Truth Discovery for Long-term Crowdsensing

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

***Abstract*—Crowdsensing has become a creative data collection paradigm owing to the popularity of mobile devices and it usually assigns each task to multiple workers for data quality. Due to different capabilities of workers and conflicts among these reported answers, truth discovery has emerged to infer the truths and estimate previously unknown reliability of workers simultaneously. The existing truth discovery works on crowdsensing focus on improving the accuracy of inferred truths. However, the long-term crowdsensing has been ignored for those works where the data collection is a long-term activity and the platform should periodically publish some aggregated data for use. If only using traditional truth discovery in each period, the estimation of reliability can be unconvincing due to the insufficient participation of workers, which finally degrades the accuracy of truths.**

To tackle this challenge, we propose a novel truth discovery mechanism to efficiently detect truths in long-term crowdsensing. We leverage a specific phenomenon in long-term crowdsensing that some workers may contribute data in previous periods and cumulatively uses history data to generate prior beliefs of workers. Then the truth discovery problem can be modelled as a maximum-a-posterior (MAP) estimation. Moreover, considering new workers without history data, we introduce the confidence interval to initialize their prior beliefs. Different from the existing works that assume the reliability of each worker is changeless, we consider the rapidly change of reliability in this paper due to the long-term working. Experiments with several state-of-the-art truth discovery approaches demonstrate that the proposed mechanism performs better than existing works in long-term crowdsensing.

***Index Terms*—long-term crowdsensing, truth discovery.**

I. Introduction

Due to the widespread popularity of mobile devices with abundant embedded sensors (e.g., camera, accelerator, microphone, compass and GPS), crowdsensing has been proposed to collect data for sensing applications. The generated sensor data can provide useful information for environmental monitoring, healthcare, smart transportation and many others. However, because the quality of reported data varies from worker to worker, multiple workers are expected to perform each task so that the quality of data can be guaranteed. Therefore, it is a challenge work to extract correct answers (truths) from the noise, conflicting and heterogeneous data contributed by various workers with mobile devices.

In crowdsensing, some straightforward approaches to get truths by aggregating data are Average and Majority Voting (MV), which treat all workers as equal. However, it is unreasonable to have this assumption because different workers may contribute data with different qualities due to a variety of mobile devices, let alone different abilities and attitudes from workers. Moreover, these naive methods can be easily attacked by malicious workers and spammers, who deliberately report the fake or noisy information. Hence, the reliability of workers should be added into truth inference and the answers from workers with high reliability naturally count heavily on final aggregated data.

Due to the lack of ground truth about each task, the reliability of each worker is difficult to estimate and usually unknown a priori. Hence, the truth discovery that estimates reliability of workers and infers truths of tasks simultaneously has been widely studied []. The principle of existing works can be summarized as: The workers that provide true answers more often will be assigned higher reliability degrees, and the answer reported by reliable workers will be regarded as truths. Based on this general principle, different truth discovery approaches have been proposed for various scenarios, where they usually make different assumptions about task types (e.g., decision-making [], single-choice [], numeric []). Moreover, some recent works also consider difficulty of tasks and assume that each worker has different quality for different tasks [].

However, a common scenario in crowdsensing has been ignored in these works, that is long-term crowdsensing where the data collection can continue for a long time and the platform should publish data periodically. That is to say, the platform would republish some new tasks at the start of each period and aggregate those data contributed by workers to infer truths about tasks at the end of that period. However, different from doing micro tasks online, workers in long-term crowdsensing can only perform a few tasks due to the movement cost or limited time during each period, which incurs collected data in this period hard to estimate the reliability of workers convincingly. To solve the problem of insufficient participation, a confidence-aware approach was proposed for truth discovery [], which considered the confidence interval of estimation for reliability so that source reliability with various participation can be reflected. In their assumption, the workers with less claims (participation) will be given a lower reliability even though their answers mostly are correct, which will degrade the accuracy of truth inference. This unneglectable problem and shortages in existing works motivate us to design a novel truth discovery mechanism built for long-term crowdsensing.

