Probabilistic Multigraph Modeling for Improving the Quality of Crowdsourced Affective Data

Jianbo Ye, Jia Li, Michelle G. Newman, Reginald B. Adams, Jr. and James Z. Wang

Abstract—We proposed a probabilistic approach to joint modeling of participants' *reliability* and humans' *regularity* in crowdsourced affective studies. Reliability measures how likely a subject will respond to a question seriously; and regularity measures how often a human will agree with other seriously-entered responses coming from a targeted population. Crowdsourcing-based studies or experiments, which rely on human self-reported affect, pose additional challenges as compared with typical crowdsourcing studies that attempt to acquire *concrete non-affective* labels of objects. The reliability of participants has been massively pursued for typical non-affective crowdsourcing studies, whereas the regularity of humans in an affective experiment in its own right has not been thoroughly considered. It has been often observed that different individuals exhibit different feelings on the same test question, which does not have a sole correct response in the first place. High reliability of responses from one individual thus cannot conclusively result in high consensus across individuals. Instead, globally testing consensus of a population is of interest to investigators. Built upon the agreement multigraph among tasks and workers, our probabilistic model differentiates subject regularity from population reliability. We demonstrate the method's effectiveness for in-depth robust analysis of large-scale crowdsourced affective data, including emotion and aesthetic assessments collected by presenting visual stimuli to human subjects.

Index Terms—Emotions, human subjects, crowdsourcing, probabilistic graphical model, visual stimuli

1 Introduction

Humans' sensitivity to affective stimuli intrinsically varies from one person to another. Differences in gender, age, society, culture, personality, social status, and personal experience can contribute to its high variability between people. Further, inconsistencies may also exist for the same individual across environmental contexts and current mood or affective state. The causal effects and factors for such affective experiences have been extensively investigated, as evident in the literature on psychological and human studies, where controlled experiments are commonly conducted within a small group of human subjects — to ensure the reliability of collected data. To complement the shortcomings of those controlled experiments, ecological psychology aims to understand how objects and things in our surrounding environments effect human behaviors and affective experiences, in which *real-world* studies are favored over those within artificial laboratory environments [1, 2]. The key ingredient of those ecological approaches is the availability of large-scale data collected from human subjects, remedying the high complexity and heterogeneity that the real-world has to offer. With the growing attention on affective computing (initiated from the seminal discussion [3] to recent communications [4]), multiple data-driven approaches have been developed to understand what particular environmental factors drive the feelings of humans [5, 6], and how those effects differ among various sociological structures and between human groups.

One crucial hurdle for those affective computing approaches is the lack of full-spectrum annotated stimuli data at a large scale. To address this bottleneck, crowdsourcing-based approaches are highly helpful for collecting uncontrolled human data from anonymous participants [7]. In a recent study reported in [8], anonymous subjects from the Internet were recruited to annotate a set of visual stimuli (images): at each time point, after being presented with an image stimulus, participants were asked to assess their personal psychological experiences using ordinal scales for each of the affective dimensions: valence, arousal, dominance and likeness (which means the degree of appreciation in our context). This study also collected demographics data to analyze individual difference predictors of affective responses. Because labeling a large number of visual stimuli can become tedious, even with crowdsourcing, each image stimulus was examined by only a few subjects. This study allowed tens of thousands of images to obtain at least one label from a participant, which created a large data set for environmental psychology and automated emotion analysis of images.

One interesting question to investigate, however, is *whether the affective labels provided by subjects are reliable*. A related question is how to separate spammers from reliable subjects, or at least to narrow the scope of data to a highly reliable subgroup. Here, spammers are defined as those participants who provide answers without serious consideration of the presented questions. No answer from a statistical perspective is known yet for crowdsourced affective data.

A great difficulty in analyzing affective data is caused by the absence of ground truth in the first place, that is, there is no *correct* answer for evoked emotion. It is generally accepted that even the most reliable subjects can naturally have varied emotions. Indeed, with variability among human responses anticipated, psychological studies often care about questions such as where humans are emotionally consistent and where they are not, and which subgroups of humans are more consistent than another. Given a population, many, if not the vast majority of stimuli may not have a consensus emotion at all. Majority

Manuscript received ; revised .
J. Ye and J. Z. Wang are with the College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA. J. Li is with the Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA. M. Newman and R. B. Adams, Jr. are with the Department of Psychology, The Pennsylvania State University, University Park, PA 16802, USA. (e-mails: {jxy198,jiali,mgn1,radams,jwang}@psu.edu)

Annotator ID	Valence	Reliability
3474	5.1/8	0.08
2500	0.0/8	0.56
3475	0.0/8	0.34
2540	8.0/8	0.04

Image Confidence: 75% ($\leq$ 90%)

Figure 1. An example illustrating one may need to acquire more reliable labels, ensuring the image confidence is more than 0.9.

voting or (weighted) averaging to force an "objective truth" of the emotional response or probably for the sake of convenience, as is routinely done in affective computing so that classification on a single quantity can be carried out, is a crude treatment bound to erase or disregard information essential for many interesting psychological studies, *e.g.*, to discover connections between varied affective responses and varied demographics.

The involvement of spammers as participating subjects introduces an extra source of variation to the emotional responses, which unfortunately is tangled with the "appropriate" variation. If responses associated with an image stimulus contain answers by spammers, the inter-annotator variation for the specific question could be as large as the variation across different questions, reducing the robustness of any analysis. An example is shown in Fig. 1. Most annotators labeling this image are deemed unreliable, and two of them are highly susceptible as spammers according to our model. Investigators may be recommended to eliminate this image or acquire more reliable labels for its use. Yet, one should not be swayed by this example into the practice of discarding images that solicited responses of a large range. Certain images are controversial in nature and will stimulate quite different emotions to different viewers. Our system acquired the reliability scores shown in Fig. 1 by examining the entire data set; the data on this image alone would not be conclusive, in fact, far from so.

Facing the intertwined "appropriate" and "inappropriate" variations in the subjects as well as the variations in the images, we are motivated to unravel the sources of uncertainties by taking a global approach. The judgment on the reliability of a subject cannot be a per-image decision, and has to leverage the whole data. Our model was constructed to integrate these uncertainties, attempting to discern them with the help of big data. In addition, due to the lack of ground truth labels, we model the relational data that code whether two subjects' emotion responses on an image agree, bypassing the thorny questions of what the true labels are and if they exist at all.

For the sake of automated emotion analysis of images, one also needs to narrow the scope to parts of data, each of which have sufficient number of qualified labels. Our work computes image confidences, which can support off-line data filtering or guide on-line budgeted crowdsourcing practices.

In summary, systematic analysis of crowdsourced affective data is of great importance to human subject studies and affective computing, while remains an open question. To substantially address the aforementioned challenges and expand the evidential space for psychological studies, we propose a probabilistic approach, called **Gated Latent Beta Allocation (GLBA)**. This method computes maximum a posteriori probability (MAP) estimates of each subject's reliability and regularity based on a variational expectation-maximization (EM) framework. With this method, investigators running affective human subject studies can substantially reduce or eliminate the contamination caused by spammers, hence improve the quality and usefulness of collected data (Fig. 2).

1.1 Related Work

Estimating the reliability of subjects is necessary in crowdsourcing-based data collection because the incentives of participants and the interest of researchers diverge. There were two levels of assumptions explored for the crowdsourced data, which we name as the first-order assumption (A1) and the second-order assumption (A2). Let a task be the provision of emotion responses for one image. Consider a task or test conducted by a number of participants. Their responses within this task form a subgroup of data.

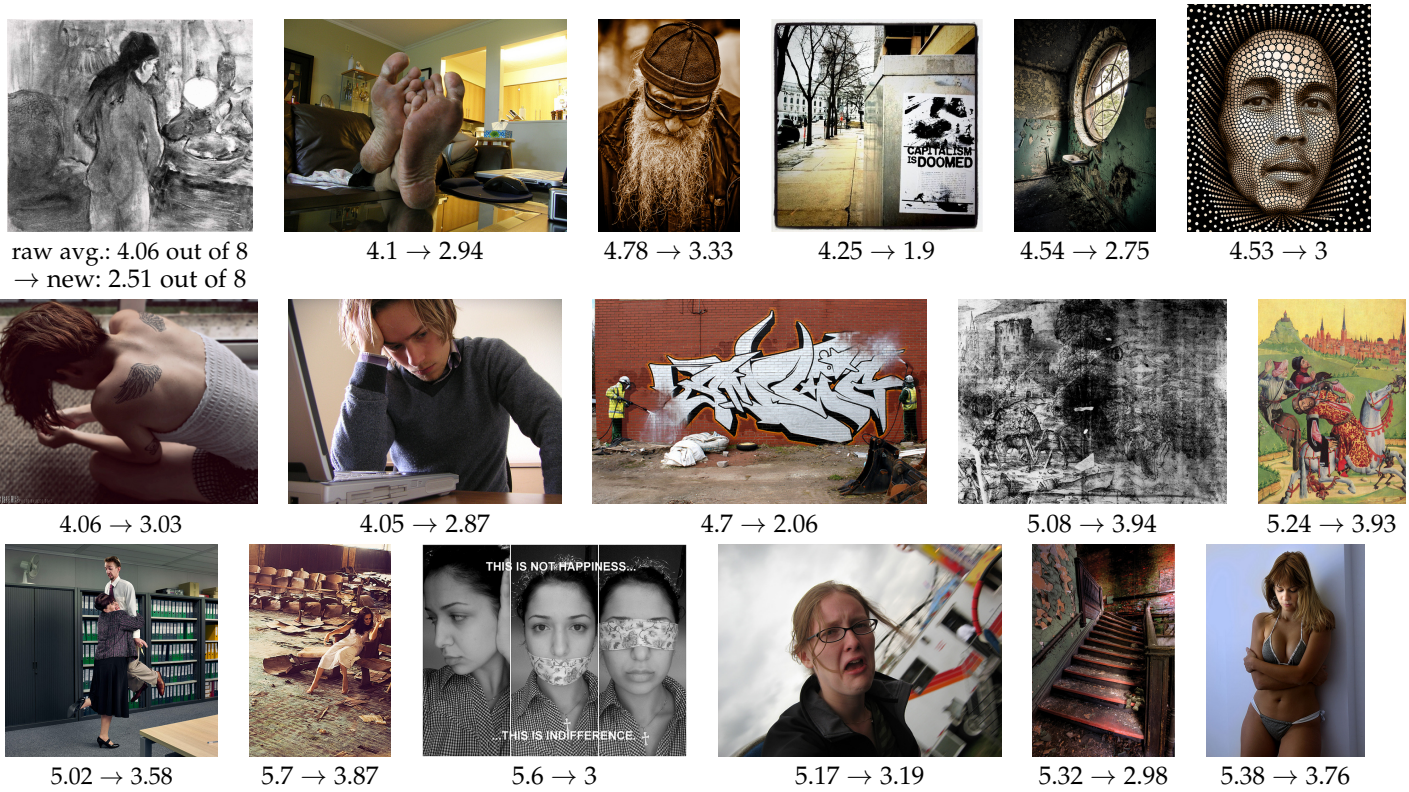
A1 There exists a true label of practical interest for each task. The dependencies between collected labels are mediated by this unobserved true label, of which noisy labels are otherwise conditionally independent.

