Appendix

A. Proof of Lemma 3

Recall that θ is a finite index. For each (i,j) such that $N_{ij} > 0$, we have

$$\begin{aligned}\frac{1}{N_{ij}} \sum_{t:(i,t),(j,t)\in A} (Y_{i,t}Y_{j,t} - x_i x_j)^2 &= \frac{1}{N_{ij}} \sum_{t:(i,t),(j,t)\in A} (1 - 2Y_{i,t}Y_{j,t}x_i x_j + x_i^2 x_j^2) \\ &= 1 - 2\tilde{C}_{ij}x_i x_j + x_i^2 x_j^2 \\ &= \tilde{C}_{ij}^2 - 2\tilde{C}_{ij}x_i x_j + x_i^2 x_j^2 + 1 - \tilde{C}_{ij}^2 \\ &= (\tilde{C}_{ij} - x_i x_j)^2 + 1 - \tilde{C}_{ij}^2.\end{aligned}$$

Therefore,

$$\frac{1}{2} \sum_{(i,t),(j,t)\in A} (Y_{i,t}Y_{j,t} - x_i x_j)^2 = \frac{1}{2} \sum_{i,j\in[W]} N_{ij}(\tilde{C}_{ij} - x_i x_j)^2 + \sum_{i,j\in[W]} N_{ij}(1 - \tilde{C}_{ij}^2)$$

Since $\sum_{i,j\in[W]} N_{ij}(1 - \tilde{C}_{ij}^2)$ is a constant, Eq.(1) is equivalent to the optimization problem $\operatorname{argmin}_{x\in[-1,+1]^W} L(x)$.

B. Proof of Theorem 1

The proof directly follows from Lemma 4 and Lemma 5. We will next prove these Lemmas.

Proof of Lemma 4: Take any two workers i, j that are connected in $G = ([W], E)$. Let $t \in \mathcal{N}$ be such that $(i,t),(j,t) \in A$. By assumption, $Y_{i,t}Y_{j,t} = g_t^2 Z_{i,t}Z_{j,t} = Z_{i,t}Z_{j,t}$. Now, by the law of large numbers,

$$\begin{aligned}C_{ij} \doteq &\lim_{T\to\infty} \frac{1}{T} \sum_{(i,t),(j,t)\in A, t\le T} Z_{i,t}Z_{j,t} = \mathbb{E}\left[Z_i Z_j\right] \\ =&\mathbb{E}\left[Z_i\right]\mathbb{E}\left[Z_j\right] = s_i s_j,\end{aligned}$$

where $(Z_i)_i \sim \Pi_{i=1}^W \mathrm{Rad}(s_i)$. Note that $C_{ij} = C_{ji}$. We define $C_{ij} = 0$ when $(i,j) \notin E$.

Now, WLOG assume that workers $1, 2, \ldots, 2k+1$ form a cycle in G: $(1,2), \ldots, (2k, 2k+1), (2k+1, 1) \in E$. Then,

$$\begin{aligned}s_1 &= C_{1,2k+1}s_{2k+1}^{-1} \\ &= C_{1,2k+1}C_{2k+1,2k}^{-1}s_{2k} \\ &= C_{1,2k+1}C_{2k+1,2k}^{-1}C_{2k,2k-1}s_{2k-1}^{-1} \\ &\vdots \\ &= C_{1,2k+1}C_{2k+1,2k}^{-1}C_{2k,2k-1}\ldots C_{2,1}s_1^{-1},\end{aligned}$$

or

$$|s_1| = \sqrt{C_{1,2k+1}C_{2k+1,2k}^{-1}C_{2k,2k-1}\ldots C_{2,1}},$$

assuming that $C_{2,3}, C_{4,5}, \ldots, C_{2k,2k+1} \neq 0$. Since G is connected, for any worker i there exists a path from worker 1 to worker i. If this path was given by the vertices $1, 2, \ldots, \ell$ then

$$\begin{aligned}|s_\ell| =&|C_{\ell,\ell-1}|\,|s_{\ell-1}^{-1}| = |C_{\ell,\ell-1}|\,|C_{\ell-1,\ell-2}^{-1}|\,|s_{\ell-2}| \\ =&\cdots = |C_{\ell,\ell-1}|\,|C_{\ell-1,\ell-2}^{-1}|\cdots|C_{2,1}^{(-1)^\ell}|\,|s_1|^{(-1)^{\ell+1}}.\end{aligned}$$

