Accounting for Confirmation Bias in Crowdsourced Label Aggregation

Meric Altug Gemalmaz , Ming Yin
Purdue University
{mgemalma, mingyin}@purdue.edu

Abstract

Collecting large-scale human-annotated datasets via crowdsourcing to train and improve automated models is a prominent human-in-the-loop approach to integrate human and machine intelligence. However, together with their unique intelligence, humans also come with their biases and subjective beliefs, which may influence the quality of the annotated data and negatively impact the effectiveness of the human-in-the-loop systems. One of the most common types of cognitive biases that humans are subject to is the *confirmation bias*, which is people's tendency to favor information that confirms their existing beliefs and values. In this paper, we present an algorithmic approach to infer the correct answers of tasks by aggregating the annotations from multiple crowd workers, while taking workers' various levels of confirmation bias into consideration. Evaluations on real-world crowd annotations show that the proposed bias-aware label aggregation algorithm outperforms baseline methods in accurately inferring the ground-truth labels of different tasks when crowd workers indeed exhibit some degree of confirmation bias. Through simulations on synthetic data, we further identify the conditions when the proposed algorithm has the largest advantages over baseline methods.

1 Introduction

Over the past decade, crowdsourcing—the act of outsourcing tasks to the crowd—has become a ubiquitous paradigm for obtaining data from people to enhance machine intelligence. However, a long-standing challenge in crowdsourcing is how to control the quality of crowd work [Allahbakhsh *et al.*, 2013]. Recently, it is recognized that an important reason that contributes to the limited work quality of individual crowd workers is that workers are prone to a wide range of *biases* in their work. For example, workers may be influenced by their social bias (e.g., racial bias, gender bias) during their annotation process [Otterbacher *et al.*, 2019; Biswas *et al.*, 2020]. The design of crowdsourcing tasks (e.g., what information is shown to workers in what order) may also have subtle impact on workers and trigger their cognitive biases such as the anchoring bias and ambiguity effect [Eickhoff, 2018; Zhuang *et al.*, 2015].

Another common type of cognitive bias that crowd workers are often subject to is their *confirmation bias*, which refers to people's tendency of favoring information that confirms their previously existing beliefs and values [Nickerson, 1998]. Indeed, researchers have showed that when judging the truthfulness of news statements, crowd workers tend to believe those statements coming from speakers off the same political party that they have recently voted for to be more true [La Barbera *et al.*, 2020]. Similarly, Hube *et al.* [2019] revealed that when crowd workers are asked to determine whether a statement is neutral or opinionated, they are more likely to label a statement as neutral if its stance aligns with their own opinions.

In practice, to obtain high-quality annotations from crowd workers, a redundancy-based strategy is often deployed. That is, the same task is completed by multiple workers, and numerous label aggregation algorithms have been proposed to infer the ground-truth answer for each task based on the collection of annotations obtained on it [Whitehill *et al.*, 2009; Welinder *et al.*, 2010; Liu *et al.*, 2012; Demartini *et al.*, 2012; Zheng *et al.*, 2017]. While these algorithms adopt various models to characterize worker behavior during the label generation process, they seldom take worker's cognitive biases, such as their confirmation bias, into account. In so doing, the current crowdsourced label aggregation algorithms might have missed the opportunity to further improve the inference accuracy by explicitly modeling how worker's cognitive biases have influenced their work quality.

Therefore, in this paper, we focus on worker's confirmation bias and propose a new label aggregation algorithm to account for it. Specifically, we formulate a probabilistic model of the label generation process by assuming that among other factors, worker's label on a task is influenced by both the values of the worker and the values expressed in the task. We then make use of the expectation-maximization algorithm to simultaneously infer the values of each worker, the values of each task, as well as the ground-truth answer for each task.

To examine the effectiveness of the proposed algorithm, we collect annotations from real crowd workers on Amazon Mechanical Turk on the subjective task of differentiating factual statements from opinion statements, for which workers are shown to indeed exhibit a degree of confirmation bias in their annotations. We find that compared to a set of baseline

label aggregation algorithms, the proposed bias-aware label aggregation algorithm achieves a higher level of accuracy in uncovering the ground-truth label for each task. We further investigate the robustness of the proposed algorithm through simulations using synthetic datasets. Our simulation results highlight several scenarios that the proposed algorithm shows the largest advantage over baseline algorithms, such as when crowd workers suffer from confirmation bias in their annotations to a larger extent and when the distribution of worker's values is more dispersed or even polarized.

