Learning from Crowds in the Presence of Schools of Thought

Yuandong Tian
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
yuandong@andrew.cmu.edu

Jun Zhu
Department of Comp. Sci. & Tech.
Tsinghua University
Beijing, 10084, China
dcszj@mail.tsinghua.edu.cn

ABSTRACT

Crowdsourcing has recently become popular among machine learning researchers and social scientists as an effective way to collect large-scale experimental data from distributed workers. To extract useful information from the cheap but potentially unreliable answers to tasks, a key problem is to identify reliable workers as well as unambiguous tasks. Although for objective tasks that have one correct answer per task, previous works can estimate worker reliability and task clarity based on the single gold standard assumption, for tasks that are subjective and accept multiple reasonable answers that workers may be grouped into, a phenomenon called *schools of thought*, existing models cannot be trivially applied. In this work, we present a statistical model to estimate worker reliability and task clarity without resorting to the single gold standard assumption. This is instantiated by explicitly characterizing the grouping behavior to form schools of thought with a rank-1 factorization of a worker-task group-size matrix. Instead of performing an intermediate inference step, which can be expensive and unstable, we present an algorithm to analytically compute the sizes of different groups. We perform extensive empirical studies on real data collected from Amazon Mechanical Turk. Our method discovers the schools of thought, shows reasonable estimation of worker reliability and task clarity, and is robust to hyperparameter changes. Furthermore, our estimated worker reliability can be used to improve the gold standard prediction for objective tasks.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Models - Statistical

General Terms

Algorithms, Experimentation

Keywords

Crowdsourcing, Schools of Thought, Pattern Analysis

KDD'12, August 12–16, 2012, Beijing, China.

1. INTRODUCTION

Crowdsourcing has emerged as an effective way to collect large-scale data to help solve challenging scientific and engineering problems. In web services such as Amazon Mechanical Turk (M-Turk)[1], human intelligence tasks (HITs) (e.g., "Does the image contain a car?") are distributed from the requestor to an *unknown* set of workers, called *crowds*, who are paid with a low cost to fulfill them. There are two kinds of applications for crowdsourcing. One application is to obtain correct labels of a dataset, used in computer vision [21], natural language processing [19], etc. In such a scenario, task is objective with one correct answer, called *gold standard*. The goal is to recover it from the noisy worker responses. The other application is to use crowdsourcing for qualitative user studies [11], demographic survey [16] or solving a hard problem [1]. In this case, each task may have multiple valid answers, called *schools of thought*, since the tasks are subjective or can be misinterpreted, and the workers come from a variety of cultural and educational background [16].

Due to the open and anonymous nature of crowdsourcing, the quality of the collected data is not guaranteed. In order to use these data well, in both scenarios we need to address two common key problems – "how to identify a small number of unreliable workers whose answers may be random or even adversary" and "how to identify tasks that may cause confusion to workers". Formally, we call these two factors worker *reliability* and task *clarity*. In crowdsourcing applications that aim to obtain ground truth labels of a dataset, only the labels from reliable workers should be trusted, and ambiguous tasks should be redesigned to remove any misunderstanding in the future. In applications that aim for user study or look for multiple opinions, one needs to distinguish whether a worker has a reasonable opinion, or just puts random answers that may ruin the data distribution.

In the former case, many previous works [15, 23, 19, 14, 10] have been presented and shown promising results compared to the "majority voting" heuristic. The worker reliability is defined either as the degree of concentration (precision) [15, 23, 22] or confusion matrix [9, 18, 5] referencing the estimated gold standard, whose existence is an essential assumption in these works. Some works also model task difficulty [22], again based on the existence of gold standard. Computationally, an iterative approach is usually adopted to estimate the gold standard and worker reliability simultaneously. The rationale is that knowing the gold standard helps to identify reliable workers and workers' reliability is useful to weigh their answers to recover the gold standard.

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

However, in the latter case where *more than one* answers could be valid and reasonable, defining worker reliability on top of a single gold standard is no longer a good idea. For some workers, their reliability can be underestimated only because they support a reasonable idea that is not the estimated single gold standard, yet in fact they may follow a unique but reasonable thinking and should be respected. On the other hand, an unambiguous task that is supposed to have alternative answers may also be estimated as confusing.

To deal with this problem, in this paper, we directly model worker *reliability* and task *clarity* without the help of gold standard. As a result, this model works in both scenarios of crowdsourcing applications. Our model is built on the following two mild assumptions on the grouping behavior that happens in schools of thought: 1) reliable workers tend to agree with other workers in many tasks; and 2) the answers to a clear task tend to form tight clusters. Following this idea, we develop a low-rank computational model to explicitly relate the grouping behavior of schools of thought, characterized by *group sizes*, to worker reliability and task clarity. To bypass the hard model selection problem of determining the unknown number of clusters (i.e., schools), we apply nonparametric Bayesian clustering techniques, which have shown great promise in statistics and machine learning [3, 6, 24]. Moreover, instead of performing a potentially expensive and unstable intermediate inference step, which is necessary for all the previous works [15, 23, 21, 22], we derive an analytical form to estimate the expected group sizes. The analytic form only depends on pairwise distances between answers, making it generalizable to different answer types. Interestingly, our model could provide a generative interpretation of latent distance model for social networks [8]. The worker reliability and task clarity are thus obtained via the rank-1 factorization of the expected group size matrix.

Different from most previous works that focus on gold standard estimation, recent work [21] also models the schools of thought for binary queries by assigning each worker a (different) linear classifier on hidden multidimensional representations of tasks, with Gaussian priors on both representations and classifiers. The worker reliability is thus defined as the precision of the linear model. However, it remains a question whether such linear representations and Gaussian priors for both tasks and workers are faithful to the data, and whether heterogeneous tasks can be represented in the same space. Our work avoids such representation issues by explicitly modeling the grouping structure of the data. This leads to fewer assumptions and parameters. Moreover, our method allows an analytic solution, while [21] uses a coordinate descent procedure to obtain a local optimum.

Finally, we apply our method to both simulation data and the real data collected from M-Turk. In the real data, we discover the group structure of schools of thought (Fig. 1), and show estimated worker reliability and task clarity, as well as a comparison with several benchmarks and sensitivity analysis. For objective tasks, we use the estimated worker reliability to select high quality workers for recovering gold standard, which outperforms previous works [15, 21].

The paper is structured as follows. Section 2 introduces our statistical model for crowdsourcing data analysis in the presence of schools of thought, together with a simple algorithm. Section 3 presents synthetic validation, and Section 4 presents analytical results on M-Turk. Finally, Section 5 concludes with future directions discussed.

2. A LOW-RANK SCHOOLS OF THOUGHT MODEL

In this section, we present a computational model to estimate worker reliability and task clarity for crowdsourcing data in the presence of schools of thought.

