Modeling Truth Existence in Truth Discovery

Shi Zhi; Bo Zhao; Wenzhu Tong; Jing Gao; Dian Yu; Heng Ji; and Jiawei Han

ABSTRACT

When integrating information from multiple sources, it is common to encounter conflicting answers to the same question. Truth discovery is to infer the most accurate and complete integrated answers from conflicting sources. In some cases, there exist questions for which the true answers are excluded from the candidate answers provided by all sources. Without any prior knowledge, these questions, named no-truth questions, are difficult to be distinguished from the questions that have true answers, named has-truth questions. In particular, these no-truth questions degrade the precision of the answer integration system. We address such a challenge by introducing source quality, which is made up of three fine-grained measures: silent rate, false spoken rate and true spoken rate. By incorporating these three measures, we propose a probabilistic graphical model, which simultaneously infers truth as well as source quality without any a priori training involving ground truth answers. Moreover, since inferring this graphical model requires parameter tuning of the prior of truth, we propose an initialization scheme based upon a quantity named truth existence score, which synthesizes two indicators, namely, participation rate and consistency rate. Compared with existing methods, our method can effectively filter out no-truth questions, which results in more accurate source quality estimation. Consequently, our method provides more accurate and complete answers to both has-truth and no-truth questions. Experiments on three real-world datasets illustrate the notable advantage of our method over existing state-of-the-art truth discovery methods.

1. INTRODUCTION

The rapid growth of web data provides an overwhelming amount of information. Though information is available from more sources, sources often disagree with each other. For example, the location of the last confirmed case of Ebola patient in the US may be reported by multiple websites with different answers. Any false report of location leads to unnecessary panic. Hence, it is crucial to identify the most accurate and complete answer among conflicting answers. This problem is commonly known as truth discovery [19].

One straightforward approach to truth discovery problem is majority voting. It collects all possible answers to one question from multiple sources and treats the most frequent answer as the truth. However, majority voting ignores an important fact: the quality of each source is often heterogeneous. Consequently, it fails to discover the truth in the setting where a majority of sources provide a wrong answer, while few high-quality sources provide the correct answer. In order to address such a challenge, one feasible approach is to incorporate source quality [5,11,19]. The rationale behind this approach is, if an answer comes from a more reliable source, it is more likely to be true; meanwhile, if a source is associated with a more trustworthy answer, it is likely to be more reliable. Naturally, one may infer the truth and source quality in an iterative way. It allows to find truth and estimate source quality in an unsupervised fashion. There exist several methods that leverage this intuition and additional heuristics [3,15,17,20].

However, existing methods do not address a critical issue: there might be questions, named no-truth questions, whose true answers are not included in the candidate answers provided by all sources. Without careful treatment, this issue can severely degrade the performance of the truth discovery system. In the sequel, we motivate this truth existence problem with an example.

Example Recently, automatic knowledge base construction is explored to build a proliferation of knowledge bases [2]. In knowledge base construction, one crucial step is to build information extracting systems to discover the answers from millions of documents, named slot filling [6]. The true answers to the questions do not necessarily exist in the corpus and can be hard to detect. In Tables 1 and 2 we provide an example of the slot filling task: Table 1 gives the questions; in Table 2 each column represents the answers that come from 13 sources to a single question, while each row gives the answers provided by a single source to the 8 questions. Each blank item suggests that the corresponding source does not provide any answer to this question. We name the blank items to be an empty answer.

For the first four questions, correct answers exist among the candidate answers. We define them as the has-truth questions. Meanwhile, for the last four questions, the answers either do not exist, or are not discovered by all sources. We define the truth to these questions as empty, and the questions as no-truth questions.

Table 1: Example Questions of Slot Filling Task

<table>
<thead>
<tr>
<th>Question</th>
<th>Has-truth</th>
<th>No-truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>q1: What’s the age of Ramazan Bashardost?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q2: What’s the province of birth of Ramazan Bashardost?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q3: What’s the age of Marc Bolland?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q4: What’s the age of Stuart Rose?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q5: What’s the age of Ramazan Bashardost?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q6: What’s the country of birth of Marc Bolland?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q7: What’s the province of death of Stuart Rose?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>q8: What’s the province of death of Stuart Rose?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DOI: http://dx.doi.org/10.1145/2783258.2783339.
Our task is to integrate the answers to all questions. Ideally, for questions whose true answers are among the candidate answers, we should identify them, while for those questions whose answers are not among the candidates, e.g., the province of death of Stuart Rose who is still alive, we should faithfully provide the empty answer. Based on this, we point out the drawbacks of some strategies.

**Strategy 1 (MajVot)** Majority voting among non-empty answers easily fails in this example. As shown in Table 2 for the first four questions majority voting will have a correct answer. However, for the other questions majority voting would randomly choose a candidate answer instead of refusing to answer it. Though majority voting answers all has-truth questions correctly, half of its answers, i.e., \( q_5 \) to \( q_8 \), are wrong. Moreover, incorporating source quality in the same way as TruthFinder \([9]\) cannot alleviate the issue, because it will still provide answers to all questions regardless of truth existence. An alternative approach is to compute a confidence score based upon source quality and use a threshold to decide whether we want to give out an answer or not. However, it is hard to determine such a suitable threshold without any extra knowledge.

**Strategy 2 (MajVotEmp)** Based on Strategy 1, a naive way to deal with the no-truth questions is to treat the empty answer as a candidate answer and apply majority voting on both empty and non-empty answers. However, this strategy is likely to output empty answer as the true answer for many questions. Consider the example, for \( q_8 \) and \( q_9 \), the most frequent answer is the empty answer. Recent work \([11]\) treats the empty answer equally as other non-empty answers and considers source quality to estimate the truth. Even in this situation, the final output is severely degraded because the estimation of source quality will be severely affected by empty answers, which in turn affects the truth inference.

In summary, previous methods will fail when no-truth questions exist. They either suffer from low accuracy when they provide answers to all questions, or low coverage when they treat empty answer equally as the others. Truth existence estimation is crucial because when a source fails to answer a has-truth question, it should be punished; when it keeps silent to a no-truth question, it should be rewarded. Source quality measures used in previous work \([17],[19]\) cannot alleviate this issue even when they consider empty answers because using single quality measure cannot depict the complete performance of sources, which will hurt the truth estimation step.

Therefore, we propose a new model called Truth Existence Model (TEM), which can leverage the correctness and the completeness of truth integrated from a mixture of correct answers, empty answers and erroneous answers. To the best of our knowledge, this is the first work modeling the existence of truth in truth discovery. For no-truth questions, the proposed unsupervised approach can confidently output an empty answer instead of randomly selecting a non-empty answer as the output. We define three source quality: silent rate, false spoken rate and true spoken rate to model a complete spectrum of source behavior. We propose a probabilistic model that can naturally incorporate the proposed source quality measures into the estimation of truth. Efficient inference and parameter setting strategies are derived. We also provide a novel cluster-based initialization scheme to help better estimate truth existence.

We evaluate TEM on three real-world datasets on both source quality and truth estimation. TEM can achieve 19.4% improvement in F1 on SF2013. Furthermore, the results on synthetic datasets with various proportion of no-truth questions demonstrate that TEM can perform consistently best on both high-quality and low-quality datasets. We also discuss different initialization schemes of truth existence and variations of TEM. The results show that TEM outperforms state-of-the-art truth discovery approaches in both accuracy and robustness with comparable time complexity.

The rest of the paper is organized as follows. We first describe our data model and problem formulation in Section 2. In Section 3 we define three types of source quality. Section 4 introduces the probabilistic graphical model and inference algorithm. Section 5 presents our experimental results. We present the related work in Section 6 and conclude the paper in Section 7.

### 2. PROBLEM FORMULATION

In this section we introduce important terms and the problem definition. We assume a single-value question type. The truth to a single-value question is unique. In this case, each source only provides a single answer to each question.

