ARTICLE IN PRESS

Mathematical and Computer Modelling ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Mathematical and Computer Modelling

journal homepage: www.elsevier.com/locate/mcm

Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms

Matthias Hirth *, Tobias Hoßfeld, Phuoc Tran-Gia

University of Würzburg, Institute of Computer Science, Germany

ARTICLE INFO

Keywords:
Crowdsourcing
Quality assurance
Cost estimation

ABSTRACT

Crowdsourcing is becoming more and more important for commercial purposes. With the growth of crowdsourcing platforms like Amazon Mechanical Turk or Microworkers, 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. In this paper, we analyze two widely used crowd-based approaches to validate the submitted work.[1] Both approaches are evaluated with regard to their detection quality, their costs and their applicability to different types of typical crowdsourcing tasks.

1. Introduction

With the tremendous growth of the Internet's user base, a huge workforce with a large amount of knowledge has been developed. This is already utilized in projects like the Wikipedia, where users created an encyclopedia by sharing their knowledge, or OpenStreetMap which offers 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. Jeff Howe defined crowdsourcing as "…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" [1]. The main differences with outsourcing are that, the entrepreneur does not know who accomplishes a task and that the workers do not form an organized group, but are members of a large anonymous crowd. 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 Amazon Mechanical Turk (MTurk) or Microworkers, crowdsourcing became also interesting for commercial usage. These platforms are specialized on very granular work [2] and offer an easy access to a huge amount of workers. Using commercial crowdsourcing, work can be done very quickly by accessing a large and relatively cheap workforce, but the results are not reliable. Some workers submit incorrect results in order to maximize their income by completing as many jobs as possible, others just do not work correctly. In this paper, we denote all of them as cheaters, as the reason for submitting invalid work is irrelevant for our analysis. Sometimes a small amount of incorrect results can be tolerated, but not in general. Therefore, techniques have to be developed to detect cheating workers and invalid work results.

This paper is an extended version of [3]. It presents two approaches to detect cheating workers by using crowdsourcing workers for the validation. Both approaches can be easily integrated in current crowdsourcing platforms. In the following,

* Corresponding author. Tel.: +49 93131 86954; fax: +49 931 31 86632.
E-mail addresses: matthias.hirth@informatik.uni-wuerzburg.de (M. Hirth), tobias.hossfeld@informatik.uni-wuerzburg.de (T. Hoßfeld), phuoc.trangia@informatik.uni-wuerzburg.de (P. Tran-Gia).

[1] This paper is an extended version of Hirth et al. (2011) [3].

0895-7177/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.mcm.2012.01.006

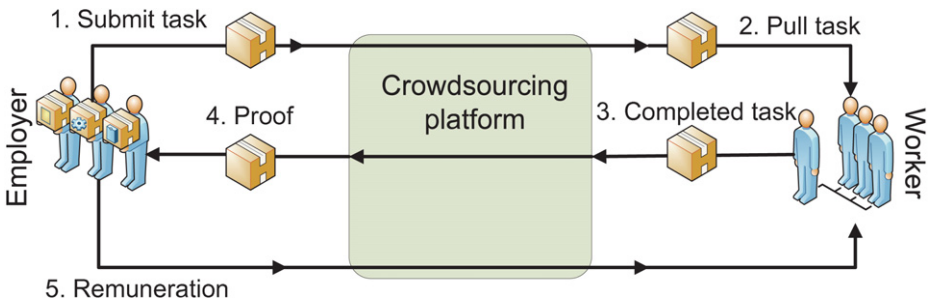

Fig. 1. The crowdsourcing scheme.

we evaluate the quality of our cheat-detection solutions, discuss their costs, and demonstrate their applicability to common crowdsourcing tasks. General guidelines are given on how our findings can be used for real crowdsourcing tasks. In addition to [3], this paper (a) introduces a new cost model, (b) gives a more sophisticated evaluation of the model and (c) shows the influence of reliable workers on the overall costs.

The paper is structured as follows. Section 2 gives a quick overview of the concept of crowdsourcing, current trends and research already done in this area. In Section 3, we present two approaches for work validation, which are evaluated in Section 4. A cost model for both approaches is developed in Section 5, and applied to different real world crowdsourcing use cases in Section 6. The paper is concluded in Section 7.

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

Every employer needs a mediator to access the worker crowd. This mediator is called a *crowdsourcing platform* which is schematically depicted in Fig. 1. Well known examples of these platforms are e.g. MTurk [4,5], or Microworkers [6].

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 provide *proof* of a completed task. 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*, *complex*, and *creative tasks*. Routine tasks are jobs that do not require any level of qualification, like bookmarking a web page using social bookmarking services such as digg, relevance evaluation [7], or creating a new YouTube account. *Complex tasks*, like text annotation [8] or rewriting a given text, need some general skills, in contrast to *creative tasks* where highly specialized skills are required. *Creative tasks* include, e.g. writing an article on a given topic or even research and development [9].

Detecting cheating workers is more difficult for complex tasks than for routine tasks. Assume a routine task, where a worker has to create a new YouTube account. The worker has to submit the login data in order to prove that the task is completed. It is easy to check automatically whether the login data is valid or not. This is exemplary for routine tasks, where verification is often simple and easy to automatize. This is different for complex or creative tasks. Assume a complex task, where a worker has 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 style. This cannot 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. Development of crowdsourcing

In recent years, a lot of applications were developed which use the crowdsourcing paradigm but differ from the aforementioned commercial web based crowdsourcing platforms. Thus, we want to give a short overview of some current trends: crowd-sourced sensing, "real-world crowdsourcing", and enterprise crowdsourcing. However, developing cheat detection and quality assurance mechanisms for these types of crowdsourcing is out of the scope of this paper due to the different task and crowd structure compared to commercial crowdsourcing platforms.

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

Crowd-sourced sensing referrers to the concept of replacing fixed installed sensor nodes by a mobile crowd. Depending on the sensing task, no equipment, smartphones or special equipment can be required. For example Demirbas et al. [10] proposed a crowdsourcing based sensing system which uses Twitter to exchange information. Using this system they performed weather monitoring by asking the contributors to indicate the current weather by returning predefined values via Twitter. Furthermore, they were able to create a noise map using GPS enabled smart-phones as sensor nodes. An example of a crowd-sensing application using specialized equipment is geigercrowd.net[2] which was launched after the nuclear disaster of 2011 in Fukushima. On this web page voluntaries submit radiation measurements from all over Japan in order to create an open data and up-to-date radiation map.

Crowd-sourced sensing shows that the crowdsourcing paradigm is not only applicable to task in the online world, however crowd-sourced sensing is still limited to tasks for data retrieval. Real-world crowdsourcing goes even a step further. Platforms like taskrabbit.com,[3] airrun.com (see footnote 3) and gigwalk.com (see footnote 3) offer means to crowd-source almost ever job from shopping to wall painting.

