A Crowdsourcing Framework for Collecting Tabular Data

Caihua Shan, Nikos Mamoulis, Guoliang Li, Reynold Cheng, Zhipeng Huang, and Yudian Zheng

Abstract—In crowdsourcing, human workers are employed to tackle problems that are traditionally difficult for computers (e.g., data cleaning, missing value filling, and sentiment analysis). In this paper, we study the effective use of crowdsourcing in filling missing values in a given relation (e.g., a table containing different attributes of celebrity stars, such as nationality and age). A task given to a worker typically consists of questions about the missing attribute values (e.g., What is the age of Jet Li?). Although this problem has been studied before, existing work often treats related attributes independently, leading to suboptimal performance. In this paper, we present T-Crowd, which is a crowdsourcing system that considers attribute relationships. Particularly, T-Crowd integrates each worker's answers on different attributes to effectively learn his/her trustworthiness and the true data values. The attribute relationship information is used to guide task allocation to workers. Our solution seamlessly supports categorical and continuous attributes. Our extensive experiments on real and synthetic datasets show that T-Crowd outperforms state-of-the-art methods, improving the quality of truth inference and reducing the monetary cost of crowdsourcing.

Index Terms—Crowdsourcing, tabular data, truth inference, task assignment

1 Introduction

Crowdsourcing is an effective way to address computer-hard problems [8], [23], [36], [37], [43] by utilizing numerous ordinary humans (called *workers* or *the crowd*). The general workflow of crowdsourcing is as follows: at first a *requester* proposes a problem, then the problem is transformed into many tasks (i.e., questions), and finally the workers complete the tasks assigned to them and they are given a monetary reward. Crowdsourcing involves two interrelated processes: *truth inference* and *task assignment*. Truth inference refers to addressing noise and errors for inferring the correct value (or truth) for each task from redundant answers [11], [39]. Task assignment refers to selecting appropriate tasks to assign to each incoming worker. Truth inference can be used as a module in task assignment, to estimate the confidence of estimated true values [5], [23].

In this paper, we focus on crowdsourcing *tabular data*, i.e., a collection of related items which are structured in a tabular form and comply to a schema. Each column represents a particular attribute or variable. Each row corresponds to an entity and includes a value for each of the variables. Table 1 illustrates an example about data collection of celebrities; given the name of a celebrity, the goal is to collect the nationality, age, and notability (range from 1 to 5) of the person from the crowd. The bold values shown in Table 1 are the unknown (ground) truth data to be collected from the workers. Each cell of this table can be considered as a task, i.e., a worker may be asked to provide a value for the nationality of a celebrity given his/her name. Our target is to complete an empty or partial-filled table by filling in the cells effectively. Crowdsourcing tabular data finds direct application in database cleaning and integration [15], [28], [29].

Most crowdsourcing systems assume that the set of tasks are homogeneous and independent. However, tasks in tabular data can be *heterogeneous* and *dependent* to each other, which makes effective crowdsourcing on them challenging.

First, the datatypes and domains of different attributes may vary. For example, in Table 1, the task "the nationality of Jet Li?" has a different datatype compared to the task "the age of DiCaprio?" (i.e., categorical vs. continuous). Even attributes of the same datatype may have different domains (e.g., Age vs. Notability). As a result, approaches for integrating the answers of a worker in different homogeneous tasks are not directly applicable. These include the popular EM algorithm [9] for categorical data and data integration models applied for continuous attributes (GTM [40] and CATD [20]), to be discussed in Section 2. As we will show, applying a different approach for each column does not transfer the knowledge from one datatype to the other, i.e., the estimation of worker quality can be inaccurate due to data sparsity.

Second, in tabular data, there are potential dependencies between rows and columns. The difficulty of a task might depend on the corresponding entity and attribute. As a result, the quality of a worker on a particular task may depend on

- C. Shan, R. Cheng, and Z. Huang are with the Department of Computer Science, University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: {chshan, ckcheng, zphuang}@cs.hku.hk.
- N. Mamoulis is with the Department of Computing Science, University of Ioannina, Ioannina, Epirus PC 45110, Greece. E-mail: nikos@cs.uoi.gr.
- G. Li is with the Department of Computing Science, TNList, Tsinghua University, Beijing 100091, China. E-mail: liguoliang@tsinghua.edu.cn.
- Y. Zheng is with Twitter Inc., San Francisco, CA 94103. E-mail: yudianz@twitter.com.

Manuscript received 11 Mar. 2018; revised 10 Nov. 2018; accepted 22 Apr. 2019. Date of publication 3 May 2019; date of current version 6 Oct. 2020. (Corresponding author: Caihua Shan.) Recommended for acceptance by K. Selcuk Candan. Digital Object Identifier no. 10.1109/TKDE.2019.2914903

TABLE 1
Ground Truth about Celebrities

	Name	Nationality	Age	Notability
1	Leonardo DiCaprio	**United States**	**42**	**5**
2	Jet Li	**China**	**54**	**4**
3	James Purefoy	**Great Britain**	**53**	**3**

TABLE 2
Answers to Questions about Celebrities

Worker	Row Id	Nationality	Age	Notability
u_1	1	**United States**	**40**	**5**
	2	**China**	**45**	**3**
u_2	1	**United States**	**42**	**5**
	3	**Great Britain**	**53**	**3**
u_3	2	**China**	**50**	**4**
	3	**United States**	**35**	**2**

her quality on other tasks in the same row or column. Take Table 2 as an example, where bold values are the answers of three workers on tasks from Table 1. Note that worker u_3 inputs a wrong nationality of James Purefoy, meaning that she might mistake this celebrity for someone else. Therefore, her answers for the age and notability of the person have high chance to be unreliable, despite the high quality of her input for the second row. This means that when we assign a new task to the coming worker, we should not only consider the worker's inherent quality, but also whether the worker is familiar with the entity (we call it the worker's structure-aware quality). Traditional task assignment methods [7], [18] focus on capturing the former but ignore the latter.

In this paper, we present *T-Crowd*, the first crowdsourcing system that considers heterogeneous and dependent of tabular data in both truth inference and task assignment. T-Crowd processes the submitted answers by each worker to infer a *unified quality* for him or her. T-Crowd seamlessly integrates the worker's answers to questions of different datatypes and domains, addressing consistency and data sparsity issues that would arise from the alternative approach of using different models for different columns. For example, the overall quality of worker u_2 can be regarded better than that of worker u_1 considering their answers to both categorical and continuous values in Table 2. Unified worker quality greatly improves truth inference and task assignment, reducing the total number of tasks to be assigned to workers until all true values can be estimated with high confidence.

T-Crowd captures the importance of tasks (i.e., how confident we are about their value estimates) in the different columns and rows, based on the collected data so far. We also define an *inherent information gain* which is a uniform measure for ranking tasks with respect to a given worker. Then we choose to assign to the worker the tasks with the highest anticipated benefit. In contrast, previous work [15], [29] on crowdsourcing tabular data performs task assignment based on only how many more answers are needed for each task, disregarding worker quality. To further improve performance, we utilize the potential correlations between tasks. We define a *structure-aware information gain* which extends the inherent information gain to also consider as a parameter the previous answers given by the worker on tasks that appear in the same row, when selecting new tasks to assign to him or her.

A preliminary version of this work, which focused on truth inference in crowdsourcing tabular data, appears in [33]. In this paper, in addition to truth inference, we study the task assignment problem. In addition, we evaluate the performance of our proposed task assignment approach on three real datasets and compare it with four competitors. Besides, we add several new and recent competitors for our proposed truth inference algorithms, such as CATD, Zencrowd, TC-onlyCate and TC-onlyCont. The latter two are the constrained versions of T-Crowd that apply only on the categorical or continuous attributes. Finally, we expand our case studies, add experiments with synthetic datasets, and include a comparison to CrowdFill [29].

To summarize, our main contributions are as follows:

- To the best of our knowledge, we are the first to study crowdsourcing tabular data with both heterogeneity and dependency.
- We unify worker quality for all tasks in crowdsourced tabular data, improving the accuracy of truth inference and the performance of task assignment, compared to models that treat each attribute independently.
- Given an incoming worker, we find suitable tasks for him/her based on *inherent information gain*, including the benefit of obtaining additional answers in tasks and the worker's inherent quality. We also extend it to *structure-aware information gain*, which considers the correlation of answer quality between tasks in the same row.
- We evaluate T-Crowd on real datasets; the results demonstrate its superiority over existing alternatives. Compared to previous work, T-Crowd has better truth inference accuracy and converges to the true values of the tasks using only about half of the answers by the workers.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 defines the problem and gives an overview of our system. In Section 4, we present our methodology for truth inference. Our task assignment policy is presented in Section 5. Section 6 includes our experimental evaluation. Finally, we conclude in Section 7.

2 Related Work

Related work falls into two categories: *truth inference* methods used to infer the truth and *task assignment* strategies for an incoming worker.

Truth Inference. The most basic truth inference methods are majority voting for multiple-choice tasks (i.e., categorical data) and taking the median for numerical tasks (i.e., continuous data). These approaches regard all workers as equal, disregarding any differences in their trustworthiness. Methods such as D&S [9], [17] use a confusion matrix to model a worker's quality, and use an Expectation-Maximization (EM) algorithm to infer the truth. More advanced approaches such as TruthFinder [39], Accusim [12], and GLAD [38] improve accuracy using different worker answering models or by considering more parameters, such as a task's difficulty. These

TABLE 3
Table of Notations

Notation	Description
c_{ij}	cell (task) in the ith row and jth column
a_{ij}^u	answer given by worker u for cell c_{ij}
$\mathcal{A}$	the set of all answers, i.e., $\mathcal{A} = \{a_{ij}^u\}$
T_{ij}	distribution of estimated truth for cell c_{ij}
T_{ij}^* $(\widehat{T}_{ij})$	ground truth (estimated truth) for cell c_{ij}
e_{ij}^u	error of a_{ij}^u with respect to $\widehat{T}_{ij}$
q_u	quality of worker u
α_i (β_j)	difficulty of row i (column j)

methods focus on answering tasks on categorical data. Other methods, such as GTM [40], are designed for continuous crowdsourced data. CRH [21], [22] and CATD [20] are two existing truth inference approaches for both categorical and continuous data. CRH [21] incorporates different distance functions between the answers and the estimated truth to recognize the characteristics of various data types. Specifically, CRH proposes an objective function and minimizes it by updating the estimated true values and source reliability (i.e., worker quality) in turns. CATD [20] considers both source reliability and the confidence interval of the estimation. Additional information of tasks or workers has also been considered in truth inference, such as the latent topics of the tasks [24] and the learn bias of workers [44].