2 Related Work

Quality Control in Crowdsourcing. To solicit high-quality work from inexpert crowd workers, researchers have proposed a variety of strategies such as providing effective incentives to workers [Ho *et al.*, 2015], training novice workers [Doroudi *et al.*, 2016], assigning tasks to workers with relevant skills [Zheng *et al.*, 2015], and enabling communication between workers on the same task [Tang *et al.*, 2019]. Yet, in practice, the most widely adopted approach for ensuring the quality of crowd work, especially for simple classification tasks, is to assign a task to multiple workers and then infer its correct answer using all annotations collected on it.

To effectively combine multiple annotations and infer the ground-truth label for a task, researchers have designed various label aggregation algorithms to improve the inference accuracy by explicitly characterizing how worker's quality in a task is affected by multiple factors. For example, Whitehill *et al.* [2009] characterized worker's labeling process using a probabilistic graphic model assuming that a worker's label on a task is influenced by the worker's skill level as well as the task difficulty. Welinder *et al.* [2010] introduced a more sophisticated model to capture worker's diverse skills on various latent topics underlying a task. More recently, Braylan and Lease [2020] extended label aggregation algorithms from simple annotations (e.g., class labels) to complex annotations (e.g., open-ended text) by modeling the distances between annotations. Moreover, Li *et al.* [2020] proposed algorithms that ensure the aggregated labels satisfy fairness constraints. For a more complete review of label aggregation algorithms in crowdsourcing, please see [Zheng *et al.*, 2017].

Bias in Crowdsourced Annotations. Recent studies showed that crowd workers could be influenced by a wide range of biases during their annotation process. Such biases can be triggered by the design of the tasks. For example, it is shown that grouping multiple data items together in a batch for workers to label may lead to the "in-batch annotation bias," that is, a worker's judgment on one data item is affected by other data items within the batch [Zhuang *et al.*, 2015]. Similarly, workers are also subject to the "sequential bias" in their labeling process such that their annotation on one task might be influenced by the previous task that they see as well as the label they provide on it [Newell and Ruths, 2016; Huang *et al.*, 2018]. Within a single task, the ways that information is presented and the order that questions are asked can also result in worker's cognitive bias which negatively impacts the work quality [Eickhoff, 2018]. In addition, workers may exhibit biases in their annotations as a result of the interaction between the characteristics of the worker and the task. For example, Biswas *et al.* [2020] showed that when crowd workers are asked to assess the recidivism risk of criminal defendants, they tend to slightly favor defendants of their own race, showing some degree of in-group bias.

Another type of bias that crowd workers are prone to, especially in subjective tasks, is the confirmation bias. Via experimental studies, it is found that crowd workers tend to label a piece of news as true rather than fake, or a statement as neutral rather than opinionated, if the information expressed in the news or statements align well with the worker's own belief and value [Hube *et al.*, 2019; La Barbera *et al.*, 2020]. Researchers have also revealed that confirmation bias may largely explain why in the real-world crowdsourcing applications of misinformation flagging on social-media platforms, the news sources flagged by the crowd tend to be the most popular (and largely reliable) ones [Coscia and Rossi, 2020].

Mitigate Confirmation Bias in Crowdsourcing. Most recently, researchers have explored different approaches to mitigate crowd worker's confirmation bias and reduce the negative impact the bias brings to work quality, which have mixed success. For example, Hube *et al.* [2019] showed that raising people's awareness of their own bias can effectively reduce worker's bias in annotations. On the other hand, it is found that enabling workers with different beliefs and values to work on the same task and interact with each other does not help reduce worker's bias [Duan *et al.*, 2020]. This paper provides a new approach in "mitigating" confirmation bias— we explicitly model how worker's confirmation bias sneak into their annotations, and then design algorithms based on such model to reduce the bias in the final, aggregated labels.

3 Bias-aware Label Aggregation

In this section, we outline our algorithmic approach for crowdsourced label aggregation which takes annotators' confirmation biases into account. We consider subjective labeling tasks in which annotators are asked to provide binary labels in each task, and importantly, one of the two candidate labels is generally perceived to be more "*preferable*" (e.g., a piece of news is "true" rather than "fake," a statement is "neutral" rather than "opinionated"). On these tasks, annotators might be subject to confirmation bias—they might favor information that confirms their previously existing beliefs or values, hence increase their chance of providing the preferable label in tasks containing the favorable information.