In this paper, we consider the different data types in long-term crowdsensing and propose two truth discovery mechanisms to efficiently infer truths of categorical data and continuous data tasks respectively at each period before publication. Generally, many workers in long-term crowdsensing would ever perform tasks and contribute data in previous periods. Hence, we leverage the history data of each worker as prior beliefs to cumulatively estimate its reliability without frequently revisiting previous data, which guarantees the efficiency. The truth discovery problem with prior beliefs can be modelled as a maximum-a-posterior (MAP) estimation and then the Expectation Maximization (EM) algorithm will be used for solving. Moreover, considering new arrival workers without history data, we initialize their prior beliefs by analyzing

history of other workers by assuming that the reliability of workers satisfies Normal distribution, where the confidence interval will be used for parameter estimation. Different from the existing works that assume the reliability of each worker is changeless, we consider the rapidly change of reliability in this paper due to the long-term working. The main contributions of this paper are summarized as follows:

- To the best of our knowledge, we are the first to focus on the truth discovery problem in long-term crowdsensing where the participation of workers is insufficient in each data collection period and the accuracy of inferred truths may be degraded.
- We propose two truth discovery mechanisms respectively for categorical data tasks and continuous data tasks that leverage previous data of workers into reliability estimation phase and the confidence interval is employed to estimate reliability of new workers.
- Due to the long-term working, the change of reliability of workers is considered in this paper and we propose a change detection method to monitor it so that the truth discovery in this period would not be controlled by previous data.
- We conduct extensive experiments to compare the proposed truth discovery mechanism with state-of-the-art methods. The experiments results show that ***

The remainder of this paper is organized as follows. We briefly discuss the existing truth discovery for crowdsensing systems in Section II. We present the long-term crowdsensing system model and formulate truth discovery problem in Section III. We introduce the truth discovery mechanism for categorical data in Section IV, and the consider the continuous data in Section V. We evaluate the performance of the proposed algorithms in Section VI and finally conclude the paper in Section VII.

II. Related Work

In crowdsensing, resolving conflicts from multiple workers and inferring the final truths of tasks have become an important research aspect of this area. Truth discovery has been proposed to address this problem which can estimate the reliability of workers and infer the truths of tasks simultaneously. In following, we will discuss the existing works on truth discovery in crowdsensing according to task types, which can be classified into categorical data tasks and continuous data tasks.

In categorical data tasks, workers need to select a single choice (multiple choices) from the candidate answers. The platform need to infer the final correct answer for each task from the reported answers from workers. In [1], [2], the authors set a value w (weight) for each worker, which represents the probability that each worker answers tasks correctly. Aydin et al. [1] modelled the truth discovery problem as a weighted distance minimization problem, where the distance measured the difference between the reported answers and inferred correct answers. Demartini et al. [2] modelled the problem as a probabilistic graphical model where the variables and conditional dependency structures can constitute a graph. In [3], the authors focused on resolving conflicts in heterogeneous data and leveraged the joint inference to reduce the error rate. Different from assuming each worker have the same quality to answer different tasks, some existing works modelled each worker have different quality for different tasks. In [4], the authors assumed different tasks have different difficulty levels and the probability of a worker answering correctly for a task was determined by its quality and difficulty level. Moreover, some works [5], [6], [7] assumed each worker has various levels of expertise for different topics, where the quality of worker was modelled as a vector according to the size of topics. In some works [8], [9], [10], confusion matrix was used to model the probability that a worker answers ith choice when the truth of task is jth.

In continuous data tasks, each worker needs to provide a value for each task. The platform will derive a final value as truth for each task by integrating values provided from workers. In [1], the authors modelled the truth inference problem as a optimization problem and aimed to minimize the weighted distance of all tasks. Some works [11][12] assumed that the value reported by each worker follows Gaussian distribution, where bias and variance were used to capture the reliability of workers. Namely, a worker with less bias has more accurate estimation for tasks and a higher variance means a larger variation of errors around bias. In [13][14][15], the authors considered confidence interval in modeling quality of each worker so that the number of performed tasks can also influence the quality of each worker. The more tasks a worker performs, the more confident the estimated quality will be.

However, the existing works did not consider the long-term crowdsensing and the characteristics in this scenario have been ignored for those works. Hence, we focus on the long-term crowdsensing and propose two novel truth discovery for two type tasks respectively.

III. Preliminaries

In this section, we first introduce the system model of data collection in long-term crowdsensing, and then describe the truth discovery problem.

A. System Model

We consider the long-term crowdsensing applications which leverages workers with mobile devices to collect sensing data in a long-term and needs to publish the aggregated data periodically (e.g., each day). In each period, the platform update a group tasks and allocate them to workers. According to the type of contributed data, the tasks can be divided into continuous data (numeric) tasks and categorical data (e.g., single choice, multiple choice) tasks. More specifically, we take the city noise monitoring as an example of continuous data tasks. In this scenario,workers move to some particular regions and measure their ambient noise value, which will be reported as a numeric to the platform. As for the categorical data tasks, the platform can leverage workers to observe that whether parking spaces are unoccupied or street lamps are broken in some locations, where the reported answer can be true or false. However, due to different sensing capabilities and

