

# Does Confidence Reporting from the Crowd Benefit Crowdsourcing Performance?

Qunwei Li

Department of EECS, Syracuse University  
NY 13210  
qli33@syr.edu

Pramod K. Varshney

Department of EECS, Syracuse University  
NY 13210  
varshney@syr.edu

## ABSTRACT

We explore the design of an effective crowdsourcing system for an  $M$ -ary classification task. Crowd workers complete simple binary microtasks whose results are aggregated to give the final classification decision. We consider the scenario where the workers have a reject option so that they are allowed to skip microtasks when they are unable to or choose not to respond to binary microtasks. Additionally, the workers report quantized confidence levels when they are able to submit definitive answers. We present an aggregation approach using a weighted majority voting rule, where each worker's response is assigned an optimized weight to maximize crowd's classification performance. We obtain a counterintuitive result that the classification performance does not benefit from workers reporting quantized confidence. Therefore, the crowdsourcing system designer should employ the reject option without requiring confidence reporting.

## CCS CONCEPTS

•Human-centered computing →Social network analysis;

## KEYWORDS

Classification, crowdsourcing, distributed inference, information fusion, reject option, confidence reporting

## ACM Reference format:

Qunwei Li and Pramod K. Varshney. 2016. Does Confidence Reporting from the Crowd Benefit Crowdsourcing Performance?. In *Proceedings of SocialSens'2017, Pittsburgh, PA, USA, Apr. 21 2017*, 6 pages. DOI: {<http://dx.doi.org/10.1145/3055601.3055607>}

## 1 INTRODUCTION

Crowdsourcing provides a new framework to utilize distributed human wisdom to solve problems that machines cannot perform well, like handwriting recognition, paraphrase acquisition, audio transcription, and photo tagging [2, 5, 18]. Despite the successful applications of crowdsourcing, the relatively low quality of output is a key challenge [1, 8, 16].

Several methods have been proposed to deal with the aforementioned problems [7, 10, 19, 21, 24–27]. A crowdsourcing task

is decomposed into microtasks that are easy for an individual to accomplish, and these microtasks could be as simple as binary distinctions [10]. A classification problem with crowdsourcing, where taxonomy and dichotomous keys are used to design binary questions, is considered in [25]. In our research group, we employed binary questions and studied the use of error-control codes and decoding algorithms to design crowdsourcing systems for reliable classification [24, 25]. A group control mechanism where the reputation of the workers is taken into consideration to partition the crowd into groups is presented in [19, 27]. Group control and majority voting are compared in [7], which reports that majority voting is more cost-effective on less complex tasks.

In past work on classification via crowdsourcing, crowd workers were required to provide a definitive yes/no response to binary microtasks. Crowd workers may be unable to answer questions for a variety of reasons such as lack of expertise. As an example, in mismatched speech transcription, i.e., transcription by workers who do not know the language, workers may not be able to perceive the phonological dimensions they are tasked to differentiate [9]. In recent work, we have investigated the design of the optimal aggregation rule when the workers have a reject option so that they are unable to or choose not to respond [13].

The possibility of using confidence scores to improve the quality of crowdsourced labels was investigated in [11]. An aggregation method using confidence scores to integrate labels provided by crowdsourcing workers was developed in [17]. A payment mechanism was proposed for crowdsourcing systems with a reject option and confidence score reporting [22]. Indeed, confidence reporting can be useful for estimating the quality of the provided responses and possibly yield better outcomes when the aggregation is not optimal. However, potential crowdsourcing performance improvement with an optimal aggregation rule resulting from confidence reporting has not yet been investigated. As is studied in this paper, when an optimal aggregation rule is developed, confidence reporting does not help to improve the performance.

In this paper, we further consider the problem investigated in [13] by studying the scenario when the workers include their confidence levels in their responses. The main contribution of this paper is the counterintuitive finding that the confidence scores of the crowd do not play a role in the optimal aggregation rule. The weight assignment scheme to ensure the maximum weight for the correct class is the same as that when there is no confidence reporting. Although confidence reporting can provide useful information for estimating the quality of the crowd, the noise introduced due to categorization of confidence makes the estimation less accurate. Since the estimation result is essential for aggregation, confidence reporting may cause performance degradation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*SocialSens'2017, Pittsburgh, PA, USA*