A2 The uncertainty model for a subgroup of data does not depend on its actual specified task. The performance of a participant is consistent across subgroups of data subject to a single fixed effect.

Existing approaches that model the complexities of tasks or reliability of participants often require one or both of these two assumptions. Under the umbrella of assumption A1, most probabilistic approaches using the observer models [9, 10, 11, 12] focus on estimating the ground truth from multiple noisy labels. For example, the modeling of one reliability parameter per subject is an established practice for estimating the ground truth label [12]. For the case of categorical labels, modeling of one free parameter per class per subject is a more general approach [9, 13]. Our approach does not model the ground truth of labels, hence it is not viable to compare our approach with other methods in this regard. Instead, we sidestep this issue to tackle whether the labels from one subject can agree with labels from another on a single task. Agreement is judged subject to a preselected criterion. Such treatment may be more realistic as a means to process sparse ordinal labels for each task.

Assumption A2 is also widely exploited among methods, often conditioned on A1. It assumes that all of the tasks have the same level of difficulty [14, 15]. Modeling one difficulty parameter per task has been explored in [16] for categorical labels. However, in our approach, task difficulty is modeled as a random effect without subscribing a task-specific parameter. Wisely choosing the modeling complexity and assumptions should be based on availability and purity of data. As suggested in [17], more complexity in a model could challenge the statistical estimation subject to the constraint of real data. Choices with respect to our model attempted to properly analyze the affective data we obtained.

If the mutual agreement rate between two participants does not depend on the actual specified task (*i.e.*, when A2 holds), we can essentially convert the resulting problem to a graph mining problem, where subjects are vertices, agreements are edges, and the proximity between subjects is modeled by how likely they agree with each other in a general sense. Probabilistic models for such relational data can be traced back to early stochastic blockmodels [18, 19], latent space model [20], and their later extensions with mixed membership [21, 22] and



raw avg.: 4.06 out of 8 → new: 2.51 out of 8 | 4.1 → 2.94 | 4.78 → 3.33 | 4.25 → 1.9 | 4.54 → 2.75 | 4.53 → 3

4.06 → 3.03 | 4.05 → 2.87 | 4.7 → 2.06 | 5.08 → 3.94 | 5.24 → 3.93

5.02 → 3.58 | 5.7 → 3.87 | 5.6 → 3 | 5.17 → 3.19 | 5.32 → 2.98 | 5.38 → 3.76

Figure 2. Images shown are considered of lower valence than their average valence ratings (*i.e.*, evoking a higher degree of negative emotions) after processing the data set using our proposed method. Our method eliminates the contamination introduced by spammers. The range of valence ratings is between 0 and 8.

2.63 → 3.77 | 2.8 → 4.14 | 3.0 → 4.7 | 4.4 → 6.21 | 4.7 → 6.26

Figure 3. Images shown are considered of higher valence than their average valence ratings (*i.e.*, evoking a higher degree of positive emotions) after processing the data set using our proposed method. Our method again eliminates the contamination introduced by spammers. The range of valence ratings is between 0 and 8.

nonparametric Bayes [23]. We adopt the idea of mixed memberships wherein two particular modes of memberships are modeled for each subject, one being the reliable mode and the other the random mode. For the random mode, the behavior is assumed to be shared across different subjects, whereas the regular behaviors of subjects in the reliable mode are assumed to be different. Therefore, we can extend this framework from graph to multigraph in the interest of crowdsourced data analysis. Specifically, data are collected as subgroups, each of which is composed of a small agreement graphs for a single task, such that the covariate within a subgroup is modeled. Our approach does not rely on A2. Instead, it models the random effects added to subjects' performance in each task via the multigraph approach. Assumption A1 and A2 implies a bipartite graph structure between tasks and subjects. In contrast, our approach starts from the multigraph structure among subjects that is coordinated by tasks. Finding the proper and flexible structure that data possess is crucial for modeling [24].

1.2 Our Contributions

To our knowledge, this is the first attempt to connect probabilistic observer models with probabilistic graphs, and to explore modeling at this complexity from the joint perspective. We summarize our contributions as follows:

- We developed a probabilistic multigraph model to analyze crowdsourced data and its approximate variational EM algorithm for estimation. The new method, accepting the intrinsic variation in subjective responses, does not assume the existence of ground truth labels, in stark contrast to previous work having devoted much effort to obtain objective true labels.
- Our method exploits the relational data in the construction and application of the statistical model. Specifically, instead of the direct labels, the pair-wise status of agreement between labels given by different subjects is used. As a result, the multigraph agreement model is naturally applicable to more flexible types of responses, easily going

beyond binary and categorical labels. Our work serves as a proof of concept for this new relational perspective.
- Our experiments have validated the effectiveness of our approach on real-world affective data. Because our experimental setup was of a larger scale and more challenging than settings addressed by existing methods, we believe our method can fill some gaps for demands in the practical world, for instance, when gold standards are not available.

2 The Method

In this section, we describe our proposed method. Let us present the mathematical notations first. A symbol with subscript omitted always indicates an array, *e.g.*, $x = (\dots, x_i, \dots)$. The arithmetic operations perform over arrays in the element-wise manner, *e.g.*, $x + y = (\dots, x_i + y_i, \dots)$. Random variables are denoted as capital English letters. The tilde sign indicates the value of parameters in the last iteration of EM, *e.g.*, $\tilde{\theta}$. Given a function f_θ, we denote $f_{\tilde{\theta}}$ by $\tilde{f}_\theta$ or simply $\tilde{f}$, if the parameter $\tilde{\theta}$ is implied. Additional notations, as summarized in Table 1, will be explained in more details later.

Table 1
Symbols and descriptions of parameters, random variables, and statistics.

Symbols	Descriptions
O_i	subject i
τ_i	rate of subject reliability
α_i, β_i	shape of subject regularity
γ	rate of agreement by chance
Θ	union of parameters
$T_j^{(k)}$	whether O_j reliably response
$J_i^{(k)}$	rate of O_i agreeing with other reliable responses
$I_{i,j}^{(k)}$	whether O_i agrees with the responses from O_j
$\omega_i^{(k)}(\cdot)$	cumulative degree of responses agreed by O_i
$\psi_i^{(k)}(\cdot)$	cumulative degree of responses
$r_j^{(k)}(\cdot)$	a ratio amplifies or discounts the reliability of O_j
$\tilde{\tau}_i^{(k)}$	sufficient statistics of posterior $T_i^{(k)}$, given $\tilde{\Theta}$
$\tilde{\alpha}_i^{(k)}, \tilde{\beta}_i^{(k)}$	sufficient statistics of posterior $J_i^{(k)}$, given $\tilde{\Theta}$

2.1 Agreement Multigraph

We represent the data as a directed multigraph, which does not assume a particular type of crowdsourced response. Suppose we have prepared m questions in the study, the answers can be binary, categorical, ordinal, and multidimensional. Given a subject pair (i, j) who are asked to look at the k-th question, one designs an agreement protocol that determines whether the answer from subject i agrees with that from subject j. If subject i's agrees with subject j's on task k, then we set $I_{i,j}^{(k)} = 1$. Otherwise, $I_{i,j}^{(k)} = 0$.

In our case, we are given ordinal data from multiple channels, we define $I_{i,j}^{(k)} = 1$ if (sum of) the percentile difference between two answers $a_i, a_j \in \{1, \dots, A\}$ satisfies

$$\frac{1}{2}\left|P\left[a_i^{(k)}\right] - P\left[a_j^{(k)}\right]\right| + \frac{1}{2}\left|P\left[a_i^{(k)}+1\right] - P\left[a_j^{(k)}+1\right]\right| \leq \delta, \tag{1}$$

The percentile $P[\cdot]$ is calculated from the whole pool of answers for each discrete value, and $\delta = 0.2$. In the above equation, we measure the percentile difference between a_i and a_j as well as that between $a_i + 1$ and $a_j + 1$ in order to reduce the effect of imposing discrete values on the answers that are by nature continuous. If the condition does not hold, they disagree and $I_{i,j}^{(k)} = 0$. Here we assume that if two scores for the same image are within a 20% percentile interval, they are considered to reach an agreement. Compared with setting a threshold on their absolute difference, such rule adapts to the non-uniformity of score distribution. Two subjects can agree with each other by chance or they indeed experience similar emotions in response to the same visual stimulus.