It remains to show that $\mathcal{P}(s)$ can be recovered. Let $i, j \in [W]$ be different workers. Then, if $\pi \subset E$ is a path in G from i to j, we have $\Pi_{(u,v)\in E}\operatorname{sgn}(C_{u,v}) = \Pi_{(u,v)\in E}\operatorname{sgn}(s_u)\operatorname{sgn}(s_v) = \operatorname{sgn}(s_i)\operatorname{sgn}(s_j)$ regardless of how π is chosen. Now, if i

and j are such that for some path π connecting them $\Pi_{(u,v)\in E}\operatorname{sgn}(C_{u,v}) = +1$, we assign i, j to the same group. Since G is connected, this creates at most two groups and the resulting "partition" must match $\mathcal{P}(s)$.

Proof of Lemma 5: Take s, α which are used in the definition of richness of Θ. We construct two other skill vectors s' and s'' as follows: We set $s'_1 = \alpha s_1$ and $s''_1 = s_1/\alpha$. Now, if worker i is at an even distance from worker 1 on some path in G then $s'_i = \alpha s_i$ and $s''_i = s_i/\alpha$, otherwise we set $s'_i = s_i/\alpha$ and $s''_i = \alpha s_i$. Note that all workers can be accessed from worker 1 because G is connected. Note that if there are multiple paths from worker 1 to some other worker then all of these have the same parity, or the graph had an odd cycle. Now, both s and s' give rise to the same products, $s_i s_j$, along any edge $(i, j) \in E$. Since both are in Θ by assumption, the result is proven.

Reverse Direction for Theorem 1: We prove this by contraposition. First, assume that (i) does not hold. We want to prove that learnability fails. If (i) does not hold, we can take $s, s' \in [-1, 1]^W$ different skill vectors such that $|s| = |s'|$ and $\mathcal{P}(s) = \mathcal{P}(s')$ and $s, s' \in S(\Theta)$. It follows that $s = -s'$. Take any $g \in \{\pm 1\}^W$. Note that the instances (s, A, g) and $(-s, A, -g)$ lead to the same joint distribution over the observed labels. Hence, no inference schema can tell these instances apart, thus any inference schema will suffer linear regret on one of these instances. Now, if (ii) does not hold, Lemma 5 gives two skill vectors s, s' which are different and $s \neq \pm s'$, which again give the same likelihood to any data. This again leads to that any inference schema will suffer a linear regret on one of these instances.

C. Proof of Theorem 2

Recall that we are given the interaction matrix N which is nonnegative, irreducible, aperiodic, with integer entries, symmetric, and with zero diagonal; and also a vector $s \in \mathbb{R}^W$. We need to argue that there does not exist a vector $x \neq \{\pm s\}$ such that, for each $i = 1, \ldots, W$, we have

$$\sum_{j=1}^{W} N_{ij}(x_i x_j - s_i s_j)x_j = 0. \tag{5}$$

We begin by adopting the following notation. For a vector x, D_x will refer to the diagonal matrix with x on the diagonal. For a matrix A, $\operatorname{diag}[A]$ will refer to the *diagonal of A stacked as a vector* (note that this is an **unusual** notation). Also, let us refer to the set of matrices which are nonnegative, irreducible, aperiodic, symmetric and with have zero diagonal as *admissible*.

