Collaborative Classification from Noisy Labels
Supplementary Material

A DERIVATION OF ALGORITHM 1

In this appendix, we provide a detailed derivation of the variational message-passing algorithm presented in Section 3 of the main text.

We begin with a brief recall of the coordinate-ascent variational inference (CAVI) algorithm. Given a set of random variables $\boldsymbol{z} = (z_1, \ldots, z_L)$ and a joint distribution $p(\boldsymbol{z})$, the algorithm finds a mean-field approximation $q(\boldsymbol{z}) \doteq \prod_\ell q(z_\ell)$ that minimizes the divergence $\mathrm{KL}(q \| p)$. Starting with an arbitrary $q(\boldsymbol{z})$, the algorithm iterates over $\ell \in [L]$ and refines the approximating distribution using the update rule

$$q(z_\ell) \propto \exp\left\{\mathbf{E}_{-\ell}[\log p(z_\ell, \boldsymbol{z}_{-\ell})]\right\}, \tag{1}$$

where $\boldsymbol{z}_{-\ell}$ denotes the set of all variables except the ℓth one, and $\mathbf{E}_{-\ell}[\cdot]$ denotes an expectation taken over $q(\boldsymbol{z}_{-\ell})$. Winn and Bishop (2005) and Blei et al. (2017) show, using different arguments, that the update (1) strictly decreases $\mathrm{KL}(q \| p)$. Thus, CAVI is guaranteed to converge to a local minimum of the KL-divergence.

We are now ready to consider the probabilistic model of Section 3. For convenience, we summarize the main symbols used throughout the paper in Table 1. In our case, $\boldsymbol{z} = (\boldsymbol{u}, \boldsymbol{v})$, and the distribution $p(\boldsymbol{u}, \boldsymbol{v})$ is a product of different types of factors,

$$p_0(u_i) = \mathrm{Dir}(u_i \mid \alpha_i), \qquad p_0(v_j) = \mathrm{Cat}(v_j \mid \beta_j), \qquad f(u_i, v_j) = \mathrm{Cat}(v_j \mid u_i).$$

Writing out the logarithm of the joint density, we obtain

$$\begin{aligned}
\log p(\boldsymbol{u}, \boldsymbol{v}) &= \sum_i \log p_0(u_i) + \sum_j \log p_0(v_j) + \sum_{(i,j) \in \mathcal{E}} \log f(u_i, v_j) + \mathrm{cst} \\
&= \sum_{j,k} \mathbf{1}\{v_j = k\} \log \beta_{jk} + \sum_{i,k} (\alpha_{ik} - 1) \log u_{ik} + \sum_{i,k} \left(\sum_{j \in \mathcal{N}_i} \mathbf{1}\{v_j = k\} \right) + \mathrm{cst} \\
&= \sum_{j,k} \mathbf{1}\{v_j = k\} \log \beta_{jk} + \sum_{i,k} \left(\alpha_{ik} - 1 + \sum_{j \in \mathcal{N}_i} \mathbf{1}\{v_j = k\} \right) \log u_{ik} + \mathrm{cst}
\end{aligned}$$

Writing out the CAVI update for the approximate marginal $q(v_j)$, we get

$$\begin{aligned}
q(v_j = k) \propto \exp\left\{\mathbf{E}_{-j}[\log p(\boldsymbol{u}, \boldsymbol{v})]\right\} &\propto \exp\left\{ \log \beta_{jk} + \sum_{i \in \mathcal{N}_j} \mathbf{E}_{-j}[\log u_{ik}] \right\} \\
&\propto \exp\left\{ \log \beta_{ik} + \sum_{i \in \mathcal{N}_j} \psi(\bar{\alpha}_{ik}) \right\},
\end{aligned} \tag{2}$$

Table 1: Table of symbols and notation.

Symbol	Domain	Description
M	$\mathbf{N}$	Number of users
N	$\mathbf{N}$	Number of items
K	$\mathbf{N}$	Number of classes
i	$[M]$	User
j, ℓ	$[N]$	Item
v_j	$[K]$	True class of item j
$\hat{v}_j$	$[K]$	Noisy label associated with item j
$\mathcal{E}$	$\mathcal{P}([M] \times [N])$	Set of edges, denotes user-item interactions
$\mathcal{N}_i$	$\mathcal{P}([N])$	Neighbors of user i, $\mathcal{N}_i = \{j : (i, j) \in \mathcal{E}\}$
$\mathcal{N}_j$	$\mathcal{P}([M])$	Neighbors of item j, $\mathcal{N}_j = \{i : (i, j) \in \mathcal{E}\}$
δ	$[0, 1]$	Corruption probability
α_i	$\mathbf{R}^K$	Parameter of the Dirichlet prior on user i's class proportions
β_j	Δ	Parameter of the categorical prior on item j's class
$\bar{\alpha}_i$	$\mathbf{R}^K$	Parameter of the variational posterior on user i's class proportions
$\bar{\beta}_j$	Δ	Parameter of the variational posterior on item j's class