3.1 Label Generation Model

Consider the scenario that N annotators are asked to complete M binary labeling tasks. An annotator i's label on task j is denoted as $l_{ij} \in \{0, 1\}$, with 0 representing the preferable label (e.g., "true news", "neutral statement"). Our goal is to determine the true label, $z_j \in \{0, 1\}$, for each task j using all the labels collected on it. To model annotators' possible confirmation bias during their label generation processes, we assume the observed labels l_{ij} depend on several causal factors: (1) the values implied by the information in the task; (2) the annotator's values; (3) the annotator's degree of bias characterizing how much the annotator is subject to confirmation

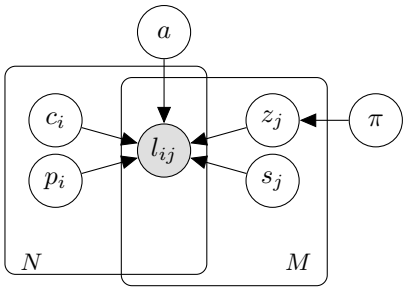


Figure 1: The probabilistic graphical model of annotators' label generation process. The shaded node is observed.

bias, (4) annotator's inherent tendency to provide the preferable label, and (5) the true label of the task. Under our model, the chance for annotator i to provide the preferable label on task j (i.e., $l_{ij} = 0$) is characterized as:

$$P(l_{ij} = 0|c_i, p_i, s_j, z_j, a) = \frac{1}{e^{a[(1-p_i)(s_j-c_i)^2+p_i z_j]}} \quad (1)$$

In Eqn. 1, for simplicity, we model the values of both the annotators and the information expressed in the tasks using a single dimensional spectrum—the values of annotator i are captured by the parameter $c_i \in [0, 1]$, while the values of the information contained in task j are captured by the parameter $s_j \in [0, 1]$[1]. For example, when considering the left–right political spectrum, $c_i = 1$ (or $s_j = 1$) could mean the values of annotator i (or the values implied by information in task j) are extremely conservative, while $c_i = 0$ (or $s_j = 0$) means the values of annotator i (or the values implied by information in task j) are extremely liberal. Annotators' confirmation bias is captured via the *distance* between c_i and s_j—holding all other variables equal, the closer c_i and s_j are to each other, the more likely annotator i will provide the preferable label in task j (i.e., $P(l_{ij} = 0)$ is larger).

We further use the parameter $p_i \in [0, 1]$ to characterize the extent to which annotator i is subject to confirmation bias. Here, $p_i = 0$ means that annotator i is heavily influenced by her confirmation bias, such that she decides her label on tasks (almost) entirely based on how much the information contained in the task aligns with her values. Conversely, when $p_i = 1$, annotator i is not influenced by her confirmation bias at all, such that she decides her label on tasks (almost) entirely based on the ground truth label z_j of the task, and $z_j \sim$ Bernoulli$(1 - \pi)$ (i.e., the prior probability for a task to have the preferable label as its ground truth is π, $P(z_j = 0) = \pi$). When $0 < p_i < 1$, the annotator is influenced by her confirmation bias to some degree, and the smaller p_i is, the more she is subject to the confirmation bias.

Finally, we use a global parameter $a \in [0, +\infty)$ to represent annotators' inherent tendency to provide the preferable label on any task, or in other words, annotators' base rate of providing the preferable label in tasks. When $a = 0$, the base rate for annotators to provide the preferable label in tasks is very high, while $a = +\infty$ means the base rate for annotators to provide the preferable label in tasks is very low.

[1]Our model can easily be extended to cases where the values of annotators and tasks are characterized in a multi-dimensional space.

Our entire label generation model for the crowdsourced annotators is shown in Figure 1. Given a set of observed labels $\mathbf{L} = \{l_{ij}\}$, the end goal of our label aggregation algorithm is to infer the most likely ground-truth label $\mathbf{z} = \{z_j\}$ for each task, as well as the values of all hidden parameters (i.e., $\mathbf{s} = \{s_j\}, \mathbf{c} = \{c_i\}, \mathbf{p} = \{p_i\}, a, \pi$).

3.2 Model Inference

We use the Expectation-Maximization (EM) algorithm to estimate the maximum likelihood estimates of the hidden parameters and infer the values of the hidden variables z_j.

In particular, in the Expectation step, we compute the posterior probabilities for each hidden variable z_j based on the current estimates of parameters and the observed labels:

$$\begin{aligned} p(z_j|\mathbf{L}, \mathbf{c}, \mathbf{p}, \mathbf{s}, a, \pi) &\propto p(z_j|\pi)p(\mathbf{L}|z_j, \mathbf{c}, \mathbf{p}, \mathbf{s}, a) \\ &\propto p(z_j|\pi) \prod_{i \in W_j} p(l_{ij}|c_i, p_i, s_j, z_j, a) \end{aligned}$$

Here, we use W_j to denote the set of all annotators who have provided labels on task j. When $l_{ij} = 0$, $p(l_{ij}|c_i, p_i, s_j, z_j, a)$ can be computed using Eqn. 1; otherwise, $p(l_{ij}|c_i, p_i, s_j, z_j, a) = 1 - P(l_{ij} = 0|c_i, p_i, s_j, z_j, a)$.