Another variation of real-world crowdsourcing is the concept of enterprise crowdsourcing as e.g. proposed by Tim Ringo, head of IBM Human Capital Management.[4] Unlike to the original concept of crowdsourcing, the work is not done by a huge anonymous crowd, but by a crowd of company employees or employees of sub-conductors. Still the work is submitted to a pool of workers instead to a designated one, but using a verified crowd even confidential tasks can be crowd-sourced.

2.4. Related work

Commercial crowdsourcing applications suffer from workers, who try to submit invalid or low quality work in order to maximize their received payment while reducing their own effort. This is the case even if the expected gain is very little [11]. Thus, numerous efforts have been made in order to improve the qualify of the results submitted by the workers and to detect cheating workers. The most easy way to test the trustworthiness and quality of a worker is to add gold standard data [12] whereof the correct task result is already known. Gold standard data can increase the quality of the task results as the worker received an immediate feedback about mistakes and as continuously cheating workers are easy to identify. In some cases gold standard data can be generated automatically [13] or even the bias of the workers can be taken into account [14].

Gold standard data is not applicable for tasks where there is no clear *correct* result, like a subjective rating. Here verification questions can be used to estimate the reliability of a worker. In [15], Kittur et al. used crowdsourcing workers to rate the quality of Wikipedia articles. The correlation between the rating obtained form crowdsourcing and a trusted reference group could be significantly improved by adding questions which test if the worker read the article. Hoßfeld et al. [16] also used verification questions, but also added task specific user monitoring to assure reliable results for a user survey, which was uses as basis for a QoE model for online video services. Hoßfeld et al. showed that using this task specific user monitoring a lot for cheating worker could be detected.

The importance of the task design is also shown in [17] and some rare task can even be designed in an almost cheat save way. For example Chen et al. [18] 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 [19], Ahn and Dabbish present a crowd-based image labeling game, which is now used in an adapted version by Google's Image Labeler. A label is added to the picture, if at least two randomly picked users suggest the same label. Ahn and Dabbish argue that cheating is not possible due to the huge number of players. Two random players are very unlikely to know each other and, hence, are not able to collaborate.

Besides the task design, the task type can influence trustworthiness of the workers. Eickhoff and Vries [20] observed that depending on the type of task more or less malicious workers are encountered and suggested to derive the quality of a worker not only from the number of completed tasks but also their type, i.e. does the worker only perform simple tasks or mainly complex ones.

Further, the complete workflow of a crowdsourcing project can be optimized in order to detect cheaters and to improve the quality. Dow et al. [21] suggest to integrate an interactive feedback systems to encourage workers and other suggest to use multiple iterative tasks [22,23] or coordination techniques [24] to improve the quality of the results.

The contribution of this work is the analysis of two generic crowd-based approaches for tasks where gold standard data is not applicable and manual re-checking by the employer is ineffective. Due to the evaluation of the approaches using a generic cost model, guidelines about the applicability of the approaches can be given.

3. Crowd based cheat detection mechanisms

We analyze two crowd-based cheat-detection approaches: 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 variables used in this and the following sections are summarized in Table 1 which can be found in the Appendix.

2 Last accessed 26.08.2011.

3 Last accessed 26.08.2011.

4 http://www.personneltoday.com/articles/2010/04/23/55343/ibm-crowd-sourcing-could-see-employed-workforce-shrink-by-three-quarters.html (last accessed 26.08.2011).

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

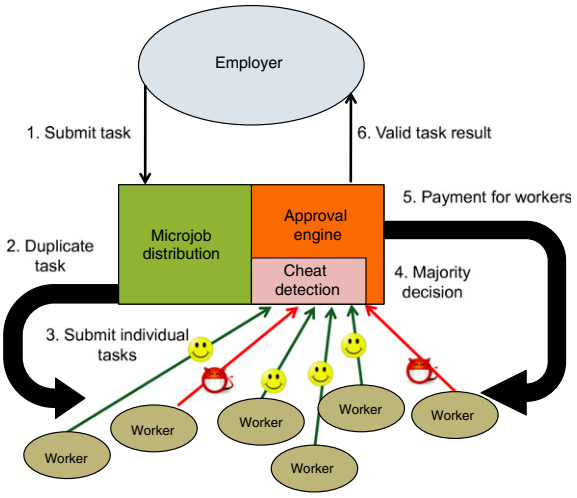

Fig. 2. *Majority decision* (MD) approach scheme.

3.1. General notion and variables

In our model, the crowd consists of N individual workers. Not every worker is honest and performs the task correctly, but we can assume that these workers do not intend to falsify the task result. They only give a random result in order to complete the task as fast as possible. 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 $p_{w|c}$. It is not possible to decide whether a worker submitted a wrong result deliberately or accidentally and the worker is treated as a cheater in both cases. Thus, in our model only cheaters submit incorrect results, i.e. $p_{w|\bar{c}} = 0$. Honest 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 five 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 $p_{w|c} = 80\%$. This results in a probability for a wrong answer $p_w = 8\%$.

3.2. The majority decision approach

The first approach (MD) uses a majority decision to eliminate incorrect results and is illustrated in Fig. 2. The employer submits his task to the crowdsourcing platform (Step 1 in Fig. 2). In order to clarify the individual steps of the approach, we have divided the platform in functional components, the job distribution component, the cheat detection component, and the task approval engine. The platforms task distribution component duplicates the task (2) and N_{md} different workers complete the tasks. They submit their individual results (3) which might be correct or incorrect. The crowdsourcing platform performs a majority decision (4) in the cheat detection engine, i.e., the result most of the workers submitted is assumed to be correct and returned to the employer (5). In this approach each worker who voted according to the majority is paid.

As an example application of the MD approach, think of 100 workers searching for relevant web pages to a given topic. If a one web page is submitted by 92 workers, it is certainly relevant for the given topic. Even if some workers are cheating, the overall result is valid.

3.3. The control group approach

Our second approach (CG) is based on a control group and is schematically depicted in Fig. 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). Now, the crowdsourcing platform generates new validation tasks for this result. The result of the main task is given to a group of N_{cg} other 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 overall rating of the main task (6). The main task is considered to be valid, if the majority of the control group decides the task is correctly done. This is necessary, because some 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). Otherwise, the task is repeated by another workers until the first result is rated valid by the control crowd. 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, while the control task is cheaper.