where $\psi(x) \doteq \Gamma'(x)/\Gamma(x)$ is the digamma function. Similarly, the update for $q(\boldsymbol{u}_i)$ is given by

$$\begin{aligned} q(u_i) &\propto \exp\left\{\mathbf{E}_{-i}[\log p(\boldsymbol{u}, \boldsymbol{v})]\right\} \propto \exp\left\{\sum_k \left(\alpha_{ik} - 1 + \sum_{j \in \mathcal{N}_i} \mathbf{E}_{-i}[\mathbf{1}\{v_j = k\}]\right)\right\} \\ &\propto \mathrm{Dir}\Big(u_i \;\Big|\; \alpha_i + \sum_{j \in \mathcal{N}_i} \bar{\beta}_j\Big) \end{aligned} \tag{3}$$

Note that, because the updates of the item (respectively, user) marginals do not depend on $\bar{\boldsymbol{\beta}}$ ($\bar{\boldsymbol{\alpha}}$), we can update the marginal over all items (users) in batch. Algorithm 1, presented in the main text, is simply a concise reformulation of (2) and (3).

Connection to LDA. Our algorithm has some parallels with the variational inference method of Blei et al. (2003) for the latent Dirichlet allocation (LDA) topic model. In particular, our model enjoys conjugacy properties similar to those of the LDA model, leading to analogous closed-form CAVI updates. The two models are however distinctly different.

Alternatives to CAVI. Most prior work on collective classification (see, e.g., Taskar et al., 2001, 2002) uses a different approach to inference in structured models called loopy belief propagation (LBP) (Pearl, 1988; Yedidia et al., 2005). A brief derivation shows that our particular probabilistic model is not amenable to inference using LBP, because the resulting messages have no closed-form representation. As an alternative, we could use the expectation propagation (EP) framework (Minka, 2001), following the approach of Minka and Lafferty (2002). This would however result in an algorithm that is significantly more complex than Algorithm 1, and that would consequently be much harder to analyze theoretically. In addition, the convergence of LBP and EP is poorly understood, in contrast to CAVI which is guaranteed to converge.

B PROOFS

In this appendix, we provide proofs for the results presented in Section 4 of the main text. We begin by presenting some well-known bounds in Section B.1. Next, we introduce three auxiliary results that characterize some aspects of the Sparse Interaction Model in the large N limit in Section B.2. We prove Theorems 1, 2 and 3 in Sections B.3, B.4 and B.5, respectively.

B.1 Useful Bounds

We recall some standard concentration inequalities. For a random variable $z \sim \mathrm{Bin}(n,p)$, the Chernoff bound yields

$$\mathbf{P}[z \geq 2np] \leq \exp\left(-\frac{np}{3}\right), \tag{4}$$

$$\mathbf{P}[z \leq np/2] \leq \exp\left(-\frac{np}{8}\right). \tag{5}$$

If $p < c/n < 1$, we can make use of the following tighter bound, given in Arratia and Gordon (1989):

$$\mathbf{P}[z \geq c] \leq \exp\left[-n\mathrm{KL}(c/n\|p)\right] \leq \exp\left(-c\log\frac{c}{np}\right). \tag{6}$$

Next, let $x_1, \dots, x_n$ be identically distributed (but not necessarily independent) random variables with support in $[a,b]$ and mean $\mathbf{E}[x_\ell] = \mu$. Define the dependency graph of $\{x_1, \dots, x_n\}$ as $\mathcal{H} = ([n], \mathcal{A})$ such that $(i,j) \in \mathcal{A} \iff x_i$ and x_j are dependent. Let χ be the chromatic number (Diestel, 2016) of the dependency graph $\mathcal{H}$, and let $z = \sum_i x_i$. Then, an application of Janson (2004, Thm. 2.1) yields, for any $q \in [0,1]$,

$$\mathbf{P}[z \leq qn\mu] \leq \exp\left(-2\frac{n(1-q)^2\mu^2}{\chi(b-a)^2}\right), \tag{7}$$

Let $\psi(x) \doteq \Gamma'(x)/\Gamma(x)$ be the digamma function. Guo and Qi (2014, Theorem 1) show that

$$\log(x+1/2) - 1/x < \psi(x) < \log(x). \tag{8}$$

B.2 Auxiliary Results

We begin with Lemma 1 of the main text, which formalizes the notion that users tend to interact with items of the same class. For convenience, we restate the lemma.

