



# Modelling Cognitive Bias in Crowdsourcing Systems

Farah Saab<sup>\*</sup>, Imad H. Elhadj, Ayman Kayssi, Ali Chehab

*Electrical and Computer Engineering Department, American University of Beirut, Riad El-Solh, Beirut 1107 2020, Lebanon*

Received 23 April 2018; received in revised form 12 March 2019; accepted 12 April 2019

Available online 19 April 2019

## Abstract

Crowdsourcing is the process of obtaining input or enlisting the services of a crowd of people with the aim of finding the best solution to a proposed problem. The method with which the collected input is aggregated affects the outcome of the crowdsourcing process. In this paper, we introduce a modelling framework through which we compare several aggregation methods for crowdsourcing systems. The work reveals a surprising result where confidence-related approaches lack in performance when compared to other approaches such as simple *plurality voting* or approaches which consider respondent competence. This inadequacy stems from a psychological phenomenon brought forth by David Dunning and Justin Kruger related to people's bias in assessing their own cognitive abilities.

© 2019 Elsevier B.V. All rights reserved.

**Keywords:** Crowdsourcing; Aggregation; Modelling; Cognitive bias; Dunning-Kruger

*“Ignorance more frequently begets confidence than does knowledge” Darwin (1871)*

## 1. Introduction

Welcome to the age of the crowd. Outsource to the crowd and the result is crowdsourcing; a term coined in 2006 by Jeff Howe and Mark Robinson from Wired magazine (Howe, 2006). Simply stated, crowdsourcing is the process of enlisting the services of a crowd of people with the aim of reaching a solution to a proposed problem. In other words, it is a way of utilizing the unused processing power of human brains. Its applications are endless. Some examples are Amazon's Mechanical Turk (MTurk), an

online market place for getting tasks done by workers (Buhrmester, Kwang, & Gosling, 2011), iStockPhoto that crowdsources stock photographs from amateurs and sells them at very competitive prices (McCurdy), InnoCentive that provides a framework for solving research problems (Singh, 2014), and TopCoder that offers a platform for developers to build software (Lakhani, Garvin, & Lonstein, 2010).

Even though crowdsourcing is still on the rise and the term was only recently introduced, human beings have been working in crowds since the beginning of civilization. Man is a social being. He dislikes solitude and has always longed for a society beyond that of his own family. His evolutionary success is mainly due to his social nature and his tendency to collaborate with his tribe (Darwin, 1871). Human beings have collaborated to build asylums for the maimed and create vaccines for the sick. They have collectively exerted their skills to ensure the continuation of their species. Their propensity to collaborate is instinctive. Spikins, Wright, and Hodgson (2016) from the University of York discuss how a subtle change in our evolutionary

<sup>\*</sup> Corresponding author.

E-mail addresses: [fws02@mail.aub.edu](mailto:fws02@mail.aub.edu) (F. Saab), [ie05@mail.aub.edu](mailto:ie05@mail.aub.edu) (I.H. Elhadj), [ayman@mail.aub.edu](mailto:ayman@mail.aub.edu) (A. Kayssi), [chehab@mail.aub.edu](mailto:chehab@mail.aub.edu) (A. Chehab).

history, thousands of years ago, has allowed individuals with autism to be integrated into society due to the emergence of a collaborative morality. On the other hand, evolution has also brought upon selfish and spiteful behavior, which is manifested through an individual's occasional preference to work alone as opposed to working within a group (Hamilton, 1970). This tradeoff between collectivism and individualism in crowdsourcing was modeled by Guazzini et al. who show how dividing a population into subgroups influences the ability of these subgroups to solve problems of varying levels of difficulty. They show both computationally and analytically how smaller groups tend to collaborate more intensely whereas larger groups create a niche for free riders who selfishly withdraw from sharing their acquired knowledge (Guazzini, Vilone, Donati, Nardi, & Levnajic, 2015).

Guazzini's model of collectivism versus individualism assumes one type of crowdsourcing tasks, mainly those that can be solved by an expert in the field. These are collaborative crowdsourcing tasks where consensus is achieved when the crowd converges to an individual solution. The larger a random crowd, the higher the probability that one individual will provide a high quality solution to the proposed task. Collectivism in this case refers to utilizing the joint intelligence of a subgroup of individuals to develop one solution provided that the contributions of the participating individuals complement rather than possibly conflict with one another. The other category of crowdsourcing tasks are those with a collective solution which is derived by means of aggregating the input of every individual in the crowd, whether this input is in agreement or disagreement with one another. Our focus in this paper is on the second type of crowdsourcing tasks, which are usually referred to as micro-tasks; those with a collective solution. Also, we do not assume a reward factor in our work. Financial rewards are the simplest ways to attract workers (Quinn & Bederson, 2011). However, once monetary rewards are included, workers have more incentives to cheat to increase their pay. Removing the reward factor thereby rids our system of the notion of free-riders.

Crowdsourcing tasks with a collective solution are similar in nature to the voting process in electoral systems. As mentioned by Conitzer et al., voting is a method of preference aggregation over a set of alternatives that can range from potential presidents to a ranked list of popular songs (Conitzer & Sandholm, 2012). Every vote in a voting scheme corresponds to a noisy perception of the correct outcome. The aggregation technique employed by the voting platform should be carefully crafted to find a compromise candidate that maximizes the voters' combined wellbeing while inferring their reliability based on their noisy votes (Dwork, Kumar, Naor, & Sivakumar, 2001). A good aggregation technique manages to extract a hidden objective ground truth that is external to human judgement given many problems such as varying expertise and task difficulty levels (Quoc Viet Hung, Tam, Tran, & Aberer, 2013). It can intelligently process crowdsourced data with

the goal of returning maximum benefit to the crowd (Barbier, Zafarani, Gao, Fung, & Liu, 2012). We focus on four of the most popular voting methods. The difference among these methods is in the utilized aggregation technique. We begin with the simplest aggregation method which is based on *plurality voting* where the most voted for choice by the crowd is selected as the final solution to a task. Besides its simplicity, another advantage of *plurality voting* is its elimination of random and incompetent replies from users much like its exclusion of extremist party representation in electoral systems. However, this exclusion is a two-edged sword. *Plurality voting* systems suffer from the tyranny of the majority. They have strong disincentive to the emergence of a new party or idea. This means that experts in the field, if not in majority, will not have decisive influence on the final output of the crowdsourcing process (Lijphart, 1991). To alleviate the disadvantage of unfair representation, weighted algorithms were introduced. One such example is the *competence-weighted* approach where more weight is given to replies from individuals with higher capabilities than others in the crowd. Giving more "competent" individuals an advantage in voting goes back to the first form of "democracy" in Athens around the sixth century B.C. (Saxonhouse, 1993). Even though participatory democracy was practiced to broaden participation, only citizens, as defined by the state, had the right to vote. This naturally meant excluding women, slaves, convicts, foreigners, and children. It was commonly known that the Athenian assembly must be a sample of the citizenship. However, this sample was not even a random one (Carter, 1986). Only individuals who were deemed worthy or competent enough were allowed to vote. Further evidence was given by Mogens Hansen who calculated the seating capacity of the Pnyx and estimated the space taken by an adult Athenian male to conclude that the Pnyx was able to hold at most 6000 people which is not even close to a fair representation as defined by modern democracy standards (Hansen, 1983).

The main limitation in *competence-weighted* approaches is the accuracy of the competence detection method that is employed in the system. In many crowdsourcing frameworks, the competence or expertise of the participants is based on a detection algorithm whose results may not be accurate as it provides in most cases a mere estimate of the actual competence of the participants, which may never be known. In some cases, the competence of a participant is based on the agreement of his reply with the "correct" reply, which raises the question of detection accuracy in cases where the ground truth is not known a priori. This variation between the estimated and true competence of participants is largely responsible for the success or failure of *competence-weighted* approaches in crowdsourcing systems. To this end, another weighting approach surfaced, which is based on ground truth information collected from the participants themselves. *Confidence-weighted* aggregation techniques are very popular and are in some cases considered an adjusted approach to *plurality voting*. The idea is

to give higher weights to the answers of participants who are more confident and lower weights otherwise. Another variation is to apply *plurality voting* on a subset of participants with maximum reported confidence. We argue that introducing confidence into the aggregation method should be studied more carefully. In most cases, people tend to overestimate or underestimate their mental capabilities and this might affect their perception of their own knowledge when answering a question or solving a task. This psychological phenomenon was in fact introduced in a very famous study by David Dunning and Justin Kruger in 1999 (Kruger & Dunning, 1999). Modelling this psychological bias and incorporating the result into confidence-related aggregation methods is a major part of this work. In a sense, we can draw an analogy between the error in competence detection in *competence-weighted* approaches and the psychological bias in assessing one's own abilities in *confidence-weighted* approaches. The success or failure of each approach in giving the correct answer depends on the error or bias size. Of course aggregation from multiple sources has been widely used in areas such as machine learning. In the following sections, we describe some of the work related to aggregating data from multiple automatic systems for example in speech recognition. However, it is worth noting that one of the main differences between aggregating data obtained from systems and aggregating data obtained from humans is that in the latter case, there is considerable bias emanating from human factors that must be coped with for better output quality. Confidence scores reported by automatic systems might suffer from inaccurate self-assessments, however, these inaccuracies differ in nature from those emanating from human factors such as the Dunning-Kruger psychological effect that is modeled in this work. There are several sources of error in machine learning systems. Two of these errors are referred to as bias and variance and the trade-off between them is commonly known as the bias-variance dilemma. The bias is an error resulting from invalid assumptions in the learning algorithm and often results in what is known as underfitting. The variance, on the other hand, is an error that results from sensitivity to minor fluctuations in the training data and results in overfitting. Modelling these two types of errors in machine learning systems depends on several factors related to the employed learning algorithm and is beyond the scope of this paper.

In what follows, we begin with a description of some of the related work where we present a study on quality control mechanisms in crowdsourcing systems and discuss some of the work on cognitive biases in crowdsourcing as well as the different types of crowdsourcing tasks. We then proceed to describe our proposed model of the Dunning-Kruger effect that will be used throughout the paper. Following that, we model the aggregation methods that were mentioned earlier: *Plurality Voting* (PR), *Confidence-Weighted* (CF), *Maximum Confidence* (MC), and *Competence-Weighted* (CP). Our aim is to provide the necessary conditions for the input variables in our system that

will render one method superior to the others. Finally, we test our findings on a crowdsourced dataset and present a discussion of the main results and conclusions that can be drawn from this work.

## 2. Related Work

We discuss some of the related research in the literature for a complete understanding of the methods used in this work. We begin with the issue of quality control mechanisms in crowdsourcing systems which have gained considerable attention recently. We describe how these mechanisms are related to the methods presented here. We then provide a summary of some of the literature on cognitive bias and how it is modeled within the context of crowdsourcing systems. Finally, the different types of crowdsourcing tasks are studied in the last subsection.

### 2.1. Quality Assurance in Crowdsourcing Systems

The benefits of crowdsourcing are realized when a large number of workers participate in solving small tasks. However, these contributing workers may try to cheat the system especially in the presence of monetary reward or they may make mistakes due to personal bias or different experience levels with the subject matter. One approach to detect such workers would be to manually verify the output quality. The problem is that manually verifying the quality of the submitted results is hard and negates many of the advantages of crowdsourcing. In some cases, verifying every submitted solution has the same cost and time as performing the task itself. We need algorithms that will accurately estimate the quality of the submitted work for maximum benefit.