While the choice of the percentile threshold δ is inevitably subjective, the selection in our experiments was guided by the desire to trade-off the preservation of the original continuous scale of the scores (favoring small values) and a sufficient level of error tolerance (favoring large values). This threshold controls the sparsity level of the multi-graph, and influences the marginal distribution of estimated parameters. Alternatively, one may assess different values of the threshold and make a selection based on some other criteria of preference (if exist) applied to the final results.

2.2 Gated Latent Beta Allocation

This subsection describes the basic probabilistic graphical model we used to jointly model subject reliability, which is independent from the supplied questions, and regularity. We refrain from carrying out a full Bayesian inference because it is impractical to end users. Instead, we use the mode(s) of the posterior as point estimates.

We assume each subject i has a reliability parameter $\tau_i \in [0, 1]$ and regularity parameters $\alpha_i, \beta_i > 0$ characterizing his or her agreement behavior with the population, for $i = 1, \dots, m$. We also use parameter γ for the rate of agreement between subjects out of pure chance. Let $\Theta = (\{\tau_i, \alpha_i, \beta_i\}_{i=1}^m, \gamma)$ be the set of parameters. Let Ω_k be the a random sub-sample from subjects $\{1, \dots, m\}$ who labeled the stimulus k, where $k = 1, \dots, n$. We also assume sets Ω_k's are created independently from each other. For each image k, every subject pair from Ω_k^2, *i.e.*, (i, j) with $i \neq j$, has a binary indicator $I_{i,j}^{(k)} \in \{0, 1\}$ coding whether their opinions agree on the respective stimulus. We assume $I_{i,j}^{(k)}$ are generated from the following probabilistic process with two latent variables. The first latent variable $T_j^{(k)}$ indicates whether subject O_j is reliable or not. Given that it is binary, a natural choice of model is the Bernoulli distribution. The second latent variable $J_i^{(k)}$, lying between 0 and 1, measures the extent subject O_i agrees with the other reliable responses. We use Beta distribution parameterized by α_i and β_i to model $J_i^{(k)}$ because it is a widely used parametric distribution for quantities on interval $[0, 1]$ and the shape of the distribution is relatively flexible. In a nutshell, $T_j^{(k)}$ is a latent switch (aka, gate) that controls whether $I_{i,j}^{(k)}$ can be used for the posterior inference of the latent variable $J_i^{(k)}$. Hence, we call our model *Gated Latent Beta Allocation* (GLBA). A graphical illustration of the model is shown in Fig. 4.

We now present the mathematical formulation of the model. For $k = 1, \dots, n$, we generate a set of random variables

independently via

$$T_j^{(k)} \quad i.i.d. \quad \sim \quad \text{Bernoulli}(\tau_j), \quad j \in \Omega_k \ , \tag{2}$$

$$J_i^{(k)} \quad i.i.d. \quad \sim \quad \text{Beta}(\alpha_i, \beta_i), \quad i \in \Omega_k \ , \tag{3}$$

$$I_{i,j}^{(k)} \left| T_j^{(k)}, J_i^{(k)} \right. \quad \sim \quad \begin{cases} \text{Bernoulli}\left(J_i^{(k)}\right) & \text{if } T_j^{(k)} = 1 \\ \text{Bernoulli}(\gamma) & \text{if } T_j^{(k)} = 0 \end{cases} \tag{4}$$

where the last random process holds for any $j \in \Omega_k^{\neg i} := \Omega_k - \{i\}$ and $i \in \Omega_k$ with $k = 1, \ldots, n$, and γ is the rate of agreement by chance if one of i, j turns out to be unreliable. Here $\{I_{i,j}^{(k)}\}$ are observed data.

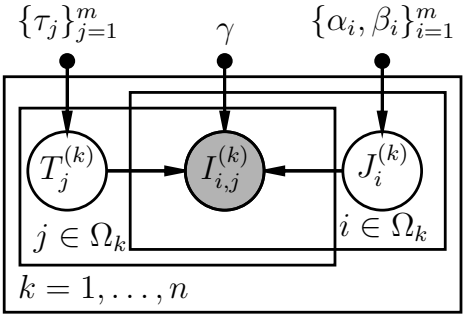


Figure 4. Probabilistic graphical model of the proposed Gated Latent Beta Allocation.

If a spammer is in the subject pool, his or her reliability parameter τ_i is zero, though others can still agree with his or her answers by chance at rate γ. On the other hand, if one is very reliable yet often provides controversial answers, his reliability τ_i can be one, while he typically disagrees with others, indicated by his high irregularity $\mathbb{E}[J_i^{(k)}] = \frac{\alpha_i}{\alpha_i + \beta_i} \approx 0$. We are interested in finding both types of subjects. However, most of subjects lie in between these two extremes.

As an interesting note, Eq. (4) is asymmetric, meaning that $I_{i,j}^{(k)} \neq I_{j,i}^{(k)}$ is possible, a scenario that should never occur by definitions of the two quantities. We propose to achieve symmetry in the final model by using the conditional distribution of $I_{i,j}^{(k)}$ and $I_{j,i}^{(k)}$ given that $I_{i,j}^{(k)} = I_{j,i}^{(k)}$, and call this model the symmetrized model. With details omitted, we state that conditioned on $T_i^{(k)}$, $T_j^{(k)}$, $J_i^{(k)}$, and $J_j^{(k)}$, the symmetrized model is still a Bernoulli distribution:

$$I_{i,j}^{(k)} \sim \text{Bernoulli}\left(H\left(\left(J_i^{(k)}\right)^{T_i^{(k)}} \gamma^{1-T_i^{(k)}}, \left(J_j^{(k)}\right)^{T_j^{(k)}} \gamma^{1-T_j^{(k)}}\right)\right), \tag{5}$$

where

$$H(p, q) = \frac{pq}{pq + (1-p)(1-q)} \ .$$

We tackle the inference and estimation of the asymmetric model for simplicity.

2.3 Variational EM

Variational inference is an optimization based strategy for approximating posterior distribution in complex distributions [25]. Since the full posterior is highly intractable, we consider to use variational EM to estimate the parameters $\Theta = (\{\tau_i, \alpha_i, \beta_i\}_{i=1}^m, \gamma)$ [26]. The parameter γ is assumed to be pre-selected by the user and does not need to be estimated. To regularize the other parameters in estimation, we use the empirical Bayes approach to choose priors. Assume the following priors

$$\tau_i \quad \sim \quad \text{Beta}(\tau_0, 1-\tau_0) \ , \tag{6}$$

$$\alpha_i + \beta_i \quad \sim \quad \text{Gamma}(2, s_0) \ . \tag{7}$$

By empirical Bayes, τ_0, s_0 are adjusted. For the ease of notations, we define two auxiliary functions $\omega_i^{(k)}(\cdot)$ and $\psi_i^{(k)}(\cdot)$:

$$\omega_i^{(k)}(x) := \sum_{j \in \Omega_k^{\neg i}} x_j I_{i,j}^{(k)}, \quad \psi_i^{(k)}(x) := \sum_{j \in \Omega_k} x_j \ . \tag{8}$$

Similarly, we define their siblings

$$\bar{\omega}_i^{(k)}(x) = \omega_i^{(k)}(1-x), \quad \bar{\psi}_i^{(k)}(x) = \psi_i^{(k)}(1-x) \ . \tag{9}$$

We also define the auxiliary function $r_j(\cdot)$ as

$$r_j^{(k)}(x) = \prod_{i \in \Omega_k^{\neg j}} \left(\frac{x_i}{\gamma}\right)^{I_{i,j}^{(k)}} \left(\frac{1-x_i}{1-\gamma}\right)^{1-I_{i,j}^{(k)}} . \tag{10}$$

Now we define the full likelihood function:

$$L_k(\Theta; T^{(k)}, J^{(k)}, I^{(k)}) := \prod_{j \in \Omega_k} \left((\tau_j)^{T_j^{(k)}} (1-\tau_j)^{1-T_j^{(k)}}\right) \cdot \prod_{i \in \Omega_k} \frac{\left(J_i^{(k)}\right)^{\alpha_i^{(k)}} \left(1-J_i^{(k)}\right)^{\beta_i^{(k)}} \phi_i^{(k)}}{B(\alpha_i, \beta_i)} , \tag{11}$$

where auxiliary variables simplifying the equations are

$$\begin{aligned} \alpha_i^{(k)} &= \alpha_i + \omega_i^{(k)}\left(T^{(k)}\right) \ , \\ \beta_i^{(k)} &= \beta_i + \psi_i^{(k)} - \omega_i^{(k)}\left(T^{(k)}\right) \ , \\ \phi_i^{(k)} &= \gamma^{\bar{\omega}_i^{(k)}(T^{(k)})} (1-\gamma)^{\bar{\psi}_i^{(k)}(T^{(k)}) - \bar{\omega}_i^{(k)}(T^{(k)})} \ , \end{aligned}$$

and $B(\cdot, \cdot)$ is the Beta function. Consequently, assume the prior likelihood is $L_\Theta(\Theta)$, the MAP estimate of Θ is to minimize

$$L(\Theta; T, J, I) := L_\Theta(\Theta) \prod_{k=1}^n L_k(\Theta; T^{(k)}, J^{(k)}, I^{(k)}) \ . \tag{12}$$

We solve the estimation using variational EM method with a fixed (τ_0, s_0) and varying γ. The idea of variational methods is to approximate the posterior by a factorizable template, whose probability distribution minimizes its KL divergence to the true posterior. Once the approximate posterior is solved, it is then used in the E-step in the EM algorithm as the alternative to the true posterior. The usual M-step is unchanged. Each time Θ is estimated, we adjust prior (τ_0, s_0) to match the mean of the MAP estimates of $\{\tau_i\}$ and $\left\{\frac{\alpha_i + \beta_i}{2}\right\}$ respective until they are sufficiently close.