For the Maximization step, we search for optimal parameter values to maximize the auxiliary function Q, i.e., the expectation of the complete data log-likelihood:

$$\begin{aligned} &Q(\mathbf{c}, \mathbf{p}, \mathbf{s}, a, \pi) = E[\ln p(\mathbf{L}, \mathbf{z}|\mathbf{c}, \mathbf{p}, \mathbf{s}, a, \pi)] \\ &= E[\ln \prod_j (p(z_j|\pi) \prod_{i \in W_j} p(l_{ij}|c_i, p_i, s_j, z_j, a))] \\ &= \sum_j E[\ln p(z_j|\pi)] + \sum_{l_{ij} \in \mathbf{L}} E[\ln p(l_{ij}|c_i, p_i, s_j, z_j, a)] \end{aligned}$$

The expectation is taken with respect to the posterior distributions of z_j that are obtained from the previous E-step. In each M-step, we use gradient descent to update hidden parameters to the values that locally optimize Q.

4 Experiment

In this section, we examine that on real-world subjective labeling tasks where annotators could suffer from confirmation bias, whether the proposed bias-aware label aggregation algorithm can help improve the accuracy of inferred labels.

4.1 Data Collection

To empirically evaluate the effectiveness of the proposed algorithm, we first collected a set of annotations generated by real crowd workers on the subjective task of differentiating factual statements from opinion statements. We used this task in our study since previous research showed that U.S. adults were more likely to label both factual and opinion statements as factual when they appealed more to their political side (i.e., "factual" is the preferable label) [Mitchell *et al.*, 2018].

Specifically, from the list of controversial topics in US politics[2], we selected "gun control" as the main topic and created

[2]https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues.

Statement	Label	Values
Gun bans alleviate intimate partner homicide.	Factual	Liberal
Active shooter events in the U.S. is sometimes associated with mental illness.	Factual	Conservative
Easy usage of the guns increases firearm related deaths.	Opinion	Liberal
Most of the problematic shooting events were led by mentally ill people.	Opinion	Conservative

Table 1: Examples of gun control related statements that we used in our study.

a set of statements related to it. We created these statements by first reviewing gun control related debate transcripts on an online debate platform DEBATE.ORG, and extracted the main talking points (e.g., gun violence, illegal guns) from both the supporters and opponents of gun control. Given a talking point, we extracted factual statements related to it from the latest Wikipedia pages, and rewrote them slightly to remove obvious cues indicating the statements as factual (e.g., statistics). To create opinion statements related to the talking point, we then adopted the Wikipedia neutral point of view (NPOV) criteria[3] to identify those opinionated arguments made by participants on DEBATE.ORG on this point that violate the NPOV criteria. In the end, we obtained a set of 12 statements, and Table 1 shows some example of the statements.

Next, we posted a human intelligence task (HIT) on Amazon Mechanical Turk (MTurk) to recruit workers to evaluate this set of statements. Our HIT was open to U.S. workers only. Each worker was asked to review all 12 statements in the HIT. For each statement, the worker was asked to decide whether it is a "factual statement," regardless of whether they think it is accurate or not, or an "opinion statement," regardless of whether they agree with it or not. We also included the "I don't know" (IDK) option in each task, in case workers are not sure about their answer. We further inserted an attention check question in the HIT, in which workers were instructed to select a pre-defined option. Finally, we asked workers to self-report their political stance on a 7-point Likert scale, with 1 representing very liberal, 4 representing neutral, and 7 representing very conservative.

In total, 110 workers completed our HIT and passed the attention check, among whom 57 were leaning liberal, 42 were leaning conservative, and 11 were neutral. Out of $110 \times 12 = 1320$ labels generated by these workers, we obtained 107 IDK labels (i.e., 8.1% of the labels are IDK)[4]. We considered the IDK labels as absent and did not include them in our further analyses.

4.2 Understanding Worker's Confirmation Bias

We start by understanding whether workers actually exhibited any confirmation bias when labeling factual and opinion statements in our HIT. To characterize the values that different statements express, we recruited another 47 MTurk workers to review these statements and determine that in a debate about gun control, whether the statement would be more likely used by a person holding liberal views or conservative views as their argument. For each statement, we took the majority answer from MTurk workers as the values of the statement (see Table 1 for examples).