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

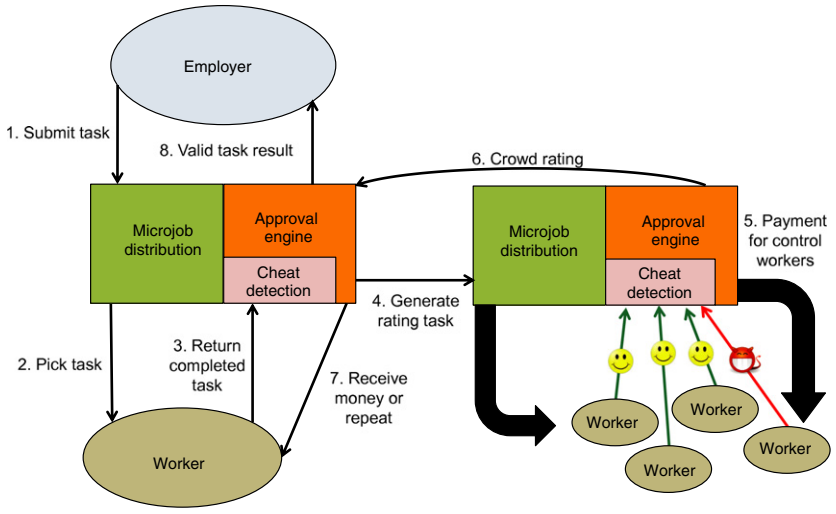

Fig. 3. *Control group* (CG) approach scheme.

An application of this approach is a task where a worker has to find a relevant article for a given topic. Depending on the topic, the article search can be time consuming and difficult. The submitted article is given to 100 low paid workers who have to judge whether the article is relevant for the topic or not. If enough workers agree that the submitted article is relevant, the searching workers is paid.

4. Evaluation of the MD and CG approaches

Both approaches use a majority decision of N_m workers in order to verify the task result. Thus, we have a closer look at how to optimize N_m, in order to minimize the costs and maximize the reliability of the results. Afterward, we evaluate the quality of the cheat-detection of both approaches.

4.1. Group size for majority decisions

For a majority decision we use N_m random workers from the total crowd of N workers. Each of their N_m results is with a probability of p_w incorrect and thus, the number of incorrect results X follows a binomial distribution $X \sim \text{BINOM}(N_m, p_w)$. To derive a correct majority decision, the number of incorrect results has to be smaller than $N_m/2$, i.e., the probability of a correct majority decision p_m is given by

$$p_m = P\left(X < \frac{N_m}{2}\right) = \sum_{k=0}^{\left\lfloor \frac{N_m-1}{2} \right\rfloor} \binom{N_m}{k} p_w^k (1-p_w)^{N_m-k}. \tag{1}$$

In order to avoid a draw while performing a majority vote, we only use odd group sizes in the following evaluations.

4.2. Quality comparison of the MD and CG approaches

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. $N_{md} = N_{cg} = N_m$.

4.2.1. The MD approach

Having a look at the MD approach, we have to evaluate whether the group made a correct or an incorrect decision. The probability for a correct result using the MD approach p_{MD} is the same as the one given in Eq. (1). Thus,

$$p_{md} = p_m. \tag{2}$$

The probability for an incorrect MD result $\overline{p_{md}}$ is given by

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

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

4.2.2. The CG approach

The CG approach is more complicated. We have to differentiate as the main worker and the control group may try to cheat. This results in four possible cases.

Correct control group decision	(1) Correct task approved (CA)	(2) Wrong task disapproved ($\overline{CA}$)
Incorrect control group decision	(3) Correct task disapproved ($C\overline{A}$)	(4) Wrong task approved ($\overline{C}A$)
	Correct main worker result	Incorrect main worker result

We assume that our crowd is very large, thus the main worker and the workers from the control group do not know each other. Further, the main task and the control tasks are very different tasks, as the main task is a complex one and the control tasks are rather simple. Hence, the cheating probabilities of the main worker and the control group are independent. Thus, the possible results of the CG approach are:

Correct control group decision	(1) $P(CA) = p_{CA} = (1 - p_w) \cdot p_m$	(2) $P(\overline{CA}) = p_{\overline{CA}} = p_w \cdot p_m$
Incorrect control group decision	(3)$P(C\overline{A}) = p_{C\overline{A}} = (1 - p_w) \cdot (1 - p_m)$	(4) $P(\overline{C}A) = p_{\overline{C}A} = p_w \cdot (1 - p_m)$
	Correct main worker result	Incorrect main worker result

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

$$p_{cg} = p_{CA} + p_{\overline{CA}} = p_m, \tag{4}$$

and the probability for an incorrect result using the CG approach $\overline{p_{\mathrm{CG}}}$ by

$$\overline{p_{cg}} = p_{C\overline{A}} + p_{\overline{C}A} = 1 - p_m. \tag{5}$$

Comparing p_{md} and p_{cg}, we can see that the both the MD and the CG approach offer the same quality of cheat-detection quality:

$$p_{md} = p_{cg} = p_m. \tag{6}$$

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

5. A cost model for the MD and CG approaches

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 important and has to be considered.

5.1. The cost model

Each worker who submits a correct task result is paid c_c, each worker submitting an incorrect result is paid c_w. Approving an invalid task does not only waste money, but has further negative impacts, like encouraging workers to continue cheating or reputation loss. To account for these negative effects, we introduce costs c_{fp} for a "false-positive approval", if an invalid task is not detected. Not paying for correct work has negative influences, too, as workers 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.

As mentioned above, the control task in the CG approach is usually easier than the main task and differently paid. Thus, we use here different costs for the control tasks, c_{cc}, c_{cw}, c_{cfp}, and c_{cfn} which we assume to be lower than their corresponding costs from the main task.

We now calculate the expected cost for both approaches. We use $N_{md} = N_{cg} = N_m$ workers, thus, the probability for a correct MD and CG approach result is p_m. This analysis helps employers to decide, which approach is cheaper for a certain use cases.

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

5.1.1. The MD approach

When performing a majority decision using N_{md} workers, we receive N_{mdc} correct results and $N_{md\bar{c}}$ wrong results from the workers, with

$$0 \leq N_{mdc} \leq N_m, \qquad 0 \leq N_{md\bar{c}} \leq N_m, \quad \text{and} \quad N_{md} = N_{mdc} + N_{md\bar{c}}.$$

In order to avoid a draw, we use always use odd group sizes.

If the majority decision is correct ($N_{md\bar{c}} < N_{md}/2$), the workers who submitted correct results are paid c_c, the workers who submitted wrong results are paid c_w. However, if the majority of the workers submits a wrong result ($N_{md\bar{c}} \geq N_{md}/2$), this result is assumed to be correct. Thus, the workers who submitted wrong results are paid c_c and each worker who submitted a correct result is paid c_w. In this case there are also additional costs for the false positive approval of the task and the rejection of the correct ones. This results in the conditional costs $C_{md,N_{md\bar{c}}}$ for $N_{md\bar{c}}$ wrong results, with

$$C_{\mathrm{MD},N_{md\bar{c}}} = \begin{cases} N_{md\bar{c}} \cdot c_w + N_{mdc} \cdot c_c, & N_{md\bar{c}} < N_{md}/2 \\ N_{md\bar{c}} \cdot (c_c + c_{fp}) + N_{mdc} \cdot (c_w + c_{fn}), & N_{md\bar{c}} \geq N_{md}/2. \end{cases}$$

Using the condition costs $C_{md,N_{md\bar{c}}}$, we can now calculate the expected costs $E[c_{md}](N_{md}, p_w)$ of the MD approach in dependency of the number of workers N_{md} involved and the probability p_w of a wrong task result. For sake of readability, we use c_{md} instead of $E[c_{md}]$.