© 2017 ACM. 978-1-4503-4977-2/17/04...\$15.00

DOI: {<http://dx.doi.org/10.1145/3055601.3055607>}

## 2 CROWDSOURCING TASK WITH A REJECT OPTION

Consider the situation where  $W$  workers take part in an  $M$ -ary object classification task. Each worker is asked  $N$  simple binary questions, termed as microtasks, and the worker's answer to a single microtask is conventionally represented by either "1" (Yes) or "0" (No), which eventually lead to a classification decision among the  $M$  classes. We assume independent microtask design and, therefore, we have  $N = \lceil \log_2 M \rceil$  independent microtasks of equal difficulty. The workers submit responses that are combined to give the final decision. Here, we consider the microtasks to be simple binary questions and the worker's answer to a single microtask is conventionally represented by either "1" (Yes) or "0" (No) [20, 25]. Thus, the  $w$ th worker's ordered answers to all the microtasks form an  $N$ -bit word, which is denoted by  $\mathbf{a}_w$ . Let  $\mathbf{a}_w(i)$ ,  $i \in \{1, 2, \dots, N\}$  represent the  $i$ th bit in this vector.

In our previous work [13], we considered a general problem setting where the worker has a reject option of skipping the microtasks. We denote this skipped answer as  $\lambda$ , whereas the "1/0" (Yes/No) answers are termed as definitive answers. Due to the variability of different worker backgrounds, the probability of submitting definitive answers is different for different workers. Let  $p_{w,i}$  represent the probability of the  $w$ th worker submitting  $\lambda$  for the  $i$ th microtask. Similarly, let  $\rho_{w,i}$  be the probability that  $\mathbf{a}_w(i)$ , the  $i$ th answer of the  $w$ th worker, is correct given that a definitive answer has been submitted. Due to the variabilities and anonymity of workers, we study crowdsourcing performance when  $p_{w,i}$  and  $\rho_{w,i}$  are realizations of certain probability distributions, which are denoted by distributions  $F_p(p)$  and  $F_\rho(\rho)$  respectively. The corresponding means are expressed as  $m$  and  $\mu$ .

Let  $H_0$  and  $H_1$  denote the hypotheses where "0" or "1" is the true answer for a single microtask, respectively. For simplicity of performance analysis,  $H_0$  and  $H_1$  are assumed equiprobable for every microtask. The crowdsourcing task manager or a fusion center (FC) collects the  $N$ -bit words from  $W$  workers and performs fusion based on an aggregation rule. We focus on finding the optimal aggregation rule and let us briefly review the results regarding the aggregation of responses from the workers for classification in our previous work [13].

• Let  $D = \{e_j, j = 1, 2, \dots, M\}$  be the set of all the object classes, where  $e_j$  represents the  $j$ th class. Based on  $w$ th worker's response to the microtasks, a subset  $D_w$  is chosen, within which the classes are associated with weight  $W_w$  for aggregation.<sup>1</sup> The fusion center FC adds up the weights for every class and chooses the one with highest overall weight as the final decision  $e_D$ , which can be expressed as

$$e_D = \arg \max_{e_j \in D} \left\{ \sum_{w=1}^W W_w I_{D_w}(e_j) \right\}, j = 1, 2, \dots, M, \quad (1)$$

where  $I_{D_w}(e_j)$  is an indicator function which equals 1 if  $e_j \in D_w$  and 0 otherwise. To derive the optimal weight  $W_w$  for each worker, one may look into the minimization of the misclassification probability, for which a closed-form expression cannot be derived without an explicit expression for  $W_w$ . Hence, it is difficult to determine the

<sup>1</sup>If all the responses from the  $w$ th worker are definitive,  $D_w$  is a singleton. Otherwise,  $D_w$  contains multiple classes.

optimal weight.

- The  $M$ -ary classification task can also be split into  $N$  binary hypothesis testing problems, by associating a classification decision with an  $N$ -bit word. Each worker votes "1" or "0" with the weight  $W_w$  for every bit. In this case, the Chair-Varshney rule gives the optimal weight as  $W_w = \log \frac{\rho_{w,i}}{1-\rho_{w,i}}$  [3]. However, this requires the prior knowledge on  $\rho_{w,i}$  for every worker, which is not available in practice.
- We proposed a novel weighted majority voting method, which was derived by solving the following optimization problem

$$\begin{aligned} & \text{maximize } E_C[\mathbb{W}] \\ & \text{subject to } E_O[\mathbb{W}] = K \end{aligned} \quad (2)$$

where  $E_C[\mathbb{W}]$  denotes the crowd's average weight contribution to the correct class and  $E_O[\mathbb{W}]$  denotes the average weight contribution to all the possible classes that is constrained to remain a constant  $K$ . Statistically, this method ensures maximum weight to the correct class and consequently maximum probability of correct classification. We showed that this method significantly outperforms the simple majority voting procedure.

In this paper, we investigate the impact of confidence reporting from the crowd on system performance. The weight assignment scheme is developed by solving problem (2) as well.

