Mining Collective Intelligence in Diverse Groups

Guo-Jun Qi[†], Charu C. Aggarwal[‡], Jiawei Han[†], Thomas Huang[†]

[†]University of Illinois at Urbana-Champaign
{qi4, hanj, t-huang1}@illinois.edu

[‡]IBM T.J. Watson Research Center
charu@us.ibm.com

ABSTRACT

Collective intelligence, which aggregates the shared information from large crowds, is often negatively impacted by unreliable information sources with the low quality data. This becomes a barrier to the effective use of collective intelligence in a variety of applications. In order to address this issue, we propose a probabilistic model to jointly assess the reliability of sources and find the true data. We observe that different sources are often not independent of each other. Instead, sources are prone to be mutually influenced, which makes them dependent when sharing information with each other. High dependency between sources makes collective intelligence vulnerable to the overuse of redundant (and possibly incorrect) information from the dependent sources. Thus, we reveal the latent group structure among dependent sources, and aggregate the information at the group level rather than from individual sources directly. This can prevent the collective intelligence from being inappropriately dominated by dependent sources. We will also explicitly reveal the reliability of groups, and minimize the negative impacts of unreliable groups. Experimental results on real-world data sets show the effectiveness of the proposed approach with respect to existing algorithms.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Data mining; Statistical databases

Keywords

Collective intelligence; Crowdsourcing; Robust classifier

1. INTRODUCTION

Collective intelligence aggregates contributions from multiple sources in order to collect data for a variety of tasks. For example, voluntary participants collaborate with each other to create a fairly extensive set of entries in *Wikipedia*, or a crowd of paid persons may perform image and news article annotations in *Amazon Mechanical Turk*. These crowdsourced tasks usually involve multiple *objects*, such as Wikipedia entries and images to be annotated. The participating sources collaborate to claim their own *observations*, such as facts and labels, on these objects. Our goal is to aggregate these collective observations to infer the *true values* (e.g., the true fact and image label) for the different objects [18, 14, 5].

We note that an important property of collective intelligence is that different sources are typically not independent of one another. For example, in the same social community, people often influence each other, where their judgments and opinions are not independent. In addition, task participants may obtain their data and knowledge from the same external information source, and their contributed information will be dependent. Thus, it may not be advisable to treat sources independently and directly aggregate the information from individual sources, when the aggregation process is clearly impacted by such dependencies. In this paper, we will infer the source dependency by revealing latent group structures among involved sources. Dependent sources will be grouped, and their reliability is analyzed at the group level. The incorporation of such dependency analysis in group structures can reduce the risk of overusing the observations made by the dependent sources in the same group, especially when these observations are unreliable. This helps prevent dependent sources from inappropriately dominating collective intelligence especially when these sources are not reliable.

Moreover, we note that groups are not equally reliable, and they may provide incorrect observations which conflict with each other, either unintentionally or maliciously. Thus, it is important to reveal the reliability of each group, and minimize the negative impact of the unreliable groups. For this purpose, we study the *general* reliability of each group, as well as its *specific* reliability on each individual object. These two types of reliability are closely related. General reliability measures the overall performance of a group by aggregating each individual reliability over the entire set of objects. On the other hand, although each object-specific reliability is distinct, it can be better estimated with a prior that a *generally reliable* group is likely to be reliable on an individual object and vice versa. Such prior can reduce the overfitting risk of estimating each object-specific reliability, especially considering that we need to determine the true value of each object at the same time [11, 1].

The remainder of this paper is organized as follows. We review the related work in Section 2. Our problem and notations are formally defined in Section 3. The probabilistic model for the problem is developed in Section 4, followed by a running example that illustrates the impact of group dependency on the model in Section 5. Section 6 presents the model inference and parameter estimation algorithms. Then Section 7 presents the application of the developed model to training classifiers from noisy crowdsourced data. We evaluate the model in Section 8 on real data sets, and summarize the paper with the conclusion in Section 9.

2. RELATED WORK

Aggregating crowdsourced knowledge and information has attracted a lot of research efforts, and yields many insightful discoveries. For example, [16] proposed an iterative truth finder algorithm by simultaneously accessing the trustworthiness of each source

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

and the correctness of claimed facts. [1] developed a probabilistic graphical model by jointly modeling the abilities of participants and the correct answers to questions in an aptitude testing setting. The work in [18] developed a latent truth model to infer the source quality and correct claims by modeling two types of false positive and false negative errors of each source. All of these algorithms estimate the performances of data sources and the impacts on the credibility of their claimed facts.

However, sources are not independent of each other in real world. Instead, their contributions are typically dependent. [16] noted this problem and used a dampening factor to compensate for excessively high confidence due to the copied content between sources. But this method did not explicitly model the dependency between sources, and how the dampening factor can reduce the dependency effect is not clear. On the other hand, [4] studied the relation between the content claimed by sources, and developed a separate weighted voting algorithm by considering the copied content between each other. However, the accuracies are accessed *independently* on the source level, which can make the accuracy of a data source overestimated if many other dependent sources repeat the same false facts.

Moreover, existing models [4, 2, 9, 6] only consider the pairwise relations between sources to their dependency, which completely ignores the higher-order dependency among sources. In contrast, we explicitly group the dependent sources to capture arbitrary orders of dependency among sources. We find that high-order dependency prevails in many real cases, and it is more effective to model them directly rather than decomposing them into separate pairwise relations. For example, sources which obtain the content from the same resource will be assigned to the same group to reflect the high order dependency among them. This yields a more compact representation to jointly assess the reliability of data sources and the correctness of the claimed facts. Moreover, we will see based on the group-level dependency, independent sources from different groups will play more important role than dependent ones in the same group in inferring the true facts. This is a desired property which can properly aggregate collective knowledge in many real world tasks.

Modeling the group dependency can be analogized to the community discovery in social networks. Community structure has been considered as a more effective data structure to capture the social relations among people than the links between pairs of persons [7]. With the similar spirit, the groups can also be more effective than pairwise dependency, and provide deeper insight into the property of high-order dependency among sources and how such property affects the aggregation of collective knowledge. However, it is worth pointing out that the groups defined in our model differ from the communities [3] in social networks. Communities are usually defined as a set of people densely linked in social networks. However, two linked people may not necessarily be influenced by one another when they report the facts and knowledge. Two close friends can express different opinions and claim conflicting truths. Therefore, we will directly investigate the data contributed by sources to find the group structure characterizing their mutual dependency that directly affects the source reliability in our collective intelligence model.

Finally, our model is motivated to explore the objective facts and knowledge. This is in contrast to the inference of individual's preference, which aims to recommend products and services based on user's ratings and opinions [12]. Instead, in this paper we aim at aggregation of collective knowledge to automatically extract the true facts, such as correct answers to questions and true categories for web pages, which do not depend on the variability of user's subjectivity.

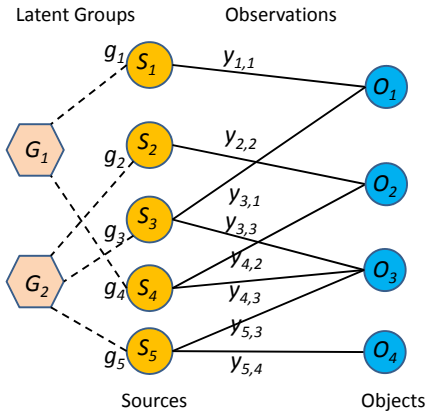


Figure 1: An example illustrating a set of five sources with their observations on four objects.

