks and their costs, as shown in the next section.

Please cite the updated version of this work:
M. Hirth, T. Hossfeld, P. Tran-Gia.
"Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms."
Mathematical and Computer Modelling, 2012. DOI: 10.1016/j.mcm.2012.01.006

5. Application of a Cost Model for different Use Cases

Before we give use cases for each of the control approaches, we specify a cost model. As the presented techniques are intended to be used in real crowdsourcing applications, the economic aspect is not negligible and has to be considered carefully.

5.1. Cost Model

When using crowdsourcing, workers are usually not paid as they submit a result, but only if the result is valid. Therefore, we define costs only for a successfully completed task. We denote these costs as c_1. In our CG approach we additionally use a second type of task for the control group. For this task, we assume different costs $c_2 \leq c_1$. Approving an invalid task does not only waste money, but might also have further negative impacts, like encouraging workers to continue cheating in one's campaigns or reputation loss. In order to account for these negative effects, we introduce additional costs c_{fp} for a "false-positive approval", if an invalid task is not detected. Also not paying for correct work has negative influences, as workers tend to stop working for this employer. Hence, we use a penalty c_{fn} for a "false-negative approval", if a correct task is assumed to be invalid.

We can now calculate the expected cost for both approaches. We use $i = j = m$ workers for our comparison, thus, the MD and CG approach have the same probability p_m to give a correct result. This analysis helps employers to decide, which of the approaches is cheaper for their current use case. But a first, we discuss the cost caused by the manual control by the employer.

5.1.1. Manual Control

If no automatized or crowd-based approach is used, the employer has to recheck the submitted tasks himself. In this case, there are no false-positive approval or false-negative approval, as the employer works correctly and knows all valid results. However, crowdsourcing tasks are generally low payed, whereas the time the employer has to spend on rechecking the task is normally quite expansive. In reality this results in two possible cases. Either a lot of time and money is spend on the controlling of the submitted tasks, or the employer approves all tasks without reviewing them. Both of them are not desirable, as the first case does not use the full cost saving potential of crowdsourcing, and the second case encourages workers to continue cheating and to submit bad quality work.

5.1.2. MD approach

There is no validation of the individual results in the MD approach, hence every worker is paid c_1 and a false-negative approval can not occur. If the majority decision is wrong, the costs are increased by c_{fp}. This results in the total expected costs c_{MD} of this approach,

$$c_{\text{MD}} = c_1 \cdot m + p_w \cdot c_{fp}. \tag{7}$$

5.1.3. CG approach

In the CG approach we assume one worker working on the main task, which costs c_1 if successfully completed, and m workers controlling the main task. Each of the m workers

is paid c_2 as these tasks are not validated. The cost of this approach varies, whether the worker on the main task is cheating or not, and whether the control crowd judges the result of the main task correctly. In the following we use the same notation as in Section 4.2.2. To calculate the total expected cost c_{CG} of this approach we have to consider the costs of four cases.

(i) Worker submits a wrong result and the control group decides the result is invalid:

$$c_{wd} = m \cdot c_2$$

(ii) Worker submits a wrong result and the control group decides the result is valid:

$$c_{w\bar{d}} = c_1 + m \cdot c_2 + c_{fp}$$

(iii) Worker submits a right result and the control group decides the result is valid:

$$c_{rd} = c_1 + m \cdot c_2$$

(iv) Worker submits a right result and the control group decides the result is invalid:

$$c_{r\bar{d}} = m \cdot c_2 + c_{fn}$$

Therefore, c_{CG} is given by

$$c_{\text{CG}} = c_{wd} \cdot p_{wd} + c_{w\bar{d}} \cdot p_{w\bar{d}} + c_{rd} \cdot p_{rd} + c_{r\bar{d}} \cdot p_{r\bar{d}}. \tag{8}$$

Using our cost model, we will now have a closer look at typical use cases of crowdsourcing. For this comparison we use five workers for the MD approach and the same amount of workers for the control group of the CG approach, i.e. six in total, to achieve the same probability for a right decision. Afterwards we give guidelines how to optimize the quality-cost ratio of the CG approach.

