FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation

Fenglong Ma[1], Yaliang Li[1], Qi Li[1], Minghui Qiu[2], Jing Gao[1], Shi Zhi[3]
Lu Su[1], Bo Zhao[4], Heng Ji[5], and Jiawei Han[3]
[1]SUNY Buffalo, USA [2]Singapore Management University, Singapore
[3]University of Illinois, Urbana, USA [4]LinkedIn, USA [5]Rensselaer Polytechnic Institute, USA
{fenglong,yaliangl,qli22,jing,lusu}@buffalo.edu, minghuiqiu@gmail.com
shizhi2@illinois.edu, bo.zhao.uiuc@gmail.com, jih@rpi.edu, hanj@illinois.edu

ABSTRACT

In crowdsourced data aggregation task, there exist conflicts in the answers provided by large numbers of sources on the same set of questions. The most important challenge for this task is to estimate source reliability and select answers that are provided by high-quality sources. Existing work solves this problem by simultaneously estimating sources' reliability and inferring questions' true answers (i.e., the truths). However, these methods assume that a source has the same reliability degree on all the questions, but ignore the fact that sources' reliability may vary significantly among different topics. To capture various expertise levels on different topics, we propose FaitCrowd, a fine grained truth discovery model for the task of aggregating conflicting data collected from multiple users/sources. FaitCrowd jointly models the process of generating question content and sources' provided answers in a probabilistic model to estimate both topical expertise and true answers simultaneously. This leads to a more precise estimation of source reliability. Therefore, FaitCrowd demonstrates better ability to obtain true answers for the questions compared with existing approaches. Experimental results on two real-world datasets show that FaitCrowd can significantly reduce the error rate of aggregation compared with the state-of-the-art multi-source aggregation approaches due to its ability of learning topical expertise from question content and collected answers.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

Truth Discovery; Source Reliability; Crowdsourcing

1. INTRODUCTION

Crowdsourcing becomes increasingly popular in recent decades, as people believe that the wisdom of the crowd can be superior to the judgements of individuals. Moreover, the development of crowdsourcing platforms, such as Amazon Mechanical Turk[1] and CrowdFlower[2], makes it more convenient to get crowdsourced data in a cheaper price. However, as the normal workers in crowdsourcing tasks are non-experts, errors are inevitable. As a result, conflicting information may be given to the same question. To obtain the final answers, one of the most important issues is how to aggregate the crowdsourced data from multiple sources so that the most trustworthy information (i.e., the truths) can be detected.

To discover the truths from conflicting data, the most intuitive approach is majority voting, which selects the majority answers from all sources as the final output. However, this approach fails to take the reliability levels of different sources into consideration, which may lead to poor performance when the number of low quality sources is large. To solve this problem, techniques for multi-source aggregation, which consider the estimation of source reliability, have been proposed to derive true answers from a collection of sources [2, 3, 4, 5, 6, 10, 11, 12, 14, 17, 18, 19, 22, 26, 28]. Despite the difference in their models, the same principle applies: The more reliable a source is, the more likely this source would provide trustworthy information, and vice versa. Based on this principle, the existing methods are trying to assign larger weights to reliable sources such that they can play a more important role when inferring the truths.

However, a common drawback of those approaches is that only one reliability degree is estimated for each source, which may not properly reflect the variation in reliability among topics in the real world. In fact, no one could be an expert in every field, and source expertise[3] usually vary among different topics. For example, *Albert Einstein* is a guru on physics but not on drawing. Therefore, it is crucial to estimate fine grained source reliability in multi-source aggregation.

Intuitively, we can directly employ topic models on question content to divide questions into topical-level groups. Then, according to source answering behavior, the aforementioned methods in multi-source aggregation are applied to estimate topical expertise for sources on each topical-level group. However, this naive approach reduces the number of answers dramatically on each topic, which may lead to an incorrect estimation of source expertise due to the fact that data is insufficient. Hence, the performance on each topic may drop, as a result, the overall performance on all the topics would drop significantly.

To tackle the aforementioned challenges, in this paper, we propose **F**ine Gr**ai**ned **T**ruth Discovery model for **Crowd**sourced data

[1]https://www.mturk.com/

[2]http://www.crowdflower.com

[3]Note that the term "expertise" and "reliability" are used interchangeably in this paper.

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2783314.

aggregation (*FaitCrowd*), which can automatically assign topics to questions, estimate topic-specific expertise for each source, and learn true answers simultaneously. To the best of our knowledge, we are the first to propose such an unsupervised probabilistic model to learn fine grained source reliability for multi-source aggregation. One important feature of FaitCrowd is the employment of latent topics, which allows us to define a distribution on source expertise for each topic. The proposed method jointly models question content and source answering behavior to learn latent topics and estimate the topical source expertise. Therefore, the proposed model can simultaneously learn topics, source expertise and true answers. We jointly sample topics and the estimated truths using Gibbs sampling and learn source expertise based on each topic using gradient descent. Compared with existing methods in multi-source data aggregation, the benefit of the proposed FaitCrowd is its ability to infer different expertise based on topics and adjust source reliability via both question content and sources' answering behavior.

The advantage of applying the proposed FaitCrowd to aggregate crowdsourced data is threefold: First, FaitCrowd can automatically learn source expertise on different topics. For crowdsourcing applications, when posting similar tasks in the future, requester can select high-quality workers on each topic, which will improve data quality and budget efficiency. Second, FaitCrowd can handle difficult tasks better using the estimated topical expertise of sources. Because FaitCrowd is capable of assigning higher topical expertise to sources who often provide correct answers on the topic, the true answers of hard questions can be correctly determined by sources with higher topical expertise. Finally, FaitCrowd can find the minority in the crowd who are truly knowledgeable in a given field.

The experiments on real world datasets show that the proposed FaitCrowd can significantly reduce the error rate comparing with the state-of-the-art approaches in multi-source aggregation. The proposed model can correctly estimate topical expertise for each source. Experiments are conducted to validate the advantage of combining topic modeling techniques and data aggregation methods. Meanwhile, we illustrate that the learned source expertise is consistent with ground truth. In summary, our main contributions are as follows:

- We recognize the difference in source reliability among topics on the truth discovery task and propose to incorporate the estimation of fine grained reliability into truth discovery.
- We propose a probabilistic model that simultaneously learns the topic-specific expertise for each source, aggregates true answers, and assigns topic labels to questions.
- We empirically show that the proposed model outperforms existing methods in multi-source aggregation with two real world datasets.

2. PROBLEM FORMULATION

In this section, we first introduce some basic terminologies used in this paper and then formally define our problem.

Input

The inputs of the proposed model are questions $\{q\}_1^Q$, sources $\{u\}_1^U$ and answers $\{a_{qu}\}_{q=1,u=1}^{Q,U}$, where Q is the number of questions and U is the number of sources.

Definition 1. A question q contains M_q words $\{w_{qm}\}_{m=1}^{M_q}$ and can be answered by sources.

Definition 2. a_{qu} denotes the answer given by source u to question q.