2.1 Basic Assumptions

In crowdsourcing, let N be the number of workers and K be the number of tasks. Each worker is required to finish all the K tasks (See the experimental design for more details). We use a scalar $\lambda_i > 0$ to model the *constant* reliability of worker i among different tasks, and a scalar $\mu_k > 0$ to model the *constant* degree of clarity of task k that holds for all workers[2].

Like any useful statistical models, we need to make appropriate assumptions in order to perform meaningful estimation and discover useful patterns. Specifically, for our problem of estimating worker reliability and task clarity in the presence of schools of thought, it suffices to make the following two mild assumptions on the behavior of workers and tasks:

1. *A worker i who is consistent with many other workers in most of the tasks is reliable, i.e. λ_i is large.*

2. *A task k whose answers form a few tight clusters is easy, well-addressed and objective (large μ_k); while a task whose answers form lots of small clusters is complicated, confusing and subjective (small μ_k).*

We argue that the above two assumptions are reasonable for characterizing crowdsourcing data. The first assumption may fail if all the workers collaborate to cheat, or the task is too hard so that most of the workers are misled towards the same wrong answer. However, since the basic idea of crowdsourcing is to ask the crowds for useful information, it is not restrictive to trust the behavior of the majority of workers. Furthermore, most previous works (e.g., majority voting, [15] and [22]) in crowdsourcing assuming the existence of gold standards implicitly make the first assumption, which is shown in both their initialization steps and their model designs.

The second assumption can be interpreted as "sense-making" that people make efforts to find interpretations from "experience" (their answers) [17]. Reliable workers are expected to use more mental resource to obtain a reasonable answer, while unreliable workers may give random nonsense answers. A sensible task only contain a few reasonable answers but random answers could be many. Thus reliable workers will form large groups, while unreliable ones are discordant. This assumption may fail if only a few candidate choices are available for a confusing task that has many potential answers. This can be avoided by concatenating multiple questions into one task, which expands the answer space to reveal the true clustering structure (See the experiment section).

Note that K tasks and N workers have to be considered jointly. A single task cannot identify whether a worker is of high quality or not. Similarly, a single worker cannot identify the ambiguity of tasks.

[2]A worker's reliability can be possibly different in various tasks and time-evolving; a task may be clear to a certain subgroup of workers but not to others. A systematic study of this more complicated phenomenon is beyond the scope of this paper. We leave it for future work.

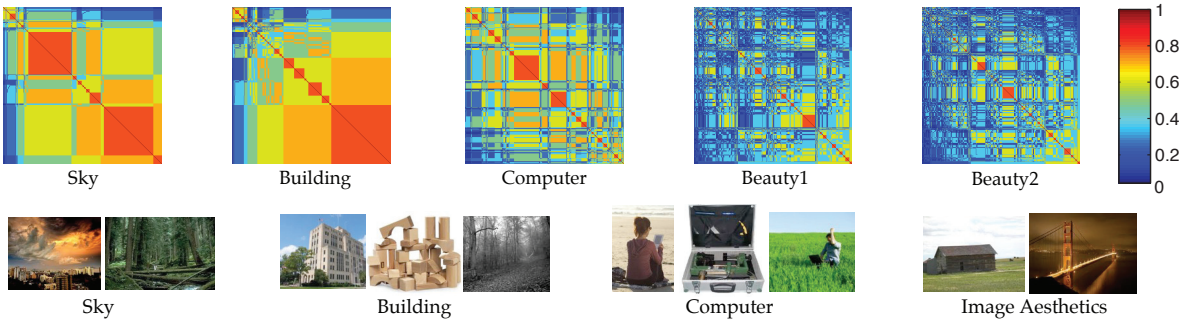

Figure 1: Depict of schools of thought in our experiment that asks 402 unique workers to categorize an object in images (*sky, building, computer*) and pick beautiful images (*beauty1, beauty2*). First row: As shown in the posterior link probability matrix computed from our model (Eqn. (7)), workers give a variety of responses that tend to cluster together, i.e., the *schools of thought*. Moreover, the scale (number) of schools of thought is task-specific. Second row: Example images in each task. *Sky:* Only positive and negative images. *Building:* Positive, ambiguous and negative images. *Computer:* Positive, ambiguous and negative images that contains elusive objects. *Image aesthetics:* Beautiful/ordinary-looking images. See Section 4 for experiment design.

2.2 The Low-rank Model

Given the two assumptions, a quantitative relation follows. Suppose we have somehow clustered the worker responses of task k into M_k clusters (the cluster model will be presented in Section 2.3), where M_k is an unknown parameter. Let z_{ik} denote the group index that worker i corresponds to in task k, and $\#z_{ik}$ be the *size* of that group. In this paper, $\#z_{ik}$ is regarded as a way to represent the scale of schools of thought. The larger $\#z_{ik}$ is, the smaller the scale is. We thus formally define our computational model by relating $\#z_{ik}$ to λ_i and μ_k as follows:

$$\#z_{ik} = \lambda_i \mu_k + \epsilon_{ik}. \tag{1}$$

where ϵ_{ik} is a zero-mean random noise (its distribution is determined by our clustering model). In matrix form,

$$\#\mathbf{Z} = \boldsymbol{\lambda}\boldsymbol{\mu}^T + \epsilon. \tag{2}$$

Under expectation, we have $\#\tilde{\mathbf{Z}} = \boldsymbol{\lambda}\boldsymbol{\mu}^\top$, which is a rank-1 factorization of the $N \times K$ *worker-task* groupsize matrix $\#\tilde{\mathbf{Z}}$. The resulting μ_k and λ_i can thus be used to rank tasks or workers. The intuition underlying the low-rank factorization is that according to the assumptions, reliable workers working on clear and well-defined tasks will give unanimous answers, yielding large group sizes in expectation. Note that we treat $\#z_{ik}$ as a continuous variable rather than an integer. Section 2.4 discusses how to estimate its expected value.

Given the assumptions, rank-1 factorization is the most straightforward way to formulate the relationship between $\#z_{ik}$, λ_i and μ_k. More complicated modeling may help, e.g. assuming a nonlinear relationship or enforcing $\#Z$ to be rank-m rather than rank-1 to find multiple factors for both workers and tasks that are related to the grouping behavior. However, we leave these extensions for future work.

2.3 The Clustering Model

To obtain worker reliability and task clarity, a key step is to estimate the expected group sizes. Although many clustering methods, such as spectral clustering [13], Kmeans and etc., can be applied to first infer the group assignment of each worker in each task and then calculate the group sizes, most of them have a difficult time in determining the unknown cluster number M_k. To bypass the hard model selection problem and also to allow the model to adaptively grow as more data are provided[3], we resort to nonparametric Bayesian techniques, which have shown great promise in machine learning, statistics, and many application areas [3, 6, 24]. More specifically, we propose to use the Dirichlet process (DP) mixture model [3], which is a nonparametric Bayesian model that can automatically resolve the unknown number of clusters M_k. Moreover, as we shall see, for our DP mixture model, we can analytically compute the expected group sizes without an intermediate inference step, which can be expensive and unstable.