5.2. Routine Task

As mentioned before, routine tasks are a typical use case for crowdsourcing. These tasks are usually low-paid. In this example, we assume the costs per task $c_1 = 1$, which is the lowest possible payment in our considered crowdsourcing platform. The task of rechecking the main task should not be higher paid, therefore, we pay $c_2 = 1$ for each control worker. The costs caused by a cheating worker are very low in this case. Usually it does not matter if one of the workers did not fulfill the task, but as he might be encouraged to continue cheating on this kind of task we impose a penalty for each approval of an invalid task of $c_{fp} = 1$. Refusing to pay a worker who completed his task, will stop him from working for this employer again. But as the crowd contains many worker who can complete this task, the costs c_{fn} are set to $c_{fn} = 1$. The resulting costs depending on p_w are displayed in Figure 6.

The costs of the MD approach are shown by the continuous line, the costs for the CG approach by the dashed line. The costs of the majority decision follow a linear growth as they are based on the fixed costs of the majority decision with an additional penalty for a

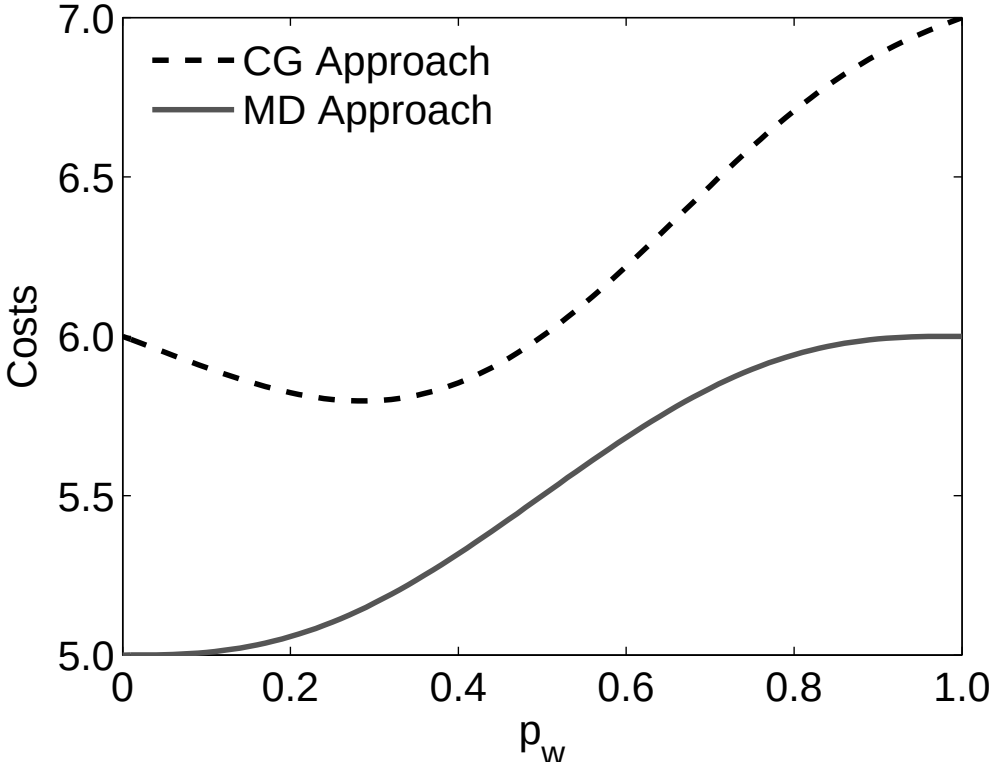

Figure 6: Costs of an unqualified task dependent on p_w

false positive approval, which becomes more likely as p_w increases. The development of the costs of the control crowd is more complex. At $p_w = 0$, the main worker surely earns $c_1 = 1$ as he is not cheating and the control group earns $5 \cdot c_2 = 5$, which results in the total costs $c_{\text{CG}} = 6$. With increasing p_w, the main worker is more likely to cheat, whereby the expected total costs decrease, as the control group detects this and the worker will not be paid. With p_w increasing further, the worker is even more likely to cheat, however p_w also influences the quality of the control group. The group makes more