There is a plethora of quality assurance (QA) techniques designed for crowdsourcing systems that can be categorized along two main dimensions: design-time and runtime approaches (Allahbakhsh et al., 2013). Worker selection based on pre-specified reputation levels or pre-specified credentials are two examples of design-time approaches. Schall et al. estimate reputations of experts based on link structure and periodically updated trust relations that capture any changes in preferences and maintain skill evolution (Schall, Skopik, & Dustdar, 2012). In some cases, domain experts check the contribution quality. The Wikipedia encyclopedia employs an expert review approach for quality control (Quinn & Bederson, 2011).

Dawid and Skene use the expectation maximization algorithm to estimate both the quality of the workers and the correct answers for each task (Dawid & Skene, 1979). Ipeirotis et al. argue that the inherent value of a worker cannot be measured from the error rate alone (Ipeirotis, Provost, & Wang, 2010). Workers may be careful to avoid error but their solutions might suffer from bias, the effects of which can be reversed. They expand on the work of Dawid and Skene and present an algorithm that separates this bias of potential high-quality workers from other unre-

coverable errors of low-quality workers. Their approach leads to better treatment of workers and allows for better quality estimation.

Amazon's Mechanical Turk is an example of a crowdsourcing marketplace that incorporates several QA methods such as reputation systems, majority consensus, contributor evaluation (Kittur, Chi, & Suh, 2008; Sorokin & Forsyth, 2008), and ground truth where a small number of tasks with an available gold standard are mixed in with other tasks so as to identify malicious workers who are deliberately attempting to sabotage the system (CrowdFlower; Sorokin & Forsyth, 2008). Requesters on MTurk can also design defensive tasks which are more difficult to cheat on than properly solve. For example, translation HITs can be redesigned so as to display images of sentences to be translated instead of text, thus making it harder for workers to copy and paste into a machine translation engine (Callison-Burch & Dredze, 2010). Another QA method it provides is based on redundancy where each task is performed by several workers and high-quality solutions are identified based on some voting scheme. The redundancy QA approach is also incorporated in the famous reCAPTCHA protection service (von Ahn, Maurer, McMillen, Abraham, & Blum, 2008). In case of any discrepancy among received answers, a word is sent to several other workers and the answer that has the highest votes is selected.

Two popular and somewhat similar approaches to QA are output agreement and input agreement (von Ahn & Dabbish, 2008). They are commonly used for labelling tasks. In output agreement, two or more randomly chosen workers are given the same input and are required to produce output based on this input. Matching output is selected as the winning label. Since workers are chosen randomly, the quality of the output is verified considering that it is based on agreement from two largely independent sources. One very popular example of a game that uses this approach is the ESP game (von Ahn & Dabbish, 2004). In input agreement, the two or more randomly selected workers are given input. The workers do not know if they are given the same or different input. They provide labels and can observe other player's labels as well. Based on all labels, the workers decide if they have the same input or not. If they correctly determine whether or not they have the same input, their labels are taken into consideration. Quality of labels is maintained by discouraging random guesses via strong penalties. This is an example of a Game With A Purpose (GWAP) where steps within a computational process are outsourced to humans in an entertaining way. A popular example is another GWAP which is TagA-Tune (Law & von Ahn, 2009). Similar to output agreement, another QA approach is the multilevel review. However unlike in output agreement, the work is not done in parallel. A group of workers first perform a task. Then a different group assesses its quality. An example of this approach is the Find-Fix-Verify crowd programming pattern proposed by Bernstein et al. (2010).

The above list presents a concise survey of the available QA methods, most of which attempt to correctly identify worker characteristics for better solution quality. Even in the absence of the ground truth, these mechanisms can be used to estimate the competence of workers to a very good degree. Our approach to detect competence based on reported worker confidence complements the work in the literature on quality control via proper user identification. In a sense, it can be categorized as a QA method that can be applied with other QA methods (ground truth seeding, contributor evaluation, etc.) for a better representation of user abilities that takes into consideration incentives, technical skills, experience with a topic, along with any personal and cognitive biases.

## 2.2. Modelling Cognitive Biases in Crowdsourcing Systems

The Dunning-Kruger effect is one of the many cognitive biases that human beings tend to fall prey to. Eric Bonabeau points out to several human biases that can be observed while making decisions. Examples of biases while generating solutions are the self-serving bias where humans tend to search for information that confirm their existing assumptions, anchoring where they tend to heavily rely on one piece of information, and stimulation bias where they only recognize a solution when they see it. Examples of biases in the evaluation phase of potential solutions are pattern obsession where humans realize patterns that do not exist, and framing where evaluation is influenced by how the solution is presented (Bonabeau, 2009). Fleischmann et al. discuss how the phenomenon of cognitive bias has been explored in psychology since the mid-seventies, but has only recently gained attention in the area of information systems (Fleischmann, Amirpur, Benlian, & Hess, 2014). According to them, interest in cognitive bias research is increasing in information systems considering how it revolves around human decision-making. They define cognitive biases in humans as systematic errors in the decision making process that result in suboptimal outcomes. They identify 120 cognitive biases in their analysis which was larger than the number of relevant papers at the time (84). They categorize the biases into perception (e.g. negativity bias), pattern recognition (e.g. confirmation bias), memory (e.g. reference point dependency), decision (e.g. cognitive dissonance), action-oriented (e.g. overconfidence), stability (e.g. anchoring), social (e.g. cultural bias), and interest (e.g. self-justification).

Roy et al. discuss a main limitation in crowdsourcing systems nowadays where inadequate representation of the uncertainty resulting from human factors results in suboptimal system design (Basu Roy, Lykourantzou, Thirumuruganathan, Amer-Yahia, & Das, 2013). The human factors that they discuss are mainly related to the availability of workers (workers may leave unexpectedly), the wage that they may request at any point in time, and their varying skill levels. They propose SmartCrowd, an interactive crowdsourcing system that takes into account

the dynamic and uncertain nature of crowdsourcing environments. Faltings et al. discuss how human computation is susceptible to systematic biases that cannot be corrected by simply aggregating multiple answers (Faltings, Jurca, Pu, & Tran, 2014). They study the case of Amazon's Mechanical Turk. The difference between an error and a bias is that the latter can be mistaken for the true value and is therefore harder to detect and cannot be treated similarly. Their examples include the anchoring effect, common beliefs, and recurring answer bias, all of which cannot be eliminated by simply increasing worker count or any of the common methods for quality control. To this end, they propose the Peer Truth Serum, a game-theoretic incentive scheme that evaluates how scaling bonuses can overcome biases in worker answers resulting in significant improvement in system accuracy. Eickhoff studies the effect of several cognitive biases in document relevance assessment tasks (Eickhoff, 2018). He demonstrates how the literature commonly assumes that noisy label submissions are due to three main reasons: unethical spammers, unqualified workers, and malicious workers. There is no consideration to cognitive biases that are systematic deviation patterns from the ground truth. The first step in countering the effects of such biases in crowdsourcing systems is to recognize them. He demonstrates how common QA methods are not enough to overcome this source of noise. To prove the effect of cognitive biases on a system, he studies the ambiguity effect which occurs when missing information in a question makes it appear more difficult and ultimately less desirable to solve. The missing information was chosen so as to have very little relevance with the document to be labelled. He showed that even when the missing information was not informative, it negatively affected workers' outcome. He then designed a two-phase experiment. In the first phase, workers are presented with information not relevant to the document, and then at a later stage, relevant information becomes available. This is known as the anchoring effect and he shows how it considerably reduced label accuracy. To study the bandwagon effect in another experiment, he discloses to the workers the prior vote statistics resulting in a drop in accuracy. Finally, he looks into the decoy effect, which is a very popular effect in the advertisement discipline where a worker's preference between two options changes with the introduction of a third option (the decoy). He shows the high risk of unintentionally suffering from this effect when several options are provided for relative ranking resulting in degraded label quality.

Gadiraju et al. present a similar work to ours (Gadiraju, Fetahu, Kawase, Siehndel, & Dietze, 2017). They demonstrate how to use worker self-assessments to derive competence levels that can be used along with their performance in the pre-screening phase to achieve better results in crowdsourcing micro-tasks. They start off by describing the Dunning-Kruger effect and investigating through several studies whether it can be observed in paid micro-task crowdsourcing systems which are different in many ways

when compared to the controlled experiments performed in the original study by Dunning and Kruger. In their first experiment, they wanted to study crowd workers' self-assessment. They deployed 8 tasks on CrowdFlower with varying difficulty levels. After solving the tasks, workers were asked questions related to their perceived test scores and those of others, as well as their perceived abilities. They observed that across all tasks of varying difficulty levels, least-competent workers were the ones to significantly overestimate their abilities and scores, whereas competent workers underestimate their abilities. In addition, competent workers overestimate other workers' performances more than incompetent workers do. In another experiment, the authors wanted to study the effects of competence on a crowdsourcing task such as tagging images. They were able to show that competent workers outperformed least-competent ones by providing better quality tags that are more diverse. Based on the results from this experiment, they suggested to employ worker self-assessments in the pre-screening phase for better worker selection. They performed another experiment where one group of workers were pre-screened in the traditional manner that considers only their performance whereas another was pre-screened based on both worker performance and self-assessment. The crowdsourcing task for both groups was sentiment analysis. Their results showed that including self-assessment in the pre-screening phase provides a better representation of actual worker competence which results in improved output quality. This work represents a starting point for research related to worker self-assessment in crowdsourcing tasks.

Part of our work complements that of Gadiraju et al. in terms of detecting worker competence through self-assessment, we have reached similar results. In addition to the common pre-screening method, we also provide a formal model of this cognitive bias which allows us to study several self-assessment-related aggregation techniques. Our model was also helpful when comparing the performance of these techniques to others that are not based on self-assessment such as the widely used *plurality voting* approach and the new *surprisingly popular* approach (Prelec, Seung, & McCoy, 2017). Our work is more generic than that of Gadiraju et al. in that it considers all types of workers in the crowd. We present the idea of a general crowd that has a confidence-competence plot as described in the original paper by Dunning and Kruger. We also study the case where the crowd is irregular meaning that the majority of its workers lie on one side of the spectrum in terms of expertise level. Gadiraju et al. showed how the Dunning-Kruger effect is in fact present in crowdsourcing systems but they failed to describe how this bias is reflected on the overall performance of a crowdsourcing system given different crowd types. Our model studies different crowds and the resulting performance given each aggregation technique.

Correctly estimating workers' actual competence levels is one of the several methods to perform quality control

in crowdsourcing systems as described before. The problem is that most of these methods do not focus on the human factor in crowdsourcing, and those that do, focus on trying to detect personal bias patterns and reversing their effects on the overall system output. The number of biases, however, is not small, and there is definitely plenty of room for research in this area. In the specific case of the Dunning-Kruger psychological bias, our work and that of Gadiraju et al. provide a starting point.