Let \( E \) denote an empty answer when a source keeps silent to a question. Let \( Q = \{q_1, ..., q_M\} \) be the set of questions where \( M \) is the total number of questions. Each question either has no truth or a single truth. Let \( S = \{s_1, ..., s_N\} \) be the set of sources where \( N \) is the total number of sources. Let \( D_i = \{d_{i,1}, ..., d_{i,N}\} \) be the set of distinct non-empty candidate answers to question \( q_i \), where \( N_i \) is the number of distinct candidate answers to question \( q_i \). \( D_i \) only contains the answers provided by sources in \( S \). Let \( A = \{a_{i,1}, ..., a_{i,N}\} \) be the set of observed answers provided by all sources to all questions. Each answer \( a_{ij} \) is associated with \( q_i \) and \( s_j \). Each source \( s_j \) will provide only one answer to one question \( q_i \). The answer can take an empty answer \( E \) or a non-empty answer in \( D_i \). Let \( T = \{t_1, ..., t_M\} \) be the set of truths where each \( t_i \) associated with \( q_i \) can either be a non-empty answer in \( D_i \) or an empty answer \( E \). We define a has-truth question as the question whose truth is in the non-empty candidate answer set \( D_i \), and a no-truth question as the question whose truth is not in \( D_i \). The truth of a no-truth question is the empty answer \( E \).

**Example 1.** Considering the data given by Table 2, the question set is \( Q = \{q_1, ..., q_8\} \). The source set is \( S = \{s_1, ..., s_{13}\} \). The non-empty candidate answer set to the first question \( q_1 \) is \( D_1 \) = \{Afghanistan, Ghazni\}, \( q_2 \) = \{Empty\}, \( q_3 \) = \{Ghazni\}, \( q_4 \) = \{50\}, \( q_5 \) = \{London\}, \( q_6 \) = \{Marks\}, \( q_7 \) = \{Russia\}, \( q_8 \) = \{Empty\}, \( q_9 \) = \{Holly\}, \( q_{10} \) = \{Kabul\}, \( q_{11} \) = \{50\}, \( q_{12} \) = \{Ghazni\}, \( q_{13} \) = \{50\}. The input of this truth discovery problem is the set of observed answers \( A = \{a_{1,1}, a_{1,2}, ..., a_{1,13}\} \), where the answer provided by \( s_1 \) to \( q_1 \) is a non-empty answer \( a_{1,1} = \text{Afghanistan} \), and the answer provided by \( s_2 \) to \( q_2 \) is \( E \), i.e., \( a_{2,1} = \text{Empty} \). The output of this truth discovery problem is the truth set of the 8 questions \( T = \{\text{Afghanistan}, \text{Ghazni}, 50, E, E, E, E, E\} \). where \( q_1 \), \( q_2 \), \( q_3 \), \( q_4 \), \( q_5 \), \( q_6 \), \( q_7 \), \( q_8 \) are has-truth questions and \( q_9 \), \( q_{10} \) are no-truth questions.

Given a set of observed answers \( A \) for \( M \) questions in \( Q \) provided by \( N \) sources in \( S \), the goal is to infer the truth \( t_i \) to each
question $q_i$ and estimate the quality of each source. Note that both truth and truth existence in the input are not known beforehand. Instead, we must infer the hidden truths by fitting the observed answers into our model.

3. **SOURCE QUALITY**

In this section, we explore source quality measures in our truth discovery model and explain why quality measures in previous methods fail in the scenarios involving truth existence problem.

3.1 Confusion Matrix

As we discussed in Section 1, empty answers are very important inputs that need to be modeled together with non-empty answers. Based on the observed answers of one source $s$ and truths of all questions, we generate the confusion matrix of source $s$ in Table 4. Here $t_i$ is the variable to represent the truth of question $q_i$, $a_i$ is the observed answer from $s$, and $d_i$ is the correct answer to question $q_i$. $E$ is an empty answer.

In Table 4, True Non-Empty (TNE) is the number of cases when source $s$ correctly answers a has-truth question. False Non-Empty (FNE) is the number of wrong answers provided by source $s$. It contains two parts: $FNE_1$ is the number of cases when source $s$ provides a wrong answer to a has-truth question; $FNE_2$ is the number of cases when source $s$ provides a non-empty answer to a no-truth question. False Empty (FE) is the number of cases when source $s$ provides an empty answer to a has-truth question. True Empty (TE) is the number of cases when source $s$ keeps silent to a no-truth question. The number of has-truth questions is $TNE + FNE_1 + FE$. The number of no-truth question is $FNE_2 + TE$ and the total number of questions is $TNE + FNE + FE + TE$, where $FNE = FNE_1 + FNE_2$.

**EXAMPLE 2.** Consider the source $s_{13}$ in the example of Table 2. $s_{13}$ gives correct answers to $q_3$ and $q_4$, thus $TNE = 2$; wrong answer to one has-truth question $q_1$, thus $FNE_1 = 1$; wrong answer to one no-truth question $q_5$, thus $FNE_2 = 1$; empty answer to one has-truth question $q_2$, thus $FE = 1$; empty answers to $q_6$, $q_7$ and $q_8$, thus $TE = 3$. The total number of questions is 8 with 4 has-truth questions and 4 no-truth questions.

3.2 Quality Measures

For each source, we define different source quality measures to describe different behaviors. We first define three new measures on has-truth questions: **Silent Rate**, **False Spoken Rate** and **True Spoken Rate**. Silent rate and false spoken rate differentiate two types of errors: false empty cases and false non-empty cases.

- **Silent Rate (SR)** is the probability that a source keeps silent to a has-truth question, i.e. $SR = \frac{FE}{FNE + FNE_1 + TNE}$
- **False Spoken Rate (FR)** is the probability of providing a wrong answer to a has-truth question, i.e. $FR = \frac{FNE_1}{FNE_1 + TNE}$
- **True Spoken Rate (TR)** is the probability of its answer being correct of a has-truth question, i.e. $TR = \frac{TNE}{FNE_1 + TNE}$
- **The relationship among them** is $SR + FR + TR = 1$.

Then for no-truth questions, two cases may happen: a source may provide a non-empty answer, which contributes to $FNE_2$, or an empty answer, which contributes to $TE$. We can define an additional false spoken rate on no-truth questions, i.e. $FR' = \frac{FE}{FNE_2 + TE}$. Then we can deduce that the quality measures on no-truth questions are:

- The probability that a source provides a wrong answer to a no-truth question is $FR'$.

<table>
<thead>
<tr>
<th>$a_i$</th>
<th>$t_i$</th>
<th>$t_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_i$</td>
<td>$E$</td>
<td>$E$</td>
</tr>
</tbody>
</table>

- The probability that a source keeps silent to a no-truth question is $1 - FR'$.

However, it is not the best way to define an additional false spoken rate on no-truth questions. Instead, we propose an assumption on the consistency of false spoken rate.

**ASSUMPTION 1.** **False spoken rate is consistent across has-truth part and no-truth part**, i.e. $FR = FR'$.

Based on this assumption, by simple algebra we can compute that $FR = \frac{FNE_1 + TNE}{FNE_2 + FE + TNE + TNE + FNE_1 + TE}$. Thus, we can define a new false spoken rate $FR = \frac{FNE}{TNE + FNE + TNE}$. In particular, the numerator of $FR$ is the number of wrong answers a source provides to both has-truth questions and no-truth questions. The denominator is the number of questions. $FR$ is the probability of a source providing a wrong answer to any question. It represents an overall false spoken rate across all questions. We can see that $FR = FR = FR'$, which means that the overall false spoken rate is consistent on both has-truth part and no-truth part.

By making this assumption, we have a single false spoken rate across all questions instead of two independent ones on each part. This assumption is reasonable because: (1) When a source provides an answer to a question, it will not consider its truth existence, so the probability of providing a wrong answer is independent of truth existence, i.e consistent across all questions. (2) It reduces the number of parameters to estimate, so it increases the effective sample size to make a more accurate estimation on overall false spoken rate. Later we will show it with experiments on real-world datasets.