3. PROBLEM DEFINITIONS

We formally define the following Multi-Source Sensing (MSS) model which abstracts the description of collective intelligence. Suppose that we have a set $\mathcal{S} := \{S_1, S_2, \cdots, S_N\}$ of N sources, and a set $\mathcal{O} := \{O_1, O_2, \cdots, O_M\}$ of M objects. Each object O_m takes a value t_m from a domain $\mathcal{X}_m$ which describes one of its attributes. Each source S_n in $\mathcal{S}$ reports its observation $y_{n,m} \in \mathcal{X}_m$ on an object O_m. Then the goal of the *MSS* model is to infer the true value t_m of each object O_m from the observations made by sources. We introduce some notations, which will be used consistently in this paper. We will use n, m, l and k in the subscript to index sources, objects, groups and values in an object domain, respectively. The variables y, t, u and r denote the observations, true values, group reliability and object-specific reliability respectively.

In this paper, we are particularly interested in categorical domain $\mathcal{X}_m = \{1, \cdots, K_m\}$ with discrete values. For example, in many crowdsourcing applications, we focus on the (binary-valued) assertion correctness in hypothesis test and (multi-valued) categories in classification problem. However, the *MSS* model can be extended to continuous domain with some effort by adopting the corresponding continuous distributions. Due to the space limitation, we leave this extension in the full version of this paper.

Figure 1 illustrates an example, where five sources make their observations on four objects. An object can be an image or a biological molecule, and an annotator or a biochemical expert (as a source) may claim the category (as the value) for each object. Alternatively, an object can be a book, and a book seller web site (as a source) claims the identity of its authors (as the values). In a broader sense, objects are even not concrete objects. They can refer to any crowdsourced tasks, such as questions (e.g., "is Peter a musician?") and assertions (e.g., "George Washington was born on February 22, 1732." and "an animal is present in an image,"), and the observations by sources are the answers to the questions, or binary-valued positive or negative claims on these assertions.

It is worth noting that each source does not need to claim the observations on all objects in $\mathcal{O}$. In many tasks, sources make claims only on small subsets of objects of interest. Thus, for notational convenience, we denote all claimed observations by $\mathbf{y}$ in bold, and use $I = \{(n,m)|\exists\ y_{n,m} \in \mathbf{y}\}$ to denote all the indices in $\mathbf{y}$. We use the notations $I_{n,\cdot} = \{m|\exists\ (n,m) \in I\}$ and

$I_{\cdot,m} = \{n|\exists\ (n,m) \in I\}$ to denote the subset of indices that are consistent with the corresponding subscripts n and m.

Meanwhile, in order to model the dependency among sources, we assume that there are a set of latent groups $\{G_1, G_2, \cdots\}$, and each source S_n is assigned to one group G_{g_n} where $g_n \in \{1, 2, \cdots\}$ is a random variable indicating its membership. For example, as illustrated in Figure 1, the five sources are inherently drawn from two latent groups, where each source is linked to the corresponding group by dotted lines. Each latent group contains a set of sources which are influenced by each other and tend to make similar observations on objects. The unseen variables of group membership will be inferred mathematically from the underlying observations. Here, we do not assume any prior knowledge on the number of groups. The composition of these latent groups will be determined with the use of a Bayesian nonparametric approach by stick-breaking construction [15], as to be presented in the next section.

To minimize the negative impact of unreliable groups, we will explicitly model the group-level reliability. Specifically, for each group G_l, we define a group reliability score $u_l \in [0, 1]$ in unit interval. This value measures the general reliability of the group over the entire set of objects. A higher value of u_l indicates the greater reliability of the group.

Meanwhile, we also specify the reliability $r_{l,m} \in \{0, 1\}$ of each group G_l on each particular object O_m. When $r_{l,m} = 1$, group G_l will have reliable performance on O_m, and otherwise it will be unreliable. The reason that we distinguish between reliability u_l and object-specific reliability $r_{l,m}$ is as follows. While a generally reliable group with a larger value of u_l, provides very useful evidence about the members of the group on a generic basis, there are likely to be natural variations within the group itself. Thus, in our model, a group reliability u_l only measures how likely it will be reliable on object set, and whether it will have a reliable performance on a particular object is given by $r_{l,m}$. In the next section, we will clarify the relationship between general reliability u_l and object-specific reliability $r_{l,m}$.

4. MULTI-SOURCE SENSING MODEL

In this section, we present a generative process for the multi-source sensing problem. The output of this model will contain the following three aspects: (1) the group membership of sources which describes their dependency when claiming their observations on a set of objects. (2) the reliability u_l associated with each group and its specific reliability $r_{l,m}$ on each object. (3) the true values t_m for each object. Our goal is to reveal the connections between these three aspects, especially how the collective observations made by sources can be explained by the latent groups and their reliability in a unified probabilistic framework.

First we define the following generative model for multi-source sensing (MSS) process below, the details of which will be explained shortly.

1. Draw $\boldsymbol{\lambda} \sim$ GEM(κ) (i.e., stick breaking construction with concentration κ).
2. For each source S_n,
   2.1. Draw its group assignment $g_n|\boldsymbol{\lambda} \sim$ Discrete($\boldsymbol{\lambda}$);
3. For each object O_m,
   3.1. Draw its true value $t_m \sim$ Uniform($\mathcal{X}_m$);
4. For each group G_l:
   4.1. Draw its group reliability $u_l \sim$ Beta(b_1, b_0);

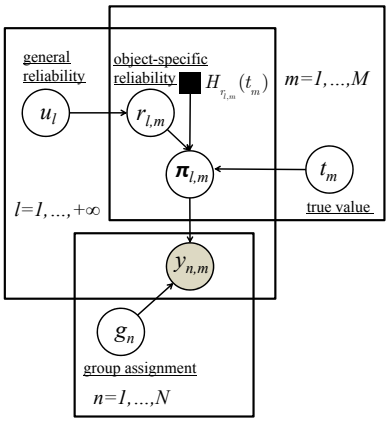


Figure 2: The graphical model for multi-source sensing. The three plates represent group reliability u_l with $l = 1, 2, \cdots$, the true values t_m for each object O_m with $m = 1, \cdots, M$, and the group assignment g_n of each source with $n = 1, \cdots, N$, respectively.

5. For each pair of group G_l and object O_m:
   5.1. Draw reliability indicator $r_{l,m} \sim$ Bernoulli(u_l);
   5.2. Draw the observation model parameter
   $$\boldsymbol{\pi}_{l,m}|r_{l,m}, t_m = z \sim H_{r_{l,m}}(t_m)$$
   for group G_l on object O_m;
6. For each $(n, m) \in I$:
   6.1. Draw observation $y_{n,m}|\boldsymbol{\pi}_{l,m}, g_n \sim F(\boldsymbol{\pi}_{g_n,m})$;

Here, $g_n|\boldsymbol{\lambda} \sim$ Discrete($\boldsymbol{\lambda}$) denotes a discrete distribution, which generates the value $g_n = l$ with probability λ_l; H and F are a pair of conjugate distributions which are determined by the type of data values on objects. For categorical values, these are Dirichlet and Multinomial distributions, respectively. Figure 2 illustrates the generative process in a graphical representation. We will explain the details later.

In Step 1, we adopt the stick-breaking construction GEM(κ) (named after Griffiths, Engen and McCloskey) with concentration parameter $\kappa \in \mathbb{R}^+$ to define the prior distribution of assigning each source S_n to a latent group G_{g_n} [15]. Specifically, in GEM(κ), a set of random variables $\boldsymbol{\rho} = \{\rho_1, \rho_2, \cdots\}$ are independently drawn from the Beta distribution $\rho_i \sim$ Beta$(1, \kappa)$. They define the mixing weights $\boldsymbol{\lambda}$ of the group membership component such that $p(g_n = l|\boldsymbol{\rho}) = \lambda_l = \rho_l \prod_{i=1}^{l-1}(1 - \rho_i)$. By the aforementioned stick-breaking process, we do not need the prior knowledge of the number of groups. This number will be determined by capturing the degree of dependency between sources.