### 2.3. Aggregation Based on Task Types

In our model, we only consider crowdsourcing tasks with a collective solution derived by means of aggregating the input of all individuals in the crowd. These tasks are similar to voting schemes where a relatively large number of workers have to participate simultaneously. Also, they are not open-ended in nature but rather have a definite answer that a group of individuals can agree on. It is worth noting, however, that there are other crowdsourcing tasks that do not fall under the model presented in this work but are worthy of mention. One example is the classical ROVER algorithm (Fiscus, 1997). The NIST Recognizer Output Voting Error Reduction system combines the output generated by multiple Automatic Speech Recognition systems resulting in a lower error rate than any of the individual systems. First, the outputs from multiple ASR systems are aligned to generate a single word transition network. Then, a voting scheme is applied to select the highest scoring word. Three benchmark evaluation submissions were used to test ROVER. Each output word was provided along with a confidence score ranging from 0 to 1. The authors investigated three voting schemes: *majority*, *confidence-weighted*, and *maximum confidence*. The best error reduction was achieved with the *maximum confidence* approach followed by the *confidence-weighted* approach and finally the *majority* approach. ROVER combines system outputs of multiple recognition systems. The difference between ROVER's voting scheme and that in crowdsourcing systems is that in the latter, the miscalibration in self-assessment leads to inaccurate confidence scores that render the *majority* score superior to *confidence-weighted* scores. The human factor does not apply in the ROVER case.

Chowdhury et al. use the ROVER tool to address the problem of cross-language transfer of domain-specific semantic annotation (Chowdhury et al., 2014). There are several issues to address. The language of interest might be under-represented due to the fact that crowdsourcing platforms have a very skewed distribution of users which can sometimes result in having a small number of speakers for the desired language. This was addressed by using targeted crowdsourcing. Also, domain-specificity of the required annotation increases complexity of the task. This was coped with via priming the annotators with the unique list of concepts from the source language. Another issue is the evaluation of crowd-annotated data without reference

target language annotation, which was coped with by applying inter-annotator agreement. It considers both segmentation and labeling agreement measures meaning that annotators have to agree on both the label and its span. The authors apply the ROVER technique with a majority voting scheme to decide on the label and its span. Their results show acceptable annotation quality. In a more recent work, they study selection and aggregation techniques for the semantic annotation task (Chowdhury et al., 2015). Their goal is to select a word mapping that is closest to the source language or to aggregate all mappings into a single one that best represents the meaning of an utterance. The baseline for selection is randomly picking one of the mappings whereas that for aggregation is using majority voting while randomly breaking ties. Both these approaches are not a good choice considering the varying expertise levels of crowd annotators. Language Models (LM) based on the maximum likelihood were used to estimate the reliability of crowd annotators. Their results show that majority-voted ROVER provides a strong baseline. However, weighting each hypothesis with respect to other annotations proves to be the best weighting scheme with an increase of 0.4 in the F-measure. Stepanov et al.'s baseline evaluation is the random re-sampling approach where the precision of one randomly selected judgment is computed and results are averaged after repeating the procedure 1000 times (Stepanov et al., 2018). Their results show that the performance of majority voting is higher than the baseline thus proving how combining the 'power of the crowd' with computational methods improves annotation quality.

In addition to what was mentioned, there are crowdsourcing tasks that have one solution only and these are modeled differently than tasks that aggregate worker solutions. In these types of tasks, it is up to the requestor to determine the winning solution based on the amount of money he is willing to pay and the minimum required solution quality. In addition, there are crowdsourcing tasks with open-ended questions. For example, Wikipedia is a crowdsourced encyclopedia. Several contributors can participate in creating Wiki pages and there is no one solution or output format for the pages. These types of tasks are modeled differently and are beyond the scope of this paper.

### 3. Modelling the Dunning-Kruger effect

The Dunning-Kruger effect is a cognitive bias in which low-ability individuals suffer from the illusion that their abilities are higher than what they really are. This effect was presented in a renowned study in 1999 by psychologists David Dunning and Justin Kruger who attributed this illusion of superiority in low-ability individuals to their metacognitive weakness in accurately evaluating their own competence. In other words, they have the propensity to overvalue their abilities when solving a task or answering a question. The study also shows how high-ability indi-

viduals generally underestimate their competence (Kruger & Dunning, 1999).

With this idea in mind, we modeled the relationship between competence and confidence, as presented in their study, as a continuous function. To the best of our knowledge, modelling this psychological effect has not been done before. The function is estimated by a quadratic equation that is concaved upwards. The choice of an upward-concaved function matches properly the decrease-increase pattern in confidence levels as a respondent's competence increases.

$$y_r = ax_r^2 + bx_r + c \quad (1)$$

The subscript  $r$  refers to the respondent,  $y_r$  represents the respondent's confidence level,  $x_r$  represents his competence level, and the coefficients  $a$ ,  $b$ , and  $c$  determine the resulting shape of the competence-confidence plot, which we will refer to from now on as the DK plot or DK function after Dunning and Kruger. Fig. 1 shows several samples of the DK plot. Depending on the type of crowd, one DK plot might be a better representation than another. Note that DK 4 is different from the other plots in that it starts with a straight line and continues with a parabola. This is another possible representation of the Dunning-Kruger effect. It shows how when a respondent knows nothing about a topic, his confidence level is very low. However, upon basic introduction to it, for example after skimming through a Wiki page, his competence increases slightly while his confidence explodes. Intuitively speaking, representing the Dunning-Kruger effect as a piecewise function as in DK 4 is a more general representation. However, it complicates the mathematical analysis significantly and has a minor effect on the obtained results. We choose to adopt a simplified general model of this effect as represented in DK 1, DK 2, and DK 3.

Choosing to model the DK effect with a polynomial function serves two purposes. First, as will be seen throughout the work, it will help us formally prove the superiority of certain aggregation techniques over others for different task difficulty levels. Second, it allows us to

model a wide variety of crowd types. The DK effect was studied in settings where both competent and incompetent workers exist. There are scenarios where the crowd is mostly competent or mostly incompetent, and in those cases, the shape of the polynomial will differ. Throughout this work, we will follow the assumption of a general crowd that has workers of all competence levels. However, it is important to keep in mind that DK1, DK2, and DK3 do not accurately represent crowds in every case. That is where the role of the coefficients  $a$ ,  $b$ , and  $c$  comes into play.

It is worth noting that the Dunning-Kruger effect is at the heart of many studies on illusory superiority. People who noticeably lack in areas of logical reasoning, emotional intelligence, grammar, financial knowledge, math, chess, fairness, job skills, driving abilities and other disciplines have the tendency to rate their expertise almost as favorably as actual experts do (Kim, Chiu, & Zou, 2010; Kruger & Dunning, 1999; Liebrand, Messick, & Wolters, 1986; Park & Santos-Pinto, 2010; Poundstone, 2017; Roy & Liersch, 2013). It is the prevalence of this effect throughout various disciplines that motivated us to formally model it in our paper.

### 3.1. Function Constraints

We define constraints on the shape of the DK function that are derived from characteristics of this psychological effect as described by Dunning and Kruger. Both confidence and competence values are in the range  $[0, 1]$ :

- (1) Respondents with the lowest competence have maximum confidence in their abilities, which means that the function starts at the point  $(0, 1)$ . From here on out, we will set the value of the coefficient  $c$  to 1.
- (2) Confidence levels decrease then increase as competence levels become higher, which means that the function is concaved upwards. From now on, we note that the coefficient  $a$  should be positive.

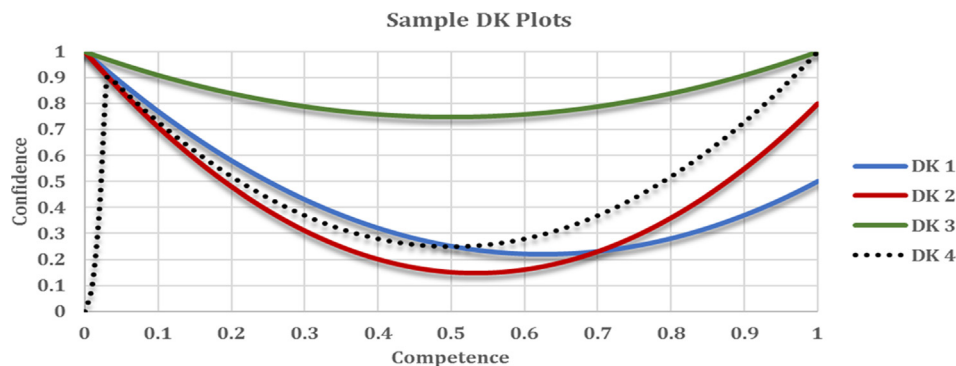


Fig. 1. Four sample DK plots. In DK 1,  $a = 2$  and  $b = -2.5$ . In DK 2,  $a = 3$  and  $b = -3.2$ . In DK 3,  $a = 1$  and  $b = -1$ . And in DK 4,  $a = 3$ ,  $b = -3$ , and the slope of the line is 1000.

- (3) Confidence levels cannot be lower than 0, which means that the point of minimum confidence either lies on the x-axis or is above it. This is achieved by setting a constraint on the function to have at most one real root.

$$b^2 - 4a \leq 0 \Rightarrow -2\sqrt{a} \leq b \leq 2\sqrt{a}$$

- (4) On the other extremity of the function, we note that respondents with maximum competence should also have a confidence value anywhere between 0 and 1.

$$x_r = 1 \Rightarrow 0 \leq y_r \leq 1 \Rightarrow -a - 1 \leq b \leq -a$$

Combining conditions (3) and (4) gives us the range of allowed values for  $b$  given  $a$ , which only applies for  $0 \leq a \leq 4$ :

$$-2\sqrt{a} \leq b \leq -a$$

*Dunning–Kruger psychological effect model.* A quadratic function representing the relationship between a participant's confidence ( $y_r$ ) and competence ( $x_r$ ) where  $y_r = ax_r^2 + bx_r + 1$  given that  $0 \leq a \leq 4$  and  $-2\sqrt{a} \leq b \leq -a$ .

### 3.2. Competence Model

We define respondent's competence as the likelihood of him or her solving a task  $k$  correctly. It is given by the Rasch psychometric measurement model, which states that the competence  $x_r^k$  of a respondent  $r$  solving a task  $k$  is a function of the respondent's ability  $\theta_r$  and the task difficulty  $\Delta_k$  (McCoy & Prelec, 2017; Rasch, 1961).

$$x_r^k = f(\theta_r) = \frac{\theta_r(1 - \Delta_k)}{\theta_r(1 - \Delta_k) + \Delta_k(1 - \theta_r)} \quad (2)$$

The competence increases with the increase in the respondent's ability and decreases with the increase in task difficulty. The ability  $\theta_r$  is a value between 0 and 1. We define this value to be normally distributed with mean  $\mu$  and standard deviation  $\sigma$ . The task difficulty  $\Delta_k$  is also a value between 0 and 1. However, in our model, we do not consider tasks with zero difficulty levels. We assume that even the easiest task will require some time to complete, i.e.  $\Delta_k \in (0, 1]$ .