Next, we re-define silent rate $SR$ and true spoken rate $TR$ based on this assumption and $FR$. Since $FR$ is the overall false spoken rate across all questions, the definitions of $SR$ and $TR$ should also be on all questions. Keeping the ratio of silent rate to true spoken rate consistent across has-truth questions and all questions, i.e. $SR = \frac{SE}{SR + FR + TR}$ and make the constraint $SR + FR + TR = 1$ stays true, we can get the following definitions.

- **SR** = $\frac{FE}{TNE + FNE + TNE} (1 - FR)$
- **FR** = $\frac{FNE}{TNE + FNE + TNE + FE + TNE}$
- **TR** = $\frac{TNE}{TNE + FE + TNE}$

The overall $SR$ and $TR$ can be interpreted as follows. (1) A no-truth question could be both valid and invalid. For example, in Table 1 $q_5$ is a valid question whose answer possibly exists in the document collection, while $q_8$ is invalid because Stuart Rose is still alive. Based on the confusion matrix in Table 4 $TE$ is the number of cases when a source keeps silent to a no-truth question. Sources generate true empty cases may result from two reasons: failing to provide an answer to a valid question, which contributes to $TE_1$, or correctly keeping silent to an invalid question, which contributes to $TE_2$. (2) We can re-write $SR$ and $TR$ as $SR = \frac{FE + TNE}{TNE + FNE_2 + SE}$ and $TR = \frac{TNE + FE}{TNE + FNE + FE + TNE}$. By definition we can compute that $TE_1 = \frac{FE}{FNE_2 + FE + TNE + TNE}$, $TE_2 = \frac{TNE}{FNE_2 + FE + TNE}$. It means that by keeping the ratio of silent rate to true spoken rate consistent across has-truth questions and all questions, true empty cases contribute to $SR$ and $TR$ proportionally to $FE$ and $TNE$, i.e. $\frac{TE_1}{TE_2} = \frac{FE}{TNE}$. (3) Keeping the ratio of silent rate to true spoken rate consistent is equivalent to making both silent rate and true spoken rate consistent across has-truth part and no-truth part, i.e. $SR = SR', TR = TR'$. 

![Table 4: Confusion Matrix of Source s](image)
where \( SR' \) and \( TR' \) are the silent rate and true spoken rate of no-truth part, \( SR' = \frac{TE_1}{FNE_2 + TE_1}, TR' = \frac{TE_2}{FNE_2 + TE_2} \). Thus, we also have \( SR = SR' = \overline{SR}, TR = TR' = TR' \).

Actually in our model, we do not differentiate \( TE_1 \) from \( TE_2 \) explicitly. By keeping the ratio of silent rate to true spoken rate consistent, the computation of \( \overline{SR} \) and \( \overline{TR} \) is dependent on the sum of \( TE_1 \) and \( TE_2 \), i.e. \( TE \), but not on each individual element.

Similarly, we can derive that \( 1 - FR = TR + SR \). It holds for both has-truth and no-truth questions. For has-truth part, when a source does not provide a wrong answer, it may either provide a correct answer or keep silent. For no-truth part, when a source gives no answer to a question, it may result from failing to provide the correct answer to a valid question, which is associated with silent rate, or by successfully predicting that there is no true answer to an invalid question, which corresponds to true spoken rate.

### 3.3 Limitation of Precision

A commonly used measure of source quality is \( Precision \). We can define it based on the confusion matrix as follows.

- **Precision** (PRE) is the probability of its non-empty answers being correct, i.e. \( \frac{TNE}{TNE + FNE} \).

Revisit the example in Table 2. Table 5 represents the source quality measures of \( s_{12} \) and \( s_{13} \). Previous work [11, 19], e.g. AverageLog and TruthFinder use precision to model the quality of sources. They only consider non-empty answers and ignore empty answers from sources. The integrating algorithms provide non-empty answers to all questions. Thus, they fail in all no-truth questions. In Table 2, the truth of \( q_3 \) will be predicted as Holland instead of E, because non-empty answer is only provided by \( s_{13} \) whose precision is 0.5.

### 3.4 Limitation of True Spoken Rate

Another alternative is to only use true spoken rate as the source quality measure. It is similar to the source quality used in probabilistic models such as LCA [12], LTM [21] and EM [17]. However, existing work do not differentiate two types of errors, i.e. false non-empty (FNE) and false empty (FE). To illustrate the necessity of it, we use the following example.

In Table 5 source \( s_{12} \) and \( s_{13} \) have the same true spoken rate but different silent rate and false spoken rate. To simplify the problem, we only consider the contribution made by \( s_{12} \) and \( s_{13} \) to \( q_2 \). With only true spoken rate as the quality measure, the probability of Afghanistan to be true is \( TR_{12} \times (1 - TR_{13}) = 0.25 \). The probability of E to be true is \( (1 - TR_{12}) \times TR_{13} = 0.25 \). These two answers get the same score because missing an answer counts to the same measure as a wrong answer does. Thus, both sources are treated as with the same quality.

However, we can conclude that \( s_{13} \) is more likely to provide a wrong answer, which is reflected by its higher error rate. Thus, when \( s_{13} \) provides an empty answer and \( s_{12} \) gives non-empty answer, empty answer made by \( s_{12} \) is more likely to be wrong while the non-empty one given by \( s_{12} \) is more likely to be the truth.

Consequently, if we differentiate empty answer from wrong answer and recompute the probabilities again, we can get the probability of Afghanistan to be true is \( TR_{12} \times SR_{13} = 0.5 \times 0.25 = 0.125 \), and the probability of an empty answer to be true is \( FR_{12} \times (1 - FR_{13}) = 0 \times 0.75 = 0 \). We can conclude that the correct answer to \( q_2 \) is Afghanistan.

### 4. COMPUTATIONAL MODEL

In this section we first describe our TEM model, a Bayesian network that naturally incorporates true spoken rate, silent rate and false spoken rate into truth estimation. We formulate it as a Maximum Likelihood Estimation (MLE) problem and apply EM algorithm to jointly estimate source quality and truth. In each iteration, source quality and truth are iteratively computed. We also discuss an initialization scheme to set the prior of truth.

#### 4.1 Model Details

We tackle the truth existence problem using Bayesian network. Figure 1 is the graphical structure of our probabilistic model. Each node represents a random variable. Each dot represents a prior parameter. The shaded nodes indicate the corresponding variables are known, and lighter nodes mean the latent variables we are going to infer. The letter on the right corner of each plate is the number of replicates for each node, e.g. \( N \) is the number of sources, \( M \) is the number of questions. A directed edge from node \( a \) to another node \( b \) means that \( b \) is generated from a distribution parameterized by values of \( a \) in addition to other related nodes.

#### Three Source Quality Measures

For each source \( s_i \in S \), we can define silent rate, false spoken rate and true spoken rate by probabilities, denoted by \( \phi_{i1}^{(1)}, \phi_{i2}^{(2)}, \phi_{i3}^{(3)} \). According to the original definitions, **silent rate** is the probability that \( s_i \) keeps silent when a question has truth, i.e. \( \phi_{i1}^{(1)} = P(a_{ij} = E|t_i = d_{ia}, t_i \neq E) \).

Based on the assumption that false spoken rate is consistent across has-truth and no-truth parts, **false spoken rate** is the probability it makes mistakes on either has-truth or no-truth questions, i.e. \( \phi_{i2}^{(2)} = P(a_{ij} = d_{ia}|t_i = d_{ia}, t_i \neq E) = P(a_{ij} \neq E|t_i = E) \).