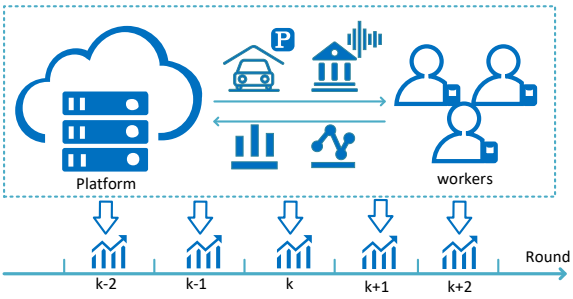

Fig. 1. The architecture of the long-term crowdsensing system.

morality levels of workers, each objective region needs multiple workers to measure for data accuracy. Once receiving the data from workers, the platform will aggregate the contributed data about each task and infer the truth before publishing.

As shown in Figure 1, the crowdsensing system can be made up of a platform that publishes tasks and massive mobile users who can participate in data collection as workers. As for the long-term crowdsensing, the data collection process can be perennial and periodic, which consists of multiple periods or rounds. At each round, the collected data from workers will be aggregated and the inferred truths will be published to the public or some organizations for use.

B. Problem Formulation

At sensing round k, the platform publishes a set of tasks which can be denoted by $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ and we assume that the task set does not change once determined at the beginning. Moreover, the truth of tasks in kth round can be denoted by $\mathcal{X}_k = \{x_1^k, x_2^k, \ldots, x_m^k\}$, which should be updated according to the collected data at each round . Let $\mathcal{W} = \{w_1, w_2, \ldots\}$ denote the set of workers who participate in each round randomly, and $\mathcal{W}_k \in \mathcal{W}$ denotes the set of workers who participate in round k and the size of $\mathcal{W}_k$ is N_k. Let $\mu_k = \{\mu_i^k | w_i \in \mathcal{W}_k\}$ be the reliability vector of workers in kth round where μ_i^k measures the accuracy of data from w_i. Generally, the higher of μ_i^k, the data contributed by w_i counts more heavily on the final truth in kth round. For each worker w_i in $\mathcal{W}_k$, the contributed data about task t_j can be denoted by x_{ij}^k, which can be null if worker w_i does not contribute data for task t_j. Given the truth discovery dataset $\mathcal{D}_k$ about m tasks from $\mathcal{W}_k$ in round k, that is $\mathcal{D}_k = \{x_{ij}^k | w_i \in \mathcal{W}_k, t_j \in \mathcal{T}\}$, the objective is to estimate the reliability vector of workers μ_k and the unknown truths $\mathcal{X}_k$.

Assuming the data are independently collected, the likelihood function of the parameters μ_k given the answers $\mathcal{D}_k$ can be factored as

$$Pr[\mathcal{D}_k | \mathcal{X}_k, \mu_k] = \prod_{j=1}^{m} \prod_{w_i \in \mathcal{W}_k} Pr[x_{ij}^k | \mathcal{X}_k, \mu_k] \quad (1)$$

where the truth $\mathcal{X}_k$ is regarded as latent variable. Hence the maximum likelihood estimation of parameter μ_k can be presented by

$$\mu_k = arg\, max_{\mu_k} \{ln(Pr[\mathcal{D}_k | \mathcal{X}_k, \mu_k])\} \quad (2)$$

where the latent variable $\mathcal{X}_k$ is also estimated along with the μ_k. However, due to the insufficient participation of workers, the $\mathcal{D}_k$ appears to be sparse on the dimension of worker, which makes the estimation of μ_k unconvincing. Consider the special feature in long-term crowdsensing that workers may participate in past rounds, the history data of workers can be used to strengthen the estimation of μ_k. However, there are some realities in long-term crowdsensing which may limit the effectiveness of using history data. Firstly, due to the long time, the reliability of some workers will happen to change because of some special states in this round such as unhappy or unskilled with new mobile devices. Secondly, workers can decide whether to participate in each round as their minds, which incurs huge differences on history data distribution among different workers and brings some new arrivals without history data. Thirdly, the volume of data will exponentially grow with the increase of rounds, which reduces the efficiency of truth discovery if revisiting data in previous rounds. Hence, the problem of reliability estimation can also be a challenge because of the above realities in long-term crowdsensing.

In the following, due to the different calculation of $Pr[\mathcal{D}_k | \mathcal{X}_k, \mu_k]$ in categorical data and continuous data and differences on reliability or truth calculation, we will respectively introduce the novel truth discovery mechanism suitable for long-term crowdsensing according to the data type.