Lemma 1. *Let $\mathcal{D} \sim \mathrm{SBM}$, and for any i, let $j, \ell \in \mathcal{N}_i$. Then, for any $k' \neq k$,*

$$p(v_\ell = k \mid v_j = k) = (1 + 1/\alpha) \cdot p(v_\ell = k' \mid v_j = k).$$

Proof. By construction, the probability that user i interacts with an item of class k is u_{ik}. Marginalizing over $u_i \sim \mathrm{Dir}(\alpha)$, we have.

$$p(v_j = k) = \int p(v_j = k \mid u_i)p(u_i)du_i = \int u_{ik}\mathrm{Dir}(u_i \mid \alpha)du_i = \frac{1}{K}.$$

Similarly, the probability that user i interacts with a first item of class k and a second item of class k' is $u_{ik}u_{ik'}$. Letting $k' \neq k$, we have

$$p(v_\ell = k, v_j = k) = \int u_{ik}^2\mathrm{Dir}(u_i \mid \alpha)du_i = \frac{1}{K} \cdot \frac{\alpha+1}{K\alpha+1},$$

$$p(v_\ell = k', v_j = k) = \int u_{ik}u_{ik'}\mathrm{Dir}(u_i \mid \alpha)du_i = \frac{1}{K} \cdot \frac{\alpha}{K\alpha+1}.$$

The claim follows by definition of conditional probability. □

The next lemma provides concentration inequalities for the number of items of class k, denoted by $|\mathcal{V}_k|$, and for the number of users interacting with each item, denoted by $|\mathcal{N}_j|$. Informally, it states that there are approximately N/K items per class, and each item is connected to approximately MS/N users in the interaction graph.

Lemma 2. *Let $S \geq 2$ and $M > 16N\log N$. With probability $1 - C/N$, we have*

$$\forall k \in [K] \qquad \frac{N}{2K} < |\mathcal{V}_k| < \frac{2N}{K} \tag{9}$$

$$\forall j \in [N] \qquad \frac{MS}{4N} < |\mathcal{N}_j| < \frac{4MS}{N} \tag{10}$$

Proof. Since $v_j \sim \mathrm{Cat}(1/K, \dots, 1/K)$ independently for all j, we have $|\mathcal{V}_k| \sim \mathrm{Bin}(N, 1/K)$. Applying inequalities (4) and (5) yields

$$\mathbf{P}\left[|\mathcal{V}_k| \leq \frac{N}{2K}\right] < \exp\left(-\frac{N}{8K}\right), \qquad \mathbf{P}\left[|\mathcal{V}_k| \geq \frac{2N}{K}\right] < \exp\left(-\frac{N}{3K}\right).$$

By using a union bound on the K classes, we obtain (9). Next, fix j and let $v_j = k$. The probability that user i interacts with item j is given by

$$r \doteq \mathbf{P}[j \in \mathcal{N}_i \mid v_j = k] = \mathbf{E}_{u_i}\left[\frac{n_{ik}}{|\mathcal{V}_k|}\right] = \frac{S}{K|\mathcal{V}_k|},$$

where we used $n_i \sim \mathrm{Mult}(S, u_i)$ and $u_i \sim \mathrm{Dir}(\alpha)$. By (9), we have

$$\frac{S}{2N} < r < \frac{2S}{N}$$

with probability at least $1 - 2\exp[-N/(8K)]$. Because each user interacts with items independently of other users, we have $|\mathcal{N}_j| \sim \mathrm{Bin}(M, r)$. As such, for any j,

$$\begin{aligned}\mathbf{P}\left[|\mathcal{N}_j| \leq \frac{MS}{4N}\right] < \mathbf{P}\left[|\mathcal{N}_j| \leq \frac{Mr}{2}\right] < \exp\left(-\frac{MS}{16N}\right) \leq \frac{1}{N^2},\\ \mathbf{P}\left[|\mathcal{N}_j| \geq \frac{4MS}{N}\right] < \mathbf{P}\left[|\mathcal{N}_j| \geq 2Mr\right] < \exp\left(-\frac{MS}{6N}\right) \leq \frac{1}{N^2},\end{aligned}$$

making use of inequalities (4) and (5). We obtain (10) by using a union bound on the N items. □

Finally, we consider $\mathcal{N}_j$, the set of users interacting with a given item j. Letting i, i' be two such users, we ask the question: How likely is it that both i and i' also "share" another item $\ell \neq j$? The next lemma states a result on the dependencies between users in $\mathcal{N}_j$, in terms of whether their interactions with other items (excluding j) overlap.