wrong decisions with a higher probability leading to increasing costs, as the wrong decisions are penalized by c_{fp} the same way as when using the MD approach. In the case of an unqualified task, the costs for the CG approach are always higher than the costs for the MD approach, since in general $c_2 = c_1$ and $c_{pf} \approx c_1$. Thus the MD approach should be preferred for routine tasks.

5.3. Complex and Creative Task

In complex and creative tasks wrong results usually have a large impact on the employer. One example for that kind of task is an advertisement campaign in forums. An employer wants to promote a product in web forums which deal with topics related to the product. But as direct advertisement is not desired in forum posts, the advertisement has to be hidden in a normal post using a recommendation which fits in the context of a forum thread. The worker has to find an appropriate forum thread and write an individual text, therefore, we assume $c_1 = 5$ in this case. As proof the worker has to give a link to the thread where he posted the advertisement. The control group only has to check the given thread, whether the post is related to the topic and includes the desired recommendation.

Hence, we assume a payment $c_2 = 1$ for each member of the control group. In order to study the influence of c_2 we also calculate the costs for $c_1 = c_2 = 5$. If the hidden advertisement campaign is recognized by the forum administrators, they are very likely to delete the post or a negative discussion about the employer will arise. Therefore, the penalty for approving wrong posts is set very high to $c_{fp} = 20$. Besides this, qualified workers for the main task are rather rare and losing one of them because of a not paid work is not desirable. Because of this, we set the penalty for not approving a correct task to $c_{fn} = 5$. The resulting costs are depicted in Figure 7 in the same manner as in Figure 6.