## 3 CROWDCOURING WITH CONFIDENCE REPORTING

We consider the case where the crowd is composed of honest workers, which means that the workers honestly observe, think, and answer the questions, give confidence levels, and skip questions that they are not confident about. We derive the optimal weight assignment for the workers and the performance of the system in a closed form. Based on these findings, we determine the potential benefits of confidence reporting in a crowdsourcing system with a reject option.

### 3.1 Confidence Level Reporting

In a crowdsourcing system where workers submit answers and report confidence, we define the  $w$ th worker's confidence about the answer to the  $i$ th microtask as the probability of this answer being correct given that this worker gives a definitive answer, which is equal to  $\rho_{w,i}$  as defined earlier. When  $\rho_{w,i}$  is bounded as  $\frac{l_{w,i}-1}{L} \leq \rho_{w,i} \leq \frac{l_{w,i}}{L}$ ,  $l_{w,i} \in \{1, \dots, L\}$ , the  $w$ th worker reports his/her confidence level as  $l_{w,i}$ . Let  $l_{w,i}$  be drawn from the distribution  $l_{w,i} \sim F_L(l)$ . Note that every worker independently gives confidence levels for different microtasks, and  $L = 1$  simply means that workers submit answers and do not report their confidence levels.

Assuming that a worker can accurately perceive the probability  $\rho_{w,i}$  and honestly report the confidence level, intuitively it is expected that it will benefit the crowdsourcing fusion center as much more information about the quality of the crowd can be extracted. However, as the confidence is quantized, which helps the workers in determining the confidence levels to be reported, quantization noise is introduced in extracting the crowd quality from confidence reporting.

As an illustrative example, consider the problem of mismatched crowdsourcing for speech transcription, which has garnered interest in the signal processing community [4, 6, 9, 12, 14, 23]. Suppose the four possibilities for a velar stop consonant to transcribe are  $R = \{ \text{क, ख, ग, घ} \}$ . The simple binary question of “whether it is aspirated or unaspirated” differentiates between  $\{ \text{ख, घ} \}$  and  $\{ \text{क, ग} \}$ , whereas the binary question of “whether it is voice or unvoiced” differentiates between  $\{ \text{ग, घ} \}$  and  $\{ \text{क, ख} \}$ . The highest confidence level is set as  $L = 4$ . Now suppose the first worker is a native Italian speaker. Since Italian does not use aspiration, this worker will be unable to differentiate between  $\{ \text{क} \}$  and  $\{ \text{ख} \}$ , or between  $\{ \text{ग} \}$  and  $\{ \text{घ} \}$ . It would be of benefit if this worker would specify the inability to perform the task through a special symbol  $\lambda$ , rather than guessing randomly, and this worker answers “Yes” with confidence level 1 to the second question. Suppose the second worker is a native Bengali speaker. Since this language makes a four-way distinction among velar stops, such a worker will probably answer both questions without a  $\lambda$ .

In the rest of this section, we address the problem “Does the confidence reporting help crowdsourcing system performance?” by performing analyses when workers report their confidences with their definitive answers.

### 3.2 Optimal Weight Assignment Scheme

We determine the optimal weight  $W_w$  for the  $w$ th worker in this section. We rewrite hereby the weight assignment problem

$$\begin{aligned} & \text{maximize } E_C [\mathbb{W}] \\ & \text{subject to } E_O [\mathbb{W}] = K \end{aligned} \quad (3)$$

where  $E_C [\mathbb{W}]$  denotes the crowd’s average weight contribution to the correct class and  $E_O [\mathbb{W}]$  denotes the average weight contribution to all the possible classes and remains a constant  $K$ . Statistically, we are looking for the weight assignment scheme such that the weight contribution to the correct class is maximized while the weight contribution to all the classes remains fixed, so as to maximize the probability of correct classification.

**PROPOSITION 3.1.** *To maximize the average weight assigned to the correct classification element, the weight for  $w$ th worker’s answer is given by*

$$W_w = \mu^{-n}, \quad (4)$$

where  $n$  is the number of definitive answers that the  $w$ th worker submits.