Clearly, we can see that the parameter κ in the above GEM construction plays the vital role of determining *a priori* the degree of dependency between sources. According to the GEM construction, we can verify that the probability of two sources S_n and S_m being assigned to the same group is given by the following:

$$\begin{aligned} P(g_n = g_m) &= \sum_{l=1}^{+\infty} \mathop{\mathbb{E}}_{\boldsymbol{\lambda}} P(g_n = l|\boldsymbol{\lambda}) P(g_m = l|\boldsymbol{\lambda}) \\ &= \sum_{l=1}^{+\infty} \mathop{\mathbb{E}}_{\lambda_l} \lambda_l^2 = \sum_{l=1}^{+\infty} \frac{2}{(1+\kappa)(2+\kappa)} \left(\frac{\kappa}{2+\kappa}\right)^{l-1} = \frac{1}{1+\kappa} \end{aligned} \tag{1}$$

It is evident that when κ is smaller, sources are more likely to be assigned to the same group where they are dependent and share the

same observation model. This will yield higher degree of dependency between sources. As κ increases, the probability that any two sources belong to the same group will decrease. In the extreme case, as $\kappa \to +\infty$, this probability approaches zero. In this case, all sources will be assigned to distinctive groups, yielding complete independence between sources. This shows that the model can flexibly capture the various degrees of dependency between sources by setting an appropriate value of κ.

In Step 3, we adopt the uniform distribution as the prior on the true value t_m of each object over its domain $\mathcal{X}_m$. The uniform distribution sets an unbiased prior so that true values will be completely determined a posteriori given observations in the model inference. In Section 7, we will show how to set a more informative prior when more knowledge about objects is available.

In Step 4, we define a Beta distribution $\text{Beta}(b_1, b_0)$ on the group reliability score u_l, where b_1 and b_0 are the soft counts which specify whether a group is reliable or not a priori, respectively. Then, in Step 5.1, object-specific reliability $r_{l,m} \in \{0, 1\}$ is sampled from the Bernoulli distribution $\text{Bern}(u_l)$ to specify the group reliability on a particular object O_m. The higher the general reliability u_l, the more likely G_l is reliable on a particular object O_m with $r_{l,m}$ being sampled to be 1. This suggests that a generally more reliable group is more likely to be reliable on a particular object. In this sense, the general reliability serves as a prior to reduce the over-fitting risk of estimating object-specific reliability in the *MSS* model.

In Step 5.2, the model parameter $\boldsymbol{\pi}_{l,m}$ for each group on a particular object is drawn from the conjugate prior $H_{r_{l,m}}(t_m)$, which depends on the true value t_m and the object-specific group reliability $r_{l,m}$. Then, given the group membership g_n, each source S_n generates its observation $y_{n,m}$ according to the corresponding group observation model $F(\boldsymbol{\pi}_{g_n,m})$ in Step 6. In the next subsection, we will detail the specification of $H_{r_{l,m}}(t_m)$ and $F(\boldsymbol{\pi}_{l,m})$ in categorical domain.

4.1 Group Observation Models

In this subsection, we discuss the specification of group observation distribution $F(\boldsymbol{\pi}_{l,m})$ and its conjugate distribution $H_{r_{l,m}}(t_m)$ for categorical values on each object. Here the group observation model on each object depends on two factors: (1) the specific reliability $r_{l,m}$ on this object, which aims to reveal the differences between reliable and unreliable observations on an object, and (2) the true value t_m for the object.

It is worth noting that although we distinguish each group observation into reliable and unreliable cases in this subsection, it does not mean that two groups are enough to capture the source dependency. These two cases are used to model the performance at the *object* level. However, given more objects, there are many possible combinations of these two cases on different objects. This is why we need more groups to capture the source dependency based on their observations on different objects. In the following, we will discuss the group observations models on each object.

In categorical domains, for each group, we choose the multinomial distribution as its observation model to generate each observation $y_{n,m}$ for its member sources on each object O_m. Thus, Step 6 in the generative process of *MSS* model becomes the following:

$$y_{n,m}|\boldsymbol{\pi}_{l,m}, g_n \sim F(\boldsymbol{\pi}_{g_n,m}) \triangleq \text{Multinomial}(\boldsymbol{\pi}_{g_n,m})$$

where $\boldsymbol{\pi}_{l,m}$ is the parameter of multinomial distribution for group G_l on object O_m. Here, all member sources in the same group share the same observation model to capture their dependency.

The model parameter $\boldsymbol{\pi}_{l,m}$ is generated by the following:

$$\begin{aligned}&\boldsymbol{\pi}_{l,m}|r_{l,m}, t_m = z \sim H_{r_{l,m}}(t_m)\\ &\triangleq \text{Dir}(\underbrace{\theta^{(r_{l,m})}, \cdots,}_{z-1} \underset{\substack{\downarrow\\ z^{\text{th}}\text{ entry}}}{\eta^{(r_{l,m})}}, \cdots, \theta^{(r_{l,m})})\end{aligned}$$

where Dir denotes Dirchlet distribution, and $\theta^{(r_{l,m})}$ and $\eta^{(r_{l,m})}$ are its soft counts for sampling the false and true values under different settings of $r_{l,m}$.

If group G_l has reliable observations for object O_m (i.e., $r_{l,m} = 1$), it should be more likely to sample the true value $t_m = z$ as its observation than sampling any other false value. Thus, we should set a larger value for $\eta^{(r_{l,m})}$ than for $\theta^{(r_{l,m})}$.

On the other hand, if group G_l has unreliable observations for object O_m, i.e., $r_{l,m} = 0$, it should *not* be more likely to claim the true value for the object than claiming the false values. Therefore, the group observation model should have $\eta^{(0)}$ no larger than $\theta^{(0)}$, i.e., $\eta^{(0)} \leq \theta^{(0)}$. Specifically, the mathematical model can distinguish between *uninformative* and *malicious* observations on the target object:

I. **Uninformative observation:** When $\eta^{(0)} = \theta^{(0)}$, sources in group G_l make uninformative observations on object O_m, since false values are equally likely to be claimed as the true value. This can be caused when these sources either carelessly claim their observations at random, or lack the knowledge about the target object.

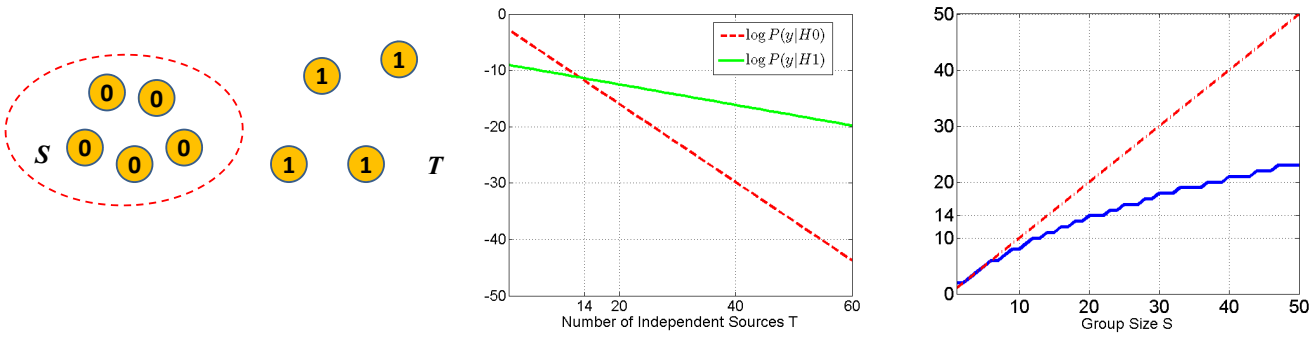
II. **Malicious observation:** When $\eta^{(0)} < \theta^{(0)}$, it suggests that the group G_l contains malicious sources which tend to claim false values for object O_m. Compared with uninformative observations, these malicious observations can even provide us with some information about the target object by interpreting the observations in a reverse manner. Actually, with $\theta^{(0)} > \eta^{(0)}$, the model gives the unclaimed observation larger weight to be evaluated as the true value.