IV. Truth Discovery for Categorical Data

In this section, due to the ubiquity of binary classification in crowdsensing (e.g., parking space monitoring, accident monitoring) and its intelligibility in mathematics, we will use the binary data ($x_i^k = \{0, 1\}$) to introduce the truth discovery mechanism in long-term crowdsensing, which is a specific situation in categorical data. The truth discovery mechanism for usual categorical data can be easily extended from the mechanism for binary data.

The biggest problem of long-term crowdsensing is that the participation of workers may be insufficient to convincingly estimate the reliability of workers (weights) at current round. Fortunately, the history data of workers can be used to constitute their prior belief in reliability. For example, worker w_i is known to contribute 10 correct answers and 5 wrong answers in previous rounds, which is prior belief. The basic idea of truth discovery is to estimate the probability of each worker answering correctly in current round based on the prior belief and infer the truths of tasks, which can be called the posterior belief. For example, if worker w_i made 5 answers in this round of which 4 were correct, the belief (posterior belief) in reliability of w_i at the end of this round becomes 14 out of 20. Moreover, the posterior belief of this round becomes the prior belief for the subsequent round, which eliminates the burden to revisit many history data and efficiently reduces the computation cost.

Hence, with the prior belief, the Bayes' theorem can be applied to our problem, that is maximum-a-posterior (MAP)

estimation

$$\begin{aligned}\mu_k &= arg\,max_{\mu_k} ln(Pr[\mu_k|\mathcal{D}_k,\mathcal{X}_k]) \\ &= arg\,max_{\mu_k} ln(\frac{Pr[\mathcal{D}_k|\mathcal{X}_k,\mu_k]Pr[\mu_k]}{Pr[\mathcal{D}_k]}) \\ &\propto arg\,max_{\mu_k}\{ln(Pr[\mathcal{D}_k|\mathcal{X}_k,\mu_k]) + ln(Pr[\mu_k])\}\end{aligned} \tag{3}$$

where $Pr[\mu_k|\mathcal{D}_k,\mathcal{X}_k])$ denotes the posterior probability of reliability. However, posterior probability is hard to calculate, which can be approximated with likelihood $Pr[\mathcal{D}_k|\mathcal{X}_k,\mu_k])$ and prior probability $Pr[\mu_k]$ as shown in Equation 3. The beta distribution $Beta(a,b)$ is known to model the success rate of a event (e.g., baseball player's batting rate), which can be used to approximate the reliability of each worker (the correctness of answers from each worker). For any $a > 0$, $b > 0$, and $\delta \in [0,1]$ the probability density function (pdf) of beta distribution $Beta(a,b)$ is given by

$$f(\delta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\delta^{a-1}(1-\delta)^{b-1} \tag{4}$$

where $\Gamma(x)$ is the gamma function. For worker w_i, the prior probability of reliability μ_i^k can be denoted by

$$Pr[\mu_i^k|\alpha_i^k,\beta_i^k] = f(\mu_i^k|\alpha_i^k+1,\beta_i^k+1) \tag{5}$$

where α_i^k and β_i^k (prior beliefs) respectively denotes the number of times worker w_i correctly or wrongly answered before round k. As for the likelihood $Pr[\mathcal{D}_k|\mathcal{X}_k,\mu_k])$, it can be decomposed as

$$\begin{aligned}Pr[\mathcal{D}_k|\mathcal{X}_k,\mu_k]) = \prod_{j=1}^{m}\{&Pr[x_j^k=1]Pr[D_k(:,j)|x_j^k=1,\mu_k] \\ &+ Pr[x_j^k=0]Pr[D_k(:,j)|x_j^k=0,\mu_k]\}\end{aligned} \tag{6}$$

where $D_k(:,j)$ denotes the jth column of D_k, that is, $D_k(:,j)$ includes all answers for task t_j. Since workers make their decision independently, given the truth answer x_j^k, we have

$$\begin{aligned}Pr[D_k(:,j)|x_j^k=1,\mu_k] &= \prod_{w_i\in\mathcal{W}_k} Pr[x_{ij}^k|x_j^k=1,\mu_k] \\ &= \prod_{w_i\in\mathcal{W}_k}(\mu_i^k)^{x_{ij}}(1-\mu_i^k)^{1-x_{ij}}\end{aligned} \tag{7}$$

Similarly, we have

$$\begin{aligned}Pr[D_k(:,j)|x_j^k=0,\mu_k] &= \prod_{w_i\in\mathcal{W}_k} Pr[x_{ij}^k|x_j^k=0,\mu_k] \\ &= \prod_{w_i\in\mathcal{W}_k}(\mu_i^k)^{1-x_{ij}}(1-\mu_i^k)^{x_{ij}}\end{aligned} \tag{8}$$