**PROOF.** See Appendix.  $\square$

**REMARK 1.** *Here the weight depends on the number of questions answered by a worker. In fact, if more questions are answered, the weight assigned to the corresponding worker’s answer is larger. This is intuitively pleasing as a high-quality worker is able to answer more questions and is assigned a higher weight. Increased weight can put more emphasis on the contribution of high-quality workers in that sense and improve overall classification performance.*

**REMARK 2.** *When  $L = \infty$ ,  $\rho_{w,i}$  associated with every worker for every microtask is reported exactly. Then the Chair-Varshney rule gives the optimal weight assignment to minimize error probability [3]. However, human decision makers are limited in their information*

*processing capacity and can only carry around seven categories [15]. Thus, the largest value of  $L$  is around 7 in practice.*

**REMARK 3.** *Note that the optimal weight assignment scheme is the same as in the case where the workers do not report confidence levels, i.e.,  $L = 1$ . Actually, the value of  $L$  does not play any role in the weight assignment, as long as  $\rho_{w,i}$  is not known exactly. Therefore, the weight assignment is universally optimal regardless of confidence reporting.*

### 3.3 Parameter Estimation

Before the proposed aggregation rule can be used,  $\mu$  has to be estimated to assign the weight for every worker’s answers. Here, we employ three approaches to estimate  $\mu$ . We refer to our previous work [13] for training-based and majority-voting based methods to estimate  $\mu$ , and give an additional method using the information extracted from the workers’ reported confidence levels.

*Confidence-based.* Note that the reported confidence levels correspond to  $\rho_{w,i}$ . We collect all the values of the submitted confidence levels and obtain the estimate of  $\mu$  from them. First, the  $w$ th worker’s confidence level for the  $i$ th microtask is represented by  $l_{w,i}$ . Considering the fact that  $\frac{l_{w,i}-1}{L} \leq \rho_{w,i} \leq \frac{l_{w,i}}{L}$  if the worker submits a definitive answer, we use  $\frac{l_{w,i}-\frac{1}{2}}{L}$  to approximate  $\rho_{w,i}$ . Let  $l_{w,i} = \frac{1}{2}$  if the  $w$ th worker skips the  $i$ th microtask. We obtain the estimate of  $\mu$  by

$$\hat{\mu} = \frac{1}{W - \epsilon} \sum_{w=1}^W \sum_{i=1}^N \frac{l_{w,i} - \frac{1}{2}}{LI(w)}, \quad (5)$$

where  $I(w)$  denotes the number of definitive answers that  $w$ th worker submits.

### 3.4 Performance Analysis

In this section, we characterize the performance of the proposed crowdsourcing classification framework in terms of the probability of correct classification  $P_c$ . Note that we have overall correct classification only when all the bits are classified correctly.

**PROPOSITION 3.2.** *The probability of correct classification  $P_c$  in the crowdsourcing system is*

$$\begin{aligned} P_c = & \left[ \frac{1}{2} + \frac{1}{2} \sum_S \binom{W}{Q} (F(Q) - F'(Q)) \right. \\ & \left. + \frac{1}{4} \sum_{S'} \binom{W}{Q} (F(Q) - F'(Q)) \right]^N, \end{aligned} \quad (6)$$

where  $Q = \left\{ (q_{-N}, q_{-N+1}, \dots, q_N) : \sum_{n=-N}^N q_n = W \right\}$  with natural numbers  $q_n$  and  $q_0$ , and  $S = \left\{ Q : \sum_{n=1}^N \mu^{-n} (q_n - q_{-n}) > 0 \right\}$ ,  $S' = \left\{ Q : \sum_{n=1}^N \mu^{-n} (q_n - q_{-n}) = 0 \right\}$ ,  $\binom{W}{Q} = \frac{W!}{\prod_{n=-N}^N q_n!}$ , and

$$F(Q) = m^{q_0} \prod_{n=1}^N (1 - \mu)^{q_{-n}} \mu^{q_n} \binom{C_{N-1}^{q_{-n}} (1 - m)^n m^{N-n}}{n}^{q_{-n} + q_n}$$

$$F'(\mathbb{Q}) = m^{q_0} \prod_{n=1}^N (1 - \mu)^{q_n} \mu^{q-n} \left( C_{N-1}^{n-1} (1 - m)^n m^{N-n} \right)^{q-n+q_n}.$$

PROOF. The proof is similar to the proof in our previous work [13] and is, therefore, omitted for brevity.  $\square$

### 4 SIMULATION RESULTS

In this section, we give the simulation results for the proposed crowdsourcing system. The workers take part in a classification task of  $N = 3$  microtasks.  $F_p(p)$  is a uniform distribution denoted as  $U(0, 1)$ .

First, we show the efficiency of the derived optimal weight assignment over the widely used simple majority voting method for crowdsourcing systems. The performance comparison is presented with the number of workers varying from 3 to 29. Here, we consider different qualities of the individual workers in the crowd which is represented by variable  $\rho_{w,i}$  with a uniform distribution  $U(0.6, 1)$ . Thus, the mean  $\mu$  is 0.8, and we give simulation results when confidence reporting is not included and the estimation of  $\mu$  is perfect in Fig. 1. It is observed that a larger crowd completes the classification task with higher quality. A significant performance improvement by the proposed method with a reject option compared with the simple majority voting is shown in the figure.