Taking the Rasch competence model into account, the DK function can then be defined in terms of respondent's ability, task difficulty, and the two coefficients  $a$  and  $b$ .

$$DK = y_r^k = a \left( \frac{\theta_r(1 - \Delta_k)}{\theta_r(1 - \Delta_k) + \Delta_k(1 - \theta_r)} \right)^2 + b \left( \frac{\theta_r(1 - \Delta_k)}{\theta_r(1 - \Delta_k) + \Delta_k(1 - \theta_r)} \right) + 1 \quad (3)$$

*Expected Competence.* The expected value of a respondent's ability ( $\theta_r$ ) is the ability mean  $\mu$ . To get the expected value of the competence ( $x_r^k = f(\theta_r)$ ), we first refer to

Jensen's inequality, which states that for a convex function  $f(\theta_r)$ ,  $\mathbb{E}[f(\theta_r)] \geq f(\mathbb{E}[\theta_r])$ , and for a concave function,  $\mathbb{E}[f(\theta_r)] \leq f(\mathbb{E}[\theta_r])$ . As defined above, the shape of the competence function according to the Rasch measurement model depends on the value of the task difficulty ( $\Delta_k$ ). For  $0 < \Delta_k \leq 0.5$ ,  $f(\theta_r)$  is concave, which gives us an upper limit  $f(\mathbb{E}[\theta_r])$  on the competence expected value. For  $0.5 \leq \Delta_k \leq 1$ ,  $f(\theta_r)$  is convex, which gives us a lower limit  $f(\mathbb{E}[\theta_r])$  on the competence expected value.

$$f(\mathbb{E}[\theta_r]) = \frac{\mu(1 - \Delta_k)}{\mu(1 - \Delta_k) + \Delta_k(1 - \mu)} \quad (4)$$

$$\begin{cases} 0 < \Delta_k \leq 0.5 & \mathbb{E}[f(\theta_r)] \leq f(\mathbb{E}[\theta_r]) \\ 0.5 \leq \Delta_k \leq 1 & \mathbb{E}[f(\theta_r)] \geq f(\mathbb{E}[\theta_r]) \end{cases}$$

The probability of one participant choosing the best option among several options is equal to the expected value of the competence. In a crowdsourcing context however, there are many participants. The probability of their aggregate answer being correct depends on the aggregation method. Modelling this probability is one of the main contributions of this paper.

Notes. Given the modelling constraints presented above, there are some observations that are worth mentioning:

- (1) Setting both coefficients in the DK function equal to 0 is equivalent to aggregating by plurality.

$$a = b = 0 \Rightarrow y_p = 1 \forall x_p$$

- (2) An extreme case is when the confidence level strictly decreases as the competence increases.

$$a = 1 \text{ and } b = -2\sqrt{a} \text{ or } a = 0 \text{ and } b = -1$$

- (3) When the two coefficients  $a$  and  $b$  are equal in absolute value to  $m$ , the confidence ( $y_r$ ) starts at 1 for minimum competence ( $x_r = 0$ ) and ends at 1 for maximum competence ( $x_r = 1$ ), i.e., the plot is symmetric about the vertical line  $x_r = 0.5$ . In addition, as  $m$  increases, the point of minimum confidence on the plot shifts downwards.

- (4) For the same values of  $a$  and  $b$ , as  $\Delta_k$  increases, the point of minimum confidence shifts towards the right.

- (5) Increasing  $a$  has the effect of increasing the confidence faster for respondents of higher competence.

- (6) Decreasing  $b$  has the effect of decreasing the confidence faster for respondents of lower competence.

## 4. Modelling Aggregation

In any crowdsourcing system, the method used to combine the large number of collected replies into a single

output is crucial in determining the success of the crowdsourcing process. Depending on many factors, most notable of which are related to the size and characteristics of the crowd, one approach may perform better than the others. In this part of the paper, we provide a general model for the most popular aggregation methods focusing on how aggregation is performed and how the probability of reaching a correct answer changes with the change in the defined system properties.

#### 4.1. Plurality

In PR voting, one of the most popular and simplest aggregation methods is used. The method selects the most voted for choice by the crowd as the final answer. The aggregated reply  $\mathbb{a}_q^{PR} \rightarrow \arg \max_{\mathbb{a} \in \mathbb{A}} \sum_{r \in R_q} \delta_{\mathbb{a}_q^r \mathbb{a}}$  for a question  $q$  is the one that agrees most with replies given by all participating respondents where  $\mathbb{a}$  is each of the possible answers from the set of answers  $\mathbb{A}$ ,  $\mathbb{a}_q^r$  is the answer for question  $q$  given by respondent  $r$ , and  $\delta_{\mathbb{a}_q^r \mathbb{a}}$  is the Kronecker delta which returns 1 when  $\mathbb{a}_q^r$  matches  $\mathbb{a}$  and 0 otherwise.

In a crowdsourcing scenario with  $N_q$  participating respondents, the probability of success of more than half of them is equivalent to independently repeating the experiment  $N_q$  number of times. PR with  $N_q$  independent respondent decisions gives an overall correct answer that follows the binomial formula (Eq. (5)) where  $\mathbb{a}_q^*$  is the correct answer to question  $q$  and  $P(\delta_{\mathbb{a}_q^* \mathbb{a}_q^*} = 1) = \mathbb{E}[f(\theta_r)]$ .

$$P_q^{PR} = P(\mathbb{a}_q^{PR} = \mathbb{a}_q^*) = \sum_{i=\frac{N_q}{2}}^{N_q} \binom{N_q}{i} (P(\delta_{\mathbb{a}_q^* \mathbb{a}_q^*} = 1))^i (1 - P(\delta_{\mathbb{a}_q^* \mathbb{a}_q^*} = 1))^{N_q-i} \quad (5)$$

For a large  $N_q$ , as is the case in most crowdsourcing systems, we can use the Chernoff bound to get a sharp bound. Let  $X$  be the number of respondents who answer correctly, we obtain a lower bound on  $P_q^{PR}$  – shown in Eq. (6) – that is only valid for  $P(\delta_{\mathbb{a}_q^* \mathbb{a}_q^*} = 1)$  greater than 0.5 ( $\mathbb{E}[f(\theta_r)] > 0.5$ ).

$$P_q^{PR} = P\left(X > \frac{N_q}{2}\right) = 1 - P\left(X \leq \frac{N_q}{2}\right) \geq 1 - e^{-\left(\frac{N_q}{2P(\delta_{\mathbb{a}_q^* \mathbb{a}_q^*} = 1)}\right) \left(P(\delta_{\mathbb{a}_q^* \mathbb{a}_q^*} = 1) - \frac{1}{2}\right)^2} \quad (6)$$

Requiring the success probability of an average respondent to be greater than 0.5 reminds us of Condorcet’s Jury Theorem on voter competence; a very popular theorem in the field of social choice theory despite its limitations which include respondents facing a binary choice. Condorcet, who was an enthusiastic supporter of democracy, believed that having an independent probability of voting for the correct decision greater than  $\frac{1}{2}$  was sufficient for majority

voting to succeed, and that adding more voters will increase the probability of reaching the correct decision (Daniel & Paroush, 1998).

Combining the constraints for low difficulty tasks where  $0 < \Delta_k < 0.5$ :

$$0.5 < \mathbb{E}[f(\theta_r)] \leq \frac{\mu(1 - \Delta_k)}{\mu(1 - \Delta_k) + \Delta_k(1 - \mu)} \Rightarrow 0.5 \leq \frac{\mu(1 - \Delta_k)}{\mu(1 - \Delta_k) + \Delta_k(1 - \mu)} \Rightarrow \mu \geq \Delta_k \quad (7)$$

Combining the constraints for higher difficulty tasks where  $0.5 < \Delta_k < 1$ :

$$\mathbb{E}[f(\theta_r)] \geq \left\{ \frac{\mu(1 - \Delta_k)}{\mu(1 - \Delta_k) + \Delta_k(1 - \mu)} \middle| 0.5 \right\} \Rightarrow \mu \geq \Delta_k \quad (8)$$

To summarize, for aggregation in PR to work, the expected value of the crowd’s ability should be at least equal to the difficulty of the proposed task. This constraint on the ability mean ensures that the expected value of the competence is at least equal to half, which agrees with the assumption of voter competence. If so, the probability of success in PR will have a lower bound and will depend on the number of respondents who attempt to solve the task. As the number of respondents grows in size, the probability of success converges at a faster rate.

In order to compute the average probability of success versus task difficulty in PR, we take the integral of the competence function over the entire range of crowd ability resulting in Eq. (9). The result is in terms of task difficulty.

$$P_{PR} = \int_0^1 \frac{\theta_r(1-\Delta_k)}{\theta_r(1-\Delta_k)+\Delta_k(1-\theta_r)} d\theta_r = F(\Delta_k) = \frac{(\Delta_k-1)(2\Delta_k+2\Delta_k \tanh^{-1}[1-2\Delta_k]-1)}{(1-2\Delta_k)^2} \quad (9)$$

#### 4.2. Confidence-Weighted

In the CF method, the notion is to give higher weights to the replies of respondents who are more confident and lower weights otherwise  $\mathbb{a}_q^{CF} \rightarrow \arg \max_{\mathbb{a} \in \mathbb{A}} \sum_{r \in R_q} y_r^k \cdot \delta_{\mathbb{a}_q^r \mathbb{a}}$  where the weights are the reported confidence values per respondent  $y_r^k \rightarrow a(x_r^k)^2 + b(x_r^k) + 1$  which are modeled based on our proposed DK function.

We compute the expected probability of success in CF by taking the integral of the *confidence-weighted* competence over the range of ability values as shown in Eq. (10). The result is in terms of task difficulty and the coefficients  $a$  and  $b$  of our modeled DK function. The full equation is provided in the [supplementary information](#).

$$P_{CF} = \frac{\int_0^1 \left( a(x_r^k)^2 + b(x_r^k) + 1 \right) (x_r^k) d\theta_r}{\int_0^1 \left( a(x_r^k)^2 + b(x_r^k) + 1 \right) d\theta_r} = F(\Delta_k, a, b) \quad (10)$$

We define a general crowd as one that includes all classes of respondents from the low end of the spectrum (non-experts) to the high end (experts). Assuming we have

a general crowd, and taking into consideration the psychological bias that accompanies respondents, which is represented in the modeled Dunning-Kruger function as a generally decreasing relationship between reported confidence and competence, *confidence-weighted* approaches will in most cases perform worse than the simple *plurality* approach. If the least competent are reporting the highest confidence and the most competent are not, then weighing by the reported confidence values will negatively affect the aggregation outcome.

**Theorem I.** *Given any DK plot, PR will perform better than CF for tasks of higher difficulty ( $\Delta_k \geq 0.5$ ).*

The full proof of Theorem I is found in the [supplementary information](#).

For different samples of the DK plot as shown by the values of the coefficients  $a$  and  $b$  in Fig. 2 we computed the difference between the probability of success of PR and that of CF. As expected, based on the model alone, PR outperforms CF for a wide range of possible DK plots. The difference between the two success probabilities is never below zero. What is interesting however, is the relationship between this difference and the coefficients  $a$  and  $b$ .

As the value of  $a$  increases, the difference is decreasing, which means that CF's performance is improving compared to PR. This agrees with our previous note on how increasing  $a$  has the effect of increasing the confidence faster for respondents of higher competence. The result of increasing  $a$  is an increased confidence value for the more competent.

As the value of  $b$  decreases, we notice a similar pattern; the difference decreases in value for all values of  $a$ . This observation agrees with another previous note on how decreasing  $b$  has the effect of decreasing the confidence fas-

ter for respondents of lower competence. The result of decreasing  $b$  is a decreased confidence value for the less competent.

