2018 IEEE 34th International Conference on Data Engineering

T-Crowd: Effective Crowdsourcing for Tabular Data

Caihua Shan [†] Nikos Mamoulis [§] Guoliang Li [#] Reynold Cheng [†] Zhipeng Huang [†] Yudian Zheng [‡]
[†] The University of Hong Kong, [§] University of Ioannina, [#] Tsinghua University, [‡] Twitter Inc.
{chshan, ckcheng, zphuang}@cs.hku.hk, nikos@cs.uoi.gr, liguoliang@tsinghua.edu.cn, yudianz@twitter.com

***Abstract*—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?*). Existing work often treats related attributes independently, leading to suboptimal performance. We present T-Crowd: a crowdsourcing system that considers attribute relationships. T-Crowd integrates each worker's answers on different attributes to effectively learn his/her trustworthiness and the true data values. Our solution seamlessly supports categorical and continuous attributes. Our experiments on real datasets show that T-Crowd outperforms state-of-the-art methods, improving the quality of truth inference.**

I. Introduction

Crowdsourcing is an effective way to address computer-hard problems 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.

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. Particularly, a column represents a particular attribute or variable; a row corresponds to an entity and includes a value for each variable. Table I 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 I are the unknown (ground) truth data to be collected from the workers. Each cell of this table can be considered as a task, where 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 applications in database cleaning and integration [6] [10] [11].

Table I: 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**

A fundamental problem in crowdsourcing is *truth inference*: how to infer the ground truth data from the answers of multiple workers. Most crowdsourcing systems assume that the set of tasks are homogeneous and independent. However, tasks in tabular data can be *heterogeneous* and *dependent* on each other, which makes effective crowdsourcing on them challenging. First, the datatypes and domains of different attributes (e.g., Nationality and Age) may vary. 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 [4], [8], [16] are not directly applicable. Second, in tabular data, there are potential dependencies between rows and columns. The difficulty of a task (which depends on the corresponding entity and attribute) affects the accuracy of inference.

In this paper, we present *T-Crowd*, a crowdsourcing system that considers the heterogeneous nature and the dependencies between missing values in a table. T-Crowd processes the submitted answers by each worker to infer a *unified quality* for him or her and 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. Besides, T-Crowd captures the difficulty of a task based on which row and column it is in. For instance, tasks about Leonardo DiCaprio (row 1) are easier than tasks about James Purefoy (row 3) because the former is more famous. Similarly tasks about age are harder than tasks about nationality. We evaluate T-Crowd on real datasets; the results demonstrate its superiority over existing alternatives. T-Crowd has better truth inference accuracy than previous work.

II. Related Work

A simple inference method is 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 [4] 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 [15], Accusim [5], and GLAD [14] improve accuracy using different worker answering models or by considering more parameters, such as a task's difficulty. These methods focus on answering tasks on categorical data. Other methods, such as GTM [16], are designed for continuous crowdsourced data. CRH [9] is one existing truth inference approach for both categorical and continuous data. It incorporates different distance functions between the answers and the estimated truth to recognize the characteristics of various data types. Specifically, CRH

2375-026X/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDE.2018.00139

proposes an objective function and minimizes it by updating the estimated true values and source reliability (i.e., worker quality) in turns. Additional information of tasks or workers has also been considered, such as the latent topics of the tasks and the learned bias of workers.

The aforementioned works do not consider tabular data. In Section IV, we present an iterative Expectation-Maximization (EM) truth inference algorithm, which improves upon previous work. The novelty of our work is that we use a probabilistic model for the answers of workers w.r.t. 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.

III. Problem Definition

Definition 1 (Tabular Data Model): We target the crowdsourcing of a two-dimensional table $C = \{c_{ij}\}$, where $i \in \{1, ..., N\}$ and $j \in \{1, ..., M\}$. 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 i-th entity in the j-th attribute, whose true value (i.e., *truth*, or *ground truth*) is denoted as T^*_{ij}.

Table I 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. A worker u will submit an answer a^u_{ij}, if cell c_{ij} is assigned to u.

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}.

Definition 3 (Truth Table): Given the set of answers $\{a^u_{ij}\}$, by workers u to cells c_{ij}, $i \in \{1, ..., N\}$, $j \in \{1, ..., M\}$, our target is to obtain the truth table including all the accurate estimates $\widehat{T}_{ij}$ for each cell c_{ij}'s true value T^*_{ij}.

IV. Approach

This section explains how T-Crowd obtains the truth table. The quality of 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 $\widehat{T}_{ij}$ (Section IV-A). 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 IV-B).

A. Worker Model

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 $L = \{l_1, l_2, ..., l_{|L|}\}$. An answer from a worker is either correct or wrong depending on whether it is the same as the ground truth. 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: $a^u_{ij} \sim \mathcal{N}(\widehat{T}_{ij}, \phi_u)$:

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

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 $\widehat{T}_{ij}$:

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

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 [2].

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

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

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 Equation 2, which means that we can use the same quality measure for categorical and continuous attributes.

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 Equation 2, we represent the quality of worker u answering cell c_{ij} as

$q_{ij}^u = \text{erf}\left(\epsilon/\sqrt{2\alpha_i\beta_j\phi_u}\right)$. Equations 1 and 3 can be changed accordingly, i.e., by replacing ϕ_u with ϕ_{ij}^u and q_u with q_{ij}^u.

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.

B. 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, 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}.

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

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 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

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

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

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

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

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

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:

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

Table II: Statistics of Real-world Datasets.