In summary, depending on $r_{l,m}$, the sources in group G_l make either reliable (when $r_{l,m} = 1$) or unreliable (when $r_{l,m} = 0$) observations on a particular object O_m. Accordingly, the corresponding parameters $\eta^{(r_{l,m})}$ and $\theta^{(r_{l,m})}$ are constrained in different ways. When $r_{l,m} = 1$, we impose a strict inequality $\eta^{(1)} > \theta^{(1)}$ to enforce that group G_l is more likely to claim the true value. On the contrary, when $r_{l,m} = 0$, we have $\theta^{(0)} \geq \eta^{(0)}$, representing that G_l will be unreliable in terms of claiming the true value for O_m. In Section 6, we will see how these parameters can be estimated by maximizing the observation likelihood of the *MSS* model subject to these constraints.

By putting together these different pieces, the *MSS* defines a complete distribution

$$\begin{aligned}&p(\mathbf{y}, \mathbf{g}, \mathbf{r}, \mathbf{u}, \mathbf{t}, \boldsymbol{\pi}|\Theta) = \prod_{m=1}^{M} p(t_m) \prod_{l=1,m=1}^{L,M} p(u_l|b_1, b_0)p(r_{l,m}|u_l)\\ &\times p(\pi_{l,m}|r_{l,m}, t_m, \eta^{(r_{l,m})}, \theta^{(r_{l,m})})\\ &\times \prod_{n=1}^{N} p(g_n|\kappa) \prod_{(n,m)\in I} p(y_{n,m}|g_n, \pi_{g_n,m})\end{aligned}$$

over $\mathbf{g} = \{g_n\}$, $\mathbf{r} = \{r_{l,m}\}$, $\mathbf{u} = \{u_l\}$, $\mathbf{t} = \{t_m\}$, $\boldsymbol{\pi} = \{\boldsymbol{\pi}_{l,m}\}$ and the source observations $\mathbf{y}$ with model parameters $\Theta = \{\eta^{(0)}, \theta^{(0)}, \eta^{(1)}, \theta^{(1)}, b_1, b_0, \kappa\}$. In Section 6, we will present how to infer (1) the true values t_m for each object, (2) group assignment g_n of each source, and (3) the general reliability u_l of each group and its specific reliability $r_{l,m}$ on each object from the *MSS* model a posteriori given the observations $\mathbf{y}$.



(a) An running example

(b) Likelihoods of two hypotheses

(c) Minimal number of independent sources to overturn the claims by S dependent sources (solid blue curve).

Figure 3: (a) A running example with S dependent sources in the same group and T independent sources. (b) Comparison of the likelihoods of two hypotheses (in Y-axis) versus varying number T of independent sources (in X-axis). The number of dependent sources in the group is fixed to $S = 20$. (c) The minimal number of independent sources (in Y-axis) to overturn the claims made by varying number of dependent sources (in X-axis). The results are obtained with $\eta^{(1)} = 10, \theta^{(1)} = 5$, and $\eta^{(0)} = \theta^{(0)} = 10$.

4.2 Multiple Attributes

In some cases, an object might have multiple attributes. There are many such examples as follows.

- A person can have many attributes. For example, she/he has a hobby of playing piano and takes "software engineer" as her/his vocation. We can consider hobby and vocation as two attributes for each person, and define their values on two different domain sets such as {playing piano, hiking, swimming, traveling $\cdots$} and {software engineer, stock trader, university faculty, $\cdots$} in MSS model, respectively.

- An image can be labeled as "tiger" as well as "forest". We can consider the presence of these two nonexclusive labels as two different attributes, and their values are boolean {Present, Not Present} for an image. In this way, we can allow an image has multiple labels simultaneously.

- A movie can have multiple actors/actresses. We can treat each actor/actress as an attribute, and use a binary value {1,0} to denote whether an actor/actress participates in a particular movie or not.

We can see in these examples, our *MSS* model is much flexible to handle multiple attributes associated with each object. Moreover, we note that different attributes often correlate with each other. For example, image labels "tiger" and "forest" often co-occur in an image, and some actors/actresses may tend to co-star a movie. Exploring these attributes together can improve the accuracy of inferring their true values.

5. DEPENDENCE VS. INDEPENDENCE: A RUNNING EXAMPLE

In this section, we show a running example that demonstrates how group reliability structure captures the dependency between sources when it infers the true value for an object. In Figure 3(a), we show a group of S sources and T independent sources. We consider an ideal case where the S sources in the group make an unanimous claim of the value 0 for an object, while the T independent sources unanimously claims the opposite value 1 for the same object. While the dependent sources in the group and the independent sources claim the different values in this example, we can investigate different values of information contributed by these sources. Especially, we wonder whether independent sources play more important roles than dependent ones in finding the true value for each object in the *MSS* model.

For this purpose, we test the following two hypotheses:

- $H0$: The true value for the object is 0, versus

- $H1$: The true value for the object is 1.

To decide which hypothesis is true, we compare the observation likelihoods given these two hypotheses in the *MSS* model. Figure 3(b) compares the two likelihoods with varying number T of independent sources. The number of dependent sources is fixed to $S = 20$. We can see with more than $T = 14$ independent sources, $H1$ has a larger likelihood than $H0$. In this case, the claims made by independent sources become more credible than that made by dependent sources. This example shows fewer independent sources can overturn the claim made by more dependent sources. This suggests that each dependent source contains less information about the true claim as compared with each independent source.

To make this point more clear, Figure 3(c) illustrates the minimum number of independent sources to ensure $p(\mathbf{y}|H1) > p(\mathbf{y}|H0)$ under varying number of dependent sources S in the group. We can see that usually fewer independent sources is needed to have its claim accepted compared with the same number of dependent sources. This shows that independent sources are more valuable than dependent sources in determining the true value for each object. This is a desired property in our model, since we would like to de-emphasize the excessive impacts of dependent sources in a group.

Of courses, in the real world, sources may not be ideally split into dependent ones in a group, and completely independent ones. The independent sources may not make unanimous claims as in this case. However, this intuitive running example explains how

the dependency encoded in group structure will affect the inference of true value on an object, and illustrates the independent claims are generally more valuable than dependent claims in the *MSS* model.

6. MODEL INFERENCE AND PARAMETER ESTIMATION

In this section, we present the inference and learning processes. We wish to infer the tractable posterior $p(\mathbf{g}, \mathbf{r}, \mathbf{u}, \mathbf{t}, \boldsymbol{\pi}|\mathbf{y})$ with a parametric family of variational distributions in the factorized form:

$$q(\mathbf{g}, \mathbf{r}, \mathbf{u}, \mathbf{t}, \boldsymbol{\pi}) = \prod_n q(g_n|\boldsymbol{\varphi}_n) \prod_{l,m} q(r_{l,m}|\boldsymbol{\tau}_{l,m}) \prod_l q(u_l|\boldsymbol{\beta}_l) \prod_m q(t_m|\boldsymbol{\nu_m}) \prod_{l,m} q(\boldsymbol{\pi}_{l,m}|\boldsymbol{\alpha}_{l,m})$$

with parameters $\boldsymbol{\varphi}_n$, $\boldsymbol{\tau}_{l,m}$, $\boldsymbol{\beta}_l$, $\boldsymbol{\nu}_m$ and $\boldsymbol{\alpha}_{l,m}$ for these factors. The distribution and the parameter for each factor can be determined by the variational approach [10]. Specifically, we aim to maximize the lower bound of the log likelihood $\log p(\mathbf{y})$, i.e.,

$$\log p(\mathbf{y}) \geq \mathop{\mathbb{E}}_q \log p(\mathbf{g}, \mathbf{r}, \mathbf{u}, \mathbf{t}, \boldsymbol{\pi}, \mathbf{y}) - \mathop{\mathbb{E}}_q (\log q(\mathbf{g}, \mathbf{r}, \mathbf{u}, \mathbf{t}, \boldsymbol{\pi})) \triangleq \mathcal{L}(q)$$

This can obtain the optimal factorized distribution. The lower bound can be maximized over one factor while the others are fixed. This is an approach which is similar to coordinate descent. In each iteration, all the factors are updated sequentially over steps by finding the fixed-point solutions until convergence. The details of these updating steps are provided in Appendix A.