Similar as the method used in [Hube *et al.*, 2019], we focused on analyzing worker's incorrect annotations to quantify the worker's confirmation bias. Specifically, for a worker i, we categorized her mistakes into four types and computed the the misclassification rates correspondingly:

- $\mathbf{ER}_{\mathbf{L}}^{\mathbf{fct}\rightarrow\mathbf{opn}}(i)$: among all factual statements with liberal values, the fraction of statements that worker i incorrectly labeled as opinion statements
- $\mathbf{ER}_{\mathbf{C}}^{\mathbf{fct}\rightarrow\mathbf{opn}}(i)$: among all factual statements with conservative values, the fraction of statements that worker i incorrectly labeled as opinion statements
- $\mathbf{ER}_{\mathbf{L}}^{\mathbf{opn}\rightarrow\mathbf{fct}}(i)$: among all opinion statements with liberal values, the fraction of statements that worker i incorrectly labeled as factual statements
- $\mathbf{ER}_{\mathbf{C}}^{\mathbf{opn}\rightarrow\mathbf{fct}}(i)$: among all opinion statements with conservative values, the fraction of statements that worker i incorrectly labeled as factual statements

If workers were indeed influenced by confirmation bias during their annotation process, we expect that for workers holding liberal (conservative) views, they have larger (smaller) $\text{ER}_{\text{C}}^{\text{fct}\rightarrow\text{opn}}$ and $\text{ER}_{\text{L}}^{\text{opn}\rightarrow\text{fct}}$, but smaller (larger) $\text{ER}_{\text{L}}^{\text{fct}\rightarrow\text{opn}}$ and $\text{ER}_{\text{C}}^{\text{opn}\rightarrow\text{fct}}$. Therefore, we define the following metric to represent the bias of worker i:

$$\begin{aligned}\text{bias}_i = \text{zscore}(\text{ER}_{\text{C}}^{\text{fct}\rightarrow\text{opn}}(i)) + \text{zscore}(\text{ER}_{\text{L}}^{\text{opn}\rightarrow\text{fct}}(i)) \\ - \text{zscore}(\text{ER}_{\text{L}}^{\text{fct}\rightarrow\text{opn}}(i)) - \text{zscore}(\text{ER}_{\text{C}}^{\text{opn}\rightarrow\text{fct}}(i))\end{aligned} \tag{2}$$

where $\text{zscore}(\cdot)$ represents the function standardizing the misclassification rates within each category (i.e., $\text{zscore}(x) = \frac{x-\bar{x}}{\sigma}$). Intuitively, the larger bias_i is, the more worker i favors information with liberal values.

To see whether workers indeed showed the tendency to favor information that was consistent with their own values, we look into the relationship between workers' self-reported political stance and the computed bias scores on them. Considering workers' annotations on all 12 statements, the average bias scores for liberal, neutral, and conservative workers are 0.18, -0.47, and -0.12, respectively, and we find a negative, albeit non-significant, correlation between workers' stance and their bias scores (Pearson correlation coefficient $\rho = -0.086; p = 0.374$). This means that compared to neutral and conservative workers, liberal workers indeed favored information with liberal values slightly more, implying some degree of confirmation bias. More interestingly, as shown in

[3]https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view.

[4]The mean and median number of tasks that a worker selected the IDK label was 0.97 and 0, respectively, and the number of IDK labels selected by individual workers showed a long-tail distribution. We also found that workers who self-reported as neutral tended to select the IDK label more frequently than workers who self-reported as leaning liberal or conservative (the percentage of IDK labels given among all labels generated by the liberal, neutral, and conservative workers were 6.0%, 16.7%, and 8.7%, respectively).

Top N	Correlation coefficient (ρ)	p-value
1	-0.192	0.044
2	-0.243	0.011
3	-0.255	0.007
4	-0.182	0.057
5	-0.221	0.021

Table 2: Considering only the N most difficult tasks (i.e., the top N statements with the lowest average labeling accuracy), the negative correlation between worker's stance and bias score is significant.

Table 2, when we restrict our attention to the subset of statements that are most difficult for workers (i.e., worker's average accuracy on the statement was the lowest among all 12 statements), we see significant negative associations between worker's stance and bias score, suggesting that workers might be influenced by their confirmation bias to the largest degree on the difficult tasks.

4.3 Evaluating Label Aggregation Performance

We now move on to compare the effectiveness of the proposed algorithm in accurately inferring the ground-truth labels of different tasks against baseline methods. In particular, we consider the following seven baseline methods:

- **Majority vote (MV)**: the ground-truth label of a task is the majority vote over all labels on that task.
- **GLAD**: the algorithm proposed in [Whitehill *et al.*, 2009] which assumes a worker's label on a task is affected by both the worker's skill and the task difficulty.
- **Multi**: the algorithm proposed in [Welinder *et al.*, 2010] that models each annotator as a multidimensional entity to capture the worker's diverse skills on various latent topics.
- **VI-BP**: the algorithm proposed in [Liu *et al.*, 2012] that transforms the label aggregation problem into a standard inference problem in graphical models and solves it via belief propagation.
- **Minimax (MM)**: the algorithm proposed in [Zhou *et al.*, 2012] which assumes a separate probabilistic distribution for each worker-task pair, and uses a minimax entropy method to infer ground-truth labels for each task.
- **ZenCrowd (ZC)**: the algorithm proposed in [Demartini *et al.*, 2012] which iteratively estimates worker reliability, removes unreliable workers, and infers ground-truth labels.
- **CBCC**: the algorithm proposed in [Venanzi *et al.*, 2014] which assumes communities exist within workers and those workers belonging to the same community share similar misclassification pattern.