$$\begin{aligned} c_{md} &= \sum_{i=0}^{N_{md}} C_{MD,N_{md\bar{c}}} \cdot P(N_{md\bar{c}} = i) \\ &= \sum_{i=0}^{\left\lfloor \frac{N_{md}-1}{2} \right\rfloor} C_{MD,N_{md\bar{c}}} \cdot P(N_{md\bar{c}} = i) + \sum_{i=\left\lceil \frac{N_{md}}{2} \right\rceil}^{N_{md}} C_{MD,N_{md\bar{c}}} \cdot P(N_{md\bar{c}} = i) \\ &= \sum_{i=0}^{\left\lfloor \frac{N_{md}-1}{2} \right\rfloor} (N_{md\bar{c}} \cdot c_w + N_{mdc} \cdot c_c) \cdot P(N_{md\bar{c}} = i) \\ &\quad + \sum_{i=\left\lceil \frac{N_{md}}{2} \right\rceil}^{N_{md}} \left(N_{md\bar{c}} \cdot (c_c + c_{fp}) + N_{mdc} \cdot (c_w + c_{fn})\right) \cdot P(N_{md\bar{c}} = i). \end{aligned} \tag{7}$$

5.1.2. The CG approach

In the CG approach one worker is working on the main task which costs c_c if the worker completed it successfully, otherwise the worker is paid c_w. The main task is controlled by N_{cg} workers. Each of the N_{cg} workers who submitted the same results the majority of the control crowd is paid c_{cc}, the rest of the workers c_{cw}. Similar to the MD approach there are penalties for approving wrong results and rejecting correct results.

The costs vary, depending on whether the worker of the main task is cheating or not, and whether the control crowd rates the result of the main task correctly. To calculate the total expected cost $E[c_{cg}]$ we have to consider four cases.

	Correct main worker result	Incorrect main worker result
Correct control group decision	(1) $c_{CA} = c_c + c_{md}$	(2) $c_{\overline{CA}} = c_w + c_{md}$
Incorrect control group decision	(3) $c_{C\overline{A}} = c_w + c_{md} + c_{fn}$	(4) $c_{\overline{C}A} = c_c + c_{md} + c_{fp}$

In each of the four cases, the control crowd is paid. As we use the same number of workers for the control crowd in the CG approach as for the majority decision in the MD approach ($N_m = N_{md} = N_{cg}$), the cost of the control crowd in the CG approach calculated using Eq. (7) and the costs c_{cc}, c_{cw}, c_{cfp}, and c_{cfn}. Now we can have a closer look at the varying costs. If the main worker did submit a correct result and the control crowd approves it, then there are only the additional costs c_c for the payment of the main worker. It is similar, if the main worker submitted a wrong result and the control crowd realizes this and disapproves the task. In that case the worker is paid c_w. If the control crowd falsely disapproves a correct main task, the main workers is paid c_w as his work is assumed to be incorrect and a penalty of c_{fn} is added. If the control approves an incorrect task, the main worker is paid c_c for his result and a penalty c_{fp} is added for the incorrect approval.

In the CG approach, the main task is repeated until the control crowd rates it to be correct. This happens, if a correct main task is approved (CA) or if a wrong task is approved ($\overline{C}A$). Thus, the probability $P(cg_{\text{approve}})$ of approving the main task is

$$P(cg_{\text{approve}}) = P(CA \cup \overline{CA}) = p_{CA} + p_{\overline{C}A} = p_m + p_w - 2 \cdot p_m p_w,$$

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

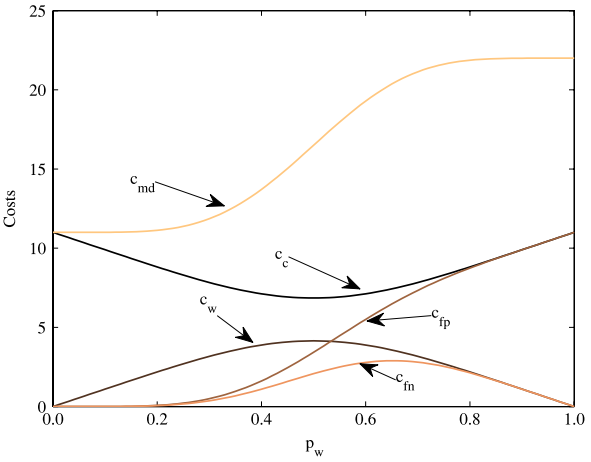

Fig. 4. Cost factors of the MD approach for $N_{md} = 11$.

and the number of repetitions R until the main task is approved follows a geometrical distribution

$$P(R = n) = P(cg_{\text{approve}}) \cdot (1 - P(cg_{\text{approve}}))^{(n-1)}.$$

The expected costs $E[c_{cg}]$ of the CG approach consist of $E[R] - 1$ times the cost $c_{\text{disapproval}}$ for a disapproving the main task and once the costs c_{approval} for approving the main task. Both the costs are again depending on whether the control crowd made a correct decision or not.

$$E[c_{\text{approval}}] = P(CA|cg_{\text{approve}}) \cdot c_{CA} + P(\overline{CA}|cg_{\text{approve}}) \cdot c_{\overline{CA}}$$
$$E[c_{\text{disapproval}}] = P(C\overline{A}|cg_{\text{disapprove}}) \cdot c_{C\overline{A}} + P(\overline{CA}|cg_{\text{disapprove}}) \cdot c_{\overline{CA}}.$$

The expected cost $E[c_{cg}]$ of the CG approach can now be calculated as follows. For sake or readability, we use R instead of $E[R]$, c_{cg} instead of $E[c_{cg}]$, c_{approval} instead of $E[c_{\text{approval}}]$, and $c_{\text{disapproval}}$ instead of $E[c_{\text{disapproval}}]$.

$$\begin{aligned} c_{cg} &= c_{\text{approval}} + (R-1) \cdot c_{\text{disapproval}} \\ &= \frac{p_{CA}}{P(cg_{\text{approve}})} \cdot c_{CA} + \frac{p_{\overline{CA}}}{P(cg_{\text{approve}})} \cdot c_{\overline{CA}} \qquad (8) \\ &\quad + (R-1) \cdot \left(\frac{p_{C\overline{A}}}{P(cg_{\text{disapprove}})} \cdot c_{C\overline{A}} + \frac{p_{\overline{CA}}}{P(cg_{\text{disapprove}})} \cdot c_{\overline{CA}} \right). \qquad (9) \end{aligned}$$

5.2. Impact of different cost factors of the MD and CG approaches

The presented cost model includes different costs which, depending on p_w, contribute more or less to the over all costs of the approach. In the following we have a look at the composition of the overall costs depending on p_w.