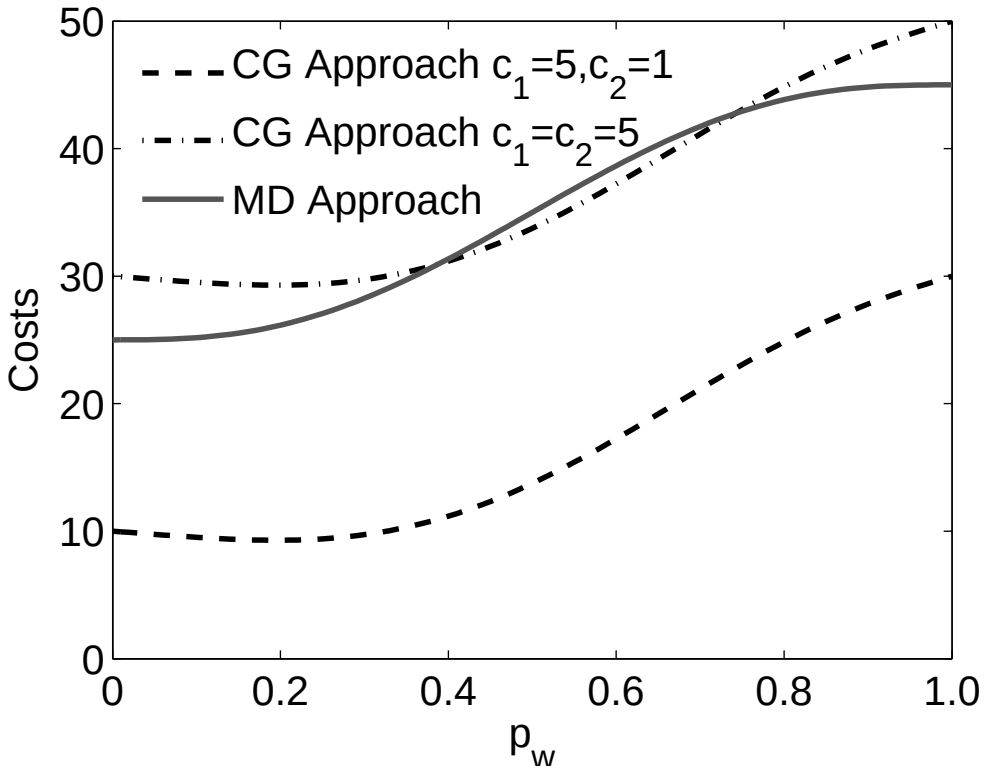

Figure 7: Costs of a qualified task dependent on p_w

The costs curves of the MD and the CG approach show a similar shape than for the unqualified task. However, in this case the CG approach with $c_1 = 5$ and $c_2 = 1$ is always cheaper than the MD approach, because of the high penalty for false positive approvals and the low costs for the control task $c_2 < c_1$. If the costs for the control task are raised $c_2 = c_1 = 5$, the CG approach performs better than MD for $p_w \in [0.20, 0.58]$ and otherwise MD outperforms CG.

We can derive two guidelines for complex and creative tasks. If cost ratio $\frac{c_2}{c_1} \ll 1$, which is the case for most complex and creative tasks, the CG approach should be favored. Otherwise the cost optimal cheat-detection mechanism for a given p_w can be found applying our cost model.

5.4. Cost-Quality Optimization Guidelines for Complex and Creative Tasks

A cost-quality optimization for complex and creative tasks is important as they are expensive compared to routine tasks. For this kind of task, the CG approach outperforms the

MD approach in terms of costs in most cases. Hence, we will focus on the CG approach in the following.

5.4.1. Optimizing Overall Costs and Cheat-Detection Quality

In order to reduce the total costs c_{CG}, a smaller control crowd can be used. However, this negatively affects the quality of the cheat-detection, as the probability of a correct CG result p_{CG} decreases with the group size, see Figure 4. Though, in many cases a trade-off between c_{CG} and p_{CG} has to be made. For our evaluation we use the example mentioned above with $c_1 = 5, c_2 = 1, c_{fp} = 20$ and $c_{fn} = 5$. Figure 8 depicts c_{CG} depending on p_{CG} for different values of p_w. As c_{CG} remains almost constant for $p_{\text{CG}} < 0.5$, we focus only on $p_{\text{CG}} \geq 0.5$.