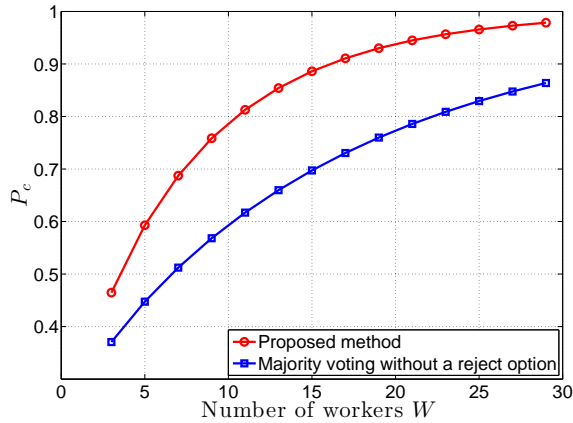


Figure 1: Performance comparison with various crowd sizes.

Since an accurate estimation of  $\mu$  is essential for applying the optimal weight assignment scheme, we next focus on the estimation results of  $\mu$  for the three estimation methods as discussed in the previous section. Let  $F_p(\rho)$  be a uniform distribution expressed as  $U(x, 1)$  with  $0 \leq x \leq 1$ , and thus we can have  $\mu$  varying from 0.5 to 1. We consider that  $W = 20$  workers participate in the classification task with a reject option and confidence reporting.

In Fig. 2a, it is observed that the training-based method has the best overall performance, which takes advantage of the gold standard questions. We can also see that the majority voting method has better performance as  $\mu$  increases. This is because a larger  $\mu$  means a better-quality crowd, which will lead to a more accurate result from majority voting, and consequently better estimation performance of  $\mu$ . When confidence is considered with  $L = 4$ , we find that the overall estimation performance is not better than the

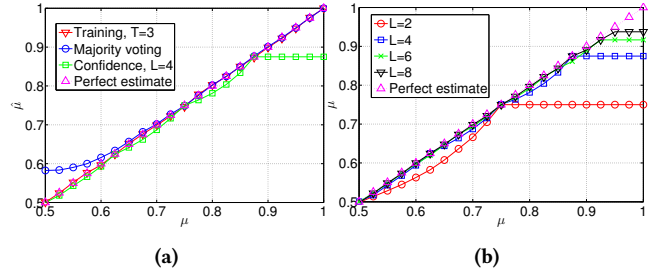


Figure 2: Estimation performance comparison. (a) Different methods. (b) Confidence-based method with different confidence levels.

other two methods because of quantization noise associated with confidence reporting in the estimation of  $\mu$ . It is also shown that the curve saturates and yields a fixed value of  $\hat{\mu} = 0.875$  when  $\mu \geq 0.9$ . This is because almost all the confidence levels submitted then are  $l_{w,i} = 4$  and the corresponding estimate result is exactly 0.875.

The estimation performance of the confidence-based method with multiple confidence levels is presented in Fig. 2b. As is expected, a larger  $L$  can help improve the estimation performance. However, it is seen that even though  $L = 8$ , the corresponding performance is still not as good as that of the other two methods. Although we can expect estimation performance improvement as the maximum number of confidence levels  $L$  increases,  $L = 8$  is pretty much the limit in practice due to the human inability to categorize beyond 7 levels. When the confidence-based estimation method is employed, the estimate value saturates at a certain fixed value when  $\mu$  is large. Therefore, it can be concluded that the confidence-based estimation method does not provide good results.

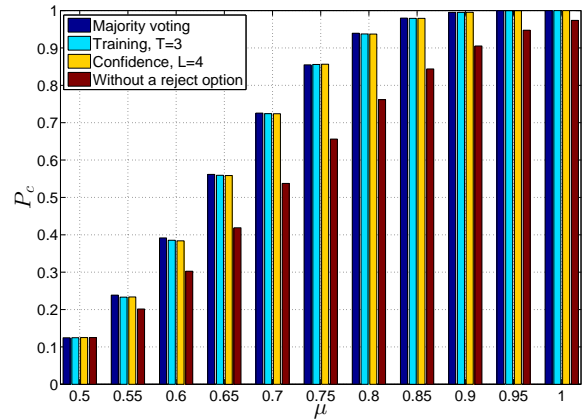


Figure 3: Robustness of the proposed system and performance comparison with simple majority voting