We analyze the computational complexity in one loop of updating all factors. Suppose that we are given N sources, M objects, and obtain L groups by the stick-breaking construction. We also denote by $K_{\max}$ the maximum size of the domain sets among all objects. Then by investigating the updating steps in Appendix A, we can find that the computational complexity is $O(NMLK_{\max})$ for one loop.

On the other hand, the model parameters Θ can be estimated by maximizing the observation likelihood. This can be done by the EM algorithm:

E-Step: Given the current parameters in Θ, apply variational inference to obtain the factorization q and their variational parameters;

M-Step: Given the factorization q, maximize the lower bound $\mathcal{L}(q)$ of the log-likelihood and obtain a new model parameter Θ. (Details of this Maximization step are given in Appendix B.)

These two steps are iterated until convergence. We obtain the variational approximation and the maximum likelihood parameter estimation results simultaneously.

7. CLASSIFICATION PROBLEMS

We are often of particular interest in the classification problem where each object takes a class as its value from a K-class domain $\mathcal{X} = \{1, 2, \cdots, K\}$. Moreover, we might be able to access the feature representations for the objects in $\mathcal{O}$. For example, if the objects are genetic sequences or text documents, we can extract their feature descriptors to describe the genetic structure and document content. Therefore, we wish to impose a more informative prior that aggregates these features into the prior distribution. For this purpose, given a feature vector $\mathbf{x}_m$ for an object, the prior on t_m becomes a conditional distribution on $\mathbf{x}_m$. For greater modeling flexibility, we choose a distribution for this prior. For example, we can choose an exponential distribution $p(t_m|\mathbf{x}_m, W)$:

$$\text{Exp}(W) := p(t_m|\mathbf{x}_m, W) = \frac{1}{Z} \exp\left\{\sum_{k=1}^{K} \delta\left[\!\left[t_m = k\right]\!\right] \langle \mathbf{w}_k, \mathbf{x}\rangle\right\} \tag{2}$$

where each coefficient vector is taken from the parameters $W = \{\mathbf{w}_k | k \in \mathcal{X}\}$, $\langle \mathbf{w}_k, \mathbf{x}\rangle$ denotes the inner product between two vectors, and Z is the normalization factor to ensure that the above exponential distribution integrates to unit value.

Accordingly, the model inference in Step A.4 in Appendix A should be changed. Each updated factor $q(t_m)$ in model inference becomes an exponential distribution:

$$q(t_m|\boldsymbol{\nu}_m) := \exp\{\sum_{k=1}^{K} \delta\left[\!\left[t_m = k\right]\!\right] \nu_{m;k}\} \tag{3}$$

with the parameter $\boldsymbol{\nu}_m$ defined as follows:

$$\nu_{m;k} = \langle \mathbf{w}_k, \mathbf{x}\rangle + \sum_l \sum_{r_l} q(r_l)\{(\eta^{(r_l)} - 1) \times \mathop{\mathbb{E}}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;k} + \sum_{k' \neq k} (\theta^{(r_l)} - 1) \mathop{\mathbb{E}}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;k'}\}$$

The other updating steps for the model inference in Appendix A stay the same.

Besides the inference, we need to learn the parameter W in $p(t_m|\mathbf{x}_m, W)$. Here, we adopt the variational EM (Expectation-Maximization) algorithm. In each iteration, the E-step (expectation) involves computing the tractable posterior distributions as in the inference step. Then, the maximization step will update W by maximizing the expected log-likelihood over q as follows:

$$\max_W \sum_{m=1}^{M} \mathbb{E}_{q(t_m|\boldsymbol{\nu}_m)} \log p(t_m|\mathbf{x_m}, W) \tag{4}$$

We can adopt any off-the-shelf optimization algorithms to solve the above problem.

The learned parameterized model $p(t_m|\mathbf{x}, W)$, as a byproduct, is a classifier conditional on the input feature vector $\mathbf{x}$. This provides us with a way to train a robust classification model with the noisy crowdsourced labels, compared with typical classifiers trained with the clean labels. On the other hand, the learned classifier enhances the *MSS* model by providing a more discriminative prior of the labeling information on objects through their feature representations. This regularizes the true classes of objects in the feature space, especially when the classes claimed by different sources on an object are too scarce or too inconsistent to make robust estimation of the true classes. In this case, the imposed prior plays a nontrivial role in determining the true class of the object.

8. EXPERIMENTAL RESULTS

In this section, we compare our approach with other existing algorithms and demonstrate its effectiveness for inferring source reliability together with the true values of objects. The comparison is performed on a book author data set from online book stores, and a user tagging data set from the online image sharing web site `Flickr.com`.

8.1 Online Book Store Data Set

The first data set is the book author data set prepared in [16]. The data set is obtained by crawling $1,263$ computer science books on *AbeBooks.com*. For each book, *AbeBooks.com* returns the book information extracted from a set of online book stores. This data set

contains a total of 877 book stores (sources), and $24,364$ listings of books (objects) and their author lists (object values) reported by these book stores. Note that each book has a different categorical domain that contains all the authors claimed by sources. Our goal is to predict the true authors for each book.

Author names are normalized by preserving the first and last names, and ignoring the middle name of each author. For evaluation purposes, the authors of 100 books are manually collected from scanned book covers [16]. We compare the returned results of each model with the ground truth author lists on this test set and report the accuracy.

We compare the proposed algorithm *MSS* with the following baselines: (1) the naive voting algorithm which counts the top voted author list for each book as the truth; (2) *TruthFinder* [16]; (3) *Accu* [4] which considers the dependency between sources; (4) *2-Estimates* as described in [5] with the highest accuracy among all the models in [5].

Table 3 compares the results of the different algorithms on the book author data set in terms of the accuracy. The *MSS* model achieves the best accuracy among all the compared models. We note that the proposed *MSS* model is an unsupervised algorithm which does not involve any training data. In other words, we do not use any true values in the *MSS* algorithm in order to produce the reliability ranking as well as other true values. Even compared with the accuracy of 0.91 of the Semi-Supervised Truth Finder (SSTF) [17] using extra training data with known true values on some objects, the *MSS* model still achieves the highest accuracy of 0.95.

Figure 4(a) illustrates the scatter plot between the predicted reliability u_l for each group and its test accuracy. From this figure, it is evident that the group reliability obtained from the *MSS* model is a good predictor of the true accuracy for each group. Meanwhile, we also report three example groups in Table 1. It is evident that within each group, the member sources have much consistent reliability as they make dependent claims. Therefore, by accurately predicting reliability of groups, the proposed *MSS* model can appropriately aggregate the contributions from differen groups based on their performances and gain the competitive accuracy as shown above.