Output

The goal is to derive true answers $\{t_q\}_{q=1}^Q$, topical expertise e, and topic labels for each question $\{z_q\}_{q=1}^Q$.

Definition 3. The estimated truth t_q[4] for question q is the most trustworthy answer provided by sources.

Definition 4. Topical expertise $e \in \mathbb{R}^{K\times U}$ is referred to as the level of reliability for sources on K topics.

Definition 5. A topic indicator z_q is the topic label of question q.

Based on these definitions, we can formally define our problem as follows:

Problem 1. Given a question set $\{q\}_1^Q$, source set $\{u\}_1^U$, answer set $\{a_{qu}\}_{q=1,u=1}^{Q,U}$, and the number of topics K , our goal is to learn topic-specific expertise e for all sources on each topic k, the true answers $T=\{t_q\}_{q=1}^Q$ and questions' topic labels $\{z_q\}_{q=1}^Q$.

3. FAITCROWD MODEL

The basic idea of the proposed model is to build a joint probabilistic model, which contains two integral components: (1) the modeling of question content, and (2) the modeling of answers given by sources. We first summarize the proposed joint model, describe the generative process, and finally demonstrate the two integral components of the proposed model in detail.

3.1 Model Overview

In contrast to existing methods in multi-source aggregation, we jointly model question content and source answering behavior using latent topics. The advantage of the proposed joint model is that modeling question content can help estimate reasonable source reliability, and in turn, modeling answers leads to the discovery of meaningful topics. In other words, the two integral components simultaneously help each other.

Figure 1 shows the proposed fine grained truth discovery model for crowdsourced data aggregation. The inputs are Q questions, K topics, M_q words $\{w_{qm}\}_{m=1}^{M_q}$ for each question q, and N_q answers $\{a_{qu}\}_{u=1}^{N_q}$ provided by sources to question q. The shaded circles represent hyper-parameters[5] except w_{qm}, a_{qu} and u, which are inputs. The outputs are source expertise e, estimated true answers $\{t_q\}_{q=1}^Q$ and topic labels $\{z_q\}_{q=1}^Q$. The remaining ones, φ, y, ϕ', ϕ, θ and b_q, are the intermediate variables learned by the proposed model.

The generative process of the proposed model is as follows:

- Draw $\theta \sim Dir(\alpha)$, $\phi' \sim Dir(\beta')$,$\varphi \sim Dir(\eta)$
- For the k-th topic ($k = 1, 2, \cdots, K$)
  - Draw a word distribution on topic k, $\phi_k \sim Dir(\beta)$
  - For the u-th source ($u = 1, 2, \cdots, U$)
    - Draw source-topic specific expertise, $e_{ku} \sim N(\mu, \sigma^2)$

[4]Note that t_q is learned by the proposed model as the estimated truth instead of the real answer of question q. Real true answers are only used in evaluation.

[5]η, β', β and α are hyper-parameters of Dirichlet distribution, μ denotes the mean of Gaussian distribution, σ^2 and $\sigma^{2'}$ are variances of Gaussian distribution, and γ_q is the parameter of Uniform distribution.

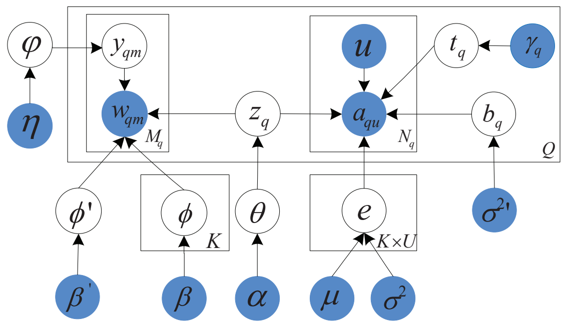

Figure 1: Fine Grained Truth Discovery Model for Crowd-sourced Data Aggregation.

- For the q-th question ($q = 1, 2, \cdots, Q$)
  - Draw a topic $z_q \sim Multi(\theta)$
  - For the m-th word ($n = 1, 2, \cdots, M_q$)
    - Draw a word category $y_{qm} \sim Bernoulli(\varphi)$
    - Draw a word $w_{qm} \sim Multi(\phi_{z_q})$ if $y_{qm} = 1$, otherwise draw $w_{qm} \sim Multi(\phi')$
  - Draw a true answer $t_q \sim U(\gamma_q)$
  - Draw a bias $b_q \sim N(0, \sigma^{2\prime})$
  - For the u-th source ($u = 1, 2, \cdots, N_q$)
    - Draw an answer $a_{qu}|t_q \sim logistic(e_{z_q u}, b_q)$

Given a topic distribution θ on a dataset, we can draw a topic z_q from Multinomial distribution θ for each question q. Next, we introduce the following generative processes, including word generation and answer generation which are all based on topic $z_q = k$. **Word generation.** We assume that topic-specific words on each topic k have a distribution ϕ_k and background words have a distribution ϕ'. There is a switch y drawn from Bernoulli distribution with parameter φ to select words' distribution. If $y_{qm} = 1$, then the word w_{qm} is drawn from topical word distribution ϕ_k; otherwise, it is drawn from background word distribution ϕ'. Based on the above assumption, we can generate words based on topic z_q. **Answer generation.** We assume that a source's answer on a question is associated with the source's expertise and the question's bias. We use a logistic function to model the answer provided by a source. According to the drawn topic z_q, the expertise of the source u who provided an answer to question q can be matched to $e_{z_q u}$. Moreover, we use t_q to denote the estimated true answer to question q. Finally, we use source u's expertise $e_{z_q u}$ on topic z_q, question bias b_q, and the estimated truth t_q to model the probability of a_{qu} using the logistic function.

The proposed model alternates between modeling question component and modeling answer component to learn the source's topical expertise and to estimate the true answers. The detailed generative processes of the two integral components are introduced in the following subsections.

3.2 Modeling Question Content

For modeling question content, we first draw the corpus topic distribution θ, the parameter of Bernoulli distribution φ, background word distribution ϕ' and topical word distribution $(\phi_k)_{k=1}^K$. Because the length of each question is short, we follow the idea of Twitter-LDA [30] for word generation. We assume that each question is about a single topic. Then, we draw a topic indicator z_q to question q. Let $n^w_{q,y=1}$ be the frequency of w (i.e., w_{qm}) as topical words in question q, $n^w_{q,y=0}$ be the frequency of w as background words, θ_k be the probability of question q on topic k, and $\phi_{kw} = p(w|k)$ be the probability of topical word w generated by topic k. Then the probability of topical word w appearing $n^w_{q,y=1}$ times in question q is defined as $(\theta_k \phi_{kw})^{n^w_{q,y=1}}$, and the probability of background word w is $(\phi'_w)^{n^w_{q,y=0}}$. We assume words are independent, and the probability of all the words in question q under topic k and word category y is

$$p(\boldsymbol{w}_q|k, y) = \prod_{w=1}^{V} (\phi'_w)^{n^w_{q,y=0}} (\theta_k \phi_{kw})^{n^w_{q,y=1}},$$

where V is the number of all the unique words in corpus and $M_q = \sum_{w=1}^{V} (n^w_{q,y=1} + n^w_{q,y=0})$. We also assume questions are independent, and the probability of observing the question set $\{q\}_1^Q$ is:

$$p(\boldsymbol{w}|\theta, \phi, \phi') = \prod_{q=1}^{Q} \prod_{w=1}^{V} (\phi'_w)^{n^w_{q,y=0}} \left(\sum_{k=1}^{K} (\theta_k \phi_{kw})^{n^w_{q,y=1}} \right) \quad (1)$$

3.3 Modeling Answers

Intuitively, most sources have the ability to provide correct answers for most questions, yet only a few sources are gurus or novices on the topic. Thus, we assume that sources' expertise is drawn from a Gaussian distribution for each topic, i.e., $e_{ku} \sim N(\mu, \sigma^2)$. The value of e_{ku} is from $-\infty$ to $+\infty$. For each answer provided by source u to question q, it depends on several factors: (1) The topic of the question. Since different sources are familiar with different topics, the question's topic influences each source's answer. (2) The expertise of the source on this topic. If source u is very skilled on this topic, u may give a correct answer to question q. (3) The number of correct answers provided by the source on the topic. If source u provides many correct answers on this topic, u may be an expert of the topic. (4) The bias on this question. A lower bias means that the question is easy. Then every source is more likely to give a correct answer.

Based on the above analysis, we give the process of generating source u's answer a_{qu} for question q and assume that there are γ_q different choices $\{1, ..., \gamma_q\}$ for each question q. We draw a true answer t_q from a Uniform distribution $U(\gamma_q)$, a topic indictor $z_q = k$ from Multinomial distribution $Multi(\theta)$, and the question's bias b_q from Gaussian distribution $N(0, \sigma^{2\prime})$[6]. Using the logistic function, given the topic $z_q = k$, the correct answer given by source u to question q is denoted as a_{qu} which is generated as follows:

$$p\left(a_{qu} = c | t_q = c, z_q, \rho_{z_q u}, e_{z_q u}, b_q\right) = \omega(-\rho_{z_q u} e_{z_q u} + b_q) \quad (2)$$

where $\omega(-\rho_{z_q u} e_{z_q u} + b_q) = \frac{1}{1+\exp\left(-\rho_{z_q u} e_{z_q u} + b_q\right)}$, $\rho_{z_q u}$ is the estimated contribution ratio[7] of source u on topic z_q, and b_q is a bias on question q. From Eq.(2), we can see that as the topical expertise and the contribution ratio of source u increase and the bias of the question q decreases (i.e., a more knowledge user answers a easier question), the probability that u's answer to q is the final true answer increases. In contrast, when the expertise and the contribution ratio of source u decrease and the bias b_q increases, the probability drops.

[6]Because the difficulty of most questions is moderate and only a small part of questions are very easy or hard, we use a Gaussian distribution on biases.

[7]The estimated contribution ratio ρ_{ku} is equal to the number of correct answers provided by source u on topic k divided by the number of questions on this topic.

Here we consider the "one-coin model", i.e., for all $c' \neq c$,

$$p\left(a_{qu} = c' | t_q = c, z_q, \rho_{z_q u}, e_{z_q u}, b_q\right) = \frac{1 - \omega\left(-\rho_{z_q u} e_{z_q u} + b_q\right)}{\gamma_q - 1} \tag{3}$$

Combining Eq.(2) and Eq.(3), the probability of a_{qu} is:

$$\begin{aligned} & p\left(a_{qu} | t_q = c, z_q, \rho_{z_q u}, e_{z_q u}, b_q\right) \\ =& \omega(-\rho_{z_q u} e_{z_q u} + b_q)^{\delta\left(a_{qu}, c\right)} \left(\frac{1 - \omega(-\rho_{z_q u} e_{z_q u} + b_q)}{\gamma_q - 1}\right)^{1-\delta\left(a_{qu}, c\right)} \end{aligned}$$

where $\delta(x, y)$ is the Kronecker delta function.

Given the topic z_q, the joint probability of a_{qu}, t_q, $\rho_{z_q u}$, $e_{z_q u}$ and b_q is:

$$\begin{aligned} & p(a_{qu}, t_q = c, \rho_{z_q u}, e_{z_q u}, b_q | z_q, \mu, \sigma^2, \sigma^{2'}, \gamma_q) \\ =& p(e_{z_q u} | \mu, \sigma^2) p(b_q | \sigma^{2'}) p(t_q = c | \gamma_q) \\ & p\left(a_{qu} | t_q = c, z_q, \rho_{z_q u}, e_{z_q u}, b_q\right) \end{aligned} \tag{4}$$

For all the observed answers $A = \{a_{qu}\}_{q=1,u=1}^{Q,U}$, the probability is:

$$p\left(A | T, e, b\right) = \prod_{q=1}^{Q} \prod_{u=1}^{U} \prod_{c=1}^{\gamma_q} p\left(a_{qu} | t_q = c, z_q, \rho_{z_q u}, e_{z_q u}, b_q\right) \tag{5}$$

4. INFERENCE AND LEARNING

In this section, we present the objective function of the proposed model and discuss how to infer parameters using Gibbs-EM [21].

4.1 Objective Function

The objective of the proposed model is to learn the hidden topics, sources' topical expertise and questions' true answers based on jointly modeling question content and answers. Hence, the objective function builds on Eq.(1) and Eq.(5). More precisely, it is the negative log posterior of $\boldsymbol{w}$ and A shown as follows:

$$J = -\log p(\boldsymbol{w} | \alpha, \beta, \beta', \eta) - \log p\left(A | \mu, \sigma^2, \gamma, \sigma^{2'}\right) \tag{6}$$

where the first term denotes the likelihood of generating the question content, and the latter denotes the likelihood of generating answers.

It is intractable to perform exact inference on the posterior distribution of all the hidden variables. Therefore, we employ a hybrid inference method combining sampling and variational optimization, named Gibbs-EM [21] which is an inference method alternating between Gibbs sampling and gradient descent. We employ Gibbs sampling to learn the hidden variables by fixing the values of ρ_{ku}, e_{ku} and b_q, and we use gradient descent to learn hidden factors.

4.2 Hidden Variable Inference

We perform Gibbs sampling to learn the hidden variables z_q and t_q by fixing the values of e and b updated in the gradient descent step. Dirichlet-Multinomial conjugacy allows Gibbs sampling to work by sampling on the topic indicator z_q, collapsing out ϕ and ϕ'. Since it is a conventional step, we omit the detailed derivations and present the derived Gibbs sampling update rules. Interested readers are referred to [7] for details.

When sampling a topic z_q, two independent parts, i.e., question content part and answer part, are considered. For the estimated true answers, we only take the answer part into consideration. We jointly sample z_q and t_q as follows:

$$\begin{aligned} & p(z_q = k, t_q = c | z_{\neg q}, \rho_{z_q u}, e_{z_q u}, b_q, \alpha, \beta, \gamma_q) \\ & \propto (n_{\neg q}^{k} + \alpha) \cdot \frac{\prod_{w=1}^{V} \prod_{i=0}^{n_{q,y=1}^{w}-1} (n_{\neg q,k,y=1}^{w} + \beta + i)}{\prod_{j=0}^{n_{q,y=1}^{w}-1} (\sum_{w=1}^{V} n_{\neg q,k,y=1}^{w} + V\beta + j)} \\ & \cdot \prod_{u=1}^{N_q} p(a_{qu} | c, k, \rho_{ku}, e_{ku}, b_q) \cdot p(c | \gamma_q), \end{aligned} \tag{7}$$

where $n_{\neg q}^{k}$ denotes the number of times that topic k is sampled in the question set without considering the current question q, and $n_{\neg q,k,y=1}^{w}$ denotes the number of times that w is sampled as a topic-specific word in topic k without considering the current word assignment.

4.3 Parameter Estimation

Though we fix z_q and t_q at this step, it is difficult to directly calculate $e_{z_q u}$ and b_q by maximizing the probability of posterior distribution. Therefore, we employ gradient descent to learn $e_{z_q u}$ and b_q. Based on Eq.(4) and Eq.(6), the objective is modified as

$$\begin{aligned} J_{qu} &= -\log p(a_{qu}, t_q, z_q, \rho_{z_q u}, e_{z_q u}, b_q | \mu, \sigma^2, \sigma^{2'}, \gamma_q) \\ &= -\log p(a_{qu} | t_q, z_q, \rho_{z_q u}, e_{z_q u}, b_q) - \log p(t_q | \gamma_q) \\ & - \log p(e_{z_q u} | \mu, \sigma^2) - \log p(b_q | \sigma^{2'}) \\ & \propto -\log p(a_{qu} | t_q, z_q, \rho_{z_q u}, e_{z_q u}, b_q) + \frac{(e_{z_q u} - \mu)^2}{2\sigma^2} + \frac{b_q^2}{2\sigma^{2'}} \end{aligned}$$

Then we can differentiate J_{qu} to obtain its gradients:

$$\frac{\partial J_{qu}}{\partial e_{z_q u}} = \rho_{z_q u}(\delta\left(a_{qu}, c\right) - \omega(-\rho_{z_q u} e_{z_q u} + b_q)) + \frac{e_{z_q u} - \mu}{\sigma^2}$$

$$\frac{\partial J_{qu}}{\partial b_q} = -\omega(-\rho_{z_q u} e_{z_q u} + b_q) + \delta\left(a_{qu}, c\right) + \frac{b_q}{\sigma^{2'}}$$

Then gradient descent method is used to update $e_{z_q u}$ and b_q based on the gradients:

$$e_{z_q u}^{new} := e_{z_q u}^{old} - \lambda \frac{\partial J_{qu}}{\partial e_{z_q u}} \tag{8}$$

$$b_q^{new} := b_q^{old} - \lambda \frac{\partial J_{qu}}{\partial b_q} \tag{9}$$

We can derive intermediate parameters and make the following parameter estimations:

$$\theta_k = \frac{n^k + \alpha}{\sum_{k'=1}^{K} n^{k'} + K\alpha} \tag{10}$$

$$\rho_{ku} = \frac{n_u^k}{n^k} \tag{11}$$

$$\phi_{kw} = \frac{n_{k,y=1}^{w} + \beta}{\sum_{w'=1}^{V} n_{k,y=1}^{w'} + V\beta} \tag{12}$$

$$\phi'_w = \frac{n_{y=0}^{w} + \beta'}{\sum_{w'=1}^{V} n_{y=0}^{w'} + V\beta'} \tag{13}$$

where n^k is the number of times topic k is sampled, n_u^k is the number of times source u provides (estimated) correct answers on topic k, $n_{k,y=1}^{w}$ is the number of times word w sampled as a topical word specific to topic k, and $n_{y=0}^{w}$ is the number of times word w sampled as background words.

4.4 Algorithm Flow

The model inference and parameter learning process are described in Algorithm 1. We first jointly sample a pair of z_q and t_q, i.e., assign a topic and select an answer as the truth to question q, by fixing source expertise e and bias b. Then, fixing z_q and t_q, we update $e_{z_q u}$ according to Eq.(8) and b_q according to Eq.(9). Finally, we estimate ρ_{ku}, θ_k, ϕ_{kw} and ϕ'_w.

Algorithm 1 FaitCrowd Learning Algorithm.

Input: Question set $\{q\}_1^Q$; Source set $\{u\}_1^U$; Answers $\{a_{qu}\}_{q=1,u=1}^{Q,U}$; Topic number K; Parameters: η, α, β, β', μ, σ^2, $\sigma^{2'}$, λ

while not convergence **do**
  for the q-th question ($q = 1, 2, \cdots, Q$)
    Joint sample (z_q, t_q) according to Eq.(7);
    for the u-th source ($u = 1, 2, \cdots, N_q$)
      Update $e_{z_q u}$ according to Eq.(8);
      Update b_q according to Eq.(9);
    end for
  end for
  for the k-th topic ($k = 1, 2, \cdots, K$)
    Update θ_k according to Eq.(10);
    for the u-th source ($u = 1, 2, \cdots, U$)
      Update ρ_{ku} accprding to Eq.(11);
    end for
    for the w-th word ($w = 1, 2, \cdots, V$)
      Update ϕ_{kw} according to Eq.(12);
    end for
  end for
  for the w-th word ($w = 1, 2, \cdots, V$)
    Update ϕ'_w according to Eq.(13);
  end for
end while

Output: Source expertise $\boldsymbol{e}$; True answers $\{t_q\}_{q=1}^Q$; Question topic labels $\{z_q\}_{q=1}^Q$ ($z_q \in (1, \cdots, K)$).

Algorithm 1 shows that FaitCrowd needs $O(QN_q + KU + KV + V)$, which is dominated by $O(QN_q)$, where QN_q is the number of answers. Therefore, FaitCrowd has linear running time.

5. EXPERIMENTS

In this section, we first describe the two real world datasets in Section 5.1 and introduce baselines and parameter settings in Section 5.2. In Section 5.3, the results of experiments show that the proposed method can significantly reduce the error rate compared with the state-of-the-art approaches in multi-source aggregation. We test the proposed FaitCrowd method against conducting topic modeling and true answer inference to show the importance of integrating question content and answers. In Section 5.4, the correctness of topical expertise is analyzed using ranking methods, and some examples are given to demonstrate that the topic expertise learned by the proposed model is reasonable. Finally, we analyze parameters' sensitivity in Section 5.5. The proposed method shows the power of learning source topical expertise accurately and reducing the error rate dramatically.

5.1 Data Description

5.1.1 *The Game Dataset*

The Game dataset [1] is collected from a crowdsourcing platform via an Android App based on a TV game show "Who Wants to Be a Millionaire". Here each user is a source. Users receive each question's content and its four corresponding candidate answers via the Android App. Then they can provide answers which would be collected by the App. For each question, the game show provides the correct answer, as well as its difficulty level drawn from 1 to 10. Level 1 questions are the easiest and Level 10 means extremely difficult. Note that correct answers and difficulty levels are not used by the proposed approach and baselines. They are only used for evaluation. The Game dataset contains 2,103 questions, 37,029 sources, 214,849 answers and 12,995 unique words.

5.1.2 *The SFV Dataset*

The SFV dataset [8] is extracted from Slot Filling Validation (SFV) task of the NITS Text Analysis Conference Knowledge Base Population (TAC-KBP) track. The SFV task aims at collecting "slot fillers" (answers) from a large-scale multi-source corpus for certain attributes of a query entity, such as a person or an organization. Assuming "Albert Einstein" is a query entity and the birthday is an attribute, the task of extracting Albert Einstein's birthday is submitted to 18 different information extraction systems. Next, 18 systems return the answers and provide sentences to support the answers. We can aggregate answers from systems' returns. This is indeed a crowdsourced data aggregation task, i.e. aggregating conflicting answers to obtain the estimate truths. TAC-KBP provides ground truth data corresponding to each query entity.

The sentence set for each pair of query entity and attribute returned by different systems is defined as the question, and a system is regarded as a source. For each question, answers from different sources may have conflicts among them.

We use KBP 2013 dataset. Since systems can resubmit their answers, we only select answers that systems submitted at the first time. The dataset contains 328 questions, 18 sources, 2,538 answers and 5,587 unique words.

5.2 Experiment Setup

We compare the proposed FaitCrowd model against several existing unsupervised algorithms commonly employed in multi-source aggregation. A naive baseline is **MV** (majority voting), which estimates true answers as the ones given by the majority of the sources. This approach regards all the sources equally in true answer estimation. We also compare the proposed method with some state-of-the-art methods that estimate source reliability, including: **TruthFinder** [26], **AccuPr** [4], **Investment** [14], **3-Estimates** [6], **CRH** [11], **CATD** [10], **D&S** [2] and **ZenCrowd** [3]. Details of these methods are discussed in related work.

We further compare two variants of FaitCrowd to show the benefit of considering biases and background words. **Fait-Crowd-b** is a variant of FaitCrowd without taking bias information into consideration. **FaitCrowd-g-b** is based on FaitCrowd-b by further removing the modeling of background words. Comparison with these two baselines can show that: (1) Question's bias is important. The bias captures the difficulty of each question. If the question is easy, any source can provide a correct answer. Thus, this will affect source expertise. (2) The number of background words is larger than the number of topical words for each question. Removing the modeling of background words will affect the accuracy of topic modeling thereby increasing the error rate of estimating true answers.

We perform 200 runs of Gibbs-EM and use grid search to select the number of topics K for the two datasets: 12 for the Game dataset and 8 for the SFV dataset. For question content modeling part, we set $\eta = 20$, $\beta' = \beta = 0.01$ and $\alpha = 50/K$. For answer modeling, we set Gaussian priors to e_{ku} with mean μ as 45 and 35, variances σ^2 as 70 and 30 for the Game and SFV dataset respectively, and set the variance $\sigma^{2'} = 50$ of biases b and the learning rate $\lambda = 0.01$. We also conduct experiments to evaluate the performance of FaitCrowd using different settings for μ and σ^2.

5.3 Performance Validation

The experimental results show that the proposed method can significantly reduce the error rate compared with baselines, perform well on difficult questions, and find knowledgeable sources even if their answers are minority. The comparison between separate models (conducting topic modeling and true answer estimation separately) and FaitCrowd show that FaitCrowd is more effective on estimating true answers by jointly modeling questions and answers.

5.3.1 Performance Metric

To evaluate the performance of each method, *Error Rate* is used as evaluation metric, which is defined as the number of incorrectly answered questions divided by the total number of questions Q. A lower error rate means that the method's estimation is closer to the ground truth, and the method is better than those with higher error rates.

5.3.2 Results on the Game and SFV Datasets

Table 1 shows experimental results of the proposed FaitCrowd and baseline methods on the Game dataset. We list the number of questions in each difficulty level in the parentheses.

From Table 1, we can see that the proposed FaitCrowd is better than all the baselines in terms of *Error Rate*. The error rates of the proposed methods, including FaitCrowd, FaitCrowd-b and FaitCrowd-g-b, are lower than those of baselines' on all question levels, especially on more difficult questions. For easy questions (from Level 1 to Level 7), all the methods can estimate most answers correctly. Most baselines make mistakes on the same few hard questions, which leads to the ties among several methods as the best. However, the error rates increase dramatically for all baseline methods on difficult questions. The error rates of FaitCrowd on difficult questions (from Level 8 to Level 10) increases slightly, but the performance is much better than that of the baseline methods. For the most difficult level (Level 10), the error rate of the proposed FaitCrowd is 11.36%, while all the baseline methods have error rates over 20.45%. The reason is that majority answers provided by sources are usually wrong for difficult questions, and baselines cannot estimate correctly because their estimation of source reliability is not accurate. However, the proposed method can estimate topic expertise accurately.

Compared with FaitCrowd-b and FaitCrowd-g-b, FaitCrowd achieves a lower error rate by adding biases on questions and modeling background words. If we do not consider biases when modeling answers, source expertise will be wrongly estimated on difficult questions. Without taking background words information into account, the overall error rate increases further. That is because the length of each question is short but duplicate words exist among questions, which would affect the results of modeling topics as well as topical expertise of sources. Therefore, adding background words information and biases is reasonable.

Overall, the error rate of FaitCrowd reduces 17.73% compared with the best baseline method CATD. For TruthFinder, the error rate is larger than other methods'. That is because this method is dramatically affected by the large number of lower quality claims. On the Game dataset, lots of sources provide low quality answers and the number of conflicts is very high, which leads to the poor performance of TruthFinder. The error rate of Investment is larger than MV because Investment estimates the probability of each claim being correct given each source's reliability without considering complement vote. Other baseline methods are all better than MV but worse than FaitCrowd.

Table 2 presents the result comparison on the SFV dataset. Note that CATD method requires that the number of choices of each question must be equal, but the SFV dataset does not satisfy this requirement. Therefore, we did not compare with CATD on this dataset. The proposed method achieves the best performance comparing with all the baseline methods. Similar to what we observe on the Game dataset, Investment has a higher error rate. However, TruthFinder performs better than several baselines because the number of conflicts in the SFV dataset is much lower than that of the Game dataset. In contrast, the error rate of ZenCrowd is much higher than MV because the number of sources and answers in the SFV dataset is so small that there is not sufficient data for ZenCrowd to learn sources' confusion matrix.

Table 2: Comparison on the SFV dataset.

Method	Error Rate
FaitCrowd	**0.0610**
FaitCrowd-b	0.0671
FaitCrowd-g-b	0.0701
MV	0.1128
TruthFinder	0.0793
AccuPr	0.0701
Investment	0.2896
3-Estimates	0.1128
CRH	0.0854
D&S	0.1098
ZenCrowd	0.1555

5.3.3 Case Study

We use question 79 in the Game dataset as an example to illustrate how the proposed method achieves better results. It contains four choices – A, B, C and D. There are 25 sources voting A, 12 sources voting B, 4 sources voting C and 7 sources voting D. The correct answer is D. Obviously, majority voting cannot provide the correct answer. However, other baselines all provided the answer A as the correct answer. That is because these methods cannot learn the accurate expertise. Though the number of sources who provide A and B are much larger than D's, the proposed method still learns the correct answer because the expertise of those sources who give answer D are higher than others'. In this case, the correct answer is determined by the sources who are more knowledgeable on this question. Therefore, the benefit of the proposed method is to derive topic expertise accurately.

5.3.4 Model Validation

Here we illustrate the importance of joint modeling question content and answers by comparing with the method that conducts topic modeling and true answer inference separately.

Firstly, we use TwitterLDA [30] to learn K topics and divide the dataset into K sub-datasets according to the learned topic labels. Then, we run all the baseline methods for each topic. Finally, we collect all the estimated true answers to calculate the *Error Rate* for all questions. In order to validate the effectiveness of the proposed model, we conduct TwitterLDA on the two datasets using the same values of parameters in FaitCrowd model.

Table 3 shows the results of model validation on the Game dataset and the SFV dataset. We can see that baselines' performance is worse or similar compared to that of the same approaches applied on the whole dataset. Dividing the whole dataset into sub-topical

Table 1: Comparison on the Game dataset.

Method	Error Rate										
	L1 (303)	L2 (295)	L3 (290)	L4 (276)	L5 (253)	L6 (218)	L7 (187)	L8 (138)	L9 (99)	L10 (44)	Overall (2103)
FaitCrowd	**0.0132**	**0.0271**	**0.0241**	**0.0254**	**0.0395**	**0.0550**	**0.0481**	**0.0870**	**0.1010**	**0.1136**	**0.0399**
FaitCrowd-b	**0.0132**	**0.0271**	0.0276	0.0290	0.0553	0.0596	**0.0481**	0.0942	0.1111	0.1363	0.0447
FaitCrowd-g-b	**0.0132**	**0.0271**	**0.0241**	0.0290	0.0435	0.0688	0.0535	0.1304	0.1111	0.1818	0.0480
MV	0.0297	0.0305	0.0414	0.0507	0.0672	0.1101	0.1016	0.3043	0.3737	0.5227	0.0980
TruthFinder	0.0693	0.0915	0.1241	0.0942	0.1581	0.2294	0.2674	0.3913	0.5455	0.5455	0.1816
AccuPr	0.0264	0.0305	0.0345	0.0507	0.0632	0.0963	0.0909	0.2826	0.3636	0.5000	0.0913
Investment	0.0330	0.0407	0.0586	0.0761	0.0870	0.1239	0.1283	0.3406	0.3838	0.5455	0.1151
3-Estimates	0.0264	0.0305	0.0310	0.0507	0.0672	0.1055	0.0963	0.2971	0.3737	0.5000	0.0942
CRH	0.0264	**0.0271**	0.0345	0.0435	0.0593	0.0872	0.0856	0.2609	0.3535	0.4545	0.0866
CATD	**0.0132**	**0.0271**	0.0276	0.0290	0.0435	0.0596	**0.0481**	0.1304	0.1414	0.2045	0.0485
D&S	0.0297	0.0305	0.0483	0.0507	0.0672	0.1101	0.0963	0.2971	0.3636	0.5227	0.0975
ZenCrowd	0.0330	0.0305	0.0345	0.0471	0.0593	0.0872	0.0856	0.2754	0.3636	0.5227	0.0899

Table 3: Results of model validation.

Method	Error Rate	
	Game	SFV
FaitCrowd	**0.0399**	**0.0610**
MV	0.1013	0.1144
TruthFinder	0.2049	0.0762
AccuPr	0.1070	0.0678
Investment	0.2477	0.2896
3-Estimates	0.1116	0.1159
CRH	0.0856	0.0762
CATD	0.0504	-
D&S	0.1012	0.1153
ZenCrowd	0.0988	0.1283

datasets will reduce the number of responses per topic, which leads to insufficient data for baseline approaches. Therefore, these methods cannot correctly estimate source reliability of each topic. In contrast, the proposed method jointly conducts question content modeling part and answering modeling part. Therefore, it can learn true answers with sufficient data, and consequently performs better than baselines.

5.4 Topical Expertise Validation

The proposed method can learn reasonable source expertise based on meaningful topics. We employ two measures to validate the correctness of topical expertise learned by FaitCrowd. Experimental results show that the topical expertise learned by the proposed method highly correlates with the ground truth. We show some interesting examples to illustrate the diverse source expertise learned by FaitCrowd on different topics.

5.4.1 Performance Measures

We adopt two common measures, *Pearson*[8] and *Kendall*[9], to evaluate the topical expertise estimated by the proposed FaitCrowd. *Pearson* and *Kendall* are used to measure the correlation between two variables – one variable is topical expertise learned by FaitCrowd, and the other is the percentage of correct answers obtained from ground truth. The higher values of *Pearson* and *Kendall*, the better performance of the proposed method.

5.4.2 Correlations on the Game and SFV Datasets

Table 4 lists the *Pearson* and *Kendall* coefficients on the Game and SFV datasets. Overall, the average values of *Pearson* and *Kendall* on all the topics are 0.8661 and 0.7072 on the Game dataset, 0.9821 and 0.8989 on the SFV dataset respectively. Consequently, we can see that the topical expertise estimated by FaitCrowd correlates with the ground truth accuracy greatly. This suggests that the topical expertise can represent the reliability of sources on the topic and also the proposed FaitCrowd is reasonable and effective in learning topical expertise for sources.

Table 4: Correlations on the Game and SFV dataset.

Topic	Game		SFV	
	Pearson	Kendall	Pearson	Kendall
1	0.8989	0.7090	0.9818	0.8721
2	0.9030	0.7727	0.9861	0.9471
3	0.8766	0.7102	0.9762	0.8750
4	0.8435	0.6894	0.9929	0.9535
5	0.8984	0.7064	0.9867	0.9373
6	0.8678	0.6970	0.9804	0.9402
7	0.7650	0.6332	0.9745	0.8525
8	0.8827	0.7310	0.9786	0.8131
9	0.8949	0.7417	-	-
10	0.8145	0.6651	-	-
11	0.8640	0.6890	-	-
12	0.8839	0.7415	-	-

Because there are 12 topics and 8 topics on the Game and SFV datasets respectively, we cannot display them all. We select one topic for each dataset as an example to show the high correlation between source expertise and ground truth accuracy. Figures 2 and 3 show two example topics of the Game and SFV dataset respectively. Each point denotes a source who answers questions on this topic. X-axis is the ground truth accuracy and Y-axis is the expertise for each source on this topic. Ideally, the expertise estimated by the proposed method is consistent with the ground truth accuracy. Therefore, all the points should lie on a straight line. If the coefficient values of *Pearson* and *Kendall* both equal to 1, the agreement between the two rankings is perfect, i.e., the two rankings are the

[8] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient
[9] http://en.wikipedia.org/wiki/Kendall_tau_rank_correlation_coefficient

same. For the two datasets, the source expertise (Y) increases when ground truth accuracy (X) increases, which means that the source expertise learned by FaitCrowd is highly correlated with the ground truth accuracy.

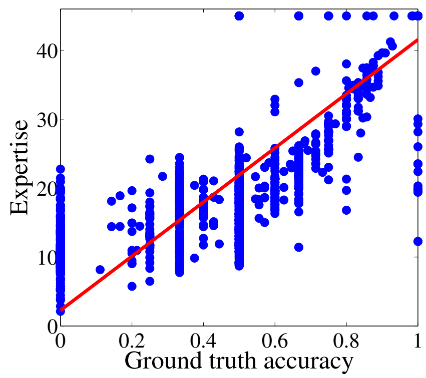

Figure 2: Correlations of Topic 2 on the Game dataset.

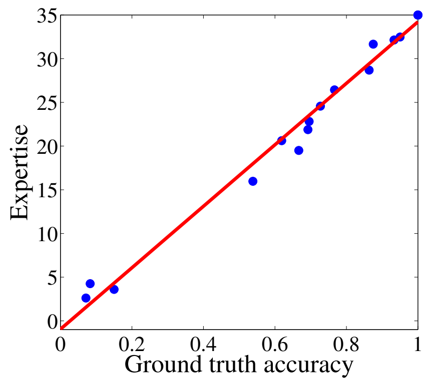

Figure 3: Correlations of Topic 4 on the SFV dataset.

5.4.3 Expertise Diversity Analysis

We now show two examples to illustrate the diverse topical expertise learned by the proposed FaitCrowd model. For each source, we compare the topical expertise obtained by the proposed model with ground truth accuracy on topics. The topical expertise for each source may vary on different topics. Ideally, it should correspond to the ground truth accuracy, i.e., the higher source expertise, the higher the ground truth accuracy. Figure 4 and Figure 5 show the statistics of Source 7 on the Game dataset and Source 16 on the SFV dataset. Each point represents a topic, X-axis is the source's ground truth accuracy and Y-axis is its expertise on each topic. From Figure 4, we can see that the topical expertise learned by the proposed FaitCrowd model is diverse, and the source with higher ground truth accuracy has higher expertise. Similar to the Game dataset, the topical expertise of Source 16 varies on different topics in Figure 5. From these two examples, we can see that the proposed FaitCrowd can estimate diverse topical expertise effectively. The proposed method uses text information to estimate expertise on different topics.

5.5 Parameter Sensitivity Analysis

To better visualize the effect of parameters, we use accuracy (1 - *Error Rate*) to validate the performance of the proposed method.

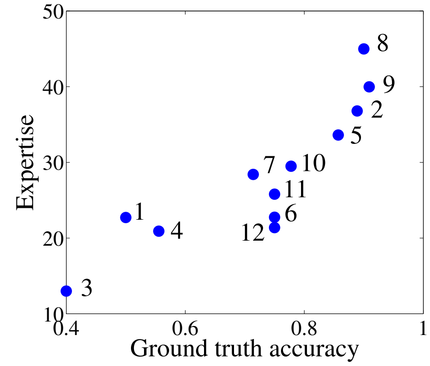

Figure 4: Source 7 on the Game dataset.

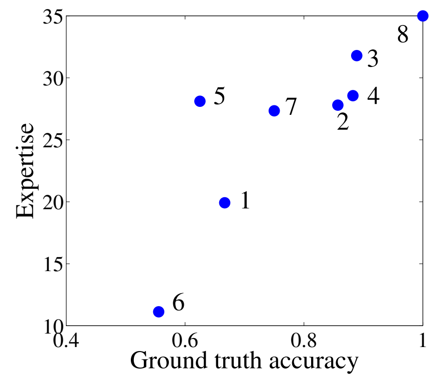

Figure 5: Source 16 on the SFV dataset.

Figure 6 shows parameter settings on the Game dataset. X-axis denotes the mean μ, Y-axis denotes the variance σ^2 of Gaussian distribution we assumed on source expertise e, Z-axis is the accuracy of the proposed method on each pair of μ and σ^2. We can see that when the value of μ increases, the accuracy has the increasing tend. When $\mu = 45$ and $\sigma^2 = 70$, the accuracy reaches the peak value. Then, the accuracy drops slightly when μ increases. However, the change is typically small, which means the proposed method is not heavily affected by parameter settings.

6. RELATED WORK

Some existing approaches conduct multi-source data aggregation by incorporating the estimation of source reliability, and thus they are relevant to the proposed approach. Yin et. at. [26] formally defined truth discovery problem and used a heuristic method, named TruthFinder, to compute the probability of each answer being correct given the estimated user reliability degrees. Pasternack et. al. [14] introduced a framework, called Investment in which sources "invest" their reliability uniformly on the observations they provide, and collect credits back from the confidence of those observations. In turn, the confidence of observations grows according to a non-linear function defined based on the sum of invested reliability from their providers. Three fixpoint algorithms (including 3-Estimates) were proposed by [6] corresponding to different levels of complexity of an underlying probabilistic model to estimate source reliability. AccuPr (a special case of Accu model) was introduced by Dong et. al. in [4]. Li et. al. [11] proposed an optimization framework, CRH, to model different data types jointly, and estimate source reliability and truth simultaneously. They also

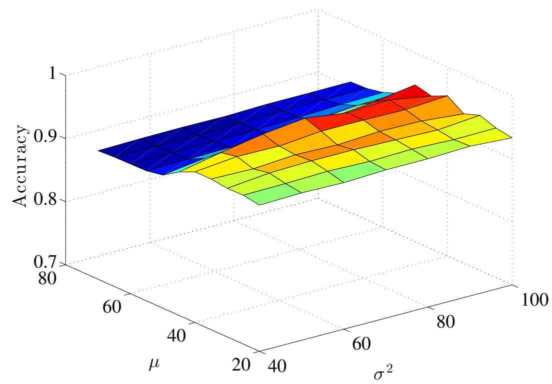

Figure 6: Performance w.r.t. parameters on the Game dataset.

proposed CATD [10] method to automatically estimate truth from conflicting data with long-tail phenomenon. Note that FaitCrowd is quite different from CRH and CATD. CRH and CATD only model users' answers, but FaitCrowd jointly models questions' text and users' answers. Also, the outputs are different. CRH and CATD only provide one reliability per source/user. Differently, FaitCrowd outputs fine grained users' expertise as well as questions' topic.

The following methods are relevant to truth discovery, but have a different problem setting. Pasternack et. al. used a set of probabilistic model parameters to estimate the source credibility in [15]. Based on the idea of "gain" and "cost", Dong et. al. [5] focused on source selection problem in truth finding. Zhao et. al. presented a probabilistic graphical model to resolve the problem of existence of multiple truths for a single entity in truth discovery tasks in [28] and designed a probabilistic graphical model to estimate source reliability on numerical data in [27]. Vydiswaran et. al. [20] and Mukherjee et. at. [13] proposed different models to estimate users' reliability and discover credible claims on unstructured data.

There are also some work related to crowdsourced data aggregation. The classic approach was named D&S [2], which used a confusion matrix for each user and a class prior to model user expertise. ZenCrowd [3] used EM to simultaneously estimate true labels and user reliability, which assumes that users act independently and simplifies the estimation of the full confusion matrix per user. These two methods have the same problem setting with the proposed model. They are used as baselines in the experiments.

A different problem setting is used in the following crowdsourced data aggregation methods. Snow et. al. [17] adopted D&S [2] model but considered the fully-supervised case of Maximum Likelihood Estimation with Laplacian smoothing. Venanzi et. al. [19] introduced CommunityBCC (Community-based Bayesian aggregation model) to estimate each user's reliability and true labels using the community's confusion matrices and employing ground truth to improve the accuracy. CommunityBCC is a semi-supervised method, which is different from the proposed unsupervised model.

GLAD [24] used the user expertise and the questions' difficulty to estimate the true answer. Raykar et. al. [16] proposed a Bayesian approach to add work specific priors for each class for binary labeling tasks. Similar to [16], Welinder et. al. [23] also added priors to each parameter used in Bayesian approach. However, this method cannot generalize to multi-choice scenario. Zhou et. al. [31] defined a separate probabilistic distribution for each user-item pair and adopted a minimax entropy principle to estimate true labels and user reliability jointly. These methods are used in binary labeling tasks, however, the proposed model is to handle on multiple-choice questions aggregation.

For the estimation of topic-level expertise in community-based question answering tasks, previous work focused on learning latent topics and topic-level user expertise. Guo et. al. [9] proposed a generative model for questions and users by using the category information. Yang et. al. [25] proposed the CQARank model to estimate both latent topics of questions and topical expertise by exploiting voting information. Zhao et. al. [29] proposed TEL model to generate experts and topics simultaneously by using users' historical contribution. Though these approaches can be used to estimate topic-level expertise, they need extra information in addition to question content and users' answers, such as categories, user votes and users' historical contributions, to help infer topical expertise accurately. Therefore, the setting is very different from the task in this paper. Note that we do not assume the availability of any other information, but only use question content and user answers to jointly learn topic-level expertise and true answers.

All the above discussed methods cannot estimate source reliability accurately for each topic when expertise significantly differs on topics. To the best of our knowledge, we are the first to build a joint model to consider both question topics and fine grained user expertise simultaneously. By modeling question content and answers alternatively, the proposed FaitCrowd can fully take advantage of available information and obtain more accurate estimation of topical expertise.

7. CONCLUSIONS

The estimation of source reliability is crucial for effective multi-source data aggregation. Many existing works in multi-source data aggregation propose various ways to estimate source reliability. These methods usually assume that source reliability is consistent across different questions. However, the expertise of sources should be topic dependent in the sense that on different topics their expertise may vary significantly. A naive adaptation of existing work is to simply split data based on topics and then apply those aggregation methods on each group defined by a topic separately. This approach faces a serious challenge that there may be insufficient data to support a good estimation of source reliability. In this paper, we propose a novel probabilistic Bayesian model to address the challenge of inferring fine grained source reliability. By jointly modeling question content and collected answers, the proposed model learns the topics of questions, topic-specific expertise of sources, and the true answers simultaneously. Experimental results on two real crowdsourced datasets prove the effectiveness of the proposed FaitCrowd model. We demonstrate that FaitCrowd can successfully detect the true answers from the expert sources on the corresponding topics even when their answers are minority in the answer set. Analysis shows that the learned topical expertise for sources is consistent with the real topical expertise.

8. ACKNOWLEDGEMENTS

This work was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1319973, IIS-1017362, IIS-1320617, and IIS-1354329, HDTRA1-10-1-0120, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), and MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

9. REFERENCES

[1] B. I. Aydin, Y. S. Yilmaz, Y. Li, Q. Li, J. Gao, and M. Demirbas. Crowdsourcing for multiple-choice question answering. In *Twenty-Sixth IAAI Conference*, pages 2946–2953, 2014.

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

[3] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st International Conference on World Wide Web*, pages 469–478, 2012.

[4] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. In *Proceedings of the VLDB Endowment*, 2(1):550–561, 2009.

[5] X. L. Dong, B. Saha, and D. Srivastava. Less is more: Selecting sources wisely for integration. In *Proceedings of the VLDB Endowment*, 6(2):37–48, 2012.

[6] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 131–140, 2010.

[7] T. L. Griffiths and M. Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[8] H. Ji, R. Grishman, and H. T. Dang. Overview of the TAC 2011 knowledge base population track. In *Third Text Analysis Conference*, 2011.

[9] J. Guo, S. Xu, S. Bao, and Y. Yu. Tapping on the potential of Q&A community by recommending answer providers. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 921–930, 2008.

[10] 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. In *Proceedings of the VLDB Endowment*, 8(4):425–436, 2014.

[11] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 1187–1198, 2014.

[12] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. A Survey on Truth Discovery. In *ArXiv Preprint ArXiv:1505.02463*, 2015.

[13] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil. People on drugs: Credibility of user statements in health communities. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 65–74, 2014.

[14] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 877–885, 2010.

[15] J. Pasternack and D. Roth. Latent credibility analysis. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1009–1020, 2013.

[16] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. In *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[17] 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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[18] L. Su, Q. Li, S. Hu, S. Wang, J. Gao, H. Liu, T. F. Abdelzaher, J. Han, X. Liu, Y. Gao, and L. Kaplan. Generalized decision aggregation in distributed sensing systems. In *2014 IEEE Real-Time Systems Symposium (RTSS)*, pages 1–10, 2014.

[19] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 155–164, 2014.

[20] V. Vydiswaran, C. Zhai, and D. Roth. Content-driven trust propagation framework. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 974–982, 2011.

[21] H. M. Wallach. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 977–984, 2006.

[22] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proceedings of the 11th International Conference on Information Processing in Sensor Networks*, pages 233–244, 2012.

[23] P. Welinder, S. Branson, P. Perona, and S. J. Belongie. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, pages 2424–2432, 2010.

[24] 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 *Advances in Neural Information Processing Systems*, pages 2035–2043, 2009.

[25] L. Yang, M. Qiu, S. Gottipati, F. Zhu, J. Jiang, H. Sun, and Z. Chen. CQARank: jointly model topics and expertise in community question answering. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pages 99–108, 2013.

[26] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, 2008.

[27] B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In *Proceedings of the 10th International Workshop on Quality in Databases*, 2012.

[28] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. In *Proceedings of the VLDB Endowment*, 5(6):550–561, 2012.

[29] T. Zhao, N. Bian, C. Li, and M. Li. Topic-level expert modeling in community question answering. In *SIAM International Conference on Data Mining*, pages 776–784, 2013.

[30] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, pages 338–349, 2011.

[31] D. Zhou, S. Basu, Y. Mao, and J. C. Platt. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems*, pages 2195–2203, 2012.