The aforementioned works do not consider tabular data. In Section 4, we present an iterative Expectation-Maximization (EM) truth inference algorithm, which improves the accuracy of truth inference from the answers compared to previous work. The novelty of our work is that we use a probabilistic model for the answers of workers wrt different data types and that we unify workers' quality on categorical data and continuous data explicitly, while methods like CRH design different distance functions for the different data types.

Task Assignment. Online task assignment selects which tasks to assign to each incoming worker, in order to achieve the maximum possible quality for the collected data. In earlier crowdsourcing systems, such as CDAS [23], the candidate tasks are randomly assigned to workers. AskIt [5] is yet another crowdsourcing platform, which assigns the tasks that have the highest uncertainty, again disregarding the quality (or expertise) of the incoming worker for these tasks. CrowdDB [15], Deco [28], and Qurk [25] are extensions of relational database systems that incorporate the crowd's knowledge into query processing. They use answers from the crowd to make up the missing values of query operators. They are similar to our approach in that they collect tabular data; however, they do not focus on the assignment strategy and simply assign random tasks to workers. CrowdFill [29] is a recent system for tabular data, which uses a non-conventional workflow that is not supported by common crowdsourcing platforms such as AMT. In CrowdFill, workers are asked to select and perform tasks from a subset of the table given to them and they can also vote for the answers to these tasks by other workers. Besides, CrowdFill does not estimate worker quality, and does not use properties of tabular data (e.g., attribute dependencies) to assign tasks to workers. Some methods [14], [26], [41] consider the case where the tasks are relevant to different domains and workers are given the tasks that match their domain expertise. In recent work, such as OptKG [7] and CrowdDQS [18], task assignment is modeled by a Markov Decision process or solved by using maximum potential gain, but the application of these models is limited to only multiple-choice tasks (categorical tasks). Other forms of online task assignment, which need explicit workers' collaboration, have been studied in [31], [32]. Different from the above works, our method focuses on crowdsourced tabular data, which is structured and heterogeneous, presenting challenges and opportunities as discussed in the Section 1.

3 Problem Definition

In this section, we formulate the problem and give an overview of T-Crowd. Our goal is to perform crowdsourcing on a two-dimensional table C, defined as follows.

Definition 1 (Tabular Data Model). *We target the crowdsourcing of a two-dimensional table $C = \{c_{ij}\}$, where $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, M\}$. C has an* entity *attribute which is the key attribute of the table. Each column is a categorical or a continuous attribute. Each cell c_{ij} represents the value of the ith entity in the jth attribute, whose true value (i.e.,* truth, *or* ground truth*) is denoted as T_{ij}^*.*

Table 1 shows an example of tabular data about celebrities that we want to crowdsource. *Age* and *Notability* are continuous attributes, while *Nationality* is categorical. The entity attribute is *Name*. To obtain the truth for the remaining attributes, we ask the crowd to provide answers.

Definition 2 (Task, Worker, Answer). *A task is related to a cell c_{ij} and the workers are asked to answer the task, by providing values for the cell. Let U be a set of workers. A worker $u \in U$ will submit an answer a_{ij}^u, if cell c_{ij} is assigned to u.*

For example, to get the age of the second entity, a task provides the name of the second entity and asks workers to input the age. Since workers may have different levels of quality (e.g., some workers are experts, while some are spammers), each task c_{ij} is often assigned to multiple workers and all acquired answers for c_{ij} are aggregated to infer the true value of c_{ij}. Next, we define the two problems that we aim to address in this paper.

Definition 3 (Truth Inference). *Given the set of answers $\{a_{ij}^u\}$, by workers $u \in U$ to cells c_{ij}, $i \in \{1, \ldots, N\}$, $j \in \{1, \ldots, M\}$, the problem of truth inference is to compute an accurate estimate $\widehat{T}_{ij}$ for each cell c_{ij}'s true value T_{ij}^*.*

Definition 4 (Task Assignment). *When a worker u requests for a task for C, decide the task to be assigned to u.*

Note that existing crowdsourcing platforms, such as the Amazon Mechanical Turk (AMT) [1], support the functionality of dynamically assigning tasks to an incoming worker (e.g., the 'external-HIT' feature in AMT [2]). Table 4 summarizes the notations used in this paper.

System Architecture. Fig. 1 gives an overview of T-Crowd, our proposed system for crowdsourcing tabular data. A *requester* (e.g., a lifestyle journal) first defines the structure (i.e., schema) of the collected data, such as the datatypes of the columns, and the key attribute. Then the requester

TABLE 4
Table of Notations in Task Assignment

Notation	Description
e^u_{ij}	error of answer a^u_{ij} with respect to $\widehat{T}_{ij}$
E_j	distribution of error in column j
$P(E_j\|E_k)$	correlation of error between column j and k
w_{jk}	correlation coefficient between column j and k

publishes tasks to a crowdsourcing platform, e.g., AMT [1]. For an incoming worker u, our Task Assignment module determines one or more cells and assigns the corresponding task(s) to u. This is based on the anticipated *information gain* of the different cells by u's answers. Intuitively, the information gain is an estimate of how much more accurate the cells' values become upon collection of u's inputs. When the worker submits an answer a^u_{ij} for a cell c_{ij} to the system, the Truth Inference module infers the estimated truth $\widehat{T}_{ij}$. To facilitate task assignment and truth inference, we also estimate the quality of worker q_u and the difficulty of cells α_i and β_j. Task(s) are assigned to workers and answers are collected until $\widehat{T}_{ij}$ converges (or a budget is exhausted).

4 Truth Inference

In this section, we explain how T-Crowd performs truth inference on tabular data. The quality of truth inference for a data cell c_{ij} depends on the quality of workers who answer c_{ij}, and the difficulty of c_{ij}. We first discuss how to model worker quality q_u and cell difficulty $\alpha_i(\beta_j)$ if we already know the truth $\widehat{T}_{ij}$ (Sections 4.1). Then, we show how to infer the true values of cells $\widehat{T}_{ij}$ and these two factors simultaneously by maximizing the likelihood of workers answers a^u_{ij} (Section 4.2).

4.1 Worker Model

4.1.1 Quality of a Worker

The challenge in modeling worker quality is that attributes may have different datatypes; the answer set of a categorical task is finite and nominal, while that of a continuous task is an integer or a real number. Hence, it is not straightforward to model the quality of a worker using a single parameter. To address this problem, we propose a unified model for both categorical and continuous attributes.

We model the truth of a categorical attribute l^* as an element in a finite unordered set of possible answers $L = \{l_1, l_2, \ldots, l_{|L|}\}$. An answer from a worker is either correct or wrong depending on whether it is the same as the ground truth. On the other hand, for a continuous attribute, the quality of the answer depends on how close it is to the ground truth. For example, if the age of Jet Li is 54, and a worker answers 53, which is close to the truth, the answer is considered to be a good one.

As discussed, our goal is to use a single parameter q_u to represent the quality of a worker u. For the ease of presentation, we first illustrate how the worker's quality for continuous datatypes can be modeled, and then show how the model can be extended for categorical datatypes.

• For *continuous* datatypes, we model the distribution of the answer given by worker u as a normal distribution: $a^u_{ij} \sim \mathcal{N}(\widehat{T}_{ij}, \phi_u)$:

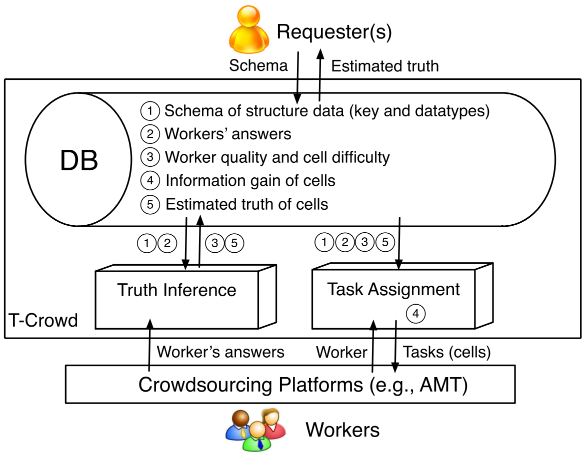


Fig. 1. System architecture.

$$P(a^u_{ij} = x) = \frac{1}{\sqrt{2\pi\phi_u}} \exp\left(-\frac{(x - \widehat{T}_{ij})^2}{2\phi_u}\right), \tag{1}$$

where $\widehat{T}_{ij}$ is the expected value of c_{ij} and ϕ_u is the variance of u. Intuitively, the higher the quality of a worker is, the smaller the variance will be, as his/her answer should have smaller difference from the truth. Inspired by this, we model $q_u \in [0, 1]$ as the probability that the answer from worker u falls into a small range (ϵ) around the truth $\widehat{T}_{ij}$:

$$q_u = P(\ a^u_{ij} \in [\widehat{T}_{ij} - \epsilon, \widehat{T}_{ij} + \epsilon]\ ) = \text{erf}(\epsilon/\sqrt{2\phi_u}). \tag{2}$$

Intuitively, q_u is the area under the normal distribution curve, where ϵ is a general parameter that controls the shape of the area and "erf" is the Gauss error function [4].

• For *categorical* attributes, $q_u \in [0, 1]$ indicates the probability that the worker u would correctly answer a task, i.e.,

$$P(a^u_{ij} = z) = (q_u)^{\mathbb{1}_{\{\widehat{T}_{ij}=z\}}} \cdot \left(\frac{1-q_u}{|L|-1}\right)^{\mathbb{1}_{\{\widehat{T}_{ij}\neq z\}}}, \tag{3}$$

where $\mathbb{1}_{\{\cdot\}}$ is an *indicator function* which returns 1 if the argument is true; 0, otherwise. For example, $\mathbb{1}_{\{5=5\}} = 1$ and $\mathbb{1}_{\{5=3\}} = 0$. Intuitively, worker u has probability q_u to give the correct answer and we evenly distribute the probability $(1 - q_u)$ to the remaining (false) answers. Note that q_u can be expressed as in Eq. (2), which means that we can use the same quality measure for categorical and continuous attributes.

4.1.2 Difficulty of a Cell

The answers from workers do not only depend on their expertise, but they are also influenced by the difficulty of tasks. Hence, in our model, the quality of answer a^u_{ij} depends on the quality of worker u, the difficulty β_j of attribute (i.e., column) j, and the difficulty α_i of entity (i.e., row) i.