A. EM algorithm

Considering the latent variable $\mathcal{X}_k$, the Expectation Maximization (EM) algorithm can be used for MAP estimation, which is summarized below:

1) We let $\pi_j = Pr[x_j^k=1|D_k(:,j),\mu_k]$, and then initialize π_j as:
$$\pi_j = \frac{1}{|D_k(:,j)|}\sum_{w_i\in\mathcal{W}_k} x_{ij}^k \tag{9}$$
where $|D_k(:,j)|$ denotes the number of answers for t_j.
2) (**E-step**) Given the π_j, we can estimate the reliability of each worker in $\mathcal{W}_k$ as follows:
$$\mu_i^k = \frac{\alpha_i^k + \sum_{t_j\in\mathcal{T}_i^k}[\pi_j x_{ij}^k + (1-\pi_j)(1-x_{ij}^k)]}{\alpha_i^k+\beta_i^k+|\mathcal{T}_i^k|} \tag{10}$$
where $\mathcal{T}_i^k$ denotes the tasks performed by w_i in round k and $|\mathcal{T}_i^k|$ is the number of tasks in $\mathcal{T}_i^k$.
3) (**M-step**) Based on the μ_k, update π_j as:
$$\pi_j = \frac{Pr[D(:,j)|x_j^k=1,\mu_k]Pr[x_j^k=1]}{\sum_{\gamma=0,1}Pr[D(:,j)|x_j^k=\gamma,\mu_k]Pr[x_j^k=\gamma]} \tag{11}$$
where $Pr[x_j^k=1]$ can be calculated as
$$Pr[x_j^k=1] = \frac{\sum_{j=1}^{m}Pr[x_j^k=1|D_k(:,j),\mu_j]}{m} \tag{12}$$
Note that the $Pr[x_j^k=1]$ is the same for any t_j since without any additional information.

Iterate the step 2 and step 3 till convergence, when the reliability μ_k and $\pi_j = Pr[x_j^k=1|D_k(:,j),\mu_k]$ can be estimated. For the truth of each task, we have

$$x_j^k = \begin{cases}1 \text{ if } \pi_j >= 0.5\\ 0 \text{ if } \pi_j < 0.5\end{cases} \tag{13}$$

After deriving the truths of tasks at round k and corresponding reliability of workers, we need to update the beta parameters α_i^{k+1} and β_i^{k+1} for each worker as prior beliefs in subsequent round. For α_i^{k+1}, we have

$$\alpha_i^{k+1} = \alpha_i^k + \sum_{t_j\in\mathcal{T}_i^k}\mathcal{I}(x_j^k, x_{ij}^k) \tag{14}$$

where $\mathcal{I}(x,y)$ is an indicator function and the its value can be 1 if $x=y$, otherwise equals to 0. Meanwhile, the β_i^{k+1} can be updated as

$$\beta_i^{k+1} = \beta_i^k + |\mathcal{T}_i^k| - \sum_{t_j\in\mathcal{T}_i^k}\mathcal{I}(x_j^k, x_{ij}^k). \tag{15}$$

Considering the randomly arrival of workers, there exist some workers who are not involved in current round. The parameters α_i^{k+1}, β_i^{k+1} can directly inherit the value of α_i^k, β_i^k, that is

$$\alpha_i^{k+1} = \alpha_i^k,\ \beta_i^{k+1} = \beta_i^k. \tag{16}$$

B. Initialization of prior parameters for new worker

In long-term crowdsenisng, some workers may newly arrive, which makes the prior beliefs of them can be missing. Leveraging the information of other workers having participated in task performance, we can initialize the parameters α_k and β_k for those new workers. By assuming the reliability of all workers follows Normal distribution, the interval estimation can be used for our problem. Here, we introduce the definition of confidence interval.

Definition 1. *Let X be a sample set from a probability distribution with statistical parameters θ. Given samples $X_1, X_2, \ldots, X_n$ from X and definite value a, if two statistical magnitude $\underline{\theta} = \underline{\theta}(X_1, X_2, \ldots, X_n)$ and $\bar{\theta} = \bar{\theta}(X_1, X_2, \ldots, X_n)$ satisfy*

$$Pr[\underline{\theta} < \theta < \overline{\theta}] = 1 - a$$

$(\underline{\theta}, \overline{\theta})$ can be called as the confidence interval of θ with confidence level $1 - a$.