Assume that x satisfies (5). First, assume that none of the components of x are zero. The case x has zero components will be dealt with later. And we also make the simplifying assumptions that $s > 0$ and that we are looking for $x > 0$. In the case of s has negative components, we can always recover the absolute value $|s|$ of s by the absolute value $|x|$of x. If we assume $s_i > 0$, we can figure out all other signs of s and get one solution x by the assumption that the graph is non-bipartite. Similarly, there is another solution $-x$ if we assume $s_i < 0$. Since we only consider skill vectors that have a positive sum, the optimal solution must be one of $\{\pm x\}$ that has a positive sum.

Then, we will argue that given $s > 0$ we cannot find $x > 0, x \neq s$ and admissible W such that (5) holds. We can multiply the ith equation of (5) by x_i. Our first observation is that we may rewrite Eq. (5) as

$$\operatorname{diag}\left[D_x N D_x (xx^T - ss^T)\right] = 0. \tag{6}$$

It suffices to argue that we cannot find $x > 0, x \neq s$ and admissible F such that

$$\operatorname{diag}\left[F(xx^T - ss^T)\right] = 0.$$

By $x > 0$ we mean that $x_i > 0$ for all i, i.e., no component of x_i is zero. We were able to drop the D_x from the equation because N is admissible if and only if $D_x N D_x$ is.

We proceed as follows. Since

$$x_i x_j - s_i s_j = s_i \left(\frac{x_i}{s_i}\frac{x_j}{s_j} - 1\right) s_j\,,$$

defining $u_i = x_i/s_i$ we have that $u > 0$ and that

$$xx^T - ss^T = D_s(uu^T - 11^T)D_s\,.$$

We must therefore argue that it is impossible to find $u > 0, u \neq 1$ and admissible F such that

$$\operatorname{diag}\left[D_s F D_s (uu^T - 11^T) D_s\right] = 0.$$

Since $s > 0$ it will suffice to argue that we cannot find $u > 0, u \neq 1$ and admissible Z such that

$$\operatorname{diag}\left[Z(uu^T - 11^T)\right] = 0. \tag{7}$$

Without loss of generality, we can assume that $u_1 \leq u_2 \leq \cdots \leq u_W$; we can always relabel indices to make this hold.

Now there are three possibilities:

1. $u_1 u_W > 1$.
2. $u_1 u_W = 1$.
3. $u_1 u_W < 1$.

We argue that in each case we cannot find a suitable u that satisfies Eq. (7). Indeed, let us consider the first possibility. In that case the last column of $uu^T - 11^T$, with entries $u_i u_W - 1$, is strictly positive, and therefore, considering that $[Z(uu^T - 11^T)]_{WW} = 0$, we obtain that the last row of Z must be zero – contradicting irreducibility. Similarly, in case 3, the first column of $uu^T - 11^T$, with entries $u_1 u_i - 1$, is negative, and, considering that $[Z(uu^T - 11^T)]_{11} = 0$, we see that the first row of Z must be zero, which can not hold true.

It remains to consider case 2. Consider any $u > 0, u \neq 1$. We may assume that $u_1 < u_W$ (ruling out the possibility that a u proportional to the all-ones vector satisfies Eq. (7) is trivial).We break up $\{1, \ldots, W\}$ into three blocks. The first block is all the indices j such that $u_j = u_1$. The third block is all the indices j such that that $u_j = u_W$. All the other indices go into block 2. Note that block 2 may be empty, for example if every entry of u is equal to u_1 or u_W.

The advantage of partitioning this way is that the matrix $uu^T - 11^T$ has the following sign structure:

$$uu^T - 11^T = \left(\begin{array}{c|c|c} - & - & 0 \\ \hline - & * & + \\ \hline 0 & + & + \end{array}\right)$$

where $-$ represents a strictly negative submatrix, $+$ represents a strictly positive submatrix, while $*$ represents a submatrix that can have elements of any sign. The strict negativity comes from the fact that $u_1 < u_W$.