Moreover, to compare the reliability between sources, we can define the reliability of each source S_n by the expected reliability score of its assigned groups as follows:

$$\text{Reliability}(S_n) = \sum_l q(g_n = l) \mathop{\mathbb{E}}_{q(u_l|\boldsymbol{\beta}_l)}[u_l]$$

where

$$\mathop{\mathbb{E}}_{q(u_l|\boldsymbol{\beta}_l)}[u_l] = \frac{\beta_{l,1}}{\beta_{l,1} + \beta_{l,2}}$$

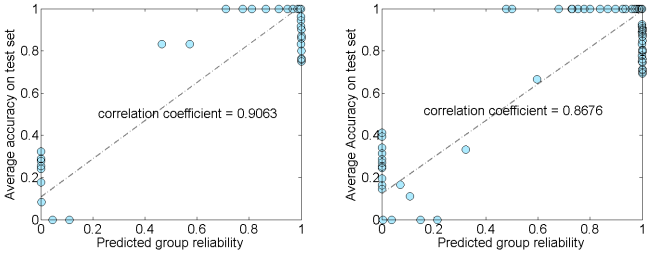
Then, sources can be ranked based on such source reliability. In Table 2, we rank the top-10 and bottom-10 book stores in this way.

In order to show the extent to which this ranking list is consistent with the real source reliability, we provide the accuracy of these bookstores on test data sets. Note that each individual bookstore may only claim on a subset of books in the test set, and the accuracy is computed based on the claimed books. From the table, we can see that the obtained rank of data sources is consistent with the rank of their accuracies on the test set. On the contrary, the accuracy of the bottom-10 bookstores is much worse compared to that of the top-10 book stores on the test set. This also partially explains the better performance of the *MSS* model.

Since κ influences the dependency modeling between sources, we study the sensitivity of the model accuracy versus κ in Figure 3. We know that when $\kappa = 0$, all sources are completely dependent, and assigned to the same group. At this point, the model has a much lower accuracy, since all sources are tied to the same level of reliability within a single group. As κ increases, the accuracy achieves the peak at $\kappa = 5.0$. After that point, it deteriorates as the model gradually stops capturing the source dependency with increased κ. This demonstrates the importance of modeling the source dependency, and the capability of the *MSS* model in capturing such dependencies with κ.

Table 4: The rounds used before convergence and computing time for each model.

Model	Bookstore		User Tagging	
	Rounds	Time(s)	Rounds	Time (s)
Voting	1	0.2	1	0.5
2-Estimates	29	21.2	32	628.1
TruthFinder	8	11.6	11	435.0
Accu	22	185.8	23	3339.7
MSS	9	10.3	12	366.2



(a) Book author data set (b) Flickr data set

Figure 4: Scatter plots on two data sets. The horizontal axis represents the predicted group reliability by u_l and the vertical axis represents the average accuracy of the member sources on the test set. The slope of each red line in the scatter is the correlation coefficient which shows the statistical correlation between u_l and the average accuracy.

8.2 Flickr Image Tagging Data Set

We also evaluate the algorithm on a user tagging data set from an online image sharing web site *Flickr.com*. This data set contains $13,528$ users (data sources) who annotate $36,280$ images (data objects) with their own tags. We consider 12 tags - "balloon," "bird," "box," "car," "cat," "child," "dog," "flower," "snow leopard," "waterfall," "guitar," "pumpkin" for evaluation purposes. Each tag is associated with a binary value 1/0 to represent its presence or not in an image. This forms a multi-attribute model with these 12 tags to find whether they are present on each image as described in Section 4.2. Different from the book author data set, we apply the extended classification model in Section 7, where the visual content of each image is represented by a $8,000$ dimensional hierarchical gaussian [19] feature vector.

Figure 4 illustrates some image examples in this data set and the tags annotated by users. It is evident that some images are wrongly tagged by users. The *MSS* model aims to correct these errors and yield accurate annotations on these images. To test accuracy, we manually annotate these 12 tags on a subset of $1,816$ images.

We follow the same experimental setup as on the book author data set. For the sake of fair comparison, we adopt the variants in [8] to incorporate visual features to enhance the original algorithms for comparison by inferring the true values based on object clusters in

Table 1: Three example groups among all 33 groups discovered by the *MSS* model on book author data set. The parenthesis after the name of each bookstore is its accuracy on test set.

Group I	Group II	Group III
FREE U.S. AIR SHIPPING (0.3750)	The Book Depository (0.3043)	DVD Legacy (0.5833)
TheBookCom (0.3556)	textbookxdotcom (0.4444)	Englishbookservice.com (0.5500)
Browns Books (0.3438)	Caiman (0.3855)	Henry's Biz Books (0.6000)
Mellon's Books (0.4000)	Bobs Books (0.4615)	Blackwell Online (0.6579)
	Books Down Under (0.4750)	Morgenstundt Buch & Kunst (0.6207)
	Limelight Bookshop (0.3896)	
	Powell's Books (0.3810)	

Table 2: Top-10 and bottom-10 book stores ranked by their posterior probability of belonging to a reliable group. We also report the accuracy of these bookstores on the test set.

top-10 bookstore	accuracy	bottom-10 bookstore	accuracy
International Books	1	textbooksNow	0.0476
happybook	1	Gunter Koppon	0.225
eCampus.com	0.9375	www.textbooksrus.com	0.3333
COBU GmbH & Co. KG	0.875	Gunars Store	0.2308
HTBOOK	1	Indoo.com	0.3846
AlphaCraze.com	0.8462	Bobs Books	0.4615
Cobain LLC	1	OPOE-ABE Books	0
Book Lovers USA	0.8667	The Book Depository	0.3043
Versandantiquariat Robert A. Mueller	0.8158	Limelight Bookshop	0.3896
THESAINTBOOKSTORE	0.8214	textbookxdotcom	0.4444

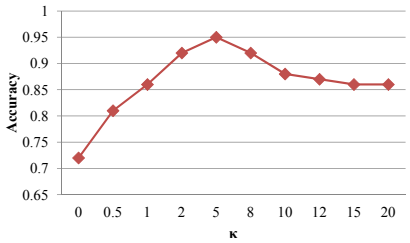


Figure 5: Parametric Sensitivity: model accuracy versus different κ on book author data set.

(a) balloon (b) snow leopard

(c) guitar (d) pumpkin

Figure 6: Examples of image and the associated user tags in Flickr data set. In each subfigure the left image is correctly tagged by users, while the right one is wrongly tagged.

the feature space. It has shown better accuracy compared with the original algorithms [8]. Table 3 shows the average precision and recall on the 12 tags by the compared algorithms. We can see that *MSS* still performs the best among these compared algorithms. The Figure 4(b) illustrates the scatter plot between the predicted reliability of each group and the average accuracy of its member sources on the test set. It is evident that the obtained group reliability is still a good predictor of the true accuracy with strong correlation coefficient 0.8676. This guarantees a competitive performance of the *MSS* model on this *Flickr* data set as on the book author data set.

We also compare the computational time used by different algorithms in Table 4. The experiments are conducted on a personal computer with Intel Core i7-2600 3.40 GHz CPU, 8 GB physical memory and Windows 7 operating system. We can see that compared with most of other algorithms, *MSS* model can converge in fewer rounds with less computational cost.

9. CONCLUSION

In this paper, we propose an integrated true value inference and group reliability approach. Dependent sources which are grouped together, and their (general and specific) reliability is assessed at the group level. The true data values are extracted from the reliable groups so that the risk of overusing the observations from dependent sources can be minimized. The overall approach is described by a probabilistic multi-source sensing model, based on which we jointly infer group reliability as well as the true values for objects *a posterior* given the observations from sources. The key to the success of this model is to capture the dependency between sources, and aggregate the collective knowledge at the group granularity. We present experimental results on two real data sets, which demonstrate the effectiveness of the proposed model over other existing algorithms.

Acknowledgements

Research was sponsored by the Army Research Laboratory and National Science Foundation and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 and Grant IIS-1144111. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is au-

Table 3: Comparison of different algorithms on book author and Flickr data set. On book author data set, the algorithms are compared by their accuracies. On Flickr data set, the algorithms are compared by their average precisions and recalls on 12 tags.