For workers with prior beliefs at round k denoted by $\mathcal{W}_{-k}$, that is $\mathcal{W}_{-k}$ represents workers who have involved in task performance before round k, the reliability μ_i^k of each worker (sample) can be approximated by the mean of beta distribution $Beta(\alpha_i^k + 1, \beta_i^k + 1)$. Generally, the reliability μ_i^k of workers follows a Normal Distribution $N(\lambda, \sigma^2)$ where λ denotes the mean and σ^2 denotes the variance. Given the samples $\mu_1^k, \mu_2^k, \ldots, \mu_\xi^k$ where ξ denotes the number of workers in $\mathcal{W}_{-k}$, we constitute a pivotal quantity to estimate the parameter λ

$$\begin{gathered} \frac{\bar{X} - \lambda}{S/\sqrt{n-1}} \sim t(n-1) \\ n = \xi, \bar{X} = \frac{\Sigma_{i=1}^n \mu_i^k}{n}, \; S = \sqrt{\frac{\Sigma_{i=1}^n (\mu_i^k - \bar{X})}{n-1}} \end{gathered} \tag{17}$$

where $\bar{X}$ denotes the sample average and S denotes the sample standard deviation. Leveraging the fractile of t-distribution, we have

$$Pr[|\frac{\bar{X} - \lambda}{S/\sqrt{n-1}}| < t_{\frac{a}{2}}(n-1)] = 1 - a \tag{18}$$

Hence, the confidence interval of λ with $1 - a$ confidence degree can be

$$(\bar{X} - \frac{S}{\sqrt{n}} t_{\frac{a}{2}}(n-1), \bar{X} + \frac{S}{\sqrt{n}} t_{\frac{a}{2}}(n-1)) \tag{19}$$

Similarly, since we have

$$Pr[\chi^2_{1-\frac{a}{2}}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\frac{a}{2}}(n-1)] = 1 - a \tag{20}$$

the confidence interval of σ^2 with $1 - a$ confidence level can be

$$(\frac{(n-1)S^2}{\chi^2_{\frac{a}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\frac{a}{2}}(n-1)}) \tag{21}$$

where $\chi^2(n-1)$ denotes the chi square distribution with $n-1$ degree of freedom. A higher confidence level will tend to produce a broader confidence interval and the 95% confidence level ($a = 0.05$) is most commonly used. For each new worker without prior beliefs, the platform can initialize its μ_i^k and $(\sigma_i^k)^2$ according to the estimated confidence interval of average λ and variance σ^2 of reliability. Considering the prosperity of network, workers without any traces are rare. The principle is to evaluate and grade each new worker based on additional knowledge (e.g., completion of other events, social media, browsing history) about it. Based on the grade $G = \{g_i | w_i \in \mathcal{W}\}$, the μ_i^k and $(\sigma_i^k)^2$ can be initialize as

$$\begin{gathered} \mu_i^k = F_\lambda(g_i), \; (\sigma_i^k)^2 = F_\sigma(g_i) \\ \mu_i^k \in [\underline{\lambda}, \overline{\lambda}], \; \sigma^2 \in [\underline{\sigma}^2, \overline{\sigma}^2] \end{gathered} \tag{22}$$

where $[\underline{\lambda}, \overline{\lambda}]$ and $[\underline{\sigma}^2, \overline{\sigma}^2]$ are confidence intervals of λ and σ^2. The design of $F_\lambda(g_i)$ and $F_\sigma(g_i)$ can be a regression problem, which is out of scope in this paper. Moreover, if there is no additional knowledge about new workers, the value of μ_i^k and $(\sigma_i^k)^2$ can be $\frac{\underline{\lambda} + \overline{\lambda}}{2}$ and $\frac{\underline{\sigma}^2 + \overline{\sigma}^2}{2}$ respectively. Since the reliability of each worker follows beta distribution, whose mean and variance are known to be

$$\begin{aligned} \mu_i^k &= \frac{\alpha_i^k + 1}{\alpha_i^k + \beta_i^k + 2} \\ (\sigma_i^k)^2 &= \frac{(\alpha_i^k + 1)(\beta_i^k + 1)}{(\alpha_i^k + \beta_i^k + 2)^2(\alpha_i^k + \beta_i^k + 3)} \end{aligned} \tag{23}$$

we can derive the α_i^k and β_i^k for each new worker by solving the above binary quadratic equation. Then, all workers can be involved into above EM algorithm for MAP estimation. The pseudocode of truth discovery for categorical (binary) data is shown in Algorithm 1.