E-step. We use the factorized Q-approximation with variational principle:

$$p_\Theta\left(T^{(k)}, J^{(k)} \left| I^{(k)} \right.\right) \approx \prod_{j \in \Omega_k} q^*_{T_j, \Theta}\left(T_j^{(k)}\right) \prod_{i \in \Omega_k} q^*_{J_i, \Theta}\left(J_i^{(k)}\right) . \tag{13}$$

- Let

$$q^*_{T_j,\Theta}\left(T_j^{(k)}\right) \propto \exp\left(\mathbb{E}_{J,T^{\neg j}}\left[\log L_k\left(\Theta; T^{(k)}, J^{(k)}, I^{(k)}\right)\right]\right) , \quad (14)$$

whose distribution can be written as

$$\text{Bernoulli}\left(\frac{\tau_j R_j^{(k)}}{\tau_j R_j^{(k)} + 1 - \tau_j}\right),$$

where $\log R_j^{(k)} = \mathbb{E}_J\left[\sum_{i\in\Omega_k^{\neg j}} \log\left(r_i^{(k)}(J^{(k)})\right)\right]$. As suggested by Johnson and Kotz [27], the geometric mean can be numerically approximated by

$$R_j^{(k)} \approx \prod_{i\in\Omega_k^{\neg j}} \frac{1}{\alpha_i^{(k)} + \beta_i^{(k)}} \left(\frac{\alpha_i^{(k)}}{\gamma}\right)^{I_{i,j}^{(k)}} \left(\frac{\beta_i^{(k)}}{1-\gamma}\right)^{1-I_{i,j}^{(k)}}, \quad (15)$$

if both $\alpha_i^{(k)}$ and $\beta_i^{(k)}$ are sufficiently larger than 1.

- Let

$$q^*_{J_i,\Theta}(J_i^{(k)}) \propto \exp\left(\mathbb{E}_{T,J^{\neg i}}\left[\log L_k\left(\Theta; T^{(k)}, J^{(k)}, I^{(k)}\right)\right]\right) , \quad (16)$$

whose distribution is

$$\text{Beta}(\alpha_i + \omega_i^{(k)}(\tau), \beta_i + \psi_i^{(k)}(\tau) - \omega_i^{(k)}(\tau)) .$$

Given parameter $\tilde{\Omega} = \{\tilde{\tau}_i, \tilde{\alpha}_i, \tilde{\beta}_i\}_{i=1}$, we can compute the approximate posterior expectation of the log likelihood, which reads

$$\begin{aligned}
&\mathbb{E}_{T,J|\tilde{\Theta},I} \log L_k(\Theta; T^{(k)}, J^{(k)}, I^{(k)}) \approx \\
&\text{const.} + \log L_\Theta(\Theta) + \\
&\sum_{j\in\Omega_k}\left(\tilde{\tau}_i^{(k)} \log\tau_j + (1-\tilde{\tau}_i^{(k)})\log(1-\tau_j)\right) + \\
&\sum_{i\in\Omega_k}\left\langle \begin{pmatrix}\alpha_i\\ \beta_i\end{pmatrix}, \frac{\nabla B(\tilde{\alpha}_i^{(k)}, \tilde{\beta}_i^{(k)})}{B(\tilde{\alpha}_i^{(k)}, \tilde{\beta}_i^{(k)})}\right\rangle - \\
&\sum_{i\in\Omega_k} \log B(\alpha_i, \beta_i) + \log\gamma \sum_{i\in\Omega_k} \bar{\omega}_i^{(k)}\left(\tilde{\tau}_i^{(k)}\right) + \\
&\log(1-\gamma)\sum_{i\in\Omega_k}\left(\bar{\psi}_i^{(k)}\left(\tilde{\tau}_i^{(k)}\right) - \bar{\omega}_i^{(k)}\left(\tilde{\tau}_i^{(k)}\right)\right), \quad (17)
\end{aligned}$$

where relevant statistics are defined as

$$\begin{aligned}
\tilde{\alpha}_i^{(k)} &= \tilde{\alpha}_i + \omega_i^{(k)}(\tilde{\tau}) , \\
\tilde{\beta}_i^{(k)} &= \tilde{\beta}_i + \psi_i^{(k)}(\tilde{\tau}) - \omega_i^{(k)}(\tilde{\tau}) , \text{ and} \quad (18)\\
\tilde{\tau}_i^{(k)} &= \frac{\tilde{R}_i^{(k)}\tilde{\tau}_i}{\tilde{R}_i^{(k)}\tilde{\tau}_i + 1 - \tilde{\tau}_i} .
\end{aligned}$$

Remark $B(\cdot,\cdot)$ is the Beta function, and $\tilde{R}_i^{(k)}$ is calculated from approximation Eq. (15)

M-step. Compute the partial derivatives of L with respect to α_i and β_i: let Δ_i be the set of images that are labeled by subject i. We set $\partial L/\partial\alpha_i = 0$ and $\partial L/\partial\beta_i = 0$ for each i, which reads

$$\begin{aligned}
&\left(\frac{\alpha_i+\beta_i}{s_0} - \log(\alpha_i+\beta_i)\right)\cdot\begin{pmatrix}1\\1\end{pmatrix} \\
&= \sum_{k\in\Delta_i}\frac{\nabla B(\tilde{\alpha}_i^{(k)}, \tilde{\beta}_i^{(k)})}{B(\tilde{\alpha}_i^{(k)}, \tilde{\beta}_i^{(k)})} - \frac{\nabla B(\alpha_i,\beta_i)}{B(\alpha_i,\beta_i)} \\
&= \sum_{k\in\Delta_i}\begin{pmatrix}\Psi(\tilde{\alpha}_i^{(k)}) - \Psi(\tilde{\alpha}_i^{(k)} + \tilde{\beta}_i^{(k)}) \\ \Psi(\tilde{\beta}_i^{(k)}) - \Psi(\tilde{\alpha}_i^{(k)} + \tilde{\beta}_i^{(k)})\end{pmatrix} \\
&\quad -|\Delta_i|\cdot\begin{pmatrix}\Psi(\alpha_i) - \Psi(\alpha_i+\beta_i)\\ \Psi(\beta_i) - \Psi(\alpha_i+\beta_i)\end{pmatrix} , \quad (19)
\end{aligned}$$

where $\Psi(x) \in [\log(x-1), \log x]$ is the Digamma function. The above two equations can be practically solved by Newton-Raphson method with a projected modification (ensuring α, β always are greater than zero).

Compute the derivatives of L with respect to τ_i and set $\partial L/\partial\tau_i = 0$, which reads

$$\tau_i = \frac{1}{|\Delta_i|+1}\left(\tau_0 + \sum_{k\in\Delta_i}\tilde{\tau}_i^{(k)}\right) . \quad (20)$$

Compute the derivatives of L w.r.t. γ and set to zero, which reads

$$\gamma = \frac{\sum_{i\in\Omega_k}\bar{\omega}_i^{(k)}(\tilde{\tau}_i^{(k)})}{\sum_{i\in\Omega_k}\bar{\psi}_i^{(k)}(\tilde{\tau}_i^{(k)})} . \quad (21)$$

In practice, the update formula for γ needs not to be used if γ is pre-fixed. See Algorithm 1 for details.

2.4 The Algorithm

We present our final algorithm to estimate all parameters by knowing the multigraph data $\{I_{i,j}^{(k)}\}$. Our algorithm is designed based on Eqs. (19), (20), and (21). In each EM iteration, there are two loops: one for collecting relevant statistics for each subgraph, and the other for re-computing the parameter estimates for each subject. Please refer to Algorithm 1 for details.

Algorithm 1 Variational EM algorithm of GLBA

Input: A multi-graph $\{I_{i,j}^k \in \{0,1\}\}_{i,j\in\Omega_k}$, $0 < \gamma < 0.5$
Output: subject parameters $\Theta = (\{(\tau_i, \alpha_i, \beta_i)\}_{i=1}^m, \gamma)$

Initialisation : $\tau_0 = 0.5, \alpha_i = \beta_i = \tau_i = 1.0, i = 1, \ldots, m$
repeat
  for $k = 1$ to n **do**
    compute statistics $\tilde{\alpha}_i^{(k)}, \tilde{\beta}_i^{(k)}, \tilde{\tau}_i^{(k)}$ by Eq. (18);
  end for
  for $i = 1$ to m **do**
    solve (α_i, β_i) from Eq. (19) (Newton-Raphson);
    compute τ_i by Eq. (20);
  end for
  (optional) update γ from Eq. (21);
until $\{(\tau_i, \alpha_i, \beta_i)\}_{i=1}^m$ are all converged.
return Θ

3 Experiments

3.1 Data Sets

We studied a crowdsourced affective data set acquired from the Amazon Mechanical Turk (AMT) platform [8]. The affective

data set is a collection of image stimuli and their affective labels including valence, arousal, dominance and likeness (degree of appreciation). Labels for each image are ordinal: {1, ... , 9} for the first three dimensions, and {1, ..., 7} for the likeness dimension. The study setup and collected data statistics have been detailed in [8], which we describe briefly here for the sake of completeness.

At the beginning of a session, the AMT study host provides the subject brief training on the concepts of affective dimensions. Here are descriptions used for valence, arousal, dominance, and likeness.

- Valence: degree of feeling happy vs. unhappy
- Arousal: degree of feeling excited vs. calm
- Dominance: degree of feeling submissive vs. dominant
- Likeness: how much you like or dislike the image

The questions presented to the subject for each image are given below in exact wording.

- Slide the solid bubble along each of the bars associated with the 3 scales (Valence, Arousal, and Dominance) in order to indicate how you ACTUALLY FELT WHILE YOU OBSERVED THE IMAGE.
- How did you like this image? (Like extremely, Like very much, Like slightly, Neither like nor dislike, Dislike slightly, Dislike very much, Dislike extremely)

Each AMT subject is asked to finish a set of labeling tasks, and each task is to provide affective labels on a single image from a prepared set, called the EmoSet. This set contains around 40,000 images crawled from the Internet using affective keywords. Each task is divided into two stages. First, the subject views the image; and second, he/she provides ratings in the emotion dimensions through a Web interface. Subjects usually spend three to ten seconds to view each image, and five to twenty seconds to label it. The system records the time durations respectively for the two stages of each task and calculates the average cost (at a rate of about 1.4 US Dollars per hour). Around 4,000 subjects were recruited in total. For the experiments below, we retained image stimuli that have received affective labels from at least four subjects. Under this screening, the AMT data have 47,688 responses from 2,039 subjects on 11,038 images. Here, one response refers to the labeling of one image by one subject conducted in one task.