Partitioning Z in a compatible manner, we have that

$$\operatorname{diag}\left[\left(\begin{array}{c|c|c} Z_{11} & Z_{12} & Z_{13} \\ \hline Z_{21} & Z_{22} & Z_{23} \\ \hline Z_{31} & Z_{32} & Z_{33} \end{array}\right)\left(\begin{array}{c|c|c} - & - & 0 \\ \hline - & * & + \\ \hline 0 & + & + \end{array}\right)\right] = 0.$$

Considering the $(1,1)$ diagonal block of the above product, noting that $Z \geq 0$, we obtain $Z_{11} = Z_{12} = 0$; and considering the $(3,3)$ diagonal block of the above product we obtain $Z_{32} = Z_{33} = 0$. By symmetry, also $Z_{21} = 0$ and $Z_{33} = 0$.

From here we can easily derive a contradiction. Indeed, if the middle block is nonempty, the matrix is reducible; and if the second block is empty, it is periodic with an even period, finishing the proof for the case when none of the components of x can have zero entries.

Finally, we can easily rule out stationary points as unstable when x has zero entries. Note that the Hessian $P(x)$ in matrix form is $D_s P(x/s) D_s$, where x/s is componentwise division, P is given in the main body of the text. Note that $P(x)$ is positive (semi)definite if and only if $P(x/s)$ is positive (semi)definite. Further, by the form of P, $P(x/s)$ is not positive semidefinite if any of the components of x are zero. This completes the proof, noting that gradient algorithms when initialized randomly will not converge to non-local minima.

D. Proof of Theorem 3

Fix $s \in [-1,1]^W$. Let

$$f(x) = \sum_{i<j} N_{ij}(x_i x_j - s_i s_j)^2. \tag{8}$$

be the "noisey-free" objective function underlying s and let $P_s(x) = \nabla^2 f(x)$ be the Hessian of f at x. By straightforward calculation,

$$\frac{\partial f}{\partial x_i}(x) = \sum_{k=1}^{W} 2N_{ik}(x_i x_k - s_i s_k)x_k,$$

and

$$(P_s(x))_{ii} = \frac{\partial^2 f}{\partial x_i^2}(x) = \sum_{k=1}^{W} 2N_{ik}x_k^2\,,$$
$$(P_s(x))_{ij} = \frac{\partial^2 f}{\partial x_i x_j}(x) = 4N_{ij}x_i x_j - 2s_i s_j N_{ij}, \quad i \neq j.$$

Therefore,

$$\nabla^2 f(x) = D_s P_1(x/s) D_s\,.$$

Note that in the main body of the paper, we denoted $P_1(u)$ by $P(u)$, dropping the 1 subindex. We will continue to use this notation here. Recall that $s_{\min} = \min_i |s_i|$. From the above, by the continuity of $P = P_1$, it follows immediately that f is strongly convex in a small enough neighborhood of s:

Lemma 6. *There exists a positive number δ such that f is strongly convex when restricted to the set*

$$B_\delta = \{x \in [-1,1]^W \,:\, \|x/s - 1\|_\infty \leq \delta\}\,.$$

In particular, for any $x \in B_\delta$,

$$\lambda_{\min}\left(\nabla^2 f(x)\right) \geq s_{\min}^2\,\mu\,,$$

with $\mu = \lambda_{\min}(P(1))/2$.

Note that $\mu > 0$ by Proposition 2.1 of (Desai & Rao, 1994).

Proof. Clearly, $\lambda_{\min}(\nabla^2 f(x)) \geq s_{\min}^2 \lambda_{\min}(P(x/s))$. In particular, $\lambda_{\min}(\nabla^2 f(s)) \geq 2s_{\min}^2\mu$. The statement then follows from that P_s, and hence also $\lambda_{\min}(P_s(\cdot))$ is a continuous function of its argument. □

As an immediate corollary we get:

Corollary 2. *For any $x \in B_\delta$ with δ as in the previous lemma,*

$$\|x - s\| \leq \|\nabla f(x)\|_2 / (s_{\min}^2\mu)\,.$$

Proof. For $x = s$ there is nothing to be shown. Hence, assume $x \in B_\delta$ and $x \neq s$. By the $(s_{\min}^2\mu)$-strong convexity of f on B_δ (cf. Theorem 2.1.9 of Nesterov (2004)), since both x and s are in B_δ,

$$(\nabla f(x) - \nabla f(s))^T(x - s) \geq s_{\min}^2\mu\,\|x - s\|^2.$$

Using that $\nabla f(s) = 0$ and applying Cauchy-Schwarz, we obtain

$$\|\nabla f(x)\|\|x - s\| \geq s_{\min}^2\mu\,\|x - s\|^2\,.$$

The result follows by dividing both sides by $\|x - s\|$. □

Type of workers	α	β	Bayes error	Prediction error (const. noise)
Adversary vs. hammer	0.5	0.5	0.0036 ± 0.0014	0.5990 ± 0.4860
Asym. with more positive skills	5	1	0.0038 ± 0.0014	0.0041 ± 0.0013
Asym. with more negative skills	2	5	0.0314 ± 0.0062	0.9667 ± 0.0067
Hammer	2	2	0.0615 ± 0.0083	0.4162 ± 0.4273
Spammer	1	3	0.0129 ± 0.0034	0.9864 ± 0.0041

Table 3. Average prediction errors with different skills distributions.

Now, consider x s.t. $\nabla L(x) = 0$ and $\min_i |x_i| \geq \epsilon$ for some $\epsilon > 0$. Denote by $|\Delta|$ the matrix $(|\Delta_{ij}|)$. Using $N_{ij} \geq 0$, the inequality $\|\mathrm{diag}[A]\|^2 = \sum_i A_i i^2 \leq \|A\|_{\mathrm{F}}^2$ which holds for any square matrix A, and that $\max_i |x_i| \leq 1$, combined with (4), we get

$$\left\|\mathrm{diag}\left[ND_x(xx^T - ss^T)\right]\right\| = \|\mathrm{diag}[ND_x\Delta]\| \leq \|N|\Delta|\,\|_{\mathrm{F}} \leq \|N\|_{\mathrm{F}}\,\|\Delta\|_{\mathrm{F}}\,. \tag{9}$$

Let

$$\pi(N) = \inf\{\left\|\mathrm{diag}\left[ND_y(yy^T - ss^T)\right]\right\| \,:\, y \notin B_\delta,\, \min_i |y_i| \geq \epsilon\}\,.$$

Call N admissible if it is positive integer valued, irreducible and non-bipartite and let $\mathcal{N}$ denote the set of such matrices. It follows that $x \in B_\delta$ when $\|\Delta\|_{\mathrm{F}} \leq c := \inf_{N\in\mathcal{N}} \pi(N)/\,\|N\|_{\mathrm{F}}$. Note that $c > 0$.

Now, since $x \in B_\delta$ holds, Corollary 2, (4) and (9) together give that

$$\|x - s\| \leq \frac{\|\nabla f(x)\|}{s_{\min}^2\mu} \leq \frac{\|\mathrm{diag}[ND_x\Delta]\|}{s_{\min}^2\mu} \leq \frac{\|N\|_{\mathrm{F}}\,\|\Delta\|_{\mathrm{F}}}{s_{\min}^2\mu}\,,$$

finishing the proof.

E. Additional Experiments and Details

E.1. Multi-Class

For the multiclass data-sets (i.e., Dogs and Web) we run our algorithm with two different approaches. One is to use one-vs-rest strategy for each class $k \in \mathcal{K}$ by assuming class-independent models determine the probability of the worker flipping the ground truth. Another method is to assume homogeneous Dawid-Skene model instead of the single coin model, which construct a confusion matrix for each worker with only a single parameter. More specifically, the diagonal elements of the confusion matrix will be identical for each worker and the off-diagonal elements of each row will be the same. After representing labels as vectors containing a 1 in the column of the class index they represented, we minimize a weighted least-squares objective $\frac{1}{2}\sum_{(i,j)\in E} N_{ij}[\tilde{C}_{ij} - |\mathcal{K}|x_ix_j - (|\mathcal{K}|-2)x_i - (|\mathcal{K}|-2)x_j]^2$ by PGD, where $\tilde{C}_{ij} = \frac{4(|\mathcal{K}|-1)}{T}\sum_{t=1}^{T}\langle Y_{i,t}, Y_{j,t}\rangle - |\mathcal{K}|$. For different data-sets, we choose the model that best fits the data. Then, in order to predict the label, a score function for class-conditional skill is calculated for each class k using $score(k) = \sum_{(i,t)\in A} \log\frac{1+s_i}{1-s_i}\mathbf{1}(Y_{i,t} = k)$, where $\mathbf{1}(\cdot)$ is a ± 1 indicator. The label is given by finding the class corresponding to the maximum of the score function.