Algorithm 1 Truth Discovery for Categorical data

Input: Dataset D_k, workers in current round $\mathcal{W}_k$ and prior beliefs $P_b = \{(\alpha_i^k, \beta_i^k) | w_i \in \mathcal{W}_{-k}\}$

Output: The truth $\mathcal{X}_k$ and updated prior beliefs P_b

1: Based on the Pb, calculate the reliability of workers $X = \{\mu_1^k, \mu_2^k, \ldots, \mu_\xi^k\}$.
2: Calculate the confidence intervals $(\underline{\lambda}, \overline{\lambda})$ and $(\underline{\sigma}^2, \overline{\sigma}^2)$ with 95% confidence level based on the samples X, which is assumed to follow normal distribution.
3: For worker w_i in $\mathcal{W}_k$ but not in $\mathcal{W}_{-k}$, initialize its μ_i^k and $(\sigma_i^k)^2$ according to the confidence intervals and then derive its beta parameters α_i^k, β_i^k.
4: Apply the EM algorithm for MAP estimation and the reliability μ_k and π_j can be estimated, while the truth $\mathcal{X}_k$ can be discovered according to the π_j.
5: Update the beta parameters α_i^{k+1}, β_i^{k+1} for all workers, which compose prior beliefs P_b for subsequent round.

C. Reliability Change Detection

V. Truth Discovery for Continuous Data

In this section, we focus on the truth discovery for continuous data tasks where the reported value x_{ij}^k and final truth x_i^k are numerical. Different from the categorical data tasks where each answer will be either correct or wrong (i.e. whether the same as or different from the truths), the correctness of answers in continuous data tasks are evaluated by the closeness to truths. Hence, the reliability of each worker is determined by the deviation of its answers from the correct answers. The smaller the deviation of answers from correct answers be, the more reliable each worker will be.

Since we cannot excavate the number of times each worker correctly or wrongly making an answer, we use the average error and the number of performed tasks as prior beliefs where the error denotes the difference between reported answer and final truth. Generally speaking, the worker is more reliable if it has lower average error and vice versa. Moreover, the worker with a higher number of performed tasks will be more reliable than others if they have the same average error. Therefore, the truth discovery problem turns to be a MAP estimation shown in Equation 3 based on the prior beliefs.

We assume the reliability of each worker follows Normal distribution $N(1/\epsilon, 1/\tau)$ where ϵ denotes the average error and

τ denotes the number of performed tasks. Then for worker w_i, the prior probability of reliability μ_i^k can be denoted by

$$Pr[\mu_i^k|\epsilon_i^k,\tau_i^k] = \frac{\tau_i^k}{\sqrt{2\pi}} exp\{-\frac{(\tau_i^k)}{2}(\mu_i^k - 1/\epsilon_i^k)^2\} \quad (24)$$

where τ_i^k and ϵ_i^k denotes the number of performed tasks and average error of worker w_i before round k respectively. Given the correct answer x_j^k and reliability μ_i^k, we assume the answer x_{ij}^k from worker w_i also follows a Normal distribution $N(x_{ij}^k, 1/\mu_i^k)$, that is

$$Pr[x_{ij}^k|x_j^k,\mu_i^k] = \frac{\mu_i^k}{\sqrt{2\pi}} exp\{-\frac{(\mu_i^k)^2}{2}(x_{ij}^k - x_j^k)^2\}. \quad (25)$$

Hence, given the answers D_k, the likelihood of parameter μ_k can be shown as

$$Pr[D_k|\mathcal{X}_k,\mu_k] = \prod_{j=1}^{m}\prod_{w_i\in\mathcal{W}_k} Pr[x_{ij}^k|x_j^k,\mu_i^k]. \quad (26)$$

A. EM Algorithm

Similarly, the EM algorithm can be applied into the MAP estimation in continuous data.

1) We use the average method to initialize the truth x_j^k for each task. That is

$$x_j^k = \frac{\sum_{w_i\in\mathcal{W}_k(j)} x_{ij}}{|\mathcal{W}_k(j)|} \quad (27)$$

where $\mathcal{W}_k(j)$ denotes workers who participate in task t_j in round k and $|\mathcal{W}_k(j)|$ denotes the number of workers in set.