Because humans can naturally feel differently from each other in their affective experiences, there was no gold standard criterion to identify spammers. Such a human emotion data set is difficult to analyze and the quality of data is hard to assess. Among several emotion dimensions, we found that participants were more consistent in the valence dimension. As a reminder, valence is the rated degree of positivity of emotion evoked by looking at an image. We call the variance of the ratings from different subjects on the same image the within-task variance, while the variance of the ratings from all the subjects on all the images the cross-task variance. For valence and likeness, the within-task variance accounts for about 70% of the cross-task variance, much smaller than for the other two dimensions. Therefore, the remaining experiments were focused on evaluating the regularity of image valences in the data.

3.2 Baselines for Comparison

We discuss below several baseline methods or models with which we compare our method.

Dawid and Skene [9]. Our method falls into the general category of consensus methods in the literature of statistics and machine learning, where the spammer filtering decision is made completely based on the labels provided by observers. Those consensus methods have been developed along the line of Dawid and Skene [9], and they mainly deal with categorical labels by modeling each observer using a designated confusion matrix. More recent developments of the observer models have been discussed in [17], where a benchmark has shown that the Dawid-Skene method is still quite competitive in unsupervised settings according to a number of real-world data sets for which ground-truth labels are believed to exist albeit unknown. However, this method is not directly applicable to our scenario. To enable comparison with this baseline method, we first convert each affective dimension into a categorical label by thresholding. We create three categories: high, neural, and low, each covering a continuous range of values on the scale. For example, high valence category implies a score greater than a neural score (*i.e.*, 5) by more than a threshold (*e.g.*, 0.5). Such a thresholding approach has been adopted in developing affective categorization systems, *e.g.* [5, 6].

Time duration. In the practice of data collection, the host filtered spammers by a simple criterion—to declare a subject spammer if he spends substantially less time on every task. The labels provided by the identified spammers were then excluded from the data set for subsequent use, and the host also declined to pay for the task. However, some subjects who were declined to be paid wrote emails to the host arguing for their cases. Under this spirit, in our experiments, we form a baseline method that uses the average time duration of each subject to red-flag a spammer.

Filtering based on gold standard examples. A widely used spammer detection approach in crowdsourcing is to create a small set with known ground truth labels and use it to spot anyone who gives incorrect labels. However, such a policy was not implemented in our data collection process because as we argued earlier, there is simply no ground truth for the emotion responses to an image in a general sense. On the other hand, just for the sake of comparison, it seems reasonable to find a subset of images that evoke such extreme emotions that ground truth labels can be accepted. This subset will then serve the role of gold standard examples. We used our method to retrieve a subset of images which evoke extreme emotions with high confidence (see Section 3.7 for confidence score and emotion score calculation). For the valence dimension, we were able to identify at most 101 images with valence score ≥ 8 (on the scale of $1 \ldots 9$) with over 90% confidence and 37 images with valence score ≤ 2 with over 90% confidence. We also looked at those images one by one (as provided in the supplementary materials) and believe that within a reasonable tolerance of doubt those images should evoke clear emotions in the valence dimension. Unfortunately, only a small fraction of subjects in our pool have labeled at least one image from this "gold standard" subset. Among this small group, their disparity from the gold standard enables us to find three susceptible spammers. To see whether these three susceptible spammers can also be detected by our method, we find that their reliability scores $\tau \in [0, 1]$ are $0.11, 0.22, 0.35$ respectively. In Fig. 9, we plot the distribution of τ of the entire subject pool. These three scores are clearly on the low end with respect to the scores of the other subjects. Thus the three spammers are also assessed to be highly susceptible by our model.

In summary, while we were able to compare our method with the first two baselines quantitatively, with results to be presented shortly, comparison with the third baseline is limited due to the way the AMT data were collected [8].

3.3 Model Setup

Since our hypotheses included a random agreement ratio γ that is pre-selected, we adjusted the parameter γ from 0.3 to 0.48 to see empirically how it affects the result in practice.

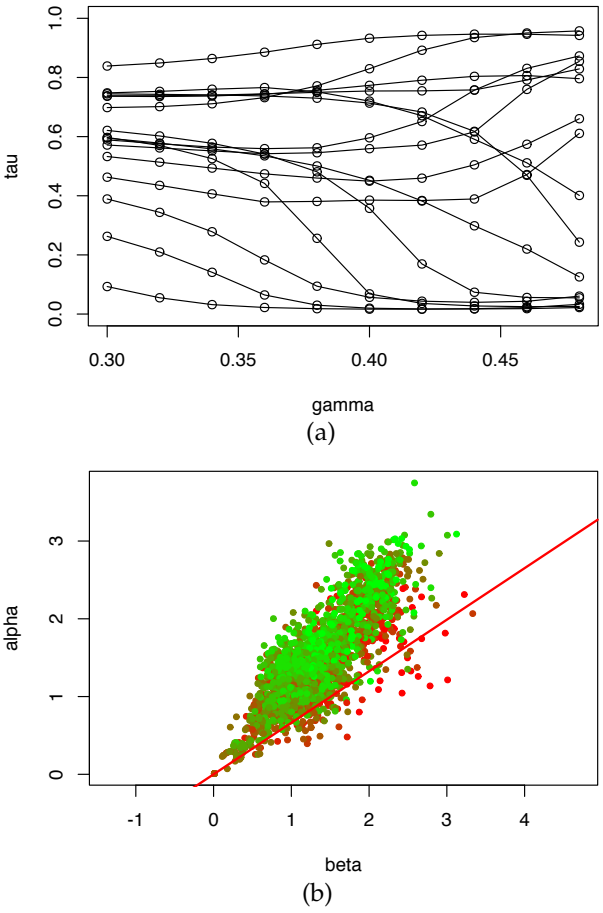


Figure 5. (a) Reliability scores versus $\gamma \in [0.3, 0.48]$ for the top 15 users who provided the most numbers of ratings. (b) Visualization of the estimated regularity parameters of each worker at a given γ. Green dots are for workers with high reliability and red dots for low reliability. The slope of the red line equals γ.

Fig. 5 depicts how the reliability parameter τ varies with γ for different workers in our data set. Results are shown for the top 15 users who provided the most numbers of ratings. Generally speaking, a higher γ corresponds to a higher chance of agreement between workers purely out of random. From the figure, we can see that a worker providing more ratings is not necessarily more reliable. It is quite possible that some workers took advantage of the AMT study to earn monetary compensation without paying enough attention to the actual questions.

In Table 2, we demonstrate the valence, arousal, and dominance labels for two categories of subjects. On the top, the first category contains susceptible spammers with low estimated reliability parameter τ; and on the bottom, the second category contains highly reliable subjects with high values of τ. Each subject takes one row. For the convenience of visualization, we represent the three-dimensional emotion scores given to any image by a particular color whose RGB values are mapped from the values in the three dimensions respectively. The emotion labels for every image by one subject are then condensed into one color bar. The labels provided by each subject for all his images are then shown as a palette in one row. For clarity, the color bars are sorted in lexicographic order of their RGB values. One can clearly see that those labels given by the subjects from these two categories exhibit quite different patterns. The palettes of the susceptible spammers are more extreme in terms of saturation or brightness. The abnormality of label distributions of the first category naturally originates from the fact that spammers intended to label the data by exerting the minimal efforts and without paying attention to the questions.

3.4 Basic Statistics of Manually Annotated Spammers

For each subject in the pool, by observing all his or her labels in different emotion dimensions, there was a reasonable chance of spotting abnormality solely by visualizing the distribution. If one were a spammer, it often happened that his or her labels were highly correlated, skewed or deviated in an extreme manner from a neural emotion along different dimensions. In such cases, it was possible to manually exclude his or her responses from the data due to his or her high susceptibility. We applied this same practice to identifying highly susceptible subjects from the pool. We found about 200 susceptible participants.

We studied several basic statistics of this subset in comparison with the whole population: total number of tasks completed, average time duration spent on image viewing and survey per task. The histograms of these quantities are plotted in Fig. 6. One can see that the annotated spammers did not necessarily spend less time or finish fewer tasks than the others, and the time duration has shown only marginal sensitivity to those annotated spammers (See Fig. 6). The figures demonstrate that those statistics are not effective criteria for spammer filtering.

We will use this subset of susceptible subjects as a "pseudo-gold standard" set for quantitative comparisons of our method and the baselines in the subsequent studies. As explained previously in 3.2, other choices of constructing a gold standard set either conflict the high variation nature of emotion responses or yield only a tiny (of size three) set of spammers.

3.5 Top-K Precision Performance in Retrieving the Real Spammers

We conducted experiments on each affective dimension, and evaluated whether the subjects with the lowest estimated τ were supposed to be real spammers according to the "pseudo-gold standard" subset constructed in Section 3.4. Since there was no gold standard to correctly classify whether one subject was truly a spammer or not, we have been agnostic here. Based on that subset, we were able to partially evaluate the top-K precision in retrieving the real spammers, especially the most susceptible ones.