#### 4.3. Maximum Confidence

In this approach, PR is applied to a subset of the population with *maximum confidence*.

$$\mathbb{a}_q^{MC} = \arg \max_{a \in \mathbb{A}} \sum_{r \in R_q^{MC}} \delta_{a,r} \quad (11)$$

We compute the expected probability of success in MC similarly to that in PR but by taking different integral boundaries, which depend on the number of defined confidence levels. Or, if the confidence is a continuous value, the choice of respondents is those in the top  $\mathcal{P}$  percentage in terms of confidence level. Given a DK plot and a percentage  $\mathcal{P}$ , boundary competence values are computed as follows. A horizontal line based on the selected  $\mathcal{P}$  is drawn on the DK plot. The line will intersect the DK function at two points. The x-values of these two intersection points are the resulting competence values that are used as boundaries for the integral that is used to compute the MC success probability. We define  $x_{min}^{\mathcal{P}}$  and  $x_{max}^{\mathcal{P}}$  as the intersection points between our modeled DK function and the horizontal line drawn at  $y^{\mathcal{P}} = 1 - \frac{\mathcal{P}}{100}$ .

The minimum point on the DK plot is defined as the point where the first derivative equals 0. Let  $\nabla$  denote the point of minimum confidence. Its coordinates are shown below.

$$\nabla = (\nabla_x, \nabla_y) = \left( -\frac{b}{2a}, 1 - \frac{b^2}{4a} \right)$$

Difference between the average success probability of PR and CF across all task difficulty levels

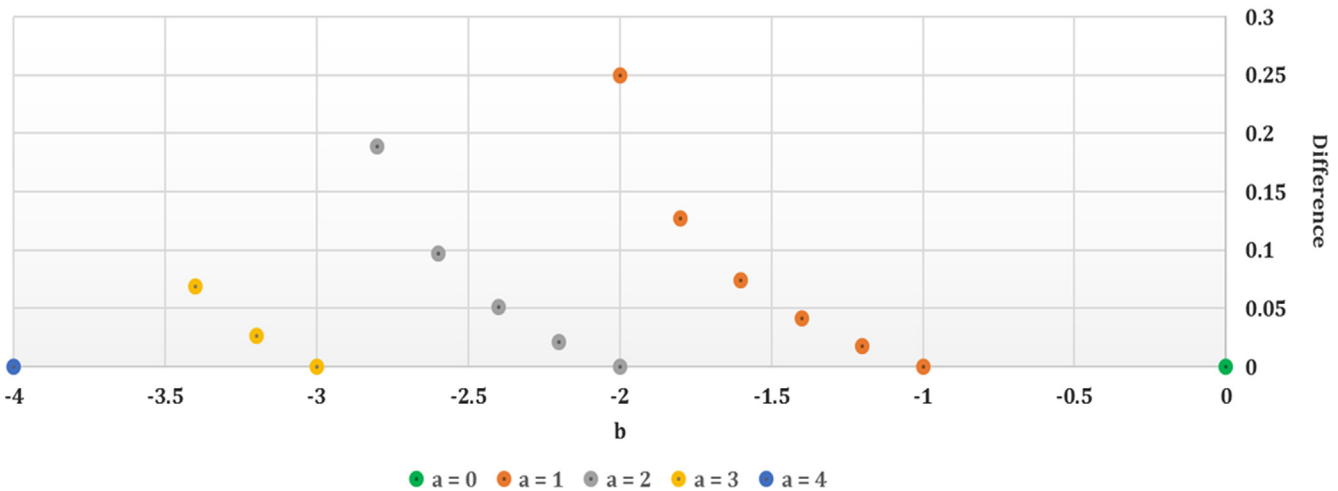


Fig. 2. Difference between the average success probability of *plurality voting* and that of *confidence-weighted* voting across all task difficulty levels and for different samples of the Dunning-Kruger function as shown by the values of  $a$  and  $b$ .

There are two cases to consider; in the first case, the point  $y^{\mathcal{P}} = (1 - \frac{\mathcal{P}}{100}) \leq \nabla_y$ , which means that the percentage covers the entire range of confidence values. In this case, applying MC has the same effect as applying PR and the two boundary points are in fact one point.

$$y^{\mathcal{P}} \leq \nabla_y \Rightarrow x_{min}^{\mathcal{P}} = x_{max}^{\mathcal{P}}$$

In the second case, the point  $y^{\mathcal{P}} = (1 - \frac{\mathcal{P}}{100}) > \nabla_y$ , which means that if we are to apply the MC method, then we should only consider the answers of the respondents whose competence levels are less than  $x_{min}^{\mathcal{P}}$  or greater than  $x_{max}^{\mathcal{P}}$  where  $x_{min}^{\mathcal{P}}$  and  $x_{max}^{\mathcal{P}}$  are computed as shown below.

$$y^{\mathcal{P}} > \nabla \Rightarrow \mathcal{P} < \frac{25b^2}{a} \text{ and } x_{min}^{\mathcal{P}} = \frac{-b - \sqrt{b^2 - \frac{a\mathcal{P}}{25}}}{2a} \text{ and } x_{max}^{\mathcal{P}} = \frac{-b + \sqrt{b^2 - \frac{a\mathcal{P}}{25}}}{2a}$$

We compute the expected probability of success in MC by taking the integral of the competence over the range of ability values as shown in Eq. (12). The result is in terms of the boundary points and the task difficulty. The full equation of  $P_{MC}$  is provided in the [supplementary information](#)

$$P_{MC} = \int_0^{x_{min}^{\mathcal{P}}} \frac{\theta_r(1-\Delta_k)}{\theta_r(1-\Delta_k)+\Delta_k(1-\theta_r)} d\theta_r + \int_{x_{max}^{\mathcal{P}}}^1 \frac{\theta_r(1-\Delta_k)}{\theta_r(1-\Delta_k)+\Delta_k(1-\theta_r)} d\theta_r = F(\Delta_k, x_{min}^{\mathcal{P}}, x_{max}^{\mathcal{P}}) \tag{12}$$

Assuming we have a general crowd and taking into consideration the psychological bias that accompanies respondents, which is represented in the modeled Dunning-Kruger function as a generally decreasing relationship between reported confidence and competence, selecting the most confident replies only will affect the aggregation outcome negatively. If the least competent respondents are reporting the highest confidence values, a *maximum confidence* approach will result in selecting a majority composed of the least competent individuals in the crowd. The outcome in MC is affected more severely than that in CF. In this method, we are considering a sub-crowd of respondents who were the most extreme when reporting their confidence levels. The MC method focuses on the least competent respondents and disregards respondents whose reported confidence values were moderate, which we argue are the majority of the most competent respondents. Accordingly, the higher the value of  $\mathcal{P}$ , the better the performance of MC when compared to CF and PR. For a value of  $\mathcal{P} = 100$ , the MC approach is the same as PR voting.

**Theorem II.** *Given any DK plot, any task difficulty, and any value of the percentage  $\mathcal{P}$ , PR performs better than MC.*

**Theorem III.** *The probability function based on MC is increasing in  $\mathcal{P}$ .*

**Theorem IV.** *At  $\Delta_k = 0.5$ , CF performs better than MC for all values of  $\mathcal{P} \leq \frac{25b^2}{a} - 10$ . As  $\mathcal{P}$  decreases, the point at which MC outperforms CF shifts further to the right at higher difficulty levels.*

Proofs of Theorems II, III, and IV are found in the [supplementary information](#).

We show a sample DK plot for a general crowd where  $a = 2$  and  $b = -2.5$  in the top right corner of [Fig. 3](#). Using this DK function, we plot the success probabilities of PR, CF, and MC for three values of  $\mathcal{P}$  (20, 30, and 75) and for values of task difficulty ranging from 0.1 to 1. These plots are based on our modeled values of PR, CF, and MC.

A first look at the figure shows how PR outperforms all other methods for difficulty levels higher than 0.5. However, for smaller values of the task difficulty, the performances of PR and CF are comparable. For the smaller values of  $\mathcal{P}$  (20 and 30), the performance of MC is very poor. For a higher value of  $\mathcal{P}$  (75), the performance of MC becomes comparable to that of CF when the task difficulty is around 0.7 and even surpasses it for the harder tasks. However, its performance remains worse than PR throughout.

Note that for the chosen  $a$  and  $b$  of 2 and  $-2.5$ , the percentage limit that was defined in Theorem IV evaluates to  $(\frac{25b^2}{a} - 10) \cong 68$ . Based on Theorem IV, this means that for  $\mathcal{P} = 20$  and  $\mathcal{P} = 30$ , at  $\Delta = 0.5$ ,  $P_{CF} \geq P_{MC}$  which is true. And as  $\mathcal{P}$  decreases,  $P_{CF}$  and  $P_{MC}$  intersect further to the right at values of  $\Delta > 0.5$  which is shown in [Fig. 3](#). At  $\mathcal{P} = 75$ , the two curves intersect at  $\Delta \cong 0.7$  and as  $\mathcal{P}$  decreases to 20% and 30%, the intersection takes place at  $\Delta = 1$ .

#### 4.4. Competence-Weighted

Another weighted approach worth discussing is the *competence-weighted* approach. In CP, rather than weighting the replies by the reported confidence of the respondents, the method weighs them by the estimated competence of the respondents. The real competence of a respondent is an unknown value. Additionally, it is affected by many factors such as his well-being and ability to focus at the time of solving the task. Instead of getting the real competence values of respondents, CP methods try to estimate these values. The literature is filled with findings related to competence detection techniques ([Attiaoui, Martin, & Ben, 2017](#); [Bougoussa, Dumoulin, & Wang, 2008](#); [Pal, Farzan, Konstan, & Kraut, 2011](#); [Welinder & Perona, 2010](#); [Whitehill, Wu, Bergsma, Movellan, & Ruvolo, 2009](#); [Zhang, Ackerman, & Adamic, 2007](#)). Some techniques are a function of the time it takes a respondent to solve a task ([Kyllonen & Zu, 2016](#)). Some are based on the performance of a respondent in previous tasks

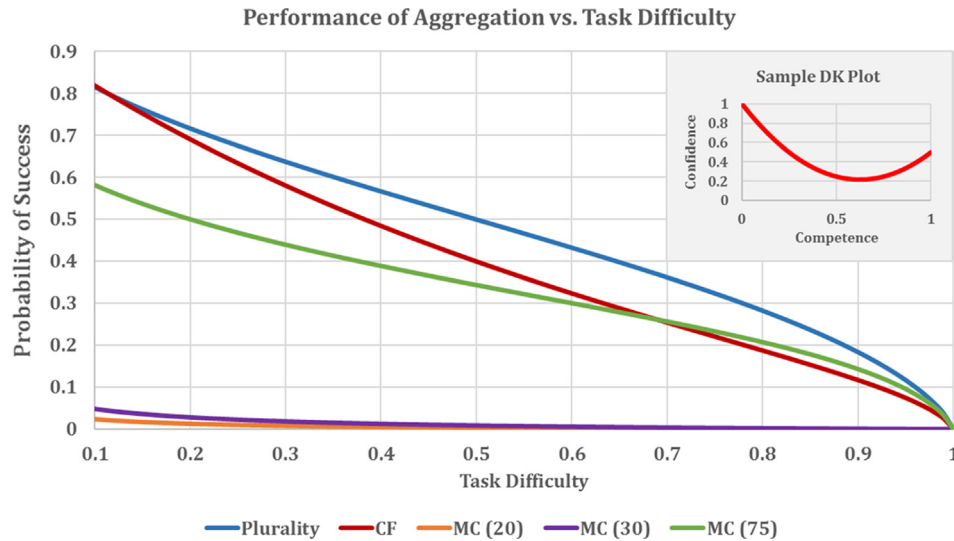


Fig. 3. Success probabilities of plurality, *confidence-weighted*, and three *maximum confidence* approaches with different percentage values (20, 30, and 75). The used Dunning-Kruger function is shown in the top right corner ( $y = 2x^2 - 2.5x + 1$ ).