Formally, let $\mathbf{x}_{ik}$ be the d-dimensional observed answers of worker i to task k. Typically, $\mathbf{x}_{ik}$ is a vector encoding the worker i's answers to a sequence of questions in task k (e.g., a worker gives answers to d questions of a survey). For a task k, answers of workers form M_k clusters. Let $\mathbf{c}_{ik}$ denote the center of the cluster that worker i belongs to. For different workers i and j, $\mathbf{c}_{ik} = \mathbf{c}_{jk}$ if they are in the same group. For each task $k = 1 \dots K$, the DP mixture with a Gaussian likelihood model can thus be written as:

$$\begin{aligned} G_k|\{\alpha_k, G_0^k\} &\sim \mathcal{DP}(\alpha_k, G_0^k) \\ \mathbf{c}_{ik}|G_k &\sim G_k \\ \mathbf{x}_{ik}|\{\mathbf{c}_{ik}, \sigma^2\} &\sim \mathcal{N}(\mathbf{x}_{ik}|\mathbf{c}_{ik}, \sigma^2\mathbf{I}), \end{aligned} \tag{3}$$

Although in principle, we could use a separate variance σ_{ik}^2 for each worker i and each task k, here we treat them as one single hyperparameter σ for simplicity and will provide sensitivity analysis of our model with respect to σ^2. Due to the fact that the distributions G_k sampled from a DP are discrete almost surely [4], there is a non-zero probability that two workers belong to the same cluster. Thus, we will have a partition of $\mathbf{x}_{ik}$ according to the sampled values $\mathbf{c}_{ik}$ and automatically infer the cluster number M_k. Alternatively, the mixture model can be equivalently represented using Chinese restaurant process (CRP)[4] prior, which is:

$$\begin{aligned} z_{ik} &\sim \mathcal{CRP}(\cdot|\alpha_k) \\ \mathbf{c}_{mk} &\sim G_0^k \\ \mathbf{x}_{ik}|\{z_{ik}, \mathbf{c}, \sigma^2\} &\sim \mathcal{N}(\cdot|\mathbf{c}_{z_{ik},k}, \sigma^2\mathbf{I}), \end{aligned} \tag{4}$$

where $\mathbf{c}_{mk}$ is the cluster center for cluster m at task k. In

[3]For example, the number of clusters can grow for a DP mixture model when more data are provided. In other words, DP mixture can have an unbounded number of clusters.
[4]A CRP is a marginalized version of a DP [6].

this work, we set the base distribution G_0^k as $\mathcal{N}(\mathbf{0}, \sigma_{k0}^2\mathbf{I})$, where σ_{k0}^2 are the task-specific variances that characterize how much the means of clusters are scattered around the origin. We will present a procedure to automatically estimate them. On the other hand, the hyperparameter σ characterizes the variance within each cluster and need to be specified. A sensitivity analysis is shown in the experiment.

For the DP mixture, exact inference of the group assignment z_{ik} is intractable. Typical solutions resort to variational or sampling [12] methods. However, these approximate inference methods often lead to local optimum solutions and can be expensive and sensitive to initialization. Thus, we need to estimate the expected group sizes in a more efficient and robust way. Here we derive an analytic solution to the expected group sizes for the clustering model, without intermediate inference. Our approach can be expected to be faster and more stable compared to those using approximate inference, analogous to what people have popularly done in collapsed sampling [7] or collapsed variational inference [20] in probabilistic latent variable models.

2.4 Expected Group Size

For each task k, we use $\mathbf{X}_k = \{\mathbf{x}_{ik}\}_{i=1}^N$ to denote all its worker responses. Let W_{ij}^k be a binary random variable that equals to 1 if workers i and j are in the same group of thought in task k ($W_{ii}^k = 1$). From W_{ij}^k, the group size $\#z_{ik}$ for worker i can be computed as $\#z_{ik} = \sum_{j=1}^N W_{ij}^k$. Thus, by linearity of expectation, the expected group size conditioned on $\mathbf{X}_k$ is:

$$\begin{aligned} \#\tilde{z}_{ik} &\equiv \mathbb{E}\left[\#z_{ik}|\mathbf{X}_k\right] = \mathbb{E}\left[\sum_{j=1}^N W_{ij}^k \middle| \mathbf{X}_k\right] \\ &= \sum_{j=1}^N \mathbb{E}\left[W_{ij}^k|\mathbf{X}_k\right] = \sum_{j=1}^N \mathbb{P}\left(W_{ij}^k|\mathbf{X}_k\right), \end{aligned} \quad (5)$$

where the last equality holds because W_{ij}^k is binary. Note that the linearity of expectation still applies even if W_{ij}^k are not independent variables (e.g., for $i \neq i'$, W_{ij}^k and $W_{i'j}^k$ may be dependent because they share the same worker j).

To compute the posterior distribution of W_{ij}^k, we make the following approximation:

$$\mathbb{P}\left(W_{ij}^k|\mathbf{X}_k\right) \approx \mathbb{P}\left(W_{ij}^k|D_{ij}^k\right), \quad (6)$$

where $D_{ij}^k = \|\mathbf{x}_{ik} - \mathbf{x}_{jk}\|^2$ is the squared Euclidean distance[5] between responses of workers i and j. The intuition is that two workers i and j being in the same group is largely due to their affinity, but is *almost* independent of other workers' responses. Our synthetic experiments verify that this approximation is very accurate (See Section 3 for details). In practice, this approximation is also reasonable in crowdsourcing services like M-Turk, where the onsite communication between workers is not allowed. After computing the likelihood $\mathbb{P}\left(D_{ij}^k|W_{ij}^k\right)$ and prior $\mathbb{P}\left(W_{ij}^k = 1\right)$, with Bayes' rule we obtain the posterior link probability:

$$\mathbb{P}\left(W_{ij}^k = 1|D_{ij}^k\right) = \frac{1}{1 + \exp\left(\beta_k D_{ij}^k + \beta_{k0}\right)} \quad (7)$$

where $\beta_k \equiv \frac{\sigma_{k0}^2}{4\sigma^2(\sigma_{k0}^2+\sigma^2)}$ and $\beta_{k0} \equiv \log\alpha_k + \frac{d}{2}\log\frac{\sigma^2}{\sigma_{k0}^2+\sigma^2}$. Please see Appendix for detailed derivation. In this work, we derive the prior from the exchangeability property of CRP.

Note that the posterior distribution in Eqn. (7) is not restricted to the CRP-Gaussian cluster model proposed in Eqn. (4). In general, we can apply other priors or explicitly define $\mathbb{P}\left(W_{ij}^k\right)$. Besides Gaussian, each cluster's noise model can follow any other unimodal distribution and Eqn. (7) still hold with D_{ij}^k with a different distance metric. The metric could also be redefined and generalized to arbitrary type of answers (e.g. binary, categorical).