Fig. 4 shows the expected costs of the MD approach depending on p_w. c_c, c_w, c_{fp}, and c_{fn} are all set to 1 and $N_{md} = 11$ workers are used. In this example, the cost c_{md} are constantly increasing with p_w. This results form all workers being paid the same amount no matter if they vote according the majority or not ($c_c = c_w = 1$) and the increasing probability for the penalties c_{fp} and c_{fn} with the increase of p_w. For different values of c_c, c_w, c_{fp}, and c_{fn}, the costs c_{md} might become minimal for $p_w \neq 0$. In the following we focus on the contribution of the different cost factors depending on p_w which is independent of the specific values of c_c, c_w, c_{fp}, and c_{fn}.

First, we have a look a c_c. As the probability for a wrong answer p_w is zero, all workers are paid c_c. With in increase of p_w, some workers are submitting wrong results and are no longer paid c_c. With a further increase of p_w the wrong results are no longer detected and again more workers are paid c_c. c_w is complementary to c_c as each worker who is not paid c_c receives c_w. With p_w increasing, also the probability of a wrong majority decision increases. This results in an larger contribution of c_{fp}, as wrong answers are more likely to be assume correct. It is similar with c_{fn}. As p_w increases, more correct results are assumed to be incorrect and the contribution of c_{fp} increases. However, with increasing p_w also the number of correct answers decreases and consequently the impact of c_{fp} decreases again for large values of p_w.

Fig. 5 shows the expected costs of the CG approach depending on p_w. c_c, c_w, c_{fp}, and c_{fn} are all set to 1 and $N_{cg} = 1$ worker is used for the control task with $c_{cc} = c_{cw} = c_{cfp} = c_{cfn} = 1$.

For the CG approach c_{fp} and c_{fn} behave similar than for the MD approach. However, in the CG approach c_c is always paid as the main task is repeated as long until a result is assumed to be correct. But with increasing p_w, more repetitions of the main task are required and thus c_w and c_{md} are contribution more to the total costs, as they are paid for each repetitions.

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

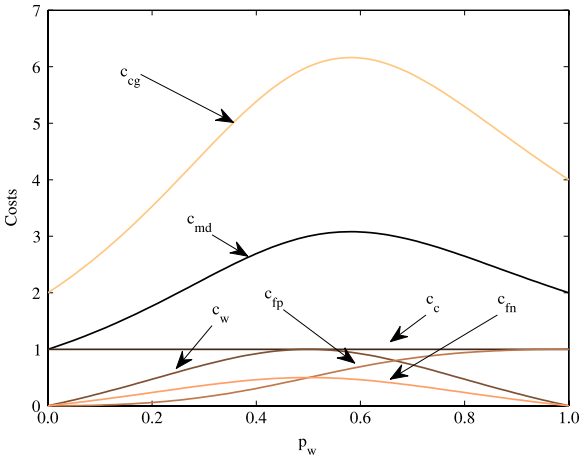

Fig. 5. Cost factors of the CG approach for $N_{cg} = 1$.

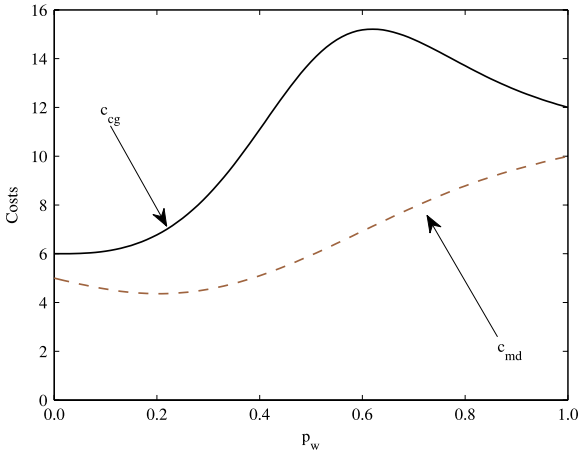

Fig. 6. Costs of an unqualified task dependent on p_w.

For large values of p_w wrong task results are again more likely to be approved and thus the impact of c_w and c_{md} decreases again. As a result, c_{cg} is not constantly increasing with p_w but has a maximum depending on the chosen parameter values.

6. Application of the cost model to different real world use cases

We now have a look at typical use cases of crowdsourcing. In the following we consider which approach, i.e. MD or CG, is optimal in terms of costs for which kind of crowdsourcing task, i.e. routine, complex and creative tasks.

It has to be noted that the costs assumed in the following sections are typical values which are taken from a large crowdsourcing platform. The costs are normalized to $c = 1$ which is the lowest payment in the crowdsourcing platform.

6.1. Routine tasks

Routine tasks are typically low paid with $c_c = 1$ for the main task. The task of re-checking the main task should not be higher paid, thus, we pay $c_{cw} = 1$ for the control task. Workers submitting wrong results are generally not paid, thus $c_w = c_{cw} = 0$. The costs caused by not detected a cheating worker are very low in this case, but as he might be encouraged to continue cheating 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. But as the crowd contains many worker who can complete simple tasks, the penalty $c_{fn} = 1$ is low. For the control crowd we also use the same penalties $c_{cfp} = c_{cfn} = 1$. The group size for both approaches is $N_{md} = N_{cg} = 5$.

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

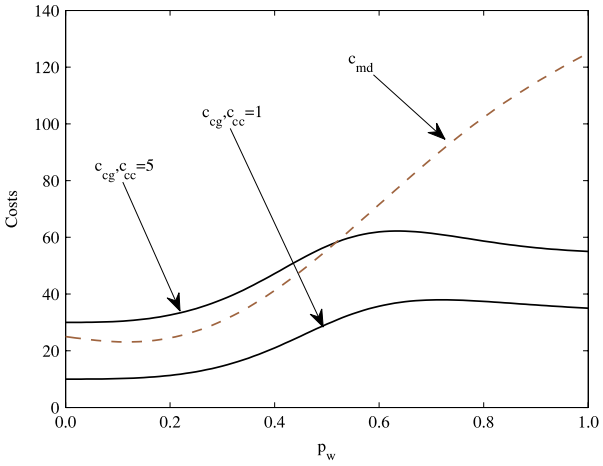

Fig. 7. Costs of a qualified task dependent on p_w.

The resulting costs depending on p_w are displayed in Fig. 6. The costs of the MD approach c_{md} are shown by the dashed line, the costs for the CG approach c_{cg} by the continuous line. At the beginning, c_{md} slightly decreases as the workers voting against the majority are no longer paid. With a further increase of p_w, we see the same effects as shown in Fig. 4 without the contribution of c_w. The development of c_{cg} differs a little from the one depicted in Fig. 5, as $c_{cw} = c_w = 0$ and $N_{cg} = 5$ but shows the same effects. Only the increase of the costs is a little later and the maximum is sharper.

But even if the difference of c_{md} and c_{cg} for small and large values of p_w is rather small, in the case of an unqualified task, the costs of the CG approach are always higher than the costs of the MD approach, since generally $c_{cc} = c_c$ and $c_{fp} \approx c_c$. Thus, the MD approach should be preferred for routine tasks.