Note that *none* of these baseline algorithms explicitly accounts for worker's confirmation bias when aggregating crowdsourced labels. We implemented these baseline algorithms using the open-sourced code repository provided by Zheng *et al.* [2017]. We further implemented the proposed bias-aware label aggregation algorithm, and we terminated the EM-based inference after convergence or 1000 iterations, whichever was reached earlier[5]. In total, we implemented eight different label aggregation algorithms.

[5]To account for the impact of parameter initialization on the performance of the algorithm, we deployed an empirically effective heuristic to restart the EM algorithm. We ran EM for three times. For all three runs, we adopted a relatively uninformative initialization for p_i, π, and a ($p_i = 0.5, \pi = 0.5$, and $a = 2$). Then, in the first EM, we initialized $c_i = 0.5$ and initialized all statements' values from one extreme (e.g., $s_j = 1$), hoping that this run of EM would return an accurate ordering of c_i. Then, in the second EM, we initialized $s_j = 0.5$ and $c_i = 1$, hoping to get an accurate ordering of s_j. In the third EM, we initialized c_i (s_j) using the final c_i (s_j) values from the first (second) EM. In the end, we report the inference results from the EM run that gives the highest likelihood of the data.

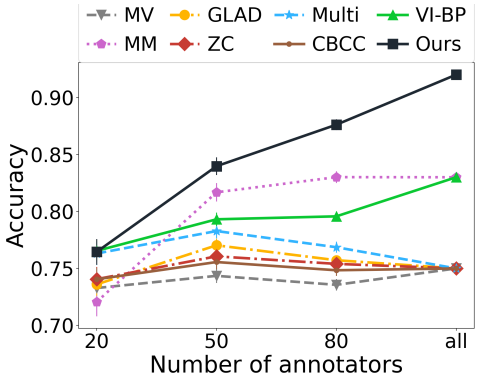


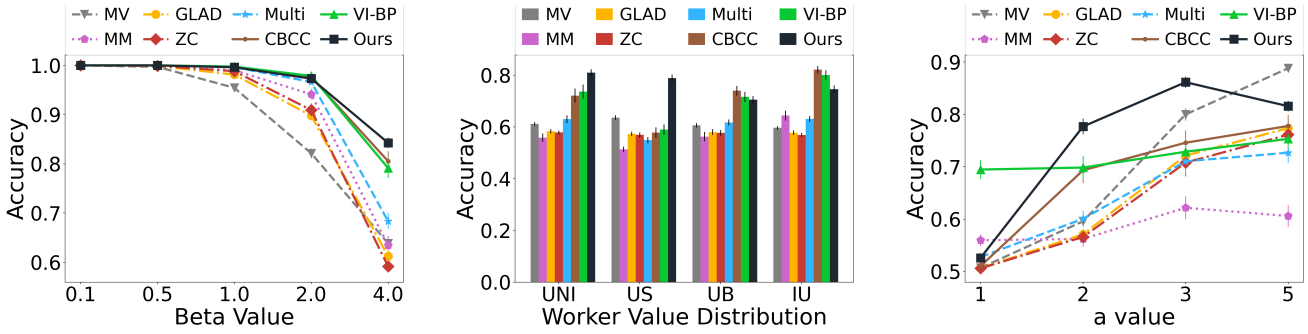
Figure 2: Comparing the performance of different algorithms in accurately inferring ground-truth labels on the real-world dataset, as the number of annotators increases. Error bars represent the standard errors of the mean. Note that uncertainty in inference accuracy due to random sampling does not exist when all worker's annotations are used in the inference (i.e., "all" in the x-axis).

We applied all these eight algorithms on the annotations that we collected from MTurk workers for differentiating factual and opinion statements, and inferred the ground-truth label for each statement. Figure 2 compares the accuracy of the inferred labels when using different algorithms. In addition to making inference using the entire set of annotations from all 110 workers, to see how the accuracy of the inference varies with the number of annotators, we also randomly sampled annotations from K ($K \in \{20, 50, 80\}$) workers and inferred the ground-truth label for each statement using only the subset of annotations provided by these K workers. For each K, we repeated the random sampling process for 100 times, and the average accuracy of the inferred labels across 100 trials is presented in Figure 2 for each algorithm. Clearly, we find that by taking worker's confirmation bias into consideration, our proposed label aggregation algorithm almost always achieves higher inference accuracy than all baseline algorithms, and its advantage over baseline algorithms becomes more salient as the number of annotators increases.