Specifically, we computed the reliability parameter τ for each subject and chose the K subjects with the lowest values as the most susceptible spammers. Because τ depends on the random agreement rate γ, we computed τ's using 10 values

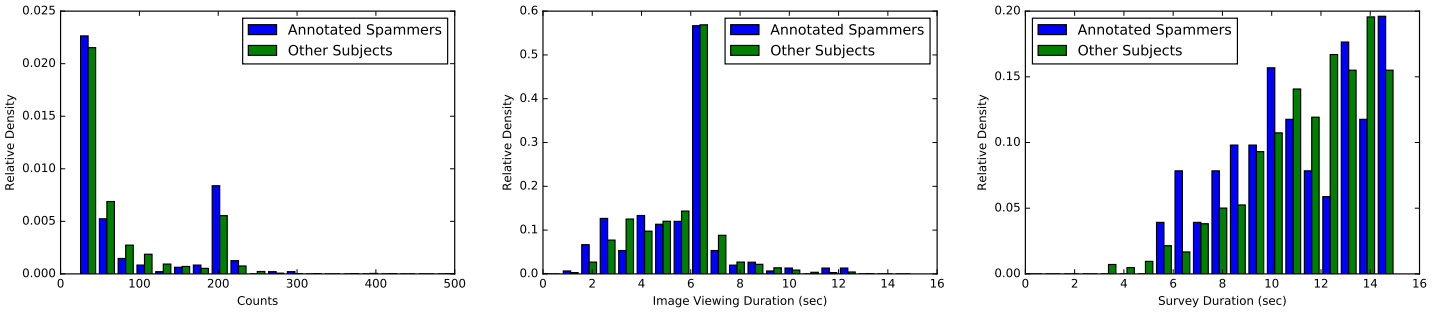


Figure 6. Normalized histogram of basic statistics including total number of tasks completed and average time duration spent at each of the two stages per task.

τ_i	α_i	β_i	reported emotions (sorted)
0.19	1.17	2.43	
0.08	0.75	2.20	
0.08	1.16	2.50	
0.09	0.67	1.70	
0.03	0.94	1.90	
0.17	0.72	1.47	
0.06	1.14	2.50	
0.17	0.86	1.79	
0.04	1.01	2.63	
0.03	1.08	2.84	
0.92	2.29	1.49	
0.94	2.55	1.98	
0.95	2.61	1.68	
0.92	2.40	1.66	
0.91	2.21	1.40	
0.92	2.45	1.97	
0.93	2.38	1.69	
0.93	1.76	1.40	
0.91	2.44	1.86	
0.92	2.30	1.85	
0.92	2.45	1.82	
0.91	1.64	1.29	
0.90	1.68	1.12	
0.91	2.72	2.22	

Table 2
Oracles in the AMT data set. Upper: malicious oracles whose α_i/β_i is among the lowest 30, meanwhile $|\Delta_i|$ is greater than 10. Lower: reliable oracles whose τ_i is among the top 30, meanwhile $\alpha_i/\beta_i > 1.2$. Their reported emotions are visualized by RGB colors. The estimates of Θ is based on the valence dimension.

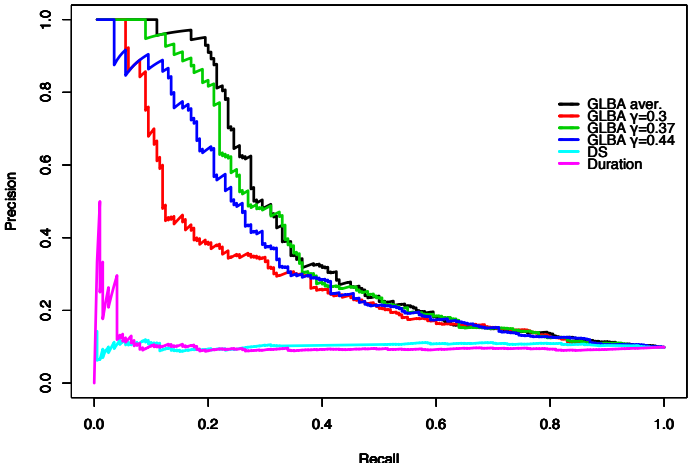


Figure 7. The agnostic Precision-Recall curve (by valence) based on manually annotated spammers. The top 20, top 40 and top 60 precision is $100\%, 95\%, 78\%$ respectively (black line). It is expected that precision drops quickly with increasing recalls, because the manually annotation process can only identify a special type of spammers, while other types of spammers can be identified by the algorithm. The PR curves at $\gamma = 0.3, 0.37, 0.44$ are also plotted. Two baselines are compared: the Dawid and Skene (DS) approach and the time duration based approach.

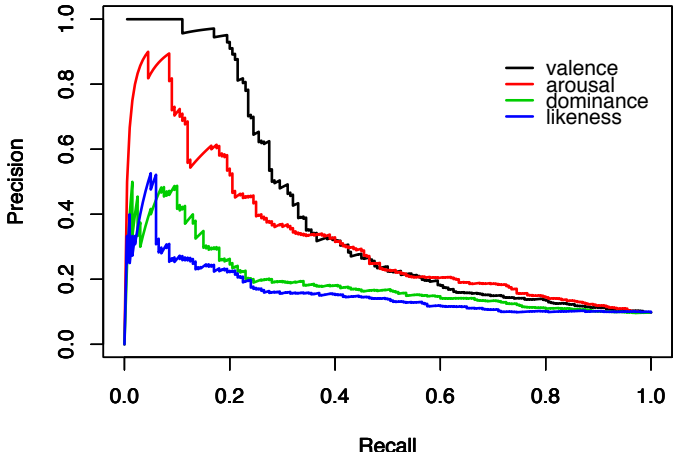


Figure 8. The agnostic Precision-Recall curve based on manually annotated spammers computed from different affective dimensions: valence, arousal, dominance, and likeness.

of γ evenly spaced out over interval $[0.3, 0.48]$. The average value of τ was then used for ranking. The Precision Recall Curves are shown in Fig. 7. Our method achieves high top-K precision by retrieving the most susceptible subjects from the pool according to the average τ. In particular, the top-20 precision is 100%, the top-40 precision is 95%, and the top-60 precision is 78%. Clearly, our algorithm has yielded results well aligned with the human judgment on the most susceptible ones. In Fig. 7, we also plot Precision Recall Curves by fixing γ to $0.3, 0.37, 0.44$ and using the corresponding τ. The result at $\gamma = 0.37$ is better than the other two across recalls, indicating that a proper level of the random agreement rate can be important for achieving the best performance. The two baseline methods are clearly not competitive in this evaluation. The Dawin-Skene method [9], widely used in processing crowdsourced data with objective ground truth labels, drops quickly to a remarkably low precision even at a low recall. The time duration method, used in the practice of AMT host, is better than the Dawin-Skene method, yet substantially worse than the performance of our method.

We also tested this same method of identifying spammers using affective dimensions other than valence. As shown in Fig. 8, the two most discerning dimensions were valence and arousal. It is not surprising that people can reach relatively higher consensus when rating images by these two dimensions than by dominance or likeness. Dominance is much more likely to draw on evidence from context and social situation in most circumstances and hence less likely to have its nature determined to a larger extent by the stimulus itself.

3.6 Recall Performance in Retrieving the Simulated Spammers

The evaluation of top-K precision was limited in two respects: (1) the susceptible subjects were identified because we could clearly observe their abnormality in terms of the multivariate distribution of provided labels. If the participant labeled the data by acting exactly the same as the distribution of the population, we could not manually identify him/her using the aforementioned methodology. (2) We still need to determine if one is a spammer, how likely we are to spot him/her.

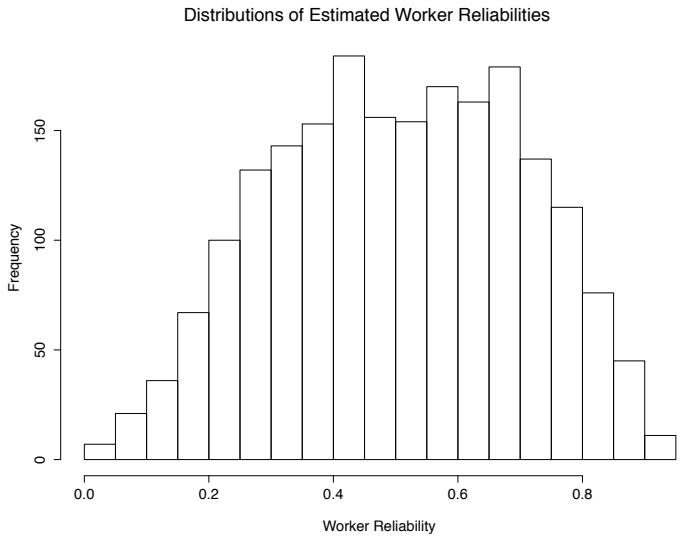
In this study, we simulated several highly "intelligent" spammers, who labeled the data by exactly following the label distribution of the whole population. Every time, we generated 10 spammers, who randomly labeled 50 images. The labels of simulated spammers were not overlapping. We mixed those labels of the simulated spammers with the existing data set, and then conducted our method again to determine how accurate our approach was with respect to finding the simulated spammers. We repeated this process 10 times in order to estimate the τ distribution of the simulated spammers. Results are reported Fig. 9. We drew the histogram of the estimated reliability of all real workers and compared them to the estimated reliability of simulated spammers (in the table included in Fig. 9). We noted that more than half of the simulated spammers were identified as highly susceptible based on the τ estimation (≤ 0.2), and none of them were supposed to have a high reliability score (≥ 0.6). This result validates that our method is robust enough to spot the "intelligent" spammers, even if they disguise themselves as random labelers within a population.

3.7 Qualitative Comparison Based on Controversial Examples

To re-rank the emotion dimensions and likenesses of stimuli with the reliability of the subject accounted for, we adopted the following formula to find the stimuli with "reliably" highest ratings. Assume each rating $a_i \in [0, 1]$. We define the following to replace the usual average:

$$b_k := \underbrace{\frac{\sum_{i\in\Omega_k} \tau_i a_i^{(k)}}{\sum_{i\in\Omega_k} \tau_i}}_{\text{est. score}} \cdot \underbrace{\left(1 - \prod_{i\in\Omega_k} (1-\tau_i)\right)}_{\text{confidence}}, \qquad (22)$$

where $\left(1 - \prod_{i\in\Omega_k}(1-\tau_i)\right) \in [0, 1]$ is the *cumulative confidence score* for image k. This adjusted rating b_k not only allows more reliable subjects to play a bigger role via the weighted average (the first term of the product) but also modulates the weighted average by the cumulative confidence score for the image.