E.2. Experiments for Different Skill-Distribution

We randomly assign binary classes to $T = 300$ tasks and select five pairs of parameters. Average prediction errors are presented in Table 3 averaged over 10 independent runs. Parameters $\alpha = 5, \beta = 1$, correspond to reliable workers leading to small prediction error; the prediction error with parameters $\alpha = 2, \beta = 2$ and $\alpha = 0.5, \alpha = 0.5$, is almost random because of $\sum_{i\in[W]} s_i$ is no longer positive, which validates our theory. Similar situation arises for $\alpha = 2, \beta = 5$ and $\alpha = 5, \beta = 1$, because the skills are all flipped relative to our assumption that the sum of the skills is positive.

E.3. Graph Size

We focus on how the graph size affects the performance of PGD algorithm. Note that graph size is associated with the number of workers. Our goal is to demonstrate that for a constant amount of noise, prediction accuracy of PGD does not

Number of workers	21	51	71	91
Bayes error	0.0425 ± 0.0042	0.0622 ± 0.0040	0.0634 ± 0.0033	0.0574 ± 0.0030
Prediction error (const. noise)	0.0425 ± 0.0042	0.0641 ± 0.0126	0.0662 ± 0.0072	0.0618 ± 0.0063

Table 4. Average prediction errors for different graph sizes.

Table 5. Prediction errors for different weightings

Worker type Assigned most tasks	$[N_{ij} > 0]$	$B(N_{ij}) = N_{ij}$	$B(N_{ij}) = N_{ij}^2$
Spammers	0.33 ± 0.03	0.33 ± 0.03	0.55 ± 0.17
Positive skill workers	0.17 ± 0.06	0.09 ± 0.02	0.09 ± 0.02

degrade with graph-size. We again consider the case when the worker-interaction graph is a star-graph with an odd-cycle of length 3. We increase the size of worker-interaction graph by adding nodes to the star-graph. Skills s are selected between 0.8 and -0.3 uniformly. To fix the noise level, we define $C_{ij} = s_i s_j + \xi_{ij}, \forall (i,j) \in E$ where ξ_{ij} is randomly selected from $[-0.2, 0.2]$. Note that the noise level is quite large relative to what we expect in terms of accuracy of correlation estimates. We iteratively run PGD for 50 times. The average prediction errors with different graph size it presented in Table 4. It can be seen that the prediction error is not sensitive to the graph size compared to the Bayes error.

E.4. Weighting function:

It is straighforward from our proof of Theorem 2 to see that PGD algorithm converges to the global optimal for any non-negative weights. Our objective is based on weighting with number of counts in Eq. 3. However, there are other options that one could consider. (Dalvi et al., 2013) has suggested using $B(N_{ij}) = N_{ij}^2$, while we use N_{ij}. Another possibility is to use binary weights. We iteratively run PGD 10 times for each weighing function with $T = 300$ tasks for different types of task assignements. If N_{ij}'s are all equal, these choices produce identical results. We consider two cases: (a) Spammers are assigned a majority of tasks; (b) Positively skilled workers are assigned most tasks. The prediction errors are compared in Table 5. Note that quadratic weighting is quite bad in this case because it tends to ignore positively skilled workers. On the other hand unweighted case does not accurately estimate spammers and also results in poor choice.