Lemma 3. *For any $C_1 > 0$, let $M \geq C_1 N \log N$. For any j, let $\mathcal{H}_j = (\mathcal{N}_j, \mathcal{A}_j)$ be a graph such that*

$$(i, i') \in \mathcal{A}_j \iff \mathcal{N}_i \cap \mathcal{N}_{i'} \setminus \{j\} \neq \varnothing.$$

Let χ_j be the chromatic number of $\mathcal{H}_j$. Then, with probability $1 - C_2/N$ we have, for all $j \in [N]$,

$$\chi_j \leq \frac{5M}{C_1 N \log N}$$

Proof. From standard results on greedy colorings (Diestel, 2016), we know that $\chi_j \leq \Delta(\mathcal{H}_j) + 1$, where $\Delta(\mathcal{H}_j)$ is the maximum degree of a vertex in $\mathcal{H}_j$. Without loss of generality, assume that all users in $\mathcal{N}_j$ interact with items of class $v_j = k$ only. For $i, i' \in \mathcal{N}_j$ we have

$$\begin{aligned} r \doteq \mathbf{P}[(i, i') \in \mathcal{A}_j] &= 1 - \mathbf{P}[\mathcal{N}_i \cap \mathcal{N}_{i'} \setminus \{j\} = \varnothing] = 1 - \prod_{\ell=1}^{S-1}\left(1 - \frac{S-1}{|\mathcal{V}_k| - \ell}\right)\\ &\leq 1 - \left[1 - \frac{S-1}{|\mathcal{V}_k| - (S-1)}\right]^{S-1} \leq \frac{(S-1)^2}{|\mathcal{V}_k| - (S-1)} \leq \frac{4KS^2}{N},\end{aligned}$$

for $N \geq S - 1$. The last inequality uses (9). Note that, in general, the edge probabilities are *not* jointly independent. However, as each user interacts with items independently of other users, the neighbors in $\mathcal{H}_j$ of any fixed $i \in \mathcal{N}_j$ are distributed independently and identically. Letting $\mathcal{Z}_{ji}$ be the set of neighbours of i in $\mathcal{H}_j$, we have $|\mathcal{Z}_{ji}| \sim \mathrm{Bin}(|\mathcal{N}_j| - 1, r)$. Without loss of generality, assume that $M \leq C_3 N \log N$. Then, inequality (6) yields

$$\begin{aligned}\mathbf{P}\left[|\mathcal{Z}_{ij}| \geq \frac{5M}{C_1 N \log N}\right] &\leq \exp\left(-5 \log \frac{N}{16 C_1 K S^3 \log N}\right)\\ &\leq \left(\frac{16 C_1 K S^3 \log N}{N}\right)^{-5} \leq \frac{C_4}{N^4},\end{aligned}$$

for N large enough. By union bound, it follows that

$$\chi_j \leq 1 + \Delta(\mathcal{H}_j) = 1 + \max_i |\mathcal{Z}_{ij}| \leq \frac{5M}{C_1 N \log N}$$

for all j with probability at least $1 - C_3 \log N / N^2$. □

The meaning of Lemma 3 is as follows. With high probability, any subset of $\chi_j + 1$ users in $\mathcal{N}_j$ contains at least two users whose neighborhoods are disjoint (except for j).

B.3 Proof of Theorem 1

We now focus on Algorithm 2. For a given j, we let $v_j = k$. From line 7 in the algorithm, we see that the output $\bar{v}_j$ is equal to k if and only if $z_k > z_{k'}$ for all $k' \neq k$. Expanding line 3 into line 6, it follows that

$$\bar{v}_j = k \iff z_k - z_{k'} = \sum_{i \in \mathcal{N}_j} \sum_{\ell \in \mathcal{N}_i \setminus \{j\}} (\mathbf{1}\{\hat{v}_\ell = k\} - \mathbf{1}\{\hat{v}_\ell = k'\}) > 0 \qquad \forall k' \neq k. \tag{11}$$

We begin by analyzing the inner sum in (11). The next lemma characterizes its expected value.

Lemma 4. *For any $j \in [N]$, let $v_j = k$. For any $i \in \mathcal{N}_j$ and any $k' \neq k$, we have*

$$\mathbf{E}\left[\sum_{\ell \in \mathcal{N}_i \setminus \{j\}} (\mathbf{1}\{\hat{v}_\ell = k\} - \mathbf{1}\{\hat{v}_\ell = k'\})\right] = \frac{S-1}{K\alpha + 1}\left(1 - \frac{K}{K-1}\delta\right)$$

where the expectation is taken over $\mathcal{N}_i$ and $\hat{v}_\ell$ for all $\ell \in \mathcal{N}_i \setminus \{j\}$

Proof. By Lemma 1 and by definition of $p(\hat{v}_j \mid v_j)$, we have, for all $\ell \in \mathcal{N}_i \setminus \{j\}$,

$$\mathbf{P}[\hat{v}_\ell = k \mid v_j = k] = (1-\delta) \cdot \frac{\alpha + 1}{K\alpha + 1} + \frac{\delta}{K-1} \cdot \frac{(K-1)\alpha}{K\alpha + 1},$$
$$\mathbf{P}[\hat{v}_\ell = k' \mid v_j = k] = (1-\delta) \cdot \frac{\alpha}{K\alpha + 1} + \frac{\delta}{K-1} \cdot \frac{(K-1)\alpha + 1}{K\alpha + 1}.$$