Interestingly, our link distribution model in Eqn. (7) for two workers being in the same group has the same logistic form as the latent distance model for social network analysis [8], where the link distribution model is directly defined.

2.5 Worker Reliability and Task Clarity

Once we have obtained the expected group size $\#\tilde{z}_{ik}$ for each worker i and each task k, the worker reliability $\boldsymbol{\lambda}$ and task clarity $\boldsymbol{\mu}$ can be estimated as the first left and right singular vector corresponding to the largest singular value of the expected group size matrix $\#\tilde{\mathbf{Z}} \equiv \{\#\tilde{z}_{ik}\}$ (Eqn. (1)). The entire algorithm is summarized in Alg. 1. Note we do not need to impose positive constraints for $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, since the first left and right singular vector of a matrix with positive entries is always positive by Perron-Frobenius theorem.

Algorithm 1 Estimation worker reliability and task clarity in the presence of schools of thought.

1: **(Input)** The worker responses $\{\mathbf{X}_k\}_{k=1}^K$ for K tasks.
2: **(Output)** Worker reliability $\boldsymbol{\lambda}$ and task clarity $\boldsymbol{\mu}$.
3: **for** k = 1:K **do**
4: Compute the posterior $\mathbb{P}\left(W_{ij}^k|D_{ij}^k\right)$ that worker i and j are in the same group (Eqn. (7)).
5: Computer the expected group size $\#\tilde{z}_{ik}$ (Eqn. (5)).
6: **end for**
7: Run SVD on expected *worker-task* groupsize matrix $\#\tilde{\mathbf{Z}}$: $\#\tilde{\mathbf{Z}} = U\Lambda V^\top$
8: Set $\boldsymbol{\lambda} = U_{\cdot 1}\sqrt{\Lambda_{11}}$ and $\boldsymbol{\mu} = V_{\cdot 1}\sqrt{\Lambda_{11}}$, where $U_{\cdot 1}$ is the first column of U, alike for $V_{\cdot 1}$.

2.6 Hyperparameters

Before ending this section, we introduce a simple procedure to estimate the hyperparameters α_k and σ_{k0}, with the assumption that σ is given. As we shall see, our model is insensitive to the only tunable hyperparameter σ. Thus, although the procedure does not estimate all hyperparameters, it is good enough for our use.

Specifically, after marginalizing cluster partition, we can obtain $\mathbb{P}\left(D_{ij}^k\right)$, a distribution of the observable pairwise squared distances parameterized by the hyperparameters α_k, σ_{k0} and σ. Similarly we compute $\mathbb{P}\left(\|\mathbf{x}_{ik}\|^2\right)$. Given σ, we can estimate α_k and σ_{k0} from the equations:

$$\mathbb{E}\left[D_{ij}^k\right] = 2d\left(\sigma^2 + \sigma_{k0}^2\frac{\alpha_k}{1+\alpha_k}\right) \quad (8)$$

$$\mathbb{E}\left[\|\mathbf{x}_{ik}\|^2\right] = d(\sigma^2 + \sigma_{k0}^2). \quad (9)$$

The derivation is simple by noticing that $\mathbb{E}\left[D_{ij}^k\right] = \sum_{l=0,1} \mathbb{E}\left[D_{ij}^k|W_{ij}^k = l\right]\mathbb{P}\left(W_{ij}^k = l\right)$. Similarly for $\mathbb{P}\left(\|\mathbf{x}_{ik}\|^2\right)$.

3. SYNTHETIC VALIDATION

Before presenting the experiments on real data, we first conduct synthetic experiments to show empirically that our

[5]The same result follows when using Euclidean distance.

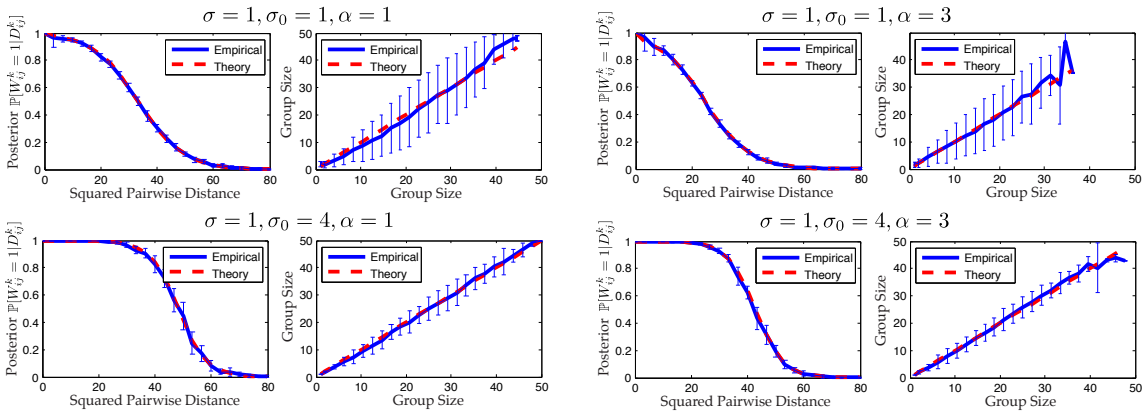

Figure 2: Validation of posterior link probability (Eqn. (7)) and expected group size (Eqn. (5)). Our theoretical estimation matches well with the simulation under different parameter settings.

approximation (Eqn. (6)) is valid and our estimation of expected group sizes is reasonably accurate for the clustering model. We investigate the performance with four sets of parameters $(\alpha, \sigma_0, \sigma)$ as shown in Fig. 2. For each set of parameters, we set $d = 12$ and generate $K = 500$ independent tasks according to Eqn. (4), each with $N = 50$ workers and $\sigma_{k0} = \sigma_0, \alpha_k = \alpha$. We compare our estimated group sizes using Eqn. (5) and posterior link probability using Eqn. (7) with those empirically computed from the simulation. We can see that our theoretical estimation is very accurate, especially when the clusters are well-separated (i.e., $\sigma_0 \gg \sigma$).

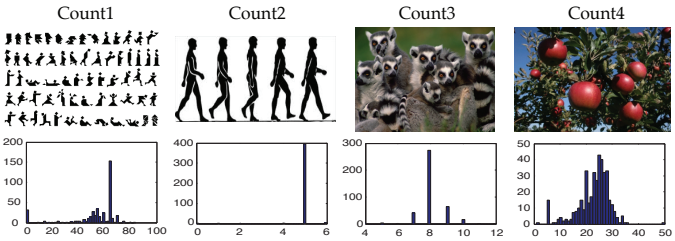

Figure 3: Object counting task. A counting histogram is shown as well as the image. In all tasks, the cluster structures of worker responses are clear. For *Count3*, an outlier response 725 is not shown.