Model	book author data set	Flickr data set	
	accuracy	precision	recall
Voting[4]	0.71	0.8499	0.8511
2-Estimates[5]	0.73	0.8545	0.8602
TruthFinder[17]	0.83	0.8637	0.8649
Accu[4]	0.87	0.8731	0.8743
MSS	**0.95**	**0.9176**	**0.9212**

thorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. The first author was also in part supported by an IBM Fellowship and National Natural Science Foundation of China under Grant 61272214.

Appendix A: Model Inference

In this Appendix, we derive the variational inference for the proposed *MSS* model, and give the detail steps to update the variational parameters in each factor.

A.1: Update each factor $q(\boldsymbol{\pi}_{l,m}|\boldsymbol{\alpha}_{l,m})$ for the group observation parameter $\boldsymbol{\pi}_{l,m}$.

By variational approach, we can verify that the optimal $q(\boldsymbol{\pi}_{l,m}|\boldsymbol{\alpha}_{l,m})$ has the form

$$\begin{aligned} q(\boldsymbol{\pi}_{l,m}|\boldsymbol{\alpha}_{l,m}) &\propto \exp\{\mathop{\mathbb{E}}_{q(\boldsymbol{r}_{l,m}),q(t_m)} \ln p(\boldsymbol{\pi}_{l,m}|r_{l,m},t_m) \\ &\quad + \sum_{n\in I_{\cdot,m}} \mathop{\mathbb{E}}_{q(\boldsymbol{g}_n)} \ln p(y_{n,m}|\boldsymbol{\pi}_{l,m},g_n)\} \\ &\propto \prod_{k\in\mathcal{X}} {\pi_{l,m;k}}^{\alpha_{l,m;k}-1} \end{aligned}$$

It still has Dirichlet distribution with the parameters

$$\begin{aligned} \alpha_{l,m;k} = &\sum_{n\in I_{\cdot,m}} q(g_n=l)\delta\left[\!\left[y_{n,m}=k\right]\!\right] \\ &+ \sum_{r_{l,m}\in\{0,1\}} q(r_{l,m})[(\eta^{(r_{l,m})}-1)q(t_m=k) \\ &+ (\theta^{(r_{l,m})}-1)(1-q(t_m=k))] + 1 \end{aligned}$$

for each $k\in\mathcal{X}_m$, where $\delta\left[\!\left[A\right]\!\right]$ is the indicator function which outputs 1 if A holds, and 0 otherwise. Here we index the element in $\boldsymbol{\alpha}_{l,m}$ and $\boldsymbol{\pi}_{l,m}$ by k after the colon. We will follow this notation convention to index the element in vectors in this paper.

A.2: Update each factor $q(u_l|\boldsymbol{\beta}_l)$ for general group reliability u_l.

We have

$$\begin{aligned} \ln q(u_l|\boldsymbol{\beta}_l) &\propto \sum_m \mathop{\mathbb{E}}_{q(r_{l,m})} \ln p(r_{l,m}|u_l) + \ln p(u_l|b_1,b_0) \\ &= (\sum_m q_1(r_{l,m}) + b_1 - 1)\ln u_l \\ &\quad + (\sum_m q_0(r_{l,m}) + b_0 - 1)\ln(1-u_l) \end{aligned}$$

where $q_i(r_{l,m})$ is short for $q(r_{l,m}=i)$ for $i=0,1$, respectively. It is evident the posterior of u_l still has Beta distribution as Beta($\boldsymbol{\beta}_l$) with parameter

$$\boldsymbol{\beta}_l = [\sum_m q_1(r_{l,m}) + b_1, \sum_m q_0(r_{l,m}) + b_0].$$

It is evident that the above updated parameter sums up the posterior reliability $q_1(r_{l,m})$ and $q_0(r_{l,m})$ over all objects. This corresponds to the intuition that the general reliability is the sum of the reliability on individual objects.

A.3: Update each factor $q(r_{l,m}|\boldsymbol{\tau}_{l,m})$ for the object-specific reliability $r_{l,m}$ of group G_l on O_m:

$$\begin{aligned} \ln q(r_{l,m}|\boldsymbol{\tau}_{l,m}) \propto &\mathop{\mathbb{E}}_{q(t_m),q(\boldsymbol{\pi}_{l,m})} \ln p(\boldsymbol{\pi}_{l,m}|r_{l,m},t_m) \\ &+ \mathop{\mathbb{E}}_{q(u_l)} \ln p(r_{l,m}|u_l) \end{aligned} \tag{5}$$

Thus, we have

$$\begin{aligned} &\ln q(r_{l,m}|\boldsymbol{\tau}_{l,m}) \\ &\propto \sum_{k\in\mathcal{X}_m} q(t_m=k)[(\eta^{(r_{l,m})}-1)\mathop{\mathbb{E}}_{q(\boldsymbol{\pi}_{l,m})}\ln\pi_{l,m;k} \\ &+ (\theta^{(r_{l,m})}-1)\sum_{j\neq k}\mathop{\mathbb{E}}_{q(\boldsymbol{\pi}_{l,m})}\ln\pi_{l,m;j}] \\ &+ r_{l,m}\mathop{\mathbb{E}}_{q(u_l)}\ln u_l + (1-r_{l,m})\mathop{\mathbb{E}}_{q(u_l)}\ln(1-u_l) \end{aligned} \tag{6}$$

for $r_{l,m}\in\{0,1\}$, respectively. Here we compute the expectation of the logarithmic Dirichlet variable as

$$\mathop{\mathbb{E}}_{q(\boldsymbol{\pi}_{l,m})}\ln\pi_{l,m;k} = \psi(\alpha_{l,m;k}) - \psi(\sum_i \alpha_{l,m;i})$$

with the digamma function $\psi(\cdot)$; the expectation of the logarithmic Beta variables

$$\mathbb{E}_{q(u_l)}\ln u_l = \psi(\beta_{l;1}) - \psi(\beta_{l;1}+\beta_{l;2})$$

and

$$\mathbb{E}_{q(u_l)}\ln(1-u_l) = \psi(\beta_{l;2}) - \psi(\beta_{l;1}+\beta_{l;2}).$$

Finally, the updated values of $q(r_{l,m})$ are normalized to be valid probabilities.

The last line of Eq. (6) reflects how the general reliability u_l affects the estimation of the object-specific reliability. This embodies the idea that a generally reliable group is likely to be reliable on a particular object and vice versa. This can reduce the overfitting risk of estimating $r_{l,m}$ especially considering that $q(t_m)$ in the second line also needs to be estimated simultaneously in the *MSS* model as in the next step.

A.4: Update each factor $q(t_m|\boldsymbol{\nu_m})$ for the true value.

We have

$$\begin{aligned} &\ln q(t_m=k|\boldsymbol{\nu_m}) \propto \ln p(t_m=k) \\ &+ \sum_l \sum_{r_{l,m}\in\{0,1\}} q(r_{l,m}) \mathop{\mathbb{E}}_{q(\boldsymbol{\pi}_{l,m})} \ln p(\boldsymbol{\pi}_{l,m}|t_m=k,r_{l,m}) \end{aligned}$$

This suggests that

$$\ln q(t_m = k|\boldsymbol{\nu_m}) \propto \sum_l \sum_{r_{l,m}} q(r_{l,m})\{(\eta^{(r_{l,m})} - 1) \mathop{\mathbb{E}}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;k} + \sum_{k' \neq k} (\theta^{(r_{l,m})} - 1) \mathop{\mathbb{E}}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;k'}\}$$

All $q(t_m = k), k \in \mathcal{X}_m$ are normalized to ensure they are validate probabilities.

A.5: Update each factor $q(g_n|\boldsymbol{\varphi}_n)$ for the group assignment of each source.