Using the linearity of expectation and elementary algebraic manipulations, we find that

$$\begin{aligned}\mathbf{E}\left[\sum_{\ell \in \mathcal{N}_i \setminus \{j\}} (\mathbf{1}\{\hat{v}_\ell = k\} - \mathbf{1}\{\hat{v}_\ell = k'\})\right] &= (S-1)(\mathbf{P}[\hat{v}_\ell = k \mid v_j = k] - \mathbf{P}[\hat{v}_\ell = k' \mid v_j = k]) \\ &= \frac{S-1}{K\alpha + 1}\left(1 - \frac{K}{K-1}\delta\right).\end{aligned}$$

□

We are now ready to prove Theorem 1, which we briefly restate here for convenience.

Theorem 1. *Let $\mathcal{D} \sim$ SBM, and let $\bar{\boldsymbol{v}}$ be the output of Algorithm 2 on $\mathcal{D}$. If $\delta < \frac{K-1}{K}$ and $M \geq \max\{16, 40\frac{(K\alpha+1)^2}{S}(1 - \frac{K}{K-1}\delta)^{-1}\} \cdot N \log N$, then for all $j \in [N]$, $\bar{v}_j = v_j$ w.h.p.*

Proof. Fix j, let $v_j = k$ and $k' \neq k$. Define auxiliary variables $\{y_i : i \in \mathcal{N}_j\}$ as follows:

$$y_i = \sum_{\ell \in \mathcal{N}_i \setminus \{j\}} (\mathbf{1}\{\hat{v}_\ell = k\} - \mathbf{1}\{\hat{v}_\ell = k'\})$$

By construction, we have $y_i \in [-(S-1), S-1]$. We use Lemma 2 to lower-bound $|\mathcal{N}_j|$, Lemma 3 to upper-bound χ_j, Lemma 4 to characterize $\mathbf{E}[y_i]$, inequality (7) with $q = 1$ and condition (11) to obtain

$$\mathbf{P}[z_k - z_{k'} \leq 0] = \mathbf{P}\left[\sum_{i \in \mathcal{N}_j} y_{ij} \leq 0\right] \leq \exp\left[-C_1 \frac{S \log N}{10(K\alpha + 1)^2}\left(1 - \frac{K}{K-1}\delta\right)\right].$$

Setting $C_1 = 20\frac{(K\alpha+1)^2}{S}(1 - \frac{K}{K-1}\delta)^{-1}$, we ensure that $\mathbf{P}[z_k - z_{k'} \leq 0] \leq 1/N^2$. By choosing $M \geq \max\{16, C_1\}N \log N$ we satisfy the conditions of Lemmas 2 and 3. By union bound,

$$\mathbf{P}[\forall j\ \bar{v}_j = v_j] \geq 1 - \sum_j \sum_{k' \neq v_j} \mathbf{P}[z_{v_j} - z_{k'}] \geq 1 - K/N,$$

and the claim follows. □

B.4 Proof of Theorem 2

We proceed in a similar way to the last section for Algorithm 1, For a given j, we let $v_j = k$ and $\bar{v}_j = \arg\max_\ell \bar{\beta}_{j\ell}$, where $\bar{\beta}_j$ is the output of the algorithm after one iteration. Expanding line 3 into line 6, it follows that

$$\begin{aligned} \bar{v}_j = k \iff \forall k' \neq k \quad & \log(\bar{\beta}_{jk}/\bar{\beta}_{jk'}) \\ &= \log(\beta_{jk}/\beta_{jk'}) + \sum_{i \in \mathcal{N}_j} \left[\psi\Big(\alpha_{ik} + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k}\Big) - \psi\Big(\alpha_{ik'} + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k'}\Big) \right] > 0. \end{aligned} \tag{12}$$

We begin by controlling the expectation of the term inside the sum in (12). The next lemma provides conditions under which the expectation can be bounded from below by a constant.

Lemma 5. *For any $j \in [N]$, let $k \doteq v_j$. For any $i \in \mathcal{N}_j$ and any $k' \neq k$, provided that $\delta \leq 1/80$, $S \geq 1 + (K-1)/\delta$ and $\alpha \leq \delta/[(1-\delta)K - 1]$, we have*

$$\mathbf{E}\left[\psi\Big(\alpha_{ik} + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k}\Big) - \psi\Big(\alpha_{ik'} + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k'}\Big) \right] \geq 1/6,$$

where the expectation is taken over $\mathcal{N}_i$ and β_ℓ for all $\ell \in \mathcal{N}_i \setminus \{j\}$.