4. EXPERIMENTS ON M-TURK

Now, we present empirical studies on the real data collected from Amazon Mechanical Turk (M-Turk). Since our main focus is to characterize the schools of thought phenomenon in crowdsourcing, most of the experiments are qualitative. At the end, we also present some quantitative results to demonstrate the potential usefulness of our statistical analysis for predicting gold standard (if exists).

4.1 Experimental setup

For each HIT (i.e., Human Intelligence Task), we design three missions, each containing several tasks with different levels of clarity listed as follows. Each worker is required to finish all the tasks only once. Following [21], all tasks are vision-related. We expect to see workers of various reliability, task-specific schools of thought due to diversity of clarity, and give insights to possible confusions in human-aided vision tasks. We emphasize that our analysis is not restricted to vision tasks and our techniques can be applied to analyze crowdsourcing data collected in other fields, including text mining, natural language processing, etc.

Mission 1: Object Categorization. In this mission, we provide three tasks. In each task, workers are asked to decide which of the 12 images contain a certain object category, namely *sky*, *building* and *computer*. Each task is designed to have a different level of clarity. For task *sky*, there are simply 6 images with sky and 6 without sky. For the less clear task *building*, there are 4 images with a typical building, 4 images without a building, and 4 images that contain a building-like structure. The task *computer* is the most confusing one, in which 6 out of 12 images are equally divided into three subsets, each containing a typical computer, a typical object that is not a computer, and a computer-like electrical appliance (e.g., Apple iPad or Amazon Kindle); and the other 6 images follow the same strategy of three-way-division but the objects in the images are elusive and require some efforts to find.

Mission 2: Object Counting. In this objective mission, workers are asked to count the number of objects in 4 images. We regard each image as one task. Among the 4 images, the simplest one contains 5 humans, one contains 65 small human-shaped objects that are laborious to count, one contains 8 animals huddling together and requires efforts to count, and the most confusing one contains 27 apples that are of various sizes and partially hidden in tree branches. All the 4 images have gold standards counted by authors.

Mission 3: Images Aesthetics. In this subjective mission, there are two tasks with comparably low clarity. In each task, workers are asked to pick 6 most beautiful images from 12 images. Among the 12 images, 3 are considered ordinary-looking, 3 are beautiful, 3 are impressively beautiful, and 3 are of astonishing beauty, according to authors' criterion.

Fig. 1 and Fig. 3 show example images. All images are manually picked online. The images are randomly shuffled when presented to workers to avoid any order bias ("Donkey vote"). Once we have obtained the responses of workers, we remove incomplete and duplicate responses from the same worker and construct a dataset that contains the responses of 402 unique workers to the 9 tasks. For each worker, the response includes a 12 dimensional binary vector for each task in object categorization and image aesthetics missions, and a 1 dimensional integer for each object counting task.

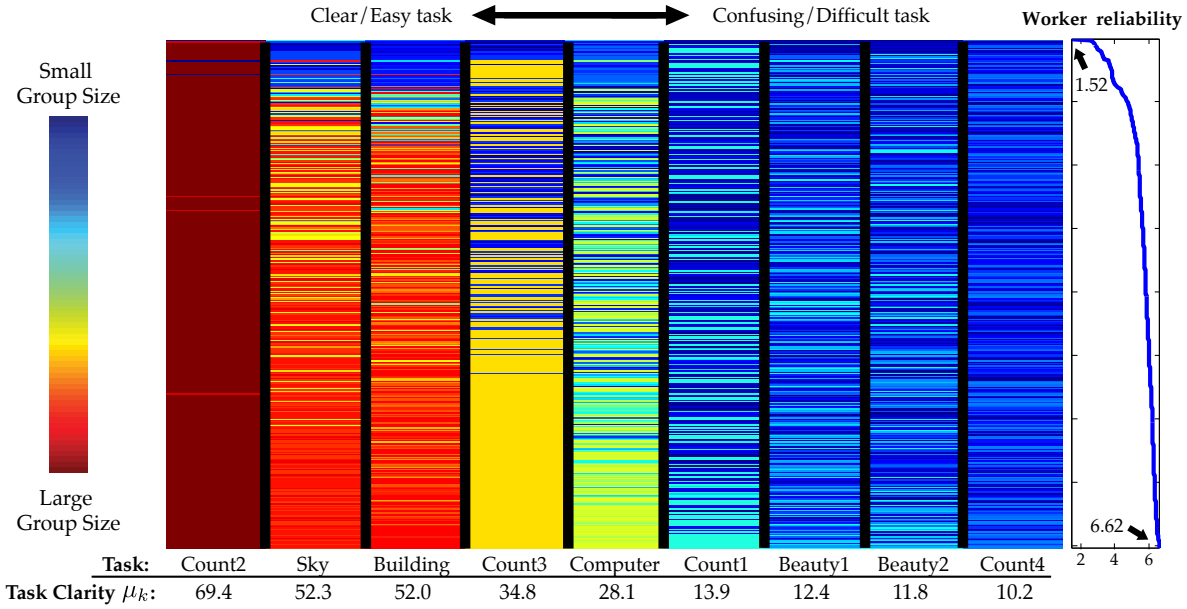

Figure 4: Expected *worker-task* groupsize matrix with rows sorted by worker reliability $\boldsymbol{\lambda}$ and columns sorted by task clarity $\boldsymbol{\mu}$. The matrix shows a strong factorizable structure. Workers: Reliable workers (the bottom part of the matrix) show their consistency across different tasks. Coinciding with our assumption, they always stay with large groups. In addition, from the distribution of reliability, a small portion of the workers can be identified as low quality (the top part of the matrix). Tasks: Tasks in object categorization mission are generally clear while tasks in image aesthetics mission are in general ambiguous. Tasks in counting missions show mixed clarity. *Count2* (Counting five humans) is the clearest task, while *Count4* (Counting apples) is the most confusing one. (This figure is best viewed in color.)

Unless explicitly mentioned, in all tasks we set the hyperparameter $\sigma = 0.2$ and estimate α and σ_0 from empirical expectation (Eqn. (8) and Eqn. (9)). We will also provide a sensitivity analysis on the hyperparameter σ.

4.2 Characterization of Schools of Thought

We apply our low-rank model to all the 9 tasks. The rank-1 residual error $\|\#\tilde{\mathbf{Z}} - \boldsymbol{\lambda}\boldsymbol{\mu}^T\|_F / \|\#\tilde{\mathbf{Z}}\|_F$ is 0.27, which means 73% of the energy in $\#\tilde{\mathbf{Z}}$ has been explained away by $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. This shows that our model can fit the data well. Although we do not jointly model the interaction between clustering and factorization, the cluster size matrix $\#\tilde{\mathbf{Z}}$ *naturally* follow the rank-1 factorization, which verifies our low-rank assumption. It may be theoretically intriguing to formulate a joint model and design an iterative procedure for model fitting. However, this may result in an improper bias on the data.