Ranges	≤ 0.2	0.2∼0.4	0.4∼0.6	0.6∼0.8	$\geq$0.8
Counts	54	34	12	0	0

Figure 9. The histogram distribution of estimated worker reliabilities τ and statistics of simulated spammers based on 10 repeated runs, each with 10 spammers injected.

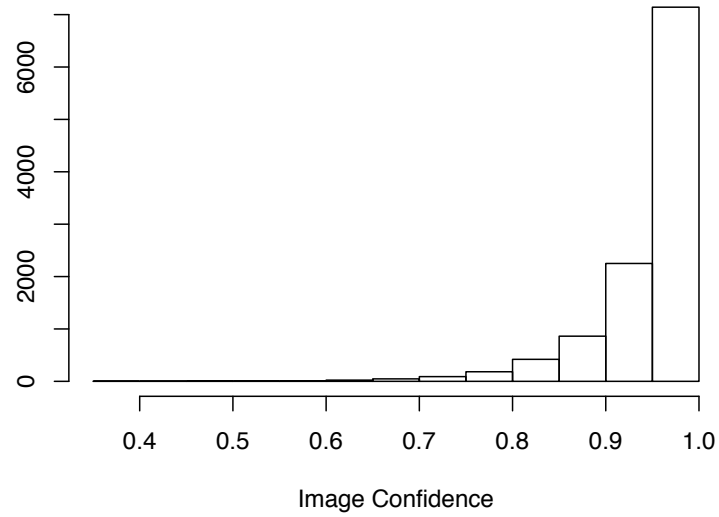


Figure 10. The histogram of image confidences estimated based on our method. About 85% of images have a confidence scores higher than 90%.

Similarly, in order to find those with "reliably" lowest ratings, we replace $a_i^{(k)}$ with $(1 - a_i^{(k)})$ in the above formula and then still seek for the images with the highest b_k's.

If b_k is higher than a neutral level, then the emotional response to the image is considered high. Fig. 10 shows the histogram of image confidence scores estimated by our method. More than 85% of images had acquired a sufficient number of quality labels. To obtain a qualitative sense of the usefulness of the reliability parameter τ, we compared our approach with the simple average-and-rank scheme by examining controversial image examples according to each emotion dimension. Here, being controversial means the assessment of the average emotion response for an image differs significantly between the methods. Despite the variability of human nature, the majority of the population were quite likely to reach consensus for a portion of the stimuli. Therefore, this investigation is meaningful. In Fig. 2 and Fig. 3, we show example image stimuli that were recognized to clearly deviate from neutral emotions by one method but not agreed upon by the other. We skipped stimuli images that were fear inducing, visually annoying or improper. Interested readers can see the complete results in the supplementary material.

3.8 Cost/Overhead Analysis

There is an inevitable trade-off between the quality of the labels and the average cost of acquiring them when screening is applied based on reliability. If we set a higher standard for reliability, the quality of the labels retained tends to improve but we are left with fewer labels to use. It is interesting to visualize the trade-off quantitatively. Let us define overhead numerically as the number of labels removed from the data set when quality control is imposed; and let the threshold on either subject reliability or image confidence used to filter labels be the index for label quality. We obtained what we call *overhead curve* in Figure 11. On the left plot, the result is based on filtering subjects with reliability scores below a threshold (all labels given by such subjects are excluded); on the right, it is based on filtering images with confidence scores below a threshold. As shown by the plots, if either the labels from subjects with reliability scores below 0.3 are discarded or those for images with confidence scores below 90% are discarded, roughly 10,000 out of 47,688 labels are deemed unusable. At an even higher standard, e.g., subject reliability $\geq$ .5 or image confidence level $\geq$ 95%, around half of the labels will be excluded from the data set. Although this means the average per label cost is doubled at the stringent quality standard, we believe the screening is worthwhile in comparison with analysis misled by wrong data. In a large-scale crowdsource environment, it is simply impractical to expect all the subjects to be fully serious. This contrasts starkly with a well-controlled lab environment for data collection. In a sense, post-collection analysis of data to ensure quality is unavoidable. It is indeed a matter of which analysis should be applied.

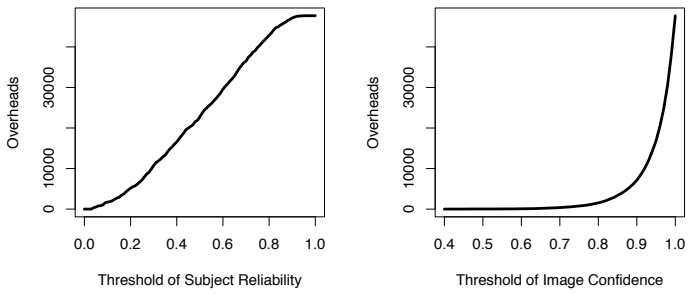


Figure 11. Left: Overhead curve based on subject filtering; Right: overhead curve based on image filtering. The overhead is quantified by the number of labels discarded after filtering.

4 Discussions

Underlying Principles: Our approach to assess the reliability of crowdsourced affective data deviates fundamentally from the standard approaches much concerned with hunting for "ground truth" emotion stimulated by an image. An individual's emotion response is expected to be naturally different

because it depends on subjective opinions rooted in the individual's lifetime exposure to images and concepts, a topic having been pursued long in the literature of social psychology. The new principle we adopted here focuses on the relational knowledge about the ratings of the subjects. Our analysis steps away from the use of "ground truth" by recasting the data as relational quantities.

As pointed out by a reviewer, such a relational perspective may be intrinsic in human cognition, going beyond our specific problem here. For instance, the same spirit of exploiting relationships has already appeared in studies to understand linguistic learning. Gentner [28, 29] proposed that one should understand linguistic learning in a relational way. Instead of assuming there are well-formed abstract language concepts to grasp, the human's cognitive ability often starts from analogical processing based on examples of a concept, and then utilizes the symbolic systems (languages) to reinforce and guide the learning, and to facilitate memory of the acquired concepts. The relationships among the examples and the abstract concept play a role in learning hand in hand, refining recursively the understanding of each other. The whole process is an interlocked and repeated improvement of one side assisted by the other. In a similar fashion, our system improves its assessment about which images evoke highly consensus emotion responses and which subjects are reliable. At the beginning, the lack of either kind of information obscures the truth about the other. Or equivalently, knowing either makes the understanding of the other easy. This is a chicken-and-egg situation. Like the proposed way of learning languages, our system pulls out of the dilemma by recursively enhancing the understanding of one side conditioned on what has been known about the other.

Results: We found that the crowdsourced affective data we examined are particularly challenging for the conventional school of observer models, developed along the line of Dawid and Skene [9]. We identified two major reasons. First, each image in our data set has a much smaller number of observers, compared with what are typically studied in the benchmarks [17]. In our data set, most images were only labeled by 4 to 8 subjects, while many existing benchmark data sets have tens of subjects per task. Second, a more profound reason is that most images do not have a ground truth affective label at the first place. This can render ineffective many statistical methods which model the user-task confusion matrix and hence count on the existence of "true" labels and the fixed characteristics of uncertainty in responses (assumptions A1 and A2).

Our experiments demonstrate that valence and arousal are the two most effective dimensions that can be used to analyze the reliability of subjects. Although subjects may not reach a consensus at local scales (say, an individual task) because the emotions are inherently subjective, consensus at a global scale can still be well justified.

Usage Scenarios: We would like to articulate on the scenarios under which our method or other traditional approaches (*e.g.*, those described in Section 3.2) are more suitable.

First, our method is not meant to replace traditional approaches that add control factors at the design stage of the experiments, for example, recording task completion time, and testing subjects with examples annotated with gold standard labels. Those methods are effective at identifying extremely careless subjects. But we argue that the reliability of a subject is often not a matter of yes or no, but can take a continum of intermediate levels. Moreover, consensus models such as Dawid-Skene methods require that each task is assigned to multiple annotators.

Second, our method can be integrated with other approaches so as to collect data most efficiently. Traditional heuristic approaches require the host to come up with a number of design questions or procedures effective for screening spammers before executing the experiments, which can be a big challenge especially for affective data. In contrast, the consensus models support post analyses of collected data and have no special requirement for the experimental designs. This suggests we may use a consensus model to carry out a pilot study which then informs us how to best design the data collection procedure.

Third, as a new method in the family of consensus models, our approach is unique in terms of its fundamental assumptions, and hence should be utilized in quite different scenarios than the other models. Methods based on modeling confusion matrix are more suitable for aggregating binary and categorical labels, while the agreement-based methods (ours included) are more suitable for continuous and multi-dimensional labels (or more complicated structures) that normally have no ground truth. The former are often evaluated quantitatively by how accurately they estimate the true labels [17], while the latter are evaluated directly by how effectively they identify unreliable annotators, a perspective barely touched in the existing literature.

Limitations and Future Work: Despite the fact that we did not assume A1 or A2 and approached the problem of assessing the quality of crowdsourced data form an unusual angle, there are interesting questions left about the statistical model we employed.

- Some choices of parameters in the model are quite heuristic. The usage of our model requires pre-set values for certain parameters, *e.g.*, γ, but we have not found theoretically pinned-down guidelines on how to choose those parameters. As a result, it is always subjective to some extent to declare a subject spammer. The ranking of reliability of subjects seems easier to accept. Where the cutoff should be will involve some manual checking on the result or will be determined by some other factors such as the desired cost of acquiring a certain amount of data.
- Although we have made great efforts to design various measures to evaluate our method, struggling to get around the issue of lacking an objective gold standard (its very existence has been questioned), these measures have limitations in one way or the other, as discussed in Section 3. We feel that due to the subjective nature of emotion responses to images, there is no simple and quick solution to this. The ultimate test of the method has to come from its usage in practice and a relatively long-term evaluation from the real-world.
- The effects of subgroup consistency, though varied from task to task, were random effects. We constructed the model this way to stretch its applicability because the number of responses collected per task in our empirical data was often small. Some related approaches (*e.g.* [16]) propose to estimate a difficulty/consistency parameter for each task, but often require a relatively large number of

annotators per task. Which kind of probabilistic assumptions is more accurate or works better calls for future exploration.