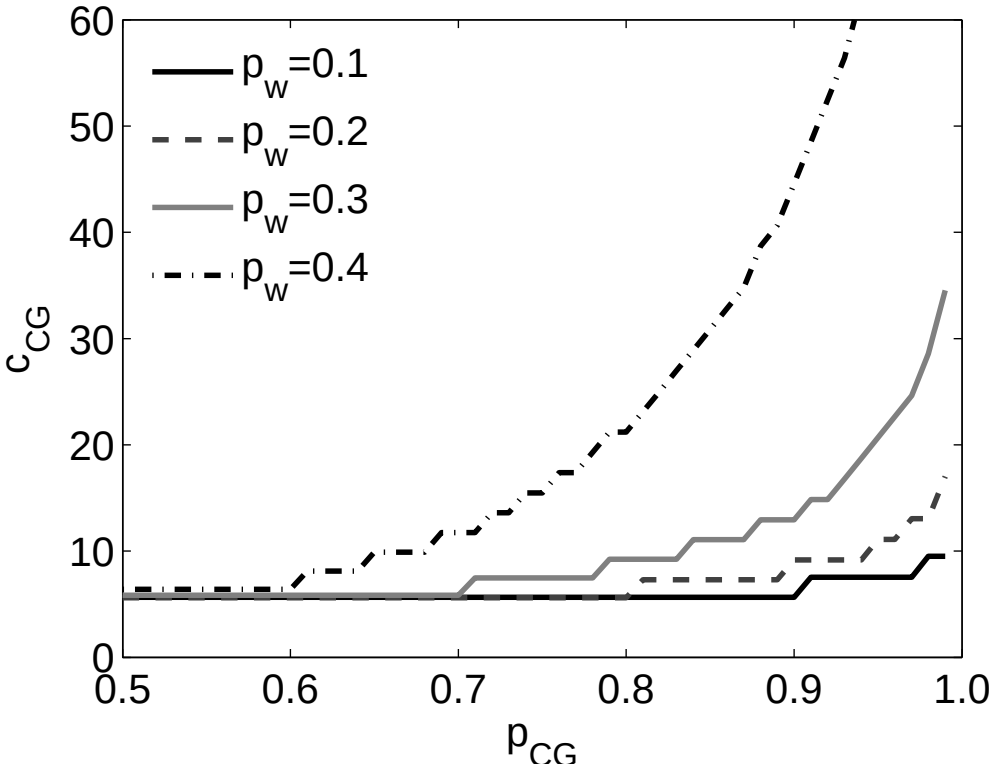

Figure 8: Total costs depending the cheat detection-quality

Figure 8 shows that c_{CG} increases with p_w and p_{CG}. More workers are needed for a better cheat-detection quality c_{CG} leading to higher costs. Also with an increase of p_w more workers are required to achieve a valid result. For a small value of p_w the influence of p_{CG} on the costs is only marginal and increasing the cheat-detection quality e.g. from 90% to 99% does not affect the costs very much. But for high values of p_w the costs increase tremendously with p_{CG}, which makes an detection improvement from 90% to 99% extremely expensive.

We can assume that an employer can approximately determine p_w based on the results of previous campaigns. Therefore, this model allows him make a trade-off between costs and result quality according to his needs.

5.4.2. Maximizing Available Salary for the Main Task

Creative and complex tasks require special skills. In order to attract skilled workers, these tasks have to be better paid than routine tasks. But the costs c_{CG} for a task using the CG approach are split between the skilled main task worker and the control group workers responsible for the cheat-detection quality. Therefore, we have to make a trade-off between the available money for the main worker and the cheat-detection quality.

By setting $c_1 = 0$ and assuming fixed values for $c_2 = 1, c_{fp} = 10, c_{fn} = 5$ and $p_w = 0.3$, we can use Equation 8 to calculate the overhead costs $c_{\text{CG}_{\text{overhead}}}$ dependent on the desired probability of a correct result p_{CG}. For a fixed budget c_{CG}, the maximal available salary c_1 can now be calculated by,

$$c_1 = c_{\text{CG}} - c_{\text{CG}_{\text{overhead}}}.$$

In the ideal case $c_1 = c_{\text{CG}}$, which means there are no other costs than the costs for the main task, but in reality $c_1 < c_{\text{CG}}$. Hence, we introduce the quotient $\frac{c_1}{c_{\text{CG}}}$ as a measure for the efficiency of the cost distribution. The smaller the value the more budget is spend on the control group, the higher the value the more salary is available for the main task. Figure 9 depicts $\frac{c_1}{c_{\text{CG}}}$ for different budgets c_{CG} and different cheat-detection qualities p_{CG}.