Proof. By using Lemma 1 and by definition of $p(\hat{v}_j \mid v_j)$, we have, for all $\ell \in \mathcal{N}_i \setminus \{j\}$,

$$\beta_{\ell k} = \begin{cases} 1 - \delta & \text{w.p. } (1-\delta)\dfrac{\alpha+1}{K\alpha+1} + \delta\dfrac{\alpha}{K\alpha+1}, \\ \dfrac{\delta}{K-1} & \text{w.p. } \delta\dfrac{\alpha+1}{K\alpha+1} + (1 - \dfrac{\delta}{K-1})\dfrac{(K-1)\alpha}{K\alpha+1}, \end{cases}$$

$$\beta_{\ell k'} = \begin{cases} 1 - \delta & \text{w.p. } (1-\delta)\dfrac{\alpha}{K\alpha+1} + \dfrac{\delta}{K-1}\dfrac{(K-1)\alpha+1}{K\alpha+1}, \\ \dfrac{\delta}{K-1} & \text{w.p. } \delta\dfrac{\alpha}{K\alpha+1} + (1 - \dfrac{\delta}{K-1})\dfrac{(K-1)\alpha+1}{K\alpha+1}. \end{cases}$$

We begin by upper-bounding the first term inside of the expectation. By using (8) and Jensen's inequality, we get

$$\begin{aligned} \mathbf{E}\left[\psi\Big(\alpha + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k'}\Big)\right] &\leq \mathbf{E}\left[\log\Big(\alpha + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k'}\Big)\right] \leq \log\Big(\alpha + \beta_{jk'} + \sum_{\ell \in \mathcal{N}_i \setminus \{j\}} \mathbf{E}[\beta_{\ell k'}]\Big) \\ &\leq \log\left[\alpha + \beta_{jk'} + (S-1)\left(\frac{\alpha}{K\alpha+1} + 2\frac{\delta}{K-1}\right)\right], \end{aligned}$$

Next, we lower-bound the second term.

$$\begin{aligned}
\mathbf{E}\left[\psi\left(\alpha + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k}\right)\right] &\geq \mathbf{E}\left[\log\left(\alpha + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k}\right)\right] - \frac{K-1}{S\delta} \\
&\geq \log\left(\alpha + \beta_{jk} + \sum_{\ell \in \mathcal{N}_i \setminus \{j\}} \exp \mathbf{E}[\log \beta_{\ell k}]\right) - \frac{K-1}{S\delta} \\
&\geq \log\left\{\alpha + \beta_{jk} + (S-1)\left[(1-\delta) \cdot \left(\frac{\delta}{K-1}\right)^{\left(\delta + \frac{(K-1)\alpha}{K\alpha+1}\right)}\right]\right\} - \frac{K-1}{S\delta} \\
&\geq \log\left\{\alpha + \beta_{jk} + (S-1)\left[\frac{(1-\delta)\delta}{\left(2\delta + \frac{(K-1)\alpha}{K\alpha+1}\right)\left(1 + (K-1)\delta + \frac{(K-1)^2\alpha}{K\alpha+1}\right)}\right]\right\} - \frac{K-1}{S\delta},
\end{aligned}$$

where we used (8) on the first line, and a variational lower bound (Paisley, 2010) on the second line. If $\alpha \leq \delta/[(1-\delta)K-1]$, the bounds simplify as follows.

$$\begin{aligned}
\mathbf{E}\left[\psi\left(\alpha + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k'}\right)\right] &\leq \log\left[\alpha + \beta_{jk'} + (S-1)\frac{3\delta}{K-1}\right], \\
\mathbf{E}\left[\psi\left(\alpha + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k}\right)\right] &\geq \log\left[\alpha + \beta_{jk} + (S-1)\frac{1-\delta}{3+6(K-1)\delta}\right] - \frac{K-1}{S\delta}.
\end{aligned}$$

Letting $\delta < 1/80$ and $S \geq (K-1)/\delta + 1$, we have $\alpha \leq 1$ and

$$\begin{aligned}
&\mathbf{E}\left[\psi\left(\alpha + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k}\right) - \psi\left(\alpha + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k'}\right)\right] \\
&\quad \geq \log\left[\exp\left(-\frac{K-1}{S\delta}\right)\frac{\alpha + \beta_{jk} + (S-1)\frac{1-\delta}{3+6(K-1)\delta}}{\alpha + \beta_{jk'} + (S-1)\frac{3\delta}{K-1}}\right] \\
&\quad \geq \log\left[e^{-1}\frac{(S-1)\frac{1-\delta}{3+6(K-1)\delta}}{1 + (1-\delta) + (S-1)\frac{3\delta}{K-1}}\right] \\
&\quad \geq \log\left[e^{-1}\frac{(K-1)(1-\delta)}{5\delta[3+6(K-1)\delta]}\right] \geq \log(4/3) > 1/6.
\end{aligned}$$

This concludes the proof. □

We are now ready to prove Theorem 2, which we briefly restate here for convenience.