To incorporate the difficulty of each cell c_{ij} into the worker's quality, we define the variance of his/her answer to a cell c_{ij} as $\phi^u_{ij} = \alpha_i\beta_j\phi_u$. Hence, the variance is positively correlated to the difficulties α_i and β_j, and the inherent variance (ϕ_u) of answers by worker u. Then, following Eq. (2), we represent the quality of worker u answering cell c_{ij} as $q^u_{ij} = \text{erf}(\epsilon/\sqrt{2\alpha_i\beta_j\phi_u})$. To model the worker's answers on categorical and continuous data, Eqs. (1) and (3) can be changed accordingly, i.e., by replacing ϕ_u with ϕ^u_{ij} and q_u with q^u_{ij}.

Note that $\widehat{T}_{ij}$, α_i, β_j and ϕ_u are unknown and we discuss how to compute them later. The worker quality q_u (q_{ij}^u) can be calculated directly if we know α_i, β_j, and ϕ_u.

4.2 Inference Process

The objective function of the truth inference problem is to maximize the likelihood of workers' answers, i.e.,

$$\arg\max_{\alpha,\beta,\phi} P(\mathcal{A}|\alpha,\beta,\phi) = \arg\max_{\alpha,\beta,\phi} \sum\nolimits_{\mathcal{T}} P(\mathcal{A},\mathcal{T}|\alpha,\beta,\phi),$$

where $\mathcal{A}$ is the current set of answers by all workers on all cells and $\mathcal{T}$ is a set of all hidden true values, i.e., $\mathcal{T}=\{T_{ij}\}$. T_{ij} denotes the estimated distribution of truth in cell c_{ij}. To optimize this non-convex function, we use the Expectation-Maximization (EM) algorithm [11], which takes an iterative approach. In each iteration of EM, the E-step computes the hidden variables in $\mathcal{T}$, and the M-step computes the parameters α_i, β_j and ϕ_u (q_u). Next, we provide details about the E-step and the M-step.

Expectation Step (E-step). In the E-step, we compute the posterior probabilities of hidden variable $T_{ij}\in\mathcal{T}$ given the values of α, β and ϕ and the observed variable $A_{ij}=\{a_{ij}^u\}$, $u\in U_{ij}$, i.e., the current answer set of cell c_{ij}.

$$P(\ T_{ij}=z|A_{ij},\alpha_i,\beta_j,\phi\ ) \propto \prod\nolimits_{u\in U_{ij}} P(a_{ij}^u|T_{ij}=z,\alpha_i,\beta_j,\phi_u)\cdot \mathrm{Prior}(T_{ij}=z). \tag{4}$$

Based on our defined worker model of $P(T_{ij}=z|A_{ij},\alpha_i,\beta_j,\phi)$ for different datatypes, the distribution is defined as follows.

(1) For cells c_{ij} of continuous type, we regard that $\mathrm{Prior}(T_{ij}=z)$ follows a normal distribution $\mathcal{N}(\mu_j^0, \phi_j^0)$, and $T_{ij}\sim\mathcal{N}(T_{ij}^{\mu},\ T_{ij}^{\phi})$, where T_{ij}^{μ} and T_{ij}^{ϕ} satisfy that

$$T_{ij}^{\mu} = \Big(\sum\nolimits_{u\in U_{ij}} \frac{a_{ij}^u}{\alpha_i\beta_j\phi_u} + \frac{\mu_j^0}{\phi_j^0}\Big)\ T_{ij}^{\phi},$$

$$T_{ij}^{\phi} = \Big(\sum\nolimits_{u\in U_{ij}} \frac{1}{\alpha_i\beta_j\phi_u} + \frac{1}{\phi_j^0}\Big)^{-1}.$$

(2) For cells c_{ij} of categorical type, we have

$$P(T_{ij}=z) = \frac{\prod_{u\in U_{ij}}\big[(q_{ij}^u)^{\mathbb{1}_{\{a_{ij}^u=z\}}}\big(\frac{1-q_{ij}^u}{|L_j|-1}\big)^{\mathbb{1}_{\{a_{ij}^u\neq z\}}}\big]}{\sum_{z\in L_j}\prod_{u\in U_{ij}}\big[(q_{ij}^u)^{\mathbb{1}_{\{a_{ij}^u=z\}}}\big(\frac{1-q_{ij}^u}{|L_j|-1}\big)^{\mathbb{1}_{\{a_{ij}^u\neq z\}}}\big]},$$

where q_{ij}^u is defined as $\mathrm{erf}(\epsilon/\sqrt{2\alpha_i\beta_j\phi_u})$ and L_j is the label set of column j. $\mathrm{Prior}(T_{ij}=z)$ is uniform so it disappears.

Intuitively, the answer given by high quality worker will be trusted more, i.e., given higher weight. To be specific, we estimate the truth distribution T_{ij} by combining the set A_{ij} of workers' answers for c_{ij}. (1) T_{ij}^{μ} can be regarded as a weighted average of answer a_{ij}^u based on the quality $\alpha_i\beta_j\phi_u$. T_{ij}^{ϕ} is a normalized term. (2) Similarly, $P(T_{ij}=z)$ is a normalized product of the qualities q_{ij}^u of the workers whose answer a_{ij}^u is z.

Maximization Step (M-step). In the M-step, we find the values of parameters α, β and ϕ that maximize the expectation of the joint log-likelihood of the observed variable $\mathcal{A}$, as shown below:

$$\begin{aligned} &Q(\alpha,\beta,\phi) = \mathrm{E}_{\mathcal{T}}[\ln P(\mathcal{A},\mathcal{T}|\alpha,\beta,\phi)] \\ &= \sum_j\sum_i\ \mathrm{E}_{T_{ij}}\big[\ln \mathrm{Prior}(T_{ij}) + \sum_{u\in U_{ij}} \ln P(a_{ij}^u|T_{ij},\alpha_i,\beta_j,\phi_u)\big]. \end{aligned} \tag{5}$$

Formula $\mathrm{E}_{T_{ij}}[\sum_{u\in U_{ij}} \ln P(a_{ij}^u|T_{ij},\alpha_i,\beta_j,\phi_u)]$ is calculated for the different datatypes, as follows.

(1) For cells c_{ij} of continuous type:

$$\sum\nolimits_{u\in U_{ij}}\Big[-\frac{1}{2}\ln(2\pi\alpha_i\beta_j\phi_u) - \frac{(a_{ij}^u - T_{ij}^{\mu})^2 + T_{ij}^{\phi}}{2\alpha_i\beta_j\phi_u}\Big].$$

(2) For cells c_{ij} of categorical type:

$$\sum\nolimits_{z\in L_j} P(T_{ij}=z)\cdot \sum\nolimits_{u\in U_{ij}} \Big(\mathbb{1}_{\{a_{ij}^u=z\}} \ln \mathrm{erf}\big(\frac{\epsilon}{\sqrt{2\alpha_i\beta_j\phi_u}}\big) + \mathbb{1}_{\{a_{ij}^u\neq z\}} \ln \frac{1-\mathrm{erf}\big(\frac{\epsilon}{\sqrt{2\alpha_i\beta_j\phi_u}}\big)}{|L_j|-1}\Big).$$

We apply gradient descent to find the values of α, β and ϕ that locally maximize $Q(\alpha,\beta,\phi)$.

Intuitively, a worker will be of high quality if his/her answers are close to the estimated truth. Thus, we compute a value ϕ_u that maximizes the expectation of the log-likelihood of worker u's answers a_{**}^u. Similarly, we also find an α_i (resp. β_j) that maximizes the expectation of the log-likelihood of answers a_{i*}^* in row i (resp. a_{*j}^* in column j).

Algorithm 1. *Truth Inference Method*

Input: workers' answers $a_{ij}^u\in\mathcal{A}$, prior distribution of truth $\mathrm{Prior}(T_{ij})$
Output: truth distribution $T_{ij}\in\mathcal{T}$, worker's quality ϕ_u, difficulty of row α_i and column β_j

```
Initialize T_ij using Prior(T_ij)
while true do
  // Step 1: Estimate Worker Quality and Cell Difficulty
  Compute α_i, β_j and φ_u maximizing Eq. (5);
  // Step 2: Infer the Truth
  for 1 ≤ i ≤ N do
    for 1 ≤ j ≤ M do
      Obtain T_ij by Eq. (4);
  // Check for Convergence
  if Converged then
      break;
return T_ij, α_i, β_j and φ_u;
```

Algorithm. By combining the two steps above, we can iteratively update the parameters until convergence. Each T_{ij} is initialized by following the distribution in $\mathrm{Prior}(T_{ij})$. At each iteration, the M-step applies gradient descent to find α_i, β_j and ϕ_u by maximizing Eq. (5) and the E-step applies Eq. (4). We identify convergence if the differences between the parameter values in subsequent iterations are below a threshold (e.g., 10^{-5}).

Finally we estimate the truth $\widehat{T}_{ij}$ of each cell c_{ij} as:

$$\widehat{T}_{ij} = \begin{cases} T_{ij}^{\mu} & ,\ c_{ij} \text{ is continuous,} \\ \arg\max_{z\in L_j}\ P(T_{ij}=z) & ,\ c_{ij} \text{ is categorical.} \end{cases}$$

Time Complexity. The total cost of the E-step is $\mathcal{O}(l \cdot |\mathcal{A}|)$, where $\mathcal{A}$ is the set of all obtained answers and $l = \max_j(|L_j|)$. In the M-step, one gradient descent needs to compute the gradient of each parameter which takes $\mathcal{O}(l \cdot |\mathcal{A}|)$. If the gradient descent takes v iterations to converge, this step takes $\mathcal{O}(vl \cdot |\mathcal{A}|)$ time in total. Assuming that the algorithm needs w iterations to converge, the total time complexity is $\mathcal{O}(wvl \cdot |\mathcal{A}|)$. In practice, l is constant, and v and w are smaller than 20, thus the time complexity is linear to the number of answers.

5 Online Task Assignment

In this section, we discuss how we select tasks for a worker u. Section 5.1 defines an *inherent information gain* function to measure the utility of assigning a task to the worker, which can handle both categorical and continuous data. The function considers the quality of the worker, the need to obtain more answers for the task, and the task's difficulty. Intuitively, we prefer to assign tasks whose gain of information will be improved the most if the incoming worker answers them. In Section 5.2, we extend this to a *structure-aware information gain* function, which also considers the correlations in the qualities of answers given by the same worker to different cells of the same row.

5.1 Inherent Information Gain

We need a uniform measure for the *utility* (or benefit) of assigning a task (either categorical or continuous) to a worker u with quality q_u. For this purpose we define an *inherent information gain* function, following the steps below.