5 Simulation

Finally, we conduct simulations on synthetic datasets to explore when the proposed bias-aware label aggregation algorithm shows the largest advantages over baseline algorithms.

The Impact of Confirmation Bias Degree. First, we examine that compared to baseline algorithms, how the performance of the proposed algorithm changes with the degree to which crowd workers are subject to confirmation bias. To



(a) (b) (c)

Figure 3: Comparing the performance of different algorithms in accurately inferring ground-truth labels on synthetic datasets as the degree that workers suffer from confirmation bias changes (3a), the distribution of worker's values changes (3b), or the tendency for workers to provide the preferable label changes (3c). Error bars represent the standard errors of the mean.

do so, we generated synthetic datasets of worker annotations following the label generation model that we describe in Section 3. In particular, for each dataset, we randomly created $M = 100$ tasks. For each task, the values it took was drawn uniformly randomly between 0 and 1 (i.e., $s_j \sim U[0, 1]$), and with 50% chance it had the preferable label (i.e., $z_j \sim$ Bernoulli(0.5)). We then simulated a group of $N = 25$ workers by setting $a = 2$, sampling each worker's values uniformly randomly between 0 and 1 (i.e., $c_i \sim U[0, 1]$), and setting $p_i \sim$ Beta$(1, \beta)$. Intuitively, the larger the value of β is, the more crowd workers suffer from confirmation bias.

To simulate different degrees of confirmation bias, we considered five different values of β: 0.1, 0.5, 1, 2, 4. For each value of β, we generated 50 synthetic datasets by simulating worker's annotation on each task according to Eqn. 1. Given a specific dataset, we next used all eight label aggregation algorithms to infer the ground-truth label for each task. Figure 3a shows how the inference accuracy of different algorithms, averaged over the 50 datasets, changes with β. It is clear that as crowd workers suffer more from confirmation bias (i.e., β increases), while the inference accuracy of all algorithms decreases, the advantage of our bias-aware algorithm over the baseline algorithms becomes larger. In other words, using the proposed algorithm to aggregate crowd-generated annotations is especially helpful when crowd workers exhibit a higher level of confirmation bias.

The Impact of the Distribution of Worker's Values. We next explore how the distribution of crowd workers' own values affects the performance comparison between different label aggregation algorithms. We again simulated annotations from $N = 25$ workers on $M = 100$ tasks. For each task, $s_j \sim U[0, 1]$ and $z_j \sim$ Bernoulli(0.5), while for each worker, $a = 2$ and $p_i \sim$ Beta$(1, 5)$. For each worker's values c_i, we considered four types of distributions from which it could be randomly drawn: (1) *Uniform (UNI)*: $c_i \sim$ Beta$(1, 1)$, reflecting the case that crowd workers' values are uniformly spread over the spectrum; (2) *U-shape (US)*: $c_i \sim$ Beta$(0.5, 0.5)$, reflecting the case that crowd workers are polarized and tend to hold divergent and extreme values; (3) *Unbalanced (UB)*: $c_i \sim$ Beta$(1, 2)$, reflecting the case that most workers lean towards one extreme on the spectrum of values; and (4) *Inverse-U shape (IU)*: $c_i \sim$ Beta$(2, 2)$, reflecting the case that most workers lean towards the middle of the spectrum of values. Again, given a specific values distribution, we simulated 50 synthetic worker annotation datasets, and the average inference accuracy of different label aggregation algorithms is shown in Figure 3b. Here, we observe that the advantage of our bias-aware algorithm is particularly salient when workers' values are widely dispersed or even polarized. When workers' values lean towards one extreme or the middle of the spectrum—that is, when most workers' values are somewhat similar—the performance of the proposed algorithm is on par with the best-performing baseline algorithms (i.e., CBCC and VI-BP).

The Impact of Base Rate of the Preferable Label. Lastly, we look into how worker's tendency of providing the preferable label (i.e., worker's "positive bias") changes the performance of various algorithms. We simulated 50 datasets, with each dataset containing $N = 25$ workers and $M = 100$ tasks. Further, we set $s_j \sim U[0, 1]$, $z_j \sim$ Bernoulli(0.5), $c_i \sim$ Beta$(1, 1)$, and $p_i \sim$ Beta$(1, 5)$. We then varied $a \in \{1, 2, 3, 5\}$, and Figure 3c presents the average inference accuracy of different label aggregation algorithms. Interestingly, we find the proposed algorithm has the largest advantage over baseline algorithms when workers have a moderate level of base rate of providing the preferable label. When workers have very high base rates to provide the preferable label (e.g., $a = 1$), the proposed algorithm performs worse than the VI-BP algorithm. On the other hand, when workers are unlikely to provide the preferable label (e.g., $a = 5$), while the proposed algorithm outperforms many baselines, a simple majority vote can be the most effective aggregation strategy if the distribution of worker's values is balanced.