- Only one "major" reliable mode was assumed at one time, and hereafter only the regularities conditioned on this mode are estimated. In another word, all the reliable users are assumed to behave consistently. One may ask whether there exist subgroups of reliable users who behave consistently within a group but differ across groups for reasons such as different demographic backgrounds. In our current model, if such "minor" reliable mode exists in a population, these subjects may be absorbed into the spammer group. Our model implicitly assumes that diversity in demography or in other aspects does not cause influential differences in emotion responses. Because of this, our method in dealing with culturally sensitive data is not well justified.

Experimentally our method is only evaluated on one particular large data set [8]. Evaluations on other affective data sets (when publicly available) are of interest.

We have focused on the post analysis of collected data. As a future direction, it is of interest to examine the capacity of our approach to reduce time and cost in the practice of crowdsourcing using A/B test. We hereby briefly discuss an online heuristic strategy to dynamically allocate tasks to more reliable subjects. Recall that our model has two sets of parameters: parameter τ_i indicating the reliability of subjects and parameter $\alpha_i; \beta_i$ capturing the regularity. We can use the variance of distribution $\text{Beta}(\alpha_i, \beta_i)$ to determine how confident we are with the estimation of τ_i. For subject i, if the variance of $\text{Beta}(\alpha_i, \beta_i)$ is smaller than a threshold while τ_i is below a certain percentile, this subject is considered *confidently* unreliable and he/she may be excluded from the future subject pool.

5 Conclusions

In this work, we developed a probabilistic model, namely Gated Latent Beta Allocation, to analyze the off-line consensus for crowdsourced affective data. Compared to the usual crowdsourcing settings, where reliable workers are supposed to have consensus, the consensus analysis of affective data is more challenging because of the innate variation in emotion responses even out of true feelings. To overcome this difficulty, our model estimates the reliability of subjects by exploiting the agreement relationships between their ratings at a global scale. The experiments show that the relational data based on the valence of human responses are more effective than the other emotion dimensions for identifying spammer subjects. By evaluating and comparing the new method with some standard methods in multiple ways, we find that the results have demonstrated clear advantages and the system seems ready for use in practice.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1110970. We are grateful to the reviewers and the Associate Editor for their constructive comments.

References

[1] R. G. Barker, *Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior*. Stanford University Press, 1968.

[2] J. J. Gibson, *The Senses Considered as Perceptual Systems*. Houghton Mifflin, 1966.

[3] R. W. Picard and R. Picard, *Affective Computing*. MIT press Cambridge, 1997, vol. 252.

[4] S. Marsella and J. Gratch, "Computationally modeling human emotion," *Communications of the ACM*, vol. 57, no. 12, pp. 56–67, 2014.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *European Conference on Computer Vision*. Springer, 2006, pp. 288–301.

[6] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *Proceedings of the 20th ACM International Conference on Multimedia*. ACM, 2012, pp. 229–238.

[7] J. Howe, "The rise of crowdsourcing," *Wired Magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[8] X. Lu, "Visual characteristics for computational prediction of aesthetics and evoked emotions," Ph.D. dissertation, The Pennsylvania State University, 2015, chapter 5. [Online]. Available: https://etda.libraries.psu.edu/catalog/28857

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

[10] S. L. Hui and S. D. Walter, "Estimating the error rates of diagnostic tests," *Biometrics*, pp. 167–171, 1980.

[11] P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," in *Advances in Neural Information Processing Systems*, 1995, pp. 1085–1092.

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

[13] 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*, vol. 11, pp. 1297–1322, 2010.

[14] Q. Liu, J. Peng, and A. T. Ihler, "Variational inference for crowdsourcing," in *Advances in Neural Information Processing Systems*, 2012, pp. 692–700.

[15] V. C. Raykar and S. Yu, "Eliminating spammers and ranking annotators for crowdsourced labeling tasks," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 491–518, 2012.

[16] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Advances in Neural Information Processing Systems*, 2009, pp. 2035–2043.

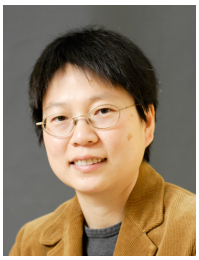
[17] A. Sheshadri and M. Lease, "Square: A benchmark for research on computing crowd consensus," in *First AAAI Conference on Human Computation and Crowdsourcing*, 2013, pp. 156–164.

[18] Y. J. Wang and G. Y. Wong, "Stochastic blockmodels for directed graphs," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 8–19, 1987.

[19] K. Nowicki and T. A. B. Snijders, "Estimation and predic-

tion for stochastic blockstructures," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 1077–1087, 2001.
[20] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, vol. 97, no. 460, pp. 1090–1098, 2002.
[21] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," in *Advances in Neural Information Processing Systems*, 2009, pp. 33–40.
[22] M. Kim and J. Leskovec, "Latent multi-group membership graph model," in *Proceedings of the 29th International Conference on Machine Learning*, 2012, pp. 1719–1726.
[23] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 2006, pp. 381–388.
[24] C. Kemp and J. B. Tenenbaum, "The discovery of structural form," *Proceedings of the National Academy of Sciences*, vol. 105, no. 31, pp. 10 687–10 692, 2008.
[25] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.
[26] J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, M. West *et al.*, "The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures," *Bayesian Statistics*, vol. 7, pp. 453–464, 2003.
[27] N. L. Johnson, S. Kotz, and N. Balakrishnan, "Chapter 21: beta distributions," *Continuous Univariate Distributions Vol. 2*, 1995.
[28] D. Gentner, "Bootstrapping the mind: Analogical processes and symbol systems," *Cognitive Science*, vol. 34, no. 5, pp. 752–775, 2010.
[29] D. Gentner and S. Christie, "Mutual bootstrapping between language and analogical processing," *Language and Cognition*, vol. 2, no. 2, pp. 261–283, 2010.

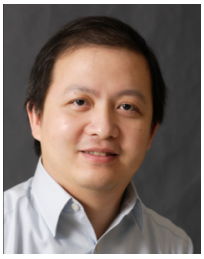
Jianbo Ye received his B. S. degree in Mathematics from the University of Science and Technology of China in 2011. He worked as a research postgraduate at The University of Hong Kong, from 2011 to 2012, and a research intern at Intel Labs, China in 2013. He is currently a PhD candidate and Research Assistant at the College of Information Sciences and Technology, The Pennsylvania State University. His research interests include statistical modeling and learning, numerical optimization and method, and affective image modeling.



Jia Li is a Professor of Statistics at The Pennsylvania State University. She received the MS degree in Electrical Engineering, the MS degree in Statistics, and the PhD degree in Electrical Engineering, all from Stanford University. She worked as a Program Director in the Division of Mathematical Sciences at the National Science Foundation from 2011 to 2013, a Visiting Scientist at Google Labs in Pittsburgh from 2007 to 2008, a researcher at the Xerox Palo Alto Research Center from 1999 to 2000, and a Research Associate in the Computer Science Department at Stanford University in 1999. Her research interests include statistical modeling and learning, data mining, computational biology, image processing, and image annotation and retrieval.

Michelle G. Newman is a Professor of Psychology and Psychiatry, and the Director of the Center for the Treatment of Anxiety and Depression at The Pennsylvania State University. She received her PhD from the State University of New York at Stony Brook in 1992 and completed a post-doctoral Fellowship at Stanford University School of Medicine in 1994. She has conducted psychotherapy outcome studies for generalized anxiety disorder, social phobia, and panic disorder, and has done basic emotion and experimental work related to these disorders. She is currently an editor for Behavior Therapy and is on the editorial boards of Psychotherapy Research, Cognitive Therapy and Research, and American Journal of Health Behavior. She is the recipient of the American Psychological Association (APA) Division 12 Turner Award for distinguished contribution to clinical research, and the APA Society of Psychotherapy (Division 29): Distinguished Publication of Psychotherapy Research Award. She is a Fellow of APA Divisions 12 and 29 and of the American Association for Behavioral and Cognitive Therapies.

Reginald B. Adams, Jr. is an Associate Professor of Psychology at The Pennsylvania State University. He received his PhD from Dartmouth College in 2002. He is interested in how we extract social and emotional meaning from nonverbal cues, particularly via the face. His work addresses how multiple social messages (*e.g.*, emotion, gender, race, age, etc.) combine across multiple modalities and interact to form unified representations that guide our impressions of and responses to others. Although his questions are social psychological in origin, his research draws upon visual cognition and affective neuroscience to address social perception at the functional and neuroanatomical levels. Before joining Penn State, he was awarded a National Research Service Award (NRSA) from the National Institute of Mental Health to train as a postdoctoral fellow at Harvard and Tufts Universities. His continuing research efforts have been funded through NSF, NIA and NIMH (NIH).



James Z. Wang is a Professor of Information Sciences and Technology at The Pennsylvania State University. He received the bachelor's degree in mathematics and computer science *summa cum laude* from the University of Minnesota, and the MS degree in mathematics, the MS degree in computer science and the PhD degree in medical information sciences, all from Stanford University. His research interests include computational aesthetics and emotions, automatic image tagging, image retrieval, and computerized analysis of paintings. He was a visiting professor at the Robotics Institute at Carnegie Mellon University (2007-2008), a lead special section guest editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence (2008), and a program manager at the Office of the Director of the National Science Foundation (2011-2012). He was a recipient of a National Science Foundation Career award (2004).