**True spoken rate** is the probability to provide a trustworthy answer when a question has truth, i.e. \( \phi_{i3}^{(3)} = P(a_{ij} = d_{ia}|t_i = d_{ia}, t_i \neq E) \). The relationship among these three source quality measures is that silent rate, false spoken rate and true spoken rate add up to 1, i.e., \( \phi_{i1}^{(1)} + \phi_{i2}^{(2)} + \phi_{i3}^{(3)} = 1 \). We can derive that \( p(a_{ij} = E|t_i = E) = 1 - p(a_{ij} \neq E|t_i = E) = 1 - \phi_{i2}^{(2)} \).

#### Prior of Source Quality Measures

For each source \( s_i \in S \), a source quality vector, denoted by \( \phi_s \), i.e. \( \phi_s = (\phi_{i1}^{(1)}, \phi_{i2}^{(2)}, \phi_{i3}^{(3)}) \). We generate \( \phi_s \) from a Dirichlet distribution with hyper-parameter \( \alpha = (\alpha_1, \alpha_2, \alpha_3) \), i.e. \( \phi_s \sim \text{Dirichlet}(\alpha) \).

Later we will see that \( \alpha \) serves as the pseudo counts of silent answers, wrong answers and correct answers when estimating the corresponding source quality. It controls the prior belief for three source quality measures. In practice, we can plug in our assumption on the overall quality of sources by using either symmetric or asymmetric Dirichlet distribution. It makes our model robust to both high-quality and low-quality data.

#### Prior of Truth

For each question \( q_i \in Q \), we define the prior of truth as \( \eta_{i0} = \theta_{i01}, \eta_{i1}, \eta_{i2}, ..., \eta_{iN_i} \), where \( \eta_{i0}(t_i = 1, ..., N_i) \) is the probability of truth \( t_i \) being one of the non-empty candidate answer \( d_{ia} \). \( \eta_{i0} = \)
\(P(t_i = E) \eta_i n = P(t_i = d_i n), n = 1, \ldots, N_i\). The probability that the truth is empty or any non-empty candidate answer should add up to 1 i.e. \(\sum_{n=1}^{N_i} \eta_i n = 1\). We will discuss the initialization of truth distribution in depth in Section 4.4.

### 4.2 Likelihood Function

The observed answers to \(q_i\) provided by \(N\) sources in \(S\) are denoted by \(A_i\), where \(A_i = \{a_{i1}, a_{i2}, \ldots, a_{iN_i}\}\) is a subset of \(A\). Based on the dependencies of random variables in TEM, for each question \(q_i\) we bring in a latent truth \(t_i \in T\) and partition the likelihood of observations of \(q_i\) into \(N_i + 1\) parts. The probability to observe \(A_i\) given source quality \(\phi_S\) is:

\[
P(A_i|\phi_S, \eta) = \sum_{n=1}^{N_i} \eta_i n P(A_i|t_i = d_i n, \phi_S) + \eta_i 0 P(A_i|t_i = E, \phi_S)
\]

Eq. 1 is expanded by the law of total probability to the combination of \(N_i + 1\) mixing components. The mixing weight for each component \(n\) is fixed to its corresponding prior truth \(\eta_i n\).

We assume sources are mutually independent. Thus, the probability \(P(A_i|t_i = d_i n)\) is the multiplication of the conditional probability of observed answer from each source, i.e. \(a_{ij}\). Thus, each component in the first \(N_i\) parts is computed by:

\[
P(A_i|t_i = d_i n, \phi_S) = \prod_{j=1}^{N} P(a_{ij}|t_i = d_i n, \phi_S)\]

Given the latent truth \(t_i\), the observed answer \(a_{ij}\) is generated from a categorical distribution parameterized by \(\phi_j\). The probability of \(a_{ij}\) given \(t_i = d_i n\) is Eq. 3 where \(I(\cdot)\) is an indicator function serving as a selector of corresponding source quality.

\[
P(\alpha_j|t_i = d_i n, \phi_S) = \phi_j(1) I(a_{ij} = E) \phi_j(2) I(a_{ij} \neq d_i n, a_{ij} \neq E) \phi_j(3) I(a_{ij} = d_i n, a_{ij} \neq E)
\]

Similarly, the last component is shown in Eq. 4.

\[
P(A_i|t_i = E, \phi_S) = \prod_{j=1}^{N} \phi_j(1) I(a_{ij} \neq E) (1 - \phi_j(2)) I(a_{ij} = E)
\]

Source quality of each source follows a Dirichlet distribution parameterized by \(\alpha\). Thus, the prior of source quality of \(S\) is:

\[
p(\phi_S|\alpha) \propto \prod_{j=1}^{N} \phi_j(1) \alpha^{-1} \phi_j(2) \alpha^{-1} \phi_j(3) \alpha^{-1}
\]

Then the complete likelihood of all observed answers and source quality given hyper-parameters \(\alpha\) and \(\eta\) is:

\[
\max P(A, S, \phi_S|\alpha, \eta) = p(\phi_S|\alpha) \prod_{i=1}^{M} P(A_i|\phi_S, \eta)
\]

Eq. 6 is the objective function. Given the observed answers \(A\), non-empty candidate answer set \(D\) of each question \(q_i\), prior of truth \(\eta\) and hyper-parameter \(\alpha\), the objective is to infer the parameters \(\Theta = \{\phi_S\}\) and estimate the posterior of latent truths in \(T\).

### 4.3 Inference

**Expectation-maximization (EM) algorithm** [2][18] is an iterative approach to find the maximum likelihood estimate of the latent variables in graphical model. EM algorithm iteratively alternates between two steps, called the expectation step (E-step) and the maximization step (M-step). In E-step, it computes the expected log-likelihood of the complete data. In M-step, it estimates parameters by maximizing the log-likelihood of the complete data. This process continues until it converges, i.e., reaching a local maxima.

**E-step:** Given observed answers \(A\) and the estimated source quality \(\phi^{(k)}\) of \(k\)-th round, for question \(q_i\), by Bayes' rule we compute the posterior probability of truth being empty of \(k+1\)-th round, \n
\[
\gamma^{(k+1)}_{\eta 0}\]

Based on the independence of sources, we derive that the posterior of truth is estimated by the multiplication of corresponding source quality and prior of truth.

\[
\gamma^{(k+1)}_{\eta 0} = P(t_i = E|\phi^{(k)}_S, \phi^{(k)}_A, \eta)
\]

\[
\propto \eta 0 \prod_{j=1}^{N} (\phi^{(1)}_j) I(a_{ij} \neq E) (1 - (\phi^{(2)}_j)) I(a_{ij} = E)
\]

Eq. 7 indicates if a question is answered by sources with high false spoken rate \(\phi^{(2)}_j\), or not answered by sources with low false spoken rate, or with high prior of \(t_i = E\), this question is unlikely to have truth. The probability that the truth of \(q_i\) being non-empty is:

\[
\gamma^{(k+1)}_{\eta 0} = P(t_i = d_i n, t_i \neq E|\phi^{(k)}_S, \phi^{(k)}_A, \eta)
\]

\[
\propto \eta 0 \prod_{j=1}^{N} (\phi^{(1)}_j) I(a_{ij} = E) (\phi^{(2)}_j) I(a_{ij} \neq d_i n, a_{ij} \neq E)
\]

\[
(\phi^{(1)}_j) I(a_{ij} = d_i n, a_{ij} \neq E)
\]

Eq. 8 shows that if an answer \(d_i n\) is provided by sources with high true spoken rate \(\phi^{(2)}_j\), or different from the answer provided by sources with high false spoken rate \(\phi^{(2)}_j\), or not provided by sources with high silent rate \(\phi^{(1)}_j\) that keeps silent to this question, or with high prior of \(t_i = d_i n\), it is prone to be the truth.