(Lee, Steyvers, De Young, & Miller, 2012). A wide variety of techniques even make use of Bayes' theorem to try to detect the competence of respondents when the ground truth of the proposed questions is not available a priori (Bachrach, Graepel, Minka, & Guiver, 2012; Lakshminarayanan & Whye Teh, 2013; Raykar et al., 2009). Even though our focus in this work is not on the deployed competence detection method, we stress that it is one of the most important steps when designing a crowd-sourcing platform. Many techniques have been derived with the aim to estimate workers' competence and reliability when attempting to solve a task. The better the detection technique, the higher the chances of reaching a correct aggregate answer. In fact, detecting the competence or reliability of workers is one of the many techniques of quality control that are employed by the system to improve performance. We provided a survey of quality control mechanisms in the literature review section where we showed that estimating the real competence of respondents is not a trivial task. If the estimated and the real values of the competence are very far off indicating a primitive detection method, then naturally, a *competence-weighted* aggregation approach will perform poorly.

Instead of focusing on detecting the competence, we make use of our modeled DK function to derive the competence of respondents based on their reported confidence values. Getting the reported confidence from a respondent is a very easy task. And this reported confidence value is an actual value rather than an estimated one. Assuming a general crowd, we can give lower competence values for highly confident respondents and higher competence values for the moderately confident ones. In other words, we can use our model of the DK function as a competence detection technique. The performance of this detection technique can be evaluated based on the performance of the *competence-weighted* aggregation method when using the

estimated competence values derived from our model of the DK function.

**Hypothesis I.** Given a general crowd, *competence-weighted* approaches using competence values estimated from the modeled DK function will outperform approaches based on *plurality voting*.

We argue that the choice of the DK function coefficients does not significantly affect the performance of the competence detection method. This, of course, only holds in the case of a general crowd where the overall confidence level decreases as the competence increases and where more competent individuals are never as confident as the least competent ones.

#### 4.5. The Irregular Crowd

If we do not consider a general crowd, the relationship between confidence and competence will no longer abide by the characteristics of the DK plot. For example, if we have a crowd of experts in the field, a sample confidence-competence relationship might be a straight line starting from 0 with a slope of 1. This is, by far, the best representation of confidence versus competence since the crowd consists of experts who have the ability to correctly estimate their cognitive skills on either side of the spectrum. In this case, CF approaches will perform better than simple PR voting due to the fact that the reported confidence values are not biased.

In fact, this is illustrated in one of the experiments by Prelec et al. who describe a new and interesting aggregation method commonly referred to as the *Surprisingly Popular* algorithm (Prelec et al., 2017). The authors argue that this algorithm can help extract the correct answer from a crowd even when the majority replies incorrectly. To accomplish this, participants are queried twice. They are asked to

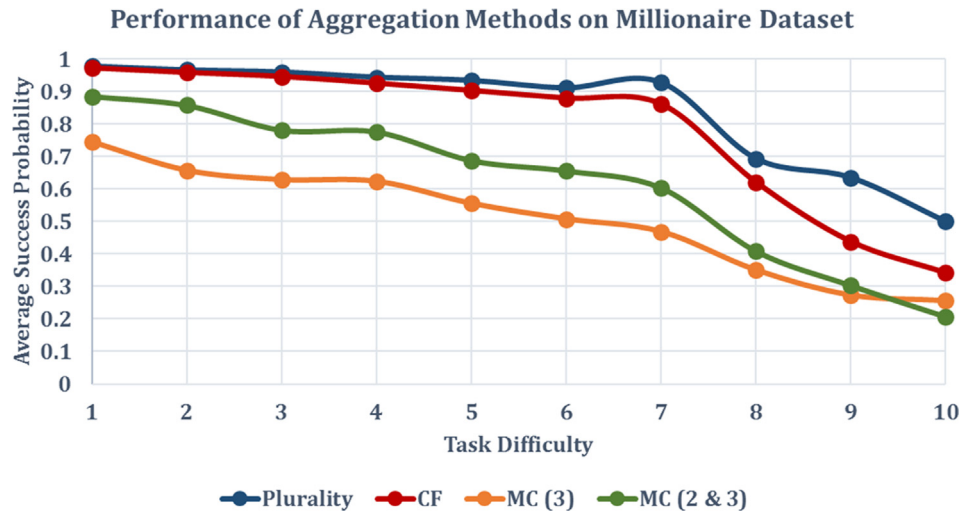


Fig. 4. Success probabilities of different aggregation methods based on data from the “Who Wants to Be a Millionaire?” dataset. In the case of MC, we considered two scenarios, one with confidence level 3 only and one with confidence levels 2 and 3.

answer the question and to predict what the majority will answer as well. The “correct” answer is based on the variation between the given answers and the predicted ones. In their paper, they describe how they performed the same crowdsourcing experiment twice. The experiment asked respondents to judge the price of 90 reproductions of modern 20th century artwork. In the first part of the experiment, the crowd was composed of people working with art in galleries or museums. In the second part, the crowd was composed of MIT master’s and doctoral students who have not taken any courses on art or art history. What is interesting is that both the *confidence-weighted* and the *maximum confidence* approaches outperformed the *majority vote* in the case of the art professionals. On the other hand, they both performed worse than the *majority vote* in the case of the MIT students who fit the description of a general crowd. The *maximum confidence* in the second part even performed worse than the *confidence-weighted* approach, which agrees with our results in this paper. The *surprisingly popular* algorithm has received considerable attention recently and is already referenced in works related to voting and crowdsourcing (Bang & Frith, 2017; Laan, Madirolas, & Polavieja, 2017). We did not include this algorithm in our study due to the lack of data. However, comparing this algorithm to other aggregation techniques and studying the effect of peoples’ psychological biases on its performance is part of our future work.

## 5. Numerical Analysis

To validate our model of the different aggregation techniques, we performed numerical analyses on crowdsourced data. We begin by describing the dataset that was used. Then, we discuss the two main experiments that we performed and the results that were drawn from each experiment.

### 5.1. Dataset

Aydin et al. developed and deployed a crowdsourcing system for playing the popular “Who Wants to Be a Millionaire?” television game show (Aydin, Yilmaz, & Demirbas, 2017). They created an Android app that allowed respondents to play the game at the same time as the show was being broadcasted. The app was downloaded over 300,000 times over a period of 9 months. The data in it is based on 1908 live game show questions from a total of 80 broadcasted episodes. Respondents gave a total of 214,658 answers to these questions. Every game show had an average respondent count of 733. The average number of answers per question was around 100. On the server side, project administrators instantly type the questions and answers as they appear on the game show. On the respondent side, every participant selects his answer to each question along with his confidence for every submitted answer. The confidence values to choose from are 1 for “no idea”, 2 for “guessing”, and 3 for “certain”. Every game has 12 questions ranked by increasing difficulty. Questions 11 and 12 are the most difficult questions. Consequently, the number of times these questions were asked was of no statistical significance and so we do not include them in our analysis. Taking into consideration the large number of respondents who are fans of the “Who Wants to Be a Millionaire?” game show and who can lie anywhere on the spectrum of respondent competence, we assume a general crowd in this case.

### 5.2. Results and Discussion

In the first experiment, we started by ranking the questions based on their difficulty level. The aggregated answer for every question in the case of PR was the most voted for choice. In the case of CF, the aggregated answer was the

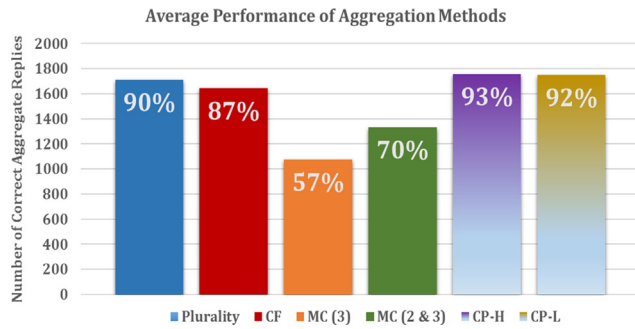


Fig. 5. Performance of different aggregation methods based on data from the “Who Wants to Be a Millionaire?” dataset. CP-L and CP-H are *competence-weighted* methods where the lower and higher competence values were selected respectively.

most voted for choice after weighing all the choices by their respective confidence values. In the case of MC, we considered two cases. In the first case, we applied *plurality voting* on the subset of respondents whose reported confidence was 3. In the second case, the subset of respondents included those whose reported confidence was either 2 or 3. Then, we compared all the aggregated answers to the gold label of every question. A question whose aggregated answer matches with the gold label gets a score of 1, otherwise, it gets a score 0. After getting the total scores for every method, we computed the average score per difficulty level and plotted the results, as shown in Fig. 4.

The first observation is related to PR. It outperforms the *confidence-weighted* approach ( $p\text{-value} < 0.0005$ ) and the *maximum confidence* approach ( $p\text{-value} < 0.0001$ ). This validates Theorem I which states that plurality outperforms *confidence-weighted* approaches for tasks with difficulty higher than 0.5. It also validates Theorem II which states that PR will always outperform MC aggregation approaches. Also note that for the easier tasks, the performance of CF is better than that of *plurality voting*. The point at which PR outperforms CF depends on the values of the coefficients  $a$  and  $b$  of the DK plot. These results agree with the conclusions of Li et al. who study an M-ary classification task via crowdsourcing where the workers report quantized confidence scores (Li & Varshney, 2017). They consider a crowd that is only composed of honest workers who answer questions and report confidence in good faith. Their simulation results demonstrate how the performance of the crowdsourcing task does not improve when incorporating workers’ confidence scores.

The second observation is that  $P_{MC}$  decreases with the decrease in  $\mathcal{P}$  except for the point at question 10, the reason of which could be due to data that is insignificant for this question. We notice that when only replies with confidence level 3 were considered, the probability of success was lower than that when confidence levels 2 and 3 were considered ( $p\text{-value} < 0.0001$ ). This validates Theorem III. CF outperforms both instances of MC regardless of the task difficulty. This does not contradict with Theorem IV

since there is no intersection whatsoever. The theorem would be contradicted had there been an intersection at a relatively small difficulty level.