Below, we first examine the existence of schools of thought and its two major latent factors – worker reliability and task clarity, and then provide detailed analysis on worker reliability.

Visualization of schools of thought. We show the patterns of schools of thought for the tasks in object categorization and image aesthetics missions as the posterior link probability matrix (Fig. 1) computed from Eqn. (7). For better visualization, we set $\alpha = 750$, $\sigma = 0.6$ and $\sigma_0 = 1$. Rather than posterior link probability matrix, histograms are shown separately for each of the 4 counting tasks (Fig. 3). Different visualization is used because each counting task is done separately, while for the other missions, workers' responses are based on 12 images at a time.

Task-specific schools of thought. From Fig. 1 and Fig. 3 we can clearly see the task-dependent schools of thought. Even for the simplest task (e.g., *sky*) and tasks with ground truth (e.g., object counting), there are still substantially diverse answers. For task *sky*, a large group of people think the outer space looking from the Moon is not sky, or a glowing bluish icy ceiling in a cave is sky. For counting, some workers think the image with 65 human-like drawings does not contain humans and give zero answer. For more complicated and confusing tasks, the number of clusters goes up, and each cluster size goes down. In subjective tasks, almost everyone has their own responses and the cluster structure is almost invisible.

Distribution of worker reliability. Fig. 4 shows the structure of the expected *worker-task* groupsize matrix $\#\tilde{\mathbf{Z}}$ as well as the estimated worker reliability $\boldsymbol{\lambda}$ and task clarity $\boldsymbol{\mu}$. From the worker reliability plot, most of the workers are comparably reliable and they tend to stay consistently in larger groups in different tasks. A small portion of workers did a very poor job, consistent with the observation in [19]. Among the three types of missions, object categorization is relatively clear, image aesthetics is in general very subjective and vague. Counting mission shows mixed results. Nearly all workers are correct in task *Count2*, making it the clearest one. One the other hand, counting apples of varied size with background clutters (task *Count4*) is extremely confusing.

4.3 Closer Examination of Worker Reliability

In this subsection, we present a closer examination of the estimated worker reliability and compare it with baselines. Besides, we also show how to use it for improving prediction of gold standard for objective tasks.

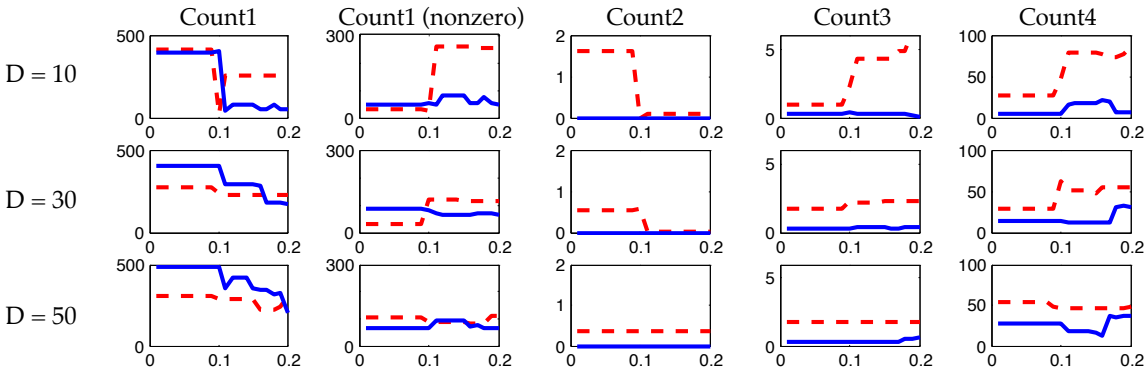

Figure 5: The variance $\bar{\sigma}_T^2$ of the responses of D most reliable workers (blue solid lines) versus the variance $\bar{\sigma}_{UT}^2$ of D most unreliable workers (red dashed lines) averaged on the 4 counting tasks. Ideally $\bar{\sigma}_T^2 \ll \bar{\sigma}_{UT}^2$ (i.e., blue solid lines are below red dashed lines). In all plots, x-axis shows hyperparameter σ ranging from 0.01 to 0.2.

Ranking workers. To verify the computed worker reliability $\boldsymbol{\lambda}$, we first estimate $\boldsymbol{\lambda}$ on the 5 tasks (*sky, building, computer, beauty1, beauty2*), then check if the estimation makes sense in the remaining four objective tasks. For checking, we rank the workers according to $\boldsymbol{\lambda}$, compute the answer variance $\bar{\sigma}_T^2$ for reliable workers, and the answer variance $\bar{\sigma}_{UT}^2$ for unreliable workers, and then compare the two variances. If $\bar{\sigma}_{UT}^2 \gg \bar{\sigma}_T^2$, which means workers labeled as "unreliable" give inconsistent answers compared to those labeled as "reliable", then the ranking is meaningful and can be generalized to other unseen tasks. Specifically, $\bar{\sigma}_T^2$ is computed from D most reliable workers, and is averaged over 4 remaining counting tasks. Similarly for $\bar{\sigma}_{UT}^2$. The result is shown in Fig. 5. We vary D from 10 to 50, and vary σ from 0.01 to 0.2 for sensitivity analysis. A subset of workers give zero responses to task *Count1*, (presumably thinking those drawings are not humans). In the second column, we also show the results after excluding zeros from ranking.

It is clear that in most of the cases, the reliable workers estimated on one set of tasks give answers that are much more consistent (i.e., with a lower variance) than the unreliable ones in a different set of tasks. This suggests that the worker reliability is generalizable from one task to another task, which is consistent with what we have observed in Fig. 4. From the results, we can also see that our approach is relatively insensitive to the change of hyperparameter σ and parameter D.

Comparison with baseline clustering models. As we have stated in Section 2.3, we can use alternative methods to perform the clustering on worker responses. Now, we compare the performance of our method ($\sigma = 0.2$) with spectral clustering [13], PCA-Kmeans and Gibbs sampling on the DP mixture model in Eqn. (3). Note that all these baselines require inference on the cluster assignment of each worker in order to compute the group sizes, while our method does not. For spectral clustering, we take the $L = 5$ to 70 smallest eigenvectors of normalized graph Laplacian and normalize them to be unit vectors as the low-dimensional embedding, on which Kmeans are performed with L clusters. For PCA-Kmeans, we first reduce the dimension of the workers' response to 5 using PCA, and run Kmeans with $L = 5$ to 70 clusters. For Gibbs sampling, we use the Algorithm 8 in [12] with the same set of hyperparameters as estimated in Eqn. (9) with $\sigma = 0.2$.

Table 1: Time cost comparison between the methods using various baseline clustering algorithms and ours.