(1) For a categorical cell c_{ij}, the distribution of truth T_{ij} has been computed by $P(T_{ij} = z)$ in Equation 4, which is the probability that label z is correct. Thus, Shannon Entropy [3] can be used to define the uncertainty of task c_{ij}:

$$H_s(T_{ij}) = -\sum_{z \in L_j} P(T_{ij} = z) \ln P(T_{ij} = z).$$

(2) For a continuous cell c_{ij}, note that for a continuous distribution, the Differential Entropy [27] is defined as:

$$-\int_{\mathbb{X}} f(x) \ln f(x)\, dx,$$

where $f(x)$ is a probability distribution. Recall that we also define the distribution of truth $T_{ij} \sim \mathcal{N}(T^{\mu}_{ij}, T^{\phi}_{ij})$ of a continuous cell c_{ij} in Equation 4, so its Differential Entropy can be computed as:

$$H_d(T_{ij}) = \frac{1}{2} \ln \left(2\, \pi\, e\, T^{\phi}_{ij}\right).$$

Given the above, we define the *uniform entropy* for task c_{ij}:

$$H(T_{ij}) = \begin{cases} H_d(T_{ij}), & \text{if } c_{ij} \text{ is continuous,} \\ H_s(T_{ij}), & \text{if } c_{ij} \text{ is categorical.} \end{cases}$$

A straightforward approach for task assignment to a worker u is to select the task c_{ij} with the largest uniform entropy. However, this is problematic, as Differential Entropy and Shannon Entropy are not comparable; hence, task assignments may be biased toward one datatype. For example, as pointed out in [27], Differential Entropy can be negative while Shannon entropy is always non-negative. Alternatively, we use Delta Entropy to measure the information gain. Suppose $\mathcal{A}_C$ is the current set of answers we have collected, we can obtain the estimated truth distribution (denoted as $T_{ij,\mathcal{A}_C}$) for each task c_{ij} by the truth inference method presented in Section 4. Specifically, for an incoming worker u, we define the *inherent information gain* of assigning task c_{ij} to her as:

$$IG_q(c_{ij}) = H(T_{ij,\mathcal{A}_C}) - \mathrm{E}_{a^u_{ij}}[H(T_{ij,\mathcal{A}_C \cup \{a^u_{ij}\}})], \quad (6)$$

where $T_{ij,\mathcal{A}_C \cup \{a^u_{ij}\}}$ is the updated distribution of the estimated truth for task c_{ij} after receiving a new answer a^u_{ij} from u.

By using the inherent information gain measure defined in Eq. (6), we alleviate the problem that the domains of the two entropy types are different. If we discretize the range of a continuous random variable X using bins of width Δ, we can compute the Shannon entropy for this new discretized random variable X^{Δ}, and we have the following formula if X's pdf is Riemann integrable:

$$H_s(X^{\Delta}) + \ln \Delta \rightarrow H_d(X),\ \ as\ \Delta \rightarrow 0.$$

Hence, if Δ is small, $H_d(X_1) - H_d(X_2) \approx H_s(X^{\Delta}_1) - H_s(X^{\Delta}_2)$, which means that the subtraction of two differential entropies can be transformed into subtraction of two Shannon entropies. As a result, for cells of different types, $IG(c_{ij})$ is comparable. Algorithm 2 describes the task assignment algorithm in detail.

Computing the Distribution of $\mathrm{E}_{a^u_{ij}}[H(T_{ij,\mathcal{A}_C \cup \{a^u_{ij}\}})]$. The distribution of an answer a^u_{ij} follows the worker model in Eqs. (1) and (3) for continuous and categorical tasks, respectively. For a categorical task c_{ij}, the domain of a^u_{ij} is a finite label set, so we use all possible values a^u_{ij} to obtain $T_{ij,\mathcal{A}_C \cup \{a^u_{ij}\}}$ using the inference method described in Section 4. For a continuous task, since the the domain of a^u_{ij} is $\mathcal{R}$, we apply sampling to approximate the value of $T_{ij,\mathcal{A}_C \cup \{a^u_{ij}\}}$. However, it is expensive to run the inference method for each possible answer. To alleviate this problem, we limit the number of iterations per answer, by only updating the parameters related to the answer and keeping the other parameters unchanged. Specifically, for a new answer a^u_{ij}, we locally update the truth distribution T_{ij}, and the qualities of workers who have answered task c_{ij}.

Time Complexity.To compute the benefit for each task c_{ij} (Eq. (6)), we should first iterate through the possible answers given by the incoming worker and compute a new distribution of truth T_{ij}. The number of possible answers for a categorical task c_{ij} is $|L_j|$ and for a continuous task is the fixed sampling number s_{cont}. Because we approximate the inference method, it only takes $\mathcal{O}(l \cdot |P|)$ where P is the set of parameters we need to update. Let $s = \max(\max_j(|L_j|), s_{\text{cont}})$; the total cost of considering one task for a certain worker is $\mathcal{O}(sl \cdot |P|)$. Then, computing the information gains of all tasks takes $\mathcal{O}(NMsl \cdot |P|)$. Since P includes the truth distribution T_{ij} and the qualities of workers who have answered task c_{ij}, P mainly depends on the average answers per task. Thus, $\mathcal{O}(NMsl \cdot |P|) \approx \mathcal{O}(sl \cdot |\mathcal{A}|)$.

Parallel or distributed computation can be used to accelerate task assignment, as the consideration of the different tasks are independent.

Algorithm 2. *Online Task Assignment Method*

```
Input: Budget B
Output: truth distribution T_ij ∈ 𝒯
Initialize each task with several answers from workers
while Budget B is not exhausted do
    // Step 1: Analyze current situation
    Run truth inference to obtain T_ij, α_i, β_j and φ_u
    // Step 2: Find task c* with highest benefit for incoming
    worker u
    for 1 ≤ i ≤ N do
        for 1 ≤ j ≤ M do
            Compute information gain IG(c_ij) by Eq. (6)
            if IG(c_ij) > IG(c*) or c* is not defined then
                c* = c_ij
    // Step 3: Collect answers
    Publish task c* and collect worker u's answer
Run truth inference to obtain the final T_ij
return T_ij
```

5.2 Structure-Aware Information Gain

The task assignment approach based on inherent information gain, described in Section 5.1, does not utilize the structural information of table C. We now propose a structure-aware task assignment method. The basic idea is to estimate correlation, i.e., the conditional distribution of the error on a task c_{ij}, given the errors on other tasks c_{i*} in the same row. For this, we consider the answer history of all workers and then use the conditional distribution to obtain a better estimation of the target worker u's error on task c_{ij}.

We have already shown how to estimate the truth $\widehat{T}_{ij}$ for each cell c_{ij} in Section 4. Based on it, we can transform answer a^u_{ij} into error e^u_{ij}. For a continuous attribute, $e^u_{ij} = a^u_{ij} - \widehat{T}_{ij}$, while for a categorical attribute, $e^u_{ij} = \begin{cases} 0 & , a^u_{ij} = \widehat{T}_{ij}, \\ 1 & , a^u_{ij} \neq \widehat{T}_{ij}. \end{cases}$ It is easy to regain answer a^u_{ij} from error e^u_{ij} by reversing the according equation.

We regard $P(E_j|E_k)$ as the correlation of error between column j and k. We estimate $P(E_j|E_k)$ with a maximum likelihood method considering all the answers a^*_{*j} and a^*_{*k} we have collected, which is discussed later. If worker u has answered task c_{ik} before, his/her error for task c_{ij} is recomputed as $P\big(E_j|E_k = e^u_{ik}\big)$. When worker u has answered multiple tasks $L^u_i = \{k|\text{ worker } u \text{ answered task } c_{ik} \text{ on row } i\}$, we need to consider all the observed errors. However, it is not practical to estimate the conditional distribution, given errors from multiple attributes, due to data sparsity. Hence, we consider a linear combination of the correlations, as follows:

$$\frac{\sum_{k\in L^u_i} w_{jk} \cdot P\big(E_j|E_k = e^u_{ik}\big)}{\sum_{k\in L^u_i} w_{jk}}, \tag{7}$$

where w_{jk} is the correlation coefficient between attribute j and k:

$$w_{jk} = \frac{(M_j - \overline{M}_j)(M_k - \overline{M}_k)}{\sqrt{(M_j - \overline{M}_j)^2}\sqrt{(M_k - \overline{M}_k)^2}}, \tag{8}$$

where M_j and M_k are the error vector on attribute j and k combined by the pair data $\{(e^u_{ij}, e^u_{ik})|$ error of answers a^u_{ij} and a^u_{ik} when they are both existed$\}$. $\overline{M}_j$ and $\overline{M}_k$ are also vectors, where each element is the mean of vector M_j and M_k, respectively.

After obtaining the conditional distribution of error e^u_{ij}, we transform error e^u_{ij} into answer a^u_{ij} by reverse operations described above. Then, we calculate $\mathrm{E}_{a^u_{ij}}[H(T_{ij,\mathcal{A}_C\cup\{a^u_{ij}\}})]$ based on new answer distribution a^u_{ij} while $H(T_{ij,\mathcal{A}_C})$ is not changed. Accordingly, the structure-aware information gain $IG_c(c_{ij})$ is calculated using Eq. (6).

Computing the Correlation $P(E_j|E_k)$. Correlation is defined as the conditional probability between column j and k and it is derived from the known errors e^*_{*j} and e^*_{*k}.

(1) *Marginal distribution $P(E_j)$.* A categorical column is regarded as a Bernoulli distribution while a continuous column is regarded as a normal distribution.