2) (**E-step**) Given the truths $\mathcal{X}_k$, the reliability of each worker μ_i^k can be updated as

$$\mu_i^k = \frac{\sqrt{\sum_{t_j\in\mathcal{T}_i^k}(x_{ij} - x_j^k)^2 + \tau\epsilon^2}}{|\mathcal{T}_i^k| + \tau} \quad (28)$$

3) (**M-step**) Based on the μ_k, the truth x_j^k for task t_j can be

$$x_j^k = \frac{\sum_{w_i\in\mathcal{W}_k(j)} \mu_i^k x_{ij}}{\sum_{w_i\in\mathcal{W}_k(j)} \mu_i^k} \quad (29)$$

The truths $\mathcal{X}_k$ and reliability of workers μ_k can be derived by iterating step 2 and 3 till convergence. Since obtaining the truths of tasks at round k, we need to update the prior beliefs ϵ_i^{k+1} and τ_i^{k+1} for each worker. For parameter τ_i^{k+1}, we have

$$\tau_i^{k+1} = \tau_i^k + |T_i^k| \quad (30)$$

As for the parameter ϵ_i^{k+1}, we can calculate it by

$$\epsilon_i^{k+1} = \frac{(\tau_i^k\epsilon_i^k + \Sigma_{t_j\in\mathcal{T}_i^k}|x_{ij} - x_j^k|)}{\tau_i^{k+1}} \quad (31)$$

For workers who disappear in round k, we reserver the former recodes, that is

$$\tau_i^{k+1} = \tau_i^k,\ \epsilon_i^{k+1} = \epsilon_i^k \quad (32)$$

B. Initialization of prior parameters for new workers

As mentioned in categorical data tasks, the reliability of workers follows a Normal distribution $N(\lambda,\sigma)$, that is workers with too high reliability or too low reliability are few. For workers with prior beliefs in $\mathcal{W}_{-k}$, the reliability of each worker μ_i^k can be estimated by $1/\epsilon_i^k$, which is the mean of Normal distribution $N(1/\epsilon_i^k, 1/\tau_i^k)$. Given the samples $\mu_1^k, \mu_2^k, \ldots, \mu_\xi^k$ where ξ denotes the number of workers in $\mathcal{W}_{-k}$, the confidence interval of mean λ and variance σ^2 shown as $[\underline{\lambda}, \overline{\lambda}]$ and $[\underline{\sigma}^2, \overline{\sigma}^2]$ can be calculated by Equation 19 and 21. Afterwards, we can initialize the μ_i^k and $(\sigma_i^k)^2$ for new workers according to the grade g_i excavated from additional information. The parameters ϵ_i^k and τ_i^k can be derived as

$$\epsilon_i^k = 1/\mu_i^k,\ \tau_i^k = 1/(\sigma_i^k)^2 \quad (33)$$

Then all workers can be involved into EM algorithm for MAP estimation.

C. Reliability Change Detection

VI. Performance Evaluation

VII. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank...

References

[1] B. I. Aydin, Y. S. Yilmaz, Y. Li, Q. Li, J. Gao, and M. Demirbas, "Crowdsourcing for multiple-choice question answering." in *AAAI*, 2014, pp. 2946–2953.

[2] G. Demartini, D. E. Difallah, and P. 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*. ACM, 2012, pp. 469–478.

[3] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, "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*. ACM, 2014, pp. 1187–1198.

[4] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han, "Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 745–754.

[5] J. Fan, G. Li, B. C. Ooi, K.-l. Tan, and J. Feng, "icrowd: An adaptive crowdsourcing framework," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, 2015, pp. 1015–1030.

[6] D. Zhou, S. Basu, Y. Mao, and J. C. Platt, "Learning from the wisdom of crowds by minimax entropy," in *Advances in neural information processing systems*, 2012, pp. 2195–2203.

[7] Z. Zhao, F. Wei, M. Zhou, W. Chen, and W. Ng, "Crowd-selection query processing in crowdsourcing databases: A task-driven approach." in *EDBT*, 2015, pp. 397–408.

[8] H.-C. Kim and Z. Ghahramani, "Bayesian classifier combination," in *Artificial Intelligence and Statistics*, 2012, pp. 619–627.

[9] Q. Liu, J. Peng, and A. T. Ihler, "Variational inference for crowdsourcing," in *Advances in neural information processing systems*, 2012, pp. 692–700.

[10] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi, "Community-based bayesian aggregation models for crowdsourcing," in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 155–164.

[11] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1297–1322, 2010.

[12] P. Welinder, S. Branson, P. Perona, and S. J. Belongie, "The multidimensional wisdom of crowds," in *Advances in neural information processing systems*, 2010, pp. 2424–2432.
[13] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han, "A confidence-aware approach for truth discovery on long-tail data," *Proceedings of the VLDB Endowment*, vol. 8, no. 4, pp. 425–436, 2014.
[14] M. Joglekar, H. Garcia-Molina, and A. Parameswaran, "Evaluating the crowd with confidence," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 686–694.
[15] H. Xiao, J. Gao, Q. Li, F. Ma, L. Su, Y. Feng, and A. Zhang, "Towards confidence interval estimation in truth discovery," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 3, pp. 575–588, 2019.

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.