Theorem 2. *Let $\mathcal{D} \sim \mathrm{SBM}$, and let $\bar{\boldsymbol{\beta}}$ be the output of Algorithm 1 on $\mathcal{D}$ after one iteration. There exist C_1, C_2, C_3, C_4 independent of N such that if $M \geq C_1 N \log N$, $\alpha < C_2$, $\delta < C_3$, $S > C_4$, then for all $j \in [N]$, $\arg\max_k\{\bar{\beta}_{jk}\} = v_j$ w.h.p.*

Proof. Fix j, let $v_j = k$ and $k' \neq k$. Define auxiliary variables $\{y_i : i \in \mathcal{N}_j\}$ as follows:

$$y_i = \psi\left(\alpha + \sum_{\ell \in \mathcal{N}_i} \beta_{\ell k}\right) - \psi\left(\alpha + \sum_{j \in \mathcal{N}_i} \beta_{\ell k'}\right)$$

By construction, we have $y_i \in [-C_5, C_5]$, where

$$C_5 \doteq \log\left(\frac{\alpha + S(1-\delta)}{1/2 + \alpha + S\frac{\delta}{K-1}}\right) + \frac{1}{\alpha + S\frac{\delta}{K-1}} \geq \psi[\alpha + S(1-\delta)] - \psi\left[\alpha + S\frac{\delta}{K-1}\right],$$

making use of (8) twice. We use Lemma 2 to lower-bound $|\mathcal{N}_j|$, Lemma 3 to upper-bound χ_j, Lemma 5 to bound $\mathbf{E}[y_i]$, inequality (7) with $q = 1/2$ and condition (12) to obtain

$$\mathbf{P}\left[\sum_{i \in \mathcal{N}_j} y_i \le \log(\beta_{jk}/\beta_{jk'})\right] \le \mathbf{P}\left[\sum_{i \in \mathcal{N}_j} y_i \le \frac{C_1 S \log N}{48}\right] \le \exp\left(-\frac{C_1 \log N}{1440 C_5^2}\right) \le 1/N^2,$$

for N large enough, if $C_1 > 2880 C_5^2$. By union bound,

$$\mathbf{P}[\forall j \ \bar{v}_j = v_j] \ge 1 - \sum_j \sum_{k' \ne v_j} \mathbf{P}\left[\sum_{i \in \mathcal{N}_j} y_i \le \log(\beta_{jv_j}/\beta_{jk'})\right] \ge 1 - K/N,$$

and the claim follows. □

B.5 Proof of Theorem 3

Before proceeding with the proof, we introduce a result from combinatorics. We define the D-uniform random hypergraph $H_D(N, M)$ as a hypergraph[1] with N vertices and M edges sampled uniformly at random among the $\binom{N}{M}$ possible subsets of $[N]$ of size M. The following lemma, adapted from Poole (2015), gives a lower bound on the number of edges necessary to connect the hypergraph.

Lemma 6 (Poole, 2015, Lemma 2.1)**.** *Let $H_D(N, M)$ be a D-uniform random hypergraph. For any $\epsilon > 0$, if $M < \frac{1-\varepsilon}{D} N \log N$ then $H_D(N, M)$ has at least $\lceil \log \log N \rceil$ vertices of degree 0 with high-probability.*

We are now ready to prove Theorem 3, which we briefly restate here for convenience.

Theorem 3. *Let $\mathcal{D} \sim$ SBM. If $M \le \frac{1}{5KS} N \log N$, then w.h.p. there exists a set of items $\mathcal{B} \subseteq [N]$ such that $|\mathcal{B}| \ge \log \log N$ and $\mathcal{N}_j = \varnothing$ for all $j \in \mathcal{B}$.*

Proof. We start by viewing the bipartite interaction graph as a hypergraph $\mathcal{H}([N], \mathcal{A})$, where $\mathcal{A} = \{\mathcal{N}_i : i \in [M]\}$. In other words, $\mathcal{H}$ is a hypergraph on the N items where every user corresponds to an edge consisting of all the items that user interacted with.

Due to the probabilistic nature of user-item interactions in SBM, $\mathcal{H}$ is a random hypergraph that is S-uniform (all users interact with exactly S items), but it is not a S-uniform random hypergraph in the sense of Lemma 6. Indeed, users have a bias towards interacting with items of the same class (see e.g., Lemma 1).

We thus fix a class $k \in [K]$, and assume that for all $k' \ne k$, if $v_j = k'$ then $\mathcal{N}_j \ne \varnothing$, i.e., all "bad" items have class k. Let $\mathcal{V}_k = \{j \in [N] : v_j = k\}$. Conditioned on class k, users choose n_{ik} items uniformly at random from $\mathcal{V}_k$. In the worst case, $n_{ik} = S$ for all $i \in [M]$, and the connectivity of items in $\mathcal{V}_k$ is determined by that of the S-uniform random hypergraph $H_S(|\mathcal{V}_k|, M)$.