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

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

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 Prior(T_{ij}). At each iteration, the M-step applies gradient descent to find α_i, β_j and ϕ_u by maximizing Equation 5 and the E-step applies Equation 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:

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

V. Experiments

We use three real datasets to perform our experiments. Their statistics are shown in Table II. For dataset Celebrity [3], workers are given the picture of a celebrity and they are requested to provide values for categorical (name, nationality, ethnicity) and continuous attributes (age, height, notability, sentiment). For dataset Restaurant [12], workers are shown restaurant reviews and they are asked to specify values for review aspect (e.g., food or location), review attribute (e.g., price or style), and review sentiment (e.g., negative or positive). The start and end position of the first occurrence of the restaurant's name in the review is also crowdsourced. For dataset Emotion [13], workers are given a small piece of text and they give a number in [0,100] for any of the following six emotions: anger, disgust, fear, joy, sadness, and surprise, and a single numeric rating in the interval [-100,100] for their overall (positive or negative) sentiment about the text.

For Celebrity and Restaurant datasets, we collected the workers' answers using AMT [1]. Each task in Celebrity and Restaurant is answered 5 and 4 times, 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 [13]; each task is answered 10 times.

A. Comparison to Previous Work

We compare T-Crowd to previous work on truth inference in crowdsourcing. Based on the guidance from [7] [17], we select Majority Voting (MV), D&S [4] and GLAD [14] to compare with T-crowd for tables including only categorical attributes. For tables including only continuous attributes, we also compare to a baseline Median approach which uses the median of workers' answers as the estimated true value and GTM [16]. For tables including both continuous and categorical attributes we compare to CRH [9].

Table III: Effectiveness of Truth Inference.

	Celebrity		Restaurant		Emotion
Method	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
Maj. Voting	0.0573	/	0.2003	/	/
EM	0.0620	/	0.2463	/	/
GLAD	0.0498	/	0.1905	/	/
Median	/	0.6998	/	0.6784	0.7026
GTM	/	0.6516	/	0.5871	0.6792

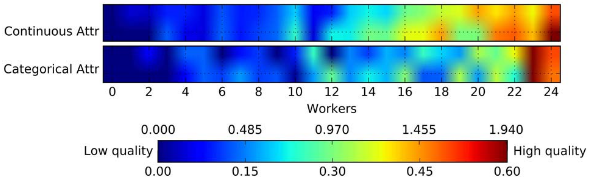

Figure 1: Uniform Worker Quality.

We compare the tested methods on categorical attribute inference by their Error Rate (percentage of mismatched values predicted truth and the ground truth). For continuous attributes we use the root of mean squared distance (RMSE) between each method's estimated truth and the ground truth. Since attributes have different scales, we normalize each attribute's RMSE by its own standard deviation and average them, which results to a mean normalized absolute distance (MNAD).

Table III summarizes the effectiveness of truth inference by all methods in terms of Error Rate and MNAD on the three real-world datasets. 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% on categorical data and the MNAD by 2.7% on continuous data compared to the best result of other methods. The corresponding reductions on Restaurant are 2.6% and 4%. On Emotion, we outperform previous work by 10%. CRH does not have stable performance as it is effective on Celebrity and Restaurant, but ineffective on Emotion. Overall, our method is more robust than them.

B. Case Study

We performed a case study on the Restaurant dataset that demonstrates the effectiveness of T-Crowd. In Figure 1, 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.

VI. Conclusions

In this paper we design a unified crowdsourcing framework for collecting multi-type tabular data. Our experiments on three datasets confirm the superiority of T-Crowd compared to the state-of-the-art.

Acknowledgement. We would like to thank the reviewers for the insightful comments. Guoliang Li was supported by the 973 program of China (2015CB358700), NSF China (61632016, 61472198, 61521002, 61661166012) and TAL education. Reynold Cheng, Caihua Shan, Zhipeng Huang, and Yudian Zheng were supported by the Research Grants Council of Hong Kong (RGC Projects HKU 17229116 and 17205115) and the University of Hong Kong (Projects 104004572, 102009508 and 104004129).

References

[1] Amazon Mechanical Turk. https://www.mturk.com/mturk/.

[2] ERF Function. http://mathworld.wolfram.com/Erf.html.

[3] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *ECCV*, pages 768–783, 2014.

[4] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28, 1979.

[5] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *PVLDB*, 2:550–561, 2009.

[6] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In *SIGMOD*, pages 61–72. ACM, 2011.

[7] G. Li, J. Wang, Y. Zheng, and M. J. Franklin. Crowdsourced data management: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 28(9):2296–2319, 2016.

[8] 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. *PVLDB*, 8(4):425–436, 2014.

[9] 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 *SIGMOD*, pages 1187–1198. ACM, 2014.

[10] H. Park, H. Garcia-Molina, R. Pang, N. Polyzotis, A. Parameswaran, and J. Widom. Deco: A system for declarative crowdsourcing. *PVLDB*, 5(12):1990–1993, 2012.

[11] H. Park and J. Widom. Crowdfill: Collecting structured data from the crowd. In *SIGMOD*, pages 577–588. ACM, 2014.

[12] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *SemEval*, pages 27–35, 2014.

[13] 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 *EMNLP*, pages 254–263. Association for Computational Linguistics, 2008.

[14] 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 *NIPS*, pages 2035–2043, 2009.

[15] X. Yin, J. Han, and S. Y. Philip. Truth discovery with multiple conflicting information providers on the web. *TKDE*, 20(6):796–808, 2008.

[16] B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. *Proc. of QDB*, 2012.

[17] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng. Truth inference in crowdsourcing: is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552, 2017.