(2) *Conditional distribution $P(E_j|E_k)$.* Since we have categorical and continuous columns, we have four cases in total. For each case, we use the maximal likelihood method to estimate the parameters in the assumed distribution. We elaborate on these cases below:

i. *both j and k are categorical:* $P(E_j = 1|E_k = 0)$, $P(E_j = 0|E_k = 0)$, $P(E_j = 1|E_k = 1)$ and $P(E_j = 0|E_k = 1)$ are counted based on the occurrences.

ii. *both j and k are continuous:* Because errors in continuous columns follow normal distributions, joint distribution $P(E_j, E_k)$ is a bivariate normal distribution. If the mean vector is $\begin{pmatrix} \mu_j \\ \mu_k \end{pmatrix}$ and the covariance matrix is $\begin{pmatrix} \sigma_j^2 & \rho\sigma_j\sigma_k \\ \rho\sigma_j\sigma_k & \sigma_k^2 \end{pmatrix}$, the conditional distribution $P(E_j|E_k)$ is also a normal distribution

$$P(E_j|E_k = e^u_{ik}) \sim \mathcal{N}(\mu_j + \frac{\sigma_j}{\sigma_k}\rho(e^u_{ik} - \mu_k), (1-\rho^2)\sigma_j^2).$$

iii. *column k is categorical and column j is continuous:* We assume that the conditional distributions $P(E_j|E_k = 0)$ and $P(E_j|E_k = 1)$ obey normal distributions. We obtain the mean and variance when $E_k = 0$ or $E_k = 1$ separately.

iv. *column j is categorical and column k is continuous:* Based on the same assumptions as in case (iii), we can estimate $P(E_k|E_j = 0)$ and $P(E_k|E_j = 1)$. Because we also know $P(E_k)$ and $P(E_j)$, the conditional distributions can be calculated using Bayes' theorem

$$P(E_j|E_k = e^u_{ik}) = \frac{P(E_k = e^u_{ik}|E_j)P(E_j)}{P(E_k = e^u_{ik})}$$

Time Complexity. To compute the correlation $P(E_j|E_k)$, we should iterate through each column and calculate the corresponding conditional distribution. Because there are

TABLE 5
Statistics of Real-World Datasets

Dataset	#Rows	#Columns	#Cells	# Ans. per Task (Avg)
Celebrity	174	7	1218	5
Restaurant	203	5	1015	4
Emotion	100	7	700	10

M columns, the total cost is $\mathcal{O}(M \cdot |\mathcal{A}|)$. The same time is needed to calculate the correlation coefficient W_{jk}. The cost of computing the benefit of each task is the same as that of computing the Inherent Information Gain, which is discussed before. In total, the cost is $\mathcal{O}((M + sl) \cdot |\mathcal{A}|)$.

Assigning Multiple Tasks to Workers. So far we focused on how to select one task to assign to the incoming worker. This does not restrict the applicability of our approach in the case that multiple tasks should be determined and given to the worker as a batch (e.g., as in a HIT on AMT [1]). Suppose that the worker is to be assigned a set $D = \{c_{i_1j_1}, c_{i_2j_2}, \ldots, c_{i_Kj_K}\}$ of K tasks. From the set $\mathcal{A}_{\mathcal{D}} = \{a^u_{i_1j_1}, a^u_{i_2j_2}, \ldots, a^u_{i_kj_k}\}$ of estimated answers to the tasks by the worker, we can update the distribution of the estimated truth $T_{ij,\mathcal{A}_C \cup \mathcal{A}_{\mathcal{D}}}$ for each task $c_{ij} \in D$. Then, we can calculate the information gain for D as:

$$IG(D) = \sum\nolimits_{c_{ij} \in D} \left(H(T_{ij,\mathcal{A}_C}) - \mathrm{E}_{\mathcal{A}_{\mathcal{D}}}[H(T_{ij,\mathcal{A}_C \cup \mathcal{A}_D})] \right). \quad (9)$$

Because the search space of D is $\binom{N \cdot M}{K}$, finding K tasks which maximize $IG(D)$ is expensive. To alleviate the cost, we can apply a greedy approach that iteratively selects the top-K tasks with the largest $IG(c_{ij})$.

6 Experiments

This section presents our experimental results. We present the datasets used in Section 6.1. In Sections 6.2 and 6.3, we compare different crowdsourcing solutions in terms of truth inference and task assignment respectively. We perform case studies in Section 6.4. Results on synthetic datasets are shown in Section 6.5. We measure the efficiency in Section 6.6 and do an extra comparison to CrowdFill in Section 6.7. We have implemented a prototype of T-Crowd and other crowdsourcing solutions in Python 2.7, on a Ubuntu server with 8-core Intel(R) Core(TM) i7-3770 CPU @ 1.60 GHz cores and 16 GB memory.

6.1 Datasets

We use three real datasets to perform our experiments. Their statistics are shown in Table 5.

Celebrity [6]. This dataset contains information about celebrities. Given a celebrity's picture, workers are requested to provide the following attribute values: name, age, height, nationality, ethnicity, notability, and sentiment of the celebrity in the picture. Name, nationality, and ethnicity are categorical, age, height, notability, and sentiment are continuous. For each entity, the ground truth for name and age are obtained from [6], while that of height, nationality and ethnicity is extracted from IMDb. Notability and Sentiment are the subjective attributes. Our marked ground truth for subjective questions is the answer that agrees with most persons' opinion. The ground truth of sentiment is labeled by three movie experts ranging from 1 to 5 based on facial expressions. The experts followed the measurement of the facial action coding system (FACS) and the emotional facial action system (EMFACS)[13], [16], [35]. This includes 44 action units and combinations of FACS action units represent prototypic expressions of emotion. The positive expressions (happy or relaxed) take 5, the neutral take 3 and the negative ones (sad, fear or angry) take 1. This measurement has been researched in field of facial expression analysis in many years and it can be regarded as objective. Notability is also a value in the range of 1-5 and is obtained from the persons rank in IMDb, i.e., we map 1-200 as 5, 201-400 as 4, 401-600 as 3, 601-800 as 2 and > 800 as 1. Since ratings in IMDb are solicited from numerous people, we believe it represents the opinions of the majority.

Restaurant [30]. This dataset contains information about restaurants. Given a review about a certain restaurant, workers are asked to specify the aspect (e.g., food or location), attribute (e.g., price or style), and sentiment (e.g., negative or positive) of the review. They are asked to identify the target (i.e., the restaurant referred by the review) by the starting and end position of its first occurrence in the text. Here, aspect, attribute, and sentiment are categorical; the starting and end positions are continuous. The reviews and the ground truth are obtained from [30].

Emotion [34]. This dataset collects scores for different emotions from a small piece of text. Each worker is asked to give a number in [0,100] for each of the following six emotions: anger, disgust, fear, joy, sadness, and surprise, and a single numeric rating in the interval [-100,100] for her overall (positive or negative) sentiment about the text. Here, all the 7 attributes are continuous. The workers' answers and the ground truth are provided by [34].

For the Celebrity and Restaurant datasets, we collected the workers' answers using AMT [1]. The average number of answers for each task in Celebrity and Restaurant is 5 and 4, respectively, by different workers. We spent $\$0.05$ per HIT where the number of tasks put in a HIT is the same as the number of columns (total cost $\$43.5$ and $\$40.6$, respectively). For Emotion, we use the workers' answers from [34]; each task is answered 10 times. We observed that for all continuous attributes the collected values (excluding spam answers) follow a normal distribution, which is consistent with our assumption in Section 4.1.

6.2 Truth Inference

We select some important other existing solutions based on the guidance from [42],[19] and study the effectiveness of our truth inference approach:

(1) For both categorical and continuous data:
- T-Crowd is our method proposed in Section 4. *TC-onlyCate* and *TC-onlyCont* are the constrained versions of T-Crowd that apply only on the categorical or continuous attributes.
- CRH [21] detects truth from heterogeneous data types by minimizing a loss function.
- CATD [20] detects truth from multi-source data that follows a long-tail distribution along with confidence intervals.

(2) For categorical data only:

TABLE 6
Effectiveness of Truth Inference

Method	Celebrity		Restaurant		Emotion
	Error Rate	MNAD	Error Rate	MNAD	MNAD
T-Crowd	**0.0441**	**0.6339**	**0.1855**	**0.5607**	**0.5961**
CRH	0.0460	0.6737	0.1921	0.5835	0.7224
CATD	0.0498	0.7113	0.1954	0.7234	0.6648
Maj. Voting	0.0573	/	0.2003	/	/
EM	0.0620	/	0.2463	/	/
GLAD	0.0498	/	0.1905	/	/
Zencrowd	0.0479	/	0.1872	/	/
TC-onlyCate	0.0498	/	0.1986	/	/
Median	/	0.6998	/	0.6784	0.7026
GTM	/	0.6516	/	0.5871	0.6792
TC-onlyCont	/	0.6400	/	0.5682	**0.5961**

- Majority Voting (MV) determines the correct labels based on the majority of answers from workers.
- D&S [9] iteratively estimates each worker's confusion matrix, which is used to infer the correct labels.
- GLAD [38] is a probabilistic approach for crowdsourcing categorical data.
- Zencrowd [10] is a variant of D&S.

(3) For continuous data only:
- Median uses the median of workers' answers as the estimated true value.
- GTM [40] is a truth-finding method specially designed for continuous data.

Effectiveness Measures. We adopt the following measures, proposed in [21], for evaluating the effectiveness of truth inference on categorical and continuous data items:

- Error Rate: For categorical data, we measure the Error Rate by computing the percentage of mismatched values between each method's predicted truth and the ground truth.
- MNAD (Mean Normalized Absolute Distance): It is the root of mean squared distance (RMSE) between each method's estimated truth and the ground truth. Since different attributes have different scales, we normalize each attribute's RMSE by its own standard deviation and average them.

Effectiveness Comparison. In Table 6, we summarize the effectiveness of truth inference by all methods in terms of Error Rate and MNAD on the three real-world datasets. We observe that our proposed approach T-Crowd is better than all other methods both on categorical data and continuous data. On Celebrity, our method reduces the error rate by 4 percent on categorical data and the MNAD by 2.7 percent on continuous data compared to the best result of other methods. The corresponding reductions on Restaurant are 2.6 and 4 percent. On Emotion, we outperform previous work by 10 percent. CRH does not have stable performance as it is effective on Celebrity and Restaurant, but ineffective on Emotion. Similarly, CATD is good in terms of error rate but not good in terms of MNAD. Overall, our method is more robust than them.

We also test constrained versions of T-Crowd that apply only on the categorical or only on the continuous attributes. Note that the effectiveness of T-Crowd is better than that of its constrained versions and that the constrained versions are competitive compared to other methods in their class. In summary, T-Crowd outperforms truth inference approaches applied on categorical and continuous data separately. This result demonstrates the benefit of modeling worker quality by a probabilistic model in a unified manner for all datatypes.

6.3 Task Assignment

We compare the effectiveness of task assignment by our approach against other crowdsourcing methods.