6.2. Complex and creative tasks

For complex and creative tasks wrong results usually have a large impact on the employer. Assume an advertisement campaign in forums. The employer wants to promote a product in web forums dealing with topics related to the product. Direct advertisement is not desired in forum posts, hence, the advertisement has to be hidden in a normal post using a recommendation which fits in the context of a forum thread. As the worker has to find an appropriate forum thread and writes an individual text, we assume $c_c = 5$ in this case and again $c_w = 0$. As proof the worker submits a link to the thread where he posted the advertisement. The control group checks the given thread, if the post is related to the topic and includes the desired recommendation. Each member of the control group is paid $c_{cc} = 1$ if it rates according to the majority, otherwise it is paid $c_{cw} = 0$. If the advertisement campaign is recognized by the forum administrators, the posts are deleted and a negative discussion about the employer will arise. Thus, the penalty for approving wrong posts is set very high to $c_{fp} = 20$. Besides this, qualified workers for the main task are rare and losing one of them is not desirable. Because of this, we assume $c_{fn} = 5$. As the control crowd workers do not require special qualifications and a few miss-ratings can be tolerated we choose $c_{cfn} = c_{cfp} = 1$. The group size for both approaches is again $N_{md} = N_{cg} = 5$.

The resulting costs are depicted in Fig. 7, with c_{md} shown by the dashed line and c_{cg} by the continuous line. The costs curves of c_{md} and c_{cg} show a similar shape to that for the unqualified task. However, in this case the CG approach with $c_c = 5$ and $c_{cc} = 1$ is always cheaper than the MD approach, because of the high penalty for false positive approvals and the low costs for the control task $c_{cc} < c_c$. If the costs for the control task are raised $c_{cc} = c_c = 5$, the CG approach only performs better than MD for $p_w > 0.52$.

Using this results we derive a guideline for complex and creative tasks. If the cost ratio $c_{cc}/c_c \ll 1$, the CG approach should be favored. Otherwise, a more detailed analysis is required. Therefore, we have a look at the impact of c_{fp} and c_{fn} on choice of the optimal validation approach.

For this analysis we use $N_{md} = N_{cg} = 5$ and normalize all costs to c_c. c_{cc} is generally smaller than c_c, thus we choose $c_{cc} = c_c/2$ for this analysis. As a few miss-rating of the control-crowd are tolerable we use $c_{cfp} = c_{cfp} = c_{cc}$, similar to the previous example and $c_w = c_{cw} = 0$. In order to analyze the impact of c_{fp} and c_{fn} we vary both penalties from c_c to $10 \cdot c_c$ and determine the costs for both approaches for different values of p_w.

Fig. 8 visualizes the cost-optimal validation approach depending on c_{fp} and c_{fn} for three exemplary values of p_w. The x-axis showing the values of c_{fn} and the y-axis showing the values of c_{fp} are normalized to c_c. For each value of p_w, the line marks the decision border for the cost-optimal approach. The decision border indicates weather the CG approach is more cost effective than the MD approach. For each combination of c_{fp} and c_{fn} above the line, the CG approach is more cost effective, for combinations below the line, the MD approach is cost optimal. It has to be noted that the steps of the curves in Fig. 8, i.e. the decision border, are caused by numerical inaccuracies.

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

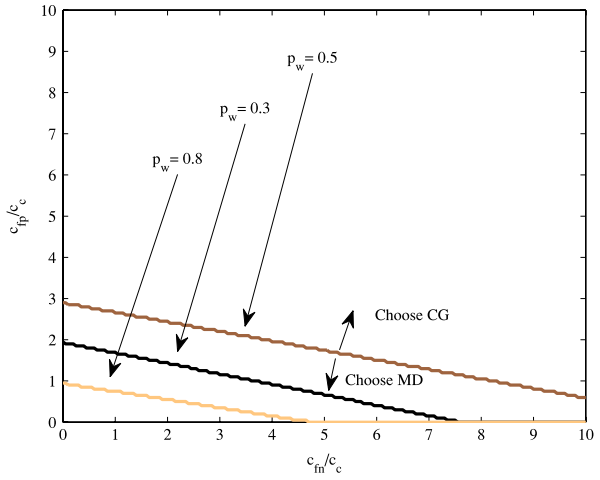

Fig. 8. Cost-optimal validation approach in dependency of c_{fp} and c_{fn}.

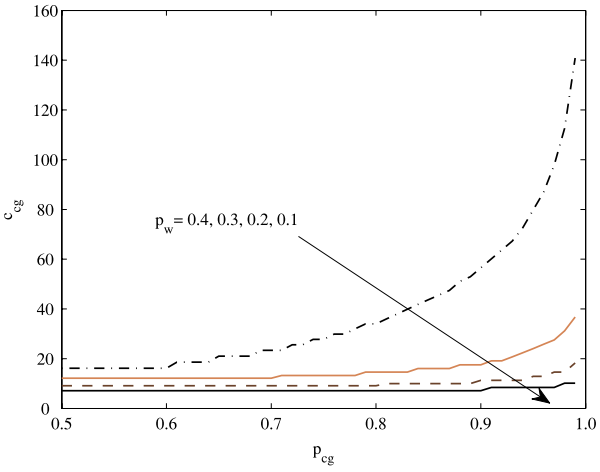

Fig. 9. Total costs c_{cg} depending the cheat-detection quality p_{cg} for different probabilities for wrong answers p_w.

The results depicted in Fig. 8 show that for $c_{cc}/c_c \ll 1$, both cost factors c_{fp} and c_{fn} have to be considered while choosing the cost-optimal validation approach. Additionally, we can see a dependency between the curse of the decision border and p_w. Depending on p_w, the optimal approach can be easily computed, as only the ratio of c_{fp} and c_{fn} determines the cost optimal approach.

6.3. Cost-quality optimization guidelines for complex and creative tasks

A cost-quality optimization,i.e. finding a trade-off between cheat-detection quality and the costs, 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.

6.3.1. Optimizing overall costs and cheat-detection quality

In order to reduce the total costs c_{cg}, a smaller control crowd can be used. But this negatively affects the quality of the cheat-detection, as p_{cg} decreases with the group size. Though, a trade-off between c_{cg} and p_{cg} exists. For our evaluation we use the example of the forum advertisement campaign with $c_c = 5$, $c_w = 0$, $c_{fp} = 20$, $c_{fn} = 5$, $c_{cc} = 1$, $c_{cw} = 0$, $c_{cfp} = 1$, and $c_{cfn} = 1$. Fig. 9 depicts c_{cg} depending on p_{cg} for different values of p_w. As c_{cg} remains almost constant for $p_{cg} < 0.5$, we focus only on $p_{cg} \geq 0.5$.

Fig. 9 shows that c_{cg} increases with p_w and p_{cg}. A better cheat-detection quality p_{cg} needs more workers 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 is rather cheap. For high values of p_w the costs increase tremendously with p_{cg}, which makes an detection improvement extremely expensive.