Even though the three methods differ in performance in the estimation of  $\mu$ , we show in Fig. 3 the robustness of the proposed system. We observe from Fig. 2a that the majority voting based method suffers from performance degradation in the low- $\mu$  regime, while the confidence based one suffers in the high- $\mu$  regime. However, when the value of  $\mu$  is low, the workers are making random

guesses even when they believe that they are able to respond with definitive answers. When the value of  $\mu$  is large, almost all the definitive answers submitted are correct. Therefore, in those two situations, the performance degradation in the estimation of  $\mu$  is negligible. From Fig. 3, we see that system performance of the proposed system with estimation results from Fig. 2a is almost the same as with the other three estimation methods, which significantly outperforms the system where simple majority voting is employed without a reject option. However, if a significant performance degradation in the estimation of  $\mu$  occurs outside the two aforementioned regimes, overall classification performance loss is expected. For example, consider the case where  $\mu$  is 0.8 while  $\hat{\mu}$  is 0.5, and  $N = 5$ , then  $P_c = 0.8$ . However, the actual  $P_c$  equals 0.89 when  $\mu$  is estimated with an acceptable error.

## 5 CONCLUSION

We have studied a novel framework of crowdsourcing system for classification, where an individual worker has the reject option and can skip a microtask if he/she has no definitive answer, and gives definitive answers with quantized confidence. We presented an aggregation approach using a weighted majority voting rule, where each worker's response is assigned an optimized weight to maximize the crowd's classification performance. However, we showed that reporting of confidence by the crowd does not benefit classification performance. One is advised to adopt the reject option without confidence indication from the workers as it does not improve classification performance and may degrade performance in some cases.

## APPENDIX

To solve problem (3), we need  $E_C[\mathbb{W}]$  and  $E_O[\mathbb{W}]$ . First, the  $w$ th worker can have weight contribution to  $E_C[\mathbb{W}]$  only if all his/her definitive answers are correct. Thus, we have the average weight assigned to the correct element as

$$E_C[\mathbb{W}] = E_{p,\rho,l} \left[ \sum_{w=1}^W \sum_{n=0}^N \sum_{N_n \in [N]} \prod_{i \in N_n} W_w (1-p_{w,i}) \rho_{w,i} \prod_{j \in [N]-N_n} p_{w,j} \right] \quad (7)$$

where  $[N]$  denotes  $\{1, \dots, N\}$  and  $N_n \in [N]$  with cardinality  $n$ . Given a known  $w$ th worker, i.e.,  $p_{w,i}$  is known, we write

$$A_w(p_{w,i}) = \sum_{n=0}^N E_{\rho,l} \left[ W_w \prod_{i \in N_n} \rho_{w,i} |l_{w,i}| P_\lambda(n) \right], \quad (8)$$

where  $P_\lambda(n) = \sum_{N_n \in [N]} \prod_{i \in N_n} (1-p_{w,i}) \prod_{j \in [N]-N_n} p_{w,j}$ .

Note that  $\sum_{n=0}^N P_\lambda(n) = 1$ , and then (8) is upper-bounded using Cauchy-Schwarz inequality as follows:

$$A_w(p_{w,i}) = \sum_{n=0}^N E_{\rho,l} \left[ W_w \prod_{i \in N_n} \rho_{w,i} |l_{w,i}| \sqrt{P_\lambda(n)} \sqrt{P_\lambda(n)} \right] \\ \leq \sqrt{\sum_{n=0}^N E_{\rho,l}^2 \left[ W_w \prod_{i \in N_n} \rho_{w,i} |l_{w,i}| \right] P_\lambda(n)} \sqrt{\sum_{n=0}^N P_\lambda(n)}. \quad (9)$$

Also note that equality holds in (9) only if

$$E_{\rho,l} \left[ W_w \prod_{i \in N_n} \rho_{w,i} |l_{w,i}| \sqrt{P_\lambda(n)} \right] = \alpha_w(p_{w,i}) \sqrt{P_\lambda(n)}, \quad (10)$$

where  $\alpha_w$  is a positive quantity independent of  $n$ , which might be a function of  $p_{w,i}$ , and

$$E_{\rho,l} \left[ W_w \prod_{i \in N_n} \rho_{w,i} |l_{w,i}| \right] = \alpha_w(p_{w,i}). \quad (11)$$

Note that  $\int_{p_{w,i}} F_p(p_{w,i} = x) dx = 1$ , and similarly we write

$$E_p[A_w(p_{w,i})] \leq \int_{p_{w,i}} \alpha_w(p_{w,i}) \Pr(p_{w,i} = x) dx \\ \leq \sqrt{\int_{p_{w,i}} \alpha_w^2(p_{w,i}) \Pr(p_{w,i} = x) dx} \sqrt{\int_{p_{w,i}} \Pr(p_{w,i} = x) dx}. \quad (12)$$