Competitors. We compare T-Crowd, which uses the truth inference method of Section 4.2 and the task assignment method in Section 5.2 with the following approaches:

- CDAS [23] measures the confidence of the currently estimated values of all tasks based on a quality-sensitive answering model. Each task for which we are already confident is "terminated" and no longer assigned to workers. At each step, CDAS selects at random a non-terminated task to assign to the incoming worker.
- AskIt! [5] uses an entropy-like method to define the uncertainty of each task, and infers the truth by Majority Voting. The task with the highest uncertainty is the next one to be assigned to the incoming worker.
- CRH [21] is an inference method suitable for heterogeneous data. It does not focus on task assignment, hence, tasks are randomly assigned to the incoming workers.
- CATD [20] is an inference method suitable for heterogeneous data, which does not focus on task assignment. Similar to CRH, we collected answers by randomly assigning tasks.

Effectiveness Measures. As in the evaluation of truth inference, we use Error Rate and MNAD to measure task assignment quality. Specifically, for each tested method, we measure the Error Rate and MNAD as a function of the average number of answers collected by task so far. A good method would be able to converge fast with fewer answers per task(i.e., by performing fewer assignments and hence spending less money). Besides, it should achieve a lower true value estimation error when it converges.

End-To-End Comparison. To perform a fair comparison with existing work, we performed experiments on AMT [1] by using the same settings for the different methods (i.e., each task costs the same). We use the 'external-HIT' [2] feature provided by AMT to dynamically assign tasks for the incoming worker. To assess the effectiveness of task assignment, we vary the budget and compare the Error Rate and MNAD of each method under the same budget. To be specific, for each budget, we record the error rate and MNAD on all real datasets as more answers are collected.

Fig. 2 shows the experimental results. Naturally, the error rate and MNAD of all assignment policies decrease as more answers are received from the workers and converge to good results after a large number of answers. Askit! uses an entropy-like method, which makes it to prefer continuous tasks first. Thus its MNAD drops fast while the error rate remains high. After selecting all continuous tasks, its error rate starts to drop. Since no task is terminated in the first few iterations, CDAS converges slowly. In addition, since its

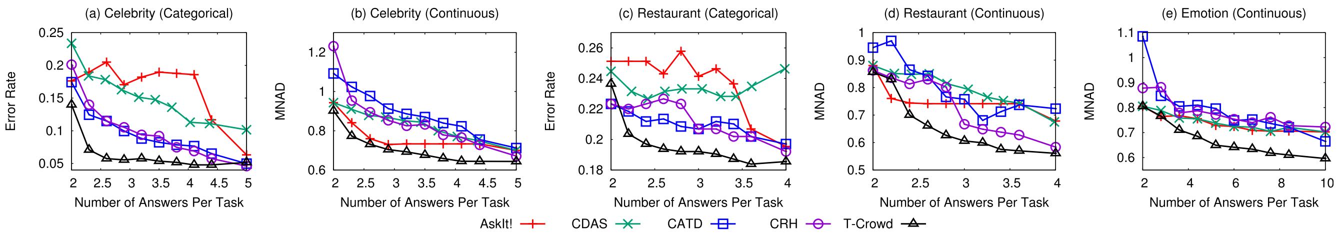


Fig. 2. End-to-end system comparison (effectiveness).

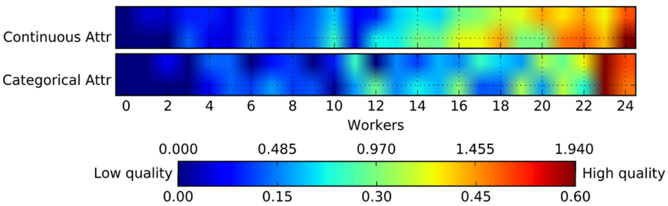
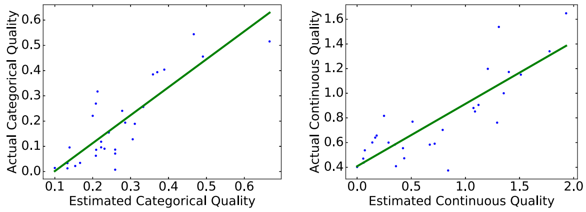


Fig. 3. Uniform worker quality.



(a) categorical (b) continuous

Fig. 4. Estimated and actual worker quality.

inference method is simple, the final inferred result is not good compared to that of other methods. CRH and CATD are not probabilistic, which do not use metrics, like entropy or information gain, as the objective for task assignment, so they do not perform as well as T-Crowd. They are superior to Askit! and CDAS because they are more effective in inferring the true values of tasks.

We observe that T-Crowd converges much faster to a low error rate and MNAD compared to the other policies. Specifically, T-Crowd converges to low values before the average number of answers per task is 3 on Celebrity and Restaurant and 6 on Emotion, which shows the effectiveness of our structure-aware information gain measure as an assignment criterion. In addition, due to our superior truth inference method, the values eventually inferred by our framework are better compared to those inferred by the other methods.

6.4 Case Studies

We performed several case studies in order to assess the quality of our system. Due to space constraints, we only report the results on Restaurant. Similar observations can be derived by experimentation on the other datasets.

6.4.1 Worker Quality

Our first study's goal is to show that (1) each worker's actual quality (computed based on the ground truth) is consistent among different attributes; (2) each worker's estimated quality can be well calibrated to the worker's actual quality.

Consistent Quality for Different Attributes. We collected statistics from the Restaurant dataset to support our assumption in truth inference: a worker has consistent quality over different datatypes of attributes. In Fig. 3, we plot a heat map, with the x-axis representing the 25 workers who have given the largest number of answers and the y-axis representing categorical attributes 'Aspect' and 'Sentiment' and continuous attributes 'StartTarget' and 'EndTarget'. Different colors are aligned to standard deviation values (above the colorbar) for continuous attributes and error rates (below the colorbar) for categorical attributes. The color of each pixel represents the average error of answers given by worker u to the tasks on attribute j. For a categorical attribute j, the error is the percentage of wrong answers. For a continuous attribute j, the error is the standard deviation of the differences between the answers and the ground truth. The red color (far right) implies larger error and lower worker quality, while the blue color (far left) means smaller error and better worker quality. Note that the workers have consistent performance for categorical and continuous attributes. In addition, the colors for the same worker are similar regardless the attribute type, i.e., each worker's actual quality is consistent among different attributes.

Calibration to the Actual Quality. Fig. 4 shows that our estimated quality of a worker is close to the actual quality. Each point represents a worker and the x-axis value is the quality estimated by our method while the y-axis value is the actual worker's quality. We also show the result of a linear regression. Observe the strong correlation between our estimation and actual quality; the correlation coefficient is 0.844 for categorical and 0.841 for continuous attributes.

6.4.2 Assignment Heuristics

We evaluate the performance of different assignment heuristics. Note that for all of them, we use our inference approach (Section 4.2). The tested heuristics are listed as follows:

- Random: it randomly chooses the task assigned to the worker.
- Looping: it selects the next task in a round-robin manner.
- Entropy: it greedily chooses the next task which has the highest uncertainty (defined as entropy).
- Inherent Information Gain: it proposed in Section 5.1.
- Structure-Aware Information Gain: it proposed in Section 5.2.

Fig. 5 presents the Error Rate and MNAD as a function of number of tasks assigned to the workers on Restaurant. The results on the other datasets are similar and omitted for the interest of space. Random and Looping select tasks without considering the answers collected so far, so they converge slowly. Entropy is biased toward selecting continuous tasks over categorical first; hence, this heuristic reduces the MNAD

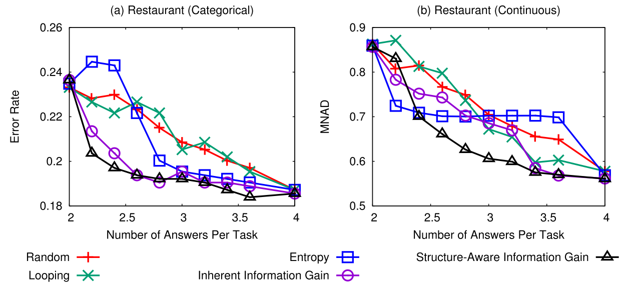


Fig. 5. Effectiveness of assignment heuristics.

Aspect \ Sentiment	correct	wrong
correct	589	90
wrong	98	35

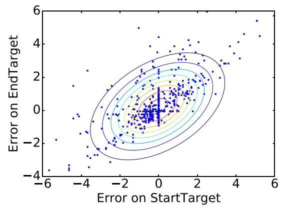
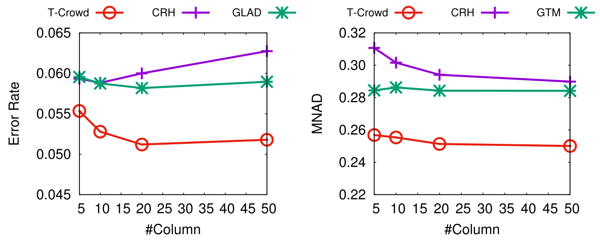


Fig. 6. Correlation among attributes.



(a) Categorical Columns (b) Continuous Columns

Fig. 7. Effect of the number of columns.

fast, but not the Error Rate. Inherent and Structure-Aware Information Gain consider the continuous and categorical tasks fairly and decrease the Error Rate and MNAD simultaneously. Besides, Structure-Aware Information Gain converges faster than Inherent Information Gain w.r.t. MNAD because it also considers the correlations between attributes. Recall that we use Structure-Aware Information Gain as our default method.

6.4.3 Correlation Among Attributes

We perform one more experiment to support our assumption that there exist correlations among attributes, by analyzing the answers of workers.

Fig. 6 shows the experimental results. In the left part of the figure observe that attributes 'Aspect' and 'Sentiment' have strong correlation. Specifically, if a worker answers 'Aspect' correctly, the probability to answer 'Sentiment' correctly is 86 percent. However, if a worker answers 'Aspect' wrongly, the probability to answer 'Sentiment' correctly is only 73 percent. In the right part of the figure, we plot a scatter diagram, with each point representing a worker's error on attributes 'StartTarget' and 'EndTarget'. We use *maximum likelihood estimation* to obtain the joint distribution of errors on these two attributes as described in Section 5.2. We observe a positive correlation between errors on attributes 'StartTarget' and 'EndTarget', which justifies our proposed Structure-Aware Information Gain method that considers correlations among attributes. For example, if the error of 'StartTarget' is 0, the distribution of 'EndTarget' error is $\mathcal{N}(0.28, 0.76)$. However, if the error of 'StartTarget' is 6, the distribution of 'EndTarget' error is $\mathcal{N}(3.75, 0.76)$. In other words, knowing the exact answer of a worker on one attribute can help to predict his/her answer distribution for other attributes better.

6.5 Synthetic Data

In this section, we use two types of synthetic data, in order to test the performance of our truth inference approach in cases not covered by the real data settings.