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

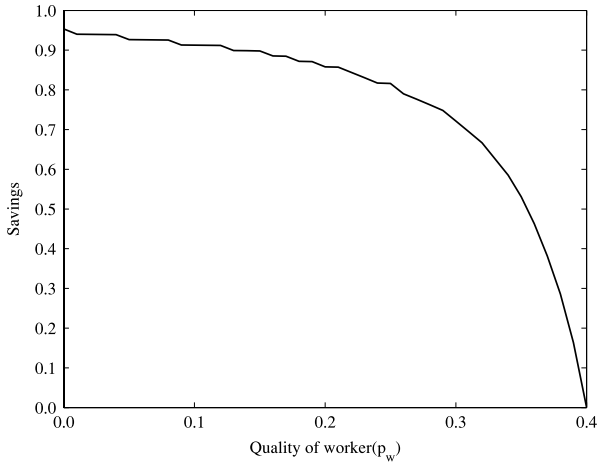

Fig. 10. Savings due to the usage of good workers.

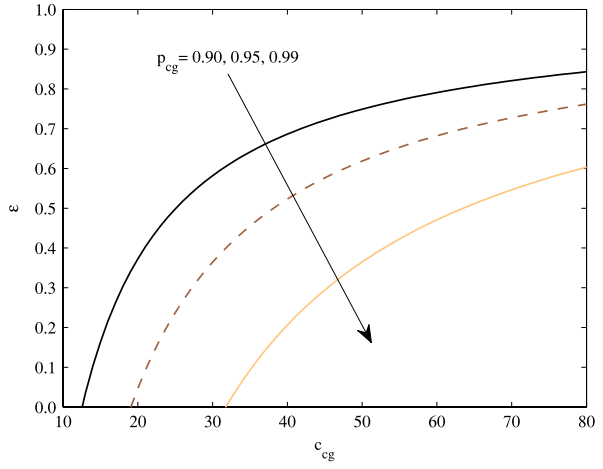

Fig. 11. Efficiency of the cost distribution ε depending on the budget b.

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

To illustrate this, we have a look at two examples with $p_w = 0.4$. Assume an employer wants to spend $c_{cg} = 30$ for the campaign. We can derive from Fig. 9 that p_{cg} will be about 0.77. For $p_w = 0.4$ Eq. (1) can be solved numerically to $N_{cg} = \frac{\ln(1-p_{cg})+1.0404}{-0.0310}$ and we can calculate the required control group size $N_{cg} = 25$. The second use case is an employer who demands $p_{cg} = 90\%$ for his campaign. We can calculate the required group size $N_{cg} = 40$ and derive $c_{cg} \approx 57$ from Fig. 9.

6.3.2. Cost saving by using specialized crowds

One possibility of influence p_w is to use only well trained workers, so called specialized crowds, for a task. In order to analyze the impact of specialized crowds on c_{cg}, we determine the 99% quantiles of the required number N_{cg} of workers for a correct decision depending on p_w. Using these group sizes, we calculate the costs of the CG approach and the savings as difference between the calculated values and the costs for the CG approach for $p_w = 0.4$. The cost savings normalized by the cost of the CG approach for $p_w = 0.4$ are shown in Fig. 10.

The figure shows that using optimal workers $p_w = 0$ can save about 95% of the costs, compared to a group of workers with $p_w = 0.4$. It also shows, that even groups with larger values of p_w still reduce the costs significantly. This results from the huge amount of workers required for the control group if p_w increases. In general, well trained workers are more expensive than untrained. However, due to the large cost saving potential of these trained workers, they are more cost effective than untrained ones.

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

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

Variable name	Relation to other variables	Meaning
General variables		
N		Number of workers in the crowd
N_m		Number of workers involved in the majority decision
p_c		Probability that a randomly chosen worker is a cheater
$p_{w\|c}$		Probability that a result submitted by a cheater is wrong
$p_{w\|\overline{c}}$	$1 - p_{w\|c}$	Probability that a result submitted by a non-cheating worker is wrong
p_w	$p_c \cdot p_{w\|c}$	Probability that a result is wrong
X		Number of wrong results
p_m	$P(X < N_m/2)$	Probability of a correct majority decision
c_c		Costs of the main task
c_w		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		
N_{md}		Number of workers used for the MD approach
p_{md}	p_m if $N_{md} = N_m$	Probability of a correct result using the MD approach
$\overline{p_{md}}$	$1 - p_{md}$	Probability of a wrong result using the MD approach
N_{mdc}		Number of workers submitting a correct result
$N_{md\overline{c}}$		Number of workers submitting an incorrect result
$C_{\text{MD},N_{md\overline{c}}}$		Costs of the MD approach if $N_{md\overline{c}}$ workers submit incorrect results
c_{md}		Expected total costs if using the MD approach
CG related variables		
N_{cg}		Number of workers used for the control group of the CG approach
p_{CA}	$(1 - p_w) \cdot p_m$ if $N_{cg} = N_m$	Probability that a correct main worker result is approved by the control group
$p_{C\overline{A}}$	$(1 - p_w) \cdot (1 - p_m)$ if $N_{cg} = N_m$	Probability that a correct main worker result is not approved by the control group
$p_{\overline{C}A}$	$p_w \cdot p_m$ if $N_{cg} = N_m$	Probability that an incorrect main worker result is approved by the control group
$p_{\overline{CA}}$	$p_w \cdot (1 - p_m)$ if $N_{cg} = N_m$	Probability that an incorrect main worker result is not approved by the control group
p_{cg}	$p_{CA} + p_{\overline{CA}} p_m$ if $N_{md} = N_m$	Probability of a correct result using the CG approach
$\overline{p_{cg}}$	$1 - \overline{p_{cg}} = p_{C\overline{A}} + p_{\overline{C}A} 1 - p_m$ if $N_{md} = N_m$	Probability of a wrong result using the CG approach
$P(cg_{\text{approve}})$		Probability that the control crowds approves the main task
$P(cg_{\text{approve}})$		Probability that the control crowds does not approve the main task
R		Number of repetitions until the main task is approved
c_{cc}		Costs for a correct control group validation task
c_{cw}		Costs for an incorrect control group validation task
c_{cfp}		Costs for a false positive correct control group validation task
c_{cfn}		Costs for a false negative control group validation task
c_{CA}		Costs if the main worker submits a correct result and the control group approves it
$c_{C\overline{A}}$		Costs if the main worker submits a correct result and the control group does not approve it
$c_{\overline{C}A}$		Costs if the main worker submits an incorrect result and the control group approves it
$c_{\overline{CA}}$		Costs if the main worker submits an incorrect result and the control group does not approve it
c_{cg}		Expected total costs if using the CG approach

6.3.3. Maximizing available salary for the main task

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

We calculate the overhead costs $c_{\text{CG-overhead}}$ using Eq. (9) and setting $c_c = 0$ dependent on the desired p_{cg}. For a fixed budget b, the maximal available salary c_c can now be calculated by,

$$c_c = b - c_{\text{CG-overhead}}.$$

We introduce the quotient $\varepsilon = c_c/b$ as a measure for the efficiency of the cost distribution. $\varepsilon = 1$ means that the entire budget is spent for the main task. Fig. 11 depicts ε for different budgets b and different cheat-detection qualities p_{cg}.