Methods	Time (sec)
Ours	1.41 ± 0.05
Spectral Clustering	3.90 ± 0.36
PCA-Kmeans	0.19 ± 0.06
Gibbs Sampling	53.63 ± 0.19

Fig. 6 shows the performances of different methods. All the baselines are repeated for 500 times with random initialization. We present their best average performances and its associated standard deviations, achieved by tuning the hyperparameter (i.e., the number of clusters). We can see that our approach is comparable with baselines but is much more stable because it does not require initialization. Tbl. 1 shows the average time cost over 50 runs. Ours is faster than spectral clustering and Gibbs sampling. PCA-Kmeans is the fastest since it does not compute pairwise distance, but its performance is worse than ours.

Table 2: Performance comparison on predicting the gold standard counts for the four counting tasks.

	Cnt1	Cnt2	Cnt3	Cnt4
Ours. $D = 5$	65	5	8	26
Ours. $D = 10$	65	5	8	26
Ours. $D = 20$	65	5	8	25.6
MV	53.7	5.0	9.9	22.9
MV(median)	60	5	8	24
LFC [15]	56	5	8	24
MDW [21] ($c = 1$) (top-10 pred., 50 init.)	63.7	5	8	25.96
MDW [21] ($c = 3$) (top-10 pred., 50 init.)	65.01	5	8	25.48
Gold Standard	65	5	8	27

Prediction of gold standard. As a specific case of schools of thought, our approach can also be applied for those tasks having one unanimous answer (e.g., the counting tasks) and predict the gold standard. Here, we provide one example that uses selective "majority voting", namely, we first select D most reliable workers based on the estimated worker reliability $\boldsymbol{\lambda}$, and then apply majority voting for

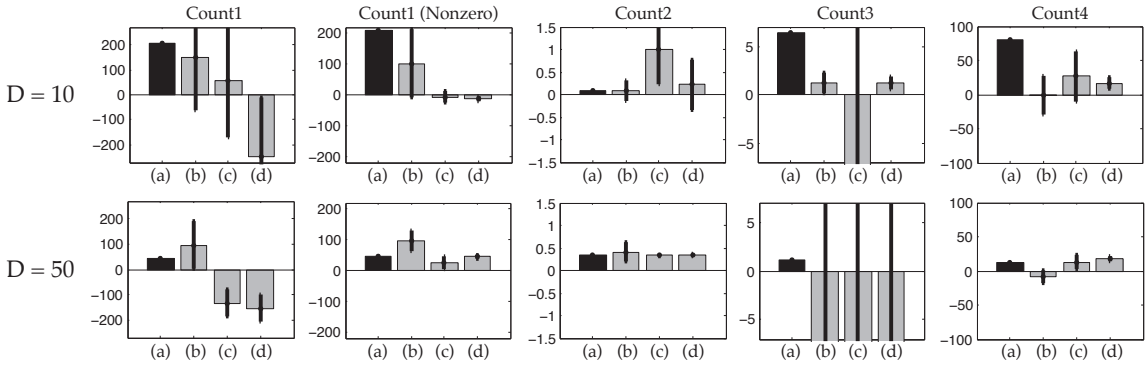

Figure 6: Comparison between (a) our method (with $\sigma = 0.2$), (b) spectral clustering, (c) PCA+Kmeans, and (d) Gibbs sampling, on the difference $\bar{\sigma}^2_{UT} - \bar{\sigma}^2_T$, where $\bar{\sigma}^2_{UT}$ is the variance of the answers of D most unreliable workers and $\bar{\sigma}^2_T$ is the variance of D most reliable workers averaged on 4 counting tasks. Ideally $\bar{\sigma}^2_{UT} \gg \bar{\sigma}^2_T$, and thus the bar should be a large positive number. The first row shows $D = 10$, and the second shows $D = 50$.

prediction. An alternative method would be weighing the responses of workers with their reliability. Yet we obtain the same performance.

Tbl. 2 shows the comparison of our method to three very competitive baseline methods on predicting the count numbers for 4 counting tasks. The baselines include the classic inclusive "majority voting" (MV) heuristic, which performs over all the 402 workers, the "learning from crowds" (LFC) method [15], which iteratively estimates workers' precisions and tasks' gold standard using an EM procedure, and the "multidimensional wisdom of crowds" (MDW) method [21] as we have discussed in the introduction. For LFC, we follow Eqn.(10)-(11) in [15] and use MV as the initial guess of gold standard as suggested in the paper. For MDW, since it does not handle the case that the gold standard is in the continuous domain, we first find maximum a posteriori estimation of the workers' precisions, treat them as workers' reliability, and estimate the gold standard by averaging the answers of top-10 reliable workers. In MDW, each task k is a hidden c-dimensional vector $\mathbf{v}_k$ and each worker is represented as a set of weights in the same dimension. Both need to be estimated from the data. In the experiment, we choose $c = 1$ and $c = 3$.

We can see that our selective MV outperforms all the three baselines, especially on *Count1* and *Count4*, and our prediction matches the gold standard very well. MDW has comparable performance yet it is quite sensitive to the initialization. When $d = 1$, MDW gives the same prediction as ours (i.e. $(65, 5, 8, 26)$) in most random initializations, however, it also gives $(0, 5, 8, 24.05)$ if not initialized properly. The reason is that, although most workers vote for 65 (called "65-voters"), there is a small group voting for 0s ("0-voters") on *Count1*. If MDW falls into this small cluster and regards 0-voters as more reliable than 65-voters, the gold standard estimation of *Count4* will also change. In contrast, ours always rate 65-voters over 0-voters since 65-voters form a larger cluster than 0-voters. Thus, ours gives a deterministic answer to worker reliability. Besides, on the four counting tasks, using the stopping criterion

$$\sqrt{\sum_k \|\mathbf{v}_k^{t+1} - \mathbf{v}_k^t\|^2} / \sqrt{\sum_k \|\mathbf{v}_k^t\|^2} \leq 10^{-4}$$

(where $\mathbf{v}_k$ is the hidden representation of task k in iteration t), MDW spends 3.25 ± 3.34 seconds for $c = 1$, and 4.26 ± 3.70 seconds for $c = 3$, averaged over 50 random initializations. The large variation in the rate of convergence is due to different initialization. In comparison, our method runs for 0.68 seconds, is insensitive to the parameter D and the value of the hyperparameter σ, and has no initialization.

5. CONCLUSIONS AND FUTURE WORK

This paper formally analyzes the schools of thought phenomenon and its two key underlying factors, worker reliability and task clarity, in crowdsourcing by presenting a computational model based on a low-rank assumption, which characterizes the relationships between the group sizes of worker responses and the two key factors. Furthermore, the expected group sizes can be estimated analytically instead of performing an expensive and unstable inference step. We report real experiments on Amazon Mechanical Turk.

In terms of time cost, the major bottleneck of our work is to compute the groupsize matrix $\#\mathbf{Z}$, which has time complexity of $O(KN^2)$ (K is the number of tasks and N is the number of workers). However, if N is large, given worker i, we can sample a few other workers to obtain an unbiased estimate of $\#z_{ik}$, yielding approximately $O(KN)$ complexity. Another interesting future direction is to handle the case of missing data that some workers may not give answer to some tasks. In such a case, how to estimate the group size for observed answers and how to factorize the group size matrix $\#\mathbf{Z}$ in the presence of missing entries deserve more exploration.