**M-step:** For source \(j\), we estimate the source quality \(\phi_j\) by maximizing the expectation of the log-likelihood of the complete data. Take derivatives of it with respect to \(\phi^{(1)}_j\) and \(\phi^{(2)}_j\) with constraint \(\phi^{(3)}_j = 1 - \phi^{(1)}_j - \phi^{(2)}_j\) and set them to zero, we get estimated source quality of \(k\)-th round with the estimation of posterior of truth of \(k\)-th round. The estimated source quality of \(s_j\) is shown in Eq. 9-11.

\[
(\phi^{(1)}_j) = \frac{a_j + (\alpha_j - 1)}{a_j + b_j + (\alpha_1 + \alpha_2 - 2)} (1 - (\phi^{(2)}_j))
\]

\[
(\phi^{(2)}_j) = \frac{c_j + (\alpha_2 - 1)}{a_j + b_j + c_j + d_j + (\alpha_1 + \alpha_2 + \alpha_3 - 3)}
\]

\[
(\phi^{(3)}_j) = \frac{b_j + (\alpha_j - 1)}{a_j + b_j + (\alpha_1 + \alpha_2 - 2)} (1 - (\phi^{(2)}_j))
\]

Eq. 12-14 are the empirical counts weighted by the posterior probability of each case being true. Eq. 15 is the weighted count of cases when \(a_{ij}\) keeps silent to a has-truth question. \(a_{ij}\) is exactly the estimation of FE defined in Section 3.

\[
a_j = \sum_{i=1}^{M} (1 - \gamma^{(k)}_{\eta 0}) I(a_{ij} = E)
\]

\[
b_j = \sum_{i=1}^{M} N_i \sum_{n=1}^{N_i} (\gamma^{(k)}_{\eta 0}) I(a_{ij} = d_i n, a_{ij} \neq E)
\]

\[
c_j = \sum_{i=1}^{M} \gamma^{(k)}_{\eta 0} I(a_{ij} \neq E) + \sum_{n=1}^{N_i} \gamma^{(k)}_{\eta 0} I(a_{ij} \neq d_i n, a_{ij} \neq E)
\]

\[
d_j = \sum_{i=1}^{M} \gamma^{(k)}_{\eta 0} I(a_{ij} = E)
\]

The sum of these weighted count is the number of questions.

\[
a_j + b_j + c_j + d_j = M
\]
Note that these weighted counts are added by corresponding pseudo counts originated from prior of source quality. Thus, the prior of source quality serves as a smoothing factor for source quality.

We can interpret Eq. 3.11 as follows. False spoken rate $\phi^{(2)}_j$ is estimated by the number of wrong answers a source provides to all questions. The ratio of silent rate $\phi^{(1)}_j$ and true spoken rate $\phi^{(3)}$ is estimated by the ratio of the number of empty answers and the number of correct answers to has-truth questions. With constraint $\phi^{(1)}_j + \phi^{(2)}_j + \phi^{(3)} = 1$, silent rate and true spoken rate have closed form solutions. We can see that Eq. 3.11 exactly match the definitions of SR, PR and TR in Section 3.

4.4 Practical Issues

4.4.1 Initialization of Prior of Truth

As mentioned in Section 4.1, for each question $q_i$, we need to set the prior of truth $\eta_i$. Two classical methods can be applied to set the prior truth distribution. (1) UNIFORM: Uniformly assign weight to each dimension of the prior truth distribution, i.e. $1/(N_i + 1)$ where $N_i$ is the number of non-empty candidate answers. It indicates that we put the same initial guess on both empty and non-empty candidate answers. (2) VOTE: Assign weight to each dimension of prior truth distribution proportionally to the number of occurrences of each candidate answer.

Take $q_2$ in Table 2 as an example. The parameter setting of prior truth distribution is shown in Table 6.

As shown in the example of Table 2, empty answers play an important role in estimating truths of questions. It is risky to treat empty answers same as the other empty ones, because a large proportion of empty answers may take the majority, making the estimation of has-truth question be no-truth. Thus, it is important to develop a new scheme to initialize the prior of empty and non-empty answers separately.

Here, we propose an initialization scheme called EXISTENCE based upon a quantity named truth existence score, which synthesizes two indicators, namely, participation rate and consistency rate. We define two types of sources and two indicators.

- **participating sources** are sources that provide a non-empty answer to a question.
- **majority sources** are sources whose non-empty answers are agreed by the largest number of sources.
- **Participating Rate (PR)** is the ratio between the number of participating sources and the number of sources.
- **Consistency Rate (CR)** is the ratio between the number majority sources and the number of participating sources.

**Example 3.** In Table 2, the participating sources of $q_2$ are $83, 84, 87, 88, 89, 812$, and the majority sources of $q_2$ are the sources who provide answer Afghanistan, i.e. $84, 87, 812$. So the participating rate of $q_2$ is $6/13$ and the consistency rate of $q_2$ is $3/6$.

These two rates can effectively reflect the truth existence of each question. High participating rate indicates that a large proportion of sources tend to provide answer to this question. High consistency rate indicates that a large proportion of sources agree on one answer. When a large proportion of sources provide an answer to a certain question and reach an agreement on one answer, the question is likely to have a correct answer within its candidate answers.

We define the truth existence score as the probability that a question has a correct answer in its candidate answers, i.e. $P(t_i \neq E)$. We may treat participating rate and consistency rate as two features and the estimation of truth existence score can be conducted in either semi-supervised or supervised method.

### Table 6: Parameter Setting of Prior Truth Distribution of $q_2$

<table>
<thead>
<tr>
<th>Candidate Answer</th>
<th>UNIFORM</th>
<th>VOTE</th>
<th>EXISTENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>1/4</td>
<td>2/15</td>
<td>3/6 · $p$</td>
</tr>
<tr>
<td>Pakistan</td>
<td>1/4</td>
<td>2/13</td>
<td>2/6 · $p$</td>
</tr>
<tr>
<td>Khost</td>
<td>1/4</td>
<td>1/13</td>
<td>1/6 · $p$</td>
</tr>
<tr>
<td>Empty</td>
<td>1/4</td>
<td>7/13</td>
<td>1 − $p$</td>
</tr>
</tbody>
</table>

However, in most real applications, the labeling information is not known in advance, or is expensive to obtain. Consequently, we introduce an unsupervised method to coarsely estimate the truth existence score. By using a Gaussian Mixture model (GMM), we can separate questions into two groups: has-truth cluster and no-truth cluster. We may use the posterior probability of each question belonging to the has-truth cluster as truth existence score. Each question is represented by PR and CR in the new features space. Intuitively, GMM tends to cluster questions into two groups centered at two peaks of the density function of new features. Thus, we can consider it as a "relative grouping". It means that the clustering result of one question is affected by the other questions, i.e. compared to other questions, how likely it is to be has-truth. To know which cluster is the has-truth cluster, we simply assume that the question whose product of participating rate and consistency rate is the largest belongs to the has-truth cluster.

For the non-empty candidate answers, we use VOTE to initialize the prior truth of each dimension, i.e. set the prior truth of this candidate answer proportionally to the number of occurrence. Table 6 shows the initialization of prior truth distribution by EXISTENCE, where $p$ is the estimated truth existence score.

EXISTENCE provides us an alternative way to initialize prior of truth. When most of sources are credible, we may just trust majority and use VOTE to initialize truth prior. When a large part of questions are no-truth, EXISTENCE can outperform other initialization schemes. Later we will see it with real-world datasets.

4.4.2 Smoothing Factor

One problem of using the posterior probability of GMM clustering is that it may be very small, i.e. close to 0, or very large, i.e. close to 1. In this case, the judgment of truth existence may be too bold. So we introduce a smoothing factor $\delta$ to compensate this judgment. We use $P(t_i \neq E) + \delta$ as truth existence score if $P(t_i \neq E) < 0.5 - \delta$, or $P(t_i \neq E) - \delta$ if $P(t_i \neq E) > 0.5 + \delta$.

Algorithm 1 presents the implementation of TEM.