The aim of the second experiment was to test our competence detection idea based on our modeled DK function. We selected three random DK functions ( $a = 2$  and  $b = -2.75$  |  $a = 2.55$  and  $b = -3.15$  |  $a = 3$  and  $b = -3.45$ ). We derived the competence value for every user from his reported confidence using each of the three DK functions. For some confidence values however, there are two possible competence values considering the general shape of our modeled DK plot. For these cases, we generated two sets of weights. In the first set, we selected the lower competence value (CP-L) and in the second, we selected the higher one (CP-H). We then weighed the answers using the derived competence values and computed the scores of CP-L and CP-H after comparing each set of aggregated answers to the gold label and taking an average score over the 3 DK functions. We present the results in terms of a bar chart in Fig. 5. Inside the bars, we show the percentage of questions out of the total questions that were correctly aggregated by each method.

Our first observation is that, as in the previous experiment, MC (3) performed worse than MC (2 & 3) and both performed worse than CF which in turn lagged behind PR. What is interesting in this experiment, however, is that both CP-L and CP-H performed better than PR ( $p\text{-value} < 0.01$ ). This indicates that the competence values that we derived using our DK model were in fact good estimates of the real competence values. This validates Hypothesis I above, which states that given a general crowd, *competence-weighted* approaches using competence values estimated from the modeled DK function will outperform approaches based on PR.

Another interesting observation is that CP-H performed slightly better than CP-L, which is expected. For high values of the reported confidence, there is only one value of the competence based on the general shape of the DK function. It is for the lower confidence values that we encounter two competence values, a low one and a high one. One of the main conclusions given by Dunning and Kruger is that the more competent individuals are those who report lower confidence values as they generally underestimate their competence. This means that giving a higher competence value to these low-confidence respondents will result in a better overall competence estimation and therefore better aggregation performance.

It is worth noting that the CP scores before averaging across the three different DK functions were very close and surpassed all other approaches. This validates how the choice of the DK coefficients  $a$  and  $b$  does not significantly affect the performance of the competence detection method. The chosen DK plots in this example revealed a pattern where the most incompetent respondents showed the greatest miscalibration in assessing their skills.

### 5.3. Comparative Analysis

Qi et al. discuss the issue of the long-tail phenomenon in crowdsourcing tasks where most workers only provide answers to a few tasks and only a few workers provide answers to plenty of tasks (Li, Li, Gao, Su, et al., 2014). They argue that existing crowdsourcing approaches clearly overlook this phenomenon which causes problems in estimating worker reliability. They propose to consider both the estimate of worker reliability along with its confidence interval in order to accurately reflect reliability levels of workers with different degrees of participation. This results in reducing the effect of less active workers who do not solve many tasks. They perform experiments on four real world crowdsourcing tasks, one of which is the previously described WWTBAM dataset. Results demonstrate how their proposed Confidence-Aware Truth Discovery (CATD) method, which takes into consideration the long-tail phenomenon, outperforms existing approaches (Dong, Saha, & Srivastava, 2013; Galland, Abiteboul, Marian, & Senellart, 2010; Li, Li, Gao, Zhao, et al., 2014; Pasternack & Roth, 2010; Yin, Han, & Yu, 2008).

More recently, Fenglong et al. addressed the issue of topic diversity in crowdsourcing systems (Ma et al., 2015). They argue that most existing systems assume that a worker will have the same reliability when answering any question. Existing approaches ignore the fact that a worker's reliability may vary significantly depending on the topic. To this end, they propose FaitCrowd which probabilistically models question content and answer generation in an attempt to assign topics to questions, learn the ground truth, and estimate workers' topic-specific expertise simultaneously. They test their method on real world datasets including WWTBAM. Their results demonstrate how FaitCrowd reduces the error rate when compared with other approaches (Dawid & Skene, 1979; Demartini, Difallah, & Cudré-Mauroux, 2012; Dong et al., 2013; Galland et al., 2010; Li, Li, Gao, Su, et al., 2014; Li, Li, Gao, Zhao, et al., 2014; Pasternack & Roth, 2010; Yin et al., 2008).

Table 1 compares our *competence-weighted* aggregation technique (CWAT) to CATD and FaitCrowd, among other methods, in terms of the Error Rate. We computed the error rate per task difficulty as well as the overall error rate. The highlighted cells are the ones that have a lower error rate than our proposed method. CWAT outperforms TruthFinder and Investment across all difficulty levels. For the more challenging tasks, it outperforms all methods except CATD and FaitCrowd. The overall performance of CATD and FaitCrowd is better than that of our method except in tasks of difficulty levels 6 and 7 where our method performs better. It is worth mentioning that the error rates of the other methods are based on the results presented by Qi and Fenglong in their papers and are not based on our own simulations of these methods.

Even though the performance of CATD was better than ours given the WWTBAM dataset, it suffers from a major drawback. The approach by Qi et al. offers no advantage over existing approaches in crowdsourcing tasks where the long-tail phenomenon is not present. Even though this phenomenon should not be overlooked when estimating worker reliability, it is not always present in crowdsourcing systems. For example, workers might not have the option of solving a subset of the questions even if they do not know the answer to some questions. Moreover, there are cases where there is only one question that workers attempt to answer. These are two crowdsourcing examples where the long-tail phenomenon does not exist and will have no effect on the overall performance of the system. This renders other methods such as ours, that work regardless of the participation level, to be more useful for better system performance. Of course, in specific cases where the long-tail phenomenon is present, the estimation of worker reliability will be affected by the total number of claims that he or she makes. A worker that makes one claim to one task will either be categorized as extremely reliable or highly unreliable based on the correctness of his single claim. We believe that this negatively affects worker reliability estimation as well as confidence reporting. Addressing this phenomenon in both reliability detection and confidence reporting is a

Table 1

Error rate from our competence-detection approach compared with approaches from the literature where the columns represent increasing task difficulty levels.

Level	1	2	3	4	5	6	7	8	9	10	Overall
<b>CWAT</b>	<b>0.0161</b>	<b>0.0341</b>	<b>0.0346</b>	<b>0.0434</b>	<b>0.0442</b>	<b>0.0585</b>	<b>0.0343</b>	<b>0.2124</b>	<b>0.2472</b>	<b>0.3</b>	<b>0.0641</b>
CATD [68]	0.0132	0.0271	0.0276	0.029	0.0435	0.0596	0.0481	0.1304	0.1414	0.2045	0.0485
FaitCrowd [74]	0.0132	0.0271	0.0241	0.0254	0.0395	0.055	0.0481	0.087	0.101	0.1136	0.0399
TruthFinder [69]	0.0693	0.0915	0.1241	0.0942	0.1581	0.2294	0.2674	0.3913	0.5455	0.5455	0.1816
AccuSim [70]	0.0264	0.0305	0.0345	0.0507	0.0632	0.0963	0.0909	0.2826	0.3636	0.5	0.0913
Investment [71]	0.033	0.0407	0.0586	0.0761	0.087	0.1239	0.1283	0.3406	0.3838	0.5455	0.1151
3-Estimates [72]	0.0264	0.0305	0.031	0.0507	0.0672	0.1055	0.0963	0.2971	0.3737	0.5	0.0942
CRH [73]	0.0264	0.0271	0.0345	0.0435	0.0593	0.0872	0.0856	0.2609	0.3535	0.4545	0.0866
D&S [22]	0.0297	0.0305	0.0483	0.0507	0.0672	0.1101	0.0963	0.2971	0.3636	0.5227	0.0975
ZenCrowd [75]	0.033	0.0305	0.0345	0.0471	0.0593	0.0872	0.0856	0.2754	0.3636	0.5227	0.0899

necessary step that we believe will result in error rates lower than both our approach and CATD.

FaitCrowd by Fenglong et al. suffers from a similar drawback. Even though the approach detects per-topic expertise of workers for better system performance, the assumption of different topics in proposed questions is not always valid. There are many examples where all proposed questions fall into the same area (e.g. technology, science, art, fashion, etc.). When this is the case, FaitCrowd will show no advantage over other approaches. This also applies in cases where the crowd is voting for the best answer for a single question. Even though our overall error rate is slightly higher than that of FaitCrowd, our method is more versatile in that it considers all types of crowds and any combination of question topics. FaitCrowd is superior only in cases where the topics of proposed questions vary which is not always the case.

## 6. Concluding Remarks

There are many factors that affect the choice of the aggregation method to adopt in crowdsourcing. Our conclusion is that the type of crowd is the most important of these factors. In the case of an incompetent crowd, no aggregation method will have a good performance. This includes simple methods such as *plurality voting*, which requires an average respondent to have a success probability above 0.5 in accordance with Condorcet's jury theorem of voter competence. When knowledge within a crowd surpasses a given threshold, the choice of aggregation technique matters as we showed in this work.

The main finding in this paper is related to the Dunning-Kruger effect. We presented a formal model of this psychological bias in terms of a quadratic relationship between respondent competence and confidence. Constraints related to our modeled function were derived from the characteristics of this effect as described by Dunning and Kruger.

Using our model of the Dunning-Kruger effect as the core of our work, we then went on to model the performance of different aggregation methods. Due to the general shape of the DK function, we were able to validate that for a general crowd of respondents, a plurality aggregation method will in most cases outperform methods based on the confidence of respondents such as the *confidence-weighted* and *maximum confidence* approaches. We also showed that the *maximum confidence* approach generally lags behind the *confidence-weighted* approach since it focuses on the least competent respondents and disregards respondents whose reported confidence values were moderate, which we argue are the majority of the most competent respondents. We also showed how the size of the subgroup of respondents in the *maximum confidence* approach affects the overall performance of the method for a general crowd.

We then went on to use the modeled DK function as a competence detection technique. There are many detection

techniques in the literature that attempt to estimate the competence of respondents in a crowdsourcing system with or without the presence of the ground truth. These techniques are complicated and in some cases, they do not give a good estimate of the real competence values of respondents. Our argument is that it is fairly easy to get the reported confidence of respondents when solving a task and we showed how we were able to use these reported values to estimate, to a good degree, the competence of respondents. We have shown that our *competence-weighted* approach outperforms previous approaches across most task difficulty levels and for the more challenging tasks, it outperforms all methods except the ones proposed by Li, Li, Gao, Su, et al. (2014) and Ma et al. (2015) where their methods' overall performances were better than ours for most levels of task difficulty. However, considering that our method is more versatile in terms of the types of crowds and question topics, we believe that an aggregation technique that aims to estimate topic-specific worker expertise (when applicable) taking into consideration the effects of the long-tail phenomenon (if present) and the cognitive biases of the workers involved in the crowdsourcing process should improve the truth discovery process considerably.

According to Dunning and Kruger, perhaps the best illustration of inflated self-appraisals of the incompetent is the tendency of the average person to rate her skills as above average, which defies the logic of statistics. On the other hand, the most competent individuals usually suffer from the false-consensus effect (Ross, Greene, & House, 1977) where they assume that because they have performed well on a task, then others must have performed well likewise, which results in these individuals underestimating their relative abilities. These two findings validate the general shape of the DK function that we adopted in our model. Dunning and Kruger also observed that it was the most incompetent individuals who showed the greatest miscalibration in assessing their skills, which validates our choice of having the DK function start at the point (0, 1) indicating the highest confidence miscalibration for the least competent.

In conclusion, this work will pave the way for better understanding of the psychological biases that accompany metacognitive skills of respondents in crowdsourcing systems and eventually lead to a more productive utilization of the wisdom of the crowd.

## Conflict of interest

The authors declared that there is no conflict of interest.