From a learning point of view, our analysis could provide insights for developing better predictive tools in several ways. For example, we have shown that high quality labels can be selected from the answers of workers who enjoy top ranking in worker reliability for building better predictive models. For future work, we can acquire new labels from those workers that have higher reliability in the context of active learning [14]. Finally, our analysis could help the market price "high-reputation" workers [2].

Acknowledgements Yuandong Tian is supported by Microsoft Research PhD Fellowship (2011-2013) and an ONR grant N00014-11-1-0295.

6. REFERENCES

[1] J. Abernethy and R.M. Frongillo. A collaborative mechanism for crowdsourcing prediction problems. In *NIPS*, 2011.

[2] E. Adar. Why I hate Mechanical Turk research (and workshops). In *CHI*, 2011.

[3] C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of statistics*, 2(6):1152–1174, 1974.

[4] D. Blackwell and J.B. MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of statistics*, 1(2):353–355, 1973.

[5] AP Dawid and AM Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

[6] S.J. Gershman and D. Blei. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology (in press)*, 2011.

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

[8] P.D. Hoff, A.E. Raftery, and M.S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[9] P.G. Ipeirotis, F. Provost, and J. Wang. Quality management on Amazon Mechanical Turk. In *ACM SIGKDD workshop on human computation*, 2010.

[10] D.R. Karger, S. Oh, and D. Shah. Iterative learning for reliable crowdsourcing systems. In *NIPS*, 2011.

[11] A. Kittur, E. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *CHI*, 2008.

[12] R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[13] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.

[14] D. Pinar, J. Carbonell, and J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *SIGKDD*, 2009.

[15] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 99:1297–1322, 2010.

[16] J. Ross, L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *CHI*, 2010.

[17] D.M. Russell, M.J. Stefik, P. Pirolli, and S.K. Card. The cost structure of sensemaking. In *CHI*, 1993.

[18] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labelling of venus images. In *NIPS*, 1995.

[19] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP*, 2008.

[20] Y.W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *NIPS*, 2006.

[21] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *NIPS*, 2010.

[22] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, 2009.

[23] Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, J.G. Dy, and PA Malvern. Modeling annotator expertise: Learning when everybody knows a bit of something. In *AISTATS*, 2010.

[24] J. Zhu, N. Chen, and E. Xing. Infinite Latent SVM for Classification and Multi-task Learning. In *NIPS*, 2011.

Appendix: Derivation of posterior distribution

$\mathbb{P}\left(W_{ij}^k = 1 | D_{ij}^k\right)$

By the Bayes' rule, we have

$$\mathbb{P}\left(W_{ij}^k = 1 | D_{ij}^k\right) = \frac{\mathbb{P}\left(D_{ij}^k | W_{ij}^k = 1\right)\mathbb{P}\left(W_{ij}^k = 1\right)}{\mathbb{P}\left(D_{ij}^k | W_{ij}^k = 0\right)\mathbb{P}\left(W_{ij}^k = 0\right) + \mathbb{P}\left(D_{ij}^k | W_{ij}^k = 1\right)\mathbb{P}\left(W_{ij}^k = 1\right)},$$

where the likelihood term $\mathbb{P}\left(D_{ij}^k | W_{ij}^k\right)$ and the prior term $\mathbb{P}\left(W_{ij}^k = 1\right)$ are computed as follows.

The likelihood term $\mathbb{P}\left(D_{ij}^k | W_{ij}^k\right)$. For $W_{ij}^k = 1$, both $\mathbf{x}_{ik}$ and $\mathbf{x}_{jk}$ are generated independently from the same cluster center with variance σ^2, thus we have $\mathbf{x}_{ik} - \mathbf{x}_{jk} \sim \mathcal{N}(\cdot|\mathbf{0}, 2\sigma^2)$. Therefore, $D_{ij}^k = \|\mathbf{x}_{ik} - \mathbf{x}_{jk}\|^2$ satisfies χ^2 distribution with the following pdf:

$$\mathbb{P}\left(D_{ij}^k | W_{ij}^k = 1\right) = \phi(D_{ij}^k; d, \sigma^2) = \frac{1}{2^{d/2}\Gamma(d/2)}\frac{1}{(\sigma^2)^{d/2}}(D_{ij}^k)^{d/2-1}\exp\left(-\frac{D_{ij}^k}{4\sigma^2}\right)$$

where d is the dimension of the workers' responses. Similarly, for $W_{ij}^k = 0$, by integrating their cluster centers out, we can show $\mathbf{x}_{ik}, \mathbf{x}_{ik}$ are independently generated from $\mathcal{N}(\cdot|\mathbf{0}, \sigma_0^2 + \sigma^2)$. Thus we have $\mathbb{P}\left(D_{ij}^k | W_{ij}^k = 0\right) = \phi(D_{ij}^k; d, \sigma^2 + \sigma_0^2)$.

The prior term $\mathbb{P}\left(W_{ij}^k = 1\right)$. By the exchangeability property of Chinese restaurant process, all workers are equal. Thus

$$\begin{aligned}\mathbb{P}\left(W_{ij}^k = 1\right) &= \mathbb{P}\left(W_{12} = 1\right) \\ &= \mathbb{P}\left(z_1 = 1\right)\mathbb{P}\left(z_2 = 1 | z_1 = 1\right) \\ &= \frac{1}{1+\alpha}\end{aligned}$$

Combining the two parts, we thus obtain the posterior distribution $\mathbb{P}\left(W_{ij}^k = 1 | D_{ij}^k\right)$:

$$\mathbb{P}\left(W_{ij}^k = 1 | D_{ij}^k\right) = \frac{1}{1 + \exp\left(\beta_k D_{ij}^k + \beta_{k0}\right)}$$

where $\beta_k \equiv \frac{\sigma_{k0}^2}{4\sigma^2(\sigma_{k0}^2 + \sigma^2)}$ and $\beta_{k0} \equiv \log\alpha_k + \frac{d}{2}\log\frac{\sigma^2}{\sigma_{k0}^2 + \sigma^2}$.

Note the same derivation follows if we use Euclidean distance $l_{ij,k} = \|\mathbf{x}_{ij} - \mathbf{x}_{ik}\|$ and

$$\begin{aligned}\mathbb{P}\left(W_{ij}^k = 1 | l_{ij,k}\right) &= \mathbb{P}\left(W_{ij}^k = 1 | D_{ij}^k\right) \\ &= \frac{1}{1 + \exp\left(\beta_k l_{ij,k}^2 + \beta_{k0}\right)},\end{aligned}$$

where is exactly the same as the one using squared Euclidean distance.