We can then lower-bound $|\mathcal{V}_k|$ using Lemma 2 and use Lemma 6 with $\varepsilon = 1/5$ to conclude the proof. □

C GENERATIVE ASSUMPTIONS

In this section, we briefly discuss the assumptions of the generative model introduced in Section 4 and sketch how relaxing them would impact the performance of CAVI and wvRN.

Constant number of edges per user. We assume that every user interacts with exactly S items. This simplifies the proofs, but it is not strictly necessary. We can expect similar results to hold if the number of interactions is S on average, but varies from user to user.

Uniform choices. We assume that users choose items within a class uniformly at random. This is clearly unrealistic in many practical applications, where we expect some items to be more popular than others. Our results rely on all items being "well-connected" in the interaction graph; If there is a popularity bias, some items might have zero or few users, and it might thus become difficult to correct their labels. In that case, we can likely develop results that depend on the popularity rank or on the size of an item's neighborhood in the interaction graph.

[1] A hypergraph is a generalization of a graph where edges are subsets of vertices of arbitrary cardinality.

Table 2: Description of classes for each dataset.

Name	Classes
Stack Overflow	`c#`, `c++`, `ios`, `java`, `javascript`, `php`, `python`, `r`, `ruby-on-rails`, `sql`
Yelp	AZ, NV, ON, OH, NC, PA, QC, AB, WI, IL
Amazon	"Books", "CDs & Vinyl", "Clothing, Shoes & Jewelry", "Electronics", "Sports & Outdoors"

Balanced item classes. We assume that items belong to one the K classes uniformly at random—in other words, that the classes are balanced. This assumption is not restrictive, and we make it for simplicity only. Our results can be extended in a straightforward way to settings where item classes are unbalanced.

Symmetric aggregate affinities. We assume that the users' class-proportion vector is sampled from a symmetric Dirichlet with concentration parameter α. That is, we assume that, when averaged over all users, class proportions are balanced (but that, individually, each user still prefers some classes more than others). In theory, we could expect that significant deviations from this assumption might be problematic for our algorithms. In practice, however, Algorithm 1 appears to work well even on highly asymmetric problems (such as the podcast dataset studied in Section 5). Algorithm 2 is more sensitive to asymmetric proportions, but it could be robustified along the lines of the *class-distribution relational neighbor* algorithm of Macskassy and Provost (2007).

D EXPERIMENTAL EVALUATION

We briefly describe each of the three public datasets we examine in Section 5.2 of the main text. Upon publication, we will release code that enables reproducing exactly the results presented in the paper.

Stack Overflow This dataset contains questions and answers from Stack Overflow, a Q&A platform for programmers. On this platform, users can ask or answer questions (items) that are annotated by tags (classes). We consider the 10 most popular programming languages discussed on the platform, and retain all questions that are tagged with exactly one of these languages.

Yelp This dataset contains reviews from Yelp, a crowd-sourced business review service. On this service, users write reviews about businesses (items), and each business is annotated with a location (class). We consider the 10 U.S. states and Canadian provinces that are the most prevalent in the dataset, and discard businesses located elsewhere.

Amazon This dataset contains product reviews from Amazon, an e-commerce platform (McAuley and Leskovec, 2013). Users write reviews about products (items) belonging to one of several categories (classes). We consider the five largest categories, and retain items annotated with exacly one of these categories.

Table 2 lists the classes we seek to distinguish in each dataset.

References

R. Arratia and L. Gordon. Tutorial on large deviations for the binomial distribution. *Bulletin of Mathematical Biology*, 51:125–131, 1989.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3 (Jan):993–1022, 2003.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

R. Diestel. *Graph Theory.* Springer, fifth edition, 2016.

B.-N. Guo and F. Qi. Sharp inequalities for the psi function and harmonic numbers. *Analysis*, 34(2):201–208, 2014.

S. Janson. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24 (3):234–248, 2004.

S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, 8(May):935–983, 2007.

J. McAuley and J. Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of RecSys'13*, Hong Kong, China, Oct. 2013.

T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

T. P. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of UAI 2002*, Edmonton, AL, Canada, Aug. 2002.

J. Paisley. Two useful bounds for variational inference. Technical report, Department of Computer Science, Princeton University, Aug. 2010. URL `http://www.columbia.edu/~jwp2128/Teaching/E6892/papers/twobounds.pdf`.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

D. Poole. On the strength of connectedness of a random hypergraph. *The Electronic Journal of Combinatorics*, 22, 2015.

B. Taskar, E. Segal, and D. Koller. Probabilistic classification and clustering in relational data. In *Proceedings of IJCAI 2001*, Seattle, WA, USA, Aug. 2001.

B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proceedings of UAI 2002*, Edmonton, AL, Canada, Aug. 2002.

J. Winn and C. M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6(Apr):661–694, 2005.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, 2005.