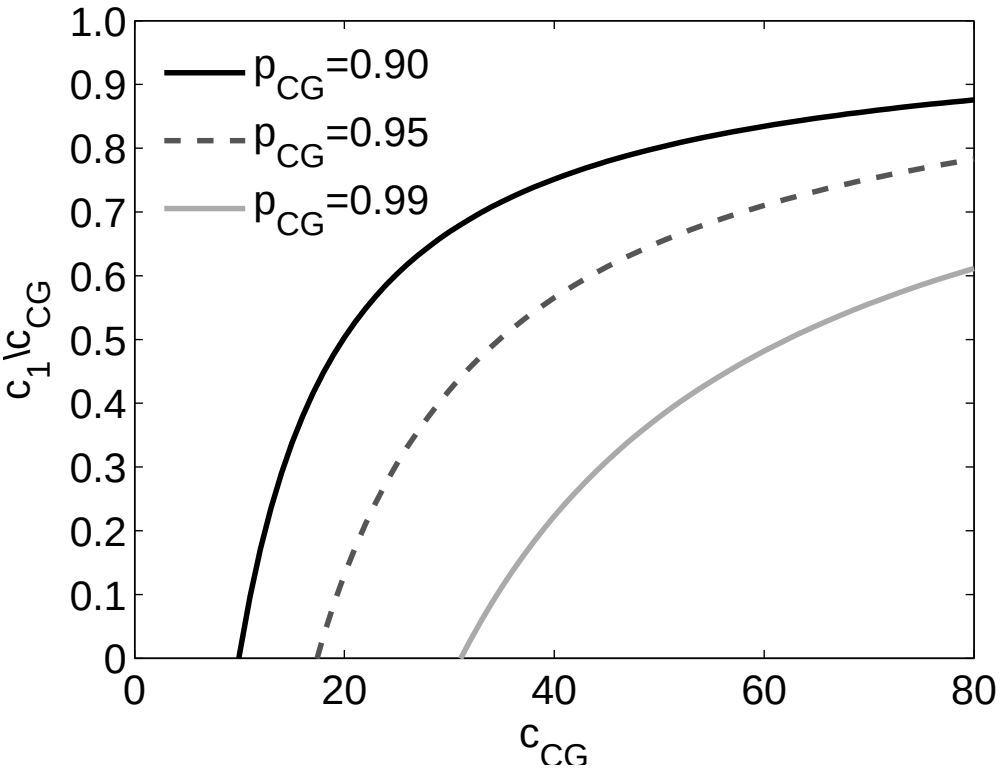

Figure 9: Efficiency of the cost distribution depending on the available budget

The intersection of the curves and the x-axis mark the minimum required task budget for the given p_{CG}. At this intersection point, no salary for the main task is available. With increasing budget c_{CG}, more salary for the main task is available as $c_{\text{CG}_{\text{overhead}}}$ remains constant. For large budgets, the main task salary is the biggest part of the total costs. The

intersections of the curves and the x-axis move to the right for higher p_{CG}, which shows that the higher the desired quality the more expensive the task. With higher p_{CG} also the efficiency of the cost distribution degrades quickly and disproportionate amount of the budget is spent on the control crowd instead of the main worker. Therefore, an employer has to consider carefully if a p_{CG} of 99% is realy needed instead of a p_{CG} of 95% or 90%.

Please cite the updated version of this work:
M. Hirth, T. Hossfeld, P. Tran-Gia.
"Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms."
Mathematical and Computer Modelling, 2012. DOI: 10.1016/j.mcm.2012.01.006

6. Conclusion

Crowdsourcing has only recently developed, but due to its various applications it is becoming an important new form of work organization. One of the major problems are untrustworthy workers trying to maximize their income by submitting as many tasks as possible even if they did not complete the task.

As manual re-checking of each task is not desirable, we proposed two different crowd-based methods, the MD and the CG approach, to verify task results. We have shown that, using the same amount of workers, both approaches offer the same significance level for detecting cheating workers. Furthermore, we have proven that it is generally better to use an odd group size instead of an even group size for the presented approaches. Using this finding helps to improve the quality and to reduce the costs of majority decisions.

All these results were used to derive a cost model for of our approaches. The costs of the MD and the CG approach were analyzed using typical types of crowdsourcing tasks. This analysis revealed that the MD approach is more suitable for low paid routine tasks, whereas the CG approach performs better for high prized tasks. In order to minimize the costs of high prized tasks, the CG approach was investigated in more detail. We showed that a slight reduce of the cheat-detection quality can significantly lower the cost for the whole task. Similarly, the overhead costs of the CG approach can be significantly decreased by slightly decreasing the cheat-detection quality.