6.5.1 Tests on Tables with Different Properties

We assess the performance of T-Crowd in terms of truth inference effectiveness by changing the following parameters of our data generator: the number of columns M, the ratio of categorical to the total number columns R and the average difficulty of tasks $\mu\{\alpha_i\beta_j\}$. The default parameters are $M = 10$, $R = 0.5$ and $\mu\{\alpha_i\beta_j\} = 1$. The rest of the settings are as follows:

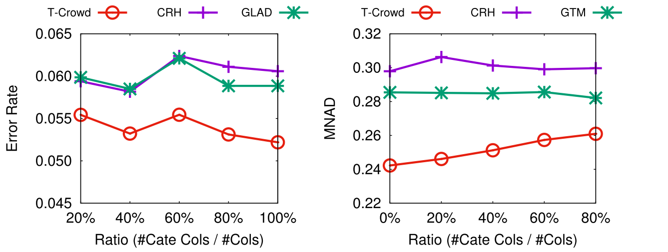
Worker Sequence and Worker Quality: We use the same number of workers as that in our real experiments for the dataset Celebrity and assume that the workers arrive in the same sequence and that they have the same quality as in the real experiment.

Data and Ground Truth Generation: We implemented a generator for a table that takes as input the number of rows N and columns M, and the datatype and domain range of each column. The number of possible answers in a categorical column is generated from a uniform distribution $U(2, 10)$. The domain of a continuous column is $[0, 1000]$. The ground truth T^*_{ij} of each cell c_{ij} is generated by selecting a value in the corresponding domain randomly.

Workers' Answers: For each worker in sequence, his answer at each cell needs to be generated. The answer a^u_{ij} of each worker u at each cell c_{ij} is created based on the ground truth T^*_{ij} and his quality q_u, based on Eqs. (1) and (3).

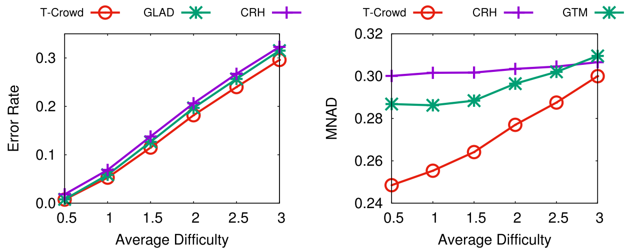
For fairness to all methods , we simulate the assignment strategy used in AMT, i.e., each task gets the same number of answers. For different parameters, we generate new datasets one hundred times and average the results to obtain the error rate and MNAD. We also run other inference methods and found that our method is dominant both on error rate and MNAD.

Results. In the first experiment, we vary the number of columns from 5 to 50. Fig. 7 shows that the error rate and MNAD decline gradually when the number of columns increases, showing that T-Crowd infers the quality of each worker and estimates truth more accurate if we have more data. Besides our method is significantly better than the other two approaches. Next, we vary the ratio of categorical attributes from 0 to 100 percent. Figs. 8a and 8b show that our method's error rate and MNAD do not change much when the ratio varies. Finally, we vary the average difficulty of each cell c_{ij} (i.e., the average $\alpha_i\beta_j$, as defined in Section 4.1.2) from 0.5 to 3. High difficulty implies that the probability that workers answer correctly decreases, hence the error rate and MNAD increase as shown in Fig. 9. For easier tasks, our method is significantly better than the others, but when the



(a) Categorical Columns (b) Continuous Columns

Fig. 8. Effect of ratio of columns.



(a) Categorical Columns (b) Continuous Columns

Fig. 9. Effect of average difficulty.

average difficulty is high, which means that the workers' answers are not credible, all methods perform badly.

6.5.2 Noise in Workers' Answers

To further demonstrate the advantage of our proposed approach T-Crowd, we conduct simulation experiments by adding noise to the original data collected for Celebrity dataset. We vary the percentage γ of altered original answers by the workers from 10 to 40 percent (i.e., γ is the percentage of answers with added noise).

For a categorical answer, we randomly select a new label from its domain and replace the original label. For a continuous answer, Gaussian noise is added. We first preprocess this answer by transforming it into its z-score. A new normalized answer is generated by adding the noise which was generated by a Gaussian distribution $\mathcal{N}(0,1)$. We finally change it to the original scale and obtain the new answer. We randomly choose $NM\gamma$ answers with replacement to add noise and keep the rest the same.

For different levels of noise γ, we generate new datasets one hundred times. For each method, we run experiments 3 times to smoothen out possible instabilities. Hence we run in total 300 simulations for each method and average them to obtain the error rate and MNAD for different levels of γ.

Figs. 10a and 10b show the results. The error rate increases while MNAD declines when γ increases. The reason for the decrease of MNAD is that the normalization denominator is the standard deviation of answers in each column. The growth rate of standard deviation is higher than that of RMSE which makes MNAD to decline.

T-Crowd performs well and stably when the level of noise γ increases both in terms of error rate and MNAD. T-Crowd has a very similar error rate and MNAD to CRH and GTM, respectively.

6.6 Efficiency

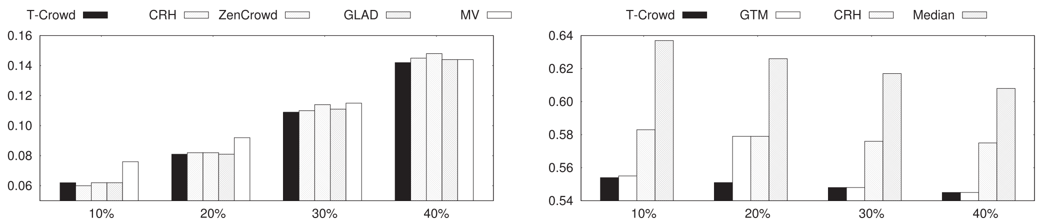
We first investigate the truth inference cost on Celebrity dataset and then show its running time on a single machine. Fig. 12a shows the change of the objective value in truth inference at each iteration. Even iteration is the objective value from M step while odd iteration is the objective value from E step. Note that our inference model converges to the estimated value, after only a few iterations.The curve of objective function is different when initial parameters $(\alpha_i, \beta_j, \phi_u)$ are different. We random the parameters several times, average the value and plot the line. We also draw the error bar in the figure to show the minimum and maximum value in each iteration.

Then, we confirm the low cost of truth inference by measuring the throughput of T-Crowd, i.e., how many answers it can process per second. For this purpose, we use synthetic data used in Section 6.5 since the number of answers collected for real data is limited. Fig. 12b shows that the runtime of T-Crowd is approximately linear to the number of answers; T-Crowd can process approximately 100 answers per second on a single machine. This performance is acceptable, given that the rate of incoming answers is much lower in a real crowdsourcing system. The performance is also consistent with our time complexity analysis at the end of Section 4.2.

Finally, we measure the time required to assign a new task to an incoming worker on the Celebrity dataset (Fig. 11). We assume that we already obtained the estimated truth using T-Crowd's truth inference module. We show the running time of computing the *structure-aware* information gain for all candidate tasks each time a new worker arrives. Because it is easy to parallelize task assignments, we run eight processes on our machine. As shown in Fig. 11, the assignment cost increases linearly with the average number of answers collected so far for each task. This is consistent to our complexity analysis at the end of Section 5.1, which suggests that the cost is linear to the total number $|\mathcal{A}|$ of answers so far. Still, as the figure shows, new assignments can be conducted in real-time, which is important for a real crowdsourcing platform.

6.7 Comparison to CrowdFill

CrowdFill [29] is a recent crowdsourcing system for tabular data. In CrowdFill, each worker is shown a fragment of a



(a) Error Rate (b) MNAD

Fig. 10. Noisy datasets.

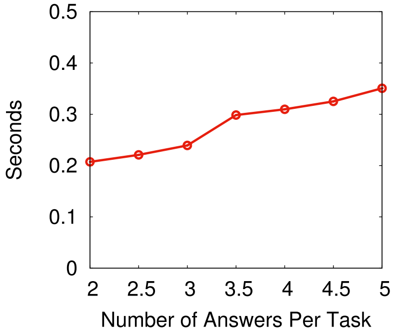
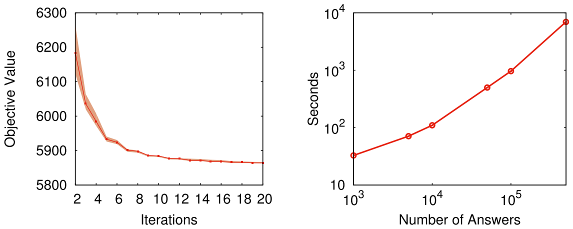


Fig. 11. Efficiency of assignment.



(a) Convergence Rate (b) Truth Inference

Fig. 12. Efficiency of inference.

partially-filled table and asked to fill in empty cells, or upvote/downvote the answers entered by other workers. Compared to T-Crowd and the other methods that we have examined, CrowdFill requires the crowdsourcing platform to include additional functions (upvote and downvote operations), which are not currently supported by AMT. Hence, we compare to CrowdFill independently.

Still, to compare the effectiveness of T-Crowd with that of CrowdFill, we conducted an experiment, following the experimental setup of [29]. We collected information about 20 Olympic champions from 5 human workers. Given the picture of an athlete, the objective is to collect information about his/her attributes {*name*, *isRetired*, *#attendedOlympiads*, *#goldMedals*, *currentAge*, *ageInPic*}. Attributes *name* and *isRetired* are categorical and the remaining ones are continuous.

To be fair to both CrowdFill and T-Crowd, each worker was requested to answer questions twice. In the first experiment, workers give their answers independently following the T-crowd setting. The answers are collected and aggregated by T-Crowd to get the final table. In the second experiment, we use Google Docs to simulate CrowdFill's collaborative process. That is, workers can view other workers' answers, and they can choose between filling an empty cell or upvoting/downvoting a completed row. When the number of votes is larger than 2, a row is accepted if its upvotes are more than its downvotes; otherwise, it is rejected and it is offered again to workers to fill in their answers. When all the rows are accepted, we obtain the final table for CrowdFill.

Table 7 shows the error of these two methods. As in the previous experiments, we show the Error Rate for categorical data and RMSE for continuous data. Observe that T-Crowd is more accurate than CrowdFill for continuous attributes and the two methods have similar accuracy for categorical attributes. As opposed to T-Crowd, CrowdFill does not compute and use the unified worker quality for continuous and categorical attributes, which negatively affects its performance on continuous attributes, for which the collected answers are sparser.