5. EXPERIMENTS

In this section we demonstrate the effectiveness and efficiency of TEM on three real-world datasets. All the experiments are conducted on a laptop with 4 GB RAM, 1.4 GHz Intel Core i5 CPU, and OS X 10.9.4. Algorithms are implemented in Python 2.7.

5.1 Experiment Setup

We provide details on datasets and the experimental settings.

5.1.1 Datasets

SF2014[1] This dataset is from TAC Knowledge Base Population 2013 slot filling validation (SFV) track [12]. In this task, 18 slot-filling systems return the answers to a given set of questions about 100 entities. We select single-valued questions and manually map answers of different representations but the same semantic meaning into a single string. Note that this problem can be solved by synonym learning or co-reference algorithms and is independent of the truth discovery problem. After that, it consists of 774 questions in total, where 329 are has-truth questions. There are 3,913

Algorithm 1 EM Algorithm for TEM inference

Input: Answers $A$ for questions in $Q$ provided by sources in $S$
Output: Truths in $T$, source quality $\phi_S$
1: [Initialization]
2: for all $q_i \in Q$ do
3: initialize truth prior $\eta_{in}, n = 0, \ldots, N_i$
4: for all $s_j \in S$ do
5: initialize $(\phi_j^{(1)}(0)), (\phi_j^{(2)}(0)), (\phi_j^{(3)}(0))$
6: [EM Algorithm]
7: $k \leftarrow 0$
8: while not converge do
9: $k \leftarrow k + 1$
10: for all $q_i \in Q$ do
11: compute $\gamma_{in}^{(k)}$, $n = 0, \ldots, N_i$ {E-Step}
12: for $s_j \in S$ do
13: compute $(\phi_j^{(1)}(k)), (\phi_j^{(2)}(k)), (\phi_j^{(3)}(k))$ {M-Step}
14: [Compute answers]
15: for all $q_i \in Q$ do
16: $n \leftarrow \max_k \gamma_{in}^{(k)}$
17: if $n = 0$ then $t_i \leftarrow E$
18: else $t_i \leftarrow s_{in}$
19: return $T, \phi_S$

non-empty answers from 18 systems. Note that we generate empty answers to questions of a certain entity only when a source answers at least one question of it. It is a common practice used in truth discovery [21]. After generating 4,591 related empty answers, there are 8504 pieces of answers in all. Ground truths are evaluated by human accessors and provided by TAC.

SF2014 [4] This dataset is from TAC-KBP 2014 SFV track. After the same pre-processing, it consists of 406 questions in total with 160 has-truth questions. The 18 systems provides 2858 answers, in which 1590 are empty and 1268 are valid answers.

Flight This dataset is crawled from 38 flight websites from Dec 1, 2011 to Jan 3, 2012 [9]. For each flight, it contains the scheduled and actual departure time, arrival time, and actual departure and arrival gate. The dataset provides ground truths for 100 flights every day which are used in our experiments. We removed those trivial questions to which none of the sources gives answers. Finally, the dataset contains 2,909 flights with 17310 questions, and 341,732 non-empty answers provided by the 38 sources. There are 1,596 no-truth questions with 80,949 empty answers.

5.1.2 Evaluation Metrics

We introduce the measures used in our experiments to evaluate the estimation of truth and source quality. We use precision, recall and F1 to measure the performance of truth discovery algorithms.

- Precision (PREC) is the ratio of the number of correct non-empty answers to the number of non-empty answers that the model returns.
- Recall (REC) is the ratio of the number of correct non-empty answers to the number of non-empty answers in ground truth.
- $F1 = \frac{2 \cdot \text{PREC} \cdot \text{REC}}{\text{PREC} + \text{REC}}$ is the geometric mean of precision and recall.

For source quality, we use Mean Root Square Error (MRSE) to measure the difference between the estimated and true source quality. We compute the true source quality based on ground truth, with the estimation of source quality in Eq. 9-11 of Section 4.3.

$$\text{MRSE} = \sqrt{\frac{\sum_{i=1}^{N} \sum_{j=1}^{3} (\phi_j^{(i)})^2 - (\phi_j^{(i)})^2)^2}{3N}}$$

5.1.3 Baselines and Parameter Settings

Most truth discovery methods adopt an iterative framework to compute the quality of sources and answer credibility. We briefly introduce them as follows.

Vot For each question, we calculate the number of occurrences of each candidate answer provided by all sources and use the answer agreed by majority of sources as truth.

TruthFinder (Find) [19] For each non-empty candidate answer, its credibility is the probability that at least one associated source is true. The quality of source is the average of credibility of answers provided by this source.

AverageLog (Ave) [11] The credibility of an answer is the average quality of associated sources. Source quality is the average answer credibility weighted by the number of answers this source provides.

Investment (Inv), PooledInvestment (PInv) [11]. Each source uniformly invests its quality among the answers they provide, and its quality is a weighted sum of the credibility of those answers.

3Estimates (3Est) [5] They introduce a factor called difficulty of the question in answer credibility and source quality.

GuessLCA (LCA) [12] Each source has a probability to tell the truth, and a probability to guess among all the candidate answers. They use an EM algorithm to compute the answer credibility and honest probability. We choose this LCA model among the four variants because it performs consistently well on reported datasets.

LTM [21] They use a Bayesian model to incorporate two-sided source quality, i.e. sensitivity and specificity, and Gibbs Sampling to infer the truth. For a certain question, if no answer is considered true, then $E$ is returned. If multiple answers are true, we select the most possible one as the true answer.

EM [17] Each source has two quality measures similar to LTM. It adopts an EM algorithm to compute these measures and answer credibility. We choose truth in the same way as LTM.

LTM and EM naturally handle no-truth cases while other baselines cannot, so we extend other baselines to run on both original and augmented datasets with empty answers. We append suffix $E$ to the methods running on augmented datasets. To avoid randomness, we run LTM and EM 10 times and report the best results.

5.2 Experimental Results

5.2.1 Initialization of Prior Truth Distribution

In Section 4.4.1, we propose a novel method called EXISTENCE to initialize the prior of truth. Here, we run EXISTENCE on SF2013 dataset to see its effectiveness. Figure 2a shows the estimation on
truth existence of questions in SF2013 dataset. The red dots represent no-truth questions, while the blue dots represent has-truth ones. Clustering centers and variance are shown by ellipsoids. From the result we can see that the dots around clustering centers have high accuracy to be correctly labeled. Figure 2b shows the accuracy of EXISTENCE. The red dots suggest these no-truth questions are labeled correctly, denoted by TN. The blue dots suggest has-truth questions are labeled correctly, denoted by TP. The green dots represent has-truth questions are labeled as no-truth, denoted by FN, while magenta ones show reversely, denoted by FP. The numbers of the four cases are shown in the legend in Figure 2b and the overall accuracy is 0.80. EXISTENCE provides us an effective way to initialize prior of truth. Later we will show that the combination of TEM and EXISTENCE will outperform significantly than other combinations.

5.2.2 Truth Inference

We first examine the performance of truth inference. For the slot filling datasets, we use EXISTENCE to initialize the prior of truth, and for flight dataset, we use VOTE because of its high quality. Table 7 presents the inference results of all methods. It shows that our TEM model outperforms existing methods in terms of F1 score on all datasets. Baseline methods either return a small number of non-empty truths, which leads to high precision but low recall, e.g. FindE, AveE, or return non-empty truths to most of the questions, which results in high recall but low precision, e.g. Ave, Vot. On the contrary, our TEM model can better infer truth existence, so it can selectively provide non-empty truths to achieve both high precision, recall and the highest F1.

To further examine the effectiveness of TEM, we conduct Student’s paired t-test on TEM and PhnE whose average F1 is the highest among all baselines. We randomly split each dataset into 10 folds, leave out 1 fold each time to run TEM and PhnE and finally obtain 30 pairs of F1 score. Then we conduct the two-tailed test on the F1 scores of two algorithms. The value of t is 2.0452 that is larger than that when p = 0.05. Therefore, we conclude that TEM is better than other baselines with statistical significance.