6 Conclusion

Crowdsourcing has become a prevalent tool for gathering data from humans. As humans are often subject to various types of biases, the challenge of how to carefully process the crowdsourced data to minimize the negative impact that people's biases bring to data quality becomes pressing. In this paper, we focus on confirmation bias, a particular type of cognitive bias, and propose a new label aggregation algorithm based on a quantitative model which characterizes how crowd workers are influenced by their confirmation bias in their annotations. The evaluation results on both real-world data and synthetic data demonstrate the effectiveness of our proposed method.

References

[Allahbakhsh *et al.*, 2013] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, 17(2):76–81, 2013.

[Biswas *et al.*, 2020] Arpita Biswas, Marta Kolczynska, Saana Rantanen, and Polina Rozenshtein. The role of in-group bias and balanced data: A comparison of human and machine recidivism risk predictions. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 97–104, 2020.

[Braylan and Lease, 2020] Alexander Braylan and Matthew Lease. Modeling and aggregation of complex annotations via annotation distances. In *Proceedings of The Web Conference 2020*, pages 1807–1818, 2020.

[Coscia and Rossi, 2020] Michele Coscia and Luca Rossi. Distortions of political bias in crowdsourced misinformation flagging. *Journal of the Royal Society Interface*, 17(167):20200020, 2020.

[Demartini *et al.*, 2012] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*, pages 469–478, 2012.

[Doroudi *et al.*, 2016] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2623–2634, 2016.

[Duan *et al.*, 2020] Xiaoni Duan, Chien-Ju Ho, and Ming Yin. Does exposure to diverse perspectives mitigate biases in crowdwork? an explorative study. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 155–158, 2020.

[Eickhoff, 2018] Carsten Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 162–170, 2018.

[Ho *et al.*, 2015] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, pages 419–429, 2015.

[Huang *et al.*, 2018] Jianrui Huang, Xianyou He, Xiaojin Ma, Yian Ren, Tingting Zhao, Xin Zeng, Han Li, and Yiheng Chen. Sequential biases on subjective judgments: Evidence from face attractiveness and ringtone agreeableness judgment. *Plos one*, 13(6):e0198723, 2018.

[Hube *et al.*, 2019] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[La Barbera *et al.*, 2020] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. Crowdsourcing truthfulness: The impact of judgment scale and assessor bias. In *European Conference on Information Retrieval*, pages 207–214. Springer, 2020.

[Li *et al.*, 2020] Yanying Li, Haipei Sun, and Wendy Hui Wang. Towards fair truth discovery from biased crowdsourced answers. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '20, page 599–607, 2020.

[Liu *et al.*, 2012] Qiang Liu, Jian Peng, and Alexander T Ihler. Variational inference for crowdsourcing. *Advances in neural information processing systems*, 25:692–700, 2012.

[Mitchell *et al.*, 2018] Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Nami Sumida. Distinguishing between factual and opinion statements in the news. https://www.journalism.org/2018/06/18/distinguishing-between-factual-and-opinion-statements-in-the-news/, June 2018. Accessed: 2021-05-21.

[Newell and Ruths, 2016] Edward Newell and Derek Ruths. How one microtask affects another. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3155–3166, 2016.

[Nickerson, 1998] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[Otterbacher *et al.*, 2019] Jahna Otterbacher, Pınar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. How do we talk about other people? group (un) fairness in natural language image descriptions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 106–114, 2019.

[Tang *et al.*, 2019] Wei Tang, Ming Yin, and Chien-Ju Ho. Leveraging peer communication to enhance crowdsourcing. In *The World Wide Web Conference*, pages 1794–1805, 2019.

[Venanzi *et al.*, 2014] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*, pages 155–164, 2014.

[Welinder *et al.*, 2010] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23:2424–2432, 2010.

[Whitehill *et al.*, 2009] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22:2035–2043, 2009.

[Zheng *et al.*, 2015] Yudian Zheng, Jiannan Wang, Guoliang Li, Reynold Cheng, and Jianhua Feng. Qasca: A quality-aware task assignment system for crowdsourcing applications. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 1031–1046, 2015.

[Zheng *et al.*, 2017] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proc. VLDB Endow.*, 10(5):541–552, January 2017.

[Zhou *et al.*, 2012] Dengyong Zhou, Sumit Basu, Yi Mao, and John Platt. Learning from the wisdom of crowds by minimax entropy. *Advances in neural information processing systems*, 25:2195–2203, 2012.

[Zhuang *et al.*, 2015] Honglei Zhuang, Aditya Parameswaran, Dan Roth, and Jiawei Han. Debiasing crowdsourced batches. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1593–1602, 2015.