TABLE 7
Error of CrowdFill

	Name	IsRetired	#attended Olympiads	#gold Medals	currentAge	ageInPic
CrowdFill	0.25	0.2	1.24	1.04	4.40	3.22
T-Crowd	0.25	0.15	0.71	0.67	3.17	2.38

7 Conclusions

In this paper we design a crowdsourcing framework for collecting multi-type tabular data. Most existing methods, which are designed for simple tasks that are all of the same datatype are not effective enough in terms of both truth inference and task assignment. Based on the characteristics of tabular data, we propose a probabilistic truth inference model that unifies worker quality on both categorical and continuous datatypes. Besides, we improve the accuracy of truth inference by considering the variance in the difficulty of different tasks. In addition, we design an information gain function which we use for selecting the tasks to assign to workers, based on the current answers and the workers' quality. We extend this function to consider the correlation in the quality of certain worker's answers for the same entity. Our experiments on three real datasets and synthetic datasets confirm the superiority of our methods, both in truth inference and task assignment compared to the state-of-the-art.

In the future, we plan to conduct experiments with larger tables compared to the ones we have used in Section 6. In addition, we plan to extend our approach to apply on tables for which entities are not known. In this case, entities should also be collected from the crowd. A third direction is the acceleration of truth inference and task assignment by parallel and/or distributed computation. Finally, we will explore the possible improvement of our approach by exploiting the possible correlations between entities (not only attributes), e.g., a worker may be more familiar to celebrities starring in a certain category of films or shows.

Acknowledgments

The authors would like to thank their reviewers for the insightful comments. Guoliang Li was supported by the 973 program of China (2015CB358700), NSF China (61632016, 61521002, 61661166012), Huawei, and TAL education. Nikos Mamoulis, Reynold Cheng, Caihua Shan, Zhipeng Huang, and Yudian Zheng were supported by the Research Grants Council of Hong Kong (RGC Projects HKU 17253616, 17229116, 106150091 and 17205115), the University of Hong Kong (Projects 104004572, 102009508, 104004129), and the Innovation and Technology Commission of Hong Kong (ITF project MRP/029/18).

References

[1] Amazon Mechanical Turk, [Online]. Available: https://www.mturk.com/mturk/, Last accessed: May. 17, 2019.
[2] AMT External Question, [Online]. Available: http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/ApiReference_ExternalQuestionArticle.html, Last accessed: May. 17, 2019.

[3] Entropy, [Online]. Available: https://en.wikipedia.org/wiki/Entropy_(information_theory), Last accessed: May. 17, 2019.

[4] ERF Function. http://mathworld.wolfram.com/Erf.html, Last accessed: May. 17, 2019.

[5] R. Boim, O. Greenshpan, T. Milo, S. Novgorodov, N. Polyzotis, and W.-C. Tan, "Asking the right questions in crowd data sourcing," in *Proc. IEEE 28th Int. Conf. Data Eng.*, 2012, pp. 1261–1264.

[6] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 768–783.

[7] X. Chen, Q. Lin, and D. Zhou, "Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 64–72.

[8] S. Das, P. S. GC, A. Doan, J. F. Naughton, G. Krishnan, R. Deep, E. Arcaute, V. Raghavendra, and Y. Park, "Falcon: Scaling up hands-off crowdsourced entity matching to build cloud services," in *Proc. ACM Int. Conf. Manage. Data*, 2017, pp. 1431–1446.

[9] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Appl. Statistics*, pp. 20–28, 1979.

[10] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, "Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 469–478.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Royal Statistical Society. Series B (methodological)*, vol. 39, pp. 1–38, 1977.

[12] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: The role of source dependence," *Proc. VLDB Endowment*, vol. 2, no. 1, pp. 550–561, 2009.

[13] P. Ekman, W. V. Freisen, and S. Ancoli, Facial signs of emotional experience, *J. personality and social psychology*. Cary, NC, USA: American Psychological Association, vol. 39, no. 6, p. 1125, 1980.

[14] J. Fan, G. Li, B. C. Ooi, K.-l. Tan, and J. Feng, "icrowd: An adaptive crowdsourcing framework," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2015, pp. 1015–1030.

[15] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin, "Crowddb: Answering queries with crowdsourcing," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2011, pp. 61–72.

[16] W. V. Friesen, P. Ekman, et al., "Emfacs-7: Emotional facial action coding system," unpublished manuscript, University California at San Francisco, vol. 2, no. 36, 1983, Art. no. 1.

[17] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *Proc. SIGKDD Workshop Human Comput.*, 2010, pp. 64–67.

[18] A. R. Khan and H. Garcia-Molina, "Crowddqs: Dynamic question selection in crowdsourcing systems," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2017, pp. 1447–1462.

[19] G. Li, J. Wang, Y. Zheng, and M. J. Franklin, "Crowdsourced data management: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 9, pp. 2296–2319, Sep. 2016.

[20] 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," *Proc. VLDB Endowment*, vol. 8, no. 4, pp. 425–436, 2014.

[21] 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 *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 1187–1198.

[22] Y. Li, Q. Li, J. Gao, L. Su, B. Zhao, W. Fan, and J. Han, "Conflicts to harmony: A framework for resolving conflicts in heterogeneous data by truth discovery," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 1986–1999, Sep. 2016.

[23] X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu, and M. Zhang, "CDAS: A crowdsourcing data analytics system," *Proc. VLDB Endowment*, vol. 5, no. 10, pp. 1040–1051, 2012.

[24] 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 *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 745–754.

[25] A. Marcus, E. Wu, D. R. Karger, S. Madden, and R. C. Miller, "Demonstration of qurk: a query processor for humanoperators," in *Proc. 2011 ACM SIGMOD Int. Conf. Management of data*, pp. 1315–1318, 2011.

[26] P. Mavridis, D. Gross-Amblard, and Z. Miklós, "Using hierarchical skills for optimized task assignment in knowledge-intensive crowdsourcing," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 843–853.

[27] J. V. Michalowicz, J. M. Nichols, and F. Bucholtz, *Handbook of Differential Entropy*. Boca Raton, FL, USA: CRC Press, 2013.

[28] H. Park, H. Garcia-Molina, R. Pang, N. Polyzotis, A. Parameswaran, and J. Widom, "Deco: A system for declarative crowdsourcing," *Proc. VLDB Endowment*, vol. 5, no. 12, pp. 1990–1993, 2012.

[29] H. Park and J. Widom, "Crowdfill: Collecting structured data from the crowd," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 577–588.

[30] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "Semeval-2014 task 4: Aspect based sentiment analysis," in *Proc. 8th Int. Workshop Semantic Evaluation*, 2014, pp. 27–35.

[31] H. Rahman, S. B. Roy, S. Thirumuruganathan, S. Amer-Yahia, and G. Das, "Task assignment optimization in collaborative crowdsourcing," in *Proc. IEEE Int. Conf. Data Mining*, 2015, pp. 949–954.

[32] S. B. Roy, I. Lykourentzou, S. Thirumuruganathan, S. Amer-Yahia, and G. Das, "Task assignment optimization in knowledge-intensive crowdsourcing," *Proc. VLDB Endowment J.*, vol. 24, no. 4, pp. 467–491, 2015.

[33] C. Shan, N. Mamoulis, G. Li, R. Cheng, Z. Huang, and Y. Zheng, "T-crowd: Effective crowdsourcing for tabular data," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, 2018, pp. 1316–1319.

[34] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2008, pp. 254–263.

[35] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," in *Handbook of Face Recognition*. Berlin, Germany: Springer, 2005, pp. 247–275.

[36] V. Verroios, H. Garcia-Molina, and Y. Papakonstantinou, "Waldo: An adaptive human interface for crowd entity resolution," in *Proc. ACM Int. Conf. Manage, Data*, 2017, pp. 1133–1148.

[37] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "Crowder: Crowdsourcing entity resolution," *Proc. VLDB Endowment*, vol. 5, no. 11, pp. 1483–1494, 2012.

[38] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. 22nd Int. Conf. Neural Inf. Process. Syst.*, 2009, pp. 2035–2043.

[39] X. Yin, J. Han, and S. Y. Philip, "Truth discovery with multiple conflicting information providers on the web," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 6, pp. 796–808, Jun. 2008.

[40] B. Zhao and J. Han, "A probabilistic model for estimating real-valued truth from conflicting sources," in *Proc. Workshop Quality Databases*, 2012.

[41] Y. Zheng, G. Li, and R. Cheng, "Docs: A domain-aware crowdsourcing system using knowledge bases," *Proc. VLDB Endowment*, vol. 10, no. 4, pp. 361–372, 2016.

[42] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, "Truth inference in crowdsourcing: Is the problem solved?" *Proc. VLDB Endowment*, vol. 10, no. 5, pp. 541–552, 2017.

[43] Y. Zheng, J. Wang, G. Li, R. Cheng, and J. Feng, "Qasca: A quality-aware task assignment system for crowdsourcing applications," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2015, pp. 1031–1046.

[44] H. Zhuang, A. Parameswaran, D. Roth, and J. Han, "Debiasing crowdsourced batches," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1593–1602.

Caihua Shan received the BE degree in computer science and technology from Zhejiang University, in 2015. She is currently working toward the PhD degree in the Department of Computer Science, The University of Hong Kong. Her research mainly focus on data management in crowdsourcing.

Nikos Mamoulis received the diploma degree in computer engineering and informatics from the University of Patras, Greece, in 1995, and the PhD degree in computer science from the Hong Kong University of Science and Technology. He is currently a faculty member at the University of Ioannina. Before, he was a professor with the Department of Computer Science, University of Hong Kong. His research focuses on the management and mining of complex data types.

Guoliang Li received the PhD degree in computer science from Tsinghua University, Beijing, China, in 2009. He is currently working as a professor with the Department of Computer Science, Tsinghua University, Beijing, China. His research interests mainly include crowdsourced data management, large-scale data cleaning, and integration.

Reynold Cheng received the PhD degree from the Department of Computer Science of Purdue University, in 2005. He is an associate professor of the Department of Computer Science, University of Hong Kong (HKU). He was granted an Outstanding Young Researcher Award in 2011-2012 by HKU.

Zhipeng Huang received the BSc degree in machine intelligence from Peking University, China in 2011. He is now working toward the PhD degree in the Department of Computer Science, University of Hong Kong. His research focuses on the mining of heterogeneous information networks.

Yudian Zheng is currently at Twitter, focusing on building large scale online distributed machine learning systems. His main research interests include data analysis, crowdsourced data management, and machine learning. He has published more than 20 papers in leading international conferences/journals, including SIGMOD, KDD, VLDB, *TKDE*, ICDE, etc.