We can see that the three real-world datasets have different proportion of no-truth questions. We define no-truth rate as the proportion of no-truth questions among all questions, which is an important factor to affect the performance of all the methods. On slot filling datasets where no-truth questions are prevalent with 57% no-truth rate on SF2013 and 56% for SF2014, our TEM model performs significantly better than all the baselines. State-of-the-art algorithms perform better when considering empty answers. This is natural because simply ignoring the empty answers will result in many no-truth questions being mistakenly answered. On flight dataset, although there are a small number of no-truth questions with only 10% no-truth rate, our TEM still achieves the highest F1 among all the methods. On this dataset, baseline methods perform better without considering empty answers due to lower no-truth rate. Note that none of the baselines perform consistently well on all datasets. This shows the strength and robustness of TEM on datasets with various no-truth rates.

To examine the effect of no-truth rate, we synthesize a set of datasets based on flight dataset. We randomly sample different number of has-truth questions and remove them to make the no-truth rate ranges from 0.1 to 0.5, and run all methods on these processed datasets. Figure 3 shows that the F1 score of all methods decreases as the no-truth rate gets larger because the removal of has-truth questions is equivalent to degrading the quality of sources. Most baselines get worse than or close to Vot when no-truth rate is 0.5. However, TEM is stable and consistently better than baseline methods. Even when half of the questions have no true answers, the F1 score of TEM is 0.70. This also proves the robustness of TEM on datasets with different no-truth rates.

5.2.3 Source Quality

Then we show the effectiveness of TEM on source quality estimation of SF2013 dataset. The estimated source quality measures are shown in Table 8. We can see that sources are different in three quality measures. SFV2013_12 has the highest false spoken rate with low silent rate and true spoken rate. This means that this aggressive source tends to give answers as many as possible while most of them are wrong. Based on the intuition we derive from Eq.7 and 8, if an answer is provided by this source, it increases the probability of this answer to be wrong for both has-truth and no-truth questions. SFV2013_14 has the highest true spoken rate and lowest false spoken rate. If it states an answer, it increases the probability of it to be correct due to its true spoken rate, and votes more to this question being has-truth because of its low false spoken rate. In all, our TEM model can represent the source quality in fine-grained measures, which helps to gain more accurate truth estimation than the state-of-the-arts methods.
When the algorithm converges, the final MRSE is 0.011, 0.013 and 0.021 on SF2013, SF2014 and flight dataset, respectively. It shows that TEM converges closely to the global optima.

5.2.4 Effectiveness of TEM with EXISTENCE

In Section 4.4.1 we propose a novel initialization method to set the prior of truth. To demonstrate the effectiveness of the combination of TEM and EXISTENCE, we run baseline methods in the following way. We first run EXISTENCE to have the estimation of has-truth questions. Then we run baseline methods on the estimated has-truth part. The evaluation metrics are the same. Table 9 shows the performance of our TEM model and the baselines.

We can see that with the estimation of has-truth questions by EXISTENCE, the performance of baseline methods is significantly improved on slot filling datasets compared to that in Table 7 which shows the power of our EXISTENCE initialization. However, our TEM still outperforms the modified baselines. It is because the misclustered questions by EXISTENCE will definitely lead to wrong answers in the baseline methods, while they could be corrected in the later stage when we iteratively estimate truths and source quality in TEM. On the other hand, the performance on flight dataset is very low. In this situation, EXISTENCE will mistakenly label many has-truth questions as no-truth.

To investigate the effect of different initializations of truth prior, we compare the performance of TEM with different initializations, i.e. EXISTENCE, UNIFORM and VOTE. As shown in Table 10 TEM with EXISTENCE is much better than that with the other two initializations on slot filling datasets. Because the quality of flight dataset is high with only 10% no-truth rate, VOTE is a better way to initialize truth prior. In all, EXISTENCE provides an effective scheme to set the prior of truth, and the combination of suitable initialization and TEM outperforms other baselines on all datasets.

5.2.5 Single FR v.s. Two FR

In Section 3 we make an assumption on the consistency of false spoken rate. Here we justify our assumption by experiments.

We first provide the false spoken rate on both has-truth and no-truth parts, i.e. FR and FR' defined in Section 3. Table 11 shows the false spoken rates of 18 sources in SF2013 dataset. The two false spoken rates are quite similar for all the sources. The minimum difference is only 0.01, and the maximum difference is 0.21. The overall difference, defined by \( \sqrt{\sum_{i=1}^{18} (FR_i - FR'_i)^2/18} \), is 0.089. It indicates that the false spoken rate on has-truth and no-truth part are consistent, and our assumption is reasonable.

In Table 12 we compare TEM model with the variation TEM2FR, which has two false spoken rate FR and FR' for has-truth questions and no-truth questions, respectively. On all datasets, TEM2FR has a higher precision because of better false spoken rate estimation on has-truth questions. However, it suffers from low recall due to inaccurate estimation of FR' on no-truth questions. The inaccurate FR' leads to a large number of truths estimated as empty, so TEM2FR has very low recall and hence low F1 score. This experiment shows that our TEM model is more robust than TEM2FR by reducing parameter size with the assumption of consistency.

5.2.6 Efficiency and Convergence

We compare the running time of all methods to show the efficiency. All algorithms except VOT are iterative. Thus, we fix the number of iterations to 100 and run all iterative algorithms for 10 times. The average running time per iteration is in Table 13.

TEM works faster than 3Est, LCA and LTM, comparable to other baselines and adaptive on datasets with various no-truth rates. Simple models like Find and Ave are faster, but they are not robust enough to consistently achieve good performance on all datasets. Other models, e.g. 3Est, LTM and LCA are much slower. 3Est and LTM consider negative claims, which increases the data size significantly. LCA is not efficient because it suffers from the SGD process to compute the source quality.

Figure 4 illustrates the F1 score change in each iteration for iterative models. LTM is a sampling-based algorithm whose number of iterations is user specified, so it is not considered here. We see that TEM converges within only 5 iterations. Find and Ave also converge fast, while 3Estimates needs about 30 iterations before convergence. Some baselines are not very stable. LCA does not converge linearly due to the randomness in SGD. Inv and Plnv update the score of truth by an exponential function, thus they may converge to a local optima with low performance, i.e. Plnv on SF2013. In summary, TEM converges fast with short running time per iteration, thus is efficient in terms of time complexity.

6. RELATED WORK

In truth discovery there exist some interesting studies handling different challenges. Yin et al. [19] are the first to formally introduce truth discovery and iteratively inferred truth and source qual-

---

### Table 8: Source Quality on SF2013 Data

<table>
<thead>
<tr>
<th>source</th>
<th>SFV2013_01</th>
<th>SFV2013_02</th>
<th>SFV2013_03</th>
<th>SFV2013_04</th>
<th>SFV2013_05</th>
<th>SFV2013_06</th>
<th>SFV2013_07</th>
<th>SFV2013_08</th>
<th>SFV2013_09</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREC</td>
<td>0.26</td>
<td>0.49</td>
<td>0.20</td>
<td>0.36</td>
<td>0.40</td>
<td>0.30</td>
<td>0.42</td>
<td>0.51</td>
<td>0.32</td>
</tr>
<tr>
<td>REC</td>
<td>0.56</td>
<td>0.13</td>
<td>0.61</td>
<td>0.20</td>
<td>0.08</td>
<td>0.17</td>
<td>0.11</td>
<td>0.13</td>
<td>0.16</td>
</tr>
<tr>
<td>F1</td>
<td>0.17</td>
<td>0.37</td>
<td>0.19</td>
<td>0.43</td>
<td>0.55</td>
<td>0.53</td>
<td>0.47</td>
<td>0.51</td>
<td>0.51</td>
</tr>
</tbody>
</table>