Our approaches showed that crowd-based cheat-detection mechanisms are cheap, reliable, and easy to implement. They can be used to reduce the cost and the time consumption currently imposed by the manual validation process of task results.

Please cite the updated version of this work:
M. Hirth, T. Hossfeld, P. Tran-Gia.
"Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms."
Mathematical and Computer Modelling, 2012. DOI: 10.1016/j.mcm.2012.01.006

A. Appendix

A.1. Variables Summary

Variable Name	Meaning
	General Variables
p_c	Probability that a randomly chosen worker is a cheater
$p_{w\|c}$	Probability that a result submitted by a cheater is wrong
$p_{w\|\overline{c}}$	Probability that a result submitted by a non-cheating worker is wrong
p_w	Probability that a result is wrong
p_m	Probability of a correct majority decision
c_1	Costs of the main task
c_{fp}	Penalty for approving an invalid result
c_{fn}	Penalty for not approving a valid result
	MD related Variables
i	Number of workers used for the MD
p_{MD}	Probability of a correct result using the MD approach
$\overline{p_{\text{MD}}}$	Probability of a wrong result using the MD approach
c_{MD}	Expected total costs if using the MD approach
	CG related Variables
j	Number of workers used for the control group of the CG approach
p_{wd}	Probability that the main worker submits a wrong result and the control group decides the result is invalid
$p_{w\overline{d}}$	Probability that the main worker submits a wrong result and the control group decides the result is valid
p_{rd}	Probability that the main worker submits a right result and the control group decides the result is valid
$p_{r\overline{d}}$	Probability that the main worker submits a right result and the control group decides the result is invalid
p_{CG}	Probability of a correct result using the CG approach
$\overline{p_{\text{CG}}}$	Probability of a wrong result using the CG approach
c_2	Costs for control group validation task
c_{wd}	Costs if the main worker submits a wrong result and the control group decides the result is invalid
$c_{w\overline{d}}$	Costs if the main worker submits a wrong result and the control group decides the result is valid
c_{rd}	Costs if the main worker submits a right result and the control group decides the result is valid
$c_{r\overline{d}}$	Costs if the main worker submits a right result and the control group decides the result is invalid
c_{CG}	Expected total costs if using the CG approach

Table 1: Variables used for the crowd model, the approach evaluation, and the cost model

A.2. Proof for Section 4.1

Assume an even number $2n, n \in \mathbb{N}$ of workers for the majority decision. The maximal number of incorrect results, which still leads to a correct majority decision is $2n/2 - 1 =$

$n-1$. Therefore, the probability for a correct majority decision of an even group is

$$P_e(X \leq n-1) = \sum_{k=0}^{n-1} \binom{2n}{k} p_w^k (1-p_w)^{2n-k}. \tag{9}$$

Assuming an odd number $2n-1, n \in \mathbb{N}$ of workers for the majority decision, the maximal number of incorrect results still leading to a correct majority decision is $\lfloor (2n-1)/2 \rfloor = \lfloor n - 1/2 \rfloor = n-1$. Hence, the probability for a correct majority decision of an even group is

$$P_o(X \leq n-1) = \sum_{k=0}^{n-1} \binom{2n-1}{k} p_w^k (1-p_w)^{2n-1-k}. \tag{10}$$

In order to poof, that a small odd group size give better results than a slightly larger even group size, we show (10) > (9) for $n \geq 2$. If $n = 0$ or $n = 1$, a majority decision is not possible.