The equality (12) holds only if

$$\alpha_w(p_{w,i}) \sqrt{\Pr(p_{w,i} = x)} = \beta \sqrt{\Pr(p_{w,i} = x)}, \quad (13)$$

WHERE  $\beta$  is a positive constant independent of  $p_{w,i}$ , and we conclude that  $\alpha_w$  is also a positive quantity independent of  $p_{w,i}$ . Then

from (11), we have  $E_{\rho,l} \left[ W_w \prod_{i \in N_n} \rho_{w,i} |l_{w,i}| \right] = \beta$ . Since  $\prod_{i \in N_n} \rho_{w,i}$  is the product of  $n$  variables, its distribution is not known *a priori*. A possible solution to weight assignment is a deterministic value given by  $W_w E_{\rho,l} \left[ \prod_{i \in N_n} \rho_{w,i} |l_{w,i}| \right] = \beta$  and, therefore, we can write the weight as

$$W_w = \frac{\beta}{E_{\rho,l} \left[ \prod_{i \in N_n} \rho_{w,i} |l_{w,i}| \right]} = \frac{\beta}{\mu^n}. \quad (14)$$

Then, we can express the crowd's average weight contribution to all the classes defined in (3) as

$$E_O[\mathbb{W}] = \sum_{w=1}^W E_{p,\rho,l} \left[ \sum_{n=0}^N \beta \mu^{-n} 2^{N-n} P_\lambda(n) \right] \\ = \sum_{w=1}^W \sum_{n=0}^N \beta \mu^{-n} 2^{N-n} \binom{N}{n} (1-m)^n m^{N-n} \\ = W \beta \left( \frac{1-m}{\mu} + 2m \right)^N = K. \quad (15)$$

Thus,  $\beta$  and the weight can be obtained accordingly. Note that the weight derived above has a term that is common for every worker. Since the voting scheme is based on comparison, we can ignore this factor and have the normalized weight as  $W_w = \mu^{-n}$ .

## ACKNOWLEDGMENTS

This work was supported in part by the Army Research Office under Grant W911NF-14-1-0339 and in part by the National Science Foundation under Grant ENG-1609916.