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 b, more salary for the main task is available as $c_{\text{CG-overhead}}$ remains constant. For large budgets, the main task salary is the biggest part b. The intersections of the curves and the x-axis move to the right for higher p_{cg}, which shows that the higher the desired cheat-detection quality the more expensive the task. With higher p_{cg} also the efficiency of the cost distribution degrades quickly and large amount of the budget is spent on the control crowd instead of the main worker. Therefore, an employer has to consider carefully the required p_{cg}.

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

7. 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 or did the task only sloppy.

As manual re-checking of each task is not desirable, we analyzed two different crowd-based methods, the *Majority Decision* and the *Control Group* 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.

A cost model was developed for both, the *Majority Decision* and the *Control Group* approach. Using this cost model the main cost factors of both approaches were identified and how the quality of the workers influences the weight of the different cost factors. The cost analysis also revealed that the *Majority Decision* approach is more suitable for low paid routine tasks, whereas the *Control Group* approach performs better for high priced tasks. In order to minimize the costs of high priced tasks, the *Control Group* 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 *Control Group* approach can be significantly decreased by slightly decreasing the cheat-detection quality. We also showed, that using better workers saves a lot of the costs, even if they are slight more expensive than other workers.

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

Acknowledgment

This work was conducted in the Internet Research Center (IRC) at the University of Würzburg. The authors alone are responsible for the content of the paper.

Appendix

A.1. Variables summary

See Table 1.

References

[1] J. Howe, The rise of crowdsourcing, June 2006. URL: http://www.wired.com/wired/archive/14.06/crowds.html.
[2] T. Hoßfeld, M. Hirth, P. Tran-Gia, Modeling of crowdsourcing platforms and granularity of work organization in future Internet, in: Proceedings of the 23rd International Teletraffic Congress, ITC, San Francisco, USA, September 2011.
[3] M. Hirth, T. Hoßfeld, P. Tran-Gia, Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms, in: Proceedings of the Workshop on Future Internet and Next Generation Networks, FINGNet, Seoul, Korea, June 2011.
[4] P. Ipeirotis, Analyzing the Amazon Mechanical Turk marketplace, CeDER Working Papers, CeDER-10-04, September 2010.
[5] J. Ross, L. Irani, M.S. Silberman, A. Zaldivar, B. Tomlinson, Who are the crowdworkers? shifting demographics in mechanical turk, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA, April 2010.
[6] M. Hirth, T. Hoßfeld, P. Tran-Gia, Anatomy of a crowdsourcing platform-using the example of microworkers.com, in: Proceedings of the Workshop on Future Internet and Next Generation Networks, FINGNet, Seoul, Korea, June 2011.
[7] O. Alonso, D.E. Rose, B. Stewart, Crowdsourcing for relevance evaluation, SIGIR Forum 42 (2) (2008).
[8] P. Hsueh, P. Melville, V. Sindhwani, Data quality from crowdsourcing: a study of annotation selection criteria, in: Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing, Boulder, Colorado, USA, May 2009.
[9] InnoCentive, Inc., InnoCentive, 2001. www.innocentive.com.
[10] M. Demirbas, M.A. Bayir, C.G. Akcora, Y.S. Yilmaz, H. Ferhatosmanoglu, Crowd-sourced sensing and collaboration using twitter, in: Proceedings of the IEEE International Symposium on "A World of Wireless, Mobile and Multimedia Networks", WoWMoM, Montreal, Canada, June 2010.
[11] S. Suri, D.G. Goldstein, W.A. Mason, Honesty in an online labor market, in: Proceedings of the 3rd Human Computation Workshop, HCOMP, San Francisco, USA, August 2011.
[12] J. Le, A. Edmonds, V. Hester, L. Biewald, Ensuring quality in crowdsourced search relevance evaluation: the effects of training question distribution, in: Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation, CSE2010, Geneva, Switzerland, July 2010.
[13] D. Oleson, A. Sorokin, G. Laughlin, V. Hester, J. Le, L. Biewald, Programmatic gold: targeted and scalable quality assurance in crowdsourcing, in: Proceedings of the 3rd Human Computation Workshop, HCOMP, San Francisco, USA, August 2011.
[14] P.G. Ipeirotis, F. Provost, J. Wang, Quality management on Amazon Mechanical Turk, in: Proceedings of the ACM SIGKDD Workshop on Human Computation. Washington, DC, USA, July 2010.
[15] A. Kittur, E.H. Chi, B. Suh, Crowdsourcing user studies with mechanical turk, in: Proceeding of the ACM SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy, April 2008.
[16] T. Hoßfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, Quantification of YouTube QoE via crowdsourcing, in: Proceedings of the IEEE International Workshop on Multimedia Quality of Experience-Modeling, Evaluation, and Directions, MQoE 2011, Dana Point, USA, December 2011.
[17] G. Kazai, J. Kamps, M. Koolen, N. Milic-Frayling, Crowdsourcing for book search evaluation: Impact of HIT design on comparative system ranking, in: Proceedings of the ACM SIGIR Conference on Research and Development in Information, Beijing, China, July, 2011.
[18] K. Chen, C. Chang, C. Wu, Y. Chang, C. Lei, C. Sinica, Quadrant of euphoria: a crowdsourcing platform for QoE assessment, IEEE Network 24 (2) (2010).
[19] L. Von Ahn, L. Dabbish, Labeling images with a computer game, in: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, Vienna, Austria, April 2004.
[20] C. Eickhoff, A. de Vries, How crowdsourcable is your task? in: Proceedings of the ACM WSDM Workshop on Crowdsourcing for Search and Data Mining, Hong Kong, China, February 2011.
[21] S. Dow, A. Kulkarni, B. Bunge, T. Nguyen, S. Klemmer, B. Hartmann, Shepherding the crowd: managing and providing feedback to crowd workers, in: Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems, CHI, Vancouver, USA, May 2011.

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006

[22] G. Little, L. Chilton, M. Goldman, R. Miller, Turkit: tools for iterative tasks on mechanical turk, in: Proceedings of the 1st Human Computation Workshop, HCOMP, Paris, France, June 2009.
[23] P. Dai, Mausam, D.S. Weld, Decision-theoretic control of crowd-sourced workflows, in: Proceedings of the 24th, AAAI Conference on Artificial Intelligence, Atlanta, USA, July 2010.
[24] A. Kittur, R.E. Kraut, Harnessing the wisdom of crowds in wikipedia: quality through coordination, in: Proceedings of the ACM Conference on Computer Supported Cooperative Work, San Diego, California, USA, November 2008.

Please cite this article in press as: M. Hirth, et al., Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms, Mathematical and Computer Modelling (2012), doi:10.1016/j.mcm.2012.01.006