Proof using induction:
Basis:$n = 2$
The probability for a correct majority decision using an even number of 4 workers is

$$\begin{aligned} P_e(X \leq n-1) &= P_e(X \leq 1) \\ &= \sum_{k=0}^{n-1} \binom{2n}{k} p_w^k (1-p_w)^{2n-k} \\ &= \sum_{k=0}^{1} \binom{4}{k} p_w^k (1-p_w)^{4-k} \\ &= -3 \cdot {p_w}^4 + 8 \cdot {p_w}^3 - 6 \cdot {p_w}^2 + 1. \end{aligned}$$

The probability for a correct majority decision using an odd number of 3 workers is

$$\begin{aligned} P_o(X \leq n-1) &= P_o(X \leq 1) \\ &= \sum_{k=0}^{n-1} \binom{2n}{k} p_w^k (1-p_w)^{2n-k} \\ &= \sum_{k=0}^{1} \binom{3}{k} p_w^k (1-p_w)^{3-k} \\ &= 2 \cdot {p_w}^3 - 3 \cdot {p_w}^2 + 1. \end{aligned}$$

The difference between $P_o(X \leq 1)$ and $P_e(X \leq 1)$ is

$$\begin{aligned} P_o(X \leq 1) - P_e(X \leq 1) &= (2 \cdot {p_w}^3 - 3 \cdot {p_w}^2 + 1) \\ &\quad - (-3 \cdot {p_w}^4 + 8 \cdot {p_w}^3 - 6 \cdot {p_w}^2 + 1) \\ &= 3 \cdot p_w^4 - 6 \cdot p_w^3 + 3 \cdot p_w^2 \\ &= 3 \cdot p_w^2 (p_w^2 - 2 \cdot p_w + 1) \\ &= 3 \cdot p_w^2 (p_w - 1)^2 \geq 0, \text{ as } 0 \leq p_w \leq 1. \end{aligned}$$

Please cite the updated version of this work:
M. Hirth, T. Hossfeld, P. Tran-Gia.
"Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms."
Mathematical and Computer Modelling, 2012. DOI: 10.1016/j.mcm.2012.01.006

Induction hypothesis (IV):for n:

$$P_o(X \leq n-1) > P_e(X \leq n-1)$$

$$\sum_{k=0}^{n-1} \binom{2n-1}{k} p_w^k (1-p_w)^{2n-1-k} > \sum_{k=0}^{n-1} \binom{2n}{k} p_w^k (1-p_w)^{2n-k}$$

Induction step:$n \to n+1$:

$$P_o(X \leq (n+1)-1) = \sum_{k=0}^{(n+1)-1} \binom{2(n+1)-1}{k} p_w^k (1-p_w)^{2(n+1)-1-k}$$

Substitute: $m = n+1$

$$\sum_{k=0}^{m-1} \binom{2m-1}{k} p_w^k (1-p_w)^{2m-1-k} >^{\text{IV}} \sum_{k=0}^{m-1} \binom{2m}{k} p_w^k (1-p_w)^{2m-k}$$
$$=^{\text{resub}} P_e(X \leq (n+1)-1) \text{q.e.d}$$

List of Figures

List of Tables

References

[1] Digg. www.digg.com.

[2] Innocentive. www.innocentive.com.

[3] Omar Alonso, Daniel E. Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[4] Amazon.com, Inc. Mechanical turk. www.mturk.com.

[5] K.T. Chen, C.J. Chang, C.C. Wu, Y.C. Chang, C.L. Lei, and C. Sinica. Quadrant of euphoria: A crowdsourcing platform for QoE assessment. *IEEE Network*, 24(2):28–35, 2010.

[6] Google Inc. Google image labeler. www.images.google.com/imagelabeler

[7] Jeff Howe. The rise of crowdsourcing. Online, June 2006.

[8] P.Y. Hsueh, P. Melville, and V. Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, 27–35, 2009.

[9] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 453–456, 2008.

[10] OpenStreetMap Foundation. Openstreetmap. www.openstreetmap.org.

[11] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 319–326, 2004.

[12] Weblabcenter, Inc. Microworkers. www.microworkers.com.

[13] Wikimedia Foundation. Wikipedia. www.wikipedia.org.

[14] Yahoo!, Inc. www.delicious.com.

Please cite the updated version of this work:
M. Hirth, T. Hossfeld, P. Tran-Gia.
"Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms."
Mathematical and Computer Modelling, 2012. DOI: 10.1016/j.mcm.2012.01.006