## REFERENCES

- [1] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. 2013. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Comput.* 2 (March 2013), 76–81.
- [2] Steven Burrows, Martin Potthast, and Benno Stein. 2013. Paraphrase acquisition via crowdsourcing and machine learning. *ACM Trans. Intell. Syst. Technol.* 4, 3 (July 2013), 43.
- [3] Zeineddin Chair and Pramod K. Varshney. 1986. Optimal data fusion in multiple sensor detection systems. *IEEE Trans. Aerosp. Electron. Syst.* AES-22, 1 (Jan. 1986), 98–101. DOI: <http://dx.doi.org/10.1109/TAES.1986.310699>
- [4] Wenda Chen, Mark Hasegawa-Johnson, and Nancy F. Chen. 2016. Mismatched Crowdsourcing based Language Perception for Under-resourced Languages. *Procedia Computer Science* 81 (2016), 23–29. DOI: <http://dx.doi.org/10.1016/j.procs.2016.04.025>
- [5] Ju Fan, Meihui Zhang, S. Kok, Meiyu Lu, and Beng Chin Ooi. 2015. CrowdOp: Query Optimization for Declarative Crowdsourcing Systems. *IEEE Trans. Knowl. Data Eng.* 27, 8 (Aug. 2015), 2078–2092. DOI: <http://dx.doi.org/10.1109/TKDE.2015.2407353>
- [6] Mark Hasegawa-Johnson, Jennifer Cole, Preethi Jyothi, and Lav R. Varshney. 2015. Models of Dataset Size, Question Design, and Cross-Language Speech Perception for Speech Crowdsourcing Applications. *Laboratory Phonology* 6, 3-4 (Oct. 2015), 381–431. DOI: <http://dx.doi.org/10.1515/lp-2015-0012>
- [7] Matthias Hirth, Tobias Hofffeld, and Phuoc Tran-Gia. 2013. Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Math. Comput. Model.* 57, 11 (July 2013), 2918–2932.
- [8] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality Management on Amazon Mechanical Turk. In *Proc. ACM SIGKDD Workshop Human Comput. (HCOMP'10)*. 64–67. DOI: <http://dx.doi.org/10.1145/1837885.1837906>
- [9] Preethi Jyothi and Mark Hasegawa-Johnson. 2015. Acquiring Speech Transcriptions Using Mismatched Crowdsourcing. In *Proc. 29th AAAI Conf. Artificial Intelligence (AAAI'15)*.
- [10] David R. Karger, Sewoong Oh, and Devavrat Shah. 2011. Iterative learning for reliable crowdsourcing systems. In *Advances in Neural Information Processing Systems (NIPS) 24*. MIT Press, Cambridge, MA, 1953–1961.
- [11] Gabriella Kazai. 2011. In search of quality in crowdsourcing for search engine evaluation. In *European Conference on Information Retrieval*. Springer, 165–176.
- [12] Xiang Kong, Preethi Jyothi, and Mark Hasegawa-Johnson. 2016. Performance Improvement of Probabilistic Transcriptions with Language-specific Constraints. *Procedia Computer Science* 81 (2016), 30–36. DOI: <http://dx.doi.org/10.1016/j.procs.2016.04.026>
- [13] Q. Li, A. Vempaty, L. R. Varshney, and P. K. Varshney. 2017. Multi-Object Classification via Crowdsourcing With a Reject Option. *IEEE Trans. Signal Process.* 65, 4 (Feb 2017), 1068–1081. DOI: <http://dx.doi.org/10.1109/TSP.2016.2630038>
- [14] Chunxi Liu, Preethi Jyothi, Hao Tang, Vimal Manohar, Rose Sloan, Tyler Kekona, Mark Hasegawa-Johnson, and Sanjeev Khudanpur. 2016. Adapting ASR for Under-Resourced Languages Using Mismatched Transcriptions.
- [15] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.
- [16] Kaixiang Mo, Erheng Zhong, and Qiang Yang. 2013. Cross-task crowdsourcing. In *Proc. ACM Int. Conf. Knowl Discovery Data Mining*. 677–685.
- [17] Satoshi Oyama, Yukino Baba, Yuko Sakurai, and Hisashi Kashima. 2013. Accurate Integration of Crowdsourced Labels Using Workers' Self-reported Confidence Scores. In *IJCAI*.
- [18] Praveen Paritosh, Panos Ipeirotis, Matt Cooper, and Siddharth Suri. 2011. The computer is the new sewing machine: Benefits and perils of crowdsourcing. In *Proc. 20th Int. Conf. World Wide Web (WWW'11)*. 325–326.
- [19] Alexander J. Quinn and Benjamin B. Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *Proc. 2011 Annu. Conf. Hum. Factors Comput. Syst. (CHI 2011)*. 1403–1412. DOI: <http://dx.doi.org/10.1145/1978942.1979148>
- [20] Jana Rucker, Christopher M Yauch, Sumanth Yenduri, LA Perkins, and F Zand. 2007. Paper-based dichotomous key to computer based application for biological identification. *J. Comput. Sci. Coll.* 22, 5 (May 2007), 30–38.
- [21] D. Sanchez-Charles, J. Nin, M. Sole, and V. Munes-Mulero. 2014. Worker ranking determination in crowdsourcing platforms using aggregation functions. In *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ 2014)*. 1801–1808. DOI: <http://dx.doi.org/10.1109/FUZZ-IEEE.2014.6891807>
- [22] Nihar B Shah and Dengyong Zhou. 2014. Double or Nothing: Multiplicative Incentive Mechanisms for Crowdsourcing. *arXiv preprint arXiv:1408.1387* (2014).
- [23] Lav R. Varshney, Preethi Jyothi, and Mark Hasegawa-Johnson. 2016. Language Coverage for Mismatched Crowdsourcing.
- [24] Lav R. Varshney, Aditya Vempaty, and Pramod K. Varshney. 2014. Assuring Privacy and Reliability in Crowdsourcing with Coding. In *Proc. 2014 Inf. Theory Appl. Workshop*. DOI: <http://dx.doi.org/10.1109/ITA.2014.6804213>
- [25] Aditya Vempaty, Lav R. Varshney, and Pramod K. Varshney. 2014. Reliable Crowdsourcing for Multi-Class Labeling Using Coding Theory. *IEEE J. Sel. Topics Signal Process.* 8, 4 (Aug. 2014), 667–679. DOI: <http://dx.doi.org/10.1109/JSTSP.2014.2316116>
- [26] Dejun Yue, Ge Yu, Derong Shen, and Xiacong Yu. 2014. A weighted aggregation rule in crowdsourcing systems for high result accuracy. In *Proc. IEEE 12th Int. Conf. Depend. Auton. Secure Comput. (DASC)*. 265–270.
- [27] Yu Zhang and Mihaela van der Schaar. 2012. Reputation-based incentive protocols in crowdsourcing applications. In *Proc. 31st IEEE Conf. Computer Commun. (INFOCOM 2012)*. 2140–2148.