## Acknowledgements

This research is supported by TELUS Corp., Canada, the National Council for Scientific Research, Lebanon, and the University Research Board, AUB, Lebanon.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cogsys.2019.04.004>.

## References

- Allahbakhsh, M., Benattallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E., & Dustdar, S. (2013). Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, 17(2), 76–81.
- Attiaoui, D., Martin, A., & Ben Yaghlane, B. (2017). Belief measure of expertise for experts detection in question answering communities: Case study stack overflow. *Procedia Computer Science*, 112, 622–631.
- Aydin, B. I., Yilmaz, Y. S., & Demirbas, M. (2017). A crowdsourced ‘Who wants to be a millionaire?’ player. *Concurrency and Computation: Practice and Experience*.
- Bachrach, Y., Graepel, T., Minka, T., & Guiver, J. (2012). How to grade a test without knowing the answers – A Bayesian graphical model for adaptive crowdsourcing and aptitude testing. *Proceedings of the 29th international conference on machine learning, ICML 2012, vol. 2*.
- Bang, D., & Frith, C. D. (2017). Making better decisions in groups. *Royal Society Open Science*, 4(8).
- Barbier, G., Zafarani, R., Gao, H., Fung, G., & Liu, H. (Sep. 2012). Maximizing benefits from crowdsourced data. *Computational and Mathematical Organization Theory*, 18(3), 257–279.
- Basu Roy, S., Lykourantzou, I., Thirumuruganathan, S., Amer-Yahia, S., & Das, G. (2013). Crowds, not drones, modeling human factors in interactive crowdsourcing. In *DBCrowd 2013 – VLDB workshop on databases and crowdsourcing* (pp. 39–42).
- Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., ... Panovich, K. (2010). Soylent: A word processor with a crowd inside. In *Proceedings of the 23rd annual ACM symposium on user interface software and technology* (pp. 313–322).
- Bonabeau, E. (2009). Decisions 2.0: The power of collective intelligence. *MIT Sloan Management Review*, 50, 45–52.
- Bougoussa, M., Dumoulin, B., & Wang, S. (2008). Identifying authoritative actors in question-answering forums – The case of Yahoo! answers. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2008)* (pp. 866–874).
- Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon’s mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(Feb.), 3–5.
- Callison-Burch, C., & Dredze, M. (2010). Creating speech and language data with Amazon’s mechanical turk. In *Proc NAACL HLT 2010 Work Creat Speech Lang Data with Amaz Mech Turk* (pp. 1–12).
- Carter, L. B. (1986). *The quiet Athenian*. Clarendon Press.
- Chowdhury, S. A., Calvo, M., Ghosh, A., Stepanov, E. A., Bayer, A. O., Riccardi, G., ... Sanchis, E. (2015). Selection and aggregation techniques for crowdsourced semantic annotation task. *INTER-SPEECH*.
- Chowdhury, S., Ghosh, A., Stepanov, E., Orkan Bayer, A., Riccardi, G., & Klasinas, I. (2014). Cross-language transfer of semantic annotation via targeted crowdsourcing. *Proc Annu Conf Int Speech Commun Assoc INTERSPEECH*.
- Conitzer, V., & Sandholm, T. (2012). Common voting rules as maximum likelihood estimators. *CoRR, abs/1207.1*.
- CrowdFlower. Available: <<https://www.figure-eight.com/>> [accessed: 18-Aug-2018].
- Daniel, Berend, & Paroush, J. (1998). When is Condorcet’s jury theorem valid? *Social Choice and Welfare*, 15(4), 481–488.
- Darwin, C. (1871). *The descent of man*. John Murray.
- Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1), 20–28.
- Demartini, G., Difallah, D. E., & Cudré-Mauroux, P. (2012). ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on world wide web* (pp. 469–478).
- Dong, X. L., Saha, B., & Srivastava, D. (2013). Less is more: Selecting sources wisely for integration. In *Proceedings of the 39th international conference on very large data bases* (pp. 37–48).
- Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. In *Proceedings of the 10th international conference on world wide web* (pp. 613–622).
- Eickhoff, C. (2018). Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 162–170).
- Faltings, B., Jurca, R., Pu, P., & Tran, B. D. (2014). Incentives to counter bias in human computation. *HCOMP*.
- Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). *1997 IEEE workshop on automatic speech recognition and understanding proceedings*, 347–354.
- Fleischmann, M., Amirpur, M., Benlian, A., & Hess, T. (2014). Cognitive biases in information systems research: A scientometric analysis. *ECIS*.
- Gadiraju, U., Fetahu, B., Kawase, R., Siehndel, P., & Dietze, S. (2017). Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction*, 24(4), p. 30 1–30: 26.
- Galland, A., Abiteboul, S., Marian, A., & Senellart, P. (2010). Corroborating information from disagreeing views. In *Proceedings of the third ACM international conference on web search and data mining* (pp. 131–140).
- Guazzini, A., Vilone, D., Donati, C., Nardi, A., & Levnajić, Z. (2015). Modeling crowdsourcing as collective problem solving. *Scientific Reports*, 5(1).
- Hamilton, W. D. (1970). Selfish and spiteful behaviour in an evolutionary model. *Nature*, 228(5277), 1218–1220.
- Hansen, M. H. (1983). *The athenian ecclesia*. Museum Tusulanum Press.
- Howe, J. (2006). The rise of crowdsourcing Conde Nast. *Wired*.
- Ipeirotis, P. G., Provost, F., & Wang, J. (2010). Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 64–67).
- Kim, Y.-H., Chiu, C.-Y., & Zou, Z. (2010). Know thyself: Misperceptions of actual performance undermine achievement motivation, future performance, and subjective well-being. *Journal of Personality and Social Psychology*, 99(3), 395–409.
- Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 453–456).
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing ones own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.
- Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, 4(4).
- Laan, A., Madirolas, G., & Polavieja, G. (2017). Rescuing collective wisdom when the average group opinion is wrong. *Frontiers in Robotics and AI*, 4(Nov.), 1–21.
- Lakhani, K. R., Garvin, D. A., & Lonstein, E. (2010). *TopCoder (A): Developing software through crowdsourcing*. Harvard Bus. Sch..
- Lakshminarayanan, B., Whye Teh, Y. (2013). Inferring ground truth from multi-annotator ordinal data: A probabilistic approach.
- Law, E., & von Ahn, L. (2009). Input-agreement: A new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1197–1206).
- Lee, M. D., Steyvers, M., De Young, M., & Miller, B. (2012). Inferring expertise in knowledge and prediction ranking tasks. *Topics in Cognitive Science*, 4(1), 151–163.
- Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., ... Han, J. (2014). A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment*, 8(4), 425–436.

- Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., & Han, J. (2014). Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on management of data* (pp. 1187–1198).
- Li, Q., & Varshney, P. K. (2017). Does confidence reporting from the crowd benefit crowdsourcing performance? *CoRR, abs/1704.0*.
- Liebrand, W. B., Messick, D. M., & Wolters, F. J. (1986). Why we are fairer than others: A cross-cultural replication and extension. *Journal of Experimental Social Psychology, 22*(6), 590–604.
- Lijphart, A. (1991). Constitutional choices for new democracies. *Journal of Democracy, 2*(1), 72–84.
- Ma, F., Li, Y., Li, Q., Qiu, M., Gao, J., Zhi, S., ... Han, J. (2015). FaitCrowd: Fine grained truth discovery for crowdsourced data aggregation. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 745–754).
- McCoy, J., Prelec, D. (2017). A statistical model for aggregating judgments by incorporating peer predictions. ArXiv e-prints.
- McCurdy, K., Crowdsourcing & iStockPhoto. DG Design Network.
- Pal, A., Farzan, R., Konstan, J. A., & Kraut, R. E. (2011). Early detection of potential experts in question answering communities. In *Proceedings of the 19th international conference on user modeling, adaption, and personalization (UMAP 2011)* (pp. 231–242).
- Park, Y. J., & Santos-Pinto, L. (2010). Overconfidence in tournaments: Evidence from the field. *Theory Decis., 69*(1), 143–166.
- Pasternack, J., & Roth, D. (2010). Knowing what to believe (when you already know something). In *Proceedings of the 23rd international conference on computational linguistics* (pp. 877–885).
- Poundstone, W. (2017). The Dunning-Kruger president. *Psychology today*.
- Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature, 532*–541.
- Quinn, A. J., & Bederson, B. B. (2011). Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1403–1412).
- Quoc Viet Hung, N., Tam, N. T., Tran, L. N., & Aberer, K. (2013). An evaluation of aggregation techniques in crowdsourcing. In *Web Information Systems Engineering – WISE 2013* (pp. 1–15).
- Rasch, G. (1961). *On general laws and the meaning of measurement in psychology*. Danmarks pædagogiske Institut.
- Raykar, V. C., Yu, S., Zhao, L. H., Jerebko, A., Florin, C., Valadez, G. H., ... Moy, L. (2009). Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *Proceedings of the 26th annual international conference on machine learning* (pp. 889–896).
- Ross, L., Greene, D., & House, P. (1977). The ‘false consensus effect’: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*(3), 279–301.
- Roy, M. M., & Liersch, M. J. (2013). I am a better driver than you think: Examining self-enhancement for driving ability. *Journal of Applied Social Psychology, 43*(8), 1648–1659.
- Saxonhouse, A. W. (1993). Athenian democracy: Modern mythmakers and ancient theorists. *American Political Science Association, 26*(3), 486–490.
- Schall, D., Skopik, F., & Dustdar, S. (2012). Expert discovery and interactions in mixed service-oriented systems. *IEEE Transactions on Services Computing, 5*(2), 233–245.
- Singh, A. K. (2014). Innocentive for crowdsourcing. *International Journal of Advanced Research in Computer Science Technology, 2*(2), 303–305.
- Sorokin, A., & Forsyth, D. (2008). Utility data annotation with Amazon Mechanical Turk. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1–8).
- Spikins, P., Wright, B., & Hodgson, D. (2016). Are there alternative adaptive strategies to human pro-sociality? The role of collaborative morality in the emergence of personality variation and autistic traits. *The Journal of Archaeology, Consciousness and Culture, 9*(4), 289–313.
- Stepanov, E. A., Chowdhury, S. A., Bayer, A. O., Ghosh, A., Klasinas, I., Calvo, M., ... Riccardi, G. (2018). Cross-language transfer of semantic annotation via targeted crowdsourcing: Task design and evaluation. *Language Resources and Evaluation, 52*(1), 341–364.
- von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 319–326).
- von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM, 51*(8), 58–67.
- von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-based character recognition via web security measures. *Science (80-.), 321*(5895), 1465–1468.
- Welinder, P., & Perona, P. (2010). Online crowdsourcing: Rating annotators and obtaining cost-effective labels. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 25–32).
- Whitehill, J., Wu, T., Bergsma, J., Movellan, J. R., & Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems* (Vol. 22, pp. 2035–2043). Curran Associates, Inc.
- Wikipedia. [Online]. Available: <<https://www.wikipedia.org/>>.
- Yin, X., Han, J., & Yu, P. S. (2008). Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering, 20*(6), 796–808.
- Zhang, J., Ackerman, M. S., & Adamic, L. (2007). Expertise networks in online communities: Structure and algorithms. In *Proceedings of the 16th international conference on world wide web* (pp. 221–230).