We can derive

$$\ln q(g_n = l|\boldsymbol{\varphi}_n) \propto \mathop{\mathbb{E}}_{q(\boldsymbol{\rho})} \ln p(g_n = l|\boldsymbol{\rho}) + \sum_{m \in I_{n,\cdot}} \mathop{\mathbb{E}}_{q(\boldsymbol{\pi}_{l,m})} \ln p(y_{n,m}|\pi_{l,m}, g_n = l) = \mathop{\mathbb{E}}_{q(\boldsymbol{\rho})} \ln p(g_n = l|\boldsymbol{\rho}) + \sum_{m \in I_{n,\cdot}} \mathop{\mathbb{E}}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;y_{n,m}}$$

This shows that $q(g_n = l|\boldsymbol{\varphi}_n)$ is a multinomial distribution with its parameter as

$$\varphi_{n;l} = q(g_n = l|\boldsymbol{\varphi}_n) = \frac{\exp(U_{n,l})}{\sum_{l=1}^{\infty} \exp(U_{n,l})} \tag{7}$$

where

$$U_{n,l} = \mathop{\mathbb{E}}_{q(\boldsymbol{\rho})} \ln p(g_n = l|\boldsymbol{\rho}) + \sum_{m \in I_{n,\cdot}} \mathop{\mathbb{E}}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;y_{n,m}}$$

As in [13], we truncate after L groups: the posterior distribution $q(\rho_i)$ after the level L is set to be its prior $p(\rho_i)$ from Beta$(1, \kappa)$; and all the expectations $\mathop{\mathbb{E}}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;k}$ after L are set to:

$$\mathbb{E}_{q(\boldsymbol{\pi}_{l,m})} \ln \pi_{l,m;k} = \mathop{\mathbb{E}}_{q(t_m), p(r_{l,m})} \{\ \mathbb{E}[\ln \pi_{l,m;k}|r_{l,m}, t_m]\ \}$$

with the prior distribution $p(r_{l,m})$ defined as Section 4 for all $l > L$, respectively. The inner conditional expectation in the above is taken with respect to the probability of $\boldsymbol{\pi}_{l,m}$ conditional on $r_{l,m}$ and t_m. Similar to the family of nested Dirichlet process mixture in [13], this will form a family of nested priors indexed by L for the *MSS* model. Thus, we can compute the infinite sum in the denominator of Eq. (7) as:

$$\sum_{l=L+1}^{\infty} \exp(U_{n,l}) = \frac{\exp(U_{n,L+1})}{1 - \exp(\mathop{\mathbb{E}}_{\rho_i \sim \text{Beta}(1,\kappa)} \ln(1 - \rho_i))}$$

A.6: Update $q(\rho_i)$ in GEM construction.

Before the truncation level L, the posterior distribution $q(\rho_i) \sim$ Beta$(\phi_{i,1}, \phi_{i,2})$ is updated as

$$\phi_{i,1} = 1 + \sum_{n=1}^{N} q(g_n = i), \quad \phi_{i,2} = \kappa + \sum_{n=1}^{N} \sum_{j=i+1}^{\infty} q(g_n = j)$$

Appendix B: Parameter Estimation

The model parameters $\Theta = \{\eta^{(0)}, \theta^{(0)}, \eta^{(1)}, \theta^{(1)}, b_1, b_0, \kappa\}$ can be estimated by maximizing the log-likelihood $\log \mathcal{L}(q)$ by the obtained factorization q with the constraints $\eta^{(1)} > \theta^{(1)}$ and $\eta^{(0)} \leq \theta^{(0)}$. Since we require $\eta^{(1)} > \theta^{(1)}$ *strictly* holds, we usually impose $\eta^{(1)} \geq (1+\epsilon)\theta^{(1)}$ with a positive value of ϵ, i.e., $\eta^{(1)}$ is larger than $\theta^{(1)}$ with a margin ϵ. This ensures the strict inequality and improves numerical stability. In the algorithm, we set $\epsilon = 0.5$. Then, the parameter estimation problem becomes the following:

$$\Theta^\star = \arg\max_{\Theta} \mathcal{L}(q) \quad s.t., 0 \leq \eta^{(0)} \leq \theta^{(0)}, \eta^{(1)} \geq (1+\varepsilon)\theta^{(1)} \geq 0, \quad b_1, b_0, \kappa \geq 0$$

This constrained optimization problem can be solved by many off-the-shelf gradient-based constrained optimization solvers with the following gradients:

$$\frac{\partial \mathcal{L}}{\partial \eta^{(r)}} = \sum_{l,m,k \in \mathcal{X}_m} \{\psi(\eta^{(r)} + (K_m - 1)\theta^{(r)}) - \psi(\eta^{(r)}) + \psi(\alpha_{l,m;k}) - \psi(\sum_i \alpha_{l,m;i})\}$$

$$\frac{\partial \mathcal{L}}{\partial \theta^{(r)}} = \sum_{k \in \mathcal{X}_m} \{\psi(\eta^{(r)} + (K_m - 1)\,\theta^{(r)}) - (K_m - 1)\psi(\theta^{(r)}) + \sum_{k'} \psi(\alpha_{l,m;k'}) - (K_m - 1)\psi(\sum_i \alpha_{l,m;i})\}$$

for $r \in \{0, 1\}$.

$$\frac{\partial \mathcal{L}}{\partial b_1} = \sum_l \psi(b_1 + b_0) - \psi(b_1) + \psi(\beta_{l,1}) - \psi(\beta_{l,1} + \beta_{l,2})$$

$$\frac{\partial \mathcal{L}}{\partial b_0} = \sum_l \psi(b_1 + b_0) - \psi(b_0) + \psi(\beta_{l,2}) - \psi(\beta_{l,1} + \beta_{l,2})$$

$$\frac{\partial \mathcal{L}}{\partial \kappa} = \sum_i \psi(1 + \kappa) - \psi(\kappa) + \psi(\phi_{i,1} + \phi_{i,2}) - \psi(\phi_{i,2})$$

10. REFERENCES

[1] Y. Bachrach, T. Minka, J. Guiver, and T. Graepel. How to grade a test without knowing the answers - a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proc. of International Conference on Machine Learning*, 2012.

[2] M. Bilgic, G. Namata, and L. Getoor. Combining collective classification and link prediction. In *Workshop on Mining Graphs and Complex Structures (at ICDM)*, 2007.

[3] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[4] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. In *Proc. of International Conference on Very Large Databases*, August 2009.

[5] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *Proc. of ACM International Conference on Web Search and Data Mining*, February 2010.

[6] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of link structure. *Journal of Machine Learning Research*, (3):679–707, 2002.

[7] M. Girvan and M. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June 2002.

[8] M. Gupta, Y. Sun, and J. Han. Trust analysis with clustering. In *Proc. of International World Wide Web Conference*, April 2011.

[9] O. Hassanzadeh and et al. A framework for semantic link discovery over relational data. In *CIKM*, 2009.

[10] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[11] G. Kasneci, J. V. Gael, D. Stern, and T. Graepel. Cobayes: Bayesian knowledge corroboration with assessors of unknown areas of expertise. In *Proc. of ACM International Conference on Web Search and Data Mining*, 2011.

[12] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009.

[13] K. Kurihara, M. Welling, and N. Vlassis. Accelerated variational dirichlet process mixtures. In *NIPS*, 2006.

[14] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *Proc. of International Conference on Computational Linguistics*, August 2010.

[15] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[16] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proc. of ACM SIGKDD conference on Knowledge Discovery and Data Mining*, August 2007.

[17] X. Yin and W. Tan. Semi-supervised truth discovery. In *Proc. of International World Wide Web Conference*, March 28-April 1 2011.

[18] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. In *Proc. of International Conference on Very Large Databases*, 2012.

[19] X. Zhou, N. Cui, Z. Li, F. Liang, and T. Huang. Hierarchical gaussianization for image classification, 2009.