### Table 9: Methods with EXISTENCE Initialization

<table>
<thead>
<tr>
<th>Method</th>
<th>SF2013</th>
<th>SF2014</th>
<th>Flight</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEM</td>
<td>0.78</td>
<td>0.65</td>
<td>0.66</td>
</tr>
<tr>
<td>VOT</td>
<td>0.67</td>
<td>0.53</td>
<td>0.67</td>
</tr>
<tr>
<td>Find</td>
<td>0.67</td>
<td>0.54</td>
<td>0.73</td>
</tr>
<tr>
<td>Ave</td>
<td>0.68</td>
<td>0.54</td>
<td>0.73</td>
</tr>
<tr>
<td>Inv</td>
<td>0.64</td>
<td>0.36</td>
<td>0.77</td>
</tr>
<tr>
<td>Plnv</td>
<td>0.62</td>
<td>0.53</td>
<td>0.86</td>
</tr>
<tr>
<td>3Est</td>
<td>0.66</td>
<td>0.54</td>
<td>0.84</td>
</tr>
<tr>
<td>LCA</td>
<td>0.63</td>
<td>0.53</td>
<td>0.90</td>
</tr>
<tr>
<td>LTM</td>
<td>0.72</td>
<td>0.55</td>
<td>0.70</td>
</tr>
<tr>
<td>EM</td>
<td>0.53</td>
<td>0.53</td>
<td>0.69</td>
</tr>
</tbody>
</table>

### Table 10: TEM with Different Initializations

<table>
<thead>
<tr>
<th>Method</th>
<th>SF2013</th>
<th>SF2014</th>
<th>Flight</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXISTENCE</td>
<td>0.78</td>
<td>0.65</td>
<td>0.66</td>
</tr>
<tr>
<td>UNIFORM</td>
<td>0.86</td>
<td>0.67</td>
<td>0.69</td>
</tr>
<tr>
<td>VOTE</td>
<td>0.89</td>
<td>0.65</td>
<td>0.71</td>
</tr>
</tbody>
</table>

### Table 11: False Spoken Rate on Has-Truth and No-Truth Part

<table>
<thead>
<tr>
<th>source</th>
<th>Has</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>Has</td>
<td>No</td>
</tr>
<tr>
<td>SFV2013_01</td>
<td>0.18</td>
<td>0.23</td>
</tr>
<tr>
<td>SFV2013_02</td>
<td>0.15</td>
<td>0.20</td>
</tr>
<tr>
<td>SFV2013_03</td>
<td>0.26</td>
<td>0.40</td>
</tr>
<tr>
<td>SFV2013_04</td>
<td>0.39</td>
<td>0.44</td>
</tr>
<tr>
<td>SFV2013_05</td>
<td>0.19</td>
<td>0.20</td>
</tr>
<tr>
<td>SFV2013_06</td>
<td>0.31</td>
<td>0.41</td>
</tr>
<tr>
<td>SFV2013_07</td>
<td>0.18</td>
<td>0.22</td>
</tr>
<tr>
<td>SFV2013_08</td>
<td>0.19</td>
<td>0.22</td>
</tr>
<tr>
<td>SFV2013_09</td>
<td>0.21</td>
<td>0.31</td>
</tr>
</tbody>
</table>

### Table 12: Comparison of TEM with TEM2FR

<table>
<thead>
<tr>
<th>source</th>
<th>Has</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>Has</td>
<td>No</td>
</tr>
<tr>
<td>SFV2013_01</td>
<td>0.18</td>
<td>0.23</td>
</tr>
<tr>
<td>SFV2013_02</td>
<td>0.15</td>
<td>0.20</td>
</tr>
<tr>
<td>SFV2013_03</td>
<td>0.26</td>
<td>0.40</td>
</tr>
<tr>
<td>SFV2013_04</td>
<td>0.39</td>
<td>0.44</td>
</tr>
<tr>
<td>SFV2013_05</td>
<td>0.19</td>
<td>0.20</td>
</tr>
<tr>
<td>SFV2013_06</td>
<td>0.31</td>
<td>0.41</td>
</tr>
<tr>
<td>SFV2013_07</td>
<td>0.18</td>
<td>0.22</td>
</tr>
<tr>
<td>SFV2013_08</td>
<td>0.19</td>
<td>0.22</td>
</tr>
<tr>
<td>SFV2013_09</td>
<td>0.21</td>
<td>0.31</td>
</tr>
</tbody>
</table>

### Table 13: Efficiency and Convergence

<table>
<thead>
<tr>
<th>Method</th>
<th>SF2013</th>
<th>SF2014</th>
<th>Flight</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEM</td>
<td>0.78</td>
<td>0.65</td>
<td>0.66</td>
</tr>
<tr>
<td>VOT</td>
<td>0.67</td>
<td>0.53</td>
<td>0.67</td>
</tr>
<tr>
<td>Find</td>
<td>0.67</td>
<td>0.54</td>
<td>0.73</td>
</tr>
<tr>
<td>Ave</td>
<td>0.68</td>
<td>0.54</td>
<td>0.73</td>
</tr>
<tr>
<td>Inv</td>
<td>0.64</td>
<td>0.36</td>
<td>0.77</td>
</tr>
<tr>
<td>Plnv</td>
<td>0.62</td>
<td>0.53</td>
<td>0.86</td>
</tr>
<tr>
<td>3Est</td>
<td>0.66</td>
<td>0.54</td>
<td>0.84</td>
</tr>
<tr>
<td>LCA</td>
<td>0.63</td>
<td>0.53</td>
<td>0.90</td>
</tr>
<tr>
<td>LTM</td>
<td>0.72</td>
<td>0.55</td>
<td>0.70</td>
</tr>
<tr>
<td>EM</td>
<td>0.53</td>
<td>0.53</td>
<td>0.69</td>
</tr>
</tbody>
</table>
ity. Using integer programming, the framework proposed in [11] could incorporate common-sense constraints into iterations. However, these iterative models do not consider truth existence, thus gain either low precision or recall. On the other hand, recent work [17] used probabilistic model to estimate quality measures by bringing latent truth. However, these models do not differentiate empty answers from wrong answers, thus cannot describe fine-grained source quality. To our best knowledge, TEM is the first to model truth existence in truth discovery problem.

Some interesting studies focused on other aspects of truth discovery. Li et al. [7,8] proposed a framework to model multiple data types in a unified optimization model by defining different loss functions. Correlation between sources estimated by similarity between answers is considered in source quality to reduce the dependency problem [3]. Guo et al. [13] alleviated source dependency problem by revealing the latent group among sources in a probabilistic model. MTM [20] incorporated the credibility of evidence into truth discovery by discovering semantic rules. Vydiswaran et al. [15] used a retrieval-based approach to find relevant articles to the answers and propagated the trustworthiness between sources, evidences and answers. Recent work [10] discovered the trustworthiness of the authors of user-generated medical statements by exploiting linguistic cues and expert sources.

7. CONCLUSION AND FUTURE WORK

In this paper, we investigate the truth existence problem of truth discovery. We show that giving an answer to every question is not acceptable and the aggregation should be able to output empty as the final answer when truth does not exist. Moreover, when a source indeed gives empty as the answer to no-truth questions, we should reward this source, and vice versa. To model this important observation, we propose three source quality measures: silent rate, false spoken rate and true spoken rate. We propose a novel probabilistic model to incorporate these measures as sources generating the answer set given true answers. Also, we proposed effective initialization approaches to initialize the prior of truth. Extensive experiments on three real-world datasets clearly show the proposed model outperforms state-of-the-art truth discovery approaches.

Interesting future work includes solving the truth existence problem when the independence assumption between sources does not hold. When two sources are dependent, the answers they agree with should be discounted. The source dependence will affect the initialization of truth prior as well.

8. ACKNOWLEDGMENT

Research was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), NSF IIS-1017362, IIS-1312061, and IIS-1354329, HDRT/A10-